Two-dimensional Proximal Constraints with Group Lasso for Disease Progression Prediction

Hao-en Sung, Mi-yen Yeh, and Shou-de Lin

Abstract—Despite the advance in medical research, there are still some diseases that cannot be cured after certain level of severity. For instance, Alzheimer's disease (AD), arguably the best known neurodegenerative disease, attracts significant attention because of its irreversible disease progression. Enormous capital and efforts have been invested into AD research for the pursuit of better prediction on its progression. While there are multiple ways to solve disease progression problem, multitask learning is the predominant strategy due to its strength of sharing partial information across time points to compensate the effect of insufficient data. In this paper, we extend algorithms for disease progression prediction from equipping one-dimensional proximal constraints to two-dimensional ones along with an assumption on the smoothness of features across time points. Our model achieves significant improvement in AD progression prediction on the basis of multitask learning model comparing with the previous works. Beside the empirical improvement, in this paper we provide theoretical analysis to show that the proposed two-dimensional proximal constraints maintains the feasibility of the decomposition procedure on derived optimization. We not only solve L2 norm models optimally, but also provide an approximation solution for models with L1 norm. Our implementation extends MALSAR library package, which is used specifically for multitask learning, and is available online.

Index Terms—Two-dimensional Proximal Constraints, Multitask Learning, Irreversible Disease, Alzheimer's Disease, Disease Progression Prediction, Fused Lasso, Group Lasso.

1 Background and Introduction

Among the documented diseases in twenty-first centuries, irreversible dementia such as Alzheimer's disease, Vascular dementia, and Lewy Body dementia have received great attention around the world [1]. Alzheimer's disease (AD), as an example, is a serious chronic neurodegenerative disease for the elders which accounts for the majority cases of dementia [2]. Some obvious symptoms include short-term memory loss, disorientation, and misbehavior. Though significant efforts have been invested in Alzheimer's research, the general cause of it is so far not well-understood.

Because of its irreversible disease progression and growing number of sufferers, Alzheimer's has received much more attention these years. Based on [3], AD resulted in roughly $486,000$ deaths in 2010, and there were about $48,000,000$ people suffered from it in 2015 according to a World Health Organization report [4]. Another rigorous research for U.S. shows that the number of Alzheimer's patients will increase from 5.1 million in 2015 to 13.5 million in 2050, while the total costs of care will escalate from 226 billion to 1101 billion, which is becoming a huge burden to the U.S. government [5].

To better understand the key factors to AD's progression and potential treatment, a worldwide project held by Alzheimer's Disease Neuroimaging Initiative (ADNI) was launched in October, 2004. In their first-stage experiment, various brain-imaging technologies, including Positron Emission Tomography (PET) and Magnetic Resonance Imaging (MRI), are used in attempt to detect Alzheimer in its early stage. Beside imaging features, a series of clinical tests, such as Cerebrospinal Fluid (CSF), neuropsychological tests, and multiple clinical questionnaires are also conducted to track the progression of AD. Other information collected includes numerous assessments to dementia, such as Mini Mental State Examination (MMSE) and Alzheimer's Disease Assessment Scale - Cognitive Subscale (ADAS-Cog).

Utilizing the above information, a variety of researches have been conducted in recent years not only on AD progression diagnosis [6], [7], [8], [9] but also in learning the hidden patterns of disease progression and predicting patients' future cognitive condition [10], [11], [12], [13], [14], so that medical staffs can provide much more accurate treatment and corresponding support. One prevalent way to model disease progression is to regard learning patient image features in the beginning $X^{(0)}$ and patient clinical scores at different time points $Y^{(j)}$ as separate tasks, as depicted in Fig. 1, so that they can be learned jointly using a multitask algorithm. There are several advantages to apply multi-task learning techniques to this problem. First of all, it solves the imbalanced instance number issue within different time

- *Hao-en Sung was with the Institute of Information Science, Academia Sinica, Taiwan. E-mail: wrangle1005@gmail.com*
- *Mi-yen Yeh was with the Institute of Information Science, Academia Sinica, Taiwan. E-mail: miyen@iis.sinica.edu.tw*
- *Shou-de Lin was with the Department of Computer Science and Information Engineering, National Taiwan University, Taiwan. E-mail: sdlin@csie.ntu.edu.tw*

Our implementation, dataset preprocessing, and experiment details are available in GitHub repository. (link)

points to prevent bias. Secondly, it guarantees a general and effective biomarkers (bio-features) across time points. Last but not the least, what is learned for each task can improve the performance of other tasks, hence, a potentially more accurate model can be learned. However, not all multi-task learning models are suitable for solving this task. We will then introduce two previous works [15], [16] arguing that a suitable multi-task learning model shall consider the smooth but non-strictly deteriorating disease condition. The main difference between these two works is that the model proposed in [16] further pursues the model sparsity.

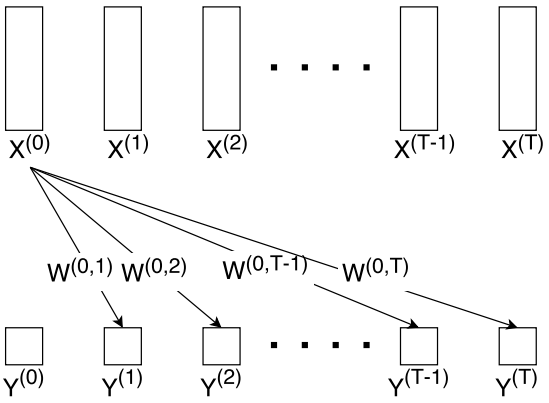


Fig. 1. Multi-task learning framework

Temporal Group Lasso (TGL) algorithm has been proposed in [15], where a novel constraint term is introduced as the regularization with the hypothesis that the clinical scores for each patient are non-strictly decreasing to the time factor. With this intuitive hypothesis, it models each weighing matrix learned from measurements at time 0, i.e. $x^{(0)}$, to clinical scores across different time points, i.e. $y^{(1)}, ..., y^{(T)}$, with squared punishment on the differences between neighboring matrices. To be formal, under the assumption that $y^{(j)} \approx y^{(j+1)}$, one can easily derive the formula $x^{(0)} \cdot w^{(j)} \approx x^{(0)} \cdot w^{(j+1)}$ with linear model. Then, $\|w^{(j)} - w^{(j+1)}\|_2^2$ is added as one-dimensional constraint during learning.

Later, an enhanced optimization formula in [16] is proposed, named Convex Fused Sparse Group Lasso (cFSGL). The main difference between these two works is that the latter substitutes L2-norms in the previous work with L1-norms for the pursuit of model sparsity. In formal, $\|w^{(j)} - w^{(j+1)}\|_1$ is instead added as one-dimensional constraint during learning.

The effectiveness of introducing constraints on proximal clinical scores in either L2-norm or L1-norm are further discussed in [17]. It is important to note that, these two models do not use all the biomarkers information after time 0, which can lead to inferior outcomes, especially when time further progresses. Considering a time-series problem with long sequence, it is crucial to re-train a model considering the latest information to avoid biased learning performance caused by hidden state transition. Possible solutions to this issue include various sliding window techniques and probabilistic-based state transformation like Markov related models [18].

In this paper, we use the sliding window technique to take into account biomarkers information across different time slots, and thus we transform the original multitask learning formulas from one-dimensional, i.e. 1D-TGL, 1D-cFSGL, to two-dimensional, i.e. 2D-TGL, 2D-cFSGL. By doing so, we address the aforementioned issue caused by using only the original features.

However, as will be shown in our experiments, directly extending 1D solution to 2D does not provide significant improvement. Theorectically, model cannot effectively learn from further information without some regularizations. In view of this, we propose to add a set of novel constraints between weighing matrices, which are learned from features across time points to a specific timing target score, as regularization terms, denoted as 2D-TGL$^+$ and 2D-cFSGL$^+$. In other words, our model needs to consider two-dimensional constraint during learning.

In this work, we mainly contribute in solving 2D-TGL$^+$ optimally and optimizing 2D-cFSGL$^+$ with introduction to *Pathwise Coordinate Optimization for Two-dimensional Fused Lasso* proposed in [19]. The general optimization procedure for both new derived algorithms is 1) use *Proximal Gradient Method (P.G.M.)* to decompose target formula into differential part and non-differential part, 2) update differential part with normal gradient descent, 3) reduce the non-differential part into multiple cases with analytical solutions or easier sub-problems, and 4) iteratively update differential and non-differential parts until convergence. All detailed mathematical proofs will be covered in later sections.

Contents of this work are arranged as follows. In Section 2, we deliver the formal notations to connect our proposed solution to the previous ones and elaborate on our framework with detailed derivations. Step-by-step optimization procedure is also demonstrated there. Experiment settings and results are shown in Section 3. We make our conclusion in Section 4. For more details, such as algorithm implementations, please refer to our GitHub repository.

2 Methodology

We deliver the problem definition in Section 2.1, the review of common norms in Section 2.2, and the definition of sign operators in Section 2.3. In Section 2.4, we will introduce the optimization formula for TGL, which is proposed in [15], in detail and describe how we make extensions. Similarly, in Section 2.5, we show how we apply our ideas to cFSGL, which is proposed in [16], and methods used to solve the extended problems.

2.1 Problem Definition

Assume there are total of $T+1$ time points starting from 0 to T with features $X^{(t)} \in \mathbb{R}^{n \times d}$ and target scores $Y^{(t)} \in \mathbb{R}^{n \times 1}$, where n is the number of instances and d is the dimension of features. We adopt the multitask learning framework here by regarding different time points as different tasks to obtain a general sets of biomarkers to improve the model performance. We use weighing matrix $W^{(i,j)} \in \mathbb{R}^{d \times 1}$ which is learned from $X^{(i)}$ to $Y^{(j)}$, where $0 \leqslant i < j \leqslant T$. For the conciseness of paper, we denote consecutive selection with

a colon. For example, $W^{(i,:)}$ selects all the weighing matrices learned from $X^{(i)}$ and $W^{(:,j)}$ selects all the weighing matrices learned for $Y^{(j)}$.

2.2 Notations: the Definition of Norm

We use the following notations for norms.

2.2.1 Frobenius Norm (L2-norm)

$$\|A\|_F = \sqrt{\sum_{i,j} A_{i,j}^2}$$

It can be written as $\|A\|_2$ when A is a vector.

2.2.2 Lasso Norm (L1-norm)

$$\|A\|_1 = \sum_{i,j} |A_{i,j}|$$

Apart from L2-norm, this norm encourages each variable moving toward zero, which improves the model sparsity.

2.2.3 L2,1-norm

$$\|A\|_{2,1} = \sum_i \sqrt{\sum_j A_{i,j}^2}$$

It is a special norm frequently used in multitask learning. Intuitively, it first calculates L2-norm for each row of the matrix and then sums up those non-negative values just like L1-norm. Because of the L1-norm sparse property, adding this term helps the learning model select a subset of general features across all time points, and achieves the purpose of feature selection and dimension reduction.

2.3 Notations: the Definition of Sign Operator

For the convenience of following proofs, we declare two kinds of sign operator definitions here.

2.3.1 Normal Sign Operator

$$\operatorname{sgn}(x) = \begin{cases} 1 & \text{if } x > 0, \\ 0 & \text{if } x = 0, \\ -1 & \text{if } x < 0. \end{cases}$$

General definition of the sign operator.

2.3.2 Extended Sign Operator

$$\operatorname{SGN}(x) = \begin{cases} \{1\} & \text{if } x > 0, \\ [-1, 1] & \text{if } x = 0, \\ \{-1\} & \text{if } x < 0. \end{cases}$$

To handle the non-smoothness issue at zero, we introduce another way to define the sign operator: the output of sign operator becomes a set of collections. Especially, it produces a set of values between -1 and 1 at 0.

2.4 Extension from 1D-TGL to 2D-TGL and 2D-TGL+

We describe TGL extensions and their formula updates in Section 2.4.1. and Section 2.4.2, respectively. The update procedure for 2D-TGL+ requires the application of *Proximal Gradient Methods* to learn differential and non-differential parts in two stages, as shown in Fig. 2. A brief summary to conclude TGL derivations is written in Section 2.4.3.

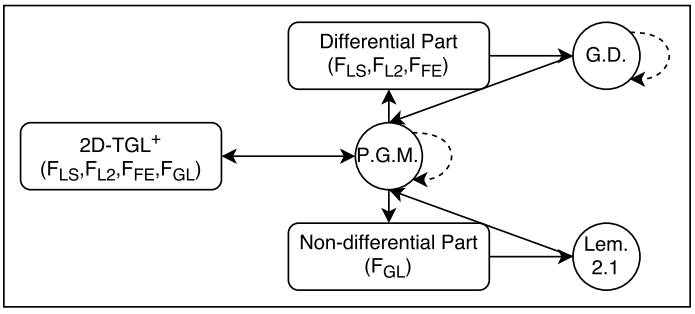


Fig. 2. Framework of 2D-TGL+ Derivations and Proofs. Rounded boxes are targets, while circles stand for lemmas and algorithms, such as Proximal Gradient Methods (P.G.M.) and Gradient Descent (G.D.). Solid lines indicate derivation directions; whereas Dotted lines show iterative updates.

2.4.1 Objective Function

According to [15], 1D-TGL uses only the information at time 0, i.e. $X^{(0)}$, and is written as Eq. 1.

$$\min_{W^{(0,j)}} \frac{1}{2} \sum_j \|X^{(0)} \cdot W^{(0,j)} - Y^{(j)}\|_2^2 + \lambda_1 \|W^{(0,:)}\|_2^2 + \lambda_2 \|W^{(0,:)} \cdot R^{(0)}\|_2^2 + \lambda_3 \|W^{(0,:)}\|_{2,1}, \quad (1)$$

where $R^{(0)} \in \mathbb{R}^{T \times (T-1)}$ is the constraint matrix for proximal clinical scores as shown in Fig. 3, with $R_{i,j}^{(0)} = +1$ if $i = j$, $R_{i,j}^{(0)} = -1$ if $i = j + 1$, or $R_{i,j}^{(0)} = 0$ otherwise.

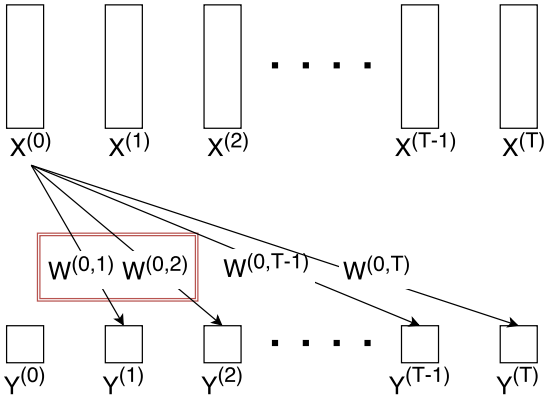


Fig. 3. Examples of constraints on proximal clinical scores. Under the assumption that $Y^{(j)} \approx Y^{(j+1)}$, with fixed $X^{(i)}$, one can derive that $W^{(i,j)} \approx W^{(i,j+1)}$. With this hypothesis, the squared difference between neighboring weighing matrices, as the red bounding box in the figure, is used as regularization.

To make use of the information provides by afterward features, we extend Eq. 1 to two-dimensional, i.e. 2D-TGL, when considering biomarker features later in time. The formula is rewritten as Eq. 2.

$$\min_{W^{(i,j)}} \frac{1}{2}\sum_{i,j}\|X^{(i)}\cdot W^{(i,j)} - Y^{(j)}\|_2^2 + \lambda_1\|W\|_F^2 + \lambda_2\|\sum_i W^{(i,:)}\cdot R^{(i)}\|_F^2 + \lambda_3\|W\|_{2,1}, \quad (2)$$

where

$$R^{(0)} = \begin{pmatrix} +1 & & & \\ -1 & +1 & & \\ & -1 & \ddots & \\ & & & +1 \\ & & & -1 \end{pmatrix},$$

$$R^{(1)} = \begin{pmatrix} 0 & & & \\ 0 & +1 & & \\ & -1 & \ddots & \\ & & & +1 \\ & & & -1 \end{pmatrix},$$

$$\vdots$$

In the multi-biomarker setting as Eq. 2, defined as 2D-TGL, we are curious about weighing matrices learned from different biomarkers to a single clinical score. Given identical target prediction, we have the formula $X^{(i)}\cdot W^{(i,j)} \approx X^{(i+1)}\cdot W^{(i+1,j)}$. Under the same assumption that Alzheimer's progression is irreversible, proximal biomarkers should have high similarity, i.e. $X^{(i)} \approx X^{(i+1)}$. It is then straightforward to derive $W^{(i,j)} \approx W^{(i+1,j)}$. Thus, another matrix S that adds constraints on proximal biomarkers, as shown in Fig. 4, are introduced to Eq. 2.

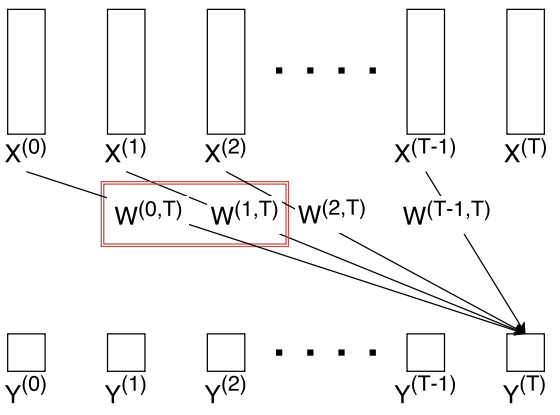


Fig. 4. Examples of constraints on proximal biomarkers. Under the assumption that $X^{(i)} \approx X^{(i+1)}$, with fixed $Y^{(j)}$, one can derive that $W^{(i,j)} \approx W^{(i+1,j)}$. With this hypothesis, the squared difference between neighboring weighing matrices, as the red bounding box in the figure, is used as regularization.

The objective function for model with new added constraint matrix S, i.e. 2D-TGL$^+$, is written as Eq. 3.

$$\min_{W^{(i,j)}} \frac{1}{2}\sum_{i,j}\|X^{(i)}\cdot W^{(i,j)} - Y^{(j)}\|_2^2 + \lambda_1\|W\|_F^2 + \lambda_2\left(\|\sum_i W^{(i,:)}\cdot R^{(i)}\|_F^2 + \|\sum_j W^{(:,j)}\cdot S^{(j)}\|_F^2\right) + \lambda_3\|W\|_{2,1}, \quad (3)$$

where

$$S^{(0)} = \begin{pmatrix} 0 & & & \\ 0 & 0 & & \\ & 0 & \ddots & \\ & & & 0 \\ & & & 0 \end{pmatrix},$$

$$S^{(1)} = \begin{pmatrix} +1 & & & \\ -1 & 0 & & \\ & 0 & \ddots & \\ & & & 0 \\ & & & 0 \end{pmatrix},$$

$$\vdots$$

To simplify the notations in the following derivations, the optimization formula of 2D-TGL$^+$ can be seen as a collection of four components, including squared loss function (F_{LS}), L2-norm regularization (F_{L2}), Fused L2-norm (Euclidean-norm) regularization (F_{FE}), and Group Lasso regularization (F_{GL}). They can be written as follows.

$$F_{LS} = \frac{1}{2}\sum_{i,j}\|X^{(i)}\cdot W^{(i,j)} - Y^{(j)}\|_2^2$$

$$F_{L2} = \lambda_1\|W\|_F^2$$

$$F_{FE} = \lambda_2\left(\|\sum_i W^{(i,:)}\cdot R^{(i)}\|_F^2 + \|\sum_j W^{(:,j)}\cdot S^{(j)}\|_F^2\right)$$

$$F_{GL} = \lambda_3\|W\|_{2,1},$$

Then, the goal becomes solving

$$\min_W F = F_{LS} + F_{L2} + F_{FE} + F_{GL}. \quad (4)$$

2.4.2 Optimization

The rewritten optimization formula, denoted Eq. 4, consists of differentiable parts (F_{LS}, F_{L2}, F_{FE}) and non-differentiable part (F_{GL}), which is proven solvable with *Proximal Gradient Method (P.G.M.)* [20], [21], [22], [23].

The optimal solution for the differentiable parts, i.e. W_D^*, is obtained by *Gradient Descent (G.D.)* iteratively updating $W^{(i,j)}$ as follows.

$$W^{(i,j)} \leftarrow (X^{(i)} \cdot W^{(i,j)} - Y^{(j)}) \cdot X^{(i)} + 2 \cdot \lambda_1 \cdot W^{(i,j)} + 2 \cdot \lambda_2 \Big((W^{(i,j)} - W^{(i,j-1)}) - (W^{(i,j+1)} - W^{(i,j)}) + (W^{(i,j)} - W^{(i-1,j)}) - (W^{(i+1,j)} - W^{(i,j)}) \Big). \quad (5)$$

After updating Eq. 5, the original problem can be transformed into finding the global optimal, i.e. W_G^*, for both differentiable and non-differentiable parts by solving

$$\min_W \frac{1}{2} \|W_G^* - W_D^*\|_F^2 + F_{GL}. \quad (6)$$

Because of the independence property between each feature, the transformed optimization formula Eq. 6 can be further simplified to

$$\min_w \frac{1}{2} \|w_G^* - w_D^*\|_F^2 + f_{GL}, \quad (7)$$

where lowercase notations indicate that we are now concerning about one specific feature.

Lemma 2.1. *Eq. 7 for solving w_G^* has an analytical solution as follows.*

$$w_G^* = \frac{\max(w_D^* - \lambda_3, 0)}{\|w_D^*\|_2} \cdot w_D^* \quad (8)$$

Proof. At optimal, it can be derived from Eq. 7:

$$w_G^* - w_D^* + \lambda_3 \cdot \frac{w_G^*}{\|w_G^*\|_2} = 0 \quad (9)$$

$$\left(1 + \frac{\lambda_3}{\|w_G^*\|_2}\right) \cdot w_G^* = w_D^* \quad (10)$$

$$\|w_G^*\|_2 + \lambda_3 = \|w_D^*\|_2. \quad (11)$$

Case 1. $\|w_D^*\| \leqslant \lambda_3$

We just let $w_G^* = 0$.

Case 2. $\|w_D^*\| > \lambda_3$

With Eq. 10 and Eq. 11, we can derive $w_G^* = \frac{w_D^* - \lambda_3}{\|w_D^*\|_2} \cdot w_D^*$.

Jointly consider Case 1 and Case 2, we further simplify the formula and obtain Eq. 8. □

2.4.3 Summary for TGL Derivations

We solve Eq. 7 with Lemma 2.1 and successively derive the update procedure for 2D-TGL and 2D-TGL$^+$.

2.5 Extension from 1D-cFSGL to 2D-cFSGL and 2D-cFSGL$^+$

We describe cFSGL extensions and their formula updates in Section 2.5.1 and Section 2.5.2, respectively. The update procedure for 2D-cFSGL$^+$ is much more challenging, as shown in Fig 5. First, we apply *Proximal Gradient Descent* to learn differential and non-differential parts separately. Later, Theorem 2.6 and Theorem 2.10 are needed to further simplify the update formulas, whose correctness are verified in Section 2.5.3 and Section 2.5.4. In Section 2.5.5, we solve two sub-problems to find out the optimal value for π^{L1} and to approximate π^{FL}. At the end, a summary to conclude the derivation of cFSGL extensions is written in Section 2.5.6.

2.5.1 Objective Function

Based on [16], cFSGL has a very similar objective function as TGL. The only difference between them is that some of the regularization terms are altered from L2-norm to L1-norm to achieve better model sparsity.

Just like the extension to TGL, the original optimization formula, denoted 1D-cFSGL, is a special case of the general multi-biomarker version, named 2D-cFSGL, which is written as Eq. 12, when fixing $i = 0$.

$$\min_{W^{(i,j)}} \frac{1}{2} \sum_{i,j} \|X^{(i)} \cdot W^{(i,j)} - Y^{(j)}\|_2^2 + \lambda_1 \|W\|_1 + \lambda_2 \|\sum_i W^{(i,:)} \cdot R^{(i)}\|_1 + \lambda_3 \|W\|_{2,1} \quad (12)$$

We are also interested in applying our ideas to Eq. 12 to pursue model sparsity. The extended 2D-cFSGL$^+$ algorithm is written as Eq. 13.

$$\min_{W^{(i,j)}} \frac{1}{2} \sum_{i,j} \|X^{(i)} \cdot W^{(i,j)} - Y^{(j)}\|_2^2 + \lambda_1 \|W\|_1 + \lambda_2 \left(\|\sum_i W^{(i,:)} \cdot R^{(i)}\|_1 + \|\sum_j W^{(:,j)} \cdot S^{(j)}\|_1 \right) + \lambda_3 \|W\|_{2,1}. \quad (13)$$

To simplify the notations used in following derivations, we further define four components in the 2D-cFSGL$^+$ algorithm, including squared loss function (F_{LS}), L1-norm regularization (F_{L1}), Fused Lasso regularization (F_{FL}), and Group Lasso regularization (F_{GL}). They can be written as follows.

$$F_{LS} = \frac{1}{2} \sum_{i,j} \|X^{(i)} \cdot W^{(i,j)} - Y^{(j)}\|_2^2$$

$$F_{L1} = \lambda_1 \|W\|_1$$

$$F_{FL} = \lambda_2 \left(\|\sum_i W^{(i,:)} \cdot R^{(i)}\|_1 + \|\sum_j W^{(:,j)} \cdot S^{(j)}\|_1 \right)$$

$$F_{GL} = \lambda_3 \|W\|_{2,1},$$

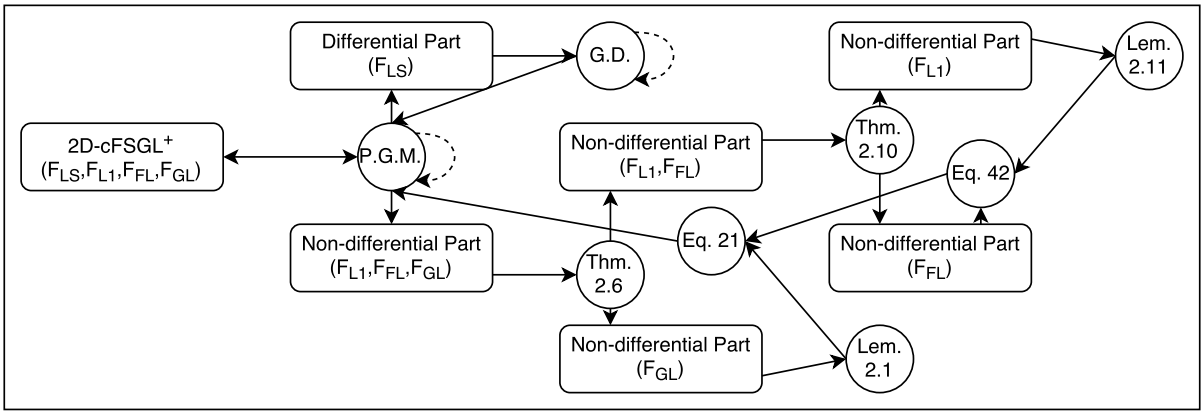


Fig. 5. Framework of 2D-cFSGL+ Derivations and Proofs. Rounded boxes are targets, while circles stand for theorems, lemmas, equations, and algorithms, such as Proximal Gradient Methods (P.G.M.) and Gradient Descent (G.D.). Solid lines indicate derivation directions; whereas dotted lines show iterative updates.

and the goal becomes solving

$$\min_W F = F_{LS} + F_{L1} + F_{FL} + F_{GL}. \tag{14}$$

2.5.2 Optimization

The update of Eq. 14 can be partitioned into differentiable (F_{LS}) and non-differentiable (F_{L1}, F_{FL}, F_{GL}) parts. We can see that the optimal W_D^* of the differentiable part can be updated as Eq. 15.

$$W^{(i,j)} \leftarrow (X^{(i)} \cdot W^{(i,j)} - Y^{(j)}) \cdot X^{(i)}. \tag{15}$$

After finding the optimal solution of the differentiable part W_D^* in Eq. 15, we can obtain the global optimal W_G^* with optimization formula

$$\min_W \frac{1}{2}\|W_G^* - W_D^*\|_F^2 + F_{L1} + F_{FL} + F_{GL}. \tag{16}$$

Considering the independence property between each feature, Eq. 16 is also reducible from uppercase notations to lowercase ones, which indicates that we are now focusing on one specific feature. Thus, we obtain Eq. 17 as follows.

$$\min_w \frac{1}{2}\|w_G^* - w_D^*\|_F^2 + f_{L1} + f_{FL} + f_{GL} \tag{17}$$

However, Eq. 17 is far more challenging to solve because of the existance of three non-smooth terms. Next we will solve it step-by-step with detailed proofs, including the decomposition of $f_{L1} + f_{FL}$ and f_{GL} as well as the decomposition of f_{L1} and f_{FL}.

2.5.3 Decomposition of $f_{L1} + f_{FL}$ and f_{GL}

This proof mainly follows the idea in [16].

Let $\pi(v)$ be defined as

$$\arg\min_w \frac{1}{2}\|w - v\|_F^2 + f_{L1} + f_{FL} + f_{GL}. \tag{18}$$

Consider two solutions that $\lambda_1 = \lambda_2 = 0$ and $\lambda_3 = 0$, i.e. $\pi^{GL}(v)$ and $\pi^{L1+FL}(v)$, respectively.

$$\pi^{L1+FL}(v) = \arg\min_w \frac{1}{2}\|w - v\|_F^2 + f_{L1} + f_{FL}, \tag{19}$$

$$\pi^{GL}(v) = \arg\min_w \frac{1}{2}\|w - v\|_F^2 + f_{GL}. \tag{20}$$

With Theorem 2.6 stated later in this section, we are able to prove the following equation holds.

$$\pi(v) = \pi^{GL}(\pi^{L1+FL}(v)). \tag{21}$$

That is to say, the original optimization problem, i.e. Eq. 17, can be solved in two separated steps without losing the optimal property.

To start with, one can derive the optimal conditions from Eq. 18-20 to Eq. 22-24 when fixing their gradients to 0.

$$\begin{aligned} 0 \in\ & \pi(v) - v + \lambda_1 \operatorname{SGN}(\pi(v)) + \\ & \lambda_2 \left(\sum_i \operatorname{SGN}(\pi(v)^{(i,:)} \cdot R^{(i)}) \cdot R^{(i)\intercal} + \right. \\ & \left. \operatorname{SGN}(\pi(v)^{(:,j)} \cdot S^{(j)}) \cdot S^{(j)\intercal} \right) + \\ & \lambda_3 \partial g(\pi(v)) \end{aligned} \tag{22}$$

$$\begin{aligned} 0 \in\ & \pi^{L1+FL}(v) - v + \lambda_1 \operatorname{SGN}(\pi^{L1+FL}(v)) + \\ & \lambda_2 \left(\sum_i \operatorname{SGN}(\pi^{L1+FL}(v)^{(i,:)} \cdot R^{(i)}) \cdot R^{(i)\intercal} + \right. \\ & \left. \operatorname{SGN}(\pi^{L1+FL}(v)^{(:,j)} \cdot S^{(j)}) \cdot S^{(j)\intercal} \right) \end{aligned} \tag{23}$$

$$\begin{aligned} 0 \in\ & \pi^{GL}(\pi^{L1+FL}(v)) - \pi^{L1+FL}(v) + \\ & \lambda_3 \partial g \left(\pi^{GL}(\pi^{L1+FL}(v)) \right), \end{aligned} \tag{24}$$

where

$$\partial g(y) = \begin{cases} \frac{y}{\|y\|_2}, \text{ if } y \neq 0, \\ y : \|y\|_2 \leqslant 1, \text{ if } y = 0. \end{cases}$$

Since our goal is to prove Eq. 21, we are interested in the relation between $\pi^{L1+FL}(v)$ and $\pi^{GL}(\pi^{L1+FL}(v))$, which will be later described in Lemma 2.4 and Lemma 2.5. Before that, we are also curious about the two possible relations between $\|\pi^{L1+FL}(v)\|_2$ and λ_3 , as shown in Lemma 2.2 and Lemma 2.3 below.

Lemma 2.2. *If $\|\pi^{L1+FL}(v)\|_2 \leqslant \lambda_3$, then one can derive $\pi^{GL}(\pi^{L1+FL}(v)) = 0$.*

Proof. Consider $\pi^{GL}(\pi^{L1+FL}(v)) \neq 0$, from Eq. 23 we know that there exists v at optimal that

$$\pi^{L1+FL}(v) = \pi^{GL}(\pi^{L1+FL}(v)) \cdot \left(1 + \frac{\lambda_3}{\|\pi^{GL}(\pi^{L1+FL}(v))\|_2}\right)$$

$$\|\pi^{L1+FL}(v)\|_2 = \|\pi^{GL}(\pi^{L1+FL}(v))\|_2 + \lambda_3.$$

Thus,

$$\because \|\pi^{GL}(\pi^{L1+FL}(v))\|_2 > 0$$

$$\therefore \|\pi^{L1+FL}(v)\|_2 > \lambda_3.$$

□

Lemma 2.3. *If $\|\pi^{L1+FL}(v)\|_2 > \lambda_3$, then $\pi^{GL}(\pi^{L1+FL}(v)) = \pi^{L1+FL}(v) \cdot \frac{\|\pi^{L1+FL}(v)\|_2 - \lambda_3}{\|\pi^{L1+FL}(v)\|_2}$.*

Proof.

Case 1. $\pi^{GL}(\pi^{L1+FL}(v)) \neq 0$

Following Eq. 24, we have

$$\|\pi^{L1+FL}(v)\|_2 = \|\pi^{GL}(\pi^{L1+FL}(v))\|_2 + \lambda_3,$$

which can be equivalently rewrite as

$$\pi^{L1+FL}(v) \cdot \frac{\|\pi^{L1+FL}(v)\|_2 - \lambda_3}{\|\pi^{L1+FL}(v)\|_2}.$$

Case 2. $\pi^{GL}(\pi^{L1+FL}(v)) = 0$

From Eq. 24, we know that

$$\pi^{L1+FL}(v) = \pi^{GL}(\pi^{L1+FL}(v)) + \lambda_3 \cdot y, \|y\|_2 \leqslant 1$$

$$\pi^{L1+FL}(v) = \lambda_3 \cdot y, \|y\|_2 \leqslant 1$$

$$\pi^{L1+FL}(v) \leqslant \lambda_3,$$

which violates the preposition of Lemma 2.3. In other words, Case 2 is not a valid vase.

□

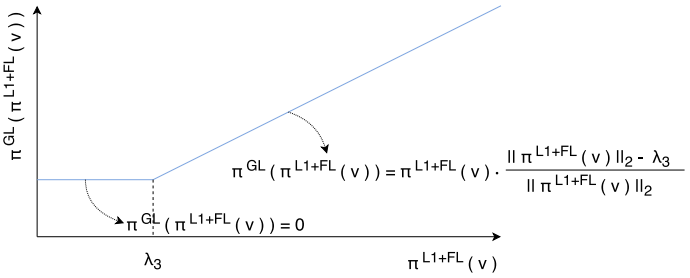


Fig. 6. Diagram for relationship between $\pi^{L1+FL}(v)$ and $\pi^{GL}(\pi^{L1+FL}(v))$. Proved in Lemma 2.2 and Lemma 2.3.

Lemma 2.4. *With Lemma 2.2 and 2.3, we have*

$$\text{SGN}\left(\pi^{L1+FL}(v)\right) \subseteq \text{SGN}\left(\pi^{GL}(\pi^{L1+FL}(v))\right) \tag{25}$$

Proof. From Fig. 6, we know that either $\|\pi^{L1+FL}(v)\|_2$ is larger than λ_3 or not, $\pi^{L1+FL}(v)$ and $\pi^{GL}(\pi^{L1+FL}(v))$ have same sign indicators. Thus, Eq. 25 holds. □

Lemma 2.5. *With Lemma 2.2 and 2.3, we also obtain*

$$\text{SGN}\left(\pi^{L1+FL}(v)^{(i,:)} \cdot R^{(i)}\right) \subseteq \text{SGN}\left(\pi^{GL}(\pi^{L1+FL}(v))^{(i,:)} \cdot R^{(i)}\right) \tag{26}$$

$$\text{SGN}\left(\pi^{L1+FL}(v)^{(:,j)} \cdot S^{(j)}\right) \subseteq \text{SGN}\left(\pi^{GL}(\pi^{L1+FL}(v))^{(:,j)} \cdot S^{(j)}\right) \tag{27}$$

.

Proof. Similar to the proof of Lemma 2.4, one can tell that pairwise subtraction in-between $\pi^{L1+FL}(v)$ and $\pi^{GL}(\pi^{L1+FL}(v))$ have same sign indicators. To be formal, if extending $R^{(i)}$ in Eq 26 as an example, then we have

$$SGN\left(\pi^{L1+FL}(v)^{(i+1,j)} - \pi^{L1+FL}(v)^{(i,j)}\right) \in SGN\left(\pi^{GL}(\pi^{L1+FL}(v)^{(i+1,j)}) - \pi^{GL}(\pi^{L1+FL}(v)^{(i,j)})\right).$$

We therefore obtain Eq. 26 and Eq. 27. □

Theorem 2.6. *$w_G^* = \pi^{GL}(\pi^{L1+FL}(v))$, where w_G^* is the global solution for both differentiable and non-differentiable parts appearing in Eq. 17.*

Proof. According to Eq. 23, Eq. 24, Lemma 2.4, and Lemma 2.5, when the optimal of $\pi^{L1+FL}(v)$ and $\pi^{GL}(v)$ occurs, we have the following formula .

$$0 \in \pi^{GL}(\pi^{L1+FL}(v)) - v + \lambda_1 \,\mathrm{SGN}\left(\pi^{GL}(\pi^{L1+FL}(v))\right) + \lambda_2 \left(\sum_i \mathrm{SGN}\left(\pi^{GL}(\pi^{L1+FL}(v))^{(i,:)} \cdot R^{(i)}\right) \cdot R^{(i)\intercal} + \sum_j \mathrm{SGN}\left(\pi^{GL}(\pi^{L1+FL}(v))^{(:,j)} \cdot S^{(j)}\right) \cdot S^{(j)\intercal}\right) + \lambda_3 \partial g\left(\pi^{GL}(\pi^{L1+FL}(v))\right) \quad (28)$$

Since there is only one solution for Eq. 18, from Eq. 22 and Eq. 28, we prove that Eq. 18 can be solved in two-steps as Eq. 19 and Eq. 20 that $\pi(v) = \pi^{GL}(\pi^{L1+FL}(v))$. □

So far, we prove the feasibility of decomposing Eq. 18 into two separated parts: $f_{L1} + f_{FL}$ and f_{GL}. We also acknowledge that Eq. 20 can be solved with Lemma 2.1, which is similar to Eq. 8 in Section 2.4; whereas, to obtain a solution to Eq. 19 needs a few more steps, which is derived in Section 2.5.4.

2.5.4 Decomposition of f_{L1} and f_{FL}

This proof mainly follows the idea in [24].

Our goal in this section is to optimize

$$\pi^{L1+FL}(v) = \arg\min_w \frac{1}{2}\|w - v\|_F^2 + f_{L1} + f_{FL} \quad (29)$$

within two steps.

Consider two solutions when $\lambda_1 = 0$ and $\lambda_2 = 0$, i.e. $\pi^{FL}(v)$ and $\pi^{L1}(v)$, respectively. They can be written as

$$\pi^{L1}(v) = \arg\min_w \frac{1}{2}\|w - v\|_F^2 + f_{L1} \quad (30)$$

$$\pi^{FL}(v) = \arg\min_w \frac{1}{2}\|w - v\|_F^2 + f_{FL}. \quad (31)$$

With Theorem 2.10 described later in this work, we can prove that

$$\pi^{L1+FL}(v) = \mathrm{sgn}(\pi^{FL}(v)) \otimes \max(|\pi^{FL}(v)| - \lambda_1, 0).$$

Before the proof of Theorem 2.10, we are curious about the derivative of Eq. 29, which is shown as follows.

$$0 \in \pi^{L1+FL}(v) - v + \lambda_1 \,\mathrm{SGN}(w) + \lambda_2 \left(\sum_i \mathrm{SGN}\left(\pi^{L1+FL}(v)^{(i,:)} \cdot R^{(i)}\right) \cdot R^{(i)\intercal} + \sum_j \mathrm{SGN}\left(\pi^{L1+FL}(v)^{(:,j)} \cdot S^{(j)}\right) \cdot S^{(j)\intercal}\right) \quad (32)$$

Due to the optimality of $\pi^{FL}(v)$ in Eq. 31, there must exist z^*, which fulfills

$$z^{(i,:)*} \in \lambda_2 \,\mathrm{SGN}\left(\pi^{FL}(v)^{(i,:)} \cdot R^{(i)}\right) \quad (33)$$

$$z^{(:,j)*} \in \lambda_2 \,\mathrm{SGN}\left(\pi^{FL}(v)^{(:,j)} \cdot S^{(j)}\right), \quad (34)$$

such that

$$\pi^{FL}(v) = v - \left(\sum_i z^{(i,:)*} \cdot R^{(i)\intercal} + \sum_j z^{(:,j)*} \cdot S^{(j)\intercal}\right). \quad (35)$$

To prove the correctness of this theorem, three more lemmas are needed and listed below.

Lemma 2.7. *Let $\pi^{L1+FL}(v)$ and g become*

$$\pi^{L1+FL}(v) = \mathrm{sgn}(\pi^{FL}(v)) \otimes \max(|\pi^{FL}(v)| - \lambda_1, 0) \quad (36)$$

$$g = \mathrm{sgn}(\pi^{FL}(v)) \otimes \min(|\pi^{FL}(v)|, \lambda_1). \quad (37)$$

then

$$0 = \pi^{L1+FL}(v) - v + g + \sum_i z^{(i,:)*} \cdot R^{(i)\intercal} + \sum_j z^{(:,j)*} \cdot S^{(j)\intercal} \quad (38)$$

Proof. To prove this lemma, we need to consider all four possible situations for Eq. 36 and Eq. 37, as listed below:

Case 1. $\pi^{FL}(v) > 0$ *and* $|\pi^{FL}(v)| > \lambda_1$

$$\pi^{FL}(v) - \lambda_1 - v + \lambda_1 + \sum_i z^{(i,:)*} \cdot R^{(i)\intercal} + \sum_j z^{(:,j)*} \cdot S^{(j)\intercal} = 0$$

Case 2. $\pi^{FL}(v) > 0$ *and* $|\pi^{FL}(v)| \leqslant \lambda_1$

$$0 - v + \pi^{FL}(v) + \sum_i z^{(i,:)*} \cdot R^{(i)\intercal} + \sum_j z^{(:,j)*} \cdot S^{(j)\intercal} = 0$$

Case 3. $\pi^{FL}(v) \leqslant 0$ *and* $|\pi^{FL}(v)| > \lambda_1$

$$-(-\pi^{FL}(v) - \lambda_1) - v + (-\lambda_1) + \sum_i z^{(i,:)*} \cdot R^{(i)\intercal} + \sum_j z^{(:,j)*} \cdot S^{(j)\intercal} = 0$$

Case 4. $\pi^{FL}(v) \leqslant 0$ *and* $|\pi^{FL}(v)| \leqslant \lambda_1$

$$0 - v + (-(-\pi^{FL}(v))) + \sum_i z^{(i,:)^*} \cdot R^{(i)^\intercal} + \sum_j z^{(:,j)^*} \cdot S^{(j)^\intercal} = 0$$

Consider all above four cases, Eq. 38 is proved. □

Lemma 2.8. *Given* $\pi^{L1+FL}(v)$ *and* g*, we have*

$$g \in \lambda_1 \operatorname{SGN}(\pi^{L1+FL}(v)). \tag{39}$$

Proof. Similarly, there are four cases for Eq. 36 and Eq. 37 needed to be discussed.

Case 1. $\pi^{FL}(v) > 0$ *and* $|\pi^{FL}(v)| > \lambda_1$

$$\because \pi^{L1+FL}(v) > 0$$

$$\therefore g = \lambda_1 \in \lambda_1 \operatorname{SGN}(\pi^{L1+FL}(v)) = \lambda_1$$

Case 2. $\pi^{FL}(v) > 0$ *and* $|\pi^{FL}(v)| \leqslant \lambda_1$

$$\because g = \pi^{FL}(v) \quad and \quad -\lambda_1 \leqslant \pi^{FL}(v) \leqslant \lambda_1$$

$$\therefore g \in \lambda_1 \operatorname{SGN}(\pi^{L1+FL}(v)) = \lambda_1 \times [-1, 1]$$

Case 3. $\pi^{FL}(v) \leqslant 0$ *and* $|\pi^{FL}(v)| > \lambda_1$

$$\because \pi^{L1+FL}(v) \leqslant 0$$

$$\therefore g = -\lambda_1 \in \lambda_1 \operatorname{SGN}(\pi^{L1+FL}(v)) = -\lambda_1$$

Case 4. $\pi^{FL}(v) \leqslant 0$ *and* $|\pi^{FL}(v)| \leqslant \lambda_1$

$$\because g = -\pi^{FL}(v) \quad and \quad -\lambda_1 \leqslant \pi^{FL}(v) \leqslant \lambda_1$$

$$\therefore g \in \lambda_1 \operatorname{SGN}(\pi^{L1+FL}(v)) = \lambda_1 \times [-1, 1]$$

Consider all above four cases, Eq. 39 is proved. □

Lemma 2.9. *The optimal solutions* $z^{(i,:)^*}$ *and* $z^{(:,j)^*}$ *also qualified*

$$z^{(i,:)^*} \in \lambda_2 \operatorname{SGN}\left(\pi^{L1+FL}(v)^{(i,:)} \cdot R^{(i)}\right) \tag{40}$$

$$z^{(:,j)^*} \in \lambda_2 \operatorname{SGN}\left(\pi^{L1+FL}(v)^{(:,j)} \cdot S^{(j)}\right). \tag{41}$$

Proof. From Eq. 33 and Eq. 34, we know $z^{(i,:)^*}$ and $z^{(:,j)^*}$ can be obtained from some subtractions between proximal entries in π^{FL} through $\operatorname{SGN}(\cdot)$. Since the conversion from $\pi^{FL}(v)$ to $\pi^{L1+FL}(v)$ can be considered linearly as shown in Fig. 7, signed pairwise subtraction correspondingly in-between π^{FL} and π^{L1+FL} produces exactly the same result. □

Now, with three lemmas stated above, we can prove the two-stage computation with added matrices $S^{(j)}, \forall\, 1 \leqslant j \leqslant T$, is still feasible with Theorem 2.10.

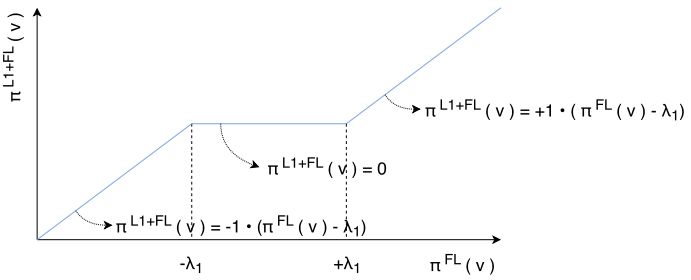


Fig. 7. Diagram for relationship between $\pi^{FL}(v)$ and $\pi^{L1+FL}(v)$. Proved in Theorem 2.10.

Theorem 2.10. *Given* $\lambda_1, \lambda_2 \geqslant 0$, $\pi^{L1+FL}(v)$ *can be solved in two stages as Eq. 30 and Eq. 31:*

$$\pi^{L1+FL}(v) = \operatorname{sgn}(\pi^{FL}(v)) \otimes \max(|\pi^{FL}(v)| - \lambda_1, 0). \tag{42}$$

Proof. From Lemma 2.7, we obtain Eq. 36. Now, together with Eq. 39 in Lemma 2.8 and Eq. 40, Eq. 41 in Lemma 2.9, they can be rewritten as

$$\begin{aligned} 0 = {} & \pi^{L1+FL}(v) - v + g + \sum_i z^{(i,:)^*} \cdot R^{(i)^\intercal} + \sum_j z^{(:,j)^*} \cdot S^{(j)^\intercal} \\ \in {} & \pi^{L1+FL}(v) - v + \lambda_1 \operatorname{SGN}(w) + \\ & \lambda_2 \left(\sum_i \operatorname{SGN}\left(\pi^{L1+FL}(v)^{(i,:)} \cdot R^{(i)}\right) \cdot R^{(i)^\intercal} + \right. \\ & \left. \sum_j \operatorname{SGN}\left(\pi^{L1+FL}(v)^{(:,j)} \cdot S^{(j)}\right) \cdot S^{(j)^\intercal} \right). \end{aligned}$$

In other words, the optimal solution for two-stage computation is the subset of optimal solution for $\pi^{L1+FL}(v)$. □

Until now, we not only prove the feasibility of decomposing 29 into f_{L1} and f_{FL}, but also finish all derivations of objective function decomposition for cFSGL extensions.

2.5.5 Optimization: Sub-problems

From now on, we are going to solve the final components of our models: the optimal value for $\pi^{L1}(v)$ and $\pi^{FL}(v)$.

Lemma 2.11. *Eq. 30 for solving* $\pi^{L1}(v)$ *has an analytical solution as follows.*

$$\pi^{L1}(v) = \operatorname{sgn}(v) \cdot \max(|v| - \lambda_1, 0) \tag{43}$$

Proof. At optimal, it can be derived from Eq. 30:

$$w - v + \lambda_1 \cdot \begin{bmatrix} \vdots \\ \cdots \operatorname{sgn}(w^{(i,j)}) \cdots \\ \vdots \end{bmatrix} = 0. \tag{44}$$

Since each entry in w is mutually independent, we can rewrite the formula as

$$w^{(i,j)} + \lambda_1 \cdot \text{sgn}(w^{(i,j)}) = v^{(i,j)}, \tag{45}$$

where $w^{(i,j)}$ and $v^{(i,j)}$ now are scalars.
Here we need to consider four possible cases for the relationship between $w^{(i,j)}$ and $v^{(i,j)}$.

Case 1. $w^{(i,j)} \geqslant 0$ *and* $v^{(i,j)} \geqslant \lambda_1$

$w^{(i,j)} = v^{(i,j)} - \lambda_1$

Case 2. $w^{(i,j)} \geqslant 0$ *and* $v^{(i,j)} < \lambda_1$

We simply let $w^{(i,j)} = 0$.

Case 3. $w^{(i,j)} < 0$ *and* $v^{(i,j)} \geqslant \lambda_1$

$w^{(i,j)} = v^{(i,j)} + \lambda_1$

Case 4. $w^{(i,j)} < 0$ *and* $v^{(i,j)} < \lambda_1$

We simply let $w^{(i,j)} = 0$.

When jointly consider Case 1 and Case 2, we obtain $w^{(i,j)} = \max(v^{(i,j)} - \lambda_1, 0)$; whereas Case 3 and Case 4 shows $w^{(i,j)} = \min(v^{(i,j)} + \lambda_1, 0)$. Thus, $w^{(i,j)} = \text{sgn}(v^{(i,j)} \cdot \max(|v^{(i,j)}| - \lambda_1, 0)$. □

Eq. 31 for finding $\pi^{FL}(v)$ is much more challenging to solve. An approximation algorithm was proposed in [19], called *Pathwise Coordinate Optimization for Two-dimensional Fused Lasso*. Its core procedures are summarized as pseudo code in Algorithm 1. The spirit of this algorithm is to gradually merge individual cells to reduce model cost while iterating through three sub-procedures *Descent Cycle*, *Fusion Cycle*, and *Smoothing Cycle*, until convergence. In the beginning, every cell in two-dimensional matrix is independent from each other. In *Descent Cycle*, the algorithm optimizes iteratively through each group of cells while keeping other cells fixed. After the convergence in *Descent Cycle*, it attempts to merge neighboring group of cells in *Fusion Cycle* if model cost reduction is feasible. *Smoothing Cycle* is designed to prevent learning model from being stuck at the local optimal.

As another key contribution of this paper, we have discovered the connection between the solution to one-dimensional problem and pathwise coordinate optimization algorithm for two-dimensional problem. With Theorem 2.10, we are finally able to solve Eq. 29.

2.5.6 Summary of cFSGL Derivations

To sum up the update procedure for the extension on cFSGL as stated in Eq. 14, we first apply *Proximal Gradient Methods (P.G.M.)* and feature reduction to obtain Eq. 17. A couple of two-step partitions on derived optimization formula are later applied as described in Theorem 2.6 and Theorem 2.10. Three separated sub-optimization formulas as stated in Eq. 19, Eq. 30, and Eq. 31 can be solved with Lemma 2.1, Lemma 2.11, and *Pathwise Coordinate Optimization for Two-dimensional Fused Lasso*, respectively.

Input : $\lambda_2; v_p = v^{(i,j)}, R^{(i)}, S^{(j)}, \ \forall\, 0 \leqslant i < j \leqslant T$
Output: $w_p = w^{(i,j)}, \ \forall\, 0 \leqslant i < j \leqslant T$

Algorithm `2D Fused Lasso()`
- **Define** $d(p, p')$ as distance from p to p'
- **Define** $D(k, k') = \min_{\substack{p \in G_k \\ p' \in G'_k}} d(p, p')$
- **Define** $c_{k,k'} = \sum_{p \in G_k} \sum_{p' \in G'_k} I[d(p, p') = 1]$
- **Define** $N_k = |G_k|$
- **Define** $\overline{w_k} = \frac{1}{N_k} \sum_{p \in G_k} w_p$
- **Define** γ_k as the agreed value in G_k
- **Init** G: $w_p \in G_p$
- **while** $w^{(i,j)}$ *not converge* **do**
  - `Descent Cycle`
  - `Fusion Cycle`
  - `Smoothing Cycle`
- **end**

Procedure `Descent Cycle`
- Iteratively consider derivatives of γ_k only and fix all other parameters:

$$N_k \cdot (\overline{w_k} - \gamma_k) + \lambda_2 \cdot \sum_{D(k,k')=1} c_{k,k'} \cdot \text{sgn}(\gamma_k - \gamma'_k).$$

- Update γ_k if the model cost reduces.

Procedure `Fusion Cycle`
- Iteratively consider to merge G_k and G'_k, where $D(k, k') = 1$, and fix all other parameters:

$$N_k \cdot (\overline{w_k} - \gamma_k) + N'_k \cdot (\overline{w'_k} - \gamma'_k) + \lambda_2 \cdot \sum_{D(k,l)=1} c_{k,l} \cdot \text{sgn}(\gamma_k - \gamma_l) + \lambda_2 \cdot \sum_{D(k',l)=1} c_{k',l} \cdot \text{sgn}(\gamma'_k - \gamma_l).$$

- If model cost reduces by merging G_k and G'_k, we introduce a new group $G_m = G_k \bigcup G'_k$, with $N_m = N_k + N'_k$, $\overline{w_m} = (N_k \overline{w_k} + N'_k \overline{w'_k})/N_m$, and $c_{m,l} = c_{k,l} + c_{k',l}$.

Procedure `Smoothing Cycle`
- To prevent model from stucking at local optimal, increase λ_2 with slightly, i.e. $\lambda_2 \leftarrow \lambda_2 + \delta$.

Algorithm 1: Pathwise Coordinate Optimization for Two-dimensional Fused Lasso.

3 EXPERIMENT

In this section, we focus on examining two hypotheses for both extended Temporal Group Lasso and extended Convex Fused Sparse Group Lasso on datasets provided by Alzheimer's Disease Neuroimaging Initiative (ADNI) [25].

We focus on the following two null hypotheses, and we attempt to apply paired P-value tests to reject both of them.

Hypothesis I: straightforwardly applying 2D concept without adding the proposed constraints cannot bring significant improvement.

In this hypothesis, we will compare our algorithm performance between using 1D information and 2D information. The null hypothesis here is that the performance of 1D models and 2D models have no significant difference.

Hypothesis II: the 2D concept with the proposed constraints cannot bring significant improvement.

Different from *Hypothesis I*, we are now interested in the comparison between 2D models without regularization and 2D models with regularization. Here, the null hypothesis is about the performance of 2D models and 2D$^+$ models have no significant difference.

3.1 Experimental Setup

In the ADNI dataset, various measurements are available from three selected subjects: Alzheimer's disease patients (AD), mild cognitive impairment patients (MCI), and normal controls (NL). The measurements, including MRI scans, PET scans, CSF measurements, and cognitive scores like MMSE and ADAS-Cog, are collected over 6-month to 1-year interval. Our experiments use the extracted features from MRI scans (M) and distributed META features (E) to predict the cognitive scores MMSE and ADAS-Cog. The detailed features are listed in Table 1. There are totally six time points, i.e. SC, M06, M12, M24, M36, and M48. The number of instances and feature dimensions for MRI and MRI+META datasets can be found in Table 2 and Table 3, respectively.

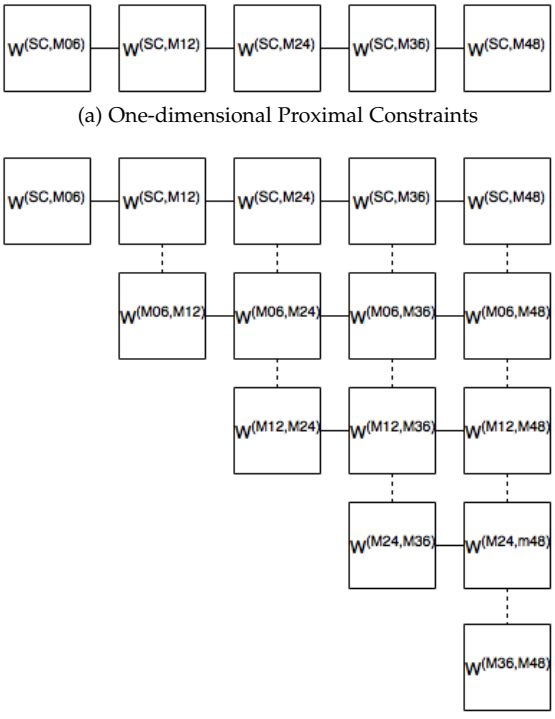
To examine the above-mentioned hypotheses, we consider all combinations between feature sets and target scores, i.e. (M, MMSE), (M+E, MMSE), (M, ADAS-Cog), and (M+E, ADAS-Cog) with two evaluation metrics — Correlated Coefficient (CC) and Root Mean Squared Error (RMSE) under five-fold cross-validation. The parameters for each competitor are grid-searched in the range of [1e-3, 1e+3]. For each experiment, performance results retrieved from five-fold cross-validations are averaged and repeated with five random seeds, while the score for each fold are weighed by the instance number at different time points in testing set. Later, five averaged results are calculated with average metric, variance metric and paired t-test metric between specific algorithms.

There are total of nine algorithms in our experiments. 1D-Ridge simply regards all information as one task and neglect the time-factor. 1D-Lasso and 2D-Lasso are baseline algorithms which omit proximal constraint matrix R. 1D-TGL and 1D-cFSGL are our main competitors proposed in [15], [16]. 2D-TGL and 2D-cFSGL use two-dimensional model framework without the derived constraint matrix S; whereas 2D-TGL$^+$ and 2D-cFSGL$^+$ equip all the proposed constraints.

From another aspect, for those 1D algorithms, there are five weighing matrices (Fig. 8a); while there are fifteen weighing matrices for 2D models (Fig. 8b). In Fig. 8b, 2D-TGL and 2D-cFSGL will only consider the constraints represented by solid lines; whereas 2D-TGL$^+$ and 2D-cFSGL$^+$ will consider both the solid lines and dashed lines.

To examine *Hypothesis I*, we target at three settings: (1D-Lasso, 2D-Lasso), (1D-TGL, 2D-TGL), and (1D-cFSGL, 2D-cFSGL), whose results in Table 4 are labeled by paired filled symbols, i.e. ■, ♣, and ♠, respectively.

On the other hand, *Hypothesis II* is tested on two settings: (2D-TGL, 2D-TGL$^+$) and (2D-cFSGL, 2D-FSGL$^+$). Their results in Table 4 are labeled by paired non-filled symbols, i.e. ♢ and ♡, respectively.



(a) One-dimensional Proximal Constraints

(b) Two-dimensional Proximal Constraints

Fig. 8. Diagrams for two hypotheses on ADNI dataset. Proximal relationship on clinical scores are shown as straight lines, which is used for general 2D models; whereas proximal biomarkers are linked with dashed lines and they are used as additional constraints for 2D-TGL$^+$ and 2D-FSGL$^+$.

We implement most of our algorithms in MATLAB as an extension to MALSAR package [26]; whereas *Pathwise Coordinate Optimization for Two-dimensional Fused Lasso*, whose pseudo code is shown as Algorithm 1, is written as a mex file in C++ with Eigen library [27] for pursuit of better runtime performance. For more details, please refer to our GitHub respository.

3.2 Prediction Performance

From Table 5a and 5b, we can observe that a model learns with MRI Image with External Information (M+E) features generally performs better than that with MRI Image Only (M) features, mainly because of the additional information.

TABLE 1
META features are collected through distributed ADNI datasets [25]. A variety of cognitive scores are included as features to predict feature cognitive scores.

Categories	Features (#num)
Demographic	age, years of education, gender; total of 3
Baseline Cognitive Scores	MMSE(1), ADAS subscores(13), ADAS-Cog(1), ADAS-MOD(1), Hachinski(1), CDR(7), FAQ(1), G.D.S(1), Neuropsychological Battery(1); total of 27
Lab tests	RCT1, RCT11, RCT12, RCT13, RCT14, RCT1407, RCT1408, RCT183, RCT19, RCT20, RCT29, RCT3, RCT392, RCT4, RCT5, RCT6, RCT8; total of 17

TABLE 2
Number of instances for different combination of MRI features and MMSE / ADAS-Cog target scores. There are overall 328 features for MRI dataset.

Features \ Targets	M06	M12	M24	M36	M48
SC	530	504	444	326	82
M06		432	384	286	74
M12			369	280	70
M24				255	67
M36					53

TABLE 3
Number of instances for different combination of MRI+META features and MMSE / ADAS-Cog target scores. There are overall 374 features for MRI+META dataset.

Features \ Targets	M06	M12	M24	M36	M48
SC	496	470	415	307	78
M06		404	361	271	70
M12			347	265	66
M24				244	64
M36					50

Ridge algorithm has the worst performance in CC and RMSE as expected, while Lasso algorithm closely defeats it in most datasets. TGL and cFSGL related algorithms have better prediction results, while in general, cFSGL defeats TGL slightly.

We conclude that we fail to reject *Hypothesis I: straightforwardly applying 2D concept without adding the proposed constraints cannot bring significant improvement*. In Table 5a and 5b, we see 2D algorithms do not consistently outperform their 1D counterparts. In other words, naively adding more information into model might lead to overfitting and thus does not guarantee a better model performance.

In terms of *Hypothesis II: the 2D concept with the proposed constraints cannot bring significant improvement*, our proposed 2D models with regularization terms pass all paired t-test significantly, as one can tell from the comparisons between (2D-TGL, 2D-TGL$^+$) and (2D-cFSGL, 2D-FSGL$^+$) across four datasets. Thus, we successivly reject this null hypothesis. That is to say, only after adding the derived proximal constraint S would make the 2D model effective.

4 Conclusion

We derive the upgrade procedures for models that equip two-dimensional proximal constraints with group lasso for disease progression, denoted as 2D-TGL$^+$ and 2D-cFSGL$^+$. We also acknowledge from experiments that the straightforward extension from one-dimensional framework to two-dimensional one, i.e. 1D-TGL → 2D-TGL and 1D-cFSGL → 2D-cFSGL, does not guarantee improvement; while adding further constraint between learned matrices as regularization terms, i.e. 2D-TGL → 2D-TGL$^+$ and 2D-cFSGL → 2D-cFSGL$^+$, significantly outperforms previous works. The solution is backed up with not only solid theoretical guarantee but also strong empirical outcomes.

References

[1] "Dementia - wikipedia." https://en.wikipedia.org/wiki/Dementia, 2016. [Online; accessed 07-Sep-2016].

[2] A. Burns and S. Iliffe, "Clinical review: Alzheimer's disease," *British Medical Journal 2009b*, vol. 338, p. b158, 2009.

[3] R. Lozano, M. Naghavi, K. Foreman, S. Lim, K. Shibuya, V. Aboyans, J. Abraham, T. Adair, R. Aggarwal, S. Y. Ahn, *et al.*, "Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the global burden of disease study 2010," *The Lancet*, vol. 380, no. 9859, pp. 2095–2128, 2013.

[4] "Dementia." http://www.who.int/mediacentre/factsheets/fs362/en/, 2016. [Online; accessed 22-June-2016].

[5] "Changing the trajectory of alzheimers disease: How a treatment by 2025 saves lives and dollars." http://www.alz.org/documents_custom/trajectory.pdf, 2016. [Online; accessed 22-June-2016].

[6] J. Morris, "Early-stage and preclinical alzheimer disease.," *Alzheimer disease and associated disorders*, vol. 19, no. 3, pp. 163–165, 2004.

[7] B. Dubois, H. H. Feldman, C. Jacova, J. L. Cummings, S. T. DeKosky, P. Barberger-Gateau, A. Delacourte, G. Frisoni, N. C. Fox, D. Galasko, *et al.*, "Revising the definition of alzheimer's disease: a new lexicon," *The Lancet Neurology*, vol. 9, no. 11, pp. 1118–1127, 2010.

[8] P. Vemuri, H. Wiste, S. Weigand, D. S. Knopman, J. Trojanowski, L. Shaw, M. A. Bernstein, P. Aisen, M. Weiner, R. C. Petersen, *et al.*, "Serial mri and csf biomarkers in normal aging, mci, and ad," *Neurology*, vol. 75, no. 2, pp. 143–151, 2010.

TABLE 4
Two hypothesis tests are examined through the comparisons between nine different algorithms on four datasets: combinations between two kinds of feature sets — MRI Image Only (M), MRI Image with External Information (M+E) and two kinds of target scores — Mini Mental State Examination (MMSE), Alzheimer's Disease Assessment Scale - Cognitive Subscale (ADAS-Cog). P-values with * indicates improvement significance under paired one-tailed t-test.

		(M, MMSE)	(M+E, MMSE)	(M, ADAS-Cog)	(M+E, ADAS-Cog)
1D	Ridge	0.742	0.777	0.745	0.795
	Lasso■	0.741	0.815	0.721	0.835
	TGL♣	0.778	0.832	0.769	0.851
	cFSGL♠	0.787	0.834	0.783	0.851
2D	Lasso■	0.755	0.827	0.731	0.834
	TGL♣♢	0.788	0.837	0.767	0.850
	cFSGL♠♡	0.799	0.842	0.773	0.849
	TGL+♢	0.818	**0.866**	0.801	0.874
	cFSGL+♡	**0.819**	**0.866**	**0.805**	**0.877**
H.T. I	P-Value■	0.0020*	0.0002*	0.0029*	0.5433
	P-Value♣	0.0765	0.0834	0.2734	0.7970
	P-Value♠	0.0054*	0.0203*	0.0022*	0.4477
H.T. II	P-Value♢	0.0018*	0.0002*	< 0.0001*	< 0.0001*
	P-Value♡	0.0001*	0.0002*	< 0.0001*	< 0.0001*

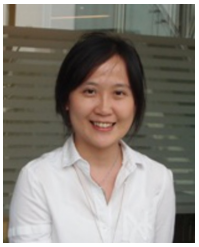
(a) Correlated Coefficient (CC) Evaluation Metric

		(M, MMSE)	(M+E, MMSE)	(M, ADAS-Cog)	(M+E, ADAS-Cog)
1D	Ridge	3.019	2.822	6.061	5.499
	Lasso■	2.952	2.533	6.143	4.855
	TGL♣	2.768	2.413	5.690	4.676
	cFSGL♠	2.739	2.412	5.639	4.595
2D	Lasso■	2.974	2.517	6.467	4.982
	TGL♣♢	2.827	2.385	5.901	4.793
	cFSGL♠♡	2.802	2.437	5.896	4.760
	TGL+♢	2.646	**2.225**	5.510	4.407
	cFSGL+♡	**2.638**	2.248	**5.378**	**4.270**
H.T. I	P-Value■	0.0904	0.3489	< 0.0001*	0.0038*
	P-Value♣	0.4042	0.1058	0.0004*	0.0031*
	P-Value♠	0.1427	0.3745	< 0.0001*	0.0026*
H.T. II	P-Value♢	0.0226*	0.0001*	< 0.0001*	< 0.0001*
	P-Value♡	0.0032*	0.0027*	< 0.0001*	< 0.0001*

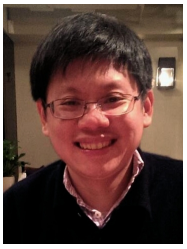
(b) Root Mean Squared Error (RMSE) Evaluation Metric

[9] R. A. Sperling, P. S. Aisen, L. A. Beckett, D. A. Bennett, S. Craft, A. M. Fagan, T. Iwatsubo, C. R. Jack, J. Kaye, T. J. Montine, *et al.*, "Toward defining the preclinical stages of alzheimers disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease," *Alzheimer's & dementia*, vol. 7, no. 3, pp. 280–292, 2011.
[10] R. J. Killiany, T. Gomez-Isla, M. Moss, R. Kikinis, T. Sandor, F. Jolesz, R. Tanzi, K. Jones, B. T. Hyman, and M. S. Albert, "Use of structural magnetic resonance imaging to predict who will get alzheimer's disease," *Annals of neurology*, vol. 47, no. 4, pp. 430–439, 2000.
[11] P. Neumann, S. Araki, A. Arcelus, A. Longo, G. Papadopoulos, K. a. Kosik, K. Kuntz, and A. Bhattacharjya, "Measuring alzheimers disease progression with transition probabilities estimates from cerad," *Neurology*, vol. 57, no. 6, pp. 957–964, 2001.
[12] S. Ray, M. Britschgi, C. Herbert, Y. Takeda-Uchimura, A. Boxer, K. Blennow, L. F. Friedman, D. R. Galasko, M. Jutel, A. Karydas, *et al.*, "Classification and prediction of clinical alzheimer's diagnosis based on plasma signaling proteins," *Nature medicine*, vol. 13, no. 11, pp. 1359–1362, 2007.
[13] Z. Zhang and B. D. Rao, "Sparse signal recovery with temporally correlated source vectors using sparse bayesian learning," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 5, no. 5, pp. 912–926, 2011.
[14] J. Wan, Z. Zhang, J. Yan, T. Li, B. D. Rao, S. Fang, S. Kim, S. L. Risacher, A. J. Saykin, and L. Shen, "Sparse bayesian multi-task learning for predicting cognitive outcomes from neuroimaging measures in alzheimer's disease," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 940–947, IEEE, 2012.
[15] J. Zhou, L. Yuan, J. Liu, and J. Ye, "A multi-task learning formulation for predicting disease progression," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 814–822, ACM, 2011.
[16] J. Zhou, J. Liu, V. A. Narayan, and J. Ye, "Modeling disease progression via fused sparse group lasso," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1095–1103, ACM, 2012.
[17] J. Zhou, J. Liu, V. A. Narayan, J. Ye, A. D. N. Initiative, *et al.*, "Modeling disease progression via multi-task learning," *NeuroImage*, vol. 78, pp. 233–248, 2013.
[18] T. G. Dietterich, "Machine learning for sequential data: A review," in *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pp. 15–30, Springer, 2002.
[19] J. Friedman, T. Hastie, H. Höfling, R. Tibshirani, *et al.*, "Pathwise coordinate optimization," *The Annals of Applied Statistics*, vol. 1, no. 2, pp. 302–332, 2007.
[20] Y. Nesterov, "A method of solving a convex programming problem with convergence rate o (1/k2)," *Soviet Mathematics Doklady*, vol. 27, no. 2, pp. 372–376, 1983.
[21] Y. Nesterov, "Smooth minimization of non-smooth functions," *Mathematical programming*, vol. 103, no. 1, pp. 127–152, 2005.
[22] P. Tseng, "On accelerated proximal gradient methods for convex-concave optimization. submitted to siam j," *J. Optim*, 2008.
[23] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.
[24] J. Liu, L. Yuan, and J. Ye, "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*, pp. 323–332, ACM, 2010.
[25] T. Iwatsubo, "Alzheimer's disease neuroimaging initiative (adni)," *Nihon rinsho. Japanese journal of clinical medicine*, vol. 69, p. 570, 2011.
[26] J. Zhou, J. Chen, and J. Y. MALSAR, "Multi-task learning via structural regularization," *There is no corresponding record for this reference*, 2011.
[27] G. Guennebaud, B. Jacob, *et al.*, "Eigen v3." http://eigen.tuxfamily.org, 2010.

Hao-en Sung received his B.S. degree in Computer Science and Information Engineering from National Taiwan University, Taipei, Taiwan in 2015. After his graduation, he first joined Intel-NTU center as a full-time researcher for driving behavior learning and present his work in Intel Asia Innovation Summit. Later, he worked as a full-time research assistant in Academia Sinica, where he focused on active learning and multi-task learning problems. His work, *A Classification model for Diverse and Noisy Labelers*, as a based model for active learning framework, is now accepted as a regular conference paper in PAKDD 2017. Hao-en is currently an M.S. student in Computer Science and Engineering of University California, San Diego.



Prof. Mi-yen Yeh Mi-Yen Yeh is currently Associate Research Fellow of Institute of Information Science (and Research Center for IT Innovation under joint appointment) at Academia Sinica, Taiwan. She received her B.S. and Ph.D. degrees in Electrical Engineering from National Taiwan University, Taiwan, in 2002 and 2009, respectively. Her main research area is on data mining and databases, with a specific focus on mining ordered data, social network analysis, and data management on non-volatile memory. She received the best paper award (in system software and security) of the 28th annual ACM Symposium on Applied Computing (SAC 2013), Distinguished Postdoctoral Fellowship in Academia Sinica, and Research Exploration Award in Pan Wen Yuan Foundation.



Prof. Shou-de Lin Shou-de Lin is currently a full professor in the CSIE department of National Taiwan University. He holds a BS in EE department from NTU, an MS-EE from the University of Michigan, and a PhD in Computer Science from the University of Southern California. He leads the Machine Discovery and Social Network Mining Lab in NTU. His international recognition includes the best paper award in IEEE Web Intelligent conference 2003, Google Research Award in 2007, Microsoft research award in 2008, merit paper award in TAAI 2010, 2014, 2016, best paper award in ASONAM 2011, US Aerospace AFOSR/AOARD research award winner for 5 years. He is the all-time winners in ACM KDD Cup, leading or co-leading the NTU team to win 5 championships. He also leads an NTU research team to win WSDM Cup 2016. He has served as the senior PC for SIGKDD and area chair for ACL.