Towards Better IncomLDL: We Are Unaware of Hidden Labels in Advance

Jiecheng Jiang[1,2*], **Jiawei Tang**[1*], **Jiahao Jiang**[1], **Hui Liu**[3], **Junhui Hou**[4†], **Yuheng Jia**[1,3,5†]

[1]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
[2]School of Computing, National University of Singapore, Singapore
[3]School of Computing Information Sciences, Saint Francis University, Hong Kong, China
[4]Department of Computer Science, City University of Hong Kong, Hong Kong, China
[5]Key Laboratory of New Generation Artificial Intelligence Technology and Its
Interdisciplinary Applications (Southeast University), Ministry of Education, China
{jcjiang, jwtang, jhjiang, yhjia}@seu.edu.cn, hliu99-c@my.cityu.edu.hk, jh.hou@cityu.edu.hk

Abstract

Label distribution learning (LDL) is a novel paradigm that describe the samples by label distribution of a sample. However, acquiring LDL dataset is costly and time-consuming, which leads to the birth of incomplete label distribution learning (IncomLDL). All the previous IncomLDL methods set the description degrees of "missing" labels in an instance to 0, but remains those of other labels unchanged. This setting is unrealistic because when certain labels are missing, the degrees of the remaining labels will increase accordingly. We fix this unrealistic setting in IncomLDL and raise a new problem: LDL with hidden labels (HidLDL), which aims to recover a complete label distribution from a real-world incomplete label distribution where certain labels in an instance are omitted during annotation. To solve this challenging problem, we discover the significance of proportional information of the observed labels and capture it by an innovative constraint to utilize it during the optimization process. We simultaneously use local feature similarity and the global low-rank structure to reveal the mysterious veil of hidden labels. Moreover, we **theoretically** give the recovery bound of our method, proving the feasibility of our method in learning from hidden labels. Extensive recovery and predictive experiments on various datasets prove the superiority of our method to state-of-the-art LDL and IncomLDL methods.

Code — https://github.com/Trisitana/HidLDL

1 Introduction

Learning with ambiguity has gained notable attention in machine learning community recently (Geng 2016). In classical machine learning, single-label learning refers to the process of assigning exactly one label to an instance. Multi-label learning (Tsoumakas and Katakis 2008; Zhang and Zhou 2014; Liu et al. 2021), on the other hand, assigns multiple labels to a single instance. However, in practical applications, for the same instance, we need to consider the relative importance of their labels. Label distribution learning (LDL) (Geng 2016) is a novel machine learning paradigm to solve the above issue by modeling the importances of multiple labels for each instance. In recent years, LDL have been widely used in many applications, such as age estimation (Geng, Yin, and Zhou 2013; Gao et al. 2018), emotion recognition (Jia et al. 2019b; Chen et al. 2020), facial beauty (Liang et al. 2018), crowd opinion prediction (Geng and Hou 2015), medical diagnosis (Wang et al. 2022), and scene text detection (Ren and Geng 2017; Ma et al. 2023) etc.

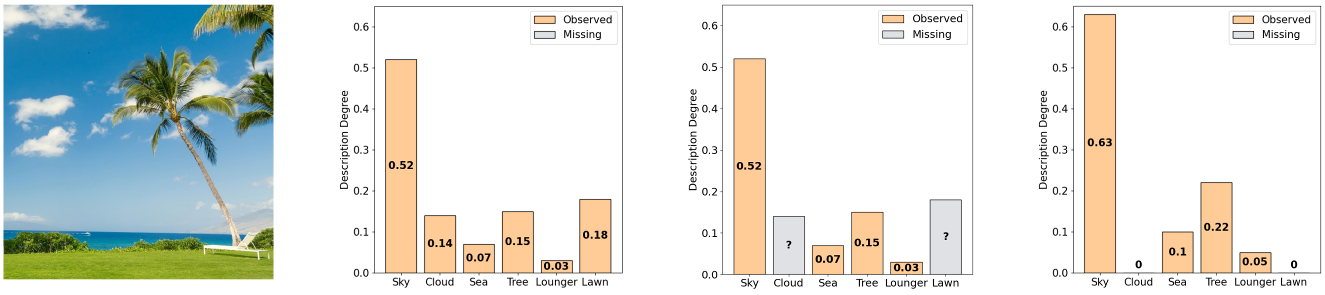
In LDL, each instance $\boldsymbol{x}$ is annotated by a real value vector $d_{\boldsymbol{x}}$, i.e., label distribution (LD). Each component of the LD vector $d_{\boldsymbol{x}}^y$ represents the description degree of the label y to the instance $\boldsymbol{x}$. Assume $d_{\boldsymbol{x}}^y \in [0,1]$ and the label set is complete, which means that all the labels in the set can always fully describe the instance. Then, we have $\sum_y d_{\boldsymbol{x}}^y = 1$.

Despite LDL being successfully applied to many scenarios in recent years, existing LDL methods are incapable of handling incomplete supervised information. In practice, the description degrees are often provided by human annotators, which can be extremely costly in terms of labor and time when dealing with a large dataset. Meanwhile, data corruption and incomplete annotations may also occur. To this end, incomplete label distribution learning (IncomLDL) (Xu and Zhou 2017) was proposed to address the problem of incomplete supervised data. In the previous research of IncomLDL, a predefined set of labels is given. The description degrees of some labels for a given instance are randomly set to 0, and those of other labels still remain unchanged.

However, the setting of IncomLDL is **unreasonable** and **unrealistic**. In real situations, when certain labels are missing (their description degrees set to 0 in IncomLDL), degrees of remaining labels should increase accordingly instead of remaining unchanged. Below, we use an illustration to better explain this unreasonable setting inherited in IncomLDL.

Fig. 1a shows a scene image containing 6 elements, and its LD $d_{\boldsymbol{x}} = \{0.52, 0.14, 0.07, 0.15, 0.03, 0.18\}$ is presented in Fig. 1b. In IncomLDL, some of the description degrees are randomly set to 0 so that incomplete data of LD are formed into $d_{\boldsymbol{x}} = \{0.52, 0, 0.07, 0.15, 0.03, 0\}$ (Fig. 1c). This is an approximation of real-world missing data. However, unless a portion of the data in the database is lost, this setting has a significant distance from the real world. Suppose we are the annotators. When annotating the image in Fig. 1a, we

[*]These authors contributed equally.
[†]Corresponding authors.



(a) image (b) LDL (c) IncomLDL (d) HidLDL

Figure 1: (a) A scene image containing 6 elements. (b) The label distribution (LD) of the image. (c) The observed LD in IncomLDL, with gray indicating missing (unobserved) labels. The sum of the description degrees of the observed labels is not 1. (d) The observed LD in HidLDL. For the unobserved labels in (c), the corresponding description degrees are 0, and the sum of the description degrees of the observed labels is 1. HidLDL is more intuitive and realistic, as the observed labels should occupy all description degree.

accidentally omit the labels for “cloud” and “lawn” (such omitted labels are referred to hereafter as hidden labels). In this case, the remaining labels ought to form a new comprehensive description of the instance, and the sum of their description degrees should equal 1. As illustrated in Fig. 1d, with “cloud” and “lawn” hidden from the annotator’s view, the description degrees of the other labels will increase to compensate and still manage to provide a complete description. Therefore, in IncomLDL, it is not a viable practice to simply set description degrees of hidden labels to 0 while keeping the description of the observed labels unchanged.

Based on the aforementioned illustration, we contend that current research on IncomLDL does not adequately reflect the true nature of incompleteness and leaves considerable room for refinement. We propose a new setting called LDL with hidden labels (HidLDL), which is closer to real situations. **We then propose a novel method that pays close attention to the crucial proportion information of observed labels, which is our most outstanding innovation.** By exploiting local feature similarity and the global low-rank structure to capture both local and global dependencies, we strive for further improvement. Moreover, we **theoretically** give the recovery bound of our method, proving the feasibility of our method in learning from hidden labels. Experimental results have shown that our method has significantly better performance compared to existing methods. The main contributions can be summarized as follows.

- We notice the unreasonable aspect of all previous incomplete LDL research, and propose a novel setting, abbreviated as HidLDL, which is closer to real situations.
- We discover the need to fully utilize the proportion information of observed labels in the new HidLDL problem. During optimization, apart from simultaneously using both local feature similarity and global low-rank structure, we innovatively use constraints to ensure the correct use of proportion information.
- We theoretically give the recovery bound of our method, in which the optimization objective and constraints of our method help determine the recovery bound, proving the feasibility of our method in learning from hidden labels.
- We empirically verify the effectiveness of our method on 12 real-world datasets in both recovery and predictive tasks. Experimental results show that our method outperforms state-of-the-art LDL and IncomLDL methods, with the improvements being statistically significant in numerous cases.

2 Related Work

2.1 Label Distribution Learning

Label distribution learning (LDL) (Xie et al. 2026; Tang and Jia 2025; Jia, Tang, and Jiang 2024; Kou et al. 2025b,a) was first proposed to solve the facial age estimation problem. Later on, (Geng 2016) discovered that in some real applications, the distribution across all labels is more desirable than the association of a single label to an instance. Since LDL was introduced, numerous methods have been proposed. These methods can be roughly divided into three categories: problem transformation, algorithm adaptation and specialized algorithms. Yang (Yang et al. 2015) proposed a deep learning based LDL method. Two classic methods SA-IIS and BFGS-LLD were then proposed in (Geng 2016). LDLLC (Jia et al. 2018) maps label distances to parameter matrix column distances and uses Pearson correlation to capture global label relationships. GD-LDL-SCL and Adam-LDL-SCL (Jia et al. 2019a) were designed to exploit local correlation in different groups which helps solve LDL problems. LDL-LDM (Wang and Geng 2021) was proposed to exploit label correlation maniford which is helpful to alleviate the overwhelm of LDL outspace. In (Wen et al. 2023), the authors proposed an auxiliary MLL process integrated into the LDL framework, which captures low-rank label correlations within this additional MLL component. Also, a new paradigm based on LDL, called OLDL (Wen et al. 2023), was proposed focusing on the internal sequential patterns of labels.

2.2 Incomplete Label Distribution Learning

The above mentioned LDL methods all assume that in the training set, each sample is annotated with a full label distribution. But in reality, the labels of some samples are either not observed, or it is expensive to annotate complete training samples (Li, Li, and Jia 2025; Jia et al. 2024, 2022). To this end, incomplete label distribution learning (IncomLDL) was introduced by (Xu and Zhou 2017) to address annotation incompleteness. To solve the LDL problem when given incomplete supervised information, they suggested a low-rank model to utilize the correlation between labels and proposed IncomLDL-prox and IncomLDL-admm. Adapting existing LDL methods to the scenario of missing annotations is another intuitive strategy for addressing incomplete LDL tasks. In (Li and Chen 2024), the authors designed a weighting scheme of label degrees to utilize prior information of label distribution itself. In (Wang and Geng 2021), the authors contended that the assumption of low-rank might not always be valid. They proposed an alternative model where the predictions are constrained to lie on a shared manifold, with the manifold's structure capturing the correlations between labels. Recently, in (Xu et al. 2025), the authors proposed an Incomplete Label Distribution Learning method via correlation decomposition by assuming a globally low-rank structure and a locally sparse structure. They developed an alternating solution with the accelerated proximal gradient descent method for optimization (IncomLDL-LCD).

3 LDL with Hidden Labels

3.1 Problem Setting

Let $\boldsymbol{x} \in R^d$ denote a feature vector, $\mathbf{X} = [\boldsymbol{x}_1; \boldsymbol{x}_2; \ldots; \boldsymbol{x}_n] \in R^{n\times d}$ denote the feature matrix and $\mathcal{Y} = \{y_1, y_2, \ldots, y_m\}$ denote the complete set of labels, where d, n and m represent the dimension of features, the number of samples, and the number of labels. In LDL, the description degree of the label y to the instance $\boldsymbol{x}$ is denoted by $d^y_{\boldsymbol{x}}$, which satisfies $d^y_{\boldsymbol{x}} \in [0,1]$ and $\sum_y d^y_{\boldsymbol{x}} = 1$, the label distribution of $\boldsymbol{x}_i$ is denoted by $\boldsymbol{d}^g_i = \left(d^{y_1}_{\boldsymbol{x}_i}, d^{y_2}_{\boldsymbol{x}_i}, \ldots, d^{y_m}_{\boldsymbol{x}_i}\right) \in R^m$, and the ground-truth label distribution matrix is denoted by $\mathbf{D}^g = [\boldsymbol{d}^g_1; \boldsymbol{d}^g_2; \ldots; \boldsymbol{d}^g_n] \in R^{n\times m}$.

In IncomLDL (Xu and Zhou 2017), for each sample $\boldsymbol{x}$, there exists a possibility that some entries in $d^y_{\boldsymbol{x}}$ may not be observed, aiming to simulate the real-world scenarios. Specially, let $\mathbf{M} \in R^{n\times m}$ denote the indices of the observed labels in $\mathbf{D}^g$:

$$M_{ij} = \begin{cases} 1 & \text{if } D^g_{ij} \text{ is observed,} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Then, the observed label distribution matrix in IncomLDL is defined as,

$$D^o_{ij} = \begin{cases} D^g_{ij} & \text{if } M_{ij} = 1, \\ 0 & \text{if } M_{ij} = 0. \end{cases} \tag{2}$$

However, this approach is not reasonable and has been discussed in detail in the previous context.

Therefore, we propose a new setting called HidLDL. Specifically, in HidLDL, due to some labels not being observed at the beginning, in the eyes of the annotator, other labels will occupy the degree of these hidden labels. We assume that the degrees of the observed labels in $\mathbf{D}^g$ will amplify proportionally. Based on this instinctive assumption, a more reasonable $\mathbf{D}^o$ is defined as,

$$D^o_{ij} = \begin{cases} \dfrac{D^g_{ij}}{\sum\limits_{k=1}^{m} D^g_{ik} \cdot M_{ik}} & \text{if } M_{ij} = 1, \\ 0 & \text{if } M_{ij} = 0. \end{cases} \tag{3}$$

The goal of HidLDL is to recover a label distribution matrix $\mathbf{D} = [\boldsymbol{d}_1; \boldsymbol{d}_2; \ldots; \boldsymbol{d}_n] \in R^{n\times m}$, which is as close as possible to the ground-truth label distribution. Due to the inaccuracy of the observed label distribution, HidLDL receives more ambiguity compared to IncomLDL and requires the use of new approach to solve it.

3.2 Proposed Method

In order to recover hidden labels, a natural idea is to learn from other instances containing these hidden labels. Based on the smoothness assumption (Zhu 2005), in the feature space, instances close to each other tend to share a common label. Even if the observed degree of labels is not accurate in HidLDL, the nearby points can still learn from each other.

Graph Dependency To capture graph dependency, we use K-nearest neighbors algorithm to determine whether $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are connected. Setting the neighbor number to the label number m, $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are connected if $\boldsymbol{x}_i$ is among K-nearest neighbors of $\boldsymbol{x}_j$ or vice versa. At first, the similarity matrix $\mathbf{A}$ is set to $\mathbf{0}_{n\times n}$, indicating that there are no initial dependencies. Then, if x_i and x_j are connected, A_{ij} is updated in Eq. (4):

$$A_{ij} = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2}{2\sigma^2}\right), \tag{4}$$

where σ is the bandwidth parameter of the Gaussian kernel. If $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are closely connected, then $\boldsymbol{d}_i$ and $\boldsymbol{d}_j$ should be close to each other. This idea leads to the following optimization objectives $\mathit{\Omega}(\mathbf{D})$:

$$\mathit{\Omega}(\mathbf{D}) = \sum_{i,j} A_{ij} \|\boldsymbol{d}_i - \boldsymbol{d}_j\|_2^2 = \operatorname{tr}\left(\mathbf{D}^\top \mathbf{G} \mathbf{D}\right), \tag{5}$$

where $\mathbf{G} = \hat{\mathbf{A}} - \mathbf{A}$ is the graph Laplacian and $\hat{\mathbf{A}}$ is a diagonal matrix whose elements are $\hat{A}_{ii} = \sum_{j=1}^n A_{ij}$.

Low-rank Assumption Global low-rank assumption has good performance in handling incomplete supervision information (Xu and Zhou 2017). To further mine the correlation between labels, we assume that $\mathbf{D}$ is low-rank, i.e., we hope that the trace norm of the recovered matrix $\mathbf{D}$ is as small as possible. Therefore, the trace norm of $\mathbf{D}$ (i.e., $\|\mathbf{D}\|_*$) is added as an optimization term, which is the sum of all singular values of the matrix.

Proposed Proportional Constraint The most important thing in HidLDL is how to utilize the observed matrix $\mathbf{D}^o$. In (Li and Chen 2024), Frobenius norm is used to minimize the difference in those observed positions, however, in HidLDL, this may not work because observed degrees of labels no longer directly reflect their final description of the instance.

Specifically, we must only use the proportional relationship between the degrees of the observed labels to avoid introducing more noise. Even if we try to incorporate this proportion information of $\mathbf{D}$ into the final optimization objective, it not only introduces additional hyperparameters but also brings in more non-smooth terms. Thus, we define the following constraint set:

$$\begin{aligned} Cons = \big\{ \mathbf{Z} \in R^{n\times m} \mid \forall i \in \{1, 2, \dots, n\}, \exists k_i \in R \\ \text{s.t. } k_i \cdot \mathbf{d}_i^o = \mathbf{z}_i \odot \mathbf{m}_i \ \big\}, \end{aligned} \tag{6}$$

where the *i*-th row of $\mathbf{Z}$ is denoted as $\mathbf{z}_i$, so is $\mathbf{m}_i$ and $\mathbf{d}_i^o$. We constrain $\mathbf{D} \in Cons$. We expect that each row of the recovered label matrix, after masking, is proportional to the initial observed matrix, with a scaling coefficient of k_i. This setting makes sense even when $\mathbf{D}_{ij}^o$ is 0, thus increasing the generality of our model.

Note that the recovered label distribution needs to be in the probability simplex. We add constraints $\mathbf{D} \times \mathbf{1}_m = \mathbf{1}_n$ and $\mathbf{D} \geq \mathbf{0}_{n\times m}$ to guarantee that all entries are negative and the sum of each row's entries is 1. Considering all the factors mentioned above, the final objective function of our method can be written as follows:

$$\begin{aligned} \min_{\mathbf{D}} \quad & \frac{1}{2} \operatorname{tr}\left(\mathbf{D}^\top \mathbf{G} \mathbf{D}\right) + \alpha \|\mathbf{D}\|_* \\ \text{s.t.} \quad & \mathbf{D} \times \mathbf{1}_m = \mathbf{1}_n,\ \mathbf{D} \geq \mathbf{0}_{n\times m},\ \mathbf{D} \in Cons, \end{aligned} \tag{7}$$

where $\operatorname{tr}(\cdot)$ denotes the trace of a matrix, $\|\cdot\|_*$ denotes the trace norm of the matrix, and α serves as a regularization parameter that controls the trade-off between graph dependency and the trace norm.

3.3 Theoretical Analysis

In this section, we provide the recovery bound of our method.

Lemma 1. *For i-th instance, the recovery bound of it can be represented by the error between the ground-truth scaling coefficient $k_i^g = \frac{d_{ij}^o}{d_{ij}^g}$ and the recovered scaling coefficient k_i. So, the target is to prove that*

$$(k_i^g - k_i)^2 \leq \epsilon_i, \tag{8}$$

where ϵ_i is a small number.

Theorem 1. *Under certain assumptions, the error of the scaling coefficients can be expressed as*

$$k_i^g - k_i \leq \frac{1 - \sum_{k=1}^m D_{ik} M_{ik}}{\sum_{k=1}^m D_{ik}^g M_{ik}}. \tag{9}$$

The detailed proof of **Theorem 1** is given in section **F** of the appendix. Let $\sigma_i = 1 - \Sigma_k D_{ik} M_{ik}$. In view of the definition of M_{ik}, σ_i represents **the sum of the labels in i-th instance that are masked**, and the recovery bound of i-th instance ϵ_i becomes

$$\epsilon_i \leq \frac{\sigma^2}{(\Sigma_k D_{ik}^g M_{ik})^2}. \tag{10}$$

From Eq. (10), we can observe that its denominator is strictly greater than zero but less than one. And the smaller the σ_i is, the tighter the recovery bound ϵ_i is. When σ_i tends to zero, ϵ_i also tends to zero. This degrades the hidden label LDL problem to the learnable traditional LDL problem, which is an intuitive conclusion. *Combining all the proofs together, we conclude that LDL with hidden labels through our method is theoretically feasible.*

3.4 Optimizing using ADMM

Considering that the optimization objective contains complex constraints and non-smooth terms, we use ADMM (Alternating Direction Method of Multipliers) (Boyd et al. 2011) to solve Eq. (7). By breaking Eq. (7) into smaller pieces, each of which will be easier to handle. Specifically, Eq. (7) can be converted into the following equivalent forms by replacing the matrix $\mathbf{D}$ with $\mathbf{A}$ and $\mathbf{B}$:

$$\begin{aligned} \min_{\mathbf{A},\mathbf{B},\mathbf{D}} \quad & \frac{1}{2} \operatorname{tr}\left(\mathbf{D}^\top \mathbf{G} \mathbf{D}\right) + \alpha \|\mathbf{A}\|_* \\ \text{s.t.} \quad & \mathbf{D} \times \mathbf{1}_m = \mathbf{1}_n,\ \mathbf{D} \geq \mathbf{0}_{n\times m},\ \mathbf{B} \in Cons, \\ & \mathbf{D} - \mathbf{A} = \mathbf{0},\ \mathbf{D} - \mathbf{B} = \mathbf{0}. \end{aligned} \tag{11}$$

Then, the solution of Eq. (21) can be accomplished by solving the following augmented Lagrange multiplier equation:

$$\begin{aligned} \min_{\mathbf{A},\mathbf{B},\mathbf{D}} \quad & \frac{1}{2} \operatorname{tr}\left(\mathbf{D}^\top \mathbf{G} \mathbf{D}\right) + \alpha \|\mathbf{A}\|_* \\ & + \langle \mathbf{\Lambda}, \mathbf{D} - \mathbf{A} \rangle + \frac{\rho}{2} \|\mathbf{D} - \mathbf{A}\|_F^2 \\ & + \langle \mathbf{\Lambda}', \mathbf{D} - \mathbf{B} \rangle + \frac{\rho}{2} \|\mathbf{D} - \mathbf{B}\|_F^2 \\ \text{s.t.} \quad & \mathbf{D} \times \mathbf{1}_m = \mathbf{1}_n,\ \mathbf{D} \geq \mathbf{0}_{n\times m},\ \mathbf{B} \in Cons, \end{aligned} \tag{12}$$

where $\mathbf{\Lambda} \in R^{n\times m}$ and $\mathbf{\Lambda}' \in R^{n\times m}$ are the Lagrange multipliers, $\rho \geq 0$ is a penalty parameter, and $\langle \cdot, \cdot \rangle$ returns the inner product of two matrices. Although ρ can increment in each loop, in our paper, we fix ρ at 2, which is sufficient for the objective to converge. To solve Eq. (12), the ADMM iteratively solves the following subproblems.

1) $\mathbf{D}$ Subproblem: Removing irrelated terms regarding $\mathbf{D}$, $\mathbf{D}$ subproblem is written as:

$$\begin{aligned} \min_{\mathbf{D}} \ & \frac{1}{2} \operatorname{tr}\left(\mathbf{D}^\top \mathbf{G} \mathbf{D}\right) + \langle \mathbf{\Lambda} \cdot \mathbf{D} - \mathbf{A} \rangle + \frac{\rho}{2} \|\mathbf{D} - \mathbf{A}\|_F^2 \\ & + \langle \mathbf{\Lambda}' \cdot \mathbf{D} - \mathbf{B} \rangle + \frac{\rho}{2} \|\mathbf{D} - \mathbf{B}\|_F^2 \\ \text{s.t.} \quad & \mathbf{D} \times \mathbf{1}_m = \mathbf{1}_n,\ \mathbf{D} \geq \mathbf{0}_{n\times m}. \end{aligned} \tag{13}$$

We use projected gradient descent (PGD) to solve Eq. (13). The gradient of the objective with respect to $\mathbf{D}$ is:

$$\nabla_{\mathbf{D}} = \mathbf{G}\mathbf{D} + \mathbf{\Lambda} + \mathbf{\Lambda}' + \rho(\mathbf{D} - \mathbf{A}) + \rho(\mathbf{D} - \mathbf{B}). \tag{14}$$

We use Stochastic Gradient Descent (SGD) to update $\mathbf{D}$ one time in a step and project it to probability simplex.

Although for every instance $\boldsymbol{d}_i$, in (Li and Chen 2024), the projection onto a probability simplex problem is solved using an $O(m \log m)$ (Wang and Carreira-Perpinán 2013) algorithm by brute force searching through $[\mathrm{m}]$ from the largest entry to the smallest one for a particular j satisfying the KKT condition, this projection method is not efficient. We use a much simpler projection function in $O(m)$:

$$\begin{aligned} &\text{Step I: Setting all the } D_{ij} \text{ to 0 if } D_{ij} < 0, \\ &\text{Step II: Normalize } \mathbf{D} \text{ as } D_{ij} = D_{ij} / \sum_{k=1}^{m} D_{ik}. \end{aligned} \tag{15}$$

We set a uniform distribution if $\sum\limits_{k=1}^{m} D_{ik} = 0$. This projection method can be realized in $O(m)$, which is much faster than (Wang and Carreira-Perpinán 2013). At the same time, it can maintain proportional information and ultimately achieve better results in experiments.

2) $\mathbf{A}$ Subproblem: Removing irrelated terms regarding $\mathbf{A}$, $\mathbf{A}$ subproblem can be rewritten into

$$\min_{\mathbf{A}} \frac{1}{2}\|\mathbf{A} - (\mathbf{D} + \frac{\mathbf{\Lambda}}{\rho})\|_F^2 + \frac{\alpha}{\rho}\|\mathbf{A}\|_*, \tag{16}$$

which has a closed-form solution by using a singular value thresholding operator (Cai, Candès, and Shen 2010).

3) $\mathbf{B}$ Subproblem: Removing irrelated terms regarding $\mathbf{B}$, $\mathbf{B}$ subproblem can be solved by optimizing each row of $\mathbf{B}$,

$$\begin{aligned} &\min_{\boldsymbol{b}_i} \langle \boldsymbol{\lambda}_i^{'}, \boldsymbol{d}_i - \boldsymbol{b}_i \rangle + \frac{\rho}{2}\|\boldsymbol{d}_i - \boldsymbol{b}_i\|_F^2 \\ &\text{s.t. } \boldsymbol{b}_i \in \{\boldsymbol{e}_i \mid \mathbf{E} \in Cons\}, \end{aligned} \tag{17}$$

where the i-th row of $\mathbf{B}$ is denoted as $\boldsymbol{b}_i$, so is $\boldsymbol{\lambda}_i^{'}$, $\boldsymbol{d}_i$, $\boldsymbol{m}_i$, and $\boldsymbol{e}_i$. Note that in Eq. (17), if $M_{ij} = 0$, no matter what k_i is, $k_i \cdot D_{ij}^o = B_{ij} \cdot M_{ij} = 0$. As a result, for those hidden positions, constraint is automatically satisfied. This problem turns out to be an unconstrained quadratic optimization, and B_{ij} is directly set as $D_{ij} + \frac{\Lambda_{ij}^{'}}{\rho}$. When it comes to observed positions, $B_{ij} = k_i \cdot D_{ij}^o$, where D_{ij}^o is a constant. Replacing all the B_{ij} with k_i, the optimization object is a quadratic equation about k_i, which is easy to deal with. Therefore, $\mathbf{B}$ can be updated as follows:

$$B_{ij} = \begin{cases} D_{ij} + \dfrac{\Lambda_{ij}'}{\rho} & M_{ij} = 0, \\ \dfrac{\sum_{k=1}^{m} (\rho D_{ik} + \Lambda_{ik}') \cdot D_{ik}^o \cdot M_{ik}}{\rho \sum_{k=1}^{m} (D_{ik}^o)^2 \cdot M_{ik}} D_{ij}^o & M_{ij} = 1. \end{cases} \tag{18}$$

Note that at least one entry in $\boldsymbol{d}_i^o$ is not 0, so $\rho \sum_{k=1}^{m} (D_{ik}^o)^2 \cdot M_{ik} > 0$. As a result, Eq. (18) is always meaningful and B_{ij} can be updated correctly.

Furthermore, we update the Lagrange multiplier matrices as follows:

$$\begin{cases} \mathbf{\Lambda} = \mathbf{\Lambda} + \rho(\mathbf{D} - \mathbf{A}), \\ \mathbf{\Lambda}' = \mathbf{\Lambda}' + \rho(\mathbf{D} - \mathbf{B}). \end{cases} \tag{19}$$

See the overall pseudocode of the solution to Eq. (7) in Algorithm 1. Besides, the stopping criterion of the algorithm is that the number of iteration reaches 100, or the maximum of all the residuals of the optimized variables is less than 10^{-3}, i.e., $\max\left(\|\mathbf{D} - \mathbf{A}\|_\infty, \|\mathbf{D} - \mathbf{B}\|_\infty\right) < 10^{-3}$, where $\|\cdot\|_\infty$ denotes the infinity norm of the matrix. The convergence of Eq. (7) is theoretically guaranteed and we provide a proof in Section **E** of the Appendix.

Algorithm 1: Optimization to Eq. (7)

```
Input: $\mathbf{D}_o$, the Laplacian matrix $\mathbf{G}$, mask $\mathbf{M}$
Initialization: $\mathbf{A} = \mathbf{B} = \mathbf{\Lambda}_1 = \mathbf{\Lambda}_2 = \mathbf{1}_{n \times m}, \mathbf{D} = \mathbf{D}_o$, $t = 1, \rho = 2$
while stopping criterion is not satisfied do
    Update $\mathbf{D}$ by Eq. (14);
    Project $\mathbf{D}$ by Eq. (15);
    Solve $\mathbf{A}$ by Eq. (16);
    Update $\mathbf{B}$ by Eq. (18);
    Update $\mathbf{\Lambda}, \mathbf{\Lambda}'$ by Eq. (19);
    $t = t + 1$;
end while
return $\mathbf{D}$
```

4 Experiment

4.1 Datasets

In this paper, we evaluate our method on 12 real-world datasets covering fields of biology, facial expression, natural scene, emotion recognition and movie. The statistics of 12 datasets are provided in Section **A** of the Appendix, and the details of them can be found in (Geng 2016; Peng et al. 2015; Li and Deng 2019).

To simulate the presence of hidden labels in the dataset, we will first make these datasets random missing (Xu and Zhou 2017) and set the missing $d_{\boldsymbol{x}}^{y}$ to 0. We vary the missing rate ω from 40% to 80%. Then, we check the missing datasets and ensure that for each instance $\boldsymbol{x}$, at least one label description has a value greater than 0 (i.e., $\exists j$ such that $d_{\boldsymbol{x}}^{y_j} > 0$). In HidLDL, the input is a seemingly complete label distribution dataset, so in the second step, we normalize missing datasets into the final hidden setting using Eq. (3).

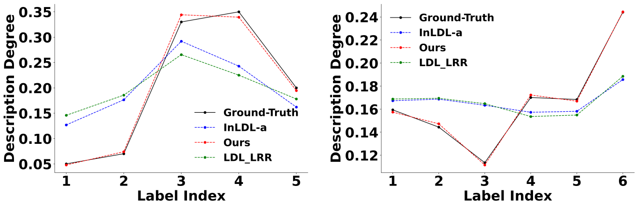
4.2 Settings and Compared Methods

To validate the effectiveness of our method, we design two experimental configurations. In the first setting, we use a hidden dataset and features of instances together to restore the true label distribution. Difference between the ground-truth and the recovered distribution matrix will be measured. In the second setting, to further verify the quality of the recovered label distribution, we train each recovered label distribution using the same simple LDL algorithm, such as SA-BFGS (Geng 2016). We then use each trained model for prediction. The training and testing sets are partitioned with ratios of 0.8 and 0.2, respectively. We repeat each experiment 5 times and report the average results.

We compare our method with six methods. Two of them are state-of-the-art methods designed for IncomLDL including IncomLDL-admm (abbreviated as InLDL-a) (Xu and Zhou 2017) and WInLDL (Li and Chen 2024), which can still be used for the hidden setting. Two state-of-the-art LDL methods named LDL-LRR (Jia et al. 2021) and LDL-DPA (Jia et al. 2023) are included. We also include two baselines named SA-IIS and PT-Bayes (Geng 2016). For comparsion methods, we use default parameters suggested in their original papers, with the exception that we adjust the regularization parameter for IncomLDL-admm from $2^{\{-10,-9,\ldots,9,10\}}$. For our method, the regularization parameter α is selected

Dataset	Ours	InLDL-a	WInLDL	SA-IIS	LDL-LRR	PT-Bayes	LDL-DPA
alpha	$\mathbf{0.4718 \pm .0011}$	$\underline{0.6878 \pm .0010}^{\bullet}$	$0.9660 \pm .0179^{\bullet}$	$0.9466 \pm .0084^{\bullet}$	$0.7078 \pm .0056^{\bullet}$	$1.9636 \pm .0673^{\bullet}$	$1.0037 \pm .0117^{\bullet}$
cdc	$\mathbf{0.3803 \pm .0044}$	$\underline{0.6493 \pm .0023}^{\bullet}$	$0.8870 \pm .0172^{\bullet}$	$0.8590 \pm .0154^{\bullet}$	$0.6607 \pm .0050^{\bullet}$	$1.6762 \pm .0708^{\bullet}$	$0.9007 \pm .0038^{\bullet}$
cold	$\mathbf{0.1746 \pm .0005}$	$\underline{0.2442} \pm .0013^{\bullet}$	$0.2800 \pm .0049^{\bullet}$	$0.2918 \pm .0045^{\bullet}$	$0.2519 \pm .0045^{\bullet}$	$0.4631 \pm .0551^{\bullet}$	$0.3068 \pm .0086^{\bullet}$
dtt	$\mathbf{0.1237 \pm .0008}$	$\underline{0.1734 \pm .0008}^{\bullet}$	$0.2200 \pm .0058^{\bullet}$	$0.2328 \pm .0063^{\bullet}$	$0.1825 \pm .0062^{\bullet}$	$0.4363 \pm .0410^{\bullet}$	$0.2499 \pm .0032^{\bullet}$
elu	$\mathbf{0.3458 \pm .0072}$	$\underline{0.5899 \pm .0011}^{\bullet}$	$0.8038 \pm .0166^{\bullet}$	$0.7791 \pm .0151^{\bullet}$	$0.6029 \pm .0073^{\bullet}$	$1.4970 \pm .0896^{\bullet}$	$0.8156 \pm .0223^{\bullet}$
spo	$\mathbf{0.3250 \pm .0064}$	$\underline{0.5162 \pm .0011}^{\bullet}$	$0.5619 \pm .0042^{\bullet}$	$0.5646 \pm .0075^{\bullet}$	$0.5202 \pm .0020^{\bullet}$	$0.8059 \pm .0204^{\bullet}$	$0.5886 \pm .0072^{\bullet}$
SJAFFE	$\mathbf{0.5983 \pm .0235}$	$0.8592 \pm .0535^{\bullet}$	$2.0855 \pm .1299^{\bullet}$	$\underline{0.8421 \pm .0339}^{\bullet}$	$0.8891 \pm .0050^{\bullet}$	$0.9838 \pm .0659^{\bullet}$	$2.0990 \pm .1017^{\bullet}$
Scene	$\mathbf{6.4190 \pm .0026}$	$\underline{6.7305 \pm .0113}^{\bullet}$	$6.7673 \pm .0186^{\bullet}$	$6.7141 \pm .0078^{\bullet}$	$6.7761 \pm .0050^{\bullet}$	$7.2195 \pm .1133^{\bullet}$	$6.7538 \pm .0139^{\bullet}$
Movie	$\mathbf{0.7494 \pm .0041}$	$\underline{1.1138 \pm .0047}^{\bullet}$	$1.5030 \pm .0083^{\bullet}$	$1.6548 \pm .0117^{\bullet}$	$1.1415 \pm .0068^{\bullet}$	$4.2956 \pm .1179^{\bullet}$	$1.4271 \pm .0389^{\bullet}$
SBU	$\mathbf{0.7028 \pm .0086}$	$0.8375 \pm .0048^{\bullet}$	$\underline{0.8278 \pm .0103}^{\bullet}$	$0.8686 \pm .0041^{\bullet}$	$0.8998 \pm .0036^{\bullet}$	$0.8737 \pm .0088^{\bullet}$	$0.8552 \pm .0062^{\bullet}$
Emo	$\mathbf{2.4372 \pm .1010}$	$\underline{3.7105 \pm .0233}^{\bullet}$	$3.7991 \pm .0087^{\bullet}$	$3.7944 \pm .0125^{\bullet}$	$3.7732 \pm .0046^{\bullet}$	$3.8714 \pm .0415^{\bullet}$	$3.7921 \pm .0150^{\bullet}$
RAF	$3.0345 \pm .0095$	$4.6553 \pm .0079^{\bullet}$	$5.2983 \pm .0065^{\bullet}$	$\mathbf{2.9245 \pm .0096}^{\circ}$	$2.9998 \pm .0054^{\circ}$	$3.1821 \pm .0440^{\bullet}$	$\underline{2.9374 \pm .0112}^{\circ}$

Table 1: Canberra (the lower the better) results for the **recovery** setting on all datasets when missing rate ω = 50%. The value is shown in mean±std form. Bold and underlined indicate the best and second best results, respectively. Under single-tailed paired t-test at a significance level of 0.05, $\bullet$ means our method's performance is statistically significantly better than the compared method and $\circ$ means the compared method's performance is statistically significantly better than ours.



(a) Movie

(b) spo

Figure 2: The visualization of two typical recovery results on the Movie (left) and spo (right) dataset.

from $2^{\{-10,-9,\ldots,9,10\}}$ based on the recover performance, while σ and ρ are set to be 1 and 2, respectively. In K-nearest neighbors algorithm, the number of neighbors is set as the number of labels. Each method will face a matrix with exactly the same missing entries in each experiment.

We use five metrics for the LDL with Hidden Labels problem, including Chebyshev, Clark, Canberra, Cosine and Intersection. Details of them can be found in (Geng 2016). Among them, Canberra, Clark and Canberra measure the distance between two vectors, thus they are the lower the better. Cosine and Intersection measure the similarity between two vectors, thus they are the higher the better.

4.3 Recover Results and Discussions

In the following, we report the recover results of different methods at different missing rate. Due to space limitation, here we only list representative results. The Canberra (the lower the better) results on all the dataset with $\omega = 50\%$ are shown in Table 1. Other results are shown in Section **B.1** of the Appendix. Moreover, different from previous research, we also study impact of missing rates on the results, which are also presented in Section **B.2** of the Appendix.

To conduct significance analysis, we conduct the paired t-test, which is a useful statistical test for comparing two methods (Box 1987). For each dataset, we perform 6 separate single-tailed paired t-tests, comparing our method with 6 compared methods individually. We mark the compared method with a black dot $\bullet$ in Table 1 if our method's performance is statistically significantly better than the compared method at a significance level of 0.05.

Metric	Method	cold	dtt	spo	SBU	Emo	RAF
Clark↓	Ours	**0.099**	**0.072**	**0.163**	**0.327**	**1.249**	**1.486**
	w/o $Cons$	0.147	0.102	0.256	0.413	1.679	1.611
	w/o TN	0.123	0.099	0.189	0.331	1.538	1.583
Cosine↑	Ours	**0.993**	**0.997**	**0.989**	**0.951**	**0.825**	**0.783**
	w/o $Cons$	0.987	0.993	0.975	0.918	0.657	0.642
	w/o TN	0.989	0.996	0.985	0.947	0.685	0.664

Table 2: Ablation Results on 6 Datasets. ↑ (↓) indicates the higher (lower) the better.

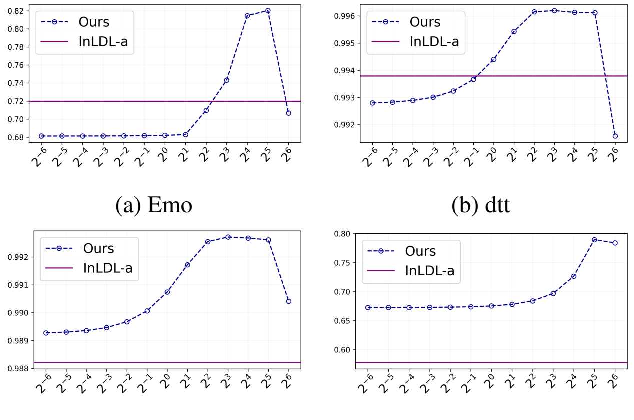
By analyzing the experimental results, we can arrive at the following three conclusions:

- Our method achieves the lowest average rank in terms of all five metrics. As shown in Table 1, we win 69 times out of 72 Canberra comparisons, with a **95.8%** rate to win, demonstrating that our method is superior to other compared methods. More importantly, our advantages are statistically significant on these 69 cases according to t-test. For detailed metrics, our improvement is particularly evident in Chebyshev, Clark and Canberra.
- IncomLDL methods are not much better than LDL methods in the recovery setting. WInLDL, although reported to be faster and more effective in the incomplete setting (Li and Chen 2024), performs quite poorly. InLDL-a, although performs relatively stably, is actually close to LDL-LRR in metrics' values.
- As the missing rate increases, the performance metrics of almost all methods become worse. Our method has an outstanding advantage when the missing rate is between 40% and 60%, as shown in Section **B.2** of the Appendix.

Fig. 2 shows typical recovery results on two datasets. We selected two methods, InLDL-a and LDL-LRR, which are reported to have relatively better performance, to compare with our method. According to the visualization, it can be seen that our method (**red line**) is closer to the ground truth (**black line**). Meanwhile, our trend has greater consistency with the ground truth, which is helpful to identify labels with the largest and smallest description degrees.

Dataset	Ours	InLDL-a	WInLDL	SA-IIS	LDL-LRR	PT-Bayes	LDL-DPA
alpha	**0.6890 ± .0012**	0.6936 ± .0022•	1.0180 ± .0094•	0.9681 ± .0138•	0.7073 ± .0101•	1.8878 ± .0996•	1.0802 ± .0173•
cdc	**0.6464 ± .0020**	0.6506 ± .0056	0.9562 ± .0250•	0.8731 ± .0175•	0.6641 ± .0095•	1.7754 ± .1194•	0.9767 ± .0178•
cold	**0.2329 ± .0005**	0.2369 ± .0021•	0.2930 ± .0117•	0.2834 ± .0071•	0.2466 ± .0038•	0.5084 ± .0815•	0.3148 ± .0128•
dtt	**0.1635 ± .0002**	0.1674 ± .0010•	0.2370 ± .0120•	0.2351 ± .0149•	0.1916 ± .0126•	0.4543 ± .0412•	0.2714 ± .0215•
elu	**0.5701 ± .0009**	0.5771 ± .0035•	0.8423 ± .0178•	0.7835 ± .0119•	0.5931 ± .0057•	1.5758 ± .1034•	0.8857 ± .0186•
spo	**0.5169 ± .0011**	0.5197 ± .0039	0.5956 ± .0137•	0.5702 ± .0132•	0.5243 ± .0048•	0.8229 ± .0539•	0.6182 ± .0193•
SJAFFE	**0.7935 ± .0182**	0.8940 ± .0263•	2.0713 ± .0657•	0.8980 ± .0248•	0.9120 ± .0093•	1.0189 ± .0281•	1.9060 ± .1365•
Scene	**6.8070 ± .0137**	6.8531 ± .0254•	6.9508 ± .0242•	6.8116 ± .0127	6.8298 ± .0083•	7.2288 ± .0596•	6.8700 ± .0238•
Movie	**1.0787 ± .0080**	1.2025 ± .0089•	1.7374 ± .0250•	1.9103 ± .0234•	1.1610 ± .0082•	4.2877 ± .0859•	1.4859 ± .0210•
SBU	0.8411 ± .0076	0.8716 ± .0053•	**0.8328 ± .0092**°	0.9017 ± .0037•	0.9142 ± .0007•	0.8988 ± .0071•	0.8892 ± .0094•
Emo	**3.7057 ± .0058**	3.9040 ± .0219•	3.8321 ± .0096•	4.2095 ± .0456•	3.8376 ± .0373•	3.8614 ± .0815•	3.9123 ± .0411•
RAF	3.0544 ± .0043	**2.8871 ± .0133**°	4.3674 ± .0360•	3.0108 ± .0123°	3.0033 ± .0108°	3.0488 ± .0245	2.9869 ± .0151°

Table 3: Canberra (the lower the better) results for the **predictive** setting on all datasets when missing rate $\omega = 50\%$. The value is shown in mean±std form. Bold and underlined indicate the best and second best results, respectively. Under single-tailed paired t-test at a significance level of 0.05, • means our method's performance is statistically significantly better than the compared method and ○ means the compared method's performance is statistically significantly better than ours.



(a) Emo

(b) dtt

(c) cold

(d) RAF

Figure 3: Comparison of our method and IncomLDL-admm on Cosine (the higher the better) and the horizontal axis represents the hyper-parameter α with the missing rate ω=50%.

Ablation Studies We perform ablation experiments on six datasets to validate the importance of our proportional constraint and global dependency. As shown in Table 2, "w/o $Cons$" method means our method without proportional constraint, while "w/o TN" means our method without the trace norm (TN) term (i.e., $\alpha = 0$). At 50% missing rate, "Ours" has a huge improvement in all the evaluation metrics compared to the "w/o $Cons$", indicating that it is critical to maintain the proportional relationship between labels. Meanwhile, compared to "w/o TN", there has also been an improvement. Other results also show indispensability of $Cons$ and TN, which are provided in Section **D** of the Appendix.

Sensitivity Analysis Our method involves one parameter: α in the trace norm regularization. The analysis of the parameter α is shown in Fig. 3. To demonstrate the performance of our method, we set the IncomLDL-admm method to achieve its best tuned results. From Fig. 3, it can be seen that when $\alpha < 2^3$, our method shows a stable growth in performance as α increases. When $\alpha > 2^5$, the performance of our method deteriorates rapidly. As a result, when α is set between 2^3 and 2^5, relatively good results can be achieved.

4.4 Predict Results and Discussions

We present a representative result in Table 3 with a 50% missing rate for the metric Canberra, while the results for other missing rates and metrics are similar, which are included in Section **C** of the Appendix. Similar to the recovery setting, the single-tailed paired t-test is conducted to analyze the significant advantages of our method.

From the reported results, we observe that:

- Our method achieves the lowest average rank in terms of all five metrics. Taking Table 3 as an example, out of 12 datasets, we rank 1st 10 times and 2nd once. Out of 72 comparisons, we win 66 times. With a winning rate of over **91.6%**, our method performs better than most compared methods. And it is worth noting that our method has a significant advantage most of the time.
- Methods focusing on the relationship between labels (such as LDL-LRR and InLDL-a) can capture advantageous information in prediction even if their recovery results are poor.
- Our method is less affected by the missing rate in the predictive scenarios. As the missing rate increases, our method is robust and deteriorates very slowly. Relevant details are in Section **C.2** of the Appendix.

5 Conclusion

This paper gives a creative answer to the question: "What does real-world incomplete data in IncomLDL actually look like and how to deal with it". Specifically, we clarify the irrationality of the training dataset in IncomLDL, and repair it by normalization. We then proposed a new method to solve this brand-new HidLDL problem, which emphasizes maintaining the observed labels' proportion during the recovery process. we then theoretically prove the feasibility of our method in learning from hidden labels. Furthermore, we theoretically provide the recovery bound of our method, demonstrating that it is feasible to learn from hidden labels. Extensive experiments validate advantages of our method against other IncomLDL and LDL methods in HidLDL.

References

Box, J. F. 1987. Guinness, Gosset, Fisher, and small samples. *Statistical science*, 45–52.

Boyd, S.; Parikh, N.; Chu, E.; Peleato, B.; Eckstein, J.; et al. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1): 1–122.

Cai, J.-F.; Candès, E. J.; and Shen, Z. 2010. A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization*, 20(4): 1956–1982.

Chen, S.; Wang, J.; Chen, Y.; Shi, Z.; Geng, X.; and Rui, Y. 2020. Label Distribution Learning on Auxiliary Label Space Graphs for Facial Expression Recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 13981–13990. Computer Vision Foundation / IEEE.

Gao, B.; Zhou, H.; Wu, J.; and Geng, X. 2018. Age Estimation Using Expectation of Label Distribution Learning. In Lang, J., ed., *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, 712–718. ijcai.org.

Geng, X. 2016. Label Distribution Learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7): 1734–1748.

Geng, X.; and Hou, P. 2015. Pre-release prediction of crowd opinion on movies by label distribution learning. In *IJCAI*, 3511–3517. Citeseer.

Geng, X.; Yin, C.; and Zhou, Z.-H. 2013. Facial age estimation by learning from label distributions. *IEEE transactions on pattern analysis and machine intelligence*, 35(10): 2401–2412.

Jia, X.; Li, W.; Liu, J.; and Zhang, Y. 2018. Label distribution learning by exploiting label correlations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Jia, X.; Li, Z.; Zheng, X.; Li, W.; and Huang, S.-J. 2019a. Label distribution learning with label correlations on local samples. *IEEE Transactions on Knowledge and Data Engineering*, 33(4): 1619–1631.

Jia, X.; Qin, T.; Lu, Y.; and Li, W. 2023. Adaptive weighted ranking-oriented label distribution learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Jia, X.; Shen, X.; Li, W.; Lu, Y.; and Zhu, J. 2021. Label distribution learning by maintaining label ranking relation. *IEEE Transactions on Knowledge and Data Engineering*, 35(2): 1695–1707.

Jia, X.; Zheng, X.; Li, W.; Zhang, C.; and Li, Z. 2019b. Facial emotion distribution learning by exploiting low-rank label correlations locally. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 9841–9850.

Jia, Y.; Liu, H.; Hou, J.; Kwong, S.; and Zhang, Q. 2022. Semisupervised Affinity Matrix Learning via Dual-Channel Information Recovery. *IEEE Trans. Cybern.*, 52(8): 7919–7930.

Jia, Y.; Tang, J.; and Jiang, J. 2024. Label Distribution Learning from Logical Label. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, 4228–4236.

Jia, Y.; Tao, S.; Wang, R.; and Wang, Y. 2024. Ensemble Clustering via Co-Association Matrix Self-Enhancement. *IEEE Trans. Neural Networks Learn. Syst.*, 35(8): 11168–11179.

Kou, Z.; Qin, S.; Wang, H.; Wang, J.; Xie, M.; Chen, S.; Jia, Y.; Liu, T.; Sugiyama, M.; and Geng, X. 2025a. Label Distribution Learning with Biased Annotations Assisted by Multi-Label Learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2025, Montreal, Canada, August 16-22, 2025*, 5545–5553.

Kou, Z.; Wang, J.; Jia, Y.; Liu, B.; and Geng, X. 2025b. Instance-Dependent Inaccurate Label Distribution Learning. *IEEE Trans. Neural Networks Learn. Syst.*, 36(1): 1425–1437.

Li, S.; and Deng, W. 2019. Blended emotion in-the-wild: Multi-label facial expression recognition using crowd-sourced annotations and deep locality feature learning. *International Journal of Computer Vision*, 127(6): 884–906.

Li, X.; and Chen, S. 2024. No regularization is needed: efficient and effective incomplete label distribution learning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 4470–4478.

Li, Y.; Li, Z.; and Jia, Y. 2025. Complementary Label Learning with Positive Label Guessing and Negative Label Enhancement. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Liang, L.; Lin, L.; Jin, L.; Xie, D.; and Li, M. 2018. SCUT-FBP5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction. In *2018 24th International conference on pattern recognition (ICPR)*, 1598–1603. IEEE.

Liu, W.; Wang, H.; Shen, X.; and Tsang, I. W. 2021. The emerging trends of multi-label learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(11): 7955–7974.

Ma, H.; Lu, N.; Mei, J.; Guan, T.; Zhang, Y.; and Geng, X. 2023. Label distribution learning for scene text detection. *Frontiers Comput. Sci.*, 17(6): 176339.

Peng, K.-C.; Chen, T.; Sadovnik, A.; and Gallagher, A. C. 2015. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 860–868.

Ren, Y.; and Geng, X. 2017. Sense Beauty by Label Distribution Learning. In Sierra, C., ed., *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, 2648–2654. ijcai.org.

Tang, J.; and Jia, Y. 2025. Concentration Distribution Learning from Label Distributions. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, Canada, July 13-19, 2025*.

Tsoumakas, G.; and Katakis, I. 2008. Multi-label classification. *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications: Concepts, Methodologies, Tools, and Applications*, 3: 64.

Wang, J.; and Geng, X. 2021. Label distribution learning by exploiting label distribution manifold. *IEEE transactions on neural networks and learning systems*, 34(2): 839–852.

Wang, J.; Zhang, F.; Jia, X.; Wang, X.; Zhang, H.; Ying, S.; Wang, Q.; Shi, J.; and Shen, D. 2022. Multi-Class ASD Classification via Label Distribution Learning with Class-Shared and Class-Specific Decomposition. *Medical Image Analysis*, 75: 102294.

Wang, W.; and Carreira-Perpinán, M. A. 2013. Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. *arXiv preprint arXiv:1309.1541*.

Wen, C.; Zhang, X.; Yao, X.; and Yang, J. 2023. Ordinal Label Distribution Learning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 23424–23434.

Xie, K. Y.; Wang, J.; Jia, Y.; Shi, B.; and Geng, X. 2026. RankMatch: A Novel Approach to Semi-Supervised Label Distribution Learning Leveraging Inter-label Correlations. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Xu, M.; and Zhou, Z.-H. 2017. Incomplete Label Distribution Learning. In *IJCAI*, 3175–3181.

Xu, S.; Shang, L.; Shen, F.; Yang, X.; and Pedrycz, W. 2025. Incomplete label distribution learning via label correlation decomposition. *Information Fusion*, 113: 102600.

Yang, X.; Gao, B.-B.; Xing, C.; Huo, Z.-W.; Wei, X.-S.; Zhou, Y.; Wu, J.; and Geng, X. 2015. Deep label distribution learning for apparent age estimation. In *Proceedings of the IEEE international conference on computer vision workshops*, 102–108.

Zhang, M.; and Zhou, Z. 2014. A Review on Multi-Label Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8): 1819–1837.

Zhu, X. 2005. *Semi-supervised learning with graphs*. Carnegie Mellon University.

Appendix

A Statistics of Datasets

Dataset	n	d	m
Yeast-alpha	2465	24	18
Yeast-cdc	2465	24	15
Yeast-cold	2465	24	4
Yeast-dtt	2465	24	4
Yeast-elu	2465	24	14
Yeast-spo	2465	24	6
SJAFFE	213	243	6
Natural Scene	2000	294	9
Movie	7755	1869	5
SBU3DFE	2500	243	6
Emotion6	1980	1000	7
RAF-ML	4908	200	6

Table A1: Statistics of the 12 datasets, where n, d and m represent the number of samples, the dimension of features, and the number of labels.

Table A1 shows the statistics of 12 datasets. Due to space limitations, we only use the abbreviations of the datasets in the paper. The following are the abbreviations of each dataset.

- Yeast-alpha $\rightarrow$ alpha
- Yeast-cdc $\rightarrow$ cdc
- Yeast-cold $\rightarrow$ cold
- Yeast-dtt $\rightarrow$ dtt
- Yeast-elu $\rightarrow$ elu
- Yeast-spo $\rightarrow$ spo
- SJAFFE $\rightarrow$ SJAFFE
- Natural Scene$\rightarrow$ Scene
- Movie $\rightarrow$ Movie
- SBU3DFE $\rightarrow$ SBU
- Emotion6 $\rightarrow$ Emo
- RAF-ML $\rightarrow$ RAF

B Recovery Results

More Recovery Results Table A2 and Table A3 show the recovery results for the evaluation metric Chebyshev at missing rate ω = 50% and the evaluation metric Clark at missing rate ω = 80%, respectively. As shown in table A2 and table A3, our method outperform comparison methods by a wide margin. Specifically, in Table A2, our method ranks first 11 times out of 12 datasets. Similarly, in Table A3, it ranks first 10 times out of 12.

B.1 Recovery Results Versus Missing Rates

Fig. A1 shows the results of all methods in the recovery experiment under different missing rates. In all of the 4 datasets, our method achieves better performance under different missing rates, which clearly indicate that our method is superior to the compared methods. Furthermore, as the missing rate decreases, the superiority of our method become more apparent. Our method has an outstanding advantage when the missing rate is between 40% and 60%.

C Predictive Results

C.1 More Predictive Results

Table A4 and Table A5 show the predictive results for the evaluation metrics of Chebyshev at missing rate ω = 50% and Clark at missing rate ω = 80%, respectively. In these two tables, our method outperforms comparison methods 21 out of 24 times, resulting in a leading rate of 87.5%.

C.2 Predictive Results Versus Missing Rates

Fig. A2 shows the results of 4 methods in the predictive experiment under different missing rates, and the results clearly indicate that our method is less affected by missing rates. Across all four datasets, it is observed that as the missing rate increases, the performance of our method declines at a relatively slow pace. However, when it comes to InLDL-a, on Movie and Yeast-cold, the InLDL-a method exhibits a more rapid rate of performance degradation.

D Further Ablation Study

We find that in most cases, the constraint we add to maintain the proportion of observed labels is the most effective. However, as shown in Table A6, in rare cases, the trace norm term is more effective, as evidenced by the performance of "w/o $Cons$" being better than "w/o". All of these occur when the missing rate is very high, reaching 80%. This is because when the missing rate is too high, there may be only one observable label in a sample, making it impossible to utilize proportional information. So the trace norm term can play a important role in extremely high missing rates.

E Convergence Analysis

Here we provide a proof of convergence of Algorithm 1.

If $\{\mathbf{\Lambda}_k, \mathbf{\Lambda}'_k\}$ are bounded and

$$\sum_{k=0}^{\infty}\left(\|\mathbf{\Lambda}_{k+1}-\mathbf{\Lambda}_k\|_{\mathrm{F}}^2+\|\mathbf{\Lambda}'_{k+1}-\mathbf{\Lambda}'_k\|_{\mathrm{F}}^2\right)<\infty, \tag{20}$$

then the sequence $\{\mathbf{Z}, \mathbf{\Lambda}, \mathbf{\Lambda}'\}$ will converge.

Optimization Problem:

$$\begin{aligned}\min_{\mathbf{A},\mathbf{B},\mathbf{D}} \quad & \frac{1}{2}\operatorname{tr}\left(\mathbf{D}^{\top}\mathbf{G}\mathbf{D}\right)+\alpha\|\mathbf{A}\|_{*}\\ \text{s.t.} \quad & \mathbf{D}\times\mathbf{1}_m=\mathbf{1}_n,\ \mathbf{D}\geq\mathbf{0}_{n\times m},\ \mathbf{B}\in Cons,\\ & \mathbf{D}-\mathbf{A}=\mathbf{0},\ \mathbf{D}-\mathbf{B}=\mathbf{0}.\end{aligned} \tag{21}$$

Augmented Lagrange multiplier equation (assume $\mathbf{Z} = \{\mathbf{A}, \mathbf{B}, \mathbf{D}\}$):

$$\begin{aligned}\min_{\mathbf{A},\mathbf{B},\mathbf{D}} \mathcal{L}\left(\mathbf{Z},\mathbf{\Lambda},\mathbf{\Lambda}'\right)=&\frac{1}{2}\operatorname{tr}\left(\mathbf{D}^{\top}\mathbf{G}\mathbf{D}\right)+\alpha\|\mathbf{A}\|_{*}\\ &+\langle\mathbf{\Lambda},\mathbf{D}-\mathbf{A}\rangle+\frac{\rho}{2}\|\mathbf{D}-\mathbf{A}\|_{\mathrm{F}}^2\\ &+\langle\mathbf{\Lambda}',\mathbf{D}-\mathbf{B}\rangle+\frac{\rho}{2}\|\mathbf{D}-\mathbf{B}\|_{\mathrm{F}}^2\\ \text{s.t. } \mathbf{D}\times\mathbf{1_m}=&\mathbf{1_n},\mathbf{D}\geq\mathbf{0}_{n\times m},\mathbf{B}\in Cons.\end{aligned} \tag{22}$$

The objective function of Eq. (22) is strongly convex w.r.t. $\mathbf{A}, \mathbf{B}, \mathbf{D}$, because

$$\begin{aligned} \mathcal{L}\left(\mathbf{Z}, \boldsymbol{\Lambda}, \boldsymbol{\Lambda}'\right) = & \frac{1}{2} \operatorname{tr}\left(\mathbf{D}^{\top} \mathbf{G} \mathbf{D}\right) + \alpha \|\mathbf{A}\|_{*} \\ & + \frac{\rho}{2} \|\mathbf{D} - \mathbf{A} + \frac{\boldsymbol{\Lambda}}{\rho}\|_{\mathrm{F}}^{2} - \frac{1}{2\rho} \|\boldsymbol{\Lambda}\|_{\mathrm{F}}^{2} \\ & + \frac{\rho}{2} \|\mathbf{D} - \mathbf{B} + \frac{\boldsymbol{\Lambda}'}{\rho}\|_{\mathrm{F}}^{2} - \frac{1}{2\rho} \|\boldsymbol{\Lambda}'\|_{\mathrm{F}}^{2}. \end{aligned} \tag{23}$$

Consequently, we have

$$\mathcal{L}\left(\mathbf{A} + \Delta\mathbf{A}\right) - \mathcal{L}\left(\mathbf{A}\right) \geq \partial_{\mathbf{A}} \mathcal{L}\left(\mathbf{A}\right)^{\top} \Delta\mathbf{A} + \rho \|\Delta\mathbf{A}\|_{\mathrm{F}}^{2}. \tag{24}$$

For $\mathbf{A}^{*}$ to be the minimizer of $\mathcal{L}(\mathbf{A}^{*})$, then we have

$$\partial_{\mathbf{A}} \mathcal{L}\left(\mathbf{A}^{*}\right)^{\top} \Delta\mathbf{A} \geq 0. \tag{25}$$

Combining Eq. (24) and Eq. (25), we have

$$\mathcal{L}\left(\mathbf{A}_{k}\right) - \mathcal{L}\left(\mathbf{A}_{k+1}\right) \geq \rho \|\mathbf{A}_{k} - \mathbf{A}_{k+1}\|_{\mathrm{F}}^{2}. \tag{26}$$

Similarly,

$$\mathcal{L}\left(\mathbf{B}_{k}\right) - \mathcal{L}\left(\mathbf{B}_{k+1}\right) \geq \rho \|\mathbf{B}_{k} - \mathbf{B}_{k+1}\|_{\mathrm{F}}^{2}. \tag{27}$$

$$\mathcal{L}\left(\mathbf{D}_{k}\right) - \mathcal{L}\left(\mathbf{D}_{k+1}\right) \geq \rho \|\mathbf{D}_{k} - \mathbf{D}_{k+1}\|_{\mathrm{F}}^{2}. \tag{28}$$

Let $\nu = \min\{1, \rho\}$ and combine Eq. (26), Eq. (27) and Eq. (28), we obtain

$$\begin{aligned} & \mathcal{L}\left(\mathbf{Z}_{k}, \boldsymbol{\Lambda}_{k}, \boldsymbol{\Lambda}_{k}'\right) - \mathcal{L}\left(\mathbf{Z}_{k+1}, \boldsymbol{\Lambda}_{k+1}, \boldsymbol{\Lambda}_{k+1}'\right) \\ & = \mathcal{L}\left(\mathbf{Z}_{k}, \boldsymbol{\Lambda}_{k}, \boldsymbol{\Lambda}_{k}'\right) - \mathcal{L}\left(\mathbf{Z}_{k+1}, \boldsymbol{\Lambda}_{k}, \boldsymbol{\Lambda}_{k}'\right) \\ & \quad + \mathcal{L}\left(\mathbf{Z}_{k+1}, \boldsymbol{\Lambda}_{k}, \boldsymbol{\Lambda}_{k}'\right) - \mathcal{L}\left(\mathbf{Z}_{k+1}, \boldsymbol{\Lambda}_{k+1}, \boldsymbol{\Lambda}_{k+1}'\right) \\ & \geq \nu \|\mathbf{Z}_{k} - \mathbf{Z}_{k+1}\|_{\mathrm{F}}^{2} - \frac{1}{\rho}\left(\|\boldsymbol{\Lambda}_{k} - \boldsymbol{\Lambda}_{k+1}\|_{\mathrm{F}}^{2} + \|\boldsymbol{\Lambda}_{k}' - \boldsymbol{\Lambda}_{k+1}'\|_{\mathrm{F}}^{2}\right) \\ & \geq \nu \|\mathbf{Z}_{k} - \mathbf{Z}_{k+1}\|_{\mathrm{F}}^{2} - \frac{1}{\nu}\left(\|\boldsymbol{\Lambda}_{k} - \boldsymbol{\Lambda}_{k+1}\|_{\mathrm{F}}^{2} + \frac{1}{\nu}\|\boldsymbol{\Lambda}_{k}' - \boldsymbol{\Lambda}_{k+1}'\|_{\mathrm{F}}^{2}\right). \end{aligned} \tag{29}$$

Recalling that $\mathcal{L}\left(\mathbf{Z}, \boldsymbol{\Lambda}, \boldsymbol{\Lambda}'\right)$ is bounded below, we have

$$\sum_{k=0}^{\infty} \nu \|\mathbf{Z}_{k} - \mathbf{Z}_{k+1}\|_{\mathrm{F}}^{2} - \sum_{k=0}^{\infty} \frac{1}{\nu}\left(\|\boldsymbol{\Lambda}_{k} - \boldsymbol{\Lambda}_{k+1}\|_{\mathrm{F}}^{2} + \frac{1}{\nu}\|\boldsymbol{\Lambda}_{k}' - \boldsymbol{\Lambda}_{k+1}'\|_{\mathrm{F}}^{2}\right) < \infty. \tag{30}$$

We have assumed the second term is bounded (in Eq. (20)), so we can immediately get

$$\sum_{k=0}^{\infty} \nu \|\mathbf{Z}_{k} - \mathbf{Z}_{k+1}\|_{\mathrm{F}}^{2} < \infty. \tag{31}$$

Recalling $\mathbf{Z} = \{\mathbf{A}, \mathbf{B}, \mathbf{D}\}$,we get the **conclusion**: As the training progresses, $\mathbf{A}, \mathbf{B}, \mathbf{D}$ in Eq. (9) will converge. i.e., $\mathbf{A}_{k+1} - \mathbf{A}_{k} \to 0$.

F Proof of Theorem 1

Theorem 1. *Under certain assumptions, the error of the scaling coefficient can be expressed as*

$$k_{i}^{g} - k_{i} = \frac{1 - \Sigma_{k} d_{ik} M_{ik}}{\Sigma_{k} d_{ik}^{g} M_{ik}}. \tag{32}$$

Proof:

From Eq. (3) and Eq. (15) in the paper, we have

$$k_{i}^{g} = \frac{D_{ij}^{o}}{D_{ij}^{g}} = \frac{1}{\sum_{k=1}^{m} D_{ik}^{g} \cdot M_{ik}} \tag{33}$$

and

$$k_{i} = \frac{\sum_{k=1}^{m}\left(\rho D_{ik} + \Lambda_{ik}'\right) \cdot D_{ik}^{o} \cdot M_{ik}}{\rho \sum_{k=1}^{m} (D_{ik}^{o})^{2} \cdot M_{ik}}. \tag{34}$$

Consider the error of the scaling coefficients:

$$\begin{aligned} k_{i}^{g} - k_{i} & = \frac{1}{\sum_{k=1}^{m} D_{ik}^{g} M_{ik}} - \frac{\sum_{k=1}^{m}\left(D_{ik} + \frac{\Lambda_{ik}'}{\rho}\right) D_{ik}^{o} M_{ik}}{\sum_{k=1}^{m} (D_{ik}^{o})^{2} M_{ik}} \\ & \leq \frac{1}{\sum_{k=1}^{m} D_{ik}^{g} M_{ik}} - \frac{\sum_{k=1}^{m}\left(D_{ik} + \frac{\Lambda_{ik}'}{\rho}\right) M_{ik}}{\sum_{k=1}^{m} D_{ik}^{o} M_{ik}} \end{aligned} \tag{35}$$

Assume that σ_{i}, i.e. the sum of the labels in i-th instance that are masked is relatively small, there holds $D_{ik}^{g} \approx D_{ik}^{o}$, and the above equation can be further rewritten as

$$\begin{aligned} k_{i}^{g} - k_{i} & \leq \frac{1}{\sum_{k=1}^{m} D_{ik}^{g} M_{ik}} - \frac{\sum_{k=1}^{m}\left(D_{ik} + \frac{\Lambda_{ik}'}{\rho}\right) M_{ik}}{\sum_{k=1}^{m} D_{ik}^{g} M_{ik}} \\ & = \frac{1 - \sum_{k=1}^{m} D_{ik} M_{ik} - \sum_{k=1}^{m} \frac{\Lambda_{ik}'}{\rho}}{\sum_{k=1}^{m} D_{ik}^{g} \, M_{ik}} \end{aligned} \tag{36}$$

In the above equation, $\frac{\Lambda_{ik}'}{\rho}$ can be considered as the gradient of D_{ik}, and should be equal to zero after optimization. So the final equation becomes

$$k_{i}^{g} - k_{i} \leq \frac{1 - \sum_{k=1}^{m} D_{ik} M_{ik}}{\sum_{k=1}^{m} D_{ik}^{g} M_{ik}}, \tag{37}$$

so **Theorem 1** in the paper is proved.

Datasets	Ours	InLDL-a	WInLDL	SA-IIS	LDL-LRR	PT-Bayes	LDL-DPA
alpha	**.0106 ± .0001**	.0136 ± .0000	.0171 ± .0003	.0169 ± .0001	.0138 ± .0001	.0356 ± .0015	.0177 ± .0003
cdc	**.0117 ± .0001**	.0164 ± .0001	.0208 ± .0002	.0204 ± .0003	.0165 ± .0002	.0394 ± .0017	.0214 ± .0002
cold	**.0386 ± .0001**	.0520 ± .0003	.0596 ± .0010	.0625 ± .0011	.0536 ± .0010	.0993 ± .0110	.0659 ± .0019
dtt	**.0273 ± .0001**	.0370 ± .0003	.0469 ± .0013	.0498 ± .0015	.0390 ± .0012	.0932 ± .0082	.0538 ± .0008
elu	**.0120 ± .0001**	.0163 ± .0000	.0213 ± .0005	.0208 ± .0005	.0166 ± .0001	.0411 ± .0029	.0217 ± .0006
spo	**.0403 ± .0009**	.0584 ± .0001	.0634 ± .0007	.0638 ± .0012	.0589 ± .0003	.0913 ± .0027	.0667 ± .0008
SJAFFE	**.0785 ± .0022**	.1143 ± .0078	.1882 ± .0108	.1089 ± .0048	.1203 ± .0011	.1256 ± .0057	.2657 ± .0149
Scene	**.2760 ± .0035**	.3276 ± .0025	.3371 ± .0029	.3299 ± .0025	.3483 ± .0011	.6317 ± .0128	.3330 ± .0032
Movie	**.0916 ± .0003**	.1301 ± .0008	.1822 ± .0013	.2298 ± .0017	.1357 ± .0014	.7735 ± .0249	.2090 ± .0072
SBU	**.0982 ± .0011**	.1272 ± .0012	.1134 ± .0009	.1322 ± .0004	.1368 ± .0004	.1331 ± .0009	.1291 ± .0010
Emo	**.2376 ± .0025**	.3061 ± .0011	.3121 ± .0024	.3201 ± .0022	.3204 ± .0019	.3220 ± .0020	.3166 ± .0020
RAF	.2715 ± .0022	.3923 ± .0016	.6962 ± .0009	**.2457 ± .0032**	.2707 ± .0012	.2963 ± .0155	.2462 ± .0027

Table A2: Chebyshev (the lower the better) results for the **recovery** setting on all datasets when missing rate $\omega = 50\%$. InLDL-a is the abbreviation for IncomLDL-admm. The value is shown in mean±std form. Bold and underlined indicate the best and second best results, respectively.

Dataset	Ours	InLDL-a	WInLDL	SA-IIS	LDL-LRR	PT-Bayes	LDL-DPA
alpha	**0.1945 ± .0010**	0.2146 ± .0004	0.9080 ± .0169	0.4548 ± .0091	0.2382 ± .0090	0.9399 ± .0663	0.5016 ± .0240
cdc	**0.1966 ± .0019**	0.2187 ± .0004	0.7998 ± .0576	0.4209 ± .0171	0.2398 ± .0063	0.8023 ± .0565	0.4809 ± .0120
cold	**0.1349 ± .0006**	0.1453 ± .0019	0.3034 ± .0113	0.1977 ± .0083	0.1515 ± .0027	0.3379 ± .0322	0.2059 ± .0094
dtt	**0.0958 ± .0004**	0.1009 ± .0007	0.2632 ± .0138	0.1551 ± .0091	0.1129 ± .0055	0.3060 ± .0211	0.1727 ± .0090
elu	**0.1830 ± .0008**	0.2039 ± .0006	0.7945 ± .0274	0.4082 ± .0077	0.2272 ± .0062	0.7924 ± .0535	0.4509 ± .0114
spo	**0.2315 ± .0009**	0.2548 ± .0015	0.4741 ± .0157	0.3181 ± .0046	0.2623 ± .0049	0.4991 ± .0385	0.3301 ± .0083
SJAFFE	0.4091 ± .0144	0.4296 ± .0026	1.5932 ± .0125	0.4966 ± .0196	0.4346 ± .0158	0.5585 ± .0355	**1.5829 ± .0480**
Scene	**2.4534 ± .0009**	2.4718 ± .0009	2.5182 ± .0039	2.4707 ± .0013	2.4707 ± .0010	2.4965 ± .0341	2.4968 ± .0048
Movie	**0.6130 ± .0029**	0.7146 ± .0077	1.0812 ± .0070	1.2111 ± .0038	0.7123 ± .0081	1.5170 ± .0846	0.9781 ± .0130
SBU	**0.3954 ± .0023**	0.4129 ± .0018	0.6107 ± .0103	0.4134 ± .0034	0.4125 ± .0006	0.4198 ± .0040	0.4393 ± .0037
Emo	**1.6046 ± .0024**	1.6895 ± .0042	1.7651 ± .0092	1.7191 ± .0069	1.7025 ± .0058	1.7355 ± .0297	1.7238 ± .0046
RAF	1.6233 ± .0021	2.1085 ± .0050	2.2144 ± .0081	1.5751 ± .0076	**1.5723 ± .0015**	1.5881 ± .0183	1.5765 ± .0028

Table A3: Clark (the lower the better) results for the **recovery** setting on all datasets when missing rate $\omega = 80\%$. InLDL-a is the abbreviation for IncomLDL-admm. The value is shown in mean±std form. Bold and underlined indicate the best and second best results, respectively.

Dataset	Ours	InLDL-a	WInLDL	SA-IIS	LDL-LRR	PT-Bayes	LDL-DPA
alpha	**.0135 ± .0001**	.0136 ± .0000	.0178 ± .0002	.0169 ± .0002	.0136 ± .0001	.0339 ± .0029	.0187 ± .0003
cdc	**.0164 ± .0001**	.0164 ± .0001	.0223 ± .0004	.0208 ± .0003	.0168 ± .0002	.0440 ± .0035	.0231 ± .0003
cold	**.0501 ± .0001**	.0510 ± .0005	.0629 ± .0026	.0612 ± .0017	.0533 ± .0009	.1114 ± .0194	.0679 ± .0028
dtt	**.0350 ± .0001**	.0360 ± .0003	.0507 ± .0030	.0507 ± .0033	.0414 ± .0028	.0983 ± .0105	.0583 ± .0048
elu	**.0159 ± .0000**	.0160 ± .0000	.0226 ± .0006	.0210 ± .0006	.0164 ± .0002	.0444 ± .0048	.0234 ± .0007
spo	**.0592 ± .0001**	.0593 ± .0003	.0674 ± .0015	.0649 ± .0017	.0599 ± .0005	.0934 ± .0041	.0708 ± .0016
SJAFFE	**.1107 ± .0040**	.1197 ± .0049	.2595 ± .0043	.1192 ± .0054	.1211 ± .0030	.1344 ± .0045	.2303 ± .0206
Scene	**.3371 ± .0036**	.3492 ± .0034	.3598 ± .0047	.3496 ± .0028	.3548 ± .0017	.6087 ± .0196	.3490 ± .0039
Movie	**.1270 ± .0004**	.1419 ± .0005	.2558 ± .0029	.2815 ± .0037	.1390 ± .0014	.7793 ± .0134	.2186 ± .0042
SBU3DFE	.1275 ± .0012	.1330 ± .0011	**.1238 ± .0015**	.1374 ± .0011	.1398 ± .0005	.1367 ± .0007	.1340 ± .0016
Emo	**.3014 ± .0012**	.3121 ± .0017	.3169 ± .0029	.3727 ± .0045	.3223 ± .0035	.3107 ± .0032	.3283 ± .0054
RAF	.2819 ± .0018	**.2196 ± .0014**	.5538 ± .0055	.2670 ± .0035	.2714 ± .0021	.2797 ± .0059	.2546 ± .0019

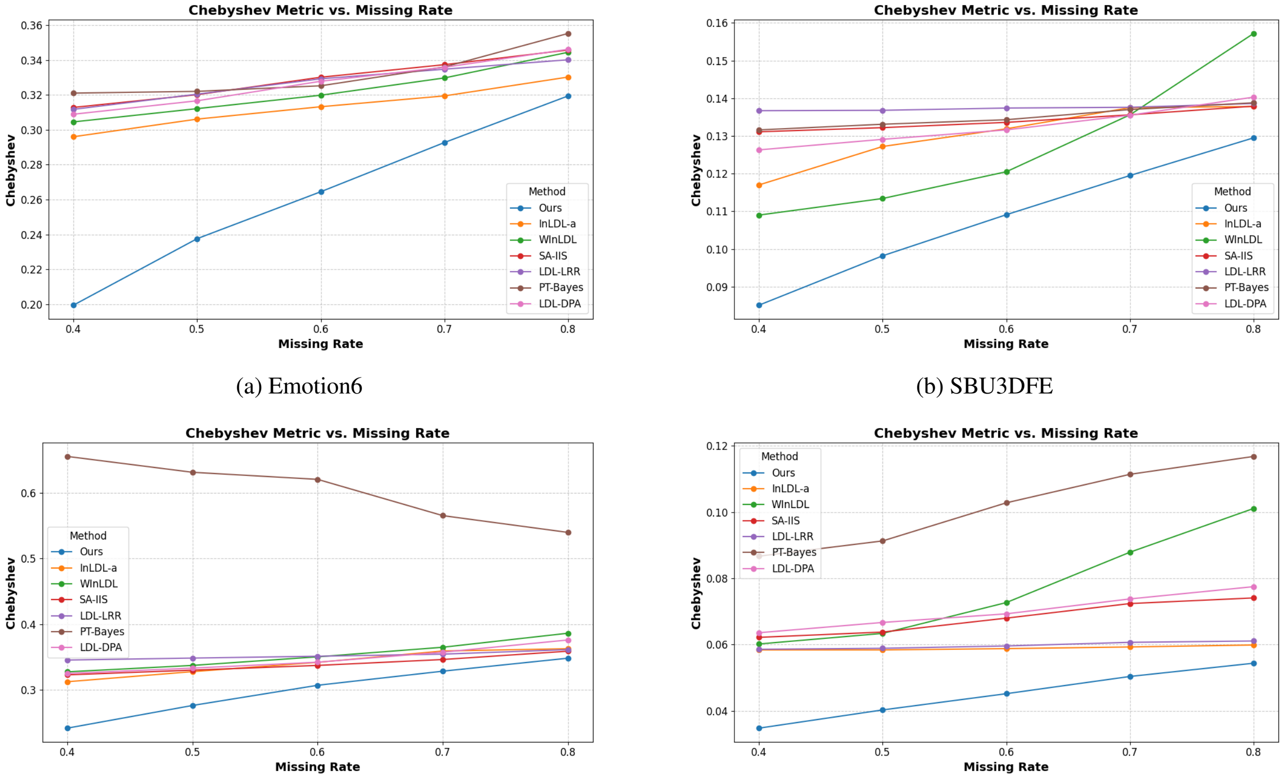
Table A4: Chebyshev (the lower the better) results for the **predictive** setting on all datasets when missing rate $\omega = 50\%$. InLDL-a is the abbreviation for IncomLDL-admm. The value is shown in mean±std form. Bold and underlined indicate the best and second best results, respectively.

Dataset	Ours	InLDL-a	WInLDL	SA-IIS	LDL-LRR	PT-Bayes	LDL-DPA
alpha	**0.2136 ± .0015**	0.2159 ± .0009	0.8084 ± .0266	0.3836 ± .0090	0.2390 ± .0090	0.8664 ± .0614	0.5579 ± .0198
cdc	**0.2158 ± .0005**	0.2181 ± .0006	0.7939 ± .0373	0.3900 ± .0111	0.2486 ± .0076	0.8847 ± .0291	0.5273 ± .0160
cold	**0.1354 ± .0005**	0.1403 ± .0012	0.3050 ± .0157	0.1769 ± .0096	0.1492 ± .0059	0.3430 ± .0498	0.2244 ± .0146
dtt	**0.0962 ± .0003**	0.0976 ± .0003	0.2897 ± .0121	0.1520 ± .0099	0.1176 ± .0048	0.3037 ± .0543	0.1872 ± .0089
elu	**0.1954 ± .0004**	0.1989 ± .0011	0.7379 ± .0249	0.3607 ± .0101	0.2271 ± .0113	0.8220 ± .0632	0.5033 ± .0204
spo	**0.1276 ± .0702**	0.1293 ± .0711	0.4749 ± .0269	0.3046 ± .0082	0.2684 ± .0048	0.5092 ± .0755	0.3572 ± .0157
SJAFFE	**0.4306 ± .0108**	0.4387 ± .0043	1.3074 ± .0488	0.4767 ± .0280	0.4481 ± .0166	0.5976 ± .0444	1.4876 ± .1105
Scene	**2.4735 ± .0017**	2.4807 ± .0016	2.5281 ± .0034	2.4840 ± .0019	2.4809 ± .0014	2.5186 ± .0187	2.5204 ± .0054
Movie	**0.6159 ± .0043**	0.6875 ± .0216	1.0864 ± .0141	1.2592 ± .0254	0.7175 ± .0070	1.4884 ± .1344	0.9911 ± .0249
SBU	**0.4085 ± .0007**	0.4214 ± .0055	0.5126 ± .0302	0.4189 ± .0042	0.4183 ± .0031	0.4260 ± .0065	0.4483 ± .0099
Emo	**1.6874 ± .0155**	1.7139 ± .0366	1.7416 ± .0195	1.8348 ± .0213	1.7227 ± .0077	1.6961 ± .0157	1.7497 ± .0078
RAF	1.5887 ± .0014	1.5872 ± .0143	1.7471 ± .0078	1.5817 ± .0089	**1.5737 ± .0046**	1.5737 ± .0024	1.5856 ± .0064

Table A5: Clark (the lower the better) results for the **predictive** setting on all datasets when missing rate ω = 80%. InLDL-a is the abbreviation for IncomLDL-admm. The value is shown in mean±std form. Bold and und erlined indicate the best and second best results, respectively.

Table A6: Ablation Results when missing rate $\omega = 80\%$ on 4 datasets. ↑ (↓) indicates the higher (lower) the better.

Metric	Method	Recovery				Predictive			
		Yeast-dtt	Yeast-elu	Yeast-spo	Emotion6	Yeast-dtt	Yeast-elu	Yeast-spo	Emotion6
$clark \downarrow$	Ours	**0.0958**	**0.1834**	**0.2299**	**1.6009**	**0.0960**	**0.1957**	**0.2532**	**1.6635**
	w/o $Cons$	0.1014	0.2043	0.2559	1.6789	0.0979	0.1994	0.2565	1.6664
	w/o TN	0.1871	0.1943	0.2653	1.8921	0.1169	0.1999	0.2638	1.7109
$cosine \uparrow$	Ours	**0.9942**	**0.9949**	**0.9797**	**0.7081**	**0.9944**	**0.9942**	**0.9763**	**0.7071**
	w/o $Cons$	0.9937	0.9938	0.9747	0.6566	0.9941	0.9941	0.9749	0.6660
	w/o TN	0.9782	0.9943	0.9732	0.4824	0.9918	0.9940	0.9738	0.6290



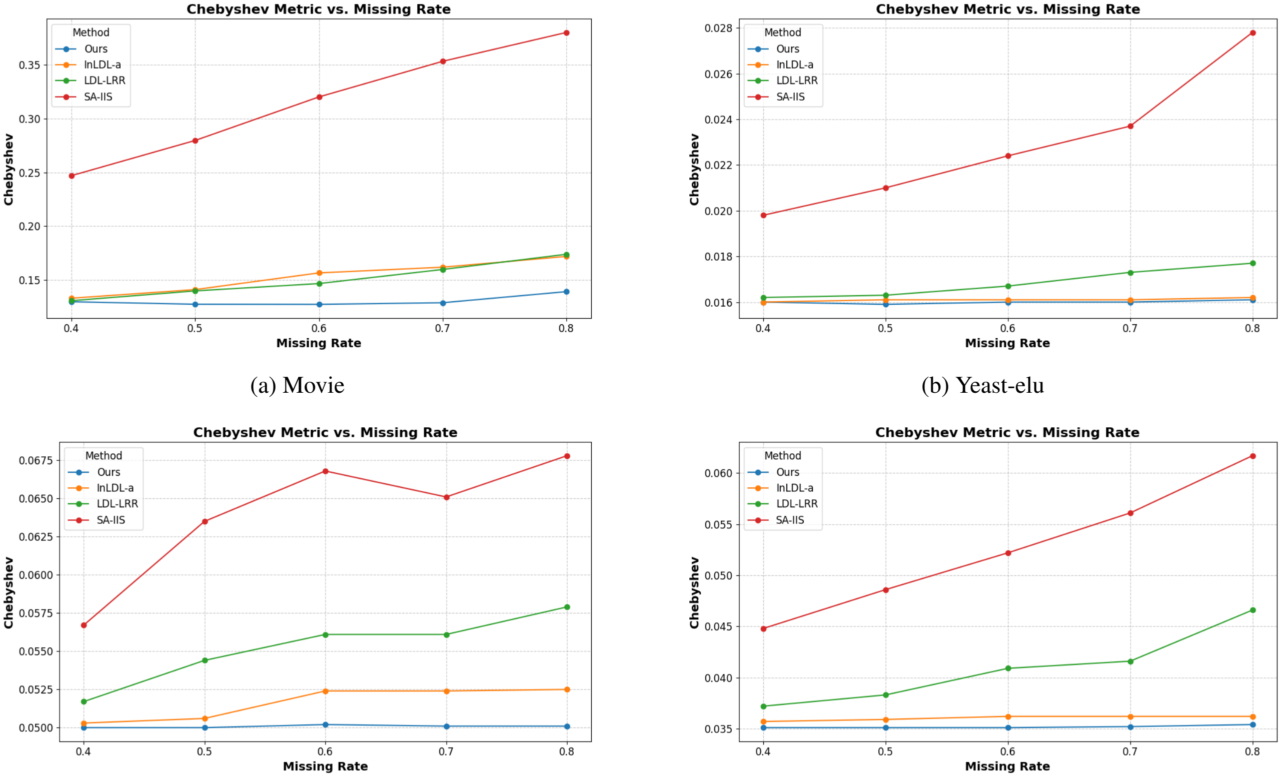
(a) Emotion6

(b) SBU3DFE

(c) Natural Scene

(d) Yeast-spo

Figure A1: Recovery performance comparison under different missing rates.



(a) Movie

(b) Yeast-elu

(c) Yeast-cold

(d) Yeast-dtt

Figure A2: Predictive performance comparison under different missing rates.