Multi-Label Learning with Global and Local Label Correlation

Yue Zhu, James T. Kwok, *Fellow, IEEE*, and Zhi-Hua Zhou, *Fellow, IEEE*

Abstract—It is well-known that exploiting label correlations is important to multi-label learning. Existing approaches either assume that the label correlations are global and shared by all instances; or that the label correlations are local and shared only by a data subset. In fact, in the real-world applications, both cases may occur that some label correlations are globally applicable and some are shared only in a local group of instances. Moreover, it is also a usual case that only partial labels are observed, which makes the exploitation of the label correlations much more difficult. That is, it is hard to estimate the label correlations when many labels are absent. In this paper, we propose a new multi-label approach GLOCAL dealing with both the full-label and the missing-label cases, exploiting global and local label correlations simultaneously, through learning a latent label representation and optimizing label manifolds. The extensive experimental studies validate the effectiveness of our approach on both full-label and missing-label data.

Index Terms—Global and local label correlation, label manifold, missing labels, multi-label learning.

1 Introduction

In real-world classification applications, an instance is often associated with more than one class labels. For example, a scene image can be annotated with several tags [5], a document may belong to multiple topics [38], and a piece of music may be associated with different genres [37]. Thus, multi-label learning has attracted a lot of attention in recent years [52].

Label correlations can provide important information in multi-label learning. For example, if labels "amusement park" and "Mickey Mouse" are present, it is very likely that label "Disney" will also appear; Similarly, if "blue sky" and "white cloud" both appear, it is very likely that label "fog" will not be present. Current studies on multi-label learning try to incorporate label correlations of different degrees [52]. However, they mostly focus on global label correlations shared by all instances [13], [20], [32]. In fact, some label correlations are only shared by a local data subset [19], [40]. For example, "apple" is related to "fruit" in gourmet magazines, but is related to "digital devices" in technology magazines. Previous studies focus on exploiting either global or local label correlations. However, considering both of them is obviously more beneficial and desirable.

Another difficulty with label correlations is that they are usually difficult to be specified manually. Usually, they are estimated from observed data. Some approaches assume that the labels are related in the form of a hierarchy, and learn the corresponding label hierarchy by hierarchical clustering [31] or Bayesian network structure learning [3], [47]. However, this assumed hierarchical structure may not exist in some applications. For example, labels "desert", "mountains", "sea", "sunset" and "trees" do not naturally belong to a hierarchy. Others estimate label correlations by label co-occurrence in the training data [30], or equivalently constructing a label kernel [50] . However, it may cause over-fitting. Moreover, co-occurrence estimation is less reliable for labels with very few positive instances.

In multi-label learning, human labelers may sometimes ignore labels they do not know or of little interest, or following the guide by some algorithm to reduce labeling costs [14], [17]. Thus, some labels may be missing from the training set, which is a kind of weakly supervised learning [51]. To address this problem, there have been attempts to recover the missing labels by exploiting label correlations [3], [41]. For example, as labels are correlated, one can assume the label correlation matrix and/or instance-label mapping matrix to have internal linear dependence structure and thus low-rank (i.e., its rank is smaller than its size) [41], [42]. A common approach to encourage this low-rank assumption during inference is by using the nuclear-norm regularizer [6], [21]. However, optimization may be computationally expensive [45]. A more direct approach to enforce this low-rank assumption on the label matrix is by approximating it as a product of two smaller matrices [15], [26].

Though this low-rank structure can be regarded as implicitly exploiting label correlations, it is still desirable to use label correlations explicitly. This has been shown to facilitate the recovery of missing labels [42]. However, estimation of label correlations becomes even more difficult in the presence of missing labels, as the observed label distribution is different from the true one. The aforementioned methods (based on hierarchical clustering and co-occurrence) will produce biased label correlation estimates, which will even do harm to the performance.

In this paper, we propose a new approach called "Multi-Label Learning with **GLO**bal and lo**CAL** Label Correlation" (GLOCAL), which simultaneously recovers the missing labels, trains the linear classifiers, explores and exploits both global and local label correlations. Classifier outputs are encouraged to be similar on highly positively correlated

- *Y. Zhu and Z.-H. Zhou are with National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, 210023, China. E-mail: {zhuy, zhouzh}@lamda.nju.edu.cn*
- *J. T. Kwok is with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong. E-mail: jamesk@cse.ust.hk*

labels, and dissimilar on highly negatively correlated labels. We do not assume the presence of external knowledge sources specifying the label correlations. Instead, these correlations are learned simultaneously with the latent label representations and instance-label mapping.

The rest of the paper is organized as follows. In Section 2, related works of multi-label learning with label correlations are introduced. In Section 3, the problem formulation and the GLOCAL approach are proposed. Experimental results are presented in Section 4. Finally, Section 5 gives some concluding remarks.

Notations For a $n \times m$ matrix $\boldsymbol{A} = [A_{i,j}]$, where $i \in \{1, \cdots, n\}$, $j \in \{1, \cdots, m\}$, $\boldsymbol{A}^\top$, denotes its transpose, $\mathrm{tr}(\boldsymbol{A}) = \sum_{i=1}^{n} A_{i,i}$ is $\boldsymbol{A}$'s trace, $\|\boldsymbol{A}\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} A_{i,j}^2}$ is its Frobenius norm, $\mathrm{diag}(\boldsymbol{A})$ returns a vector containing the diagonal elements of $\boldsymbol{A}$, and $\mathrm{Diag}(\boldsymbol{c})$ returns a diagonal matrix with $\boldsymbol{c}$ on the diagonal. For two matrices of the same size, $\boldsymbol{A}$ and $\boldsymbol{B}$, $\boldsymbol{A} \circ \boldsymbol{B}$ denotes the Hadamard (element-wise) product. For a k-dimensional vector $\boldsymbol{c} = [c_i]$, $\|\boldsymbol{c}\| = \sqrt{\sum_{i=1}^{k} c_i^2}$ is its ℓ_2-norm. $\mathbf{1}$ is an all-one vector.

2 Related Work

2.1 Multi-Label Learning

In multi-label learning, an instance can be associated with multiple class labels. Let $\boldsymbol{C} = \{c_1, \ldots, c_l\}$ be the set of l class labels. We denote the d-dimensional feature vector of an instance by $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^d$, and denote the ground-truth label vector by $\tilde{\boldsymbol{y}} \in \mathcal{Y} \subseteq \{-1, 1\}^l$, where $[\tilde{\boldsymbol{y}}]_j = 1$ if $\boldsymbol{x}$ is with class label c_j, and -1 otherwise.

Multi-label learning has been widely studied in recent years. Based on the degree of label correlations used, it can be divided into three categories [49]: (i) first-order; (ii) second-order; and (iii) high-order. For the first group, label correlations are not considered, and the multi-label problem is transformed into multiple independent binary classification problems. A well-known example is the *binary relevance* (BR) algorithm [5], which trains a classifier independently for each label. For the second group, pairwise label relations are considered. For example, *calibrated label ranking* (CLR) [13] transforms the multi-label learning problem into a pairwise label ranking problem. Finally, for the third group, all other labels' influences imposed on each label are taken into account. For example, *classifier chain* (CC) [32] transforms the multi-label learning problem into a chain of binary classification problems, with the ground-truth labels encoded into the features. Another way to consider all label correlations together is by learning a latent label space to capture higher-level label semantics. Usually, this is obtained by low-rank decomposition of the label matrix [16] (low-rank modeling will be reviewed in Section 2.2). Analogously, Jing et al. [22] used dictionary learning to obtain embedded labels. Recently, Yeh et al. [44] also proposed a deep learning approach to learn a joint feature and label embedding. These methods are highly related to *canonical correlation analysis* (CCA) [12], which learns a latent subspace to align the instance representations with the corresponding labels. An extensive experimental study on various approaches can be found in [28].

Most previous studies focus on global label correlations. However, sometimes label correlations may only be shared by a local data subset. To alleviate this problem, *multi-label learning using local correlation* (MLLOC) [19] extends the feature representation of each instance by embedding a code, which encodes the influence of the instance's labels to the local label correlations. MLLOC achieved success in exploiting local correlations. Its performance may be enhanced if global and local correlations are exploited together, and its application area can be expanded if high-dimensional data and partially observed labels (i.e., some labels may be missing [3], [42], as Section 1 mentioned) can be handled.

Recently, multi-label learning with missing labels has attracted much attention. In this paper, we adopt the general setting that both positive and negative labels can be missing [15], [42], [45]. The observed label vector is denoted by $\boldsymbol{y}$, where $[\boldsymbol{y}]_j = 0$ if class label c_j is not labeled (i.e., missing), and $[\boldsymbol{y}]_j = [\tilde{\boldsymbol{y}}]_j$ otherwise. Given the training data $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$ with n labeled instances, our goal is to learn the mapping $\Psi : \mathcal{X} \to \mathcal{Y}$.

In order to deal with missing labels, *Matrix completion using side information* (MAXIDE) [42] is based on fast low-rank matrix completion, and has strong theoretical guarantees. It focuses on the transductive learning setting, and a label correlation matrix is needed as an input. It would be desirable, if it could be adapted to an inductive setting and learn the label correlation matrix automatically from data. *Low-rank empirical risk minimization for multi-label learning* (LEML) [45] is very efficient, which also relies on a low-rank structure, and works in an inductive learning setting. It does not explicitly use label correlations. *Learning low-rank label correlations for multi-label classification* (ML-LRC) [41] achieves success in capturing global label correlations by adopting a low-rank structure on the label correlation matrix, and addresses the missing labels by introducing a supplementary label matrix. It focuses on global label correlations only, and local label correlations are not considered. Obviously, it would be more desirable to learn both global and local label correlations simultaneously.

2.2 Low-Rank Modeling in Multi-Label Learning

An $n \times m$ matrix is low-rank if its matrix rank is smaller than $\min(n, m)$. Because the labels are correlated in multi-label learning, the label matrix is often assumed to be low-rank [16], [44]. Specifically, let $\tilde{\boldsymbol{Y}} = [\tilde{\boldsymbol{y}}_1, \ldots, \tilde{\boldsymbol{y}}_n] \in \{-1, 1\}^{l \times n}$ be the ground-truth label matrix, where each $\tilde{\boldsymbol{y}}_i$ is the label vector for instance i. Let the rank of $\tilde{\boldsymbol{Y}}$ be $k < l$. $\tilde{\boldsymbol{Y}}$ can then be written as the product

$$\tilde{\boldsymbol{Y}} \simeq \boldsymbol{U}\boldsymbol{V} \tag{1}$$

of two smaller matrices, where $\boldsymbol{U} \in \mathbb{R}^{l \times k}$ and $\boldsymbol{V} \in \mathbb{R}^{k \times n}$. Intuitively, $\boldsymbol{V}$ represents the latent labels capturing higher level concepts that are more compact and semantically abstract than the original labels; while $\boldsymbol{U}$ reflects how the original labels are correlated to the latent labels.

In general, labels may only be partially observed. Low-rank modeling plays a central role in matrix completion [7], and the low-rank decomposition on the observed labels provides a natural solution for missing label recovery. Specificially, let the observed label matrix be $\boldsymbol{Y} =$

$[\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n] \in \{-1, 0, 1\}^{l \times n}$, and Ω be the set containing indices of the observed labels in $\boldsymbol{Y}$ (i.e., indices of the nonzero elements in $\boldsymbol{Y}$). Assume that $\boldsymbol{Y}$ is the same as the ground-truth label matrix $\tilde{\boldsymbol{Y}}$ on the observed labels, and $\tilde{\boldsymbol{Y}}$ can be modeled as $\boldsymbol{UV}$. We focus on minimizing the reconstruction error on the observed labels. In other words, we minimize $\|\Pi_\Omega(\boldsymbol{Y} - \boldsymbol{UV})\|_F^2$, where $[\Pi_\Omega(\boldsymbol{A})]_{i,j} = A_{i,j}$ if $(i, j) \in \Omega$, and 0 otherwise. The full-label task (i.e., all elements observed) can be regarded as a special case with $\boldsymbol{Y} = \tilde{\boldsymbol{Y}}$ and so $\Pi_\Omega(\boldsymbol{Y} - \boldsymbol{UV}) = \tilde{\boldsymbol{Y}} - \boldsymbol{UV}$. After $\boldsymbol{U}$ and $\boldsymbol{V}$ are obtained, a missing label at $(i, j) \notin \Omega$ in $\boldsymbol{Y}$ can be recovered as $\text{sign}(\boldsymbol{u}_{i,:}\boldsymbol{v}_{:,j})$, where $\boldsymbol{u}_{i,:}$ is the ith row of $\boldsymbol{U}$ and $\boldsymbol{v}_{:,j}$ is the jth column of $\boldsymbol{V}$.

2.3 Manifold Regularization

Manifold regularization [1] exploits instance similarity by encouraging predictions on similar instances to be similar. Specifically, let $S_{i,j}$ be the similarity between the ith and jth instances. This is often defined by a Gaussian function or as cosine similarity between the ith and jth instances [10]. The similarities on a set of n instances can then be stored in a $n \times n$ matrix $\boldsymbol{S} = [S_{i,j}]$, which is also known to be positive semidefinite[1]. In a binary classification problem, let f_i and f_j be the label prediction on the ith and jth instances, respectively. The manifold regularizer tries to minimize $\sum_{i,j=1}^n S_{i,j}\|f_i - f_j\|^2 = \boldsymbol{f}^\top \boldsymbol{L}\boldsymbol{f}$, where $\boldsymbol{f}$ is the vector containing predictions on all n instances, and $\boldsymbol{L} = \text{Diag}(\boldsymbol{S}\boldsymbol{1}) - \boldsymbol{S}$ is the Laplacian matrix.

3 The Proposed Approach

In this section, we propose the GLOCAL algorithm, which explores and exploits both global and local label correlations in learning the classifier. It can be used on tasks with either full labels or missing labels. Its success is mainly due to four factors: (1) It uses the low-rank structure of the label matrix to obtain a more compact and abstract latent label representation, which also provides a natural solution to missing label recovery (Section 3.1); (2) It exploits both global and local label correlations, and so the label classifier can utilize information from all labels (Section 3.2); (3) It learns the label correlations directly from data, without the need for mundane and difficult manual specification of the correlation matrix (Section 3.3); (4) It integrates the above into one joint learning problem, and adopts an efficient alternating minimization strategy for optimization (Section 3.4).

3.1 Basic Model

The basic GLOCAL model applies low-rank decomposition on the label matrix to obtain latent labels, and learns a mapping from the feature space to the latent labels. Hence, we can obtain a more compact and abstract latent label representation, which is dense, real-valued, and lower-dimensional. Learning the mapping from feature space to latent label space is also much easier than learning the one to the original label space (which is sparse, binary-valued and higher dimensional). Besides, it directly provides the solution to missing label recovery.

Specifically, we use (1) to decompose the label matrix $\tilde{\boldsymbol{Y}}$ to two low-rank matrices $\boldsymbol{U}$ and $\boldsymbol{U}$, in which $\boldsymbol{V}$ represents the latent labels and $\boldsymbol{U}$ reflects how the original labels are correlated to the latent labels. Matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ can be obtained via minimizing the reconstruction error $\|\tilde{\boldsymbol{Y}} - \boldsymbol{UV}\|_F^2$.

To map instances to the latent labels, we learn a matrix $\boldsymbol{W} \in \mathbb{R}^{d \times k}$. This $\boldsymbol{W}$ can be obtained by minimizing the square loss $\|\boldsymbol{V} - \boldsymbol{W}^\top \boldsymbol{X}\|_F^2$, where $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{d \times n}$ is the matrix containing all the instances. Subsequently, the label predicted for $\boldsymbol{x}$ is $\text{sign}(\boldsymbol{f}(\boldsymbol{x}))$, where $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{W}^\top \boldsymbol{x}$. Let $\boldsymbol{f} = [f_1, \ldots, f_l]^\top$, where $f_j(\boldsymbol{x})$ is the jth predicted label for $\boldsymbol{x}$. We can concatenate $\boldsymbol{f}(\boldsymbol{x})$ for all $\boldsymbol{x} \in \boldsymbol{X}$ together as $\boldsymbol{F}_0 = [\boldsymbol{f}(\boldsymbol{x}_1), \ldots, \boldsymbol{f}(\boldsymbol{x}_n)] = \boldsymbol{W}^\top \boldsymbol{X}$.

Combining reconstruction error minimization for low-rank matrix decomposition and square loss minimization for learning the linear mapping from instances to latent labels, we obtain the following optimization problem for the basic GLOCAL model:

$$\begin{aligned} \min_{\boldsymbol{U},\boldsymbol{V},\boldsymbol{W}} \quad & \|\Pi_\Omega(\boldsymbol{Y} - \boldsymbol{UV})\|_F^2 + \lambda\|\boldsymbol{V} - \boldsymbol{W}^\top \boldsymbol{X}\|_F^2 \\ & + \lambda_2 \mathcal{R}(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}), \end{aligned} \tag{2}$$

where $\mathcal{R}(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W})$ is a regularizer and λ, λ_2 are tradeoff parameters. While the square loss is used in problem (2), it can be replaced by any differentiable loss function.

3.2 Global and Local Manifold Regularizers

Exploiting label correlations is essential to multi-label learning. Here, we use label correlations to regularize the model. Note that global and local label correlations may coexist. In this section, we introduce label manifold regularizers to incorporate both of them.

The basic idea of the global manifold regularizer is adapted from the instance-level manifold regularizer (Section 2.3) [1]. Specifically, the more positively correlated two labels are, the closer the corresponding classifier outputs should be, and vice versa. In other words, positively correlated labels will encourage their corresponding classifier outputs to be similar to each other, while negatively correlated labels will push the corresponding outputs in opposite directions.

Recall that predictions on all n instances are stored in the $l \times n$ matrix $\boldsymbol{F}_0$, with its ith row $\boldsymbol{f}_{i,:}$ containing predictions for the ith label. If the ith and jth labels are more positively correlated, $\boldsymbol{f}_{i,:}$ should be more similar to $\boldsymbol{f}_{j,:}$, and vise versa. Analogous to the instance-level manifold regularizer [1], [29], the label manifold regularizer can be defined as:

$$\sum_{i,j} [\boldsymbol{S}_0]_{i,j} \|\boldsymbol{f}_{i,:} - \boldsymbol{f}_{j,:}\|_2^2, \tag{3}$$

where $\boldsymbol{S}_0$ is the $l \times l$ global label correlation matrix. If labels i and j are positively correlated, $[\boldsymbol{S}_0]_{i,j}$ is also positive. By minimizing (3), $\|\boldsymbol{f}_{i,:} - \boldsymbol{f}_{j,:}\|_2^2$ will be small. Let $\boldsymbol{D}_0$ be the diagonal matrix with diagonal $\boldsymbol{S}_0\boldsymbol{1}$, where $\boldsymbol{1}$ is the vector of ones. The manifold regularizer in (3) can be equivalently written as $\text{tr}(\boldsymbol{F}_0^\top \boldsymbol{L}_0 \boldsymbol{F}_0)$ [27], where $\boldsymbol{L}_0 = \boldsymbol{D}_0 - \boldsymbol{S}_0$ is the $l \times l$ label Laplacian matrix of $\boldsymbol{S}_0$.

As label correlations may vary from one local region to another, we introduce the local manifold regularizer.

1. A matrix $\boldsymbol{M}$ is positive semidefinite (psd) if $\boldsymbol{z}^\top \boldsymbol{M}\boldsymbol{z} \geq 0$ for any nonzero vector $\boldsymbol{z}$.

Assume that the dataset $\boldsymbol{X}$ is partitioned into g groups $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_g\}$, where $\boldsymbol{X}_m \in \mathbb{R}^{d \times n_m}$ has n_m instances. This partitioning can be obtained by domain knowledge (e.g., gene pathways [33] and networks [8] in bioinformatics applications) or clustering. Let $\boldsymbol{Y}_m$ be the label submatrix in $\boldsymbol{Y}$ corresponding to $\boldsymbol{X}_m$, and $\boldsymbol{S}_m \in \mathbb{R}^{l \times l}$ be the local label correlation matrix of group m. Similar to global label correlations, we encourage the classifier outputs to be similar (resp. dissimilar) on the positively (resp. negatively) correlated labels, and minimize $\mathrm{tr}(\boldsymbol{F}_m^\top \boldsymbol{L}_m \boldsymbol{F}_m)$, where $\boldsymbol{L}_m$ is the Laplacian matrix of $\boldsymbol{S}_m$ and $\boldsymbol{F}_m = \boldsymbol{U}\boldsymbol{W}^\top \boldsymbol{X}_m$ is the classifier output matrix for group m.

Adding global and local manifold regularizers to problem (2), we have the following optimization problem:

$$\begin{aligned}\min_{\boldsymbol{U},\boldsymbol{V},\boldsymbol{W}} \quad & \|\Pi_\Omega(\boldsymbol{Y}-\boldsymbol{U}\boldsymbol{V})\|_F^2+\lambda\|\boldsymbol{V}-\boldsymbol{W}^\top\boldsymbol{X}\|_F^2 \\ & +\lambda_2\mathcal{R}(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W}) + \lambda_3\mathrm{tr}(\boldsymbol{F}_0^\top \boldsymbol{L}_0 \boldsymbol{F}_0) \\ & + \sum_{m=1}^{g} \lambda_4 \mathrm{tr}(\boldsymbol{F}_m^\top \boldsymbol{L}_m \boldsymbol{F}_m), \end{aligned} \tag{4}$$

where $\lambda, \lambda_2, \lambda_3, \lambda_4$ are tradeoff parameters.

Global label correlations are encoded in the Laplacian matrix $\boldsymbol{L}_0$ and local label correlations are encoded in the $\boldsymbol{L}_m$'s. Intuitively, a large local group contributes more to the global label correlations. In particular, the following Lemma shows that when the cosine similarity is used to compute $\boldsymbol{S}_{ij}$, we have $\boldsymbol{S}_0 = \sum_{m=1}^{g} \frac{n_m}{n} \boldsymbol{S}_m$.

Lemma 1. *Let* $[\boldsymbol{S}_0]_{ij} = \frac{\boldsymbol{y}_{i,:}\boldsymbol{y}_{j,:}^\top}{\|\boldsymbol{y}_{i,:}\|\|\boldsymbol{y}_{j,:}\|}$ *and* $[\boldsymbol{S}_m]_{ij} = \frac{\boldsymbol{y}_{m,i,:}\boldsymbol{y}_{m,j,:}^\top}{\|\boldsymbol{y}_{m,i,:}\|\|\boldsymbol{y}_{m,j,:}\|}$, *where* $\boldsymbol{y}_{i,:}$ *is the* i*th row of* $\boldsymbol{Y}$, *and* $\boldsymbol{y}_{m,i,:}$ *is the* i*th row of* $\boldsymbol{Y}_m$. *Then,* $\boldsymbol{S}_0 = \sum_{m=1}^{g} \frac{n_m}{n} \boldsymbol{S}_m$.

In general, when the global label correlation matrix is a linear combination of the local label correlation matrices, the following Proposition shows that the corresponding global label Laplacian matrix is also a linear combination of the local label Laplacian matrices with the same combination coefficients.

Proposition 1. *If* $\boldsymbol{S}_0 = \sum_{m=1}^{g} \beta_m \boldsymbol{S}_m$, *then* $\boldsymbol{L}_0 = \sum_{m=1}^{g} \beta_m \boldsymbol{L}_m$.

Using Lemma 1 and Proposition 1, problem (4) can then be rewritten as follows:

$$\begin{aligned}\min_{\boldsymbol{U},\boldsymbol{V},\boldsymbol{W}} \ & \|\Pi_\Omega(\boldsymbol{Y}-\boldsymbol{U}\boldsymbol{V})\|_F^2+\lambda\|\boldsymbol{V}-\boldsymbol{W}^\top\boldsymbol{X}\|_F^2 \\ & + \sum_{m=1}^{g}\left(\frac{\lambda_3 n_m}{n}\mathrm{tr}(\boldsymbol{F}_0^\top \boldsymbol{L}_m \boldsymbol{F}_0) + \lambda_4 \mathrm{tr}(\boldsymbol{F}_m^\top \boldsymbol{L}_m \boldsymbol{F}_m)\right) \\ & +\lambda_2\mathcal{R}(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W}). \end{aligned} \tag{5}$$

3.3 Learning Label Correlations

The success of label manifold regularization hinges on a good label correlation matrix (or equivalently, a good label Laplacian matrix). In multi-label learning, one rudimentary approach is to compute the correlation coefficient between two labels by the cosine distance [39]. However, the estimate can be noisy since some labels may only have very few positive instances in the training data. When labels can be missing, this estimate may even become misleading, since the observed label distribution can be very different from the true label distribution.

In this paper, instead of specifying any correlation metric or label correlation matrix, we learn the Laplacian matrices directly. Note that the Laplacian matrices are symmetric positive definite. Thus, for $m \in \{1, \ldots, g\}$, we decompose $\boldsymbol{L}_m$ as $\boldsymbol{Z}_m \boldsymbol{Z}_m^\top$, where $\boldsymbol{Z}_m \in \mathbb{R}^{l \times k}$. For simplicity, k is set to the dimensionality of the latent representation $\boldsymbol{V}$. As a result, learning Laplacian matrices is transformed to learning $\boldsymbol{\mathcal{Z}} \equiv \{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_g\}$. Note that optimization w.r.t. $\boldsymbol{Z}_m$ may lead to the trivial solution $\boldsymbol{Z}_m = \boldsymbol{0}$. ~~To avoid this problem, we add the constraint that the diagonal elements in each $\boldsymbol{Z}_m \boldsymbol{Z}_m^\top$ are 1~~. This also enables us to obtain a normalized Laplacian matrix [9] of $\boldsymbol{L}_m$.

Let $\boldsymbol{J} = [J_{ij}]$ be the indicator matrix with $J_{ij} = 1$ if $(i,j) \in \Omega$, and 0 otherwise. $\Pi_\Omega(\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{V})$ can be rewritten as the Hadamard product $\boldsymbol{J} \circ (\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{V})$. Combining the decomposition of Laplacian matrices and the diagonal constraints of $\boldsymbol{Z}_m$, we obtain the optimization problem:

$$\begin{aligned}\min_{\boldsymbol{U},\boldsymbol{V},\boldsymbol{W},\boldsymbol{\mathcal{Z}}} & \|\boldsymbol{J}\circ(\boldsymbol{Y}-\boldsymbol{U}\boldsymbol{V})\|_F^2+\lambda\|\boldsymbol{V}-\boldsymbol{W}^\top\boldsymbol{X}\|_F^2 \\ & + \sum_{m=1}^{g}\left(\frac{\lambda_3 n_m}{n}\mathrm{tr}\left(\boldsymbol{F}_0^\top \boldsymbol{Z}_m \boldsymbol{Z}_m^\top \boldsymbol{F}_0\right) + \lambda_4 \mathrm{tr}(\boldsymbol{F}_m^\top \boldsymbol{Z}_m \boldsymbol{Z}_m^\top \boldsymbol{F}_m)\right) \\ & +\lambda_2\mathcal{R}(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W}) \\ \text{s.t.} \quad & \mathrm{diag}(\boldsymbol{Z}_m \boldsymbol{Z}_m^\top) = \boldsymbol{1}, m = 1, 2, \ldots, g. \end{aligned} \tag{6}$$

Moreover, we will use the standard regularizer $\mathcal{R}(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W}) = \|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2 + \|\boldsymbol{W}\|_F^2$ in the sequel.

3.4 Learning by Alternating Minimization

Problem (6) can be solved by alternating minimization (Algorithm 1), which enables us to iteratively adjust the variables to find a satisfying solution. In each iteration, we update one of the variables in $\{\boldsymbol{\mathcal{Z}}, \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}\}$ with gradient descent, and fix the others. The whole optimization problem is then reduced to several simpler subproblems that are easier to solve. Specifically, the MANOPT toolbox [4] is utilized to implement gradient descent with line search on the Euclidean space for the update of $\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}$, and on the manifolds for the update of $\boldsymbol{\mathcal{Z}}$. The detailed update procedures for $\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}$, and $\boldsymbol{\mathcal{Z}}$ will be discussed in the following sections.

3.4.1 Updating $\boldsymbol{Z}_m$ (Line 4 in Algorithm 1)

With $\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}$ fixed, problem (6) reduces to

$$\begin{aligned}\min_{\boldsymbol{Z}_m} \quad & \frac{\lambda_3 n_m}{n}\mathrm{tr}\left(\boldsymbol{F}_0^\top \boldsymbol{Z}_m \boldsymbol{Z}_m^\top \boldsymbol{F}_0\right) + \lambda_4 \mathrm{tr}\left(\boldsymbol{F}_m^\top \boldsymbol{Z}_m \boldsymbol{Z}_m^\top \boldsymbol{F}_m\right) \\ \text{s.t.} \quad & \mathrm{diag}(\boldsymbol{Z}_m \boldsymbol{Z}_m^\top) = \boldsymbol{1}, \end{aligned} \tag{7}$$

for each $m \in \{1, \ldots, g\}$. Due to the constraint $\mathrm{diag}(\boldsymbol{Z}_m \boldsymbol{Z}_m^\top) = \boldsymbol{1}$, it has no closed-form solution, and we solve it with projected gradient descent. The gradient of the objective w.r.t. $\boldsymbol{Z}_m$ is

$$\begin{aligned}\nabla_{\boldsymbol{Z}_m} = \ & \frac{\lambda_3 n_m}{n}\boldsymbol{U}\boldsymbol{W}^\top\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{U}^\top\boldsymbol{Z}_m \\ & +\lambda_4\boldsymbol{U}\boldsymbol{W}^\top\boldsymbol{X}_m\boldsymbol{X}_m^\top\boldsymbol{W}\boldsymbol{U}^\top\boldsymbol{Z}_m. \end{aligned}$$

To satisfy the constraint $\mathrm{diag}(\boldsymbol{Z}_m \boldsymbol{Z}_m^\top) = \boldsymbol{1}$, we project each row of $\boldsymbol{Z}_m$ onto the unit norm ball after each update:

$$\boldsymbol{z}_{m,j,:} \leftarrow \boldsymbol{z}_{m,j,:} / \|\boldsymbol{z}_{m,j,:}\|,$$

Algorithm 1 Multi-Label Learning with **GLO**bal and lo**CAL** Correlation" (GLOCAL).

Input: data matrix $\boldsymbol{X}$, label matrix $\boldsymbol{Y}$, observation indicator matrix $\boldsymbol{J}$, and the group partition
Output: $\boldsymbol{U}, \boldsymbol{W}, \boldsymbol{\mathcal{Z}}$.

```
initialize U, V, W, Z;
repeat
    for m = 1, ..., g
        //learn label correlations
        fix V, U, W, update Z_m by solving (7);
    end for
    //learn latent labels
    fix U, W, Z, update V by solving (8);
    //learn mapping from original labels to latent labels
    fix V, W, Z, update U by solving (9);
    //learn mapping from instance to latent labels
    fix U, V, Z, update W by solving (10);
until convergence;
output U, : bmV, W, and Z ≡ {Z_1, ..., Z_g}.
```

where $\boldsymbol{z}_{m,j,:}$ is the jth row of $\boldsymbol{Z}_m$.

3.4.2 Updating $\boldsymbol{V}$ (Line 6 in Algorithm 1)

With $\boldsymbol{Z}_m$'s and $\boldsymbol{U}, \boldsymbol{W}$ fixed, problem (6) reduces to

$$\min_{\boldsymbol{V}} \|\boldsymbol{J} \circ (\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{V})\|_F^2 + \lambda \|\boldsymbol{V} - \boldsymbol{W}^\top \boldsymbol{X}\|_F^2 + \lambda_2 \|\boldsymbol{V}\|_F^2. \quad (8)$$

Notice that the columns of $\boldsymbol{V}$ are independent of each other, and thus $\boldsymbol{V}$ can be obtained column by column. Let $\boldsymbol{j}_i$ and $\boldsymbol{v}_i$ be ith column of $\boldsymbol{J}$ and $\boldsymbol{V}$, respectively. The optimization problem for $\boldsymbol{v}_i$ can be written as:

$$\min_{\boldsymbol{v}_i} \|\mathrm{Diag}(\boldsymbol{j}_i)\boldsymbol{y}_i - \mathrm{Diag}(\boldsymbol{j}_i)\boldsymbol{U}\boldsymbol{v}_i\|^2 + \lambda \|\boldsymbol{v}_i - \boldsymbol{W}^\top \boldsymbol{x}_i\|^2 + \lambda_2 \|\boldsymbol{v}_i\|^2.$$

Setting the gradient w.r.t. $\boldsymbol{v}_i$ to 0, we obtain the following closed-form solution of $\boldsymbol{v}_i$:

$$\boldsymbol{v}_i = \left(\boldsymbol{U}^\top \mathrm{Diag}(\boldsymbol{j}_i)\boldsymbol{U} + (\lambda + \lambda_2)\mathbf{I}\right)^{-1} \left(\lambda \boldsymbol{W}^\top \boldsymbol{x}_i + \boldsymbol{U}^\top \mathrm{Diag}(\boldsymbol{j}_i)\boldsymbol{y}_i\right).$$

This involves computing a matrix inverse for each i. If this is expensive, we can perform gradient descent on (8) instead. The gradient of the objective in (8) w.r.t. $\boldsymbol{V}$ is

$$\nabla_{\boldsymbol{V}} = \boldsymbol{U}^\top \left(\boldsymbol{J} \circ (\boldsymbol{U}\boldsymbol{V} - \boldsymbol{Y})\right) + \lambda(\boldsymbol{V} - \boldsymbol{W}^\top \boldsymbol{X}) + \lambda_2 \boldsymbol{V}.$$

3.4.3 Updating $\boldsymbol{U}$ (Line 7 in Algorithm 1)

With $\boldsymbol{Z}_m$'s and $\boldsymbol{V}, \boldsymbol{W}$ fixed, problem (6) reduces to

$$\min_{\boldsymbol{U}} \|\boldsymbol{J} \circ (\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{V})\|_F^2 + \lambda_2 \|\boldsymbol{U}\|_F^2 + \sum_{m=1}^{g} \Big(\frac{\lambda_3 n_m}{n} \mathrm{tr}(\boldsymbol{F}_0^\top \boldsymbol{Z}_m \boldsymbol{Z}_m^\top \boldsymbol{F}_0) + \lambda_4 \mathrm{tr}(\boldsymbol{F}_m^\top \boldsymbol{Z}_m \boldsymbol{Z}_m^\top \boldsymbol{F}_m) \Big). \quad (9)$$

Again, we use gradient descent, and the gradient w.r.t. $\boldsymbol{U}$ is:

$$\nabla_{\boldsymbol{U}} = \left(\boldsymbol{J} \circ (\boldsymbol{U}\boldsymbol{V} - \boldsymbol{Y})\right)\boldsymbol{V}^\top + \lambda_2 \boldsymbol{U} + \sum_{m=1}^{g} \boldsymbol{Z}_i \boldsymbol{Z}_i^\top \boldsymbol{U} \Big(\frac{\lambda_3 n_m}{n} \boldsymbol{W}^\top \boldsymbol{X}_m \boldsymbol{X}_m^\top \boldsymbol{W} + \lambda_4 \boldsymbol{W}^\top \boldsymbol{X} \boldsymbol{X}^\top \boldsymbol{W} \Big).$$

3.4.4 Updating $\boldsymbol{W}$ (Line 8 in Algorithm 1)

With $\boldsymbol{Z}_m$'s and $\boldsymbol{U}, \boldsymbol{V}$ fixed, problem (6) reduces to

$$\min_{\boldsymbol{W}} \lambda \|\boldsymbol{V} - \boldsymbol{W}^\top \boldsymbol{X}\|_F^2 + \lambda_2 \|\boldsymbol{W}\|_F^2 + \sum_{m=1}^{g} \Big(\frac{\lambda_3 n_m}{n} \mathrm{tr}(\boldsymbol{F}_0^\top \boldsymbol{Z}_m \boldsymbol{Z}_m^\top \boldsymbol{F}_0) + \lambda_4 \mathrm{tr}(\boldsymbol{F}_m^\top \boldsymbol{Z}_m \boldsymbol{Z}_m^\top \boldsymbol{F}_m) \Big). \quad (10)$$

The gradient w.r.t. $\boldsymbol{W}$ is:

$$\nabla_{\boldsymbol{W}} = \lambda \boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{W} - \boldsymbol{V}^\top \right) + \lambda_2 \boldsymbol{W} + \sum_{m=1}^{g} \Big(\frac{\lambda_3 n_m}{n} \boldsymbol{X}\boldsymbol{X}^\top + \lambda_4 \boldsymbol{X}_m \boldsymbol{X}_m^\top \Big) \boldsymbol{W}\boldsymbol{U}^\top \boldsymbol{Z}_m \boldsymbol{Z}_m^\top \boldsymbol{U}.$$

4 Experiments

In this section, extensive experiments are performed on text and image datasets. Performance on both the full-label case and the missing-label case are discussed

4.1 Learning with Full Labels

In this experiment, we consider the simplest case that all elements in the training label matrix are observed, i.e. $\boldsymbol{J}$ in problem (6) is an all one matrix.

4.1.1 Data sets

In our experiments, we conduct comparisons on several commonly used benchmark datasets for multi-label learning tasks on text and image classification to validate the effectiveness of our approach. Specifically, on text, eleven Yahoo datasets[2] (Arts, Business, Computers, Education, Entertainment, Health, Recreation, Reference, Science, Social and Society) and the Enron dataset[3] are used. On images, the Corel5k[3] and Image[4] datasets are used. In the sequel, each dataset is denoted by its first three letters.[5] These datasets have been widely used in multi-label learning literatures [2], [16], [18], [19], [25], [28], [32], [34], [41], [42], [43], [46], [47], [48] etc. Detailed information of the datasets are shown in Table 1. For each dataset, we randomly select 60% of the instances for training, and the rest for testing.

4.1.2 Performance Evaluation

Let p be the number of test instances, $\boldsymbol{C}_i^+, \boldsymbol{C}_i^-$ be the sets of positive and negative labels associated with the ith instance; and $\boldsymbol{Z}_j^+, \boldsymbol{Z}_j^-$ be the sets of positive and negative instances belonging to the jth label. Given input $\boldsymbol{x}$, let $\mathrm{rank}_{\boldsymbol{f}}(\boldsymbol{x}, y)$ be the rank of label y in the predicted label ranking (sorted in descending order).

For performance evaluation, we use the following popular metrics in multi-label learning [52]:

1) Ranking loss (Rkl): This is the fraction that a negative label is ranked higher than a positive label. For instance i, define $\boldsymbol{Q}_i = \{(j', j'') \mid f_{j'}(\boldsymbol{x}_i) \le f_{j''}(\boldsymbol{x}_i), (j', j'') \in \boldsymbol{C}_i^+ \times \boldsymbol{C}_i^-\}$. Then, Rkl $= \frac{1}{p} \sum_{i=1}^{p} \frac{|\boldsymbol{Q}_i|}{|\boldsymbol{C}_i^+||\boldsymbol{C}_i^-|}$.
2) Average Area Under the ROC Curve (Auc): This is the fraction that a positive instance is ranked higher than

2. http://www.kecl.ntt.co.jp/as/members/ueda/yahoo.tar
3. http://mulan.sourceforge.net/datasets-mlc.html
4. http://cse.seu.edu.cn/people/zhangml/files/Image.rar
5. "Society" is denoted "Soci", so as to distinguish it from "Social".

Table 1
Datasets used in the experiments ("#instance" is the number of instances, "#dim" is the feature dimensionality, "#label" is the total size of the class label set, and "#label/instance" is the average number of labels possessed by each instance).

	#instance	#dim	#label	#label/instance		#instance	#dim	#label	#label/instance
Arts	5,000	462	26	1.64	Business	5,000	438	30	1.59
Computers	5,000	681	33	1.51	Education	5,000	550	33	1.46
Entertainment	5,000	640	21	1.42	Health	5,000	612	32	1.66
Recreation	5,000	606	22	1.42	Reference	5,000	793	33	1.17
Science	5,000	743	40	1.45	Social	5,000	1,047	39	1.28
Society	5,000	636	27	1.69	Enron	1,702	1,001	53	3.37
Corel5k	5,000	499	374	3.52	Image	2,000	294	5	1.24

a negative instance, averaged over all labels. Specifically, for label j, define $\boldsymbol{Q}_j = \{(i', i'') \mid f_j(\boldsymbol{x}_{i'}) \geq f_j(\boldsymbol{x}_{i''}), (\boldsymbol{x}_{i'}, \boldsymbol{x}_{i''}) \in \boldsymbol{Z}_j^+ \times \boldsymbol{Z}_j^-\}$. Then, $\text{Auc} = \frac{1}{l}\sum_{j=1}^{l} \frac{|\tilde{\boldsymbol{Q}}_j|}{|\boldsymbol{Z}_j^+||\boldsymbol{Z}_j^-|}$.

3) Coverage (Cvg): This counts how many steps are needed to move down the predicted label ranking so as to cover all the positive labels of the instances. $\text{Cvg} = \frac{1}{p}\sum_{i=1}^{p} \max\{\text{rank}_{\boldsymbol{f}}(\boldsymbol{x}_i, j) \mid j \in \boldsymbol{C}_i^+\} - 1$.
4) Average precision (Ap): This is the average fraction of positive labels ranked higher than a particular positive label. For instance i, define $\hat{\boldsymbol{Q}}_{i,c} = \{j \mid \text{rank}_{\boldsymbol{f}}(\boldsymbol{x}_i, j) \leq \text{rank}_{\boldsymbol{f}}(\boldsymbol{x}_i, c), j \in \boldsymbol{C}_i^+\}$. Then, $\text{Ap} = \frac{1}{p}\sum_{i=1}^{p} \frac{1}{|\boldsymbol{C}_i^+|} \sum_{c \in \boldsymbol{C}_i^+} \frac{|\hat{\boldsymbol{Q}}_{i,c}|}{\text{rank}_{\boldsymbol{f}}(\boldsymbol{x}_i, c)}$.

For Auc and Ap, the higher the better; whereas for Rkl and Cvg, the lower the better. To reduce statistical variability, results are averaged over 10 independent repetitions.

4.1.3 Baselines

In the GLOCAL algorithm, we use the kmeans clustering algorithm to partition the data into local groups. The solution of problem (2) is used to warm-start $\boldsymbol{U}, \boldsymbol{V}$ and $\boldsymbol{W}$. The $\boldsymbol{\mathcal{Z}}_m$'s are randomly initialized. GLOCAL is compared with the following state-of-the-art multi-label learning algorithms:

1) Binary relevance (BR) [5]: It trains a binary linear SVM (using the LIBLINEAR package[6] [11]) for each label independently;
2) Hierarchy of multi-label classifiers (HOMER) [35]: It transforms a multi-label classification task into a tree hierarchy of simpler multi-label classification subtasks, with each subtask handling only a smaller number of labels. In the experiments, we use the implementation in the Mulan package[7] [36].
3) Random forest with predictive clustering trees (RF-PCT) [23], [24]: It is an ensemble of predictive clustering trees that maximizes cluster homogeneity in the partition at each internal node. We use the implementation in the Clus package[8].
4) Multi-label learning using local correlation (MLLOC) [19]: It exploits local label correlations by encoding them into the instance's feature representation;
5) Low-rank empirical risk minimization for multi-label learning (LEML) [45]: It learns a linear instance-to-label mapping with low-rank structure, and implicitly takes advantage of global label correlation;
6) Learning low-rank label correlations for multi-label classification (ML-LRC) [41]: It learns and exploits low-rank global label correlations for multi-label classification with missing labels.

On the aspect of exploiting label correlations, BR and RF-PCT do not take label correlation into account; HOMER exploits global label correlations by building a hierarchy of meta-labels (which are subsets of the original labels); MLLOC considers only local label correlations; LEML implicitly uses global label correlations; ML-LRC models global label correlation matrix directly; and GLOCAL exploits both global and local label correlations. On the ability to handle missing labels, BR, MLLOC, HOMER, and RF-PCT can only learn with full labels, while LEML, ML-LRC, and GLOCAL can handle missing labels.

For simplicity, we set $\lambda = 1$ in GLOCAL. The other parameters, as well as those of the baseline methods, are selected via 5-fold cross-validation on the training set. HOMER and RF-PCT are implemented in JAVA, while the other approaches are implemented in Matlab (with some C++ code for BR and LEML).

4.1.4 Results

Performance on the test data is shown in Table 2. As can be seen, GLOCAL is better than the other compared approaches in general. On most datasets, GLOCAL is among the best two approaches on all measures. The success of GLOCAL is due to simultaneous optimizing the low-rank decomposition of label matrix, feature space mapping to latent labels, and Laplacian matrices encoding both global and local label correlations. With the low-rank decomposition of label matrix, we obtain more compact and informative latent labels. It is easier to learn the mapping from feature space to the dense, real-valued, lower-dimensional latent label space than that to the sparse, binary-valued, higher-dimensional original label space. This is especially the case in the presence of minority labels with few positive instances. Besides, the global label manifold provides information on how labels are correlated as a whole, and helps learning of the minority labels. If a minority label is positively (resp. negatively) correlated to the other labels, we can encourage its label classifier outputs to be more similar (resp. dissimilar) to those of the other labels. The local label manifold further allows local adaptation of the label classifiers. The learning of Laplacian matrices discovers label correlations that best fits the global and local data subsets, and avoids the often mandane and difficult task of manually specifying label correlations. GLOCAL obtains slightly worse results on the datasets of Computers, Health, and Entron. We speculate

6. https://www.csie.ntu.edu.tw/∼ cjlin/liblinear
7. http://mulan.sourceforge.net
8. http://clus.sourceforge.net

Table 2
Results for learning with full labels on ranking loss(Rkl), average auc(Auc), coverage(Cvg) and average precision(Ap). Rkl and Cvg are the smaller the better, Auc and Ap are the larger the better. The best results are highlighted. The italics indicates that GLOCAL is significantly better (paired t-tests at 95% significance level). The number in brackets shows ranking of the algorithm.

		BR	RF-PCT	HOMER	MLLOC	LEML	ML-LRC	GLOCAL
Art	Rkl	*0.201±0.005*(7)	*0.190±0.003*(5)	*0.199±0.004*(6)	*0.177±0.013*(4)	*0.170±0.005*(3)	*0.157±0.002*(2)	**0.138±0.002**(1)
	Auc	*0.799±0.006*(6)	*0.827±0.001*(4)	*0.756±0.008*(7)	*0.823±0.013*(5)	*0.833±0.005*(3)	0.843±0.001(2)	**0.846±0.005**(1)
	Cvg	*7.347±0.196*(6)	*6.115±0.074*(3)	*8.068±0.187*(7)	*6.762±0.344*(5)	*6.337±0.243*(4)	*5.529±0.037*(2)	**5.347±0.146**(1)
	Ap	*0.594±0.006*(5)	*0.516±0.006*(7)	*0.596±0.009*(4)	*0.606±0.006*(3)	*0.590±0.005*(6)	*0.600±0.007*(2)	**0.619±0.005**(1)
Bus	Rkl	*0.072±0.005*(6)	*0.070±0.002*(5)	*0.075±0.004*(7)	*0.055±0.009*(3)	*0.056±0.005*(4)	**0.044±0.002**(1)	**0.044±0.002**(1)
	Auc	*0.928±0.005*(5)	*0.944±0.002*(4)	*0.895±0.006*(6)	*0.944±0.008*(4)	*0.945±0.005*(3)	0.950±0.005(2)	**0.955±0.003**(1)
	Cvg	*4.087±0.268*(6)	*2.810±0.111*(3)	*5.110±0.219*(7)	*3.265±0.464*(5)	*3.187±0.270*(4)	2.560±0.059(2)	**2.559±0.169**(1)
	Ap	*0.861±0.007*(6)	*0.862±0.004*(5)	*0.824±0.005*(7)	0.878±0.011(2)	*0.867±0.007*(4)	*0.870±0.005*(3)	**0.883±0.004**(1)
Com	Rkl	0.146±0.007(6)	**0.104±0.005**(1)	*0.164±0.004*(7)	*0.134±0.014*(4)	*0.138±0.004*(5)	0.107±0.002(2)	0.107±0.002(2)
	Auc	0.854±0.007(6)	**0.901±0.004**(1)	0.796±0.005(7)	0.866±0.014(5)	0.895±0.002(2)	0.894±0.002(4)	0.895±0.002(2)
	Cvg	*6.654±0.236*(6)	**4.852±0.316**(1)	*8.663±0.195*(7)	*6.224±0.480*(5)	*6.148±0.183*(4)	4.893±0.142(3)	4.889±0.058(2)
	Ap	0.680±0.007(5)	**0.708±0.002**(1)	0.647±0.006(7)	0.689±0.009(3)	0.669±0.007(6)	0.689±0.005(3)	0.698±0.004(2)
Edu	Rkl	*0.203±0.010*(6)	0.100±0.007(3)	*0.240±0.026*(7)	*0.158±0.021*(5)	*0.145±0.008*(4)	0.099±0.002(2)	**0.095±0.002**(1)
	Auc	*0.797±0.102*(6)	**0.898±0.001**(1)	*0.732±0.028*(7)	*0.842±0.022*(5)	*0.859±0.008*(4)	0.868±0.006(3)	0.878±0.006(2)
	Cvg	*8.979±0.487*(6)	4.863±0.259(3)	*10.932±0.936*(7)	*7.381±0.765*(5)	*6.711±0.364*(4)	4.531±0.104(2)	**4.529±0.206**(1)
	Ap	*0.580±0.010*(7)	0.626±0.009(2)	*0.587±0.029*(6)	*0.613±0.004*(3)	*0.596±0.009*(5)	*0.600±0.007*(4)	**0.628±0.009**(1)
Ent	Rkl	*0.185±0.006*(6)	**0.108±0.005**(1)	*0.207±0.011*(7)	*0.146±0.013*(4)	*0.154±0.005*(5)	*0.130±0.005*(3)	**0.108±0.004**(1)
	Auc	*0.815±0.006*(6)	**0.874±0.006**(1)	*0.777±0.011*(7)	*0.854±0.013*(4)	*0.852±0.005*(5)	0.871±0.003(3)	**0.874±0.005**(1)
	Cvg	*5.006±0.160*(6)	**3.043±0.123**(1)	*6.445±0.157*(7)	*4.293±0.344*(5)	*4.193±0.139*(4)	*3.505±0.125*(3)	3.114±0.110(2)
	Ap	*0.662±0.009*(4)	**0.683±0.006**(1)	*0.591±0.008*(7)	0.670±0.005(3)	*0.647±0.007*(6)	*0.661±0.012*(5)	0.681±0.008(2)
Hea	Rkl	*0.113±0.001*(6)	**0.059±0.007**(1)	*0.236±0.011*(7)	*0.093±0.005*(5)	*0.091±0.003*(4)	*0.071±0.003*(3)	0.065±0.002(2)
	Auc	*0.886±0.003*(6)	0.916±0.005(3)	*0.778±0.010*(7)	*0.907±0.005*(5)	0.913±0.004(4)	**0.929±0.009**(1)	0.923±0.007(2)
	Cvg	*6.193±0.059*(6)	**3.614±0.521**(1)	*6.467±0.190*(7)	*5.403±0.157*(5)	*5.063±0.128*(4)	3.751±0.128(2)	3.858±0.131(3)
	Ap	*0.763±0.002*(4)	**0.785±0.008**(1)	*0.631±0.010*(7)	*0.777±0.004*(3)	*0.750±0.003*(6)	*0.755±0.006*(5)	0.782±0.001(2)
Rec	Rkl	*0.197±0.003*(6)	0.159±0.004(2)	*0.223±0.010*(7)	*0.184±0.015*(4)	*0.185±0.001*(5)	*0.170±0.004*(3)	**0.155±0.002**(1)
	Auc	0.802±0.003(6)	0.838±0.003(2)	0.753±0.008(7)	0.816±0.015(5)	0.822±0.002(4)	*0.833±0.004*(3)	**0.840±0.000**(1)
	Cvg	*5.506±0.089*(6)	4.543±0.085(3)	*6.747±0.291*(7)	*5.268±0.333*(5)	*5.110±0.040*(4)	4.515±0.045(2)	**4.431±0.048**(1)
	Ap	*0.609±0.005*(4)	0.623±0.005(2)	*0.589±0.013*(7)	0.620±0.004(3)	*0.595±0.004*(6)	*0.604±0.003*(5)	**0.625±0.004**(1)
Ref	Rkl	*0.155±0.005*(7)	*0.093±0.003*(3)	*0.143±0.007*(6)	*0.138±0.008*(5)	*0.137±0.004*(4)	*0.092±0.003*(2)	**0.086±0.003**(1)
	Auc	*0.845±0.005*(7)	*0.873±0.001*(3)	*0.867±0.007*(5)	*0.862±0.008*(6)	*0.872±0.004*(4)	**0.900±0.006**(1)	0.894±0.004(2)
	Cvg	*6.171±0.219*(7)	*4.521±0.422*(3)	*5.349±0.208*(5)	*5.514±0.309*(6)	*5.277±0.171*(4)	3.438±0.133(2)	**3.387±0.118**(1)
	Ap	0.685±0.005(3)	*0.671±0.004*(5)	0.684±0.004(4)	**0.688±0.003**(1)	*0.667±0.003*(6)	*0.667±0.007*(6)	**0.688±0.007**(1)
Sci	Rkl	*0.197±0.009*(6)	*0.165±0.004*(3)	*0.225±0.012*(7)	*0.166±0.017*(4)	*0.170±0.005*(5)	*0.131±0.002*(2)	**0.118±0.003**(1)
	Auc	*0.802±0.010*(6)	*0.849±0.003*(3)	*0.748±0.016*(7)	0.834±0.018(4)	*0.834±0.005*(4)	**0.860±0.003**(1)	0.853±0.010(2)
	Cvg	*10.189±0.435*(6)	*8.616±0.125*(3)	*12.337±0.648*(7)	*8.867±0.751*(4)	*8.885±0.197*(5)	*6.704±0.122*(2)	**6.434±0.137**(1)
	Ap	*0.568±0.012*(4)	0.573±0.003(3)	*0.520±0.012*(7)	**0.581±0.009**(1)	*0.551±0.008*(6)	*0.561±0.009*(5)	0.580±0.009(2)
Soc	Rkl	*0.112±0.001*(6)	**0.075±0.006**(1)	*0.129±0.009*(7)	*0.094±0.013*(4)	*0.106±0.006*(5)	**0.075±0.005**(1)	**0.075±0.005**(1)
	Auc	*0.888±0.002*(6)	0.907±0.005(3)	*0.869±0.007*(7)	0.906±0.013(4)	*0.894±0.006*(5)	**0.917±0.005**(1)	0.915±0.005(2)
	Cvg	*6.036±0.125*(7)	*5.064±0.176*(3)	*5.994±0.298*(6)	*5.147±0.401*(4)	*5.521±0.301*(5)	4.651±0.102(2)	**4.537±0.258**(1)
	Ap	0.724±0.005(6)	0.743±0.003(4)	0.744±0.011(3)	**0.764±0.008**(1)	0.731±0.005(5)	0.719±0.003(7)	0.758±0.008(2)
Soci	Rkl	*0.204±0.004*(6)	*0.149±0.003*(3)	*0.233±0.011*(7)	*0.182±0.006*(4)	*0.182±0.007*(4)	0.142±0.002(2)	**0.136±0.005**(1)
	Auc	*0.796±0.005*(6)	0.840±0.002(2)	*0.730±0.009*(7)	*0.818±0.006*(5)	*0.822±0.008*(4)	0.840±0.006(2)	**0.844±0.006**(1)
	Cvg	*8.048±0.108*(6)	*6.443±0.217*(3)	*9.499±0.314*(7)	*7.392±0.216*(4)	*7.438±0.162*(5)	5.973±0.108(2)	**5.852±0.194**(1)
	Ap	*0.610±0.007*(3)	*0.608±0.003*(4)	*0.605±0.009*(5)	0.623±0.004(2)	*0.599±0.006*(7)	*0.605±0.006*(5)	**0.633±0.009**(1)
Enr	Rkl	*0.194±0.006*(7)	**0.125±0.003**(1)	*0.187±0.008*(6)	*0.169±0.012*(5)	*0.159±0.005*(4)	*0.133±0.004*(3)	**0.125±0.004**(1)
	Auc	*0.806±0.006*(6)	**0.879±0.002**(1)	*0.788±0.013*(7)	*0.831±0.009*(5)	*0.851±0.006*(4)	*0.869±0.004*(3)	0.877±0.005(2)
	Cvg	*23.618±0.450*(6)	**15.954±0.253**(1)	*24.412±0.736*(7)	*21.724±0.950*(5)	*18.531±0.707*(4)	16.654±0.198(2)	16.737±0.622(3)
	Ap	*0.575±0.006*(7)	**0.653±0.005**(1)	*0.599±0.015*(4)	*0.586±0.009*(6)	*0.600±0.004*(3)	*0.591±0.004*(5)	0.647±0.006(2)
Cor	Rkl	*0.271±0.006*(6)	*0.180±0.004*(3)	*0.353±0.007*(7)	*0.230±0.012*(4)	*0.246±0.004*(5)	**0.170±0.002**(1)	0.173±0.005(2)
	Auc	*0.699±0.006*(6)	*0.788±0.002*(3)	*0.640±0.007*(7)	*0.757±0.012*(4)	*0.754±0.005*(5)	0.825±0.005(2)	**0.827±0.005**(1)
	Cvg	*261.99±3.15*(7)	*182.83±1.71*(3)	*207.87±0.01*(6)	*201.80±6.71*(5)	*184.58±1.72*(4)	137.31±2.49(2)	**136.91±3.21**(1)
	Ap	*0.153±0.001*(7)	0.197±0.002(3)	*0.180±0.007*(6)	*0.182±0.005*(5)	*0.188±0.004*(4)	0.198±0.003(2)	**0.200±0.004**(1)
Ima	Rkl	0.181±0.011(6)	0.183±0.006(7)	0.180±0.001(2)	0.180±0.008(2)	0.181±0.012(5)	0.180±0.009(2)	**0.179±0.004**(1)
	Auc	0.812±0.011(5)	**0.819±0.006**(1)	0.815±0.020(3)	0.810±0.012(4)	*0.786±0.005*(6)	*0.748±0.010*(7)	**0.819±0.009**(1)
	Cvg	1.004±0.050(7)	0.982±0.023(3)	0.987±0.037(4)	**0.975±0.060**(1)	1.000±0.027(5)	1.000±0.019(5)	**0.975±0.054**(1)
	Ap	0.788±0.008(7)	0.793±0.008(3)	0.790±0.020(4)	0.794±0.010(2)	0.790±0.008(4)	0.790±0.010(4)	**0.795±0.007**(1)

Table 3
Results for learning with full labels on the small clusters (each containing fewer than 5% of the samples). Rkl and Cvg are the smaller the better, Auc and Ap are the larger the better. The italics indicates that GLOCAL is significantly better (paired t-tests at 95% significance level).

		GLObal	loCAL	GLOCAL			GLObal	loCAL	GLOCAL
Art	Rkl	*0.137±0.003*	*0.137±0.002*	**0.130±0.005**	Bus	Rkl	0.040±0.002	0.040±0.002	**0.040±0.003**
	Auc	*0.863±0.003*	*0.863±0.002*	**0.870±0.005**		Auc	0.958±0.003	0.958±0.003	**0.958±0.003**
	Cvg	*5.286±0.046*	*5.286±0.046*	**5.197±0.065**		Cvg	2.529±0.035	2.528±0.040	**2.528±0.040**
	Ap	*0.602±0.013*	*0.602±0.010*	**0.631±0.011**		Ap	*0.882±0.002*	*0.882±0.002*	**0.886±0.003**
Com	Rkl	*0.095±0.002*	*0.095±0.002*	**0.092±0.002**	Edu	Rkl	*0.101±0.002*	*0.101±0.002*	**0.097±0.002**
	Auc	*0.905±0.002*	*0.905±0.002*	**0.908±0.001**		Auc	*0.899±0.002*	*0.899±0.002*	**0.903±0.002**
	Cvg	*4.482±0.032*	*4.486±0.040*	**4.364±0.055**		Cvg	*4.803±0.033*	*4.805±0.036*	**4.672±0.051**
	Ap	0.677±0.003	0.676±0.003	**0.678±0.005**		Ap	*0.605±0.003*	*0.605±0.003*	**0.624±0.005**
Ent	Rkl	*0.091±0.002*	*0.091±0.002*	**0.086±0.003**	Hea	Rkl	0.054±0.002	0.054±0.003	**0.053±0.004**
	Auc	*0.909±0.002*	*0.909±0.002*	**0.914±0.002**		Auc	0.945±0.003	0.946±0.003	**0.947±0.003**
	Cvg	*2.817±0.027*	*2.797±0.035*	**2.709±0.059**		Cvg	3.508±0.036	3.506±0.049	**3.504±0.041**
	Ap	*0.748±0.003*	*0.749±0.004*	**0.759±0.006**		Ap	0.810±0.004	0.810±0.004	**0.812±0.006**
Rec	Rkl	*0.124±0.002*	*0.124±0.002*	**0.118±0.002**	Ref	Rkl	*0.060±0.002*	*0.061±0.003*	**0.054±0.004**
	Auc	0.871±0.003	0.870±0.003	**0.872±0.004**		Auc	*0.940±0.003*	*0.939±0.004*	**0.946±0.004**
	Cvg	3.704±0.033	3.700±0.037	**3.700±0.042**		Cvg	*2.552±0.043*	*2.559±0.057*	**2.325±0.060**
	Ap	0.670±0.004	0.670±0.004	**0.672±0.005**		Ap	*0.739±0.004*	*0.739±0.004*	**0.783±0.005**
Sci	Rkl	0.107±0.004	0.108±0.004	**0.107±0.004**	Soc	Rkl	*0.063±0.002*	*0.063±0.002*	**0.060±0.002**
	Auc	0.893±0.004	0.892±0.004	**0.893±0.005**		Auc	*0.930±0.002*	*0.930±0.002*	**0.934±0.002**
	Cvg	*5.937±0.041*	*5.941±0.049*	**5.845±0.054**		Cvg	3.558±0.033	3.559±0.038	**3.552±0.049**
	Ap	0.608±0.003	0.608±0.003	**0.610±0.003**		Ap	0.797±0.002	0.797±0.003	**0.798±0.003**
Soci	Rkl	*0.126±0.003*	*0.126±0.005*	**0.113±0.005**	Enr	Rkl	*0.117±0.002*	*0.119±0.003*	**0.105±0.005**
	Auc	*0.874±0.003*	*0.874±0.004*	**0.887±0.005**		Auc	*0.883±0.004*	*0.881±0.004*	**0.895±0.004**
	Cvg	*5.554±0.047*	*5.553±0.053*	**5.208±0.059**		Cvg	*19.440±0.833*	*19.372±0.915*	**17.511±1.231**
	Ap	*0.670±0.004*	*0.670±0.005*	**0.711±0.005**		Ap	*0.685±0.005*	*0.673±0.005*	**0.706±0.007**
Cor	Rkl	*0.163±0.002*	*0.163±0.002*	**0.160±0.002**	Ima	Rkl	*0.197±0.003*	*0.199±0.004*	**0.190±0.004**
	Auc	*0.837±0.002*	*0.837±0.002*	**0.840±0.002**		Auc	*0.803±0.003*	*0.801±0.003*	**0.810±0.003**
	Cvg	*130.84±1.01*	*131.13±1.21*	**128.40±1.30**		Cvg	*1.064±0.015*	*1.066±0.021*	**1.027±0.027**
	Ap	0.212±0.003	0.212±0.003	**0.214±0.005**		Ap	*0.764±0.003*	*0.763±0.004*	**0.771±0.005**

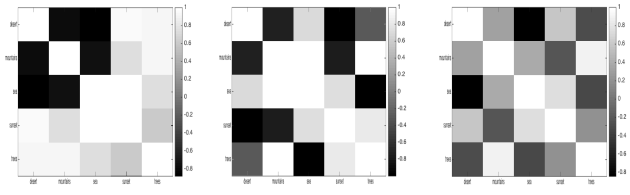
that this is because GLOCAL involves nonconvex optimization, and may get stuck in local minimum.

On the other hand, the other approaches under comparison only exploit some of the above aspects. BR does not consider label correlation, and its performance is almost always the worst or second-worst. For labels with few positive instances, it is hard for BR to obtain a good classifier without the help of label correlations (e.g., in the Coral5k dataset, there are over 100 labels each with fewer than 10 positive instances). RF-PCT, though it does not explicitly consider label correlations, achieves the best performance on Computers, Entertainment, Health, and Enron. This is mainly due to strong discriminating power of the random forest. When many labels are imbalanced, for each base classifier, labels with few positive instances may be dominated by those with many positive instances (e.g., the Coral5k dataset). As a result, though RF-PCT is an ensemble while GLOCAL is only one single classifier, RF-PCT still performs worse than GLOCAL overall. HOMER only uses the label correlations to build meta-labels via label clustering. At the leaf nodes, there may not be sufficient instances to build good models. Besides, obtaining good label clusters requires a large enough number of labels. On datasets such as Entertainment, there are only 21 labels and 6 of them have fewer than 10 positive instances. Hence, HOMER is outperformed by GLOCAL. LEML uses the low-rank structure, and does not explicitly exploit label correlations. MLLOC learns local label correlations only, while ML-LRC learns global label correlations only. However, both global and local label correlations can be potentially useful.

To show the example correlations learned by GLOCAL, we use two local groups extracted from the Image dataset.

(a) Group 1.

(b) Group 2.



(c) Local (group 1). (d) Local (group 2). (e) Global.

Figure 1. Example images from two local groups in the *Image* data set, and the corresponding 5×5 label correlation matrices. The labels are (top-to-down, left-to-right) "desert", "mountains", "sea", "sunset" and "trees".

Figure 1 shows that local label correlation does vary from group to group, and is different from global correlation. For group 1, "sunset" is highly correlated with "desert" and "sea" (Figure 1(c)). This can also be seen from the images in Figure 1(a). Moreover, "trees" sometimes co-occurs with "deserts" (first and last images in Figure 1(a)). However,

in group 2 (Figure 1(d)), "mountain" and "sea" often occur together and "trees" occurs less often with "desert" (Figure 1(b)). Figure 1(e) shows the learned global label correlation: "sea" and "sunset", "mountain" and "trees" are positively correlated, whereas "desert" and "sea", "desert" and "trees" are negatively correlated. All these correlations are consistent with intuition.

To further validate the effectiveness of global and local label correlations, we study two degenerate versions of GLOCAL: (i) GLObal, which uses only global label correlations; and (ii) loCAL, which uses only local label correlations. Note that the local groups obtained by clustering are not of equal sizes. For some datasets, the largest cluster contains more than 40% of instances, while some small ones contain fewer than 5% each. Global correlation is then dominated by the local correlation matrix of the largest cluster (Proposition 1), making the performance difference on the whole test set obscure. Hence, we focus on the performance of the small clusters. As can be seen from Table 3, using only global or local correlation may be good enough on some data sets (such as Health). On the other hand, considering both types of correlation as in GLOCAL achieves comparable or even better performance.

4.2 Learning with Missing Labels

In this section, experiments are conducted on datasets with missing labels, and evaluations are performed to validate the effectiveness on both the missing label recovery and the prediction tasks.

4.2.1 Setting

The datasets are the same as those in the full label experiments. To generate missing labels, we randomly sample $\rho\%$ of the elements in the label matrix as observed, and the rest as missing. When $\rho = 100$, it reduces to the full-label case. The performance measures are again the same as those in the full label experiments. Evaluation is performed both on the prediction of labels on the unseen testing instances, and also on the recovery of missing labels on the training set.

Among the baseline methods, only LEML and ML-LRC (together with the proposed GLOCAL) can directly handle missing labels. For the other baseline methods (BR, MLLOC, HOMER and RF-PCT), we have to first recover the full labels (here, we use the *matrix completion using side information* (MAXIDE) algorithm in [42]) before they can be used. The resultant combinations of MAXIDE+BR and MAXIDE+MLLOC are denoted MBR and MMLLOC, respectively. We do not compare with MAXIDE+HOMER because of the poor performance of HOMER in the full label case and its inability to directly handle missing labels. We also do not include MAXIDE+RF-PCT, because RF-PCT often has to ensemble many trees[9] and is very slow in both training and testing.

4.2.2 Results

Table 4 shows the results on recovering the training data's missing labels at different ratios of observed training labels. As discussed above, BR and MLLOC cannot directly handle missing labels, and so we compare with MAXIDE instead. Table 5 shows the label prediction results on the test data. To fit the tables onto one page, we do not report the standard deviation. Moreover, MBR performs the worst, and so is not shown in Table 5 because of the lack of space.

As can be seen from both tables, performance improves with more observed labels. This agrees with the intuition that as more elements in the label matrix are observed, more supervised information can be provided.

Overall, GLOCAL performs best at different ρ's on both missing-label recovery and test label prediction. The reasons for its superiority are similar to those discussed in Section 4.1.4, namely the joint learning of latent labels, instance-label mapping, and exploitation of both global and local label correlations. Moreover, as can be seen from the results on MMLLOC in Table 5, the two-stage approach of first recovering the missing training labels and then classification is not effective in predicting the test set labels. This is because MAXIDE will induce errors in the label recovery process (as can be seen from Table 4), which then propagate to the classification procedure.

4.3 Convergence

In this section, we empirically study the convergence of GLOCAL. Figure 2 shows the objective value w.r.t. the number of iterations for the full-label case on the Arts, Business, Enron and Image datasets. As can be seen, the objective converges quickly in a few iterations. A similar phenomenon can be observed on the other datasets.

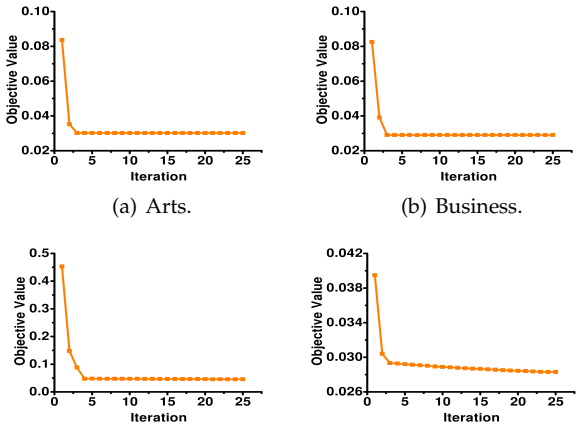


Figure 2. Convergence of GLOCAL on the Arts, Business, Enron and Image datasets.

Table 6 shows the timing results (in second) on learning with missing labels (with $\rho = 70$). GLOCAL and LEML train a classifier for all the labels jointly, and also can take advantage of the low-rank structure of either the model or label matrix during training. Thus, they are the fastest. However, GLOCAL has to be warm-started by Eqn. (2), and requires an additional clustering step to obtain local groups of the instances. Hence, it is slower than LEML. ML-LRC uses a low-rank label correlation matrix. However, it does not reduce the size of the label matrix or model involved in each iteration, and so is slower than GLOCAL. MBR and

9. On many datasets, the number of trees required is between 500 and 1000.

Table 4
Recovery results for missing label data on ranking loss(Rkl), average auc(Auc), coverage(Cvg) and average precision(Ap). Rkl and Cvg are the smaller the better, Auc and Ap are the larger the better. The italics indicates that GLOCAL is significantly better (paired t-tests at 95% significance level). The number in brackets shows ranking of the algorithm.

		ρ	MAXIDE	LEML	ML-LRC	GLOCAL			ρ	MAXIDE	LEML	ML-LRC	GLOCAL
Art	Rkl	30	*0.131*(2)	*0.133*(3)	*0.137*(4)	**0.103**(1)	Bus	Rkl	30	*0.044*(2)	*0.046*(3)	*0.046*(3)	**0.029**(1)
		70	*0.083*(2)	*0.090*(4)	*0.083*(2)	**0.074**(1)			70	*0.026*(3)	*0.027*(4)	*0.024*(2)	**0.021**(1)
	Auc	30	*0.871*(3)	*0.848*(4)	*0.879*(2)	**0.897**(1)		Auc	30	*0.956*(2)	*0.954*(3)	*0.954*(3)	**0.971**(1)
		70	*0.918*(2)	*0.912*(3)	*0.910*(4)	**0.928**(1)			70	*0.974*(2)	*0.973*(4)	*0.974*(2)	**0.979**(1)
	Cvg	30	*5.195*(3)	*5.231*(4)	*5.161*(2)	**4.189**(1)		Cvg	30	*2.550*(2)	*2.622*(3)	*2.622*(3)	**1.830**(1)
		70	*3.616*(2)	*3.733*(3)	*3.778*(4)	**3.234**(1)			70	*1.742*(2)	*1.783*(4)	*1.746*(3)	**1.477**(1)
	Ap	30	*0.645*(2)	*0.634*(4)	*0.640*(3)	**0.652**(1)		Ap	30	*0.876*(3)	*0.878*(2)	*0.876*(3)	**0.893**(1)
		70	**0.720**(1)	**0.720**(1)	*0.709*(4)	**0.720**(1)			70	*0.905*(2)	*0.901*(4)	*0.903*(3)	**0.908**(1)
Com	Rkl	30	*0.101*(4)	*0.098*(3)	*0.097*(2)	**0.073**(1)	Edu	Rkl	30	*0.097*(4)	*0.093*(3)	*0.089*(2)	**0.069**(1)
		70	*0.059*(2)	*0.063*(4)	*0.061*(3)	**0.052**(1)			70	*0.061*(2)	*0.061*(2)	*0.061*(2)	**0.058**(1)
	Auc	30	*0.905*(4)	*0.908*(3)	*0.909*(2)	**0.933**(1)		Auc	30	*0.902*(4)	*0.907*(3)	*0.911*(2)	**0.932**(1)
		70	*0.947*(2)	*0.943*(4)	*0.945*(3)	**0.955**(1)			70	*0.938*(3)	*0.938*(3)	0.940(2)	**0.942**(1)
	Cvg	30	*4.627*(4)	*4.586*(3)	*4.565*(2)	**3.511**(1)		Cvg	30	*4.672*(4)	*4.372*(3)	*3.914*(2)	**3.171**(1)
		70	*2.912*(2)	*3.100*(4)	*3.095*(3)	**2.586**(1)			70	*3.113*(4)	*3.106*(3)	3.000(2)	**2.815**(1)
	Ap	30	*0.709*(2)	*0.700*(4)	*0.705*(3)	**0.726**(1)		Ap	30	0.653(2)	*0.648*(4)	0.653(2)	**0.655**(1)
		70	**0.787**(1)	**0.787**(1)	**0.787**(1)	**0.787**(1)			70	**0.711**(1)	*0.702*(4)	0.710(3)	**0.711**(1)
Ent	Rkl	30	*0.104*(3)	*0.103*(2)	*0.106*(4)	**0.085**(1)	Hea	Rkl	30	*0.060*(4)	*0.057*(3)	*0.054*(2)	**0.041**(1)
		70	*0.063*(2)	*0.063*(2)	*0.063*(2)	**0.062**(1)			70	*0.037*(4)	*0.036*(3)	0.032(2)	**0.030**(1)
	Auc	30	*0.898*(3)	*0.899*(2)	*0.899*(2)	**0.916**(1)		Auc	30	*0.941*(4)	*0.943*(3)	*0.947*(2)	**0.960**(1)
		70	**0.940**(1)	0.938(4)	**0.940**(1)	**0.940**(1)			70	*0.964*(3)	*0.964*(3)	0.968(2)	**0.971**(1)
	Cvg	30	*3.058*(4)	*2.994*(2)	*3.022*(3)	**2.512**(1)		Cvg	30	*3.577*(4)	*3.462*(2)	*3.465*(3)	**2.567**(1)
		70	1.987(2)	2.051(3)	2.080(4)	**1.957**(1)			70	*2.524*(4)	*2.465*(3)	*2.450*(2)	**2.152**(1)
	Ap	30	0.704(1)	*0.698*(3)	*0.698*(3)	**0.704**(1)		Ap	30	*0.796*(3)	*0.794*(4)	0.798(2)	**0.801**(1)
		70	*0.763*(4)	0.765(2)	0.765(2)	**0.768**(1)			70	**0.848**(1)	*0.842*(4)	**0.848**(1)	**0.848**(1)
Rec	Rkl	30	*0.130*(2)	*0.133*(3)	*0.135*(4)	**0.110**(1)	Ref	Rkl	30	*0.083*(2)	*0.083*(2)	*0.083*(2)	**0.063**(1)
		70	*0.078*(2)	*0.080*(3)	*0.080*(3)	**0.068**(1)			70	**0.048**(1)	0.049(3)	0.049(3)	**0.048**(1)
	Auc	30	*0.873*(2)	*0.870*(3)	*0.869*(4)	**0.895**(1)		Auc	30	*0.919*(2)	*0.919*(2)	*0.918*(4)	**0.939**(1)
		70	*0.925*(2)	*0.923*(3)	*0.920*(4)	**0.934**(1)			70	**0.955**(1)	0.953(3)	0.953(3)	**0.955**(1)
	Cvg	30	*3.899*(2)	*3.919*(3)	*4.048*(4)	**3.291**(1)		Cvg	30	*3.436*(4)	*3.392*(3)	*3.372*(2)	**2.520**(1)
		70	*2.560*(2)	*2.607*(3)	*2.620*(4)	**2.262**(1)			70	*2.039*(2)	*2.103*(3)	*2.195*(4)	**1.972**(1)
	Ap	30	*0.680*(2)	*0.663*(3)	*0.660*(4)	**0.681**(1)		Ap	30	**0.681**(1)	*0.664*(4)	0.674(3)	0.679(2)
		70	*0.767*(2)	*0.763*(3)	*0.760*(4)	**0.770**(1)			70	0.745(4)	**0.746**(1)	**0.746**(1)	**0.746**(1)
Sci	Rkl	30	*0.110*(2)	*0.111*(4)	*0.110*(2)	**0.086**(1)	Soc	Rkl	30	*0.069*(3)	*0.069*(3)	*0.063*(2)	**0.042**(1)
		70	0.063(1)	*0.071*(4)	*0.070*(3)	**0.063**(1)			70	*0.041*(4)	*0.040*(2)	*0.040*(2)	**0.026**(1)
	Auc	30	*0.889*(2)	*0.889*(2)	*0.889*(2)	**0.913**(1)		Auc	30	*0.930*(3)	*0.930*(3)	*0.936*(2)	**0.957**(1)
		70	**0.935**(1)	*0.928*(3)	*0.923*(4)	**0.935**(1)			70	*0.964*(3)	*0.959*(4)	*0.966*(2)	**0.973**(1)
	Cvg	30	*6.193*(3)	*6.141*(2)	*6.271*(4)	**4.845**(1)		Cvg	30	*3.865*(3)	*3.920*(4)	*3.304*(2)	**2.443**(1)
		70	3.771(2)	*3.914*(4)	*3.878*(3)	**3.751**(1)			70	*2.103*(2)	*2.386*(4)	*2.373*(3)	**1.663**(1)
	Ap	30	0.615(1)	0.613(4)	0.614(3)	**0.615**(1)		Ap	30	*0.780*(3)	*0.780*(3)	*0.784*(2)	**0.802**(1)
		70	*0.689*(2)	*0.647*(4)	*0.650*(3)	**0.691**(1)			70	*0.854*(4)	**0.865**(1)	**0.865**(1)	**0.865**(1)
Soci	Rkl	30	*0.129*(4)	*0.128*(3)	*0.123*(2)	**0.102**(1)	Enr	Rkl	30	*0.091*(3)	*0.115*(4)	*0.085*(2)	**0.075**(1)
		70	0.074(3)	*0.081*(4)	**0.073**(1)	**0.073**(1)			70	0.042(3)	*0.060*(4)	**0.040**(1)	**0.040**(1)
	Auc	30	*0.871*(4)	*0.872*(3)	*0.877*(2)	**0.898**(1)		Auc	30	*0.910*(3)	*0.887*(4)	*0.918*(2)	**0.926**(1)
		70	0.926(3)	*0.919*(4)	0.928(2)	**0.929**(1)			70	0.960(3	*0.942*(4)	**0.962**(1)	**0.962**(1)
	Cvg	30	*5.557*(4)	*5.459*(3)	*5.167*(2)	**4.496**(1)		Cvg	30	*14.24*(3)	*16.65*(4)	*13.45*(2)	**12.05**(1)
		70	*3.641*(3)	*3.824*(4)	*3.608*(2)	**3.442**(1)			70	*7.961*(3)	*10.33*(4)	**7.480**(1)	7.510(2)
	Ap	30	0.646(3)	*0.629*(4)	0.650(2)	**0.652**(1)		Ap	30	**0.739**(1)	*0.711*(4)	**0.739**(1)	**0.739**(1)
		70	**0.719**(1)	0.717(4)	**0.719**(1)	**0.719**(1)			70	0.854(3)	*0.842*(4)	**0.855**(1)	**0.855**(1)
Cor	Rkl	30	*0.226*(4)	*0.214*(3)	*0.206*(2)	**0.185**(1)	Ima	Rkl	30	*0.302*(4)	*0.184*(3)	0.175(2)	**0.173**(1)
		70	*0.138*(4)	*0.131*(3)	**0.123**(1)	0.125(2)			70	*0.251*(4)	**0.148**(1)	**0.148**(1)	**0.148**(1)
	Auc	30	*0.773*(4)	*0.786*(3)	*0.794*(2)	**0.814**(1)		Auc	30	*0.820*(4)	**0.828**(1)	0.826(3)	**0.828**(1)
		70	0.874(1)	0.874(1)	0.874(1)	0.874(1)			70	*0.834*(4)	**0.857**(1)	0.855(2)	0.855(2)
	Cvg	30	*204.90*(4)	*182.76*(3)	*178.60*(2)	**153.82**(1)		Cvg	30	*1.493*(4)	*1.104*(3)	0.967(2)	**0.950**(1)
		70	103.63(4)	102.42(3)	**102.30**(1)	**102.30**(1)			70	*0.790*(4)	**0.760**(1)	0.770(3)	**0.760**(1)
	Ap	30	**0.275**(1)	*0.259*(4)	**0.275**(1)	**0.275**(1)		Ap	30	*0.739*(4)	*0.776*(2)	*0.775*(3)	**0.785**(1)
		70	0.279(1)	0.279(1)	0.279(1)	0.279(1)			70	*0.768*(4)	**0.841**(1)	0.834(3)	**0.841**(1)

Table 5
Prediction results for missing label data on ranking loss(Rkl), average auc(Auc), coverage(Cvg) and average precision(Ap). Rkl and Cvg are the smaller the better, Auc and Ap are the larger the better. The italics indicates that GLOCAL is significantly better (paired t-tests at 95% significance level). The number in brackets shows ranking of the algorithm.

		ρ	MMLLOC	LEML	ML-LRC	GLOCAL			ρ	MMLLOC	LEML	ML-LRC	GLOCAL
Art	Rkl	30	*0.225*(4)	*0.204*(3)	*0.184*(2)	**0.144**(1)	Bus	Rkl	30	*0.083*(4)	*0.063*(3)	*0.061*(2)	**0.054**(1)
		70	*0.193*(4)	*0.181*(3)	*0.159*(2)	**0.139**(1)			70	*0.064*(4)	*0.058*(3)	**0.046**(1)	**0.046**(1)
	Auc	30	*0.781*(4)	*0.801*(3)	0.828(2)	**0.831**(1)		Auc	30	*0.917*(4)	*0.928*(3)	**0.937**(1)	**0.937**(1)
		70	*0.819*(4)	*0.825*(3)	0.838(2)	**0.840**(1)			70	*0.935*(4)	*0.942*(3)	0.950(2)	**0.952**(1)
	Cvg	30	*9.033*(4)	*7.369*(3)	*6.281*(2)	**5.867**(1)		Cvg	30	*4.643*(4)	*3.954*(3)	*3.279*(2)	**2.863**(1)
		70	*7.262*(4)	*6.431*(3)	5.432(2)	**5.352**(1)			70	*3.670*(4)	*3.303*(3)	2.580(2)	**2.579**(1)
	Ap	30	*0.529*(2)	*0.503*(4)	*0.517*(3)	**0.572**(1)		Ap	30	*0.843*(4)	*0.866*(2)	*0.858*(3)	**0.879**(1)
		70	*0.583*(4)	*0.589*(2)	*0.588*(3)	**0.607**(1)			70	*0.861*(4)	*0.870*(2)	*0.870*(2)	**0.881**(1)
Com	Rkl	30	*0.201*(4)	*0.179*(3)	**0.152**(1)	0.154(2)	Edu	Rkl	30	*0.187*(4)	*0.176*(3)	*0.144*(2)	**0.137**(1)
		70	*0.150*(4)	*0.141*(3)	0.115(2)	**0.113**(1)			70	*0.165*(4)	*0.151*(3)	0.113(2)	**0.111**(1)
	Auc	30	*0.849*(4)	0.880(2)	*0.873*(3)	**0.883**(1)		Auc	30	*0.815*(4)	*0.817*(3)	0.845(2)	**0.846**(1)
		70	*0.868*(4)	0.894(3)	0.895(2)	**0.896**(1)			70	*0.844*(3)	*0.842*(4)	**0.860**(1)	**0.860**(1)
	Cvg	30	*8.808*(4)	*7.392*(3)	*6.052*(2)	**5.798**(1)		Cvg	30	*11.089*(4)	*9.672*(3)	6.350(2)	**6.338**(1)
		70	*6.871*(4)	*6.306*(3)	5.000(2)	**4.976**(1)			70	*8.096*(4)	*7.595*(3)	5.075(2)	**5.070**(1)
	Ap	30	*0.631*(4)	*0.646*(2)	*0.636*(3)	**0.669**(1)		Ap	30	*0.538*(3)	*0.537*(4)	*0.543*(2)	**0.592**(1)
		70	*0.674*(2)	*0.665*(4)	*0.667*(3)	**0.691**(1)			70	*0.586*(4)	*0.591*(3)	*0.600*(2)	**0.622**(1)
Ent	Rkl	30	*0.229*(4)	*0.175*(3)	*0.152*(2)	**0.122**(1)	Hea	Rkl	30	*0.137*(4)	*0.095*(3)	**0.085**(1)	**0.085**(1)
		70	*0.164*(4)	*0.159*(3)	*0.129*(2)	**0.109**(1)			70	*0.109*(4)	*0.074*(3)	*0.071*(2)	**0.065**(1)
	Auc	30	*0.832*(3)	*0.826*(4)	*0.849*(2)	**0.859**(1)		Auc	30	*0.894*(4)	*0.896*(3)	**0.907**(1)	0.906(2)
		70	*0.842*(4)	*0.850*(3)	0.870(2)	**0.871**(1)			70	*0.901*(4)	**0.920**(1)	**0.920**(1)	**0.920**(1)
	Cvg	30	*6.029*(4)	*5.755*(3)	4.170(2)	**4.153**(1)		Cvg	30	*7.104*(4)	*6.248*(3)	4.924(2)	**4.814**(1)
		70	*4.857*(4)	*4.643*(3)	*3.483*(2)	**3.117**(1)			70	*5.866*(4)	*5.167*(3)	**3.960**(1)	3.963(2)
	Ap	30	*0.601*(2)	*0.601*(2)	*0.601*(2)	**0.645**(1)		Ap	30	*0.727*(2)	*0.715*(4)	*0.720*(3)	**0.752**(1)
		70	*0.635*(4)	*0.645*(2)	*0.643*(3)	**0.670**(1)			70	*0.762*(4)	*0.770*(2)	*0.766*(3)	**0.775**(1)
Rec	Rkl	30	*0.266*(4)	*0.245*(3)	*0.202*(2)	**0.165**(1)	Ref	Rkl	30	*0.199*(4)	*0.187*(3)	*0.137*(2)	**0.098**(1)
		70	*0.204*(4)	*0.196*(3)	*0.167*(2)	**0.156**(1)			70	*0.155*(4)	*0.145*(3)	*0.098*(2)	**0.086**(1)
	Auc	30	*0.785*(4)	*0.828*(2)	*0.802*(3)	**0.839**(1)		Auc	30	*0.851*(3)	*0.847*(4)	*0.868*(2)	**0.886**(1)
		70	*0.800*(4)	*0.837*(2)	*0.836*(3)	**0.845**(1)			70	*0.861*(4)	*0.869*(3)	0.895(2)	**0.898**(1)
	Cvg	30	*7.084*(4)	*6.842*(3)	*5.397*(2)	**4.545**(1)		Cvg	30	*7.549*(4)	*6.463*(3)	*5.052*(2)	**3.367**(1)
		70	*5.952*(4)	*5.685*(3)	4.490(2)	**4.430**(1)			70	*6.419*(4)	*6.130*(3)	*3.694*(2)	**3.348**(1)
	Ap	30	*0.547*(2)	*0.540*(3)	*0.540*(3)	**0.573**(1)		Ap	30	0.631(2)	*0.609*(4)	*0.611*(3)	**0.638**(1)
		70	*0.597*(3)	*0.567*(4)	*0.600*(2)	**0.614**(1)			70	**0.675**(1)	*0.653*(3)	*0.653*(3)	0.672(2)
Sci	Rkl	30	*0.257*(4)	*0.203*(3)	*0.169*(2)	**0.144**(1)	Soc	Rkl	30	*0.149*(4)	*0.089*(2)	*0.095*(3)	**0.075**(1)
		70	*0.189*(4)	*0.174*(3)	0.134(2)	**0.129**(1)			70	*0.108*(4)	*0.079*(3)	*0.076*(2)	**0.073**(1)
	Auc	30	*0.827*(3)	*0.827*(3)	*0.830*(2)	**0.837**(1)		Auc	30	*0.906*(2)	*0.906*(3)	*0.905*(3)	**0.913**(1)
		70	*0.840*(4)	0.849(3)	**0.850**(1)	**0.850**(1)			70	*0.910*(3)	*0.900*(4)	**0.914**(1)	**0.914**(1)
	Cvg	30	*12.805*(4)	*10.587*(3)	*8.794*(2)	**6.809**(1)		Cvg	30	*7.652*(4)	*7.567*(3)	6.308(2)	**6.088**(1)
		70	*9.960*(4)	*9.501*(3)	*6.900*(2)	**6.416**(1)			70	*5.886*(4)	*5.386*(3)	5.103(2)	**4.929**(1)
	Ap	30	*0.503*(2)	*0.479*(4)	*0.485*(3)	**0.531**(1)		Ap	30	*0.712*(2)	*0.682*(4)	*0.700*(3)	**0.738**(1)
		70	0.569(3)	*0.551*(4)	0.570(2)	**0.574**(1)			70	*0.748*(2)	*0.719*(4)	*0.728*(3)	**0.761**(1)
Soci	Rkl	30	*0.252*(4)	*0.202*(3)	*0.175*(2)	**0.139**(1)	Enr	Rkl	30	*0.179*(4)	*0.172*(2)	*0.173*(3)	**0.149**(1)
		70	*0.208*(4)	*0.194*(3)	*0.141*(2)	**0.136**(1)			70	*0.170*(4)	*0.162*(3)	*0.152*(2)	**0.129**(1)
	Auc	30	*0.804*(4)	*0.808*(3)	**0.826**(1)	**0.826**(1)		Auc	30	*0.820*(4)	*0.830*(3)	*0.843*(2)	**0.853**(1)
		70	*0.816*(3)	*0.816*(3)	**0.840**(1)	**0.840**(1)			70	*0.829*(4)	*0.839*(3)	*0.849*(2)	**0.872**(1)
	Cvg	30	*9.550*(4)	*8.637*(3)	*6.944*(2)	**5.816**(1)		Cvg	30	*22.72*(4)	*21.41*(3)	*20.42*(2)	**19.01**(1)
		70	*8.227*(4)	*7.638*(3)	**5.750**(1)	**5.750**(1)			70	*21.90*(4)	*19.53*(3)	*18.17*(2)	**17.16**(1)
	Ap	30	*0.569*(2)	*0.563*(4)	*0.565*(3)	**0.601**(1)		Ap	30	*0.580*(3)	*0.582*(2)	*0.580*(3)	**0.589**(1)
		70	*0.606*(2)	*0.589*(4)	*0.590*(3)	**0.625**(1)			70	*0.585*(4)	*0.601*(3)	*0.607*(2)	**0.635**(1)
Cor	Rkl	30	*0.332*(4)	*0.308*(2)	*0.331*(3)	**0.285**(1)	Ima	Rkl	30	*0.224*(4)	*0.204*(2)	*0.220*(3)	**0.200**(1)
		70	*0.248*(3)	*0.250*(4)	0.199(2)	**0.194**(1)			70	*0.195*(3)	0.188(2)	*0.197*(4)	**0.187**(1)
	Auc	30	*0.673*(3)	*0.693*(2)	*0.670*(4)	**0.714**(1)		Auc	30	*0.796*(3)	*0.795*(4)	0.800(2)	**0.801**(1)
		70	*0.747*(4)	*0.749*(3)	0.801(2)	**0.805**(1)			70	0.812(2)	0.811(3)	0.810(4)	**0.813**(1)
	Cvg	30	*275.41*(4)	*233.83*(2)	*240.17*(3)	**211.84**(1)		Cvg	30	*1.160*(4)	*1.103*(2)	*1.131*(3)	**1.070**(1)
		70	*212.84*(4)	*190.83*(3)	*160.59*(2)	**151.23**(1)			70	*1.066*(4)	1.030(2)	*1.040*(3)	**1.025**(1)
	Ap	30	*0.158*(4)	*0.166*(2)	*0.165*(3)	**0.174**(1)		Ap	30	*0.745*(3)	*0.752*(2)	*0.744*(4)	**0.760**(1)
		70	*0.176*(4)	*0.185*(3)	0.188(2)	**0.192**(1)			70	*0.768*(4)	0.772(2)	*0.770*(3)	**0.777**(1)

Table 6
Timing results (in seconds) for learning with missing labels ($\rho = 70$). F is the time for missing label recovery. C is the time for clustering, I is the time for initialization, and R is the time of the main learning procedure. A is the total time. Note that some algorithms may not need F, C or I.

	MBR			MMLLOC					LEML			ML-LRC			GLOCAL			
	A	F	R	A	F	C	I	R	A	I	R	A	I	R	A	C	I	R
Arts	109	8	101	107	8	1	0	98	34	0	34	87	0	87	47	1	20	26
Business	38	6	32	104	6	1	0	97	35	0	35	82	0	82	49	1	24	24
Computers	78	11	67	121	11	1	0	109	46	0	46	94	0	94	53	1	31	21
Education	60	8	52	115	8	1	0	106	45	0	45	64	0	64	45	1	29	15
Entertainment	66	6	60	91	6	1	0	84	42	0	42	73	0	73	53	2	22	29
Health	64	11	53	116	11	1	0	104	41	0	41	75	0	75	67	1	32	34
Recreation	63	4	59	97	5	1	0	91	46	0	46	55	0	55	51	2	22	27
Reference	75	14	61	131	15	9	0	107	38	0	38	91	0	91	78	8	32	38
Science	101	15	86	133	15	1	0	117	53	0	53	103	0	103	77	2	32	43
Social	163	36	127	149	33	8	0	108	37	0	37	147	0	147	90	7	35	48
Society	83	8	75	106	8	1	0	97	32	0	32	117	0	117	44	2	18	24
Enron	47	10	37	59	10	1	0	48	38	0	38	78	0	78	69	1	25	43
Corel5k	458	272	186	1529	268	1	0	1260	307	0	307	709	0	709	413	1	78	344
Image	5	1	4	25	2	1	0	22	28	0	28	14	0	14	15	1	5	9

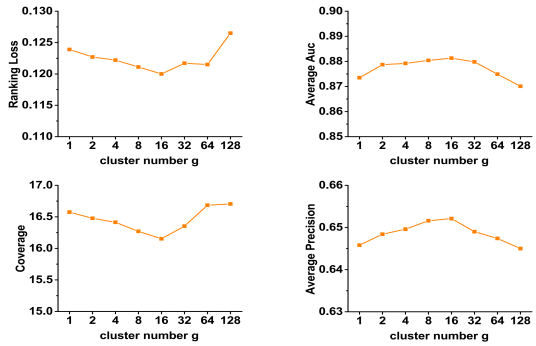


Figure 3. Varying the number of clusters g on the Enron dataset.

MMLLOC require training a classifier for each label, and also an additional step to recover the missing labels. Thus, they are often the slowest, especially when the number of class labels is large. Similar results can be observed with $\rho = 30$, which are not reported here.

4.4 Sensitivity to Parameters

In this experiment, we study the influence of parameters, including the number of clusters g, regularization parameters λ_3 and λ_4 (corresponding to the global and the local manifold regularizers, respectively), regularization parameter λ_2 for the Frobenius norm regularizer, and dimensionality k of the latent representation. We vary one parameter, while keeping the others fixed at their best setting.

4.4.1 *Varying the Number of Clusters g*

Figure 3 shows the influence on the Enron dataset. When there is only one cluster, no local label correlation is considered. With more clusters, performance improves as more local label correlations are taken into account. When too many clusters are used, very few instances are placed in each cluster, and the local label correlations cannot be reliably estimated. Thus, the performance starts to deteriorate.

4.4.2 *Influence of Label Manifold Regularizers (λ_3 and λ_4)*

A larger λ_3 means higher importance of global label correlation, whereas a larger λ_4 means higher importance of local label correlation. Figures 4 and 5 show their effects on the Enron dataset. When $\lambda_3 = 0$, only local label correlations are considered, and the performance is poor. With increasing λ_3, performance improves. However, when λ_3 is very large, performance deteriorates as the global label correlations dominate. A similar phenomenon can be observed for λ_4.

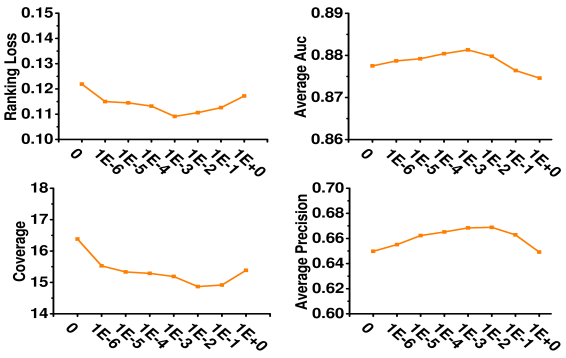


Figure 4. Varying the global label manifold regularization parameter λ_3 on the Enron dataset.

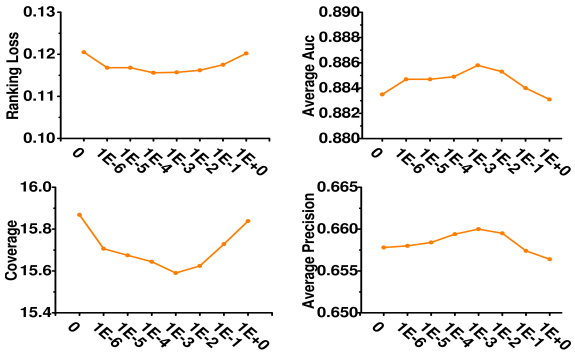


Figure 5. Varying the local label manifold regularization parameter λ_4 on the Enron dataset.

4.4.3 *Varying the Latent Representation Dimensionality k*

Figure 6 shows the effect of varying k on the Enron dataset. As can be seen, when k is too small, the latent labels cannot capture enough information. With increasing k, performance improves. When k is too large, the low-rank structure is not fully utilized, and performance starts to get worse.

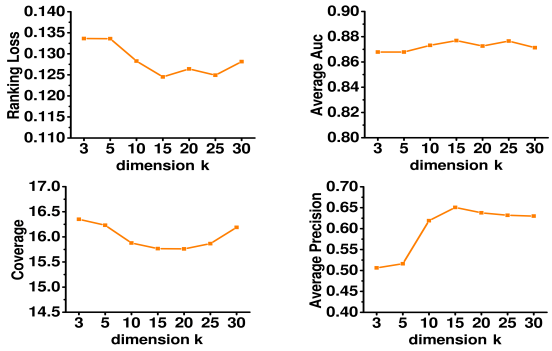


Figure 6. Varying the latent representation dimensionality on the Enron dataset.

4.4.4 Influence of λ_2

Figure 7 shows the effect of varying λ_2 on the Enron dataset. As can be seen, GLOCAL is not sensitive to this parameter.

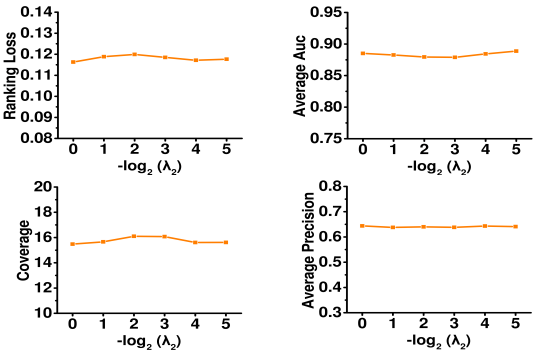


Figure 7. Varying λ_2 on the Enron dataset.

5 Conclusion

In this paper, we proposed a new multi-label correlation learning approach GLOCAL, which simultaneously recovers the missing labels, trains the classifier and exploits both global and local label correlations, through learning a latent label representation and optimizing the label manifolds. Compared with the previous work, it is the first to exploit both global and local label correlations, which directly learns the Laplacian matrix without requiring any other prior knowledge on label correlations. Moreover, GLOCAL provides a unified solution for both full-label and missing-label multi-label learning. Experimental results show that our approach outperforms the state-of-the-art multi-label learning approaches on learning with both full labels and missing labels. In our work, we handle the case that label correlations are symmetric. It has been disclosed in [18] that, label correlations can be asymmetric in many situations. So it is desirable to study the asymmetric label correlations in our future work. Besides, it is also interesting to extend our work to a multi-instance multi-label setting [53], where each object is represented by several instances, and structures among them can be taken into account.

Acknowledgments

This research was supported by NSFC (61333014), the Collaborative Innovation Center of Novel Software Technology and Industrialization, and the program B for Outstanding PhD candidate of Nanjing University .

References

[1] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

[2] W. Bi and J. T. Kwok. Efficient multi-label classification with many labels. In *Proceedings of the 30th International Conference on Machine Learning*, pages 405–413, 2013.

[3] W. Bi and J. T. Kwok. Multilabel classification with label correlations and missing labels. In *Proceedings of 28th AAAI Conference on Artificial Intelligence*, pages 1680–1686, 2014.

[4] N. Boumal., B. Mishra, P.-A. Absil., and R. Sepulchre. Manopt, a Matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 15:1455–1459, 2014.

[5] M. Boutell, J. Luo, X. Shen, and C. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004.

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

[7] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[8] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker. Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, 3(1):140–149, 2007.

[9] F. Chung. *Spectral Graph Theory*, volume 92. American Mathematical Society, 1997.

[10] Z. Dong, W. Liang, Y. Wu, M. Pei, and Y. Jia. Nonnegative correlation coding for image classification. *Science China Information Sciences*, 59(1):1–14, 2016.

[11] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[12] J. Friedman, T. Hastie, and R. Tibshirani. *The Elements of Statistical Learning*, volume 1. Springer, New York, 2001.

[13] J. Fürnkranz, E. Hüllermeier, E. Mencía, and K. Brinker. Multilabel classification via calibrated label ranking. *Machine Learning*, 73(2):133–153, 2008.

[14] N. Gao, S.-J. Huang, and S. Chen. Multi-label active learning by model guided distribution matching. *Frontiers of Computer Science*, 10(5):845–855, 2016.

[15] A. Goldberg, B. Recht, J. Xu, R. Nowak, and X. Zhu. Transduction with matrix completion: Three birds with one stone. In *Advances in Neural Information Processing Systems 23*, pages 757–765. 2010.

[16] K. H. Huang and H. T. Lin. Cost-sensitive label embedding for multi-label classification. *Machine Learning*, 106:1725–1746, 2017.

[17] S.-J. Huang, S. Chen, and Z.-H. Zhou. Multi-label active learning: Query type matters. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, pages 946–952, 2015.

[18] S.-J. Huang, Y. Yu, and Z.-H. Zhou. Multi-label hypothesis reuse. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 525–533, 2012.

[19] S.-J. Huang and Z.-H. Zhou. Multi-label learning by exploiting label correlations locally. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, pages 949–955, 2012.

[20] S. Ji, L. Tang, S. Yu, and J. Ye. Extracting shared subspace for multi-label classification. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 381–389, 2008.

[21] S. Ji and J. Ye. An accelerated gradient method for trace norm minimization. In *Proceedings of the 26th International Conference on Machine Learning*, pages 457–464, 2009.

[22] X.-Y. Jing, F. Wu, Z. Li, R. Hu, and D. Zhang. Multi-label dictionary learning for image annotation. *IEEE Transactions on Image Processing*, 25(6):2712–2725, 2016.

[23] D. Kocev, C. Vens, J. Struyf, and S. Dzeroski. Ensembles of multi-objective decision trees. In *Proceedings of th 18th European Conference on Machine Learning*, pages 624–631, 2007.

[24] D. Kocev, C. Vens, J. Struyf, and S. Dzeroski. Tree ensembles for predicting structured outputs. *Pattern Recognition*, 46(3):817–833, 2013.

[25] X. Kong, M. K. Ng, and Z.-H. Zhou. Transductive multilabel learning via label set propagation. *IEEE Transactions on Knowledge and Data Engineering*, 25(3):704–719, 2013.

[26] J. Lee, S. Kim, G. Lebanon, Y. Singer, and S. Bengio. Llorma: Local low-rank matrix approximation. *The Journal of Machine Learning Research*, 17(1):442–465, 2016.

[27] D. Luo, C. Ding, H. Huang, and T. Li. Non-negative laplacian embedding. In *Proceedings of the 9th IEEE International Conference on Data Mining*, pages 337–346, 2009.
[28] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Dzeroski. An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*, 45(9):3084–3104, 2012.
[29] S. Melacci and M. Belkin. Laplacian Support Vector Machines Trained in the Primal. *Journal of Machine Learning Research*, 12:1149–1184, 2011.
[30] J. Petterson and T. Caetano. Submodular multi-label learning. In *Advances in Neural Information Processing Systems 24*, pages 1512–1520. 2011.
[31] K. Punera, S. Rajan, and J. Ghosh. Automatically learning document taxonomies for hierarchical classification. *Proceedings of the 14th International Conference on WWW*, pages 1010–1011, 2005.
[32] J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. *Machine Learning*, 85(3):333–359, 2011.
[33] A. Subramanian, P. Tamayo, V. Mootha, S. M. B. Ebert, M. Gillette, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005.
[34] L. Sun, S. Ji, and J. Ye. Hypergraph spectral learning for multi-label classification. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 668–676, 2008.
[35] G. Tsoumakas, I. Katakis, and I. Vlahavas. Effective and efficient multi-label classification in domains with large number of labels. In *Proceedings of ECML/PKDD 2008 Workshop on Mining Multidimensional Data*, pages 30–44, 2008.
[36] G. Tsoumakas, I. Katakis, and I. Vlahavas. Mining multi-label data. In *Data Mining and Knowledge Discovery Handbook*, pages 667–685. Springer, 2009.
[37] D. Turnbull, L. Barrington, D. Torres, and C. Lanckriet. Semantic annotation and retrieval of music and sound effects. *IEEE Transactions on Audio, Speech and Language Processing*, 16(2):467–476, 2008.
[38] N. Ueda and K. Saito. Parametric mixture models for multi-labeled text. In *Advances in Neural Information Processing Systems 15*, pages 721–728. 2002.
[39] H. Wang, H. Huang, and C. Ding. Image annotation using multi-label correlated green's function. In *Proceedings of the 12th International Conference on Computer Vision*, pages 2029–2034, 2009.
[40] W. Weng, Y. Lin, S. Wu, Y. Li, and Y. Kang. Multi-label learning based on label-specific features and local pairwise label correlation. *Neurocomputing*, 2017.
[41] L. Xu, Z. Wang, Z. Shen, Y. Wang, and E. Chen. Learning low-rank label correlations for multi-label classification with missing labels. In *Proceedings of the 14th IEEE International Conference on Data Mining*, pages 1067–1072, 2014.
[42] M. Xu, R. Jin, and Z.-H. Zhou. Speedup matrix completion with side information: Application to multi-label learning. In *Advances in Neural Information Processing Systems 26*, pages 2301–2309. 2013.
[43] B. Yang, J.-T. Sun, T. Wang, and Z. Chen. Effective multi-label active learning for text classification. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 917–926, 2009.
[44] C.-K. Yeh, W.-C. Wu, W.-J. Ko, and Y.-C. F. Wang. Learning deep latent space for multi-label classification. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 2838–2844, 2017.
[45] H.-F. Yu, P. Jain, P. Kar, and I. Dhillon. Large-scale multi-label learning with missing labels. In *Proceedings of the 31th International Conference on Machine Learning*, pages 593–601, 2014.
[46] M.-L. Zhang, Y.-K. Li, and X.-Y. Liu. Towards class-imbalance aware multi-label learning. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, pages 4041–4047, 2015.
[47] M.-L. Zhang and K. Zhang. Multi-label learning by exploiting label dependency. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 999–1008, 2010.
[48] M.-L. Zhang and Z.-H. Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048, 2007.
[49] M.-L. Zhang and Z.-H. Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2014.
[50] Y. Zhang and Z.-H. Zhou. Multilabel dimensionality reduction via dependence maximization. *ACM Transactions on Knowledge Discovery from Data*, 4(3):14:1–14:21, 2010.
[51] Z.-H. Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 2018.
[52] Z.-H. Zhou and M.-L. Zhang. Multi-label learning. In *C. Sammut, G. I. Webb, eds. Encyclopedia of Machine Learning and Data Mining, Berlin: Springer*, pages 875–881, 2017.
[53] Z. H. Zhou, M. L. Zhang, S. J. Huang, and Y. F. Li. Multi-instance multi-label learning. *Artificial Intelligence*, 176(1):2291–2320, 2012.

Yue Zhu received the BSc degree in Computer Science and Technology from Nanjing Normal University, Nanjing, China, in 2011. In the same year, he was admitted to further study in Nanjing University, Nanjing, China. Now, he is currently a PhD candidate and a member of LAMDA Group. His research interests mainly include multi-instance learning, multi-label learning, multi-view learning and incremental learning.

James T. Kwok James Kwok is a Professor in the Department of Computer Science and Engineering, Hong Kong University of Science and Technology. He is an IEEE Fellow. Prof Kwok received his B.Sc. degree in Electrical and Electronic Engineering from the University of Hong Kong and his Ph.D. degree in computer science from the Hong Kong University of Science and Technology. He then joined the Department of Computer Science, Hong Kong Baptist University as an Assistant Professor. He returned to the Hong Kong University of Science and Technology in 2000 and is now a Professor in the Department of Computer Science and Engineering. He served as an Associate Editor for the IEEE Transactions on Neural Networks and Learning Systems from 2006-2012, and is currently serving as Associate Editor for the Neurocomputing journal. He is a Governing Board Member of the Asia Pacific Neural Network Society (APNNS).

Zhi-Hua Zhou (S'00-M'01-SM'06-F'13) received the BSc, MSc and PhD degrees in computer science from Nanjing University, China, in 1996, 1998 and 2000, respectively, all with the highest honors. He joined the Department of Computer Science & Technology at Nanjing University as an Assistant Professor in 2001, and is currently Professor and Standing Deputy Director of the National Key Laboratory for Novel Software Technology; he is also the Founding Director of the LAMDA group. His research interests are mainly in artificial intelligence, machine learning and data mining. He has authored the books *Ensemble Methods: Foundations and Algorithms* and *Machine Learning* (in Chinese), and published more than 150 papers in top-tier international journals or conference proceedings. He has received various awards/honors including the National Natural Science Award of China, the PAKDD Distinguished Contribution Award, the IEEE ICDM Outstanding Service Award, the Microsoft Professorship Award, etc. He also holds 22 patents. He is an Executive Editor-in-Chief of the *Frontiers of Computer Science*, Associate Editor-in-Chief of the *Science China Information Sciences*, Action or Associate Editor of the *Machine Learning*, *IEEE Transactions on Pattern Analysis and Machine Intelligence* , *ACM Transactions on Knowledge Discovery from Data* , etc. He served as Associate Editor-in-Chief for *Chinese Science Bulletin* (2008-2014), Associate Editor for *IEEE Transactions on Knowledge and Data Engineering* (2008-2012), *IEEE Transactions on Neural Networks and Learning Systems* (2014-2017), *ACM Transactions on Intelligent Systems and Technology* (2009-2017), Neural Networks (2014-2016), *Knowledge and Information Systems* (2003-2008), etc. He founded ACML (Asian Conference on Machine Learning), served as Advisory Committee member for IJCAI (2015-2016), Steering Committee member for ICDM, PAKDD and PRICAI, and Chair of various conferences such as General co-chair of PAKDD 2014 and ICDM 2016, Program co-chair of SDM 2013 and IJCAI 2015 Machine Learning Track, and Area chair of NIPS, ICML, AAAI, IJCAI, KDD, etc. He is/was the Chair of the IEEE CIS Data Mining Technical Committee (2015-2016), the Chair of the CCF-AI(2012-), and the Chair of the Machine Learning Technical Committee of CAAI (2006-2015). He is a foreign member of the Academy of Europe, and a Fellow of the ACM, AAAI, AAAS, IEEE, IAPR, IET/IEE, CCF, and CAAI.