Uncorrelated Group LASSO

Deguang Kong[1], **Ji Liu**[2], **Bo Liu**[3] **and Xuan Bao**[4]
[1]Samsung Research America, [2]University of Rochester, [3]Philips Research North America, [4]Google Inc.
doogkong@gmail.com, jliu@cs.rochester.edu, boliu@cs.umass.edu, xbxuanbao8@gmail.com

Abstract

$\ell_{2,1}$-norm is an effective regularization to enforce a simple group sparsity for feature learning. To capture some subtle structures among feature groups, we propose a new regularization called *exclusive group $\ell_{2,1}$-norm*. It enforces the sparsity at the intra-group level by using $\ell_{2,1}$-norm, while encourages the selected features to distribute in different groups by using ℓ_2 norm at the inter-group level. The proposed *exclusive group $\ell_{2,1}$-norm* is capable of eliminating the feature correlations in the context of feature selection, if highly correlated features are collected in the same groups. To solve the generic *exclusive group $\ell_{2,1}$-norm* regularized problems, we propose an efficient iterative re-weighting algorithm and provide a rigorous convergence analysis. Experiment results on real world datasets demonstrate the effectiveness of the proposed new regularization and algorithm.

Introduction

Sparse coding starts from Lasso (R.Tibshirani 1994) using ℓ_1 norm for 2-class feature selection, and grows for $\ell_{2,1}$ norm based multi-class feature selection. Group Lasso (Yuan and Lin 2006) incorporates feature group information into the feature learning process. The sparsity term is effective due to the inherent sparse structures of the real world data. Other extensions of spare coding includes exclusive Lasso (Zhou, Jin, and Hoi 2010), fused Lasso (Tibshirani et al. 2005), and generalized Lasso (Roth 2004), (Liu, Yuan, and Ye 2013). Due to the intuitive interpretation of sparse learning results, structural sparsity based methods have been widely applied to solve many practical problems, such as medical image analysis (Yang et al. 2010), cancer prediction (Gao and Church 2005), and gene-expression analysis (Ji et al. 2009).

Among all the above methods, $\ell_{2,1}$-norm based method as well as its variants and extensions (Liu, Ji, and Ye 2012), (Nie et al. 2010), (Yuan and Lin 2006) are considered as one of the most effective feature selection method. $\ell_{2,1}$ norm of a matrix $\mathbf{W} \in \Re^{p\times K}$ is defined as

$$\|\mathbf{W}\|_{2,1} = \sum_{i=1}^{p} \sqrt{\sum_{j=1}^{K} W_{ij}^2} = \sum_{i=1}^{p} \|\mathbf{w}^i\|_2, \qquad (1)$$

which enforces sparsity at row-level (corresponding to features). If each row of $\mathbf{W}$ is viewed as a group (*i.e.,* corresponding to each feature), $\ell_{2,1}$-norm can be viewed as a special case of group Lasso (Yuan and Lin 2006).

However, in standard $\ell_{2,1}$-norm based approach (such as $f(\mathbf{W}) + \lambda\|\mathbf{W}\|_{2,1}$), the correlations among different features are simply ignored. Thus, some strongly correlated variables tend to be in or out of the model together. Moreover, the correlated variables may share similar properties and thus reveal overlapped or redundant information. Especially when the number of selected variables are limited, more discriminant information with minimum correlations are desirable for prediction or classification tasks. Therefore, it is natural to eliminate the correlations in feature selection process using $\ell_{2,1}$-norm.

In this paper, we propose an *exclusive group $\ell_{2,1}$ term* for uncorrelated feature selection. In "*exclusive group $\ell_{2,1}$*" regularizer, highly correlated features are put into the same group such that the most discriminant features can survive using standard $\ell_{2,1}$ regularization at the intra-group level. Meanwhile, it enforces ℓ_2 norm at the inter-group level since there is no need for sparsity if different groups are jointly considered. Essentially, this term can be viewed as a $\ell_{\{2,1\};2}$ norm. It excels standard $\ell_{2,1}$ method because it eliminates the correlated features.

To the best of our knowledge, this is the *first* work that eliminates the correlated features in $\ell_{2,1}$ method through a novel regularization term: *exclusive group $\ell_{2,1}$-norm*. This can be viewed as an extension of exclusive feature learning work (Zhou, Jin, and Hoi 2010), (Kong et al. 2014) on multi-class/multi-task settings, which can be potentially applied in medical image analytics, bio-marker recognition, and fine-grained image classification, etc.

Contribution The main contributions of this paper are summarized as follows.

- A new regularizer known as *exclusive group $\ell_{2,1}$-norm* is proposed to eliminate the feature correlations for multi-class/multi-task feature learning;
- An effective iterative re-weighting algorithm is proposed to tackle the complicated non-smoothness in the exclusive group $\ell_{2,1}$ term;
- Promising results on real world datasets suggest potential applications of the proposed method.

Notation Throughout the paper, all matrices are written in boldface uppercase, vectors are written in boldface lowercase, and scalars are denoted by lower-case letters. n is the number of data points, p is the dimension of data, and K is the number of classes in the dataset. A group of variables is a subset $g \subset \{1, 2, \cdots, p\}$. Thus, the set of possible groups is the power set of $\{1, 2, \cdots, p\}$: $\mathcal{P}(\{1, 2, \cdots, p\})$. $\mathcal{G}_g \in \mathcal{P}(\{1, 2, \cdots, p\})$ denotes a set of group g, which is known in advance depending on the applications. If two groups have one common variable, we say that they are overlapped. For a matrix $\mathbf{W} \in \Re^{p\times K}$, $\mathbf{w}^i$ is the i-th row of $\mathbf{W}$, while $\mathbf{w}_j$ is the j-th column of $\mathbf{W}$, *i.e.,* $\mathbf{W} = [\mathbf{w}^1; \mathbf{w}^2; \cdots; \mathbf{w}^p]$. For any group variable $\mathbf{W}_{\mathcal{G}_g} \in \Re^{p\times K}$, only entries in the group g are preserved, which are the same as those in $\mathbf{W}$, and the other entries are set to zeros. For example, if $\mathcal{G}_g = \{1, 2, 4\}$, $\mathbf{W}_{\mathcal{G}_g} = [\mathbf{w}^1; \mathbf{w}^2; \mathbf{0}; \mathbf{w}^4; \mathbf{0}; \cdots, \mathbf{0}]$. For any differentiable function $f : \Re^{p\times K} \longrightarrow \Re$, $\bigtriangledown f(\mathbf{W})$ is the gradient of f at $\mathbf{W} \in \Re^{p\times K}$.

Exclusive Group $\ell_{2,1}$-Regularizer

In this paper, we propose to optimize,

$$\min_{\mathbf{W}\in\Re^{p\times K}} J_1(\mathbf{W}) = f(\mathbf{W}) + \alpha \sum_g \|\mathbf{W}_{\mathcal{G}_g}\|_{2,1}^2, \quad (2)$$

where $f(\mathbf{W})$ is a loss function involving data matrix $\mathbf{X} \in \Re^{p\times n}$ and class label matrix $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n] \in \Re^{K\times n}$. Group $\mathcal{G}_g$ is generated such that the highly correlated features are put in the same group g. $\ell_{2,1}$ norm of a matrix $\mathbf{W}$ is defined as that in Eq.(1) (Argyriou, Evgeniou, and Pontil 2008). We call $\sum_g \|\mathbf{W}_{\mathcal{G}_g}\|_{2,1}^2$ as "*exclusive group $\ell_{2,1}$-term*".

Let group indicator $\mathtt{I}_{\mathcal{G}_g} \in \{0, 1\}^p$, then $\mathbf{W}_{\mathcal{G}_g} \in \Re^{p\times K}$ preserves the feature values in group $\mathcal{G}_g$, *i.e.*,

$$(\mathbf{W}_{\mathcal{G}_g})_{\mathrm{i}:} = \begin{cases} \mathbf{W}_{\mathrm{i}:}; & \text{if } (\mathtt{I}_{\mathcal{G}_g})_{\mathrm{i}} = 1 \\ \mathbf{0}; & \text{otherwise} \end{cases} \quad (3)$$

where $\mathbf{W}_{i:}$ is the i-th row of $\mathbf{W}$.

In the "*exclusive group $\ell_{2,1}$*" regularization, highly correlated features are put into the same group (*i.e.*, group $\mathcal{G}_g$ in Eq.(2)). Thus the most discriminant features are expected to survive via "$\ell_{2,1}$ norm" at the *intra-group* level. However, for feature from different groups , there is no competitions among them. Thus non-sparsity is achieved via ℓ_2 norm at the *inter-group* level. Essentially, it can be regarded as a $\ell_{\{2,1\};2}$-operator, which can be viewed as a natural extension of $\ell_{1,2}$-norm in the multi-class setting.

Proposition 1. *Let* $\Omega^{\mathcal{G}} := \sqrt{\sum_g \|\mathbf{W}_{\mathcal{G}_g}\|_{2,1}^2}$, *then* $\Omega^{\mathcal{G}}$ *is a non-smooth convex formulation. If* $\bigcup_{g\in\mathcal{G}} = \{1, 2, \cdots, p\}$, $\Omega^{\mathcal{G}}$ *is a norm.*

How to construct group $\mathcal{G}_g$?

We construct group $\mathcal{G}_g$ such that the highly correlated features are put in the same group. Let the feature correlation matrix be $\mathbf{R} = (R_{st}) \in \Re^{p\times p}$. Clearly, $\mathbf{R} = \mathbf{R}^T$. R_{st} represents the pearson correlation between features s and t, i.e.,

$$R_{st} = \frac{|\sum_i X_{si}X_{ti}|}{\sqrt{\sum_i X_{si}^2}\sqrt{\sum_i X_{ti}^2}}, \quad (4)$$

where feature vectors are centered, *i.e.,* $\sum_i X_{si} = 0$. To make the selected features uncorrelated as much as possible, for any two features s, t, if their correlation $R_{st} > \theta$(threshold), then we put them in the same group $\mathcal{G}_g$.

Take the House dataset[1] (n=506, p=14) as an example. After computing the feature correlation matrix using 14 features, we observed that many features are highly correlated. For example, feature 5 is highly correlated with feature 6, 7, 11, and 12, *etc*. Let threshold $\theta = 0.93$, according to the definition of exclusive group $\ell_{2,1}$ term, 8 pairs will be generated such that

$$\begin{aligned}\sum_g \|\mathbf{W}_{\mathcal{G}_g}\|_{2,1}^2 = &\left\|[\mathbf{w}^3; \mathbf{w}^{10}]\right\|_{2,1}^2 + \left\|[\mathbf{w}^5; \mathbf{w}^6]\right\|_{2,1}^2 \\ &+ \left\|[\mathbf{w}^5; \mathbf{w}^7]\right\|_{2,1}^2 + \left\|[\mathbf{w}^5; \mathbf{w}^{11}]\right\|_{2,1}^2 \\ &+ \left\|[\mathbf{w}^6; \mathbf{w}^{11}]\right\|_{2,1}^2 + \left\|[\mathbf{w}^6; \mathbf{w}^{12}]\right\|_{2,1}^2 \\ &+ \left\|[\mathbf{w}^6; \mathbf{w}^{14}]\right\|_{2,1}^2 + \left\|[\mathbf{w}^7; \mathbf{w}^{11}]\right\|_{2,1}^2 .\end{aligned}$$

An illustration

We use an example to show the differences between exclusive group $\ell_{2,1}$ and $\ell_{2,1}$ term. For example, let data $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \Re^{p\times n}$ (shown in Eq.5), class indicator $\mathbf{Y} \in \Re^{k\times n}$, where $p = 7, n = 8, k = 3$, i.e.,

$$\mathbf{Y} = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \end{pmatrix}$$

By solving the standard $\ell_{2,1}$ norm using least square loss (*i.e.,* $f(\mathbf{W}) + \lambda\|\mathbf{W}\|_{2,1}$), and λ is adjusted such that 4 features are non-zeros[2], we obtain the following global optimal solution:

$$\mathbf{W}^*_{\ell_{2,1}} = \begin{pmatrix} \mathbf{0.098} & \mathbf{-0.086} & \mathbf{-0.081} \\ \mathbf{0.055} & \mathbf{0.019} & \mathbf{0.098} \\ -0.000 & 0.000 & -0.000 \\ -0.000 & 0.000 & 0.000 \\ -0.000 & 0.000 & 0.000 \\ \mathbf{0.001} & \mathbf{0.001} & \mathbf{0.002} \\ \mathbf{-0.001} & \mathbf{-0.003} & \mathbf{-0.002} \end{pmatrix} \quad (6)$$

To get optimal solution using our method of Eq.(2), we first compute the feature correlations via Eq.(4). Let $\theta = 0.3$. We obtain the following 13 groups (corresponding to $\mathcal{G}_g$ in Eq.(2)), *i.e.,*

$$\begin{aligned}&\{1, 3\}; \{2, 3\}; \{1, 4\}; \{2, 4\}; \{1, 5\}; \{2, 5\}; \{3, 6\}; \\ &\{1, 7\}; \{2, 7\}; \{3, 7\}; \{4, 7\}; \{5, 7\}; \{6, 7\}. \end{aligned} \quad (7)$$

We tune the penalty parameter α in Eq. (2) to achieve the same sparsity ratio. However, the optimal solution provided by Eq. (2) indicates different sparsity patterns from standard

[1] http://archive.ics.uci.edu/ml/datasets/housing

[2] Any number of feature can be selected. Number 4 is only for illustration purpose

$$\mathbf{X} = \begin{pmatrix} 1.1985 & -0.5955 & -0.1827 & -0.7212 & 1.0420 & -0.7221 & -0.2868 & -1.4018 \\ 0.8886 & 0.5759 & 0.9365 & 1.6189 & 0.6780 & 1.0434 & -0.6525 & 0.0545 \\ -0.8880 & -0.9235 & -0.9443 & 0.6196 & -0.8831 & -1.2441 & -0.1170 & 1.3762 \\ -0.7152 & -0.6884 & 0.8502 & -0.7252 & 0.2169 & 0.1620 & 0.3381 & 0.8394 \\ -1.4138 & 1.2350 & 0.7874 & 0.4038 & -0.6772 & 1.0096 & -1.5076 & 0.4874 \\ 0.3189 & 1.2898 & 0.2272 & -1.3298 & 1.0563 & 0.8152 & 1.6670 & -0.3058 \\ 0.6110 & -0.8934 & -1.6744 & 0.1338 & -1.4329 & -1.0640 & 0.5588 & -1.0499 \end{pmatrix} \tag{5}$$

$\ell_{2,1}$:

$$\mathbf{W}^*_{\ell_{\{2,1\};2}} = \begin{pmatrix} \mathbf{0.005} & \mathbf{-0.004} & \mathbf{-0.004} \\ \mathbf{0.003} & \mathbf{0.001} & \mathbf{0.006} \\ -0.000 & 0.000 & -0.000 \\ -0.000 & 0.000 & 0.000 \\ \mathbf{-0.003} & \mathbf{0.003} & \mathbf{0.001} \\ \mathbf{0.006} & \mathbf{0.005} & \mathbf{0.007} \\ -0.000 & -0.000 & -0.000 \end{pmatrix}. \tag{8}$$

In both solutions, a few number of features (corresponding to non-zeros rows) are selected. Clearly, the highly correlated feature pairs are selected in $\ell_{2,1}$ results, *i.e.,*

$$R_{2,7} = 0.5083, \quad R_{3,7} = 0.3859, \quad R_{4,7} = 0.4858,$$
$$R_{5,7} = 0.5555, \quad R_{6,7} = 0.3143.$$

In contrast, most highly correlated feature pairs are depressed in exclusive group $\ell_{2,1}$ results, except $R_{2,5}$ and $R_{1,5}$.

Feature selection error Feature selection is to select r features, *i.e.,* r rows of $\mathbf{W}$, such that $J(\mathbf{X}_r)$ (the residue of the selected features) is smaller. The most general loss function is least square loss, and therefore we compute the residue given the least square loss function, *i.e.,*

$$\begin{aligned} J(\mathbf{X}_r) &= \min_{\mathbf{W}} \|\mathbf{Y} - \mathbf{W}^T\mathbf{X}_r\|_F^2 \\ &= \mathrm{Tr}(\mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{Y}\mathbf{X}_r^T(\mathbf{X}_r\mathbf{X}_r^T)^{-1}\mathbf{X}_r) \end{aligned} \tag{9}$$

is minimized, where $\mathbf{X}_r$ denotes the r features selected from original data $\mathbf{X}$. This definition was the same as in . The smaller the residue, the better the feature selection results.

In the above example, $r = 4$ features are selected. Then $J(\mathbf{X}_{r=4}^{\ell_{2,1}}) = 0.4956$, $J(\mathbf{X}_{r=4}^{\ell_{\{2,1\};2}}) = 0.4259$, where $\mathbf{X}_{r=4}^{\ell_{2,1}}$ represents result using $\ell_{2,1}$, and $\mathbf{X}_{r=4}^{\ell_{\{2,1\};2}}$ represents result using Eq.(2). When $r = 3$ features are selected, $J(\mathbf{X}_{r=3}^{\ell_{2,1}}) = 0.3978$, $J(\mathbf{X}_{r=3}^{\ell_{\{2,1\};2}}) = 0.3437$. Clearly,

$$J(\mathbf{X}_{r=3}^{\ell_{\{2,1\};2}}) < J(\mathbf{X}_{r=3}^{\ell_{2,1}}), \quad J(\mathbf{X}_{r=4}^{\ell_{\{2,1\};2}}) < J(\mathbf{X}_{r=4}^{\ell_{2,1}}),$$

which further validates the effectiveness of the proposed model of Eq.(2).

Optimization Algorithm

We now consider solving Eq.(2) as an optimization problem. The challenge of solving Eq.(2) is to tackle the exclusive group $\ell_{2,1}$ regularizer. It is generally felt that exclusive group $\ell_{2,1}$ regularizer is much more difficult to solve than the Lasso term (shrinkage thresholding) or $\ell_{2,1}$ term. Existing algorithms can formulate it as a quadratic programming problem (Nocedal 2006), which can be solved by interior point method or active set method. However, the computational cost is expensive, which limits its use in practice. Recently, a primal-dual algorithm (Yang et al. 2012) is proposed to solve the similar problem, which casts the non-smooth problem into a min-max problem. However, the algorithm is a gradient descent type method and converges slowly.

We *first* derive a much more effective yet simple algorithm which can handle this non-smooth exclusive generic group $\ell_{2,1}$ term. Moreover, the proposed algorithm is a general algorithm, which allows arbitrary structure on feature space, irrespective of specific feature organizations based on different applications, i.e., linear structure (Yuan, Liu, and Ye 2011), tree structure (Liu and Ye 2010), graph structure (Jacob, Obozinski, and Vert 2009), *etc*.

Method

The general idea of the proposed algorithm is to find an auxiliary function for Eq.(2) that can be easily solved. Then the updating rules for $\mathbf{W}$ are derived. Finally, we prove the solution is exactly the optimal solution we are seeking for the original problem. Since it is a convex problem, the optimal solution is the global solution. Instead of directly optimizing Eq.(2), we propose to optimize the following objective (the reasons will be seen immediately below), i.e.,

$$J_2(\mathbf{W}) = f(\mathbf{W}) + \alpha \mathrm{Tr}(\mathbf{W}^T\mathbf{F}\mathbf{W}), \tag{10}$$

where $\mathbf{F} \in \Re^{p\times p}$ is a diagonal matrices which encodes the exclusive group information, and its diagonal elements are given by[3]

$$\mathbf{F}_{ii} = \Big(\sum_g \frac{(\mathtt{I}_{\mathcal{G}_g})_i \|\mathbf{W}_{\mathcal{G}_g}\|_{2,1}}{\|\mathbf{w}^i_{\mathcal{G}_g}\|_2}\Big), \tag{11}$$

where $\mathbf{w}^i_{\mathcal{G}_g}$ is the i-th row of $\mathbf{W}_{\mathcal{G}_g} (1 \le i \le p)$. Let $\mathtt{I}_{\mathcal{G}_g} \in \{0,1\}^{p\times 1}$ be the group index indicator for group $g \in \mathcal{G}$. For example, group $\mathcal{G}_1$ is $\{1,2\}$, then $\mathtt{I}_{\mathcal{G}_1} = [1,1,0,\cdots,0]$. Thus the group variable $\mathbf{W}_{\mathcal{G}_g}$ can be explicitly expressed as $\mathbf{W}_{\mathcal{G}_g} = \mathrm{diag}(\mathtt{I}_{\mathcal{G}_g}) \times \mathbf{W}$.

In the following, we propose an effective *iteratively reweighted* algorithm to find out the optimal global solution for $\mathbf{W}$, where in each iteration, $\mathbf{W}$ is updated along the gradient descent direction. Take the derivative of Eq.(10) w.r.t

[3]When $\|\mathbf{w}^i_{\mathcal{G}_g}\| = 0$, F_{ii} is related to the subgradient of $\mathbf{W}$ w.r.t to $\mathbf{w}^i$. We can not set $F_{ii} = 0$, otherwise, the derived algorithm cannot be guaranteed to converge. We can regularize $F_{ii} = \Big(\sum_g \frac{(\mathtt{I}_{\mathcal{G}_g})_i \|\mathbf{W}_{\mathcal{G}_g}\|_{2,1}}{\sqrt{\|\mathbf{w}^i_{\mathcal{G}_g}\|_2^2 + \epsilon}}\Big)$, then the derived algorithm can be proved to minimize the regularized $\sum_g \|(\mathbf{W}+\epsilon)_{\mathcal{G}_g}\|_{2,1}^2$. It is easy to see the regularized exclusive $\ell_{2,1}$ norm of $\mathbf{W}$ approximates exclusive $\ell_{2,1}$ norm of $\mathbf{W}$ when $\epsilon \to 0^+$.

$\mathbf{W}$ and set $\frac{\partial J_2}{\partial \mathbf{W}} = 0$. We have

$$\nabla_{\mathbf{W}} f(\mathbf{W}) + 2\alpha \mathbf{F}\mathbf{W} = 0. \quad (12)$$

Then the complete algorithm is:
(1) Updating $\mathbf{W}^t$ via Eq.(12);
(2) Updating $\mathbf{F}^t$ via Eq.(11).

The above two steps are iterated until the algorithm converges. Using the above updating rules, we can obtain the global optimal solution for Eq.(10). We can prove the obtained optimal solution is exactly the global optimal solution for Eq.(2).

Convergence Analysis

In the following, we prove the convergence of our algorithm.

Theorem 2. *Under the updating rule of Eq.(12),* $J_1(\mathbf{W}^{t+1}) - J_1(\mathbf{W}^t) \leq 0$.

To prove Theorem 2, we need two lemmas.

Lemma 1. *Under the updating rule of Eq.(12),* $J_2(\mathbf{W}^{t+1}) < J_2(\mathbf{W}^t)$.

Lemma 2. *Under the updating rule of Eq.(12),*

$$\Big(J_1(\mathbf{W}^{t+1}) - J_1(\mathbf{W}^t)\Big) \leq \Big(J_2(\mathbf{W}^{t+1}) - J_2(\mathbf{W}^t)\Big) \quad (13)$$

Proof of Theorem 2

From Lemma 1 and Lemma 2, it is easy to see $\Big(J_1(\mathbf{W}^{t+1}) - J_1(\mathbf{W}^t)\Big) \leq 0$. This completes the proof. □

Proof of Lemma 1

Eq.(10) is a convex function, and the optimal solution of Eq.(12) is obtained by taking the derivative $\frac{\partial J_2}{\partial \mathbf{W}} = 0$, thus the obtained $\mathbf{W}^*$ is the global optimal solution, $J_2(\mathbf{W}^{t+1}) < J_2(\mathbf{W}^t)$. □

Before the proof of Lemma 2, we need the following **Proposition**.

Proposition 3.

$$Tr(\mathbf{W}^T\mathbf{F}\mathbf{W}) = \sum_{g=1}^{G} \|\mathbf{W}_{\mathcal{G}_g}\|_{2,1}^2.$$

Proof of Proposition 3

$$\begin{aligned}
\mathrm{Tr}(\mathbf{W}^T\mathbf{F}\mathbf{W}) &= \sum_i (\sum_j W_{ij}^2) \sum_g (\mathbf{F}_{\mathcal{G}_g})_{ii} \\
&= \sum_g \sum_i \frac{(\mathtt{I}_{\mathcal{G}_g})_i (\sum_j W_{ij}^2) \|\mathbf{W}_{\mathcal{G}_g}\|_{2,1}}{\|\mathbf{w}^i_{\mathcal{G}_g}\|_2} \\
&= \sum_g \|\mathbf{W}_{\mathcal{G}_g}\|_{2,1} \sum_i \frac{(\mathtt{I}_{\mathcal{G}_g})_i (\sum_j W_{ij}^2)}{\|\mathbf{w}^i_{\mathcal{G}_g}\|_2} \\
&= \sum_g \sum_i \frac{(\mathtt{I}_{\mathcal{G}_g})_i \|\mathbf{w}^i_{\mathcal{G}_g}\|^2 (\sum_j \|\mathbf{w}^j_{\mathcal{G}_g}\|)}{\|\mathbf{w}^i_{\mathcal{G}_g}\|_2} \\
&= \sum_g \|\mathbf{W}_{\mathcal{G}_g}\|_{2,1}^2. \quad (14)
\end{aligned}$$

where $\mathbf{F}_{\mathcal{G}_g} \in \Re^{p\times p}$ is a diagonal matrix, and $(\mathbf{F}_{\mathcal{G}_g})_{ii} = \frac{(\mathtt{I}_{\mathcal{G}_g})_i \|\mathbf{W}_{\mathcal{G}_g}\|_{2,1}}{\|\mathbf{w}^i_{\mathcal{G}_g}\|_2}$.

Proof of Lemma 2 Let Δ = LHS-RHS of Eq.(13), and $\mathbf{F}_{\mathcal{G}_g}$ be a matrix w.r.t group $\mathcal{G}_g$, *i.e.,*

$$\mathbf{F} = \sum_g \mathbf{F}_{\mathcal{G}_g}, \quad (\mathbf{F}_{\mathcal{G}_g})_{ii} = \frac{(\mathtt{I}_{\mathcal{G}_g})_i \|\mathbf{W}_{\mathcal{G}_g}\|_{2,1}}{\|\mathbf{w}^i_{\mathcal{G}_g}\|_2}.$$

Therefore, we have

$$\begin{aligned}
\Delta &= \sum_g \|\mathbf{W}^{t+1}_{\mathcal{G}_g}\|_{2,1}^2 - \mathrm{Tr}\Big[\sum (\mathbf{W}^{t+1}_{\mathcal{G}_g})^T (\mathbf{F}^t_{\mathcal{G}_g}) \mathbf{W}^{t+1}_{\mathcal{G}_g}\Big] \\
&\quad - \sum_g \|\mathbf{W}^t_{\mathcal{G}_g}\|_{2,1}^2 + \mathrm{Tr}\Big[\sum_g (\mathbf{W}^t_{\mathcal{G}_g})^T (\mathbf{F}^t_{\mathcal{G}_g}) (\mathbf{W}^t_{\mathcal{G}_g})\Big] \\
&= \sum_g \|\mathbf{W}^{t+1}_{\mathcal{G}_g}\|_{2,1}^2 - \mathrm{Tr}\Big[\sum (\mathbf{W}^{t+1}_{\mathcal{G}_g})^T (\mathbf{F}^t_{\mathcal{G}_g}) \mathbf{W}^{t+1}_{\mathcal{G}_g}\Big] \quad (15) \\
&= \sum_g \Big(\sum_{i\in\mathcal{G}_g} \|(\mathbf{w}^i_{\mathcal{G}_g})^{(t+1)}\|\Big)^2 \\
&\quad - \sum_g \sum_{i\in\mathcal{G}_g} \frac{(\mathtt{I}_{\mathcal{G}_g})_i \|(\mathbf{w}^i_{\mathcal{G}_g})^{(t+1)}\|^2 \Big(\sum_{i\in\mathcal{G}_g} \|(\mathbf{w}^i_{\mathcal{G}_g})^{(t)}\|\Big)}{\|(\mathbf{w}^i_{\mathcal{G}_g})^{(t)}\|} \quad (16) \\
&= \sum_g \Big(\sum_{i\in\mathcal{G}_g} \|(\mathbf{w}^i_{\mathcal{G}_g})^{(t+1)}\|\Big)^2 \\
&\quad - \sum_g \Big(\sum_{i\in\mathcal{G}_g} \frac{\|(\mathbf{w}^i_{\mathcal{G}_g})^{(t+1)}\|^2}{\|(\mathbf{w}^i_{\mathcal{G}_g})^{(t)}\|}\Big)\Big(\sum_{i\in\mathcal{G}_g} \|(\mathbf{w}^i_{\mathcal{G}_g})^{(t)}\|\Big) \quad (17) \\
&= \sum_g \Big[\Big(\sum_{i\in\mathcal{G}_g} a_i b_i\Big)^2 - \Big(\sum_{i\in\mathcal{G}_g} a_i^2\Big)\Big(\sum_{i\in\mathcal{G}_g} b_i^2\Big)\Big] \leq 0, \quad (18)
\end{aligned}$$

where $a_i = \frac{\|(\mathbf{w}^i_{\mathcal{G}_g})^{(t+1)}\|}{\sqrt{\|(\mathbf{w}^i_{\mathcal{G}_g})^{(t)}\|}}$, $b_i = \sqrt{\|(\mathbf{w}^i_{\mathcal{G}_g})^{(t)}\|}$. Due to proposition 3, Eq.(15) is equivalent to Eq.(16). Eq.(18) holds due to cauchy inequality: for any scalar a_i, b_i, $(\sum_i a_i b_i)^2 \leq (\sum_i a_i^2)(\sum_i b_i^2)$. Clearly, $\Delta \leq 0$. □

Discussion In practice, for the linear loss function or a simple loss function, e.g., least square loss, we can directly get a closed-form solution of the gradient descend $\nabla_{\mathbf{W}} f(\mathbf{W})$ w.r.t $\mathbf{W}$. Thus, we can get a closed-form solution of $\mathbf{W}$ in Eq.(12) in each updating process.

For example, if $f(\mathbf{W}) = \frac{1}{2}\|\mathbf{W} - \mathbf{A}\|_2^2$, we can get the optimal solution of Eq.(10) given by $\mathbf{W} = (\mathbf{I} + 2\alpha\mathbf{F})^{-1}\mathbf{A}$. Notice that $\mathbf{F}$ is a diagonal matrix, therefore $\mathbf{W}_{ij} = \frac{1}{1+2\alpha\mathbf{F}_{ii}}\mathbf{A}_{ij}$.

When it is difficult to get the closed-form solution of the gradient descend $\nabla_{\mathbf{W}} f(\mathbf{W})$, we may refer to the accelerated proximal gradient method (a.k.a FISTA method) (Nesterov 2007; Beck and Teboulle 2009) to transform the original problem into:

$$\mathbf{W}^{t+1} = \arg\min_{\mathbf{U}} \frac{1}{2}\|\mathbf{U} - \mathbf{B}\|_F^2 + \beta \sum_g \|\mathbf{U}_{\mathcal{G}_g}\|_{2,1}^2, \quad (19)$$

where $\mathbf{B} = \mathbf{W}^t - \frac{1}{L_t}\nabla f(\mathbf{W}^t)$, $\beta = \alpha/L_t$, and L_t is a step size chosen at each iteration using certain searching strategy. Eq.(19) can be easily solved via our algorithm. One advantage of our algorithm is that it is very fast, since $\mathbf{F}$ is a diagonal matrix. In practice, we find the the above algorithm

(a) A sample image from MSRC (b) A sample image from Trevid

Figure 1: Sample images and feature values of the datasets.

Table 1: Multi-Label Datasets

Dataset	#Size	#Dimension	#Class
Trevid	3721	512	39
MSRC	591	384	23
Barcelona	139	384	4

converges very fast. Typically, it converges to global optimal solution within 10-20 times[4]. It is necessary to emphasize here that the proposed algorithm is a *generic* algorithm. This indicates that our algorithm can be easily extended to handle any loss function $f(\mathbf{W})$ (*e.g.*, logistic loss or hinge loss) with respect to any *composite group* structure (Wang and Ye 2015).

Experiments

To validate the effectiveness of our method, we conduct experiments on multi-label datasets for image annotation.

Datasets Descriptions

Table 1 describes the datasets, and Fig.(1) shows sample images and feature values.

Barcelona[5] dataset contains 139 images from 4 categories (i.e., "building", "flora","people" and "sky"). We extract the 384-dimensional color moment features.

MSRC[6] dataset contains 591 images from 23 classes (i.e., "car", "road", "building", "grass", etc.). Around 80% of the images are annotated with at least one class, and around 3 classes per images on average. We extract the 384-dimensional color moment features.

TREVID2005[7] dataset contains 137 broadcast news videos, which are segmented into 61901 sub-shots and labeled with 39 concepts(i.e., "people", "table", "animal", etc.). We randomly sample the dataset such that each concept (label) has at least 10 video key frames. We extract the 512-dimensional GIST (Oliva and Torralba 2001) features as recommended by previous researches, which encode rough geometry and spatial structures within an image.

[4]Overall, FISTA method can achieve ϵ-optimal solution in $\mathcal{O}(1/\sqrt{\epsilon})$ iterations. We adopt our method in the proximal operator step that converges fast. The formal proof in iterative step is left for future work.

[5]http://mlg.ucd.ie/content/view/61

[6]http://research.microsoft.com/en-us/projects/\\objectclassrecognition/

[7]http://www-nlpir.nist.gov/projects/tv2005/

Feature selection error

Regarding selection error of Eq.(9), we compare our method against the standard $\ell_{2,1}$-method (Liu, Ji, and Ye 2012; Argyriou, Evgeniou, and Pontil 2008; Nie et al. 2010) using *least square loss*. We adjust the parameter α such that the number of nonzero rows in $\mathbf{W}$ (*i.e.,* optimal solution) is r. In our method, group $\mathcal{G}_g$ is generated by setting the threshold $\theta = 0.3$ in Eq.(4). After we get the r features, we use the value of $J(\mathbf{X}_r)$ in Eq.(9) to validate the effectiveness of the selected r features. Fig. 2 demonstrates $J(\mathbf{X}_r)$ values on three real world datasets, where exclusive group $\ell_{2,1}$ method gives smaller feature selection errors on all datasets.

Image annotation via exclusive group $\ell_{2,1}$

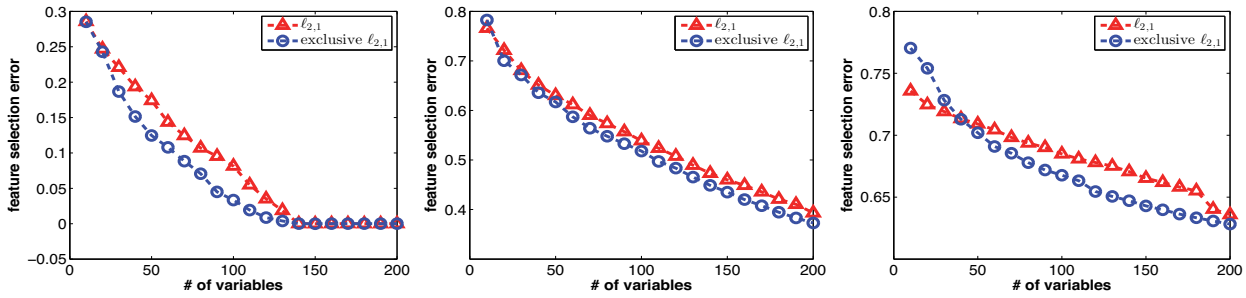
Essentially, image annotation can be viewed as a multi-label classification problem, *i.e.,* each image is given several labels simultaneously. In our approach of Eq.(2), we use *logistic loss function*, group $\mathcal{G}_g$ is generated according to the feature correlation defined in Eq.(9), where $\theta = 0.3$. We use the learned $\mathbf{W}$ from the training dataset, and predict the labels on the testing dataset.

Experiment settings In all the experiments, we use 5-fold cross validation, where 4-fold data are used for training and the remaining ones are used for testing purpose. The above procedure is repeated for 5 times and the averages of classification performances are reported in Table.2. We compare our method of Eq.(2) against the following state-of-art multi-label classification methods:

- feature learning via $\ell_{2,1}$ (Liu, Ji, and Ye 2012; Argyriou, Evgeniou, and Pontil 2008; Nie et al. 2010) (shown as $\ell_{2,1}$);
- feature learning via $\ell_{1,\infty}$ (Quattoni et al. 2009) (shown as $\ell_{1,\infty}$);
- exclusive task learning (Zhou, Jin, and Hoi 2010) (shown as "exclusive");
- SVM using multi-label ReliefF based feature selection (Kong et al. 2012) (shown as "ReliefF") ;
- Multi-Label Latent Semantic Indexing (MLSI) (Yu, Yu, and Tresp 2005);
- Multi-Label Subspace Learning (MLLS) (Ji et al. 2008);
- Multi-Label Dimension Reduction via dependence maximization(MDDM) (Zhang and Zhou 2010);

For all the other methods, we either download the codes from the authors' websites or implement the methods according to the descriptions in their papers.

Evaluation metrics We adopt the metrics defined in (Yang 1999) to evaluate the multi-label classification performance. F_1 score is a good measure of the classification accuracy, because it considers both precision and recall and can be viewed as a harmonic mean of the precision and the recall. In the multi-label case, both macro-average and micro-average F_1 scores are considered. The macro-average F_1 score is the mean of the F_1 scores of all the labels. The micro-average F_1 score is the F_1 score obtained from the summation of contingency matrices for all binary classifiers.



(a) Barcelona (b) MSRC (c) Trevid

Figure 2: (a-c) Feature selection errors of Eq.(9) on three datasets regarding different number of features. x-axis: different number of features; y-axis: feature selection error.

Table 2: Classification performance comparisons of different multi-label classification methods with 5-fold cross validation on 3 datasets. Classification methods: multi-label ReliefF (Kong et al. 2012). Multi-Label Latent Semantic Indexing (MLSI) (Yu, Yu, and Tresp 2005); Multi-Label Subspace Learning (MLLS) (Ji et al. 2008); Multi-Label Dimension Reduction via dependence maximization (MDDM) (Zhang and Zhou 2010); feature learning via $\ell_{2,1}$ (Liu, Ji, and Ye 2012), (Argyriou, Evgeniou, and Pontil 2008), (Nie et al. 2010); feature learning via $\ell_{1,\infty}$ (Quattoni et al. 2009); exclusive task learning (Zhou, Jin, and Hoi 2010); exclusive $\ell_{2,1}$ (our method). Evaluation metrics: Macro-Precision (Mac-P), Macro-F1 (Mac-F), Micro-Precision (Mic-P), Micro-F1 (Mic-F). All results shown are in percentage.

Dataset	Metrics	ReliefF	MLSI	MDDM	MLLS	$\ell_{2,1}$	$\ell_{1,\infty}$	exclusive	**exclusive $\ell_{2,1}$ (our method)**
Trevid	Mac-P	45.78	50.71	50.73	50.37	53.27	52.85	51.98	**55.12**
	Mac-F1	46.41	51.43	51.54	52.10	53.17	52.35	49.93	**54.43**
	Mic-P	33.04	39.97	40.78	41.93	41.78	42.13	39.98	**44.78**
	Mic-F1	31.13	40.23	39.78	40.51	40.93	41.53	41.21	**43.19**
MSRCv2	Mac-P	56.78	57.23	58.12	56.43	60.32	61.18	61.09	**64.35**
	Mac-F1	57.34	52.18	54.93	53.13	59.17	58.94	58.12	**62.19**
	Mic-P	47.17	48.53	46.24	48.73	51.34	52.78	49.73	**54.08**
	Mic-F1	48.23	49.01	47.25	46.52	52.15	53.20	50.33	**55.31**
Barcelona	Mac-P	75.53	74.38	70.47	71.48	76.43	77.23	75.67	**81.32**
	Mac-F1	71.14	73.19	71.28	71.30	70.23	72.14	71.45	**74.47**
	Mic-P	76.94	70.08	65.48	66.29	78.43	78.20	77.34	**81.19**
	Mic-F1	71.27	70.93	66.34	67.48	67.91	69.86	69.32	**71.89**

Explanation From Table.2, clearly, exclusive $\ell_{2,1}$ method generally achieves better performance compared to the standard $\ell_{2,1}$ method and other multi-label feature learning methods. In particular, exclusive $\ell_{2,1}$ has 10.01% and 6.30% performance improvements in macro-average and micro-average F_1 scores respectively over MLSI, one of the state-of-the-art multi-label learning methods, on dataset MSRCv2. This indicates that the elimination of redundant features helps in multi-label classification, and thus improves the performance. The proposed exclusive $\ell_{2,1}$ model can predict labels more accurately due to the capturing of inherent feature structures, instead of treating all the features equally. On Barcelona dataset, the precision of our method is around 5% better than the other methods, however, the final F_1 score of our method is only slightly better than the other methods, because the recall of our method is lower that offsets the performance gain in precision.

Applications in Business Unit (BU) The proposed algorithm can be applied for personalized health-treatment in health-care data analytics, to shorten magnetic resonance imaging scanning sessions on conventional hardware, and to identify the privacy risks of images (Tran et al. 2016) (He et al. 2015) from object detection, *etc*. The thorough discussions and comparisons are left for future work.

Conclusion

We propose an exclusive group $\ell_{2,1}$ regularizer for multi-class feature learning, which eliminates feature correlations at the intra-group level. We provide an effective algorithm to solve the new formulation, and analyze the convergence property of the proposed method. On several real world datasets, consistently smaller feature selection errors of the proposed method further validates its effectiveness. We apply the proposed uncorrelated feature learning method for the image annotation task, which further demonstrates the validity of our method. Our algorithm can be easily extended to handle other non-convex non-smooth loss functions, *e.g.*, ℓ_p loss, which is left for future work.

Acknowledgement Ji Liu is supported by the NSF grant CNS-1548078, the NEC fellowship, and the startup funding

at University of Rochester.

References

Argyriou, A.; Evgeniou, T.; and Pontil, M. 2008. Convex multi-task feature learning. *Machine Learning* 73(3):243–272.

Beck, A., and Teboulle, M. 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* 2(1):183–202.

Gao, Y., and Church, G. 2005. Improving molecular cancer class discovery through sparse non-negative matrix factorization. *Bioinformatics* 21(21):3970–3975.

He, J.; Liu, B.; Kong, D.; Bao, X.; Wang, N.; Jin, H.; and Kesidis, G. 2015. Puppies: Transformation-supported personalized privacy preserving partial image sharing. In *The Pennsylvania State University Technical Report*, CSE–2015–007.

Jacob, L.; Obozinski, G.; and Vert, J.-P. 2009. Group lasso with overlap and graph lasso. In *ICML*, 55.

Ji, S.; Tang, L.; Yu, S.; and Ye, J. 2008. Extracting shared subspace for multi-label classification. In *KDD08*.

Ji, S.; Yuan, L.; Li, Y.; Zhou, Z.; Kumar, S.; and Ye, J. 2009. Drosophila gene expression pattern annotation using sparse features and term-term interactions. In *KDD*, 407–416.

Kong, D.; Ding, C. H. Q.; Huang, H.; and Zhao, H. 2012. Multi-label relieff and f-statistic feature selections for image annotation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, 2352–2359.

Kong, D.; Ryohei, F.; Liu, J.; Nie, F.; and Ding, C. 2014. Exclusive feature learning on arbitrary structure. In *NIPS 2014*.

Liu, J., and Ye, J. 2010. Moreau-yosida regularization for grouped tree structure learning. In *NIPS*, 1459–1467.

Liu, J.; Ji, S.; and Ye, J. 2012. Multi-task feature learning via efficient l2,1-norm minimization. *CoRR* abs/1205.2631.

Liu, J.; Yuan, L.; and Ye, J. 2013. Dictionary lasso: Guaranteed sparse recovery under linear transformation. *arXiv preprint arXiv:1305.0047*.

Nesterov, Y. 2007. Gradient methods for minimizing composite objective function. *ECORE Discussion Paper*.

Nie, F.; Huang, H.; Cai, X.; and Ding, C. H. Q. 2010. Efficient and robust feature selection via joint ;2, 1-norms minimization. In *NIPS*, 1813–1821.

Nocedal, J.; Wright, S. 2006. *Numerical Optimization*. Berlin, New York: Springer-Verlag.

Oliva, A., and Torralba, A. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision* 42(3):145–175.

Quattoni, A.; Carreras, X.; Collins, M.; and Darrell, T. 2009. An efficient projection for $l_{1,\text{infinity}}$ regularization. In *ICML*, 108.

Roth, V. 2004. The generalized lasso. *IEEE Transactions on Neural Networks* 15(1):16–28.

R.Tibshirani. 1994. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* 58:267–288.

Tibshirani, R.; Saunders, M.; Rosset, S.; Zhu, J.; and Knight, K. 2005. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B* 91–108.

Tran, L.; Kong, D.; Jin, H.; and Liu, J. 2016. Privacy-cnh: A framework to detect photo privacy with convolutional neural network using hierarchical features. In *AAAI 2016*.

Wang, J., and Ye, J. 2015. Multi-layer feature reduction for tree structured group lasso via hierarchical projection. In *NIPS*.

Yang, J.; Wright, J.; Huang, T. S.; and Ma, Y. 2010. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing* 19(11):2861–2873.

Yang, T.; Jin, R.; Mahdavi, M.; and Zhu, S. 2012. An efficient primal-dual prox method for non-smooth optimization. *CoRR* abs/1201.5283.

Yang, Y. 1999. An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval* 1:67–88.

Yu, K.; Yu, S.; and Tresp, V. 2005. Multi-label informed latent semantic indexing. In *SIGIR '05*.

Yuan, M., and Lin, Y. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B* 68:49–67.

Yuan, L.; Liu, J.; and Ye, J. 2011. Efficient methods for overlapping group lasso. In *NIPS*, 352–360.

Zhang, Y., and Zhou, Z.-H. 2010. Multilabel dimensionality reduction via dependence maximization. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 4.

Zhou, Y.; Jin, R.; and Hoi, S. C. H. 2010. Exclusive lasso for multi-task feature selection. *Journal of Machine Learning Research - Proceedings Track* 9:988–995.