Robust Flexible Feature Selection via Exclusive L21 Regularization

Di Ming and **Chris Ding**
Department of Computer Science and Engineering, University of Texas at Arlington, USA
initialdiming@yahoo.com, chqding@uta.edu

Abstract

Recently, exclusive lasso has demonstrated its promising results in selecting discriminative features for each class. The sparsity is enforced on each feature across all the classes via $\ell_{1,2}$-norm. However, the exclusive sparsity of $\ell_{1,2}$-norm could not screen out a large amount of irrelevant and redundant noise features in high-dimensional data space, since each feature belongs to at least one class. Thus, in this paper, we introduce a novel regularization called "exclusive $\ell_{2,1}$", which is short for "$\ell_{2,1}$ with exclusive lasso", towards robust flexible feature selection. The exclusive $\ell_{2,1}$ regularization is the mix of $\ell_{2,1}$-norm and $\ell_{1,2}$-norm, which brings out joint sparsity at inter-group level and exclusive sparsity at intra-group level simultaneously. An efficient augmented Lagrange multipliers based optimization algorithm is proposed to iteratively solve the exclusive $\ell_{2,1}$ regularization in a row-wise fashion. Extensive experiments on twelve benchmark datasets demonstrate the effectiveness of the proposed regularization and the optimization algorithm as compared to state-of-the-arts.

1 Introduction

Feature selection plays an important role in many machine learning tasks. The main purpose is to remove irrelevant and redundant noise features in high-dimensional data space. The selected features will help to reduce the computation cost and improve the performance on real-world applications.

There are many research works on feature selection over the years. Generally, feature selection methods can be divided into three main categories [Guyon and Elisseeff, 2003]: wrapper method, filter method, and sparse coding based method (also known as embedded method). The most representative wrapper method is support vector machine recursive feature elimination (SVM-RFE) [Guyon *et al.*, 2002], but the computation cost is extremely high. Contrarily, filter method is very efficient such as F-statistic [Ding and Peng, 2003], ReliefF [Robnik-Šikonja and Kononenko, 2003], minimum redundancy maximum relevance (mRMR) [Peng *et al.*, 2005].

Recently, sparse coding based methods have been widely investigated, and applied to the study of feature selections. Least absolute shrinkage and selection operator (LASSO) [Tibshirani, 1996] is a regression based analysis method that incurs the sparsity on weights via ℓ_1-norm. ℓ_1-SVM [Zhu *et al.*, 2003] and hybrid huberized SVM (HHSVM) [Wang *et al.*, 2007] are introduced to further improve performance on two-class problem. LASSO can be derived from probabilistic selection on ridge regression [Ming *et al.*, 2019].

To solve multi-class problem, researchers search a subset of features shared by all the classes, also known as multi-task feature learning (MTFL). In this area, $\ell_{2,1}$-norm is the most widely used regularization developed in [Liu *et al.*, 2009; Nie *et al.*, 2010; Gui *et al.*, 2017]. In [Quattoni *et al.*, 2009], authors propose $\ell_{1,\infty}$-norm regularization, which shares same property of row-sparsity as $\ell_{2,1}$-norm. As compared to class-shared feature selection, exclusive lasso (eLASSO) [Zhou *et al.*, 2010; Campbell and Allen, 2017] proposes to capture the negative correlation among different classes via $\ell_{1,2}$-norm, which is first introduced in [Zhao *et al.*, 2009] called composite absolute penalties (CAP). In exclusive feature learning, discriminative features are selected for each class to provide certain flexibility. Based on this, Kong et al. of [Kong *et al.*, 2014] propose to solve the mix of ℓ_1-norm and $\ell_{1,2}$-norm, towards minimizing the feature correlation.

Motivated by previous works, in this paper, we introduce a novel regularization called "exclusive $\ell_{2,1}$", which is short for "$\ell_{2,1}$ with exclusive lasso". The exclusive $\ell_{2,1}$ regularization brings out joint sparsity at inter-group level and exclusive sparsity at intra-group level simultaneously. Thus, the proposed regularization can combine the advantages from different sparsity-induced terms, which not only removes irrelevant noise features (i.e. increase the robustness via $\ell_{2,1}$-norm) but also selects discriminative features for each class (i.e. provide the flexibility via $\ell_{1,2}$-norm).

The main contribution of this paper includes: (i) a novel "exclusive $\ell_{2,1}$" regularization is proposed to conduct robust flexible feature selection; (ii) we point out some interesting properties of $\|\mathbf{w}\|_1^2$ regularization as compared to $\|\mathbf{w}\|_1$ regularization; (iii) a sorting based explicit approach is introduced to directly solve the $\ell_{1,2}$-norm regularization; (iv) an efficient augmented Lagrange multipliers (ALM) based optimization algorithm is proposed to iteratively solve the "exclusive $\ell_{2,1}$" regularization in a row-wise fashion; (v) experimental results on twelve benchmark datasets demonstrate that the proposed regularization outperforms state-of-the-arts.

$$\mathbf{X}^T = \begin{bmatrix} 0.463 & 0.319 & -0.100 & 0.526 & 0.535 & 0.329 & 0.475 \\ 0.296 & 0.192 & 0.058 & -0.076 & 0.152 & 0.313 & -0.114 \\ 0.196 & 0.189 & 0.167 & -0.280 & 0.267 & -0.246 & 0.164 \\ 0.330 & 0.357 & 0.027 & -0.001 & 0.118 & 0.058 & 0.191 \\ 0.332 & 0.035 & -0.002 & 0.280 & 0.111 & -0.043 & 0.104 \\ -0.022 & -0.026 & 0.770 & 0.189 & 0.196 & -0.146 & -0.121 \\ -0.217 & 0.028 & 0.404 & 0.359 & 0.335 & -0.282 & -0.235 \\ 0.396 & 0.297 & 0.260 & 0.241 & 0.193 & 0.038 & 0.101 \end{bmatrix} \quad \mathbf{Y} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \tag{1}$$

$$\mathbf{W}_{21} = \begin{bmatrix} 0.764 & 0.587 & 0.378 \\ 0.097 & 0.033 & 0.082 \\ 0.054 & 0.531 & 1.003 \\ \textbf{-0.000} & \textbf{0.000} & \textbf{0.000} \\ 0.151 & 0.030 & 0.126 \\ \textbf{0.000} & \textbf{0.000} & \textbf{-0.000} \\ \textbf{0.000} & \textbf{-0.000} & \textbf{0.000} \end{bmatrix} \quad \mathbf{W}_{12} = \begin{bmatrix} 0.336 & 0.352 & \textbf{0.000} \\ 0.287 & \textbf{0.000} & 0.358 \\ \textbf{0.000} & 0.070 & 0.758 \\ -0.009 & 0.173 & \textbf{0.000} \\ 0.326 & \textbf{0.000} & 0.298 \\ \textbf{0.000} & \textbf{0.000} & -0.344 \\ 0.333 & \textbf{-0.000} & \textbf{0.000} \end{bmatrix} \quad \mathbf{W}_{\text{ex21}} = \begin{bmatrix} 0.192 & 0.114 & \textbf{0.000} \\ 0.132 & 0.014 & 0.090 \\ \textbf{0.000} & 0.041 & 0.358 \\ \textbf{0.000} & \textbf{0.000} & \textbf{0.000} \\ 0.133 & \textbf{-0.000} & 0.137 \\ 0.017 & 0.004 & -0.024 \\ \textbf{0.000} & \textbf{0.000} & \textbf{0.000} \end{bmatrix} \tag{2}$$

2 Notations and Definitions

Throughout this paper, scalars, vectors, and matrices are denoted as lower-case/capital letters, boldface lower-case letters, and boldface capital letters, respectively.

The i-th element of vector $\mathbf{w}$ is represented by w_i. Given a matrix $\mathbf{W} = (W_{ij}) \in \mathbb{R}^{d\times k}$, the i-th row is represented by $\mathbf{w}^i$ (i.e. $\mathbf{W} = [\mathbf{w}^1; \cdots; \mathbf{w}^d]$), and the j-th column is represented by $\mathbf{w}_j$ (i.e. $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_k]$). The Frobenius norm of $\mathbf{W}$ is $\|\mathbf{W}\|_F = \sqrt{\sum_{i=1}^d \sum_{j=1}^k W_{ij}^2}$. $\ell_{2,1}$-norm of $\mathbf{W}$ is $\|\mathbf{W}\|_{2,1} = \sum_{i=1}^d \left\|\mathbf{w}^i\right\|_2 = \sum_{i=1}^d \left(\sum_{j=1}^k W_{ij}^2\right)^{1/2}$. $\ell_{1,2}$-norm of $\mathbf{W}$ is $\|\mathbf{W}\|_{1,2}^2 = \sum_{i=1}^d \left\|\mathbf{w}^i\right\|_1^2 = \sum_{i=1}^d \left(\sum_{j=1}^k |W_{ij}|\right)^2$.

$\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$ represents n data points, where $\mathbf{x}_i \in \mathbb{R}^d$, and corresponding class labels are defined as $\mathbf{Y} = [\mathbf{y}^1; \cdots; \mathbf{y}^n] \in \mathbb{R}^{n\times k}$, where $\mathbf{y}^i \in \mathbb{R}^k$ is one-hot vector and $y_j^i = 1$ or $Y_{ij} = 1$ means i-th sample belonging to j-th class.

3 Exclusive $\ell_{2,1}$ Regularization

Generally, sparse coding based methods can be formulated as $\min_{\mathbf{W}} \{f(\mathbf{W}) + \lambda\Omega(\mathbf{W})\}$, where $f(\mathbf{W})$ is the loss function, $\Omega(\mathbf{W})$ is the regularization, and λ is the hyperparameter.

Our work is motivated from the following observations. The $\ell_{2,1}$ norm based feature selection (i.e. $f(\mathbf{W}) + \lambda\|\mathbf{W}\|_{2,1}$) incurs joint sparsity on rows. A selected non-zero row could still have some elements with small (in magnitude) numerical values. Suppose one of them is W_{ij}. This implies i-feature is not highly correlated with j-th class. Thus $\ell_{2,1}$ alone is too rigid for feature selection.

On the other end, exclusive lasso (i.e. $f(\mathbf{W}) + \lambda\|\mathbf{W}\|_{1,2}^2$) selects discriminative features for each class. Here, as λ increases, different elements in squared ℓ_1-norm of i-th row $\mathbf{w}^i$ are competing with each other to survive. Thus, at least one element in row $\mathbf{w}^i$ survive (remaining non-zero). The problem with exclusive lasso in this context is: all features/rows will be selected, because for each feature/row i, there will be some non-zero elements even at large regularization strength.

Towards resolving above main concerns for using $\ell_{2,1}$ regularization alone or using exclusive lasso alone, we propose to combine them together as a new regularization defined as $\Omega(\mathbf{W}) = \alpha\|\mathbf{W}\|_{2,1} + \beta\|\mathbf{W}\|_{1,2}^2$, which will be called "exclusive $\ell_{2,1}$" short for "$\ell_{2,1}$ with exclusive lasso". As a result, $\ell_{2,1}$-norm will increase the robustness to help $\ell_{1,2}$-norm, and $\ell_{1,2}$-norm will provide the flexibility to help $\ell_{2,1}$-norm.

3.1 An Illustration

The synthetic data $\mathbf{X}$, $\mathbf{Y}$ is given in Eq. (1), where $d = 7$, $n = 8$, $k = 3$. The loss function is $f(\mathbf{W}) = \|\mathbf{X}^T\mathbf{W} - \mathbf{Y}\|_F^2$. The learned matrices are given in Eq. (2), where the number of non-zero elements in $\mathbf{W}$ is enforced to 12 for each regularization. The difference is explained as follows:

(i) $\mathbf{W}_{21}$ ($\ell_{2,1}$): a feature can be selected by all the classes (e.g. *3rd* row is selected for *1st*, *2nd*, *3rd* classes), or can be discarded (e.g. *4th* row is a zero vector).

(ii) $\mathbf{W}_{12}$ (*exclusive lasso*): a feature can be selected by some classes (e.g. *5th* row is selected for *1st*, *3rd* classes; *6th* row is selected only for *3rd* class), but can not be discarded since the matrix has no zero rows.

(iii) $\mathbf{W}_{\text{ex21}}$ (*the proposed "exclusive $\ell_{2,1}$"*): a feature can be selected by all the classes (e.g. *2nd* row is selected for *1st*, *2nd*, *3rd* classes), or can be selected by some classes (e.g. *3rd* row is selected for *2nd*, *3rd* classes), or can be discarded (e.g. *4th* row is a zero vector).

4 Understanding the Exclusive Sparsity

4.1 Interesting Property of $\|\mathbf{w}\|_1^2$ Regularization

In this paper we use $\|\mathbf{w}\|_1^2$ regularization for flexible feature selection. Here, we point out some interesting properties of this regularization.

Consider $\|\mathbf{w}\|_1^2$ regularization first. We investigate the following simple proximal operator-type problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{w} - \mathbf{a}\|_2^2 + \lambda\|\mathbf{w}\|_1^2. \tag{3}$$

This is very similar to the standard ℓ_1-norm regularization problem

$$\min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{w} - \mathbf{a}\|_2^2 + \lambda\|\mathbf{w}\|_1, \tag{4}$$

which has been thoroughly studied in connection to lasso [Tibshirani, 1996].

There exists a widely held belief that optimization problems Eq. (3) and Eq. (4) behave very similarly and their solutions have identical sparsity pattern.

This belief comes from the following reasoning. Problem (3) is equivalent to

$$\min_{\mathbf{w}} \|\mathbf{w}-\mathbf{a}\|_2^2, \ \text{s.t.} \ \|\mathbf{w}\|_1^2 \leq t \tag{5}$$

for some parameter t. And problem (4) is equivalent to

$$\min_{\mathbf{w}} \|\mathbf{w}-\mathbf{a}\|_2^2, \ \text{s.t.} \ \|\mathbf{w}\|_1 \leq t \tag{6}$$

for some parameter t.

However, this widely held belief is incorrect.

Let $\mathbf{w}^*_{\ell_{12}}$ be the optimal solution for problem (3). Let $\mathbf{w}^*_{\ell_1}$ be the optimal solution for problem (4). We illustrate their significant differences in two simple cases.

Case 1 is a simple problem in 2-dim. $\mathbf{a}=(2,1)$. Optimal solutions are (computed using algorithm explained later[1])

$\lambda=0.1$, $\mathbf{w}^*_{\ell_1}=(1.95,0.95)$, $\mathbf{w}^*_{\ell_{12}}=(1.75,0.75)$.
$\lambda=1$, $\mathbf{w}^*_{\ell_1}=(1.5,0.5)$, $\mathbf{w}^*_{\ell_{12}}=(1,\mathbf{0})$.
$\lambda=10$, $\mathbf{w}^*_{\ell_1}=(\mathbf{0},\mathbf{0})$, $\mathbf{w}^*_{\ell_{12}}=(0.1818,\mathbf{0})$.
$\lambda=1000$, $\mathbf{w}^*_{\ell_1}=(\mathbf{0},\mathbf{0})$, $\mathbf{w}^*_{\ell_{12}}=(0.0020,\mathbf{0})$.

Clearly as λ increases above 1, $\mathbf{w}^*_{\ell_1}$ is all zeros, but $\mathbf{w}^*_{\ell_{12}}$ has non-zero component.

Case 2. Consider the dimension is one with $\mathbf{a}=1$. These problems can be solved analytically. The solutions are

$$\mathbf{w}^*_{\ell_1}=\left[1-\frac{\lambda}{2}\right]_+, \ \mathbf{w}^*_{\ell_{12}}=\frac{1}{1+\lambda}.$$

Clearly, when $\lambda>2$, $\mathbf{w}^*_{\ell_1}=0$, but $\mathbf{w}^*_{\ell_{12}}$ is never zero not matter how large λ is.

These two cases show that as λ increases to large values, $\mathbf{w}^*_{\ell_1}$ will become exact zero for all components, while $\mathbf{w}^*_{\ell_{12}}$ will become zero for $d-1$ components and one component approaches $\frac{1}{1+\lambda}$ asymptotically.

4.2 Solving $\ell_{1,2}$-Norm Regularization

In [Zhou *et al.*, 2010], authors illustrate the sparsity of $\ell_{1,2}$-norm from a projection point of view, then solve a min-max optimization problem. Kong et al of [Kong *et al.*, 2014] use a re-weight strategy to solve $\ell_{1,2}$-norm regularization.

However, both methods are inefficient in high-dimensional data space. Inspired by non-negative shrinkage thresholding operator [Cavazza *et al.*, 2018], we introduce a sorting based explicit approach to solve $\ell_{1,2}$-norm regularization. Here, we focus on its simplified formulation in Eq. (3), which then can be applied to solve multi-class problem in section 5.

Lemma 1. *The optimal solution $\mathbf{w}^*$ of Eq. (3) has the following property of its sign: for $i=1,\cdots,d$, (i) if $a_i=0$, $w^*_i=0$; (ii) if $a_i\neq 0$, sign$(w^*_i)=$ sign(a_i).*

Proof of Lemma 1. If $a_i=0$, $w^*_i=0$ can be easily verified. If $a_i\neq 0$, suppose $w^*_i=c$ and sign$(c)\neq$ sign(a_i). However, $w^*_i=-c$ gives the lower objective value, since $|c|=|-c|$ and $(c-a_i)^2>(-c-a_i)^2$. Thus, sign$(w^*_i)=$ sign(a_i). □

[1]For standard $\|\mathbf{w}\|_1$ regularization, Eq. (4) has the closed-form solution as $\mathbf{w}^*_{\ell_1}=\text{sign}(\mathbf{a})\odot[|\mathbf{a}|-\lambda/2]_+$. For $\|\mathbf{w}\|_1^2$ regularization, we propose a sorting based explicit approach (see Theorem 5) to solve Eq. (3), and the optimal solution $\mathbf{w}^*_{\ell_{12}}$ is given in Eq. (10).

Lemma 2. *The optimal solution $\mathbf{w}^*$ of Eq. (3) has the following property of its magnitude: for $i=1,\cdots,d$,*

$$|w^*_i|-|a_i|+\lambda\|\mathbf{w}^*\|_1=0, \ \textit{if} \ |w^*_i|>0, \tag{7}$$
$$-|a_i|+\lambda\|\mathbf{w}^*\|_1\xi_i=0, \ \xi_i\in[0,1], \ \textit{if} \ |w^*_i|=0, \tag{8}$$

where ξ_i is the subgradient of $f(x)=|x|, x\geq 0$ at $x=0$.

Proof of Lemma 2. Eq. (3) can be rewritten equivalently as

$$\min_{\mathbf{w}\in\mathbb{R}^d} J(\mathbf{w})=\sum_{i=1}^{d}\left(|w_i|-|a_i|\right)^2+\lambda\left(\sum_{i=1}^{d}|w_i|\right)^2, \tag{9}$$

since $[\text{sign}(w_i)]^2=[\text{sign}(a_i)]^2=1$, according to Lemma 1.

Taking derivative of $J(\mathbf{w})$ in Eq. (9) w.r.t $|w_i|$ and setting $\frac{\partial J(\mathbf{w})}{\partial |w_i|}=0$, we will have the same first-order optimality conditions defined in Eq. (7) and Eq. (8). □

Proposition 3. *As λ increases to large values, at least one element w_i in $\mathbf{w}$ will survive (i.e. $|w_i|>0$), given $\mathbf{a}\neq\mathbf{0}$. Otherwise, $\mathbf{w}=\mathbf{0}$ will lead to $\mathbf{a}=\mathbf{0}$ according to Eq. (8).*

Definition 4. *Given $\mathbf{a}=(a_1,\cdots,a_d)\in\mathbb{R}^d$, $\mathcal{S}$ denotes a d-dimensional vector with $\mathcal{S}_i\neq\mathcal{S}_j (i\neq j)$, $\bigcup_{i=1}^{d}\mathcal{S}_i=\{1,\cdots,d\}$, and each $\mathcal{S}_i$ represents the indexes of a descending order with respect to $\mathbf{a}$, such as $|a_{\mathcal{S}_1}|\geq|a_{\mathcal{S}_2}|\geq\cdots\geq|a_{\mathcal{S}_d}|$.*

Theorem 5. *The optimal solution of Eq. (3) is given by*

$$\mathbf{w}^*=\textit{sign}(\mathbf{a})\odot\left[|\mathbf{a}|-\frac{\lambda\tau}{1+\lambda\tau}\mu_\tau\right]_+, \tag{10}$$

where $\odot$ is the Hadamard product, i.e. $[\mathbf{x}\odot\mathbf{y}]_i=x_iy_i$, $[\cdot]_+=\max(\cdot,0)$, $\mu_\tau=\frac{1}{\tau}\sum_{i=1}^{\tau}|a_{\mathcal{S}_i}|$, and τ is the largest coordinate of $\mathcal{S}$ satisfying $|a_{\mathcal{S}_\tau}|-\frac{\lambda\tau}{1+\lambda\tau}\mu_\tau>0$.

Proof of Theorem 5. Suppose that $w^*_{\mathcal{S}_1}, w^*_{\mathcal{S}_2}, \cdots, w^*_{\mathcal{S}_\tau}$ are non-zeros. By adding Eq. (7) for $\mathcal{S}_1,\mathcal{S}_2,\cdots,\mathcal{S}_\tau$ (i.e. the first τ indexes saved in $\mathcal{S}$), we have

$$\sum_{i=1}^{\tau}|w^*_{\mathcal{S}_i}|-\sum_{i=1}^{\tau}|a_{\mathcal{S}_i}|+\lambda\tau\|\mathbf{w}^*\|_1=0, \tag{11}$$

which can be equivalently rewritten as $\|\mathbf{w}^*\|_1=\frac{\tau}{1+\lambda\tau}\mu_\tau$, where $\mu_\tau=\frac{1}{\tau}\sum_{i=1}^{\tau}|a_{\mathcal{S}_i}|$. Thus, Lemma 2 and Eq. (11) give the optimal solution $\mathbf{w}^*$ w.r.t its magnitude as follows

$$|w^*_{\mathcal{S}_i}|=|a_{\mathcal{S}_i}|-\frac{\lambda\tau}{1+\lambda\tau}\mu_\tau>0, \ \text{for} \ i=1,\cdots,\tau, \tag{12}$$
$$|w^*_{\mathcal{S}_i}|=0, \ \text{for} \ i=\tau+1,\cdots,d, \tag{13}$$

which is equivalent to the definition of $\mathbf{w}^*$ in Eq. (10), since $w^*_j=\text{sign}(w^*_j)|w^*_j|=\text{sign}(a_j)|w^*_j|$ for $j=\mathcal{S}_1,\cdots,\mathcal{S}_\tau$, and $w^*_j=0, |a_j|-\frac{\lambda\tau}{1+\lambda\tau}\mu_\tau<0$ for $j=\mathcal{S}_{\tau+1},\cdots,\mathcal{S}_d$. □

Theorem 6. *When τ is the largest coordinate of $\mathcal{S}$ satisfying $|a_{\mathcal{S}_\tau}|-\frac{\lambda\tau}{1+\lambda\tau}\mu_\tau>0$, the solution $\mathbf{w}^*$ defined in Eq. (10) achieves the global minimum of $J(\mathbf{w})$.*

Proof of Theorem 6. If $\tau=d$, we have $|w^*_{\mathcal{S}_i}|>0$ for $i=1,\cdots,d$, and each $w^*_{\mathcal{S}_i}$ given by Eq. (12) satisfies optimal condition Eq. (7). Thus, $\mathbf{w}^*$ is the global minimizer of $J(\mathbf{w})$.

If $\tau<d$, we have $|w^*_{\mathcal{S}_i}|>0$ for $i=1,\cdots,\tau$, and $|w^*_{\mathcal{S}_i}|=0$ for $i=\tau+1,\cdots,d$. Since τ is the largest coordinate of $\mathcal{S}$

Algorithm 1 Search the largest coordinate τ of $\mathcal{S}$.

Input: $\mathbf{a} \in \mathbb{R}^d$, λ, $\mathcal{S}$.
Output: τ, μ_τ.

1: **Initialize:** $\tau = d$, $\mu_\tau = \frac{1}{d}\sum_{i=1}^{d} |a_{\mathcal{S}_i}|$.
2: **while** $\tau > 1$ **and** $|a_{\mathcal{S}_\tau}| - \frac{\lambda\tau}{1+\lambda\tau}\mu_\tau < 0$ **do**
3: $\quad \mu_\tau = \frac{\tau}{\tau-1}\mu_\tau - \frac{1}{\tau-1}|a_{\mathcal{S}_\tau}|$.
4: $\quad \tau = \tau - 1$.
5: **end while**
6: **return** τ, μ_τ.

satisfying $|a_{\mathcal{S}_\tau}| - \frac{\lambda\tau}{1+\lambda\tau}\mu_\tau > 0$, for $(\tau{+}1)$-th coordinate of $\mathcal{S}$, we have $|a_{\mathcal{S}_{\tau+1}}| - \frac{\lambda(\tau+1)}{1+\lambda(\tau+1)}\mu_{\tau+1} < 0$, which can be rewritten equivalently as $|a_{\mathcal{S}_{\tau+1}}| - \frac{\lambda\tau}{1+\lambda\tau}\mu_\tau < 0$, i.e. $|a_{\mathcal{S}_{\tau+1}}| < \lambda\|\mathbf{w}^*\|_1$. This implies $w^*_{\mathcal{S}_{\tau+1}}$ satisfies Eq. (8). For $i = \tau{+}2, \cdots, d$, $w^*_{\mathcal{S}_i}$ satisfies Eq. (8), since $|a_{\mathcal{S}_i}| \le |a_{\mathcal{S}_{\tau+1}}| < \lambda\|\mathbf{w}^*\|_1$. Besides, $w^*_{\mathcal{S}_i}$ satisfies Eq. (7) for $i = 1, \cdots, \tau$. Thus, $\mathbf{w}^*$ is the global minimizer of $J(\mathbf{w})$, which completes the proof. □

Since $\mathbf{w}^*$ depends on τ, μ_τ, here we introduce an efficient algorithm (given in Algorithm 1) to search the largest coordinate τ of $\mathcal{S}$ satisfying $|a_{\mathcal{S}_\tau}| - \frac{\lambda\tau}{1+\lambda\tau}\mu_\tau > 0$ in linear time.

5 Optimization Algorithm

To select robust and flexible features, we are interested in the following optimization problem

$$\min_{\mathbf{W}} \; J_{\text{ex21}}(\mathbf{W}) = \left\|\mathbf{X}^T\mathbf{W} - \mathbf{Y}\right\|_F^2 + \alpha\|\mathbf{W}\|_{2,1} + \beta\|\mathbf{W}\|_{1,2}^2 \tag{14}$$

where the least square loss is penalized by the proposed "exclusive $\ell_{2,1}$" regularization, and α, β are hyperparameters.

First, we add an auxiliary variable $\mathbf{Z}$ to make the optimization separable between $\ell_{2,1}$-norm and $\ell_{1,2}$-norm. Thus, original problem (14) becomes

$$\begin{aligned} \min_{\mathbf{W},\mathbf{Z}} \quad & \left\|\mathbf{X}^T\mathbf{W} - \mathbf{Y}\right\|_F^2 + \alpha\|\mathbf{W}\|_{2,1} + \beta\|\mathbf{Z}\|_{1,2}^2 \\ \text{s.t.} \quad & \mathbf{Z} = \mathbf{W}. \end{aligned} \tag{15}$$

Then, augmented Lagrange multipliers (ALM) method is applied to enforce the constraint in problem (15) explicitly

$$\begin{aligned} \min_{\mathbf{W},\mathbf{Z}} \quad & \left\|\mathbf{X}^T\mathbf{W} - \mathbf{Y}\right\|_F^2 + \alpha\|\mathbf{W}\|_{2,1} + \beta\|\mathbf{Z}\|_{1,2}^2 \\ & + \langle\mathbf{\Lambda}, \mathbf{Z} - \mathbf{W}\rangle + \tfrac{\nu}{2}\left\|\mathbf{Z} - \mathbf{W}\right\|_F^2 \end{aligned} \tag{16}$$

where $\langle\cdot,\cdot\rangle$ is the inner product, i.e. $\langle\mathbf{A}, \mathbf{B}\rangle = \sum_{ij} A_{ij}B_{ij}$, $\mathbf{\Lambda}$ is the Lagrange multiplier, and ν is the penalty parameter. Problem (16) can be rewritten equivalently as

$$\begin{aligned} \min_{\mathbf{W},\mathbf{Z}} \quad & \left\|\mathbf{X}^T\mathbf{W} - \mathbf{Y}\right\|_F^2 + \alpha\|\mathbf{W}\|_{2,1} + \beta\|\mathbf{Z}\|_{1,2}^2 \\ & + \tfrac{\nu}{2}\left\|\mathbf{Z} - \mathbf{W} + \mathbf{\Lambda}/\nu\right\|_F^2. \end{aligned} \tag{17}$$

Thus, our task is to solve the variables $\mathbf{Z}$, $\mathbf{W}$ and update the parameters $\mathbf{\Lambda}$, ν.

5.1 Solving for Z

Firstly, we solve $\mathbf{Z}$ while fixing $\mathbf{W}$. Then, problem (17) w.r.t $\mathbf{Z}$ becomes

$$\mathbf{Z}_{t+1} = \arg\min_{\mathbf{Z}} \; \frac{\nu_t}{2}\left\|\mathbf{Z} - \mathbf{W}_t + \mathbf{\Lambda}_t/\nu_t\right\|_F^2 + \beta\|\mathbf{Z}\|_{1,2}^2. \tag{18}$$

Since the optimizations of each row $\mathbf{z}^i$ in $\mathbf{Z}$ are separable, we can minimize problem (18) in a row-wise fashion. Thus, the optimization of Eq. (18) w.r.t $\mathbf{z}^i$ becomes

$$\mathbf{z}^i_{t+1} = \arg\min_{\mathbf{z}^i} \; \frac{\nu_t}{2}\left\|\mathbf{z}^i - \mathbf{e}\right\|_2^2 + \beta\left\|\mathbf{z}^i\right\|_1^2, \tag{19}$$

where $i = 1, \cdots, d$ is the feature/row index, $\mathbf{e} = \mathbf{w}^i_t - \mathbf{\lambda}^i_t/\nu_t$, $\mathbf{w}^i_t$ is the i-th row of $\mathbf{W}_t$, and $\mathbf{\lambda}^i_t$ is the i-th row of $\mathbf{\Lambda}_t$.

Using Theorem 5, the optimal solution of Eq. (19) is

$$\mathbf{z}^i_{t+1} = \text{sign}(\mathbf{e}) \odot \left[|\mathbf{e}| - \frac{2\beta\tau}{\nu_t + 2\beta\tau}\mu_\tau\right]_+ \tag{20}$$

where τ, μ_τ are computed using Algorithm 1, given the input $(\mathbf{e}, 2\beta/\nu_t, \mathcal{S})$, and $\mathcal{S}$ is a k-dimensional vector representing the indexes of descending order $|e_{\mathcal{S}_1}| \ge |e_{\mathcal{S}_2}| \ge \cdots \ge |e_{\mathcal{S}_k}|$.

5.2 Solving for W

Secondly, we solve $\mathbf{W}$ while fixing $\mathbf{Z}$. Then, problem (17) w.r.t $\mathbf{W}$ becomes

$$\begin{aligned} \mathbf{W}_{t+1} = \arg\min_{\mathbf{W}} \; & \left\|\mathbf{X}^T\mathbf{W} - \mathbf{Y}\right\|_F^2 + \alpha\|\mathbf{W}\|_{2,1} \\ & + \tfrac{\nu_t}{2}\|\mathbf{Z}_{t+1} - \mathbf{W} + \mathbf{\Lambda}_t/\nu_t\|_F^2. \end{aligned} \tag{21}$$

Since $\ell_{2,1}$-norm is defined on each row $\mathbf{w}^i$ in $\mathbf{W}$, here we can solve $\mathbf{W}$ in the similar way as $\mathbf{Z}$.

To solve $\mathbf{W}$ in a row-wise fashion, we decompose the least square loss w.r.t $\mathbf{w}^i$ as follows

$$\begin{aligned} \left\|\mathbf{X}^T\mathbf{W} - \mathbf{Y}\right\|_F^2 &= \left\|\sum_{i=1}^{d}(\mathbf{x}^i)^T\mathbf{w}^i - \mathbf{Y}\right\|_F^2 \\ &= \left\|(\mathbf{x}^i)^T\mathbf{w}^i - \mathbf{Y}^{-i}\right\|_F^2 \\ &= a\|\mathbf{w}^i\|_2^2 - 2\mathbf{w}^i\mathbf{b}^T + c \end{aligned} \tag{22}$$

where $\mathbf{Y}^{-i} = \mathbf{Y} - \sum_{j\neq i}(\mathbf{x}^j)^T\mathbf{w}^j$, $a = \|\mathbf{x}^i\|_2^2$, $\mathbf{b} = \mathbf{x}^i\mathbf{Y}^{-i}$, and $c = \text{Tr}((\mathbf{Y}^{-i})^T\mathbf{Y}^{-i})$.

Thus, the optimization of Eq. (21) w.r.t $\mathbf{w}^i$ becomes

$$\mathbf{w}^i_{t+1} = \arg\min_{\mathbf{w}^i} \; \frac{2a + \nu_t}{2}\left\|\mathbf{w}^i - \mathbf{d}\right\|_2^2 + \alpha\left\|\mathbf{w}^i\right\|_2 \tag{23}$$

where $i = 1, \cdots, d$ is the feature/row index, $\mathbf{d} = \frac{1}{2a+\nu_t}(2\mathbf{b} + \nu_t\mathbf{z}^i_{t+1} + \mathbf{\lambda}^i_t)$, $\mathbf{z}^i_{t+1}$ is the i-th row of $\mathbf{Z}_{t+1}$, and $\mathbf{\lambda}^i_t$ is the i-th row of $\mathbf{\Lambda}_t$.

The optimal solution of Eq. (23) is given by[2]

$$\mathbf{w}^i_{t+1} = \left[1 - \frac{\alpha}{(2a + \nu_t)\|\mathbf{d}\|_2}\right]_+ \mathbf{d}, \tag{24}$$

where $[\cdot]_+ = \max(\cdot, 0)$.

5.3 Updating Parameters

Finally, we update parameters $\mathbf{\Lambda}$, ν at the end of t-th iteration as the following

$$\mathbf{\Lambda}_{t+1} = \mathbf{\Lambda}_t + \nu_t(\mathbf{Z}_{t+1} - \mathbf{W}_{t+1}), \tag{25}$$

$$\nu_{t+1} = \rho\nu_t \tag{26}$$

where $\rho > 1$ is a constant.

[2]For standard $\|\mathbf{w}\|_2$ regularization, the optimization problem $\min_{\mathbf{w}}\{\|\mathbf{w} - \mathbf{a}\|_2^2 + \lambda\|\mathbf{w}\|_2\}$ has the closed-form solution as $\mathbf{w}^*_{\ell_2} = \max(1 - \frac{\lambda}{2\|\mathbf{a}\|_2}, 0)\mathbf{a}$.

Algorithm 2 ALM based optimization algorithm for solving the "exclusive $\ell_{2,1}$" regularization in problem (14).

Input: data matrix $\mathbf{X} \in \mathbb{R}^{d\times n}$, class labels $\mathbf{Y} \in \mathbb{R}^{n\times k}$, hyperparameters α, β.
Output: weight matrix $\mathbf{W} \in \mathbb{R}^{d\times k}$.

Initialize: $t = 0$, $\nu_t = 1/\|\mathbf{X}\|_F$, $\rho = 1.1$, $\epsilon_1 = 1e{-}8$, $\epsilon_2 = 1e{-}5$, $\mathbf{\Lambda}_t = \mathbf{0}$, random initialization weights $\mathbf{W}_t$.
repeat
  for $i \in \{1, \cdots, d\}$ **do**
    Compute $\mathbf{e}$ via Eq. (19).
    Compute the descending order $\mathcal{S}$ of $|e_1|, \cdots, |e_k|$.
    Compute τ, μ_τ via Algorithm 1 given ($\mathbf{e}$,$2\beta/\nu_t$,$\mathcal{S}$).
    Compute $\mathbf{z}^i_{t+1}$ via Eq. (20).
  end for
  for $i \in \{1, \cdots, d\}$ **do**
    Compute $\mathbf{Y}^{-i}$, a, $\mathbf{b}$ via Eq. (22).
    Compute $\mathbf{d}$ via Eq. (23).
    Compute $\mathbf{w}^i_{t+1}$ via Eq. (24).
  end for
  Update $\mathbf{\Lambda}_{t+1}$ via Eq. (25).
  Update ν_{t+1} via Eq. (26).
  Set $t = t + 1$.
until convergence condition is satisfied:
  $\left|J_{\text{ex21}}(\mathbf{W}^{t+1}) - J_{\text{ex21}}(\mathbf{W}^t)\right| / J_{\text{ex21}}(\mathbf{W}^t) \le \epsilon_1$,
  $\left\|\mathbf{Z}^{t+1} - \mathbf{W}^{t+1}\right\|_\infty \le \epsilon_2$.
return the optimal solution: $\mathbf{W}^*$.

5.4 The Summary of Optimization Algorithm

The complete framework of the proposed augmented Lagrange multipliers (ALM) based optimization algorithm is summarized in Algorithm 2.

6 Experiments

6.1 Benchmark Datasets

Experiments on twelve benchmark datasets are conducted to evaluate the performance of feature selection methods on classification. Among those benchmarks, there are 4 image datasets: MNIST[3] [Lecun *et al.*, 1998], Yale[4], YaleB[5] , PIE [Sim *et al.*, 2002]; 1 spoken letter recognition dataset: ISOLET[6]; 5 bio-microarray datasets: Carcinomas [Yang *et al.*, 2006], Lung [Bhattacharjee *et al.*, 2001], Glioma [Nutt *et al.*, 2003], TOX[6], Tumor-14 [Ramaswamy *et al.*, 2001]; and 2 text datasets: CNAE-9 [Ciarelli and Oliveira, 2009], 20-Newsgroups[7]. The details of all the benchmark datasets is summarized in Table 1.

Dataset	k	n	d
MNIST	10	1000	784
Yale	15	165	1024
YaleB	38	2414	1024
PIE	10	210	2420
ISOLET	26	1560	617
Carcinomas	11	174	9182
Lung	5	203	3312
Glioma	4	50	4434
TOX	4	171	5748
Tumor-14	14	190	16063
CNAE-9	9	1080	856
20-Newsgroups	20	2000	5000

Table 1: The summary description of twelve benchmark datasets. k, n, d denote the number of classes, the number of data instances, the number of features for each dataset, respectively.

6.2 Evaluation Metrics

In subsequent experiments, the proposed exclusive $\ell_{2,1}$ regularization is compared to five state-of-the-arts, including three filter methods: *F-statistic* [Ding and Peng, 2003], *ReliefF* [Robnik-Šikonja and Kononenko, 2003], *minimum redundancy maximum relevance (mRMR)* [Peng *et al.*, 2005], and two sparse coding based methods: *multi-task feature learning via $\ell_{2,1}$-norm* ($\ell_{2,1}$) [Liu *et al.*, 2009; Nie *et al.*, 2010; Gui *et al.*, 2017], *exclusive Lasso* (*eLASSO*) [Zhou *et al.*, 2010; Campbell and Allen, 2017].

To evaluate the performance on classification, 5-fold cross-validation accuracy with SVM as classifier are computed on average. LIBSVM [Chang and Lin, 2011] is used as the practical implementation of SVM, where kernel is set as linear and parameter C is set as 1 for all the experiments.

When training different models, hyperparameters are adjusted to enforce the same level of sparsity on learned weight matrices. Then we select the top features with largest weights for each class. For testing, an SVM classifier is built for each class separately, by using the selected features. The final classification result is obtained via majority voting.

6.3 Analysis of the Results

Convergence Study

Convergence of our proposed ALM based optimization algorithm is shown in Fig. 1, where x-axis and y-axis denote the number of iterations and the objective value respectively.

We use the same hyperparameter setting, i.e. $\alpha = 1$, $\beta = 1$, for four benchmark datasets. As it can be seen in Fig. 1, the proposed optimization algorithm takes around 100$\sim$150 iterations to converge. This shows our ALM based optimization algorithm is efficient and converge fast in real applications.

Classification Results Comparison

Experimental results of our proposed exclusive $\ell_{2,1}$ regularization versus five state-of-the-arts are shown in Fig. 2, where x-axis denotes the number of selected features ranging from 10 to 80, and y-axis denotes the average of 5-fold cross-validation classification accuracy.

In general, sparse coding based methods ($\ell_{2,1}$, *eLASSO*, *Ours*) achieve better performances than filter methods (*F-Statistic*, *ReliefF*, *mRMR*). Among filter methods, *mRMR* has relatively higher classification accuracy, since it takes consideration of minimizing the correlation between features.

[3] In MNIST, 100 images are randomly selected out of each digit.
[4] http://vision.ucsd.edu/content/yale-face-database
[5] http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html
[6] http://featureselection.asu.edu/datasets.php
[7] http://qwone.com/~jason/20Newsgroups/ In 20-Newsgroups, 100 documents are randomly selected out of each newsgroup, and F-statistic method is used to prescreen 5,000 keywords.

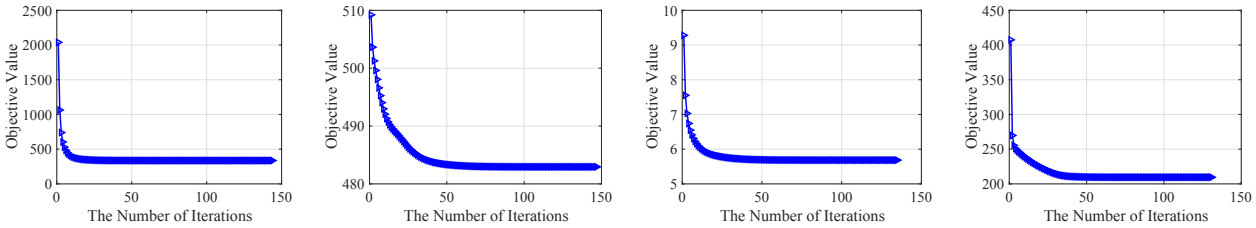

Figure 1: Convergence analysis of our proposed optimization algorithm on four benchmark datasets.

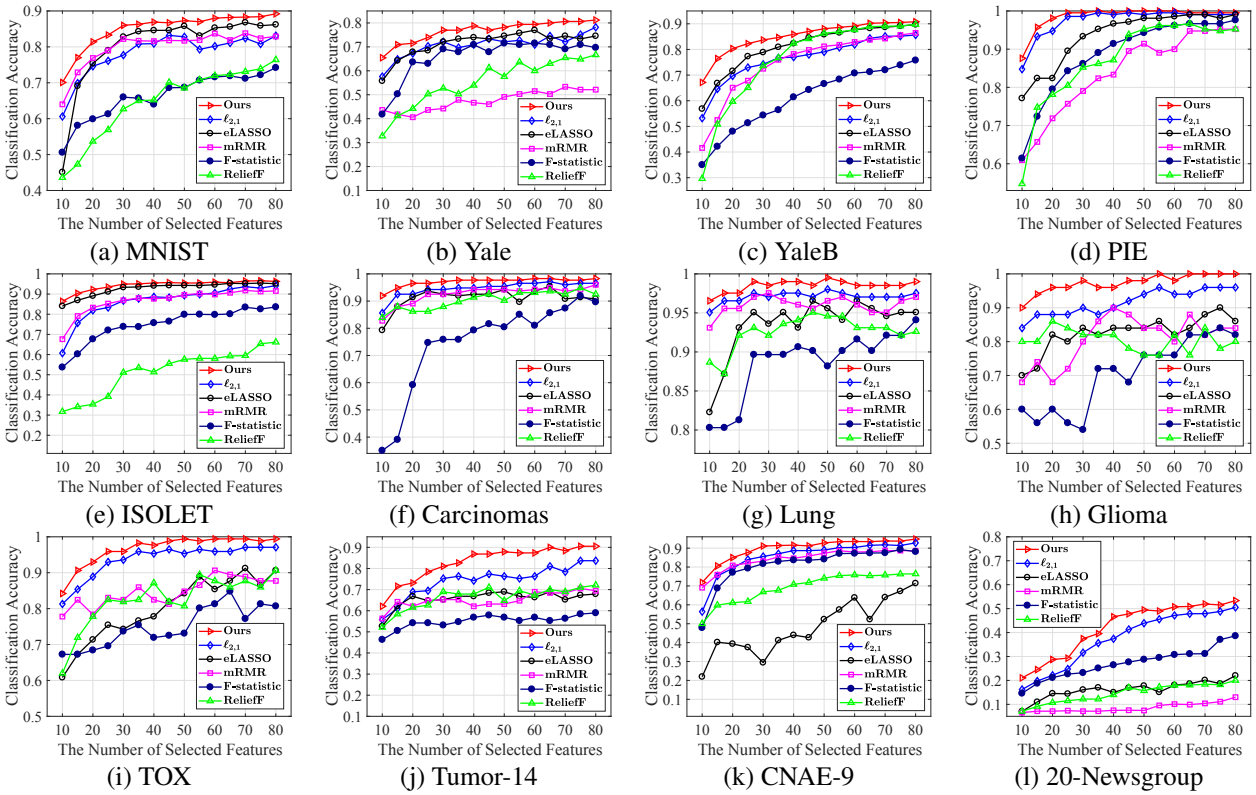

Figure 2: 5-fold cross-validation accuracy of our proposed feature selection method versus state-of-the-arts on twelve benchmark datasets.

eLASSO performs well in image and spoken letter recognition datasets. However, its performance has a great degradation in bio-microarray and text datasets, since $\ell_{1,2}$-norm cannot remove a large amount of irrelevant noise features in high-dimensional data space. $\ell_{2,1}$ has a very stable performance in all datasets via selecting class-shared features. In some cases, $\ell_{2,1}$ performs even close to *our method* around top 60$\sim$80 features. Overall, *our method* obtains the best classification result on twelve benchmark datasets. Additionally, in the small number of selected features setting, e.g. top 10$\sim$20, *our method* has an overwhelming advantage over other methods, with around 5%$\sim$10% improvement on accuracy.

7 Conclusion

In this paper, we introduce a novel "exclusive $\ell_{2,1}$" regularization for robust flexible feature selection. Besides, we point out some interesting property of $\|\mathbf{w}\|_1^2$ regularization, which can be solved directly by a sorting based explicit approach. Then, an efficient augmented Lagrange multipliers based optimization algorithm is proposed to iteratively solve the "exclusive $\ell_{2,1}$" regularization in a row-wise fashion. Extensive experiments validate the effectiveness of the proposed robust flexible feature selection, which outperforms state-of-the-arts on twelve benchmark datasets.

References

[Bhattacharjee *et al.*, 2001] Arindam Bhattacharjee, William G. Richards, Jane Staunton, Cheng Li, et al. Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences*, 98(24):13790–13795, 2001.

[Campbell and Allen, 2017] Frederick Campbell and Genevera I. Allen. Within group variable selection through the exclusive lasso. *Electronic Journal of Statistics*, 11(2):4220–4257, 2017.

[Cavazza *et al.*, 2018] Jacopo Cavazza, Pietro Morerio, Benjamin Haeffele, Connor Lane, et al. Dropout as a low-rank regularizer for matrix factorization. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, pages 435–444, 2018.

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[Ciarelli and Oliveira, 2009] Patrick M. Ciarelli and Elias Oliveira. Agglomeration and elimination of terms for dimensionality reduction. In *2009 Ninth International Conference on Intelligent Systems Design and Applications*, pages 547–552, Nov 2009.

[Ding and Peng, 2003] Chris Ding and Hanchuan Peng. Minimum redundancy feature selection from microarray gene expression data. In *Computational Systems Bioinformatics. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003*, pages 523–528, Aug 2003.

[Gui *et al.*, 2017] Jie Gui, Zhenan Sun, Shuiwang Ji, Dacheng Tao, and Tieniu Tan. Feature selection based on structured sparsity: A comprehensive study. *IEEE Transactions on Neural Networks and Learning Systems*, 28(7):1490–1507, July 2017.

[Guyon and Elisseeff, 2003] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, March 2003.

[Guyon *et al.*, 2002] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1):389–422, Jan 2002.

[Kong *et al.*, 2014] Deguang Kong, Ryohei Fujimaki, Ji Liu, Feiping Nie, and Chris Ding. Exclusive feature learning on arbitrary structures via $\ell_{1,2}$-norm. In *Advances in Neural Information Processing Systems 27*, pages 1655–1663, 2014.

[Lecun *et al.*, 1998] Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998.

[Liu *et al.*, 2009] Jun Liu, Shuiwang Ji, and Jieping Ye. Multi-task feature learning via efficient $\ell_{2,1}$-norm minimization. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 339–348, 2009.

[Ming *et al.*, 2019] Di Ming, Chris Ding, and Feiping Nie. A probabilistic derivation of lasso and l12-norm feature selections. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, 2019.

[Nie *et al.*, 2010] Feiping Nie, Heng Huang, Xiao Cai, and Chris Ding. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *Advances in Neural Information Processing Systems 23*, pages 1813–1821, 2010.

[Nutt *et al.*, 2003] Catherine L. Nutt, D. R. Mani, Rebecca A. Betensky, Pablo Tamayo, et al. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Research*, 63(7):1602–1607, 2003.

[Peng *et al.*, 2005] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, Aug 2005.

[Quattoni *et al.*, 2009] Ariadna Quattoni, Xavier Carreras, Michael Collins, and Trevor Darrell. An efficient projection for $\ell_{1,\infty}$ regularization. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 857–864, 2009.

[Ramaswamy *et al.*, 2001] Sridhar Ramaswamy, Pablo Tamayo, Ryan Rifkin, Sayan Mukherjee, et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences*, 98(26):15149–15154, 2001.

[Robnik-Šikonja and Kononenko, 2003] Marko Robnik-Šikonja and Igor Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine Learning*, 53(1):23–69, Oct 2003.

[Sim *et al.*, 2002] Terence Sim, Simon Baker, and Maan Bsat. The cmu pose, illumination, and expression (pie) database. In *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, pages 53–58, 2002.

[Tibshirani, 1996] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[Wang *et al.*, 2007] Li Wang, Ji Zhu, and Hui Zou. Hybrid huberized support vector machines for microarray classification. In *Proceedings of the 24th International Conference on Machine Learning*, pages 983–990, 2007.

[Yang *et al.*, 2006] Kun Yang, Zhipeng Cai, Jianzhong Li, and Guohui Lin. A stable gene selection in microarray data analysis. *BMC Bioinformatics*, 7(1):228, Apr 2006.

[Zhao *et al.*, 2009] Peng Zhao, Guilherme Rocha, and Bin Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A):3468–3497, 2009.

[Zhou *et al.*, 2010] Yang Zhou, Rong Jin, and Steven C.H. Hoi. Exclusive lasso for multi-task feature selection. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 988–995, 2010.

[Zhu *et al.*, 2003] Ji Zhu, Saharon Rosset, Trevor Hastie, and Rob Tibshirani. 1-norm support vector machines. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, pages 49–56, 2003.