Incomplete Label Distribution Learning *

Miao Xu and **Zhi-Hua Zhou**
National Key Laboratory for Novel Software Technology, Nanjing University
Nanjing 210023, China
{xum,zhouzh}@lamda.nju.edu.cn

Abstract

Label distribution learning (LDL) assumes labels can be associated to an instance to some degree, thus it can learn the relevance of a label to a particular instance. Although LDL has got successful practical applications, one problem with existing LDL methods is that they are designed for data with *complete* supervised information, while in reality, annotation information may be *incomplete*, because assigning each label a real value to indicate its association with a particular instance will result in large cost in labor and time. In this paper, we will solve LDL problem when given *incomplete* supervised information. We propose an objective based on trace norm minimization to exploit the correlation between labels. We develop a proximal gradient descend algorithm and an algorithm based on alternating direction method of multipliers. Experiments validate the effectiveness of our proposal.

1 Introduction

Classical machine learning tasks assume that one label is either associated with an instance or not. However, in some real applications, labels may be associated with an instance to some degree, thus each instance is annotated by soft labels rather than a single label or a set of labels. Label Distribution Learning (LDL) [Geng, 2016], which learns a mapping from a particular instance to a distribution across all the labels, can assign the relevance of a label to a particular instance. In recent years, LDL has been successfully used in facial age estimation [Geng *et al.*, 2013], action detection in videos [Geng and Ling, 2017], facial expression recognition [Zhou *et al.*, 2015], crowd opinion prediction [Geng and Hou, 2015] et al.

Despite the fact that LDL has been applied successfully in recent years, one problem with existing LDL methods is that they are designed for data with *complete* supervised information. Nevertheless, in reality, annotation information may be *incomplete*. In practice, annotations are often given by human annotators, thus assigning each label a real value to indicate its association with a particular instance will result in a large cost in labor and time, especially when there is a large number of labels and instances. On the other hand, for some labels, it may be difficult to give an accurate value to indicate how they are related to a particular instance. All these phenomena will result in training data with incomplete supervised information, thus one question is put forward, that is, how to do LDL with incomplete annotation (**IncomLDL**).

At first glance, it may be trivial to adapt existing LDL algorithms to fit for the IncomLDL problem. Some of the LDL algorithms are based on maximum entropy model [Geng *et al.*, 2010; 2013; Geng, 2016]. This type of algorithms can be adapted by optimizing the sum of entropy of *observed* labels only. There are some other LDL algorithms transforming the LDL problem into binary classification by sampling labels according to their relevance degree [Geng, 2016]. They can also be adapted to incomplete case by using the *observed* supervised information as sampling weights. Even if these algorithms can be adapted to solve the IncomLDL problem, note that they treat each label *separately*, ignoring the fact that LDL is used in scenarios when labels are *interlaced* with each other [Geng and Ling, 2017]. Without considering the correlation between labels, when facing the severe incomplete annotation, training instances for each single label will be tremendously reduced, thus we need much more training instances to learn a classifier as good as that learned from complete data. Thus label correlations should be exploited to reduce the effect of lacking training data in IncomLDL. Moreover, although one advantage of LDL is that this learning paradigm can take label correlation into consideration by assigning similar relevances to similar labels, however, when annotations are missing, such similarities would also lost. Without such similarity, we need find new ways to characterize the *label correlation* in IncomLDL.

Multi-label learning (MLL) [Zhou and Zhang, 2017], which assumes each instance is associated with multiple labels, is highly related to LDL problem. However, there is a fundamental difference between MLL and LDL. MLL assumes one label is either related to an instance or not, while in LDL, the relative importance of each label is used in describing the instances. Similar to LDL, MLL also faces the incomplete annotation problem, and various algorithms are proposed based on low-rank assumption [Xu *et al.*, 2013], instance-level smoothness [Wu *et al.*, 2016], and label em-

*This work was supported by the NSFC (61333014), the Collaborative Innovation Center of Novel Software Technology and Industrialization, and the Huawei Fund (YBN2017030027).

bedding [Bi and Kwok, 2014]. However, these algorithms are required to give the hard $0/1$ label for multi-label data, while in LDL, the prediction is constrained to be within a probability simplex for a particular instance. Furthermore, adding the probability simplex constraint to the non-smooth unconstrained incomplete MLL problem will result in difficulty in optimizing, thus new optimization algorithms are need to be exploited to solve the new problem.

In this paper, we will solve the IncomLDL problem by considering the correlation between labels, which has never been considered by previous LDL algorithms. Considering labels are correlated and determined by a few factors, we will assume the label distribution matrix is low-rank and propose an optimization objective based on trace norm minimization [Cai *et al.*, 2010]. Our optimization objective will require the entries in the observed positions of the recovered label distribution matrix to be close to those observed values, under the constraint that the recovered label distribution for every instance should form a probability simplex. By showing that following the standard analysis of convex optimization, it is non-trivial to get the optimum solution using the accelerated proximal gradient descend technique [Tseng, 2008], we will additionally develop an alternating direction method of multipliers [Boyd *et al.*, 2011] algorithm. Experiments on 15 real label distribution data sets with various missing percentages validate the effective of the proposed algorithms.

The paper is organized as follows. In Section 2 we will briefly review related work. Our proposed two algorithms IncomLDL-prox and IncomLDL-admm will be introduced in Section 3. We will show experimental results in Section 4, followed by conclusion in Section 5.

2 Related Work

Label distribution learning (LDL) [Geng, 2016] is first proposed to solve the facial age estimation problem [Geng *et al.*, 2010; 2013] by noticing the fact that instances at neighboring ages are similar. In such problems, a distribution across all ages is more desirable than a single age for a face. Later on, [Geng, 2016] discovers that in some real applications the distribution across all labels is more desirable than the association of a single label to an instance. For example, in biology experiments, the gene expression level across all time period is more desirable than the expression level at a particular time point [Geng, 2016]; in facial expression recognition, we can hardly use a particular pure emotion to describe an expression, but a mixture of several basic emotions [Zhou *et al.*, 2015]. Additionally, for some applications, there is natural uncertainty in annotation, thus the instance is annotated by a distribution across labels rather than a single label or a set of labels. One example is crowdsourcing data such as movie ratings [Geng and Hou, 2015], in which different people may have different attitudes, thus the crowd opinion will naturally form a distribution across labels.

Various algorithms have been proposed for LDL, divided into three groups. One group is based on optimizing the sum of log-likelihood of all training labels and instances. Two representative algorithms, IIS-LDL [Geng *et al.*, 2013] and BFGS-LDL [Geng, 2016] belong to this group, using improved iterative scaling [Pietra *et al.*, 1997] and BFGS [Nocedal and Wright, 2006] respectively to do optimization. Another group of algorithms is based on problem transformation [Geng, 2016], transforming LDL into binary classification problem by sampling from the original LDL data using the description degree of that label as sampling weight. After sampling, base learners such as SVM or Naive Bayes are used to do binary classification, forming algorithms PT-SVM and PT-Bayes respectively. The final group of LDL algorithms are those based on algorithm adaptation. Existing algorithms such as kNN [Geng, 2016] and boosting [Xing *et al.*, 2016] are adapted to fit the LDL schema.

When facing the training data with incomplete annotation, some of these algorithms can be adapted to deal with it. However, these algorithms treat each label *separately*, ignoring the fact that LDL is used in scenarios when labels are *interlaced* with each other [Geng and Ling, 2017]. Without considering the correlation between labels, when facing the incomplete annotation, training data for a particular label will be dramatically reduced, thus we will require much more training data to give a satisfactory classifier. Moreover, although LDL can naturally exploit label correlation [Geng, 2016] by assigning similar description degrees to similar labels, we want to state that when there are missing data in the label distribution matrix (for example up to 90% in our experiments), most of the annotation will be missing. It is hard for LDL to acquire the similarity information from the incomplete annotations, thus the classical way how LDL exploit label correlation cannot be used, we need to find new ways to characterize label correlation for IncomLDL.

Multi-label learning (MLL) [Zhou and Zhang, 2017] assumes each instance is associated with a set of labels. Label distribution learning is a natural extension of multi-label learning, by extending the crisp $0/1$ label belongingness to soft probabilistic label belongingness. There are abundant studies about multi-label learning with incomplete label assignments [Sun *et al.*, 2010; Yang *et al.*, 2013; Bi and Kwok, 2014; Wu *et al.*, 2016]. Trace norm minimization [Goldberg *et al.*, 2010; Zhao and Guo, 2015; Xu *et al.*, 2013; Yu *et al.*, 2014] has been popularly used, for that it equals to the low-rank assumption on label matrix, which implicitly exploits the label correlation. These approaches are not directly applicable to LDL, because rather than crisp 0/1 label outputs, the LDL outputs are constrained to follow a probability simplex.

To summarize, we need to propose new LDL algorithms which can deal with incomplete supervised information. By considering the label correlation, our proposed algorithms should give a better result than existing LDL algorithms.

3 The IncomLDL Methods

3.1 Formalization

We will give a more formal definition of LDL. LDL assigns a value d_x^y called description degree to instance x for a particular label y, which indicates the relevance of the label y to the instance x. In LDL, all d_x^y-s for a particular instance form a probability simplex. Note that d_x^y is not the probability that y correctly labels x, but the proportion that y accounts for in a

full description of x [Geng, 2016]. All labels with non-zero d_x^y-s are actually the correct labels to describe the instance, where their relative importance is measured by d_x^y.

Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be the feature matrix, where n is the number of instances and d is the number of features, thus the ith row of $\mathbf{X}$ will be instance x_i. $\mathbf{D} \in \mathbb{R}^{n\times m}$ is the label distribution matrix, where m is the number of labels and D_{ij} is $d_{x_i}^{y_j}$. Since the supervised information for one instance should follow the probability simplex, we have $\sum_{j=1}^{m} D_{ij} = 1\ \forall i \in [n]$ and $D_{ij} \geq 0\ \forall (i,j) \in [n] \times [m]$.

When we are dealing with the incomplete annotation, we assume that entries in label distribution matrix $\mathbf{D}$ are uniformly random missing. Let $\Omega \in [n] \times [m]$ denote the indices of observed entries sampled uniformly at random from $\mathbf{D}$. We denote the observed label distribution matrix by $\tilde{\mathbf{D}}$, which has equal size of $\mathbf{D}$, with entries in the observed positions same as $\mathbf{D}$ and entries in unobserved positions 0. Specially, we will have linear operator $\mathcal{R}_\Omega : \mathbb{R}^{n\times m} \to \mathbb{R}^{n\times m}$ defined as,

$$[\mathcal{R}_\Omega(\mathbf{M})]_{ij} = \begin{cases} M_{ij} & \text{if } (i,j) \in \Omega \\ 0 & \text{otherwise} \end{cases}$$

Finally, we will use $\|\cdot\|_F$ to denote the Frobenius norm of a matrix, and $\|\cdot\|_{tr}$ to denote the trace norm of the matrix, which is the sum of all singular values.

Based on the above notation, to solve the IncomLDL problem, we will propose a learning objective based on squared loss and a regularizer exploiting label correlations. Specially, we will have the following optimization objective,

$$\begin{aligned} \min_{\mathbf{W}} \quad & \frac{1}{2}\|\mathcal{R}_\Omega(\mathbf{XW} - \tilde{\mathbf{D}})\|_F^2 + \lambda\|\mathbf{XW}\|_{tr} \\ \text{s.t.} \quad & \mathbf{XW} \times \mathbf{1}_m = \mathbf{1}_n,\ \ \mathbf{XW} \geq \mathbf{0}_{n\times m} \end{aligned} \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{d\times m}$ is our learning objective. $\mathbf{1}_n$ and $\mathbf{1}_m$ are the length n and m vectors with all its entries equaling one, respectively. $\mathbf{0}_{n\times m}$ is an all-zero matrix of size $n \times m$.

In Eq. 1, we assume that there is a linear relationship between the label matrix $\mathbf{D}$ and feature matrix $\mathbf{X}$ since linear classifiers have acquired good performance in previous studies [Geng, 2016]. In this way, the *recovered* label matrix will be $\hat{\mathbf{D}} = \mathbf{XW}$, where $\mathbf{W}$ is the linear coefficients and will be our learning objective.

To exploit the correlation between labels, we assume that $\mathbf{D}$ is low-rank, i.e. $\mathbf{XW}$ has a small trace norm. λ is the regularization parameter trading off the importance between trace norm and the difference in Frobenius norm in those observed positions. Because the recovered supervised information for each matrix need to be in the probability simplex, that is, nonnegative and summing up to one, we add constraint $\mathbf{XW} \times \mathbf{1}_m = \mathbf{1}_n$ to make sure that each row of $\hat{\mathbf{D}}$ will sum up to one and $\mathbf{XW} \geq \mathbf{0}_{n\times m}$ to constraint all entries to be nonnegative. Combining all these aspects, we will have Eq. 1 as our learning objective for IncomLDL problem.

3.2 Optimizing using Proximal Gradient Descend

Following [Xu *et al.*, 2013], we will first use Accelerated Proximal Gradient Descend to optimize Eq. 1, which can achieve a convergence rate of $O(1/T^2)$ [Tseng, 2008], where T is the number of iterations. Since it is difficult to handle the trace norm of $\mathbf{XW}$, we will assume $\mathbf{X}$ is orthonormal without losing of generality. If $\mathbf{X}$ is not orthonormal, we can do SVD on $\mathbf{X}$ and use the top right singular vectors as a replacement for $\mathbf{X}$. After assuming $\mathbf{X}$ is orthonormal, the optimization will become,

$$\begin{aligned} \min_{\mathbf{W}} \quad & \frac{1}{2}\|\mathcal{R}_\Omega(\mathbf{XW} - \tilde{\mathbf{D}})\|_F^2 + \lambda\|\mathbf{W}\|_{tr} \\ \text{s.t.} \quad & \mathbf{XW} \times \mathbf{1}_m = \mathbf{1}_n,\ \ \mathbf{XW} \geq \mathbf{0}_{n\times m} \end{aligned} \tag{2}$$

Ignoring the constraints, Accelerated Proximal Gradient Descend will optimize Eq. 2 iteratively. In the tth iteration, it will introduce an auxiliary variable $\mathbf{Y}_t$, which is,

$$\mathbf{Y}_t = \mathbf{W}_t + \theta_t(\frac{1}{\theta_{t-1}} - 1)(\mathbf{W}_t - \mathbf{W}_{t-1}) \tag{3}$$

After introducing $\mathbf{Y}_t$, we have the following subproblem in the tth iteration,

$$\begin{aligned} \min_{\mathbf{W}} \quad & \lambda\|\mathbf{W}\|_{tr} + \\ & \frac{L}{2}\left\|\mathbf{W} - \left(\mathbf{Y}_t - \frac{1}{L}\mathbf{X}^\top\mathcal{R}_\Omega(\mathbf{XY}_{t-1} - \tilde{\mathbf{D}})\right)\right\|_F^2 \end{aligned} \tag{4}$$

which has closed form solution by SVT [Cai *et al.*, 2010].

Note that here L is the Lipschitz constant which can be found by linear search, i.e., we will initialize L and increase it until the following is violated.

$$\begin{aligned} \ell(\mathbf{W}_{t+1}) \leq \ell(\mathbf{Y}_t) + \\ \langle\nabla f(\mathbf{Y}_t), \mathbf{W}_{t+1} - \mathbf{Y}_t\rangle + \frac{L}{2}\|\mathbf{W}_{t+1} - \mathbf{Y}_t\|_F^2 \end{aligned} \tag{5}$$

where $\ell(\mathbf{W}_{t+1}) = \|\mathcal{R}_\Omega(\mathbf{XW} - \tilde{\mathbf{D}})\|_F^2/2$.

One problem with the above procedure is that, it is used for unconstrained problem, while in our setting, the problem is constrained. One straightforward solution is to project the solution $\mathbf{W}_t$ onto the set defined by the constraints, that is, we will project each row of $\hat{\mathbf{D}}_t = \mathbf{XW}_t$ onto the probability simplex. Assuming the ith row of $\hat{\mathbf{D}}_t$ is $\hat{\mathbf{d}}_t^i$, to do projection, we will solve the following problem,

$$\min_{\mathbf{d}} \quad \|\mathbf{d} - \hat{\mathbf{d}}_t^i\| \ \text{ s.t. } \mathbf{d}^\top\mathbf{1}_m = 1,\ \ \mathbf{d} \geq \mathbf{0}_m, \tag{6}$$

while without misunderstanding, we will omit i and t in $\hat{\mathbf{d}}_t^i$.

The above problem has an $O(m \log m)$ solution by sorting all the elements in $\hat{\mathbf{d}}$ in descending order into $\mathbf{u}$. Then we find the maximum position index ρ such that $\theta = u_\rho + (1 - \sum_{i=1}^{\rho} u_i)/\rho > 0$. With ρ, we will get the optimal $\mathbf{d}$ in which $d_j = \max\{\hat{d}_j + \theta, 0\}, \forall j \in [m]$. The proof of correctness of the projection procedure can be found in [Duchi *et al.*, 2008; Wang and Carreira-Perpiñán, 2013].

After we get the $\mathcal{R}_{prob}(\hat{\mathbf{D}}_t)$ which project each row of $\hat{\mathbf{D}}_t$ onto a probability simplex, we will recover $\mathbf{W}_t$ as follows,

$$\min_{\mathbf{W}} \|\mathbf{XW}^\top - \mathcal{R}_{prob}(\hat{\mathbf{D}}_t)\|_F^2 \tag{7}$$

which has closed-form solution as $(\mathbf{X}^\top\mathbf{X})^\dagger[\mathbf{X}^\top\mathcal{R}_{prob}(\hat{\mathbf{D}}_t)]$.

Algorithm 1 IncomLDL-prox

Initialization: $\theta_1 = \theta_2 \in (0, 1]$, $\mathbf{W}_1 = \mathbf{W}_2$, L, $\gamma > 1$, and stopping criterion ϵ
$t = 2$;
while stopping criterion is not satisfied **do**
  Calculate $\mathbf{Y}_t$ by Eq. 3
  Calculate $\mathbf{W}_{t+1}$ by Eq. 4
  while Eq. 5 is satisfied **do**
    $L = L * \gamma$
    Update $\mathbf{W}_{t+1}$ by Eq. 4 with the new L
  end while
  $\theta_{t+1} = (\sqrt{\theta_t^4 + 4\theta_t^2} - \theta_t^2)/2$
  $t = t + 1$
  $\hat{\mathbf{D}}_t = \mathbf{X}\mathbf{W}_t$
  Calculate each row of $\mathcal{R}_{prob}(\hat{\mathbf{D}}_t)$ by solving Eq. 6
  Update $\mathbf{W}_t$ by Eq. 7
end while

The IncomLDL-prox algorithm which uses accelerated proximal gradient descend with a projection onto the probability simplex is shown in Alg. 1.

There remains one question, that is, does the projection procedure really lead to a minimum of the following constrained optimization problem in the tth iteration?

$$\begin{aligned} \min_{\mathbf{W}} \quad & \lambda\|\mathbf{W}\|_{tr} + \frac{L}{2}\|\mathbf{W} - \mathbf{Q}_t\|_F^2 \\ \text{s.t.} \quad & \mathbf{X}\mathbf{W} \times \mathbf{1}_m = \mathbf{1}_n, \ \ \mathbf{X}\mathbf{W} \geq \mathbf{0}_{n\times m} \end{aligned} \tag{8}$$

where $\mathbf{Q}_t = \mathbf{Y}_t - \mathbf{X}^\top \mathcal{R}_\Omega(\mathbf{X}\mathbf{Y}_{t-1} - \tilde{\mathbf{D}})/L$

Note the Lagrange dual problem of Eq. 8 is as following,

$$\begin{aligned} \max_{\alpha,\mathbf{B}} \quad & \mathcal{D}_{\lambda/L}[\mathbf{Q}_t - \mathbf{X}^\top\mathbf{A} + \mathbf{X}^\top\mathbf{B}] + \alpha^\top \times \mathbf{1}_n \\ & -\frac{1}{2}\|\mathbf{Q}_t - \mathbf{X}^\top\mathbf{A} + \mathbf{X}^\top\mathbf{B}\|_F^2 \end{aligned}$$

where α and $\mathbf{B}$ are Lagrange multipliers associated with the equality constraint and inequality constraint respectively. $\mathbf{A} \in \mathbb{R}^{n\times m}$ is the catenation of m α-s. $\mathcal{D}_{\lambda/L}[\cdot]$ is the SVT solver [Cai *et al.*, 2010].

To solve Eq. 8, we will first solve α and $\mathbf{B}$ for the Lagrange dual problem. With the optimal solution α^* and $\mathbf{B}^*$, the original unconstrained optimizer $\mathbf{W}^*$ can be calculated. However, maximizing $\mathcal{D}_{\lambda/L}[\cdot]$ does not have closed-form solution for α and $\mathbf{B}$. Although Algo. 1 solves the problem by calculating $\mathcal{D}_{\lambda/L}[\mathbf{Q}_t]$ and then projecting the solution to a probability simplex, it cannot be guaranteed to find the optimal solution following the standard analysis of convex optimization [Boyd and Vandenberghe, 2004]. Thus we need to find new optimization strategy to optimize Eq. 1.

3.3 Optimizing using ADMM

ADMM (Alternating Direction Method of Multipliers) [Boyd *et al.*, 2011] which is suitable to those objectives summing up a smooth function and a non-smooth function, is proper for solving Eq. 1. To use ADMM, we first rewrite our objective into the following equivalent form,

$$\begin{aligned} \min_{\mathbf{W}\in\mathcal{C}} \quad & \frac{1}{2}\left\|\mathcal{R}_\Omega(\mathbf{X}\mathbf{W} - \tilde{\mathbf{D}})\right\|_F^2 + \lambda\|\mathbf{Z}\|_{tr} \\ \text{s.t.} \quad & \mathbf{X}\mathbf{W} - \mathbf{Z} = \mathbf{0} \end{aligned} \tag{9}$$

where

$$\mathcal{C} = \{\mathbf{W} | \mathbf{X}\mathbf{W} \times \mathbf{1}_m = \mathbf{1}_n \text{ and } \mathbf{X}\mathbf{W} \geq \mathbf{0}_{n\times m}\}.$$

Eq. 9 can be solved by the following alternative methods in iteration t,

$$\begin{aligned} \mathbf{W}^{t+1} &= \arg\min_{\mathbf{W}\in\mathcal{C}} \frac{1}{2}\|\mathcal{R}_\Omega(\mathbf{X}\mathbf{W} - \tilde{\mathbf{D}})\|_F^2 \\ &\quad + \langle\mathbf{\Lambda}^t, \mathbf{X}\mathbf{W} - \mathbf{Z}^t\rangle + \frac{\rho_1}{2}\|\mathbf{X}\mathbf{W} - \mathbf{Z}^t\|_F^2 \end{aligned} \tag{10}$$

$$\begin{aligned} \mathbf{Z}^{t+1} &= \arg\min_{\mathbf{Z}} \lambda\|\mathbf{Z}\|_{tr} + \langle\mathbf{\Lambda}^t, \mathbf{X}\mathbf{W}^{t+1} - \mathbf{Z}\rangle \\ &\quad + \frac{\rho_1}{2}\|\mathbf{X}\mathbf{W}^{t+1} - \mathbf{Z}\|_F^2 \end{aligned} \tag{11}$$

$$\mathbf{\Lambda}^{t+1} = \mathbf{\Lambda}^t + \rho_1(\mathbf{X}\mathbf{W}^{t+1} - \mathbf{Z}^{t+1}) \tag{12}$$

in which Eq. 11 can be rewritten into

$$\mathbf{Z}^{t+1} = \arg\min_{\mathbf{Z}} \frac{1}{2}\|\mathbf{Z} - (\mathbf{X}\mathbf{W}^{t+1} + \frac{\mathbf{\Lambda}^t}{\rho_1})\|_F^2 + \frac{\lambda}{\rho_1}\|\mathbf{Z}\|_{tr}$$

which has closed form solution. Thus the problem remained is how to solve Eq. 10.

Assuming $\mathbf{M} = \mathbf{X}\mathbf{W}$, we rewrite Eq. 10 here,

$$\begin{aligned} \min_{\mathbf{M}} \quad & \frac{1}{2}\|\mathcal{R}_\Omega(\mathbf{M} - \tilde{\mathbf{D}})\|_F^2 + \langle\mathbf{\Lambda}^t, \mathbf{M} - \mathbf{Z}^t\rangle \\ & + \frac{\rho_1}{2}\|\mathbf{M} - \mathbf{Z}^t\|_F^2 \\ \text{s.t.} \quad & \mathbf{M} \times \mathbf{1}_m = \mathbf{1}_n, \ \ \mathbf{M} \geq \mathbf{0}_{n\times m} \end{aligned}$$

We can decompose the above problem into optimizing each row of $\mathbf{M}$, while the ith row of $\mathbf{M}$ is denoted as M_i, so is Λ_i, $\tilde{D}_i$, Z_i and Ω_i,

$$\begin{aligned} \min_{M_i} \quad & \frac{1}{2}\|\mathcal{R}_{\Omega_i}(M_i - \tilde{D}_i)\|^2 \\ & + \langle\Lambda_i^t, M_i - Z_i^t\rangle + \frac{\rho_1}{2}\|M_i - Z_i^t\|^2 \\ \text{s.t.} \quad & M_i \times \mathbf{1}_m = 1, \ \ M_i \geq \mathbf{0}_m \end{aligned} \tag{13}$$

Although in [Duchi *et al.*, 2008; Wang and Carreira-Perpiñán, 2013], the projection onto a probability simplex problem is solved using an $O(m \log m)$ algorithm by brute force searching through $[m]$ from the largest entry to the smallest one for a particular j satisfying the KKT condition, the algorithm is inefficient to be used here. The reason is the m entries in M_i are divided into Ω_i and $\bar{\Omega}_i$, and the non-zero M_{ij}s will have different weights and different projection objectives due to the KKT condition, thus we need to search all possible positions in Ω_i and $\bar{\Omega}_i$ for a pair of perfect solution, which will cost a lot of time ($O(m^2)$ for the worst case) and cannot guarantee a unique optimal solution. Thus we will solve Eq. 13 by forming it into a standard QP problem, and use state-of-the-art QP solvers, such as interior-point-method.

Algorithm 2 IncomLDL-admm

Initialization: $\mathbf{W}^1$, $\mathbf{\Lambda}^1$ and $\mathbf{Z}^1$, λ and ρ_1, $t = 1$
while stopping criterion is not satisfied **do**
  Calculate each row of $\mathbf{M}^*$ by Eq. 13
  $\mathbf{W}^{t+1} = (\mathbf{X}^\top\mathbf{X})^\dagger[\mathbf{X}^\top\mathbf{M}^*]$
  Solve $\mathbf{Z}^{t+1}$ by Eq. 11
  Update $\mathbf{\Lambda}^{t+1}$ by Eq. 12
  $t = t + 1$
end while

Table 1: Statistics of the 15 data sets, where n is number of instance, d is number of features and m is number of labels.

Dataset	n	d	m
Yeast-alpha	2465	24	18
Yeast-cdc	2465	24	15
Yeast-elu	2465	24	14
Yeast-diau	2465	24	7
Yeast-heat	2465	24	6
Yeast-spo	2465	24	6
Yeast-cold	2465	24	4
Yeast-dtt	2465	24	4
Yeast-spo5	2465	24	3
Yeast-spoem	2465	24	2
Human Genes	20,542	36	68
Natural Scene	2,000	294	9
SJAFFE	213	243	6
SBU_3DFE	2,500	243	6
Movie	7,755	1,869	5

After we get $\mathbf{M}^*$, we can project it back into the space defined by $\mathbf{X}$ through $(\mathbf{X}^\top\mathbf{X})^\dagger[\mathbf{X}^\top\mathbf{M}^*]$. We will summarize the proposed IncomLDL-ADMM algorithm in Alg. 2

According to [He and Yuan, 2012], our IncomLDL-admm will converge at $O(1/T)$ rate to the optimum solution. Although it is slow compared to the $O(1/T^2)$ rate by accelerated proximal gradient descend method, in practice, a good approximate solution is sufficient to obtain satisfactory performance [Boyd *et al.*, 2011].

4 Experiment

We evaluate the proposed algorithms for IncomLDL problem on 15 real data sets. We implement our approaches in `Matlab`. All the results were obtained on a Linux server with CPU 2.53GHz and 48GB memory.

The algorithm is evaluated on 15 data sets covering fields of biochemistry, natural scene recognition, facial expression and movie-rating. Details of them can be found in [Geng, 2016]. Here we summarize their statistics in Table 1.

Settings and Baselines To make these data sets incomplete, we will use two kinds of settings. In the first setting, we will make all elements in the label distribution matrix uniformly random missing and call it the (general) *incomplete* setting. We vary the observed rate $\omega\%$ from 10% to 40%, and measure the difference between the groundtruth and the predicted label distribution matrix. We will then test our proposed algorithm in the second setting, the *transductive* setting, in which we have 10% test data with no supervised information, accompanied by incomplete training data, while the observed rate $\omega\%$ will also vary from 10% to 40%. We will use the incomplete training data and features of the test data together to give a prediction. Difference between the ground truth and the predicted distribution matrix for *test data* will be measured. We will repeat each experiments 10 times and report the results averaged over 10 trials.

In IncomLDL-prox, the regularization parameter is selected from $2^{\{-10,-9,\ldots,9,10\}}$ by cross-validation on training data. Parameters γ and ϵ are set to be 2 and 10^{-5} respectively. The maximum iteration is set to be 100. In IncomLDL-admm, regularization parameter λ and number of maximum iteration are selected in the same way as IncomLDL-prox. ρ_1 is simply set as 1 and all the variables are initialized to be all-zero. The stopping criterion parameters ϵ^{abs} and ϵ^{rel} are set as 10^{-4} and 10^{-2} as suggested in the survey [Boyd *et al.*, 2011].

We will compare our proposed IncomLDL algorithm with several baselines. They are all adapted from existing label distribution algorithms to fit for the incomplete situation. Note that although these algorithms can solve the incomplete label distribution problem, they consider each label separately thus are expected to work worse than our proposal. These algorithms include two maximum entropy algorithms IIS-LDL [Geng *et al.*, 2013], BFGS-LDL [Geng, 2016] and two problem transformation algorithms PT-Bayes and PT-SVM [Geng, 2016]. All the codes are shared by original authors, and we use the default parameter suggested there, except that we tune the regularization parameter for the PT-SVM algorithm using 10-folder cross-validation in the same way as in IncomLDL-prox.

Following [Geng, 2016], we will use *five* measurements for incomplete label distribution problem. Among them, `Chebyshev`, `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. Details of these measurements can be found in [Geng, 2016]. Note that there is one additional measurement proposed in [Geng, 2016], which measures the KL-divergence between two vectors. For KL-divergence is calculated by $\log(d_x^y/\hat{d}_x^y)$, and it will be meaningless when $\hat{d}_x^y$ is zero, we will not use it here.

Results Due to space limitation, here we only present representative results. Other results are similar and we will put them in a longer version. Note that we have done experiments on 4 different $\omega\%$ within both incomplete and transductive setting, measured on 5 measurements. Here we will present the `Chebyshev` (the lower the better) results for *incomplete* setting on *all* the data with $\omega = 10$ in Table 2 and the `Intersection` (the higher the better) results for *transductive* setting on *test* data only with $\omega = 30$ in Table 3.

We can see in both these two scenarios, our proposed two algorithms are superior to the baselines. The results are as expected since these two methods exploit the label correlation when there are insufficient training data facing the

Table 2: Chebyshev (the lower the better) results for incomplete setting on all data when $\omega\% = 10\%$. Each column corresponds to an algorithm, while IncomLDL-a and IncomLDL-p are abbreviation for IncomLDL-admm and IncomLDL-prox respectively. Each row corresponds to a data set. The value is measured by 10-folder cv shown in mean±std form. The best results on each row are bolded, with its comparable ones (pairwise single-tailed t-test at 95% confidence level) marked by $^\bullet$.

Data Set	IncomLDL-a	IncomLDL-p	IIS-LDL	BFGS-LDL	PT-Bayes	PT-SVM
Yeast-alpha	$.0135 \pm .0000^\bullet$	$\mathbf{.0135 \pm .0000}$	$.0214 \pm .0003$	$.0361 \pm .0000$	$.4401 \pm .0329$	$.0178 \pm .0012$
Yeast-cdc	$\mathbf{.0161 \pm .0001}$	$.0162 \pm .0000^\bullet$	$.0247 \pm .0006$	$.0427 \pm .0000$	$.4543 \pm .0359$	$.0221 \pm .0021$
Yeast-cold	$\mathbf{.0513 \pm .0001}$	$.0514 \pm .0002^\bullet$	$.0636 \pm .0013$	$.0945 \pm .0000$	$.4624 \pm .0298$	$.0672 \pm .0085$
Yeast-diau	$\mathbf{.0370 \pm .0003}$	$.0371 \pm .0001^\bullet$	$.0479 \pm .0009$	$.0751 \pm .0000$	$.4917 \pm .0408$	$.0510 \pm .0064$
Yeast-dtt	$\mathbf{.0361 \pm .0000}$	$.0362 \pm .0001^\bullet$	$.0518 \pm .0011$	$.0851 \pm .0000$	$.4718 \pm .0373$	$.0501 \pm .0096$
Yeast-elu	$\mathbf{.0162 \pm .0000}$	$.0163 \pm .0000^\bullet$	$.0255 \pm .0005$	$.0441 \pm .0000$	$.4431 \pm .0254$	$.0223 \pm .0019$
Yeast-heat	$\mathbf{.0422 \pm .0002}$	$.0425 \pm .0002^\bullet$	$.0545 \pm .0013$	$.0802 \pm .0000$	$.4590 \pm .0315$	$.0530 \pm .0049$
Yeast-spo	$.0584 \pm .0000^\bullet$	$\mathbf{.0582 \pm .0001}$	$.0671 \pm .0015$	$.0927 \pm .0000$	$.4847 \pm .0271$	$.0684 \pm .0058$
Yeast-spo5	$\mathbf{.0912 \pm .0002}$	$.0913 \pm .0001^\bullet$	$.0989 \pm .0014$	$.1327 \pm .0000$	$.4478 \pm .0426$	$.0997 \pm .0063$
Yeast-spoem	$.0875 \pm .0003^\bullet$	$\mathbf{.0874 \pm .0004}$	$.0939 \pm .0018$	$.1190 \pm .0000$	$.3565 \pm .0269$	$.0980 \pm .0126$
Human Gene	$.0533 \pm .0000^\bullet$	$\mathbf{.0533 \pm .0000}$	$.0535 \pm .0000$	$.0543 \pm .0000$	$.5453 \pm .1388$	$.0537 \pm .0001$
Natural Scene	$\mathbf{.3360 \pm .0031}$	$.3380 \pm .0025^\bullet$	$.3576 \pm .0024$	$.7179 \pm .0000$	$.3690 \pm .0000$	$.4168 \pm .0213$
SJAFFE	$\mathbf{.1078 \pm .0021}$	$.1083 \pm .0024^\bullet$	$.1279 \pm .0089$	$.7771 \pm .0000$	$.1204 \pm .0000$	$.1417 \pm .0185$
SBU_3DFE	$\mathbf{.1170 \pm .0000}$	$.1185 \pm .0009^\bullet$	$.1351 \pm .0012$	$.2301 \pm .0000$	$.1389 \pm .0000$	$.1414 \pm .0028$
Movie	$.1257 \pm .0000^\bullet$	$\mathbf{.1237 \pm .0011}$	$.3697 \pm .0038$	$.4952 \pm .0000$	$.1807 \pm .0000$	$.2510 \pm .0278$

Table 3: Intersection (the higher, the better) results for transductive setting on test data when $\omega\% = 30\%$. Each column corresponds to an algorithm, while IncomLDL-a and IncomLDL-p are abbreviation for IncomLDL-admm and IncomLDL-prox respectively. Each row corresponds to a data set. The value is measured by 10-folder cv shown in mean±std form. The best results on each row are bolded, with its comparable ones (pairwise single-tailed t-test at 95% confidence level) marked by $^\bullet$.

Data Set	IncomLDL-a	IncomLDL-p	IIS-LDL	BFGS-LDL	PT-Bayes	PT-SVM
Yeast-alpha	$.9621 \pm .0001^\bullet$	$\mathbf{.9625 \pm .0005}$	$.9417 \pm .0016$	$.8845 \pm .0038$	$.5749 \pm .0180$	$.9548 \pm .0027$
Yeast-cdc	$.9567 \pm .0013^\bullet$	$\mathbf{.9575 \pm .0011}$	$.9386 \pm .0011$	$.8862 \pm .0030$	$.5888 \pm .0136$	$.9512 \pm .0027$
Yeast-cold	$\mathbf{.9406 \pm .0015}$	$.9398 \pm .0011^\bullet$	$.9264 \pm .0019$	$.8902 \pm .0035$	$.6292 \pm .0199$	$.9279 \pm .0050$
Yeast-diau	$.9391 \pm .0008^\bullet$	$\mathbf{.9394 \pm .0013}$	$.9249 \pm .0018$	$.8825 \pm .0019$	$.5974 \pm .0183$	$.9249 \pm .0053$
Yeast-dtt	$\mathbf{.9605 \pm .0014}$	$.9582 \pm .0015^\bullet$	$.9406 \pm .0025$	$.8971 \pm .0041$	$.6338 \pm .0186$	$.9514 \pm .0044$
Yeast-elu	$.9573 \pm .0008^\bullet$	$\mathbf{.9584 \pm .0010}$	$.9383 \pm .0014$	$.8831 \pm .0034$	$.5724 \pm .0292$	$.9508 \pm .0015$
Yeast-heat	$.9393 \pm .0014^\bullet$	$\mathbf{.9402 \pm .0015}$	$.9233 \pm .0021$	$.8858 \pm .0031$	$.6166 \pm .0187$	$.9324 \pm .0046$
Yeast-spo	$\mathbf{.9175 \pm .0028}$	$.9161 \pm .0031^\bullet$	$.9048 \pm .0032$	$.8680 \pm .0037$	$.6114 \pm .0233$	$.9022 \pm .0064$
Yeast-spo5	$\mathbf{.9154 \pm .0031}$	$.9086 \pm .0043^\bullet$	$.9026 \pm .0049$	$.8727 \pm .0052$	$.6644 \pm .0224$	$.9033 \pm .0055$
Yeast-spoem	$.9112 \pm .0036^\bullet$	$\mathbf{.9133 \pm .0035}$	$.9063 \pm .0028$	$.8893 \pm .0058$	$.7430 \pm .0218$	$.9054 \pm .0080$
Human Gene	$\mathbf{.7868 \pm .0007}$	$.7868 \pm .0042^\bullet$	$.7840 \pm .0042$	$.7761 \pm .0040$	$.2707 \pm .0769$	$.7821 \pm .0042$
Natural Scene	$.4753 \pm .0092^\bullet$	$\mathbf{.5073 \pm .0129}$	$.4635 \pm .0127$	$.2529 \pm .0170$	$.2927 \pm .0172$	$.3686 \pm .0408$
SJAFFE	$\mathbf{.8555 \pm .0070}$	$.8481 \pm .0100^\bullet$	$.8440 \pm .0115$	$.2014 \pm .0156$	$.8461 \pm .0098$	$.8346 \pm .0156$
SBU_3DFE	$\mathbf{.8588 \pm .0037}$	$.8574 \pm .0038^\bullet$	$.8385 \pm .0028$	$.7044 \pm .0081$	$.8382 \pm .0037$	$.8351 \pm .0044$
Movie	$\mathbf{.8256 \pm .0021}$	$.8211 \pm .0033^\bullet$	$.6841 \pm .0055$	$.4500 \pm .0098$	$.7391 \pm .0027$	$.6719 \pm .0475$

incomplete annotation. However, although we cannot give a guarantee that IncomLDL-prox does optimize our objective Eq. 1, IncomLDL-prox's performance is comparable with that of IncomLDL-admm, even though IncomLDL-admm get the best results most of the time. We plan to study this phenomenon in our future work.

For those baseline methods, PT-SVM and IIS-LDL perform the best. BFGS-LDL, although reported to be better than IIS-LDL when data are *complete* [Geng, 2016], however, are not suitable for incomplete case, especially on Natural Scene and SJAFFE data sets. Comparing Table 2 and Table 3, we can see that for the incomplete setting, all algorithms except PT-Bayes are more stable. Thus it is much difficult to predict the total unlabeled test data in the transductive setting.

5 Conclusion

In this paper, we solve the problem of Incomplete Label Distribution Learning (IncomLDL) where the supervised information are incomplete. To solve the problem of data deficiency when facing incomplete data, we propose to use the trace norm minimization technique, thus we can exploit label correlation, reducing the effect of lacking training data. Two algorithms are then proposed based on proximal gradient descend and alternative direction method of multipliers. Experiments on all 15 data sets show the merits of the proposed algorithms compared to baselines.

References

[Bi and Kwok, 2014] Wei Bi and James T. Kwok. Multilabel classification with label correlations and missing labels. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI'14), Canada*, pages 1680–1686, 2014.

[Boyd and Vandenberghe, 2004] Stephen P. Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, 2004.

[Boyd *et al.*, 2011] Stephen P. Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[Cai *et al.*, 2010] Jian-Feng Cai, Emmanuel J. Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[Duchi *et al.*, 2008] John C. Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the l_1-ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning (ICML'18), Finland*, pages 272–279, 2008.

[Geng and Hou, 2015] Xin Geng and Peng Hou. Pre-release prediction of crowd opinion on movies by label distribution learning. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI'15), Argentina*, pages 3511–3517, 2015.

[Geng and Ling, 2017] Xin Geng and Miaogen Ling. Soft video parsing by label distribution learning. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI'17), CA*, 2017.

[Geng *et al.*, 2010] Xin Geng, Kate Smith-Miles, and Zhi-Hua Zhou. Facial age estimation by learning from label distributions. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI'10), GA*, 2010.

[Geng *et al.*, 2013] Xin Geng, Chao Yin, and Zhi-Hua Zhou. Facial age estimation by learning from label distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(10):2401–2412, 2013.

[Geng, 2016] Xin Geng. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748, 2016.

[Goldberg *et al.*, 2010] Andrew B. Goldberg, Xiaojin Zhu, Ben Recht, Jun-Ming Xu, and Robert D. Nowak. Transduction with matrix completion: Three birds with one stone. In *Proceedings of 24th Annual Conference on Neural Information Processing Systems (NIPS'10), Canada*, pages 757–765, 2010.

[He and Yuan, 2012] Bingsheng He and Xiaoming Yuan. On the o(1/n) convergence rate of the douglas-rachford alternating direction method. *SIAM J. Numerical Analysis*, 50(2):700–709, 2012.

[Nocedal and Wright, 2006] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, New York, 2nd edition, 2006.

[Pietra *et al.*, 1997] Stephen D. Pietra, Vincent J. D. Pietra, and John D. Lafferty. Inducing features of random fields. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 19(4):380–393, 1997.

[Sun *et al.*, 2010] Yu-Yin Sun, Yin Zhang, and Zhi-Hua Zhou. Multi-label learning with weak label. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI'10), GA*, 2010.

[Tseng, 2008] Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. Technical report, University of Washington, Seattle, 2008.

[Wang and Carreira-Perpiñán, 2013] Weiran Wang and Miguel Á. Carreira-Perpiñán. Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. *CoRR*, abs/1309.1541, 2013.

[Wu *et al.*, 2016] Baoyuan Wu, Siwei Lyu, and Bernard Ghanem. Constrained submodular minimization for missing labels and class imbalance in multi-label learning. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI'16), AZ*, pages 2229–2236, 2016.

[Xing *et al.*, 2016] Chao Xing, Xin Geng, and Hui Xue. Logistic boosting regression for label distribution learning. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16), NV*, pages 4489–4497, 2016.

[Xu *et al.*, 2013] Miao Xu, Rong Jin, and Zhi-Hua Zhou. Speedup matrix completion with side information: Application to multi-label learning. In *Proceedings of 27th Annual Conference on Neural Information Processing Systems (NIPS'13), NV*, pages 2301–2309, 2013.

[Yang *et al.*, 2013] Shu-Jun Yang, Yuan Jiang, and Zhi-Hua Zhou. Multi-instance multi-label learning with weak label. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI'13), China*, pages 1862–1868, 2013.

[Yu *et al.*, 2014] Hsiang-Fu Yu, Prateek Jain, Purushottam Kar, and Inderjit S. Dhillon. Large-scale multi-label learning with missing labels. In *Proceedings of the 31th International Conference on Machine Learning (ICML'14), China*, pages 593–601, 2014.

[Zhao and Guo, 2015] Feipeng Zhao and Yuhong Guo. Semi-supervised multi-label learning with incomplete labels. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI'15) Argentina*, pages 4062–4068, 2015.

[Zhou and Zhang, 2017] Zhi-Hua Zhou and Min-Ling Zhang. Multi-label learning. In Claude Sammut and Geoffrey I. Webb, editors, *Encyclopedia of Machine Learning and Data Mining*, pages 875–881. Springer, 2017.

[Zhou *et al.*, 2015] Ying Zhou, Hui Xue, and Xin Geng. Emotion distribution recognition from facial expressions. In *Proceedings of the 23rd Annual ACM Conference on Multimedia (MM'15), Australia*, pages 1247–1250, 2015.