# CorrLog: Correlated Logistic Models for Joint Prediction of Multiple Labels

**Wei Bian**
Centre for Quantum Computation and Intelligent Systems,
University of Technology, Sydney
brian.weibian@gmail.com

**Bo Xie**
Georgia Tech Center for Music Technology,
Georgia Institute of Technology
bo.xie@gatech.edu

**Dacheng Tao**
Centre for Quantum Computation and Intelligent Systems,
University of Technology, Sydney
dacheng.tao@gmail.com

## Abstract

In this paper, we present a simple but effective method for multi-label classification (MLC), termed Correlated Logistic Models (Corrlog), which extends multiple Independent Logistic Regressions (ILRs) by modeling the pairwise correlation between labels. Algorithmically, we propose an efficient method for learning parameters of Corrlog, which is based on regularized maximum pseudo-likelihood estimation and has a linear computational complexity with respect to the number of labels. Theoretically, we show that Corrlog enjoys a satisfying generalization bound which is independent of the number of labels. The effectiveness of Corrlog on modeling label correlations is illustrated by a toy example, and further experiments on real data show that Corrlog achieves competitive performance compared with popular MLC algorithms.

## 1 Introduction

Multi-label classification (MLC) extends conventional single label classification (SLC) by allowing an instance to be simultaneously assigned to multiple labels from a label set. It occurs naturally from a wide range of practical problems, from document categorization, image classification, to music annotation. Due to its great generality and potential applications, MLC has received increasing attentions from researchers in various domains in the past few years (Zhang and Zhou, 2007)(Cheng and Hüllermeier, 2009)(Hsu et al., 2009)(Tsoumakas et al., 2010)(Petterson and Caetano, 2010).

The key problem with MLC is how to utilize label correlations to boost the classification performance, motivated by which a dozen of MLC algorithms have been proposed in recent years. For example, multi-label $k$-nearest neighbor (MLkNN) (Zhang and Zhou, 2007) and instance based logistic regression (IBLR) (Cheng and Hüllermeier, 2009) use label correlations within the neighborhood of an instance for posterior inference. Classifier Chain (CC) (Read et al., 2009), as well as its ensemble and probabilistic variants (Dembczyński et al., 2010a), incorporate label correlations into a chain of binary classifiers, where the prediction of a label uses previous labels as features. Another group of algorithms are built upon concurrence or structure information extracted from the label set, including pruned problem transformation (PPT) (Read, 2008), hierarchical binary relevance (HBR) (Bianchi et al., 2006) and hierarchy of multi-label classifiers (HOMER) (Tsoumakas et al., 2010). It is impossible to refer to all the relevant literature. The recent surveys (Tsoumakas et al., 2010), (Tsoumakas and Katakis, 2007) contain many references omitted from this paper.

Besides above algorithmic studies, some theoretical properties of MLC have also been investigated. (Dembczyński et al., 2010b) and (Dembczyński et al., 2010a) give an in-depth discussion on label dependence, by which they show the difference between marginal and conditional dependence of labels and categorize popular MLC algorithms accordingly. And, most recently, Gao and Zhou (2011) studies the conditions for the consistency of MLC algorithms with surrogate loss functions.

In this paper, we present a very simple but effective method for MLC, termed Correlated Logistic Models (CorrLog),which is built upon Independent Logistic

Appearing in Proceedings of the $15^{th}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX. 

Regressions (ILRs) and explicitly models the pairwise correlation between labels. We propose an efficient learning algorithm for Corrlog, based on regularized maximum pseudo-likelihood estimation. In particular, the computational complexity of the learning procedure is linear with respect to the number of labels and thus nearly the same with ILRs. Theoretically, we show that Corrlog enjoys a satisfying generalization, which is independent of the number of labels. This indicates the generalization bound holds with high confidence even with a large number of labels. Toy example is used to illustrate how Corrlog improves ILRs in modeling label correlations. And thorough experimental results on MLC benchmark datasets suggest that CorrLog performs competitively with popular MLC algorithms.

## 2 Correlated Logistic Models

We study the problem of learning a joint prediction $\mathbf{y} = d(\mathbf{x}) : \mathcal{X} \mapsto \mathcal{Y}$, where the instance space $\mathcal{X} = \{\mathbf{x} : \|\mathbf{x}\| \leq 1, \mathbf{x} \in \mathbb{R}^D\}$ and the label space $\mathcal{Y} = \{-1, 1\}^m$. By assuming the conditional independence among labels, we can model MLC by a set of independent logistic regressions (ILRs). The posterior probability of ILRs is given by

$$\begin{aligned} p_{lr}(\mathbf{y}|\mathbf{x}) &= \prod_{i=1}^{m} p_{lr}(\mathbf{y}_i|\mathbf{x}) \\ &= \prod_{i=1}^{m} \frac{\exp\left(\mathbf{y}_i \beta_i^T \mathbf{x}\right)}{\exp\left(\beta_i^T \mathbf{x}\right) + \exp\left(-\beta_i^T \mathbf{x}\right)}, \end{aligned} \tag{1}$$

where $\beta_i \in \mathbb{R}^D$ is the coefficients for the $i$-th logistic regression (LR) in ILRs. For convenience, the bias of standard LR is omitted here, which is equivalent to augmenting $\mathbf{x}$ with a constant. Although (1) can be learned efficiently and the probabilistic formulation helps handle the imbalance problem encountered in MLC, it ignores the label correlations and tends to underfit the true posterior $p_0(\mathbf{y}|\mathbf{x})$, especially when the number of labels $m$ is large.

### 2.1 Correlated Logistic Regressions

To overcome this problem, it is essential to consider the label correlations in (1). In this paper, we propose to augment (1) with a simple function $q(\mathbf{y})$, and reformulate the posterior probability as

$$p(\mathbf{y}|\mathbf{x}) \propto p_{lr}(\mathbf{y}|\mathbf{x}) q(\mathbf{y}). \tag{2}$$

Since $q(\mathbf{y})$ is independent of $\mathbf{x}$, (2) models the label correlations in an average sense. This is similar to the concept of "marginal dependence " in MLC (Dembczyński et al., 2010b). However, they are intrinsically different, because we integrate this correlation into the posterior probability, which directly aims at prediction. In addition, the idea used in (2) for correlation modeling is also distinct from the "Curds and Whey" procedure in (Breiman and Friedman, 1997) which corrects outputs of multivariate linear regression by reconsidering their correlations to the true responses.

In this paper, we define $q(\mathbf{y})$ as a quadratic form of $\mathbf{y}$,

$$q(\mathbf{y}) = \exp\left\{\sum_{i<j} \alpha_{ij} \mathbf{y}_i \mathbf{y}_j\right\}. \tag{3}$$

where $\mathbf{y}_i$ and $\mathbf{y}_j$ are positively correlated if $\alpha_{ij} > 0$ and negatively correlated if $\alpha_{ij} < 0$. It is possible to define $\alpha_{ij}$ as functions of $\mathbf{x}$, but this will drastically increase number of model parameters by $\mathcal{O}(m^2 D)$ (and thus the model complexity) even by linear functions. Substituting (3) to (2), we obtain the Correlated Logistic models (CorrLog)

$$p(\mathbf{y}|\mathbf{x}; \Theta) \propto \exp\left\{\sum_{i=1}^{m} \mathbf{y}_i \beta_i^T \mathbf{x} + \sum_{i<j} \alpha_{ij} \mathbf{y}_i \mathbf{y}_j\right\}, \tag{4}$$

where the model parameter $\Theta$ contains $\boldsymbol{\beta} = [\beta_1, ..., \beta_m]$ and $\boldsymbol{\alpha} = [\alpha_{12}, ..., \alpha_{(m-1)m}]^T$. CorrLog is a simple modification of (1), by using quadratic term to adjust the joint prediction, in order to explore label correlations.

CorrLog is closely related to popular statistical models for joint binary responses modeling. For example, (4) can be regarded as an Ising model (Ravikumar et al., 2010) for $\mathbf{y}$ conditioned on $\mathbf{x}$. Besides, it also defines a very simple conditional random fields (CRFs) (Lafferty et al., 2001). However, it should be distinguished from the Collective multi-label classification method (Ghamrawi and McCallum, 2005), where the later directly applies CRFs and intuitively defines features $\phi(\mathbf{y}_i, \mathbf{y}_j)$ taking values from $\{0, 1, 2, 3\}$, but (4) has distinct motivations, which improves ILRs by modeling label dependence. Moreover, the classical model multivariate Probit (MP) (Ashford and Sowden, 1970) also models pairwise correlations in $\mathbf{y}$. However, it utilizes Gaussian latent variables for correlation modeling, which is essentially different from CorrLog.

### 2.2 Approximate Learning via Pseudo Likelihood Maximization

Exact learning of CorrLog (4) needs the calculation of the partition function, which is computationally intractable. An effective method to avoid this calculation is using the pseudo likelihood method (Besag, 1975), which is developed for spatial dependence analysis and has been widely applied to the estimation of

various models, such as the Ising models (Ravikumar et al., 2010) and CRFs (Sutton and Mccallum, 2007). The pseudo likelihood of (4) is given by

$$\widetilde{p}(\mathbf{y}|\mathbf{x};\Theta)=\prod_{i=1}^{m}p(\mathbf{y}_i|\mathbf{y}_{-i},\mathbf{x};\Theta), \tag{5}$$

where $\mathbf{y}_{-i}=[\mathbf{y}_1,...,\mathbf{y}_{i-1},\mathbf{y}_{i+1},...,\mathbf{y}_m]$ and the conditional probability $p(\mathbf{y}_i|\mathbf{y}_{-i},\mathbf{x};\Theta)$ can be directly obtained from (4),

$$p(\mathbf{y}_i|\mathbf{y}_{-i},\mathbf{x};\Theta)=\frac{1}{1+\exp\left\{-2\mathbf{y}_i\left(\beta_i^T\mathbf{x}+\sum_{j=i+1}^{m}\alpha_{ij}\mathbf{y}_j+\sum_{j=1}^{i-1}\alpha_{ji}\mathbf{y}_j\right)\right\}}. \tag{6}$$

Given a set of i.i.d. training data $\mathcal{D}=\{(\mathbf{x}^{(l)},\mathbf{y}^{(l)}),l=1,2,...,n\}$, the negative log pseudo likelihood is given by

$$\mathcal{L}(\Theta)=-\frac{1}{n}\sum_{l=1}^{n}\sum_{i=1}^{m}\log p(\mathbf{y}_i^{(l)}|\mathbf{y}_{-i}^{(l)},\mathbf{x}^{(l)};\Theta). \tag{7}$$

Then, the optimal parameter can be obtained by the following regularized minimization

$$\widehat{\Theta}=\arg\min_{\Theta}\mathcal{L}_r(\Theta)=\arg\min_{\Theta}\mathcal{L}(\Theta)+\lambda_1\|\boldsymbol{\beta}\|^2+\lambda_2\|\boldsymbol{\alpha}\|^2. \tag{8}$$

The two $\ell_2$ regularizations on $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ control their respective search spaces to avoid overfitting. In particular, the former one penalizes the coefficients for the prediction of individual labels while the latter penalizes the coefficients for correlations between labels. An alternative choice of the regularization on $\boldsymbol{\alpha}$ is the $\ell_1$ norm, which is intensively studied in recent year (Ravikumar et al., 2010). However, empirically we found that, given the satisfying performance of $\ell_2$ regularization, the $\ell_1$ regularization offers limited improvement compared to its heavy computational cost when the dataset is large. Thus, we prefer to adopt the $\ell_2$ regularization in this paper.

Problem (8) is convex and smooth, and thus can be solved efficiently by using the gradient based method, where the gradient of $\mathcal{L}_r(\Theta)$ is calculated by

$$\begin{cases}\nabla\mathcal{L}_{r\beta_i}=\frac{1}{n}\sum\limits_{l=1}^{n}\xi_{li}\mathbf{x}^{(l)}+2\lambda_1\beta_i\\ \nabla\mathcal{L}_{r\alpha_{ij}}=\frac{1}{n}\sum\limits_{l=1}^{n}\left(\xi_{li}\mathbf{y}_j^{(l)}+\xi_{lj}\mathbf{y}_i^{(l)}\right)+2\lambda_2\alpha_{ij}\end{cases} \tag{9}$$

with

$$\xi_{li}=\frac{-2\mathbf{y}_i^{(l)}}{1+\exp\left\{\begin{array}{l}2\mathbf{y}_i^{(l)}\left(\beta_i^T\mathbf{x}^{(l)}+\sum\limits_{j=i+1}^{m}\alpha_{ij}\mathbf{y}_j^{(l)}\right.\\ \left.+\sum\limits_{j=1}^{i-1}\alpha_{ji}\mathbf{y}_j^{(l)}\right)\end{array}\right\}}. \tag{10}$$

**Algorithm 1** Learning CorrLog by Regularized Pseudo Likelihood Maximization

**Input:** Training data $\mathcal{D}$, initialization $\boldsymbol{\beta}^{(0)}=\mathbf{0}$, $\boldsymbol{\alpha}^{(0)}=\mathbf{0}$, $\mathbf{B}^{(1)}=\boldsymbol{\beta}^{(0)}$, $\mathbf{A}^{(1)}=\boldsymbol{\alpha}^{(0)}$, $t=1$, and learning rate $\eta$, where $1/\eta$ is set larger than the Lipschitz constant of $\nabla\mathcal{L}_r(\Theta)$.
**Output:** Model parameters $\widehat{\Theta}=(\boldsymbol{\beta}^{(t)},\boldsymbol{\alpha}^{(t)})$.
**repeat**
$\boldsymbol{\beta}^{(t)}=\mathbf{B}^{(t)}-\eta\nabla\mathcal{L}_{r\boldsymbol{\beta}}(\mathbf{B}^{(t)},\mathbf{A}^{(t)})$
$\boldsymbol{\alpha}^{(t)}=\mathbf{A}^{(t)}-\eta\nabla\mathcal{L}_{r\boldsymbol{\alpha}}(\mathbf{B}^{(t)},\mathbf{A}^{(t)})$
$c_{t+1}=\left(1+\sqrt{1+4c_t^2}\right)/2$
$\mathbf{B}^{(t+1)}=\boldsymbol{\beta}^{(t)}+\frac{c_t-1}{c_{t+1}}\left(\boldsymbol{\beta}^{(t)}-\boldsymbol{\beta}^{(t-1)}\right)$
$\mathbf{A}^{(t+1)}=\boldsymbol{\alpha}^{(t)}+\frac{c_t-1}{c_{t+1}}\left(\boldsymbol{\alpha}^{(t)}-\boldsymbol{\alpha}^{(t-1)}\right)$
$t=t+1$
**until** Converged

It can been seen that the calculation of gradients with respect to $\beta_i$ and $\alpha_{ij}$ share (10). Accordingly, the computational cost of learning CorrLog is linear with respect to the number of labels and nearly the same as that of learning ILRs. This is favorable for largescale MLC problems. In particular, we use the accelerated gradient descent method developed in (Beck and Teboulle, 2009) to solve (8), which is proved to have an optimal convergence rate in the sense of Nemirovsky and Yudin (Nemirovsky and Yudin, 1983). Algorithm 1 summarizes the pseudo code.

### 2.3 Joint Prediction

Given the model parameter $\widehat{\Theta}$, the joint prediction of $\widehat{\mathbf{y}}$ on a new instance $\mathbf{x}$ can be obtained by maximum a posteriori (MAP) estimation

$$\begin{aligned}\widehat{\mathbf{y}}&=\arg\max_{\mathbf{y}\in\mathcal{Y}}p(\mathbf{y}|\mathbf{x};\widehat{\Theta})\\ &\propto\exp\left\{\sum_{i=1}^{m}\mathbf{y}_i\widehat{\beta}_i^T\mathbf{x}+\sum_{i<j}\widehat{\alpha}_{ij}\mathbf{y}_i\mathbf{y}_j\right\}.\end{aligned} \tag{11}$$

We solve (11) by using the max-product algorithm (Bishop, 2006). Although sophisticated decoding algorithms are available, such as loopy belief propagation (Murphy et al., 1999), our empirical studies show that max-product algorithm performs well for joint prediction.

## 3 Generalization Analysis

In the design of effective learning algorithms, an important issue is the estimation of the accuracy. To this end, we derive a generalization bound for CorrLog. Based upon the stability analysis introduced in

(Bousquet and Elisseeff, 2002), we first show the learning of CorrLog by maximizing the pseudo likelihood is stable, and then prove that the generalization error can be bounded by the empirical error plus a term related to the stability but independent of the number of labels.

### 3.1 The stability of CorrLog

The stability of a learning algorithm indicates how much the learned model changes according to a small change of the training data set. Denote by $\mathcal{D}^k$ a training data set the same with $\mathcal{D}$ but replace the $k$-th training sample $(\mathbf{x}^{(k)}, \mathbf{y}^{(k)})$ with another independent sample $(\mathbf{x}', \mathbf{y}')$. Suppose $\widehat{\Theta}$ and $\widehat{\Theta}^k$ are learned from (8) on $\mathcal{D}$ and $\mathcal{D}^k$, respectively. We intend to show that

$$\|\widehat{\Theta}^k - \widehat{\Theta}\| \triangleq \sum_{i=1}^{m} \|\widehat{\boldsymbol{\beta}}_i^k - \widehat{\boldsymbol{\beta}}_i\| + \sum_{i<j} |\widehat{\boldsymbol{\alpha}}_{ij}^k - \widehat{\boldsymbol{\alpha}}_{ij}|, \ \forall\, 1 \le k \le n, \tag{12}$$

is bounded by $\mathcal{O}(1/n)$. The following auxiliary model $\widehat{\Theta}^{\setminus k}$ on $\mathcal{D}$ will be used

$$\begin{aligned} \widehat{\Theta}^{\setminus k} &= \arg\min_{\Theta} \mathcal{L}_r^{\setminus k}(\Theta) \\ &= \mathcal{L}^{\setminus k}(\Theta) + \lambda_1 \|\boldsymbol{\beta}\|_2^2 + \lambda_2 \|\boldsymbol{\alpha}\|_2^2, \end{aligned} \tag{13}$$

where

$$\mathcal{L}^{\setminus k}(\Theta) = -\frac{1}{n} \sum_{l \neq k} \sum_{i=1}^{m} \log p(\mathbf{y}_i^{(l)} | \mathbf{y}_{-i}^{(l)}, \mathbf{x}^{(l)}; \Theta). \tag{14}$$

We start by bounding $\mathcal{L}_r(\widehat{\Theta}^{\setminus k}) - \mathcal{L}_r(\widehat{\Theta})$, which particularly results in the following two Lemmas.

**Lemma 1.** *For $1 \le k \le n$,*

$$\begin{aligned} \mathcal{L}_r(\widehat{\Theta}^{\setminus k}) - \mathcal{L}_r(\widehat{\Theta}) \le \frac{1}{n} \Bigg( & \sum_{i=1}^{m} \log p(\mathbf{y}_i^{(k)} | \mathbf{y}_{-i}^{(k)}, \mathbf{x}^{(k)}; \widehat{\Theta}^{\setminus k}) \\ & - \sum_{i=1}^{m} \log p(\mathbf{y}_i^{(k)} | \mathbf{y}_{-i}^{(k)}, \mathbf{x}^{(k)}; \widehat{\Theta}) \Bigg) \end{aligned} \tag{15}$$

*Proof.* Denote by RHS the righthand side of (15), we have

$$\text{RHS} = \left( \mathcal{L}_r(\widehat{\Theta}^{\setminus k}) - \mathcal{L}_r^{\setminus k}(\widehat{\Theta}^{\setminus k}) \right) - \left( \mathcal{L}_r(\widehat{\Theta}) - \mathcal{L}_r^{\setminus k}(\widehat{\Theta}) \right). \tag{16}$$

Furthermore, the definition of $\widehat{\Theta}^{\setminus k}$ implies

$$\mathcal{L}_r^{\setminus k}(\widehat{\Theta}^{\setminus k}) \le \mathcal{L}_r^{\setminus k}(\widehat{\Theta}). \tag{17}$$

Combining (16) and (17), we have (15). This completes the proof. □

**Lemma 2.** *For $1 \le k \le n$,*

$$\mathcal{L}_r(\widehat{\Theta}^{\setminus k}) - \mathcal{L}_r(\widehat{\Theta}) \ge \lambda_1 \|\widehat{\boldsymbol{\beta}}^{\setminus k} - \widehat{\boldsymbol{\beta}}\|^2 + \lambda_2 \|\widehat{\boldsymbol{\alpha}}^{\setminus k} - \widehat{\boldsymbol{\alpha}}\|^2. \tag{18}$$

*Proof.* Define a function

$$f(\Theta) = \mathcal{L}_r(\Theta) - \lambda_1 \|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|^2 - \lambda_2 \|\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}\|^2. \tag{19}$$

Then it is sufficient to show that $f(\widehat{\Theta}^{\setminus k}) \ge f(\widehat{\Theta})$. By using (8), we have

$$\begin{aligned} f(\Theta) = \mathcal{L}(\Theta) &+ 2\lambda_1 \text{trace}(\boldsymbol{\beta}^T \widehat{\boldsymbol{\beta}}) - \lambda_1 \|\widehat{\boldsymbol{\beta}}\|^2 \\ &+ 2\lambda_2 \boldsymbol{\alpha}^T \widehat{\boldsymbol{\alpha}} - \lambda_2 \|\widehat{\boldsymbol{\alpha}}\|^2. \end{aligned} \tag{20}$$

Clearly, $f(\Theta)$ is convex, and

$$\nabla f(\Theta) = \nabla \mathcal{L}(\Theta) + 2\lambda_1 \widehat{\boldsymbol{\beta}} + 2\lambda_2 \widehat{\boldsymbol{\alpha}}. \tag{21}$$

Since $\widehat{\Theta}$ minimizes $\mathcal{L}_r(\widehat{\Theta})$, we have

$$\begin{aligned} \nabla f(\widehat{\Theta}) &= \nabla \mathcal{L}(\widehat{\Theta}) + 2\lambda_1 \widehat{\boldsymbol{\beta}} + 2\lambda_2 \widehat{\boldsymbol{\alpha}} \\ &= \nabla \mathcal{L}_r(\widehat{\Theta}) = 0, \end{aligned} \tag{22}$$

which implies $f(\widehat{\Theta}) \le f(\widehat{\Theta}^{\setminus k})$. This completes the proof. □

Afterward, by checking the Lipschitz continuous property of $\log p(\mathbf{y}_i | \mathbf{y}_{-i}, \mathbf{x}; \Theta)$, we have the following Lemma 3.

**Lemma 3.** *For $\forall\ (\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$ and $1 \le k \le n$*

$$\begin{aligned} & \Big| \sum_{i=1}^{m} \log p(\mathbf{y}_i | \mathbf{y}_{-i}, \mathbf{x}; \widehat{\Theta}) - \sum_{i=1}^{m} \log p(\mathbf{y}_i | \mathbf{y}_{-i}, \mathbf{x}; \widehat{\Theta}^{\setminus k}) \Big| \\ & \le 2 \sum_{i=1}^{m} \|\widehat{\beta}_i - \widehat{\beta}_i^{\setminus k}\| + 4 \sum_{i<j} |\widehat{\alpha}_{ij} - \widehat{\alpha}_{ij}^{\setminus k}|. \end{aligned} \tag{23}$$

*Proof.* The proof is completed by verifying that

$$\|\partial \log p(\mathbf{y}_i | \mathbf{y}_{-i}, \mathbf{x}; \Theta) / \partial \beta_i\| \le 2\|\mathbf{x}\| \le 2 \tag{24}$$

and

$$|\partial \log p(\mathbf{y}_i | \mathbf{y}_{-i}, \mathbf{x}; \Theta) / \partial \alpha_{ij}| \le 4 |\mathbf{y}_i \mathbf{y}_j| = 4. \tag{25}$$

□

By combining the above three Lemmas, we have the following Theorem 1 that shows the stability of CorrLog.

**Theorem 1.** *For $\forall\ \mathcal{D}$ and $\mathcal{D}^k$, $1 \le k \le n$, it holds that*

$$\sum_{i=1}^{m} \|\widehat{\boldsymbol{\beta}}_i^k - \widehat{\boldsymbol{\beta}}_i\| + \sum_{i<j} |\widehat{\boldsymbol{\alpha}}_{ij}^k - \widehat{\boldsymbol{\alpha}}_{ij}| \le \frac{16}{\min(\lambda_1, \lambda_2) n}. \tag{26}$$

*Proof.* By rearranging the combination of (15), (18) and (23), we have

$$\|\widehat{\boldsymbol{\beta}}^{\setminus k}-\widehat{\boldsymbol{\beta}}\|^2+\|\widehat{\boldsymbol{\alpha}}^{\setminus k}-\widehat{\boldsymbol{\alpha}}\|^2 \leq \frac{4}{\min(\lambda_1,\lambda_2)n}\left(\sum_{i=1}^m\|\widehat{\beta}_i-\widehat{\beta}_i^{\setminus k}\|+\sum_{i<j}|\widehat{\alpha}_{ij}-\widehat{\alpha}_{ij}^{\setminus k}|\right). \tag{27}$$

Since

$$\|\widehat{\boldsymbol{\beta}}^{\setminus k}-\widehat{\boldsymbol{\beta}}\|^2+\|\widehat{\boldsymbol{\alpha}}^{\setminus k}-\widehat{\boldsymbol{\alpha}}\|^2 \geq \frac{1}{2}\left(\sum_{i=1}^m\|\widehat{\beta}_i-\widehat{\beta}_i^{\setminus k}\|+\sum_{i<j}|\widehat{\alpha}_{ij}-\widehat{\alpha}_{ij}^{\setminus k}|\right)^2, \tag{28}$$

we have

$$\sum_{i=1}^m\|\widehat{\beta}_i-\widehat{\beta}_i^{\setminus k}\|+\sum_{i<j}|\widehat{\alpha}_{ij}-\widehat{\alpha}_{ij}^{\setminus k}| \leq \frac{8}{\min(\lambda_1,\lambda_2)n}. \tag{29}$$

Further, since $\mathcal{D}$ and $\mathcal{D}$ differ from each other only on the $k$-th training sample, by using the same proof strategy, we have

$$\sum_{i=1}^m\|\widehat{\beta}_i^k-\widehat{\beta}_i^{\setminus k}\|+\sum_{i<j}|\widehat{\alpha}_{ij}^k-\widehat{\alpha}_{ij}^{\setminus k}| \leq \frac{8}{\min(\lambda_1,\lambda_2)n}. \tag{30}$$

Then, (26) is obtained immediately. This completes the proof. □

### 3.2 Generalization Bound

We first define a loss function to measure the generalization error. Considering that CorrLog predicts labels by MAP estimation, we define the loss function by using the log probability

$$\ell(\mathbf{x},\mathbf{y};\Theta)=\begin{cases}1, & f(\mathbf{x},\mathbf{y},\Theta)<0\\ 1-f(\mathbf{x},\mathbf{y},\Theta)/\gamma, & 0\leq f(\mathbf{x},\mathbf{y},\Theta)<\gamma\\ 0, & f(\mathbf{x},\mathbf{y},\Theta)\geq\gamma,\end{cases} \tag{31}$$

where the constant $\gamma>0$ and

$$\begin{aligned}f(\mathbf{x},\mathbf{y},\Theta)&=\log p(\mathbf{y}|\mathbf{x};\Theta)-\max_{\mathbf{y}'\neq\mathbf{y}}\log p(\mathbf{y}'|\mathbf{x};\Theta)\\&=\left(\sum_{i=1}^m\mathbf{y}_i\beta_i^T\mathbf{x}+\sum_{i<j}\alpha_{ij}\mathbf{y}_i\mathbf{y}_j\right)\\&\quad-\max_{\mathbf{y}'\neq\mathbf{y}}\left(\sum_{i=1}^m\mathbf{y}'_i\beta_i^T\mathbf{x}+\sum_{i<j}\alpha_{ij}\mathbf{y}'_i\mathbf{y}'_j\right).\end{aligned} \tag{32}$$

The loss function (31) is defined analogue to the loss function used in binary classification, where $f(\mathbf{x},\mathbf{y},\Theta)$ is replaced with the margin $y\mathbf{w}^T\mathbf{x}$ if a linear classifier $\mathbf{w}$ is used. Besides, (31) gives a 0 loss only if all dimensions of $\mathbf{y}$ are correctly predicted, which emphasizes the joint prediction in MLC. By using this loss function, the generalization error and the empirical error are given by

$$\mathcal{R}(\widehat{\Theta})=\mathbb{E}_{\mathbf{xy}}\ell(\mathbf{x},\mathbf{y};\widehat{\Theta}), \tag{33}$$

and

$$\widehat{\mathcal{R}}(\widehat{\Theta})=\frac{1}{n}\sum_{l=1}^n\ell(\mathbf{x}^{(l)},\mathbf{y}^{(l)};\widehat{\Theta}). \tag{34}$$

According to (Bousquet and Elisseeff, 2002), an exponential bound exists for $\mathcal{R}(\widehat{\Theta})$ if CorrLog has uniform stability with respect to the loss function (31). The following Theorem 2 shows this condition holds.

**Theorem 2.** *For $\forall(\mathbf{x},\mathbf{y})\in\mathcal{X}\times\mathcal{Y}$, $\mathcal{D}$ and $\mathcal{D}^k$, $1\leq k\leq n$, it holds that*

$$|\ell(\mathbf{x},\mathbf{y};\widehat{\Theta})-\ell(\mathbf{x},\mathbf{y};\widehat{\Theta}^k)|\leq\frac{32}{\gamma\min(\lambda_1,\lambda_2)n}. \tag{35}$$

*Proof.* First, by using (31), we have

$$\gamma|\ell(\mathbf{x},\mathbf{y};\widehat{\Theta})-\ell(\mathbf{x},\mathbf{y};\widehat{\Theta}^k)|\leq|f(\mathbf{x},\mathbf{y},\widehat{\Theta})-f(\mathbf{x},\mathbf{y},\widehat{\Theta}^k)|. \tag{36}$$

Then, by introducing notation $A(\mathbf{x},\mathbf{y},\boldsymbol{\beta},\boldsymbol{\alpha})=\sum_{i=1}^m\mathbf{y}_i\beta_i^T\mathbf{x}+\sum_{i<j}\alpha_{ij}\mathbf{y}_i\mathbf{y}_j$ and rewriting $f(\mathbf{x},\mathbf{y},\Theta)=A(\mathbf{x},\mathbf{y},\boldsymbol{\beta},\boldsymbol{\alpha})-\max_{\mathbf{y}'\neq\mathbf{y}}A(\mathbf{x},\mathbf{y}',\boldsymbol{\beta},\boldsymbol{\alpha})$, (36) gives rise to

$$\begin{aligned}&\gamma|\ell(\mathbf{x},\mathbf{y};\widehat{\Theta})-\ell(\mathbf{x},\mathbf{y};\widehat{\Theta}^k)|\\&\leq\left|A(\mathbf{x},\mathbf{y},\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\alpha}})-A(\mathbf{x},\mathbf{y},\widehat{\boldsymbol{\beta}}^k,\widehat{\boldsymbol{\alpha}}^k)\right|\\&+|\max_{\mathbf{y}'\neq\mathbf{y}}A(\mathbf{x},\mathbf{y}',\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\alpha}})-\max_{\mathbf{y}'\neq\mathbf{y}}A(\mathbf{x},\mathbf{y}',\widehat{\boldsymbol{\beta}}^k,\widehat{\boldsymbol{\alpha}}^k)|.\end{aligned} \tag{37}$$

Since it holds for general functions $h_1(u)$ and $h_2(u)$ that[1]

$$|\max_u h_1(u)-\max_u h_2(u)|\leq\max_u|h_1(u)-h_2(u)|, \tag{38}$$

we have

$$\begin{aligned}&\gamma|\ell(\mathbf{x},\mathbf{y};\widehat{\Theta})-\ell(\mathbf{x},\mathbf{y};\widehat{\Theta}^k)|\\&\leq\left|A(\mathbf{x},\mathbf{y},\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\alpha}})-A(\mathbf{x},\mathbf{y},\widehat{\boldsymbol{\beta}}^k,\widehat{\boldsymbol{\alpha}}^k)\right|\\&+\max_{\mathbf{y}'\neq\mathbf{y}}\left|A(\mathbf{x},\mathbf{y}',\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\alpha}})-A(\mathbf{x},\mathbf{y}',\widehat{\boldsymbol{\beta}}^k,\widehat{\boldsymbol{\alpha}}^k)\right|\\&\leq 2\max_{\mathbf{y}}\left(\sum_{i=1}^m|\mathbf{y}_i(\widehat{\beta}_i-\widehat{\beta}_i^k)^T\mathbf{x}|+\sum_{i<j}|(\widehat{\alpha}_{ij}-\widehat{\alpha}_{ij}^k)\mathbf{y}_i\mathbf{y}_j|\right)\\&\leq 2\left(\sum_{i=1}^m\|\widehat{\beta}_i-\widehat{\beta}_i^k\|+2\sum_{i<j}|\widehat{\alpha}_{ij}-\widehat{\alpha}_{ij}^k|\right).\end{aligned} \tag{39}$$

[1]Suppose $u_1^\star$ and $u_2^\star$ maximize $h_1(u)$ and $h_2(u)$ respectively, and without loss of generality $h_1(u_1^\star)\geq h_2(u_2^\star)$, we have $|h_1(u_1^\star)-h_2(u_2^\star)|=h_1(u_1^\star)-h_2(u_2^\star)\leq h_1(u_1^\star)-h_2(u_1^\star)\leq\max_u|h_1(u)-h_2(u)|$.

Afterward, the proof is completed by applying Theorem 1. □

Finally, we have the following theorem on the generalization bound of Corrlog.

**Theorem 3.** *Given i.i.d. training data* $\mathcal{D} = \{(\mathbf{x}^{(l)}, \mathbf{y}^{(l)}), l = 1, 2, ..., n\}$ *from* $\mathcal{X} \times \mathcal{Y}$ *and regularization parameters* $\lambda_1$, $\lambda_2$, *we have with at least probability* $1 - \delta$,

$$\mathcal{R}(\widehat{\Theta}) \leq \widehat{\mathcal{R}}(\widehat{\Theta}) + \frac{32}{\gamma \min(\lambda_1, \lambda_2) n} + \left( \frac{64}{\gamma \min(\lambda_1, \lambda_2)} + 1 \right) \sqrt{\frac{\log 1/\delta}{2n}}. \quad (40)$$

*Proof.* Given the stability result from above Theorem 2, the proof of (40) is similar to that of the Theorem 12 in (Bousquet and Elisseeff, 2002), and thus we omit the details due to space limitation and leave the readers to (Bousquet and Elisseeff, 2002). □

**Remarks** A notable point of Theorem 3 is that the generalization bound (40) is independent of the number of labels. This is of great importance, since it indicates that the generalization error of CorrLog can be bounded with a high confidence even the number of labels is large. Moreover, the regularization parameters result in a form of $\min(\lambda_1, \lambda_2)$, because these two regularization items penalize different aspects of CorrLog and one cannot be relieved by the other.

## 4 A Toy Example

We design a simple toy example to illustrate the capacity of CorrLog on label correlation modeling. In particular, we show that when ILRs fails drastically due to ignoring the label correlations (underfitting), CorrLog performs well. Consider a two-label classification problem on a 2-D plane, where each instance $\mathbf{x}$ is sampled uniformly from the unit disc $\|\mathbf{x}\| \leq 1$ and the corresponding labels $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2)$ are defined by

$$\mathbf{y}_1 = \text{sign}(\eta_1^T \tilde{\mathbf{x}}) \quad \text{and} \quad \mathbf{y}_2 = \text{OR}\left(\mathbf{y}_1, \text{sign}(\eta_2^T \tilde{\mathbf{x}})\right), \quad (41)$$

where $\eta_1 = (1, 1, -0.5)$, $\eta_2 = (-1, 1, -0.5)$ and the augmented feature is $\tilde{\mathbf{x}} = [\mathbf{x}^T, 1]^T$. The sign$(\cdot)$ function takes value 1 or $-1$, and the OR$(\cdot, \cdot)$ operation outputs 1 if either of its input is 1. The definition of $\mathbf{y}_2$ makes the two labels correlated. We generate 1,000 random samples according to above setting and split them into training and test sets, each of which contains 500 samples.

Figure 1 shows that true labels of test data, the predictions of ILRs and the predictions of CorrLog, where different labels are marked by different colors. In (a), the disc is divided into three regions, $-/-$, $-/+$ and $+/+$, where the two black boundaries are specified by $\eta_1$ and $\eta_2$, respectively. In (b), the first boundary $\eta_1$ properly learned by ILRs, while the second one is learned wrongly. This is because the second label is highly correlated to the first label, but ILRs ignores such correlation. The misclassification rate measured by 0-1 loss is 0.197. In contrast, CorrLog predicts correct labels for most instances with a 0-1 loss 0.068. Besides, it is interesting to note that the correlation between the two labels are "asymmetric", for the first label is not affected by the second. This asymmetry contributes the most to the misclassification of CorrLog, because previous definition implies that only symmetric correlations are modeled in CorrLog.

## 5 Real Data Experiments

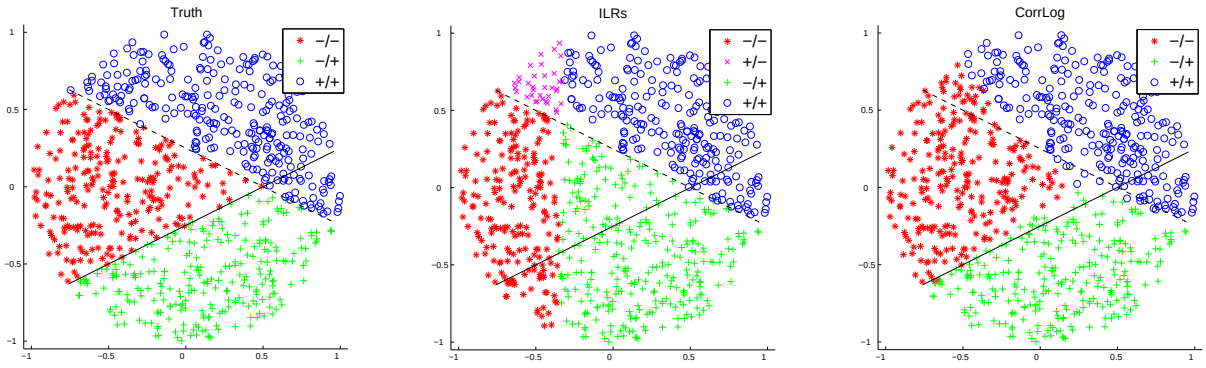
We evaluated the proposed method on six datasets, spanning different application domains. They are Enron (text), Slashdot (text), Emotions (music) and three sub-datasets selected from the Yahoo dataset (web data). The datasets were obtained from Mulan and Meka's website [2]. Summary of the basic information of the datasets is illustrated in Table 1. We can see that they vary in feature dimension and number of labels.

Table 1: Dataset summary. # train stands for the number of training instances and # test for the number of test instances. $d$ is the dimension of the features and $m$ is the number of labels. Education, Recreation and Science are from Yahoo web datasets.

| Datasets | # train | # test | $d$ | $m$ |
|---|---|---|---|---|
| Education | 2000 | 3000 | 550 | 33 |
| Emotions | 391 | 202 | 72 | 6 |
| Enron | 1123 | 579 | 1001 | 53 |
| Recreation | 2000 | 3000 | 606 | 22 |
| Science | 2000 | 3000 | 743 | 40 |
| Slashdot | 2338 | 1444 | 1079 | 22 |

**Experimental Settings:** To demonstrate the effectiveness of utilizing label correlation, we compared our algorithm's performance with ILRs. Also, we compared with three state-of-the-arts methods – Instance-Based Learning by Logistic Regression (IBLR) Cheng and Hüllermeier (2009), Multi-label k-Nearest Neighbor (MLkNN) Zhang and Zhou (2007) and Classifier Chains (CC) Read et al. (2009). We used six different measures to evaluate the performance. These in-

[2] (Mulan) http://mulan.sourceforge.net/ and (Meka) http://meka.sourceforge.net/




(a) (b) (c)

Figure 1: Comparing of ILRs and Corrlog on the two-label toy example: (a) true labels of test data; (b) predictions given by ILRs; (c) predictions given by CorrLog. The dash and solid black boundaries are specified by $\eta_1$ and $\eta_2$. In the legend, "+" and "-" stand for positive and negative labels, respectively, e.g., "$-/+$" means $\mathbf{y}_1 = -1$ and $\mathbf{y}_2 = 1$, and so on.

clude different loss functions (Hamming loss and zero-one loss) and other popular measures (precision, recall, F-1 score and accuracy). The details of these evaluation measures can be found in Read et al. (2009)Dembczyński et al. (2010a)Tsoumakas and Vlahavas (2007). We used the train-test split given in the website. The parameter pair was chosen according to the best accuracy via 5-fold cross validation on the training set. All experiments are implemented in MatLab on a dual-core laptop. The training time is less than 50 seconds while the test time is less than 30 seconds for all six datasets .

**MLC Results and Discussion:** Table. 2 summarizes the experimental results of all five algorithms evaluated by the six measures. By comparing the results of CorrLog and ILRs, we can clearly see the improvements on joint prediction. Except the Hamming loss, CorrLog outperforms ILRs on all datasets. Especially, the reduction of zero-one loss is significant on "Recreation" and "Emotions". This confirms the value of correlation modeling to joint prediction. However, it should be noticed that CorrLog performs only comparable with ILRs when the performance is measured by Hamming loss. This is because Hamming loss treats the prediction of each label individually. Besides, compared with the other three algorithms, CorrLog also shows promising performance.

## 6 Conclusion

In this paper, we have presented a new model CorrLog for MLC. Built upon IRLs, CorrLog explicitly models the pairwise correlation between labels, and thus improves the effectiveness for MLC. However, due to the learning algorithm based on regularized maximum pseudo-likelihood estimation, the computational complexity of Corrlog is linear with respect the number of labels and thus nearly the same with IRLs. This is particular favorable for MLC. Further, we proved a generalization bound for CorrLog, which is independent of the number of labels and thus suggests that satisfying generalization holds with high confidence even the number of labels is large. Thorough empirical studies on a toy dataset and several MLC benchmark datasets show the effectiveness of Corrlog by comparing with representative MLC algorithms, such as MLkNN, IBLR, and CC.

## References

Ashford, J., Sowden, R., 1970. Multi-variate probit analysis. Biometrics 26, 535–546.

Beck, A., Teboulle, M., 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Img. Sci. 2, 183–202.

Besag, J., 1975. Statistical Analysis of Non-Lattice Data. Journal of the Royal Statistical Society. Series D (The Statistician) 24 (3), 179–195.

Bianchi, N. C., Gentile, C., Zaniboni, L., 2006. Incremental algorithms for hierarchical classification. JMLR 7, 31–54.

Bishop, C. M., 2006. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Bousquet, O., Elisseeff, A., March 2002. Stability and generalization. Journal of Machine Learning Research 2, 499–526.

Table 2: Performance comparison of CorrLog, ILRs, and three other state-of-the-art methods. IBLR stands for Instance-Based Learning by Logistic Regression Cheng and Hüllermeier (2009), MLkNN for Multi-label k-Nearest Neighbor Zhang and Zhou (2007) and CC for Classifier Chains Read et al. (2009).

| | | Ham | 0-1 loss | Prec | Recall | F-1 Score | Accu |
|---|---|---|---|---|---|---|---|
| Education | CorrLog | 0.049 | **0.701** | **0.456** | **0.445** | **0.425** | **0.389** |
| | ILRs | 0.042 | 0.774 | 0.366 | 0.341 | 0.337 | 0.307 |
| | IBLR | **0.041** | 0.853 | 0.215 | 0.191 | 0.195 | 0.183 |
| | MLkNN | **0.041** | 0.866 | 0.201 | 0.170 | 0.178 | 0.166 |
| | CC | 0.045 | 0.707 | 0.447 | 0.398 | 0.405 | 0.376 |
| Recreation | CorrLog | 0.075 | **0.683** | **0.428** | **0.418** | **0.405** | **0.378** |
| | ILRs | 0.063 | 0.771 | 0.335 | 0.326 | 0.318 | 0.294 |
| | IBLR | **0.061** | 0.882 | 0.147 | 0.133 | 0.136 | 0.131 |
| | MLkNN | **0.061** | 0.883 | 0.147 | 0.131 | 0.135 | 0.130 |
| | CC | 0.071 | 0.704 | 0.408 | 0.378 | 0.380 | 0.358 |
| Emotions | CorrLog | 0.206 | **0.683** | **0.677** | **0.680** | **0.655** | **0.572** |
| | ILRs | 0.235 | 0.797 | 0.617 | 0.604 | 0.573 | 0.481 |
| | IBLR | **0.196** | 0.688 | 0.654 | 0.618 | 0.609 | 0.537 |
| | MLkNN | 0.212 | 0.802 | 0.645 | 0.551 | 0.567 | 0.480 |
| | CC | 0.238 | 0.718 | 0.629 | 0.619 | 0.597 | 0.519 |
| Science | CorrLog | 0.041 | **0.721** | **0.391** | **0.393** | **0.372** | **0.345** |
| | ILRs | 0.043 | 0.816 | 0.322 | 0.351 | 0.317 | 0.280 |
| | IBLR | **0.035** | 0.907 | 0.137 | 0.128 | 0.127 | 0.118 |
| | MLkNN | **0.035** | 0.948 | 0.073 | 0.063 | 0.066 | 0.062 |
| | CC | 0.041 | 0.724 | 0.389 | 0.366 | 0.364 | 0.340 |
| Enron | CorrLog | **0.048** | **0.855** | **0.653** | **0.554** | **0.577** | **0.469** |
| | ILRs | 0.054 | 0.891 | 0.588 | 0.510 | 0.518 | 0.408 |
| | IBLR | 0.059 | 0.929 | 0.515 | 0.385 | 0.414 | 0.317 |
| | MLkNN | 0.053 | 0.945 | 0.585 | 0.363 | 0.421 | 0.318 |
| | CC | 0.059 | 0.869 | 0.560 | 0.514 | 0.511 | 0.405 |
| Slashdot | CorrLog | **0.046** | **0.576** | **0.524** | 0.520 | **0.509** | **0.486** |
| | ILRs | **0.046** | 0.670 | 0.440 | 0.480 | 0.445 | 0.415 |
| | IBLR | 0.047 | 0.814 | 0.210 | 0.202 | 0.204 | 0.199 |
| | MLkNN | 0.048 | 0.826 | 0.194 | 0.184 | 0.187 | 0.184 |
| | CC | 0.050 | 0.609 | 0.505 | **0.527** | 0.503 | 0.474 |

Breiman, L., Friedman, J. H., 1997. Predicting multivariate responses in multiple linear regression (with discussion). The Journal of the Royal Statistical Society Series B 54, 5–54.

Cheng, W., Hüllermeier, E., 2009. Combining instance-based learning and logistic regression for multilabel classification. Machine Learning 76 (2-3), 211–225.

Dembczyński, K., Cheng, W., Hüllermeier, E., 2010a. Bayes optimal multilabel classification via probabilistic classifier chains. In: The 27th International Conference on Machine Learning (ICML 2010).

Dembczyński, K., Waegeman, W., Cheng, W., Hüllermeier, E., 2010b. On label dependence in multi-label classification. In: ICML 2010 Workshop on Learning from Multi-label data (MLD 10). pp. 5–13.

Gao, W., Zhou, Z.-H., 2011. On the consistency of multi-label learning. In: The 24th Annual Conference on Learning Theory (COLT'11).

Ghamrawi, N., McCallum, A., 2005. Collective multi-label classification. In: CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management. ACM, pp. 195–200.

Hsu, D., Kakade, S., Langford, J., Zhang, T., 2009. Multi-label prediction via compressed sensing. In: Bengio, Y., Schuurmans, D., Lafferty, J., Williams,

C. K. I., Culotta, A. (Eds.), Advances in Neural Information Processing Systems 22. pp. 772–780.

Lafferty, J. D., McCallum, A., Pereira, F. C. N., 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. San Francisco, CA, USA, pp. 282–289.

Murphy, K. P., Weiss, Y., Jordan, M. I., 1999. Loopy belief propagation for approximate inference: An empirical study. In: Uncertainty in Artificial Intelligence. pp. 467–475.

Nemirovsky, A. S., Yudin, D. B., 1983. Problem Complexity and Method Efficiency in Optimization. Wiley-Interscience Series in Discrete Mathematics. John Wiley & Sons, New York, USA.

Petterson, J., Caetano, T., 2010. Reverse multi-label learning. In: Lafferty, J., Williams, C. K. I., Shawe-Taylor, J., Zemel, R., Culotta, A. (Eds.), Advances in Neural Information Processing Systems 23. pp. 1912–1920.

Ravikumar, P., Wainwright, M. J., Lafferty, J. D., 2010. High-dimensional ising model selection using l 1 -regularized logistic regression. Annals of Statistics 38, 1287–1319.

Read, J., 2008. A Pruned Problem Transformation Method for Multi-label classification. In: Proc. 2008 New Zealand Computer Science Research Student Conference (NZCSRS 2008). pp. 143–150.

Read, J., Pfahringer, B., Holmes, G., Frank, E., 2009. Classifier chains for multi-label classification. In: ECML/PKDD. pp. 254–269.

Sutton, C. A., Mccallum, A., 2007. Piecewise pseudolikelihood for efficient training of conditional random fields. In: International Conference on Machine Learning. pp. 863–870.

Tsoumakas, G., Katakis, I., 2007. Multi-label classification: An overview. International Journal of Data Warehousing and Mining 3 (3), 1–13.

Tsoumakas, G., Katakis, I., Vlahavas, I. P., 2010. Mining multi-label data. In: Data Mining and Knowledge Discovery Handbook. pp. 667–685.

Tsoumakas, G., Vlahavas, I., Sep. 2007. Random k-Labelsets: An Ensemble Method for Multilabel Classification. In: Proceedings of the 18th European Conference on Machine Learning (ECML 2007). Warsaw, Poland, pp. 406–417.

Zhang, M. L., Zhou, Z. H., 2007. Ml-knn: A lazy learning approach to multi-label learning. Pattern Recognition 40 (7), 2038–2048.