Limited-Supervised Multi-Label Learning with Dependency Noise

Yejiang Wang[1], **Yuhai Zhao**[1*], **Zhengkui Wang**[2], **Wen Shan**[3], **Xingwei Wang**[1]

[1]School of Computer Science and Engineering, Northeastern University, China
[2]InfoComm Technology Cluster, Singapore Institute of Technology, Singapore
[3]Singapore University of Social Sciences, Singapore
wyejiang@gmail.com, {zhaoyuhai,wangxw}@mail.neu.edu.cn,
zhengkui.wang@singaporetech.edu.sg, viviensw@suss.edu.sg

Abstract

Limited-supervised multi-label learning (LML) leverages weak or noisy supervision for multi-label classification model training over data with label noise, which contains missing labels and/or redundant labels. Existing studies usually solve LML problems by assuming that label noise is independent of the input features and class labels while ignoring the fact that noisy labels may depend on the input features (*instance-dependent*) and the classes (*label-dependent*) in many real-world applications. In this paper, we propose limited-supervised Multi-label Learning with Dependency Noise (MLDN) to simultaneously identify the instance-dependent and label-dependent label noise by factorizing the noise matrix as the outputs of a mapping from the feature and label representations. Meanwhile, we regularize the problem with the manifold constraint on noise matrix to preserve local relationships and uncover the manifold structure. Theoretically, we bound noise recover error for the resulting problem. We solve the problem by using a first-order scheme based on proximal operator, and the convergence rate of it is at least sublinear. Extensive experiments conducted on various datasets demonstrate the superiority of our proposed method.

Introduction

Multi-label learning (MLL) framework has been widely studied because of its success in fitting multiple semantic meaning problems. In MLL, each instance is associated with multiple labels simultaneously. Due to its wide suitability, multi-label learning techniques have been adopted for many applications, and a number of multi-label learning algorithms have been developed (Xu et al. 2023; Dahiya et al. 2021; Lin 2023; Zhao et al. 2022, 2021).

Typically, multi-label learning methods require the training data with complete and accurate label information. However, in the real-world environment, the labels or annotations are often noisy and imperfect, where labels are usually missing or/and noisy in the training set. To model this problem, a new setting called *limited-supervised multi-label learning* (LML) has attracted enormous attention. LML assumes that there might be incompletely-labeled data and redundantly-labeled data in the dataset. The former implies that only a subset of ground-truth labels can be obtained in training step, which is called multi-label learning with missing labels (MLML) (Wu et al. 2014; Kumar and Rastogi 2022); The latter assumes that the labels assigned to the training samples may have redundantly irrelevant labels, which named partial multi-label learning (PML) (Xie and Huang 2018; Gong, Yuan, and Bao 2022). The missing labels and redundant labels can be treated as label noise in limited-supervised multi-label learning setting.

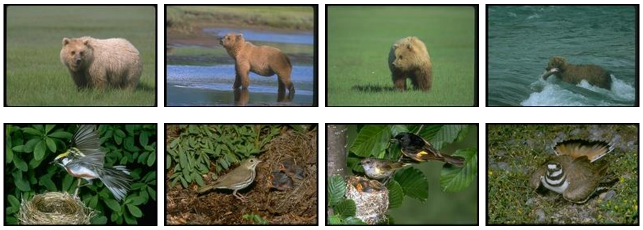


Figure 1: Examples of label Bear (first row) and Bird (second row) in data corel-5k. It is problematic to assume a same probability of mislabeling for diverse samples and labels.

There are *two* main types of methods for dealing with the label noise in limited-supervised multi-label learning. 1) Implicit noise elimination: These works implicitly correct the noisy labels before learning. For example, (Xu et al. 2014; Sun et al. 2019) employ the low-rank assumption of the label matrix to recover the true label from the assigned label matrix; 2) Explicit noise elimination: Another type of method decomposed explicitly the assigned label matrix into a true label matrix and noise matrix. For example, (Xie and Huang 2021; Li, Lyu, and Feng 2020) maintains a small set of clean data to reduce the noise in the training dataset.

Existing algorithms all assume that the label noise is *not dependent on specific* features and labels (Xie and Huang 2021; Schultheis et al. 2022; Ma and Chen 2021). However, in real-world applications, this assumption may not hold. First, there are a large number of instances with the same label and the instances may have very different feature representations. Thus, the probability of being mislabeled is highly dependent on the specific instance, which is referred as *instance-dependent label noise*. For example, although all the images in the first row of Fig.1 include bear, the second

[*] Corresponding author.

left image has a higher chance to be mislabeled (e.g., into a horse) than the first left one. And, the bear label of rightmost one has a higher probability of being missed as well due to ambiguity. Second, some labels may have higher noise probability than others. The probability of noisy labeling may depend on the specific class label, which is referred as *label-dependent label noise*. As shown in the first and second row of Fig.1, the bird label may have more noise than the bear label, as the birds are more difficult to identify. Note that the term "**dependent**" refers to noise that is only associated with a few instances or labels. This does not imply that any noise that occurs with an instance/label is dependent noise. The key is whether the noise is limited to a specific minority. Therefore, the instance-dependent and label-dependent label noise actually exist in real-world, but have been ignored from existing studies.

To tackle the above issues, in this paper, we propose *limited-supervised Multi-label Learning with Dependency Noise* (MLDN) for handling dependent noise. Specifically, it first decomposes assigned label matrix into ground-truth label matrix and dependent noise matrix, which factorizes the noise matrix as the outputs of a mapping from the feature and label representations and sets group sparsity constraints on them to model instance and label-dependent noise. Second, to preserve local relationships within the feature data and uncover its essential manifold structure, we regularize the minimization problem with the manifold constraint on noise matrix, i.e., neighboring instances should also share a similar set of noises. Thus, based on this manifold constrain, the proposed framework can provide a natural way of exploiting the intrinsic relationship among the data. Finally, a noise recovery error bound is given and the unified prediction model is optimized by first-order proximal strategies, where the convergence rate is at least sublinear. The main contributions in this paper are summarized as follows.

- We propose a novel limited-supervised Multi-label Learning with Dependency Noise (MLDN), which simultaneously identify the instance- and label-dependent label noise, where existing limited-supervised MLL solutions ignore the usage of label and feature information to identify the noise label.
- We regularize the problem with the manifold constraint on noise matrix to preserve local relationships within the features and uncover its manifold structure. We bound the noise recovery error and develop efficient proximal descent algorithms to solve the proposed formulations.
- Extensive experiments have demonstrated that MLDN significantly outperforms the state-of-the-art MLML and PML approaches on various benchmark data sets.

Related Work

Label noise is frequently observed in real-world data (Lin et al. 2023; Li et al. 2023; Wang et al. 2023a,b). Recently, to mitigate this problem, plenty of works have studied different settings of multi-label learning with limited supervision. These algorithms can be roughly categorized into two groups based on the different assumptions about the noise label (Gibaja and Ventura 2014; Liu et al. 2021).

The first line of methods focuses on learning with missing labels (MLML) (Cheng, Qian, and Min 2022; Huang et al. 2021; Schultheis et al. 2022). Most of the MLML methods attempt to complete the missing labels first, and then train the classifiers with the complete labels. For example, in (Xu, Jin, and Zhou 2013), the MLML problem is regarded as a low-rank matrix completion problem with the existence of side features information.

The second line of multi-label learning methods concentrates on addressing the extra redundant labels, which is called partial multi-label learning (PML) (Liu, Jia, and Zhang 2023; Gong, Yuan, and Bao 2022). To identify the extra labels, (Xie and Huang 2018) first utilizes the label confidence to measure the probability of being the ground-truth label for each candidate label, and obtains the ground-truth labels according to label ranking. However, existing methods assume that the label noise is not dependent on specific features and labels, and do not consider the dependent label noise that exists in real-world datasets.

The Proposed Method

Preliminaries

In this paper, we regard both missing and redundant label as noise label. Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be the given training feature matrix for n instances $\mathbf{x}_i$ in the d-dimensional feature space. The corresponding unknown ground-truth label of $\mathbf{X}$ is denoted as matrix $\mathbf{Y} \in \{-1,+1\}^{n\times q}$, where each row $\mathbf{y}_i$ corresponds to an instance and each column corresponds to a label. Here, the element $\mathbf{Y}_{ic} = +1$ indicates the c-th label is relevant to the instance $\mathbf{x}_i$; $\mathbf{Y}_{ic} = -1$, otherwise. And let $\mathbf{\Upsilon} \in \mathfrak{Y}^{n\times q}$ denote the corrupted label matrix with label noises. Generally, $\mathfrak{Y} := \mathbb{R}$ due to the existence of noise. The learning problem we are interested in is to identify noisy labels of corrupted label matrix and find a decision mapping $\mathfrak{F} : \mathbb{R}^{n\times d} \to \{-1,+1\}^{n\times q}$ from the training set $\{\mathbf{X}, \mathbf{\Upsilon}\}$, which should reproduces well the outputs of the instances (i.e., $\mathfrak{F}(\mathbf{X}_{i:}) \approx \mathbf{Y}_{i:}$).

Algorithm

To simultaneously identify the label- and instance- dependent label noise, in this paper, we propose a MLDN method that decomposes the assigned label matrix to ground-truth label matrix and noise matrix, and explicitly model the noise matrix as the outputs of linear mapping from the feature and label representations simultaneously to capture the dependent structure noise. In addition, a manifold constrain is applied on noise matrix to preserve local relationships within the features and uncover its essential manifold structure. To solve the optimization objective effectively, we develop an first-order optimization scheme which minimizes the loss based on proximal operator and the convergence rate of it is at least sublinear.

Predictor Inducing. Our goal is to use the instance matrix $\mathbf{X}$ and the assigned label matrix $\mathbf{\Upsilon}$ for training a new LML model and to predict the ground-truth labels for unseen data. To solve this problem, we assume the linear regression prediction model. Therefore, we can optimize the weight matrix $\mathbf{W} \in \mathbb{R}^{d\times q}$ by minimizing the square loss with the ℓ_1-norm regularization to learn sparse representation for each label

$$\min_{\mathbf{W}} \frac{1}{2}\|\boldsymbol{\Upsilon} - \mathbf{XW}\|_F^2 + \alpha\|\mathbf{W}\|_1$$

where α is the trade-off parameter and $\|\cdot\|_F$ and $\|\cdot\|_1$ are Frobenius-norm and ℓ_1-norm, respectively.

However, the assigned label matrix $\boldsymbol{\Upsilon}$ includes noisy labels in real application, and thus it can be decomposed

$$\boldsymbol{\Upsilon} = \mathbf{Y} + \mathbf{E}$$

where $\mathbf{Y} \in \mathbb{R}^{n\times q}$ denotes the ground-truth label matrix and $\mathbf{E} \in \mathbb{R}^{n\times q}$ denotes the noise matrix. Most existing LML methods assumed that the noise label is independent of both features and label. However, in real-world applications, the probability of mislabeling/missing label is highly dependent on the specific instance and labels. For depicting noise accurately, we model the noise matrix as the output of a linear map from feature and label representations simultaneously

$$\mathbf{E} = \mathbf{XP} + \mathbf{YQ}$$

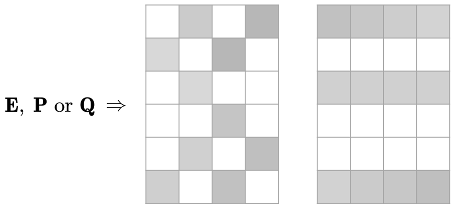
where $\mathbf{P} \in \mathbb{R}^{d\times q}$ and $\mathbf{Q} \in \mathbb{R}^{q\times q}$ are the instance- and label-specific coefficient matrices. Since the noise labels are usually caused by a few of ambiguous content, the specific coefficient matrices $\mathbf{P}$ and $\mathbf{Q}$ are sparse which indicates that the noise is dependent on only key instances and labels. Motivated by previous research (Simon et al. 2013; Huang and Zhang 2010), to handle the conditional sparse noise labels, we consider the following problem

$$\min_{\mathbf{W},\mathbf{P},\mathbf{Q}} \frac{1}{2}\|\boldsymbol{\Upsilon} - \mathbf{XW} - (\mathbf{XP} + \mathbf{XWQ})\|_F^2 \\ + \alpha\|\mathbf{W}\|_1 + \beta\|\mathbf{P}\|_{2,1} + \gamma\|\mathbf{Q}\|_{2,1}$$

where α, β and γ are the trade-off parameters, $\|\cdot\|_{2,1}$ denotes the $\ell_{2,1}$-norm, defined as $\|\mathbf{A}\|_{2,1} = \sum_i(\sum_j \mathbf{A}_{ij})^{1/2}$, encouraging group sparsity. The $\ell_{2,1}$-norm is performed to select instances (labels) across all data with group sparsity structure, i.e. the particular instances (labels) tend to have noisy labels. Our approach to identifying dependent noise involves the use of $\ell_{2,1}$-norm to constrain the noise matrices $\mathbf{P}$ and $\mathbf{Q}$ (i.e., Fig.2b). This constraint ensures that the learned matrix is row-sparse, which corresponds precisely to the form of dependent noise. In contrast, recent studies (Xie and Huang 2021; Sun et al. 2021) impose ℓ_1-norm penalty on the matrix $\mathbf{P}$ (or $\mathbf{E}$) (i.e., Fig.2a), which allow for the learned matrix to be globally sparse and not dependent on specific instances. We also provide theoretical analysis to guarantee that the noise recovery error can be small with high probability if the sample size n is large enough in the case of using the $\ell_{2,1}$-norm.

To preserve local relationships within the feature data and uncover its essential manifold structure, we regularize the minimization problem with the manifold constraint on noise matrix $\mathbf{E}$, i.e., neighboring instances should also share a similar set of noises. Specifically, we define $\mathbf{S} \in \mathbb{R}^{n\times n}$ as a pairwise similarity matrix, where $\mathbf{S}_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2/\varrho)$ if instance i and instance j are the mutually k-nearest neighbors. Otherwise, $\mathbf{S}_{ij} = 0$, where ϱ adjust the degree of proximity. Then, we can get the following regularization term

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{S}_{ij}(\frac{\mathbf{E}_{i:}}{\sqrt{\mathbf{D}_{ii}}} - \frac{\mathbf{E}_{j:}}{\sqrt{\mathbf{D}_{jj}}})^2 = tr(\mathbf{E}^\top\mathbf{LE})$$



(a) ℓ_1- norm (b) $\ell_{2,1}$- norm

Figure 2: Comparison of ℓ_1- and $\ell_{2,1}$-norm on noise.

where $(\cdot)^\top$ denotes the transpose and $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{S}\mathbf{D}^{-\frac{1}{2}}$ is the graph laplacian matrix, $\mathbf{D}$ is a diagonal matrix with $\mathbf{D}_{ii} = \sum_{j=1}^{n}\mathbf{S}_{ij}$ and $\mathbf{I}$ is an identity matrix. $tr(\cdot)$ denotes the trace operator. Based on above assumptions, we formulate the MLDN problem as follows

$$\min_{\mathbf{W},\mathbf{P},\mathbf{Q}} \frac{1}{2}\|\boldsymbol{\Upsilon} - \mathbf{XW} - (\mathbf{XP} + \mathbf{XWQ})\|_F^2 \\ + \lambda tr\left((\mathbf{XP} + \mathbf{XWQ})^\top\mathbf{L}(\mathbf{XP} + \mathbf{XWQ})\right) \\ + \alpha\|\mathbf{W}\|_1 + \beta\|\mathbf{P}\|_{2,1} + \gamma\|\mathbf{Q}\|_{2,1} \qquad (\mathcal{L})$$

where α, β, γ and λ are the trade-off parameters.

We also provide theoretical analysis as follow to demonstrate that the recovery error can be reduced to arbitrarily small values if the sample size n is sufficiently large.

Theorem 1. *Assume that the real noise $\mathbf{E}^\natural$ are instance- and label-dependent (i.e., both $\mathbf{P}^\natural$ and $\mathbf{Q}^\natural$ are group sparse, and we use s_x and s_y to denote the degree of sparsity, g_x and g_y to denote the number of groups for them, respectively). Fix $\mathbf{W}$ in Eq.($\mathcal{L}$), and $\mathbf{E}^\circ := \boldsymbol{\Upsilon} - \mathbf{XW}$ can be seen as an observation of the real noise $\mathbf{E}^\natural$. we assume that the observation error $\mathbf{E}^\circ - \mathbf{E}^\natural$ obeys the subGaussian distribution. Our goal is to recover the true matrices $\mathbf{P}^\natural$ and $\mathbf{Q}^\natural$ from $\mathbf{E}^\circ$. This is similar to a group lasso problem. If $\beta \geq 2\sigma\left(\sqrt{n} + \sqrt{6n\log n}\right)$ and $\gamma \geq 2\sigma\left(\sqrt{n} + \sqrt{6n\log q}\right)$, where $\sigma > 0$ is a constant associated with the observation error, let $g = \min\{g_x, g_y\}$, $p_x = n$ and $p_y = q$, then with probability at least $1 - 2/g^2$ the recovery error bound with*

$$\left\|\mathbf{E}^* - \mathbf{E}^\natural\right\|_2 \leq \sigma \sum_{i\in\{x,y\}} \sqrt{s_i}\left(\sqrt{\frac{1}{n}} + \sqrt{\frac{\log p_i}{n}}\right) \qquad (\mathcal{B})$$

where $\mathbf{P}^*$ is the optimal recovery matrix for Eq.($\mathcal{L}$).

This implies that our algorithm asymptotically converges to the optimal solution. The above result indicates that the parameters β and γ do not depend on the noise rate (i.e., the degree of sparsity). If they satisfy the condition and the number of samples n is large enough, the recovery error can be arbitrarily small with high probability. *This is contrary to the intuition that the parameters β and γ determine the noise rate!* Therefore this theorem ensures that we do not need to manually adjust the parameters in a complicated way.

Optimization

To solve the problem in Eq.($\mathcal{L}$), we iteratively update $\mathbf{W}$, $\mathbf{Q}$, $\mathbf{P}$. We summarize the key steps of MLDN in Algorithm 1.

Update W. When $\mathbf{Q}$ and $\mathbf{P}$ are fixed, the optimization problem in Eq.($\mathcal{L}$) w.r.t $\mathbf{W}$ can be reformulated as follows

$$\begin{aligned}\min_{\mathbf{W},\mathbf{P},\mathbf{Q}} \frac{1}{2}\|\mathbf{\Upsilon} - \mathbf{XP} - \mathbf{XW}(\mathbf{I}+\mathbf{Q})\|_F^2 + \alpha\|\mathbf{W}\|_1 \\ + \lambda tr\left((\mathbf{XP}+\mathbf{XWQ})^\top \mathbf{L}(\mathbf{XP}+\mathbf{XWQ})\right)\end{aligned} \qquad (\mathcal{W})$$

The minimization of Eq.($\mathcal{W}$) is convex, but non-smooth due to the ℓ_1-norm terms. We use the accelerated proximal gradient (*APG*) method to solve it (Beck and Teboulle 2009). Let $g(\mathbf{W}) = \alpha\|\mathbf{W}\|_1$ and $f(\mathbf{W}) = \|\mathbf{\Upsilon} - \mathbf{XP} - \mathbf{XW}(\mathbf{I}+\mathbf{Q})\|_F^2/2 + \lambda tr\left((\mathbf{XP}+\mathbf{XWQ})^\top \mathbf{L}(\mathbf{XP}+\mathbf{XWQ})\right)$. The derivation of f is denoted as

$$\begin{aligned}\nabla_{\mathbf{W}} f =& \mathbf{X}^\top(\mathbf{XW}(\mathbf{I}+\mathbf{Q}) - \mathbf{\Upsilon} + \mathbf{XP})(\mathbf{I}+\mathbf{Q}^\top) \\ &+ 2\mathbf{X}^\top\mathbf{LXPQ}^\top + 2\mathbf{X}^\top\mathbf{LXWQQ}^\top\end{aligned}$$

We approximate $f(\mathbf{W})$ by its first order taylor expansion at the solution $\dot{\mathbf{W}}$ of previous iteration

$$f(\mathbf{W}) \leq f(\dot{\mathbf{W}}) + \langle \mathbf{W} - \dot{\mathbf{W}}, \nabla_{\mathbf{W}} f(\dot{\mathbf{W}})\rangle + \frac{L_f^{\mathbf{W}}}{2}\|\mathbf{W} - \dot{\mathbf{W}}\|_2^2$$

where $L_f^{\mathbf{W}}$ is the Lipschitz constant of $\nabla_{\mathbf{W}} f$. We therefore turn to optimize the following *quadratic programming (QP)* problem

$$\min_{\mathbf{W}} \frac{1}{2} L_f^{\mathbf{W}} \left\| \mathbf{W} - (\dot{\mathbf{W}} - \nabla_{\mathbf{W}} f(\dot{\mathbf{W}})/L_f^{\mathbf{W}}) \right\|_F^2 + g(\mathbf{W})$$

An extrapolation step is included in the *APG* method, then the optimizing rules are given

$$\mathbf{Z}^* = \dot{\mathbf{W}} + \frac{\ddot{\omega} - 1}{\dot{\omega}}(\dot{\mathbf{W}} - \ddot{\mathbf{W}}) \qquad (\mathcal{Z}^\dagger)$$

$$\mathbf{W}^* = \mathcal{S}_{\alpha/L_f^{\mathbf{W}}}(\mathbf{Z}^* - \nabla_{\mathbf{W}} f(\mathbf{Z}^*)/L_f^{\mathbf{W}}) \qquad (\mathcal{W}^\dagger)$$

where $\ddot{\mathbf{W}}$ denotes the optimal solution at the iteration before last. $\dot{\omega}, \ddot{\omega} \in [0, 1)$ are extrapolation parameter at previous iteration and before last iteration, respectively. In practice, we update $\omega = 1/2 + \sqrt{4\dot{\omega}^2 + 1}/2$. $\mathcal{S}$ denotes the elementwise soft-thresholding operator defined by $\mathcal{S}_\tau(v) = (v - \tau)_+ - (-v - \tau)_+$, where $(x)_+$ replaces x with zero if $x < 0$, otherwise unchanged.

Update Q. Fix $\mathbf{W}$ and $\mathbf{P}$, the problem in Eq.($\mathcal{L}$) w.r.t $\mathbf{Q}$ is

$$\begin{aligned}\min_{\mathbf{Q}} \frac{1}{2}\|\mathbf{\Upsilon} - \mathbf{Y} - (\mathbf{XP}+\mathbf{YQ})\|_F^2 + \gamma\|\mathbf{Q}\|_{2,1} \\ + \lambda tr\left((\mathbf{XP}+\mathbf{YQ})^\top \mathbf{L}(\mathbf{XP}+\mathbf{YQ})\right)\end{aligned} \qquad (\mathcal{Q})$$

where we use $\mathbf{Y} = \mathbf{XW}$ for compactness. Defining an objective function for Eq.($\mathcal{Q}$), we have

$$\begin{aligned}\mathcal{L}(\mathbf{Q}) =& \frac{1}{2}\|\mathbf{R}_1 - \mathbf{YQ}\|_F^2 + \gamma\|\mathbf{Q}\|_{2,1} \\ &+ \lambda tr\left(\mathbf{Q}^\top\mathbf{R}_2 + \mathbf{R}_3\mathbf{Q} + \mathbf{Q}^\top\mathbf{R}_4\mathbf{Q}\right) \\ =& \frac{1}{2} tr\left((\mathbf{R}_1 - \mathbf{YQ})^\top(\mathbf{R}_1 - \mathbf{YQ})\right) + \gamma tr(\mathbf{Q}^\top\mathbf{R}_5\mathbf{Q}) \\ &+ \lambda tr\left(\mathbf{Q}^\top\mathbf{R}_2 + \mathbf{R}_3\mathbf{Q} + \mathbf{Q}^\top\mathbf{R}_4\mathbf{Q}\right) \\ =& tr\left(\frac{1}{2}\mathbf{R}_1^\top\mathbf{R}_1 + \mathbf{Q}^\top(\lambda\mathbf{R}_2 - \frac{1}{2}\mathbf{Y}^\top\mathbf{R}_1) + (\lambda\mathbf{R}_3\right. \\ &\left. -\frac{1}{2}\mathbf{R}_1^\top\mathbf{Y})\mathbf{Q} + \mathbf{Q}^\top(\frac{1}{2}\mathbf{Y}^\top\mathbf{Y} + \gamma\mathbf{R}_5 + \lambda\mathbf{R}_4)\mathbf{Q}\right)\end{aligned}$$

Algorithm 1: **Mldn**

```
Input: train data $\mathbf{X}$, assigned label matrix $\mathbf{\Upsilon}$
Output: weight matrix $\mathbf{W}$
$\dot{\omega} \leftarrow 1$ and initialize $\dot{\mathbf{W}}, \mathbf{Q}, \mathbf{P}$
while not converged
    Update $\mathbf{W}$ using Eq.($\mathcal{Z}^\dagger$) and Eq.($\mathcal{W}^\dagger$);
    $\omega \leftarrow 1/2 + \sqrt{4\dot{\omega}^2 + 1}/2$;
    $\ddot{\mathbf{W}} \leftarrow \dot{\mathbf{W}}$, $\dot{\mathbf{W}} \leftarrow \mathbf{W}$ and $\dot{\omega} \leftarrow \omega$;
    Update $\mathbf{Q}$ using Eq.($\mathcal{Q}^\dagger$);
    Update $\mathbf{P}$ using Eq.($\mathcal{P}^\dagger$);
return $\mathbf{W}$
```

where $\mathbf{R}_1 = \mathbf{\Upsilon} - \mathbf{Y} - \mathbf{XP}$, $\mathbf{R}_2 = \mathbf{Y}^\top\mathbf{LXP}$, $\mathbf{R}_3 = \mathbf{P}^\top\mathbf{X}^\top\mathbf{LY}$, $\mathbf{R}_4 = \mathbf{Y}^\top\mathbf{LY}$ and $(\mathbf{R}_5)_{ii} = \frac{1}{2\|\mathbf{Q}_{i:}\|_2}$ is the diagonal elements of the diagonal matrix $\mathbf{R}_5 \in \mathbb{R}^{q\times q}$. Taking derivative of $\mathcal{L}(\mathbf{Q})$ and set it to 0, we have

$$\nabla_{\mathbf{Q}} = \lambda(\mathbf{R}_2 + \mathbf{R}_3^\top) - \mathbf{Y}^\top\mathbf{R}_1 + (\mathbf{Y}^\top\mathbf{Y} + 2\gamma\mathbf{R}_5 + 2\lambda\mathbf{R}_4)\mathbf{Q}$$

namely the solution to the label coefficient matrix is given

$$\mathbf{Q}^* = \frac{\mathbf{Y}^\top\mathbf{Y} + 2\gamma\mathbf{R}_5 + 2\lambda\mathbf{R}_4}{\mathbf{Y}^\top\mathbf{R}_1 - \lambda(\mathbf{R}_2 + \mathbf{R}_3^\top)} \qquad (\mathcal{Q}^\dagger)$$

Update P. Similar to $\mathbf{Q}$, when $\mathbf{W}$ and $\mathbf{Q}$ are fixed, the problem in Eq.($\mathcal{L}$) w.r.t $\mathbf{P}$ is

$$\begin{aligned}\min_{\mathbf{P}} \frac{1}{2}\|\mathbf{S}_1 - \mathbf{XP}\|_F^2 + \beta\|\mathbf{P}^\top\|_{2,1} \\ + \lambda tr\left(\mathbf{P}^\top\mathbf{S}_2 + \mathbf{S}_3\mathbf{P} + \mathbf{P}^\top\mathbf{S}_4\mathbf{P}\right)\end{aligned} \qquad (\mathcal{P})$$

where $\mathbf{S}_1 = \mathbf{\Upsilon} - \mathbf{Y} - \mathbf{YQ}$, $\mathbf{S}_2 = \mathbf{X}^\top\mathbf{LYQ}$, $\mathbf{S}_3 = \mathbf{Q}^\top\mathbf{Y}^\top\mathbf{LX}$, $\mathbf{S}_4 = \mathbf{X}^\top\mathbf{LX}$. We define an objective function for Eq.($\mathcal{P}$),

$$\begin{aligned}\mathcal{L}(\mathbf{P}) = tr&\left(\frac{1}{2}\mathbf{S}_1^\top\mathbf{S}_1 + \mathbf{P}^\top(\lambda\mathbf{S}_2 - \frac{1}{2}\mathbf{X}^\top\mathbf{S}_1) + (\lambda\mathbf{S}_3\right. \\ &\left. -\frac{1}{2}\mathbf{S}_1^\top\mathbf{X})\mathbf{P} + \mathbf{P}^\top(\frac{1}{2}\mathbf{X}^\top\mathbf{X} + \beta\mathbf{S}_5 + \lambda\mathbf{S}_4)\mathbf{P}\right)\end{aligned}$$

where $(\mathbf{S}_5)_{ii} = \frac{1}{2\|\mathbf{P}_{i:}\|_2}$ is diagonal elements of the diagonal matrix $\mathbf{S}_5 \in \mathbb{R}^{d\times d}$. Taking the derivative of $\mathcal{L}(\mathbf{P})$ we have

$$\nabla_{\mathbf{P}} = \lambda(\mathbf{S}_2 + \mathbf{S}_3^\top) - \mathbf{X}^\top\mathbf{S}_1 + (\mathbf{X}^\top\mathbf{X} + 2\beta\mathbf{S}_5 + 2\lambda\mathbf{S}_4)\mathbf{P} = 0$$

the solution to the instance coefficient matrix is given by

$$\mathbf{P}^* = \frac{\mathbf{X}^\top\mathbf{X} + 2\beta\mathbf{S}_5 + 2\lambda\mathbf{S}_4}{\mathbf{X}^\top\mathbf{S}_1 - \lambda(\mathbf{S}_2 + \mathbf{S}_3^\top)} \qquad (\mathcal{P}^\dagger)$$

Experiments

Experimental Setting

Datasets. We conduct the experiments on 10 datasets including slashdot, medical, enron, scene, yeast, 20ng, corel5k, mirflickr, eurlex_dc and m_emotion (abbr. of music_emotion) (Zhang and Zhou 2013; Trohidis et al. 2008). Motivated by a similar setup in *single-label learning* (Xia et al. 2020; Yao et al. 2021), we constructed synthetic noisy datasets to generate instance- and label-dependent

Datasets	MLDN	fPML	PML-NI	PART-VLS	PART-MAP	PML-LC	PML-FP	LMNNE
RankingLoss (↓)								
slashdot	**.124±.012**	.179±.011	.175±.008	.137±.005	.133±.006	.177±.007	.181±.001	.132±.017
medical	**.008±.003**	.069±.003	.011±.004	.081±.009	.078±.013	.055±.006	.052±.013	.016±.012
enron	**.010±.004**	.172±.009	.205±.011	.017±.006	.013±.011	.018±.005	.017±.011	.015±.009
scene	**.078±.005**	.115±.003	.124±.011	.175±.002	.179±.004	.200±.008	.193±.013	.097±.009
yeast	**.179±.006**	.185±.003	.180±.015	.189±.006	.189±.013	.194±.004	.196±.011	.187±.005
20ng	.073±.005	.094±.013	.095±.015	.078±.001	**.072±.013**	.103±.007	.103±.003	.078±.017
corel5k	**.273±.008**	.298±.018	.314±.001	.286±.005	.281±.002	.395±.011	.393±.015	.298±.008
mirflickr	**.126±.005**	.133±.005	.128±.008	.133±.015	.134±.011	.155±.011	.154±.004	.137±.012
eurlex_dc	**.051±.011**	.077±.013	.054±.003	.058±.005	.053±.001	.069±.004	.064±.012	.060±.010
m_emotion	**.246±.003**	.250±.007	.251±.001	.274±.011	.266±.004	.281±.005	.277±.009	.253±.009
OneError (↓)								
slashdot	**.439±.012**	.581±.021	.516±.029	.451±.027	.441±.026	.597±.021	.536±.021	.445±.018
medical	**.201±.002**	.233±.013	.346±.003	.252±.031	.248±.016	.361±.021	.345±.014	.220±.008
enron	**.207±.003**	.466±.021	.619±.011	.498±.005	.463±.005	.569±.005	.544±.003	.412±.006
scene	**.218±.001**	.285±.015	.323±.004	.356±.011	.351±.009	.369±.021	.369±.023	.229±.011
yeast	.224±.015	**.207±.013**	.233±.005	.277±.034	.286±.006	.281±.004	.288±.003	.226±.010
20ng	**.314±.005**	.353±.016	.345±.004	**.314±.021**	.319±.021	.413±.011	.414±.002	.320±.008
corel5k	**.703±.018**	.719±.025	.818±.034	.773±.004	.774±.003	.823±.012	.819±.017	.713±.013
mirflickr	**.311±.001**	.312±.003	**.311±.004**	.921±.016	.943±.021	.793±.021	.805±.002	.319±.009
eurlex_dc	**.284±.002**	.308±.011	.315±.003	.325±.005	.323±.003	.333±.012	.331±.011	.296±.020
m_emotion	**.466±.007**	.483±.003	.469±.011	.526±.004	.513±.020	.586±.011	.579±.005	.471±.011
AveragePrecision (↑)								
slashdot	**.656±.013**	.530±.013	.571±.022	.631±.025	.639±.028	.553±.018	.585±.013	.641±.014
medical	**.821±.036**	.793±.012	.713±.016	.753±.021	.771±.011	.703±.012	.708±.001	.797±.010
enron	**.676±.004**	.661±.009	.457±.016	.591±.010	.660±.015	.555±.014	.561±.013	.668±.008
scene	**.869±.012**	.822±.012	.794±.014	.753±.014	.755±.005	.711±.016	.712±.002	.835±.006
yeast	.764±.003	.765±.015	.751±.011	.710±.004	.713±.005	.705±.012	.710±.014	**.767±.008**
20ng	.777±.017	.745±.009	.744±.002	.777±.012	**.780±.007**	.671±.009	.680±.015	.769±.003
corel5k	**.196±.001**	.192±.004	.133±.014	.170±.012	.176±.006	.133±.001	.128±.011	.183±.016
mirflickr	**.790±.010**	.775±.015	.782±.034	.722±.015	.723±.014	.576±.011	.581±.023	.779±.011
eurlex_dc	**.722±.015**	.622±.014	.720±.009	.633±.031	.631±.045	.591±.007	.591±.007	.711±.001
m_emotion	**.611±.010**	.602±.005	.604±.007	.587±.005	.601±.017	.567±.003	.566±.004	.600±.007

Table 1: Experimental results on *partial multi-label data*. ↑ indicates the larger, the better; ↓ indicates the smaller, the better.

label noise. Given a noise rate τ, we sample instance flip rates $\mathbf{r} \in \mathbb{R}^n$ from the truncated normal distribution $\mathcal{N}\left(\tau, 0.1^2, [0,1]\right)$, and independently sample $\mathbf{w}_1 \in \mathbb{R}^{d\times q}, \mathbf{w}_2 \in \mathbb{R}^{q\times q}$ from the standard normal distribution $\mathcal{N}\left(0, 1^2\right)$. For each $i \in [n]$, we generate instance- and label-dependent flip rates $\mathbf{p}_i = \mathbf{r}_i \times softmax(\mathbf{x}_i\mathbf{w}_1 + \mathbf{y}_i\mathbf{w}_2)$. Lastly, we get the set of noise label $\mathcal{E}_i$ by randomly choosing label q times from the label space according to the possibilities $\mathbf{p}_i$. 1.) For partial multi-label learning comparison purpose, there are two datasets (m_emotion and mirflickr) contain real redundant label noise, and we generate synthetic datasets for other data by adding the set of noise labels $\mathcal{E}_i$ to the original labels set as the redundant noise. 2.) For multi-label learning with missing label comparison, we generate synthetic datasets by removing the labels that appear in $\mathcal{E}_i$. In our work, both the excess and missing rate of label noise are set to 20%.

Baselines. To evaluate the performance of our proposed method, we compare MLDN with 14 state-of-the-art LML methods. 1) For the partial multi-label learning, we select 7 PML state-of-the-art approaches: fPML (Yu et al. 2018), PML-NI (Xie and Huang 2021), PART-VLS, PART-MAP (Zhang and Fang 2021), PML-LC, PML-FP (Xie and Huang 2018) and LMNNE (Gong, Yuan, and Bao 2022). 2) For multi-label learning with missing labels, we select 7 state-of-the-art MLML methods: D2ML-L, D2ML-NL (Ma and Chen 2021), LEML (Yu et al. 2014), MLMLFS (Zhu et al. 2018), GLOCAL (Zhu, Kwok, and Zhou 2017), MLMLV1 (Wu et al. 2014) and MAXIDE (Xu, Jin, and Zhou 2013).

Experimental Setup We use ten-fold cross-validation with a training/test set ratio of 8:2. The convergence condition is determined when the loss difference between the two iterations is less than 10^{-3}.

Comparison Results

we conduct a thorough experimental study on 10 different datasets and compare the proposed MLDN with 14 existing methods that are considered as state-of-the-art in this field. We use 3 widely adopted evaluation metrics for multi-label learning, namely Ranking Loss, One Error, and Average Precision, as defined in (Zhang and Zhou 2013).

Comparison on Partial Mutli-Label Learning. We first

Datasets	MLDN	D2ML-L	D2ML-NL	LEML	MLMLFS	GLOCAL	MLML	MAXIDE
RankingLoss (↓)								
slashdot	**.100±.003**	.223±.003	.191±.007	.164±.001	.209±.005	.189±.010	.179±.003	.226±.002
medical	**.045±.002**	.201±.004	.181±.006	.067±.010	.210±.005	.119±.011	.058±.002	.267±.011
enron	**.082±.004**	.238±.016	.229±.010	.270±.012	.178±.003	.128±.003	.083±.005	.262±.010
scene	.117±.004	.176±.011	.151±.003	.199±.011	.183±.002	.145±.001	**.098±.002**	.356±.006
yeast	**.174±.023**	.238±.007	.230±.007	.186±.006	.393±.023	.230±.005	.193±.002	.313±.005
20ng	**.075±.006**	.349±.014	.348±.013	.199±.003	.355±.001	.286±.003	.198±.002	.475±.004
corel5k	.167±.010	.213±.002	.205±.007	.272±.007	.208±.012	.159±.003	**.145±.006**	.225±.001
mirflickr	**.064±.006**	.071±.003	.076±.013	.088±.003	.134±.007	.080±.003	.077±.002	.224±.011
eurlex_dc	**.041±.014**	.078±.010	.045±.003	.046±.003	.059±.015	.068±.008	.049±.005	.063±.003
m_emotion	.211±.004	.277±.001	.228±.009	**.209±.006**	.324±.009	.224±.008	.222±.011	.367±.007
OneError (↓)								
slashdot	**.408±.001**	.845±.007	.635±.009	.497±.029	.750±.006	.695±.021	.683±.016	.828±.002
medical	**.166±.007**	.878±.003	.718±.006	.272±.006	.731±.014	.553±.023	.361±.023	.733±.024
enron	**.195±.012**	.656±.024	.623±.022	.407±.005	.511±.022	.309±.002	.324±.021	.501±.012
scene	.342±.005	.422±.013	.419±.014	.561±.025	.440±.015	.368±.013	**.334±.013**	.666±.003
yeast	**.233±.012**	.356±.008	.352±.009	.240±.020	.235±.007	.303±.012	.299±.004	.340±.003
20ng	**.331±.010**	.858±.032	.856±.011	.588±.009	.909±.006	.701±.010	.458±.012	.906±.026
corel5k	**.686±.005**	.926±.038	.923±.006	.689±.037	.758±.017	.745±.008	.754±.005	.749±.004
mirflickr	**.116±.007**	.154±.048	.147±.033	.126±.004	.235±.032	.155±.010	.145±.006	.303±.043
eurlex_dc	**.235±.002**	.432±.003	.245±.010	.299±.006	.339±.016	.439±.015	.328±.005	.438±.009
m_emotion	.383±.006	.497±.027	.477±.005	**.376±.021**	.614±.024	.424±.021	.406±.025	.623±.034
AveragePrecision (↑)								
slashdot	**.680±.003**	.353±.026	.461±.020	.606±.004	.429±.014	.469±.022	.474±.003	.463±.013
medical	**.831±.025**	.436±.024	.520±.011	.773±.006	.370±.006	.512±.009	.717±.007	.340±.008
enron	**.687±.009**	.327±.014	.350±.013	.500±.025	.503±.018	.612±.014	.626±.012	.449±.004
scene	.792±.036	.733±.009	.745±.009	.663±.012	.710±.024	.772±.005	**.802±.013**	.658±.024
yeast	**.750±.008**	.672±.018	.684±.013	**.750±.011**	.691±.010	.683±.016	.715±.036	.719±.004
20ng	**.770±.006**	.317±.012	.313±.013	.537±.019	.469±.020	.429±.012	.633±.006	.458±.012
corel5k	**.277±.014**	.156±.005	.200±.020	.238±.018	.196±.017	.232±.016	.237±.012	.219±.017
mirflickr	**.899±.031**	.874±.010	.880±.012	.867±.009	.804±.007	.867±.012	.864±.013	.723±.021
eurlex_dc	**.742±.007**	.694±.032	.733±.014	.700±.031	.654±.018	.629±.005	.685±.016	.635±.017
m_emotion	**.671±.022**	.592±.005	.616±.016	.663±.021	.535±.012	.644±.002	.660±.005	.588±.011

Table 2: Experimental results on *missing multi-label data*. ↑ indicates the larger, the better; ↓ indicates the smaller, the better.

study the performance difference between MLDN and other baseline algorithms in PML setting for label prediction. Table 1 provides the experimental result of 8 methods on 10 different datasets. As shown in Table 1, we can observe the following: i) For the real-word PML datasets m_emotion and mirflickr, MLDN achieves the best performance in all cases. ii) For the synthetic datasets, out of 24 (8 data sets × 3 evaluation metrics) statistical tests MLDN ranks in 1st place at 83.3% cases and in 2nd place at 16.6% cases, which prove the necessity of accounting for dependent noise.

Comparison on Missing Mutli-Label Learning. Furthermore, we study the performance difference in MLML setting for missing labels. Under 20% missing labels, the inherent label structure and correlations are damaged to some extent, and MLDN outperforms other algorithms in most cases. For example, MLDN significantly outperforms MLMLFS, D2ML-L, D2ML-NL, MAXIDE and GLOCAL in terms of all the evaluation metrics on all data, which proves the necessity to model the manifold structure from feature space.

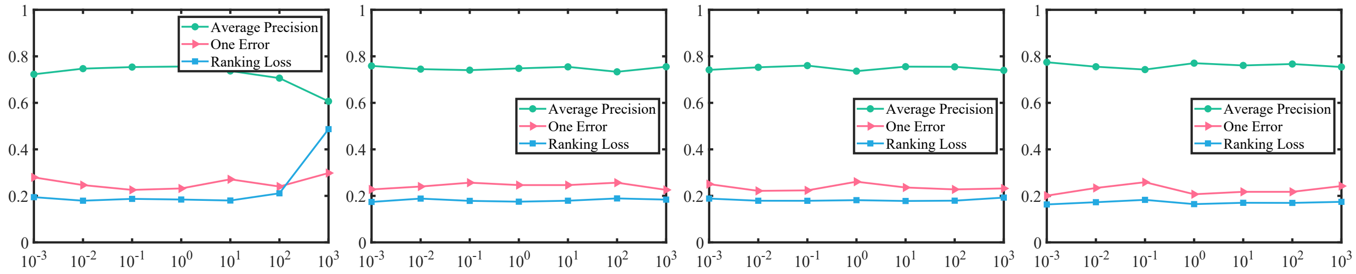
Parameter Sensitivity Analysis. In this experiment, we conduct the sensitivity analysis for our method on the yeast data set over the parameters including α, β, γ and λ. We choose them from $\{10^{-3}, \cdots, 10^{2}, 10^{3}\}$. The 4 sub-figures of Fig.3 show the performance of MLDN changes as each parameter increases with other parameter fixed. From the results, we see that the parameters β, γ and λ are not sensitive to the proposed method, which *is consistent with theoretical analysis*: if β and γ meet the conditions in Theorem 1, they have little impact on the results.

Running Time Analysis. It is crucial to study the efficiency of the compared LML approaches. The running time costs of each method on synthetic and real-world PML (MLML) data sets are recorded in Table 3. In the PML setting, obviously, our approach is the most efficient one on all the data sets. The superiority of our approach is more distinguished on larger datasets. The time costs of existing methods increase dramatically as the data size increases. In contrast, MLDN takes only 107 seconds even for the largest size for eurlex_dc. In the MLML setting, MLDN significantly outperforms most MLML methods on all datasets. For example, for the largest data eurlex_dc, MLDN is more than 30 times faster than LEML, which is the most efficient existing multi-

Datasets	MLDN	fPML	PML-NI	PART-VLS	PART-MAP	PML-LC	PML-FP	LMNNE
slashdot	**1.65s**	75.42s	7.12s	99.32s	92.64s	1931.70s	1968.43s	28.48s
medical	**1.56s**	68.51s	13.18s	25.26s	14.91s	1453.55s	1342.25s	108.07s
enron	**2.48s**	64.56s	8.19s	22.48s	11.67s	1064.18s	1133.54s	131.73s
scene	**0.28s**	1.80s	1.16s	6.87s	4.82s	245.36s	284.69s	8.41s
yeast	**0.84s**	0.95s	1.56s	2.80s	1.29s	108.11s	210.25s	1.21s
20ng	**12.57s**	229.72s	21.15s	2873.17s	2623.56s	>1d	>1d	497.11s
corel5k	**1.92s**	201.26s	132.71s	397.30s	392.75s	1391.39s	1422.35s	338.17s
mirflickr	**0.86s**	2.37s	4.16s	11.45s	7.59s	317.29s	312.65s	6.28s
eurlex_dc	**107.50s**	42945.16s	3218.15s	>1d	>1d	>1d	>1d	5413.43s
m_emotion	**0.70s**	3.13s	3.87s	32.16s	20.56s	194.27s	184.68s	4.38s

Datasets	MLDN	D2ML-L	D2ML-NL	LEML	MLMLFS	GLOCAL	MLML	MAXIDE
slashdot	5.62s	7.77s	138.74s	7.13s	15.32s	9.27s	**1.79s**	2.22s
medical	3.47s	1.23s	22.13s	7.43s	1.15s	16.21s	**0.37s**	2.41s
enron	2.17s	18.71s	96.25s	9.76s	2.58s	14.43s	**0.26s**	2.34s
scene	**0.17s**	6.43s	143.24s	0.38s	1.36s	3.99s	0.22s	0.76s
yeast	**0.09s**	11.57s	214.18s	0.28s	0.97s	1.43s	0.39s	0.18s
20ng	12.14s	1042.15s	8714.12s	31.93s	462.19s	43.02s	54.15s	**8.99s**
corel5k	7.37s	14.36s	3715.39s	103.48s	12.98s	21.91s	**1.89s**	17.56s
mirflickr	**0.47s**	151.21s	5919.94s	1.16s	17.37s	2.66s	10.26s	0.88s
eurlex_dc	**431.13s**	16666.75s	>1d	14531.04s	72122.44s	8830.92s	439.17s	1221.50s
m_emotion	**0.43s**	84.61s	11622.05s	2.33s	6.79s	3.64s	3.80s	0.68s

Table 3: The comparison of *time cost* for MLDN and PML (MLML) baselines on all the datasets with varying data size.



(a) *Vary* α (b) *Vary* β (c) *Vary* γ (d) *Vary* λ

Figure 3: The performance of MLDN changes as each parameter increases with other parameters fixed at data yeast.

AP (↑)	MLDN	MLDN-X	MLDN-Y
medical	**.821±.036**	.793±.020	.805±.016
scene	**.869±.012**	.828±.016	.845±.009

Table 4: Ablation experiment in partial multi-label data.

label learning with missing label algorithm.

Ablation Analysis

Ablation Analysis on Dependent Penalty. In this section we conduct an ablation experiment with dependent noise on data medical and scene with redundant noise. In the ablation experiment, we remove the instance-dependent penalty (refer to MLDN-Y) and label-dependent penalty (MLDN-X), respectively. The evaluation result Average Precision (AP) is given in Table 4. As shown in Table 4, both penalty terms contribute to the performance of MLDN and removing either one of them (instance- or label-dependent penalty) leads to a worse result, which also confirms the rationality of using both constraints together in our model.

Conclusion

In this paper, we proposed a novel LML framework named MLDN, which trains a robust model by considering instance-dependent and label-dependent label noise simultaneously. Specially, we factorized the noise matrix as the outputs of a mapping from the feature and label representations. Meanwhile, we regularized the problem with the manifold constraint on noise matrix to preserve local relationships and a noise recovery error bound is given. Finally, extensive experiments on 10 datasets demonstrate that our proposed method MLDN outperforms the state-of-the-art algorithms.

Acknowledgments

This research acknowledges the generous support provided by the National Natural Science Foundation of China under Grants No. 92267206 and No. 62032013, the Singapore Institute of Technology Ignition Grant under Grant No. R-IE2-A405-0001, and the AcRF Tier 1 Grant under Grant No. R-R12-A405-0009.

References

Beck, A.; and Teboulle, M. 2009. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences*, 2(1): 183–202.

Cheng, Y.; Qian, K.; and Min, F. 2022. Global and Local Attention-based Multi-Label Learning with Missing Labels. *Information Sciences*, 594: 20–42.

Dahiya, K.; Saini, D.; Mittal, A.; Shaw, A.; Dave, K.; Soni, A.; Jain, H.; Agarwal, S.; and Varma, M. 2021. DeepXML: A Deep Extreme Multi-Label Learning Framework Applied to Short Text Documents. *Proceedings of the ACM International Conference on Web Search and Data Mining*.

Gibaja, E.; and Ventura, S. 2014. Multi-Label Learning: A Review of the State of the Art and Ongoing Research. *Wiley Interdisciplinary Reviews: DMKD*, 4(6): 411–444.

Gong, X.; Yuan, D.; and Bao, W. 2022. Partial Multi-Label Learning via Large Margin Nearest Neighbour Embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(6): 6729–6736.

Huang, J.; Xu, L.; Qian, K.; Wang, J.; and Yamanishi, K. 2021. Multi-Label Learning with Missing and Completely Unobserved Labels. *Data Mining and Knowledge Discovery*, 35: 1061–1086.

Huang, J.; and Zhang, T. 2010. The Benefit of Group Sparsity. *The Annals of Statistics*, 38(4): 1978–2004.

Kumar, S.; and Rastogi, R. 2022. Low Rank Label Subspace Transformation for Multi-Label Learning with Missing Labels. *Information Sciences*, 596: 53–72.

Li, L.; Luo, S.; Zhao, Y.; Shan, C.; Wang, Z.; and Qin, L. 2023. COCLEP: Contrastive Learning-based Semi-Supervised Community Search. In *39th IEEE International Conference on Data Engineering, ICDE 2023, Anaheim, CA, USA, April 3-7, 2023*, 2483–2495. IEEE.

Li, Z.; Lyu, G.; and Feng, S. 2020. Partial Multi-Label Learning via Multi-Subspace Representation. In *Proceedings of International Joint Conference on Artificial Intelligence*, 2612–2618.

Lin, D. 2023. Probability Guided Loss for Long-Tailed Multi-Label Image Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1577–1585.

Lin, Y.; Pi, R.; Zhang, W.; Xia, X.; Gao, J.; Zhou, X.; Liu, T.; and Han, B. 2023. A Holistic View of Label Noise Transition Matrix in Deep Learning and Beyond. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Liu, B.-Q.; Jia, B.-B.; and Zhang, M.-L. 2023. Towards Enabling Binary Decomposition for Partial Multi-Label Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liu, W.; Wang, H.; Shen, X.; and Tsang, I. 2021. The Emerging Trends of Multi-Label Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ma, Z.; and Chen, S. 2021. Expand Globally, Shrink Locally: Discriminant Multi-Label Learning with Missing Labels. *Pattern Recognition*, 111: 107675.

Schultheis, E.; Wydmuch, M.; Babbar, R.; and Dembczynski, K. 2022. On Missing Labels, Long-Tails and Propensities in Extreme Multi-Label Classification. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1547–1557.

Simon, N.; Friedman, J.; Hastie, T.; and Tibshirani, R. 2013. A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics*, 22(2): 231–245.

Sun, L.; Feng, S.; Wang, T.; Lang, C.; and Jin, Y. 2019. Partial Multi-Label Learning by Low-Rank and Sparse Decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 5016–5023.

Sun, L.; Lyu, G.; Feng, S.; and Huang, X. 2021. Beyond Missing: Weakly-Supervised Multi-Label Learning with Incomplete and Noisy Labels. *Applied Intelligence*, 51: 1552–1564.

Trohidis, K.; Tsoumakas, G.; Kalliris, G.; Vlahavas, I. P.; et al. 2008. Multi-Label Classification of Music into Emotions. In *Intelligent Information Processing and Web Mining*, volume 8, 325–330.

Wang, Y.; Zhao, Y.; Wang, D. Z.; and Li, L. 2023a. GALOPA: Graph Transport Learning with Optimal Plan Alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Wang, Y.; Zhao, Y.; Wang, Z.; and Wang, M. 2023b. Robust self-supervised multi-instance learning with structure awareness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 10218–10225.

Wu, B.; Liu, Z.; Wang, S.; Hu, B.-G.; and Ji, Q. 2014. Multi-Label Learning with Missing Labels. In *2014 22nd International Conference on Pattern Recognition*, 1964–1968. IEEE.

Xia, X.; Liu, T.; Han, B.; Wang, N.; Gong, M.; Liu, H.; Niu, G.; Tao, D.; and Sugiyama, M. 2020. Part-Dependent Label Noise: Towards Instance-Dependent Label Noise. *Advances in Neural Information Processing Systems*, 33: 7597–7610.

Xie, M.-K.; and Huang, S.-J. 2018. Partial Multi-Label Learning. In *AAAI Conference on Artificial Intelligence*, volume 32.

Xie, M.-K.; and Huang, S.-J. 2021. Partial Multi-Label Learning with Noisy Label Identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xu, L.; Wang, Z.; Shen, Z.; Wang, Y.; and Chen, E. 2014. Learning Low-Rank Label Correlations for Multi-Label Classification with Missing Labels. In *2014 IEEE International Conference on Data Mining*, 1067–1072. IEEE.

Xu, M.; Jin, R.; and Zhou, Z.-H. 2013. Speedup Matrix Completion with Side Information: Application to Multi-Label Learning. In *Advances in Neural Information Processing Systems*, 2301–2309.

Xu, P.; Xiao, L.; Liu, B.; Lu, S.; Jing, L.; and Yu, J. 2023. Label-Specific Feature Augmentation for Long-Tailed Multi-Label Text Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 10602–10610.

Yao, Y.; Liu, T.; Gong, M.; Han, B.; Niu, G.; and Zhang, K. 2021. Instance-Dependent Label-Noise Learning Under a Structural Causal Model. *Advances in Neural Information Processing Systems*, 34: 4409–4420.

Yu, G.; Chen, X.; Domeniconi, C.; Wang, J.; Li, Z.; Zhang, Z.; and Wu, X. 2018. Feature-Induced Partial Multi-Label Learning. In *2018 IEEE International Conference on Data Mining (ICDM)*, 1398–1403. IEEE.

Yu, H.-F.; Jain, P.; Kar, P.; and Dhillon, I. 2014. Large-Scale Multi-Label Learning with Missing Labels. In *International Conference on Machine Learning*, 593–601. PMLR.

Zhang, M.-L.; and Fang, J.-P. 2021. Partial Multi-Label Learning via Credible Label Elicitation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43: 3587–3599.

Zhang, M.-L.; and Zhou, Z.-H. 2013. A Review on Multi-Label Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8): 1819–1837.

Zhao, X.; An, Y.; Xu, N.; and Geng, X. 2022. Fusion Label Enhancement for Multi-Label Learning. In Raedt, L. D., ed., *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 3773–3779.

Zhao, Y.; Wang, Y.; Wang, Z.; and Zhang, C. 2021. Multi-Graph Multi-Label Learning with Dual-Granularity Labeling. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*.

Zhu, P.; Xu, Q.; Hu, Q.; Zhang, C.; and Zhao, H. 2018. Multi-Label Feature Selection with Missing Labels. *Pattern Recognition*, 74: 488–502.

Zhu, Y.; Kwok, J. T.; and Zhou, Z.-H. 2017. Multi-Label Learning with Global and Local Label Correlation. *IEEE Transactions on Knowledge and Data Engineering*, 30(6): 1081–1094.