A General Framework for Sparsity Regularized Feature Selection via Iteratively Reweighted Least Square Minimization

Hanyang Peng,[1] **Yong Fan**[2]

[1]National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences; University of Chinese Academy of Sciences, 100190, Beijing, P.R. China
[2]Department of Radiology, Perelman School of Medicine, University of Pennsylvania,
Philadelphia, PA, 19104, USA
hanyang.peng@nlpr.ia.ac.cn, yong.fan@ieee.org

Abstract

A variety of feature selection methods based on sparsity regularization have been developed with different loss functions and sparse regularization functions. Capitalizing on the existing sparsity regularized feature selection methods, we propose a general sparsity feature selection (GSR-FS) algorithm that optimizes a $\ell_{2,r}$-norm $(0 < r \leq 2)$ based loss function with a $\ell_{2,p}$-norm $(0 < p \leq 1)$ sparse regularization function. The $\ell_{2,r}$-norm $(0 < r \leq 2)$ based loss function brings flexibility to balance data-fitting and robustness to outliers by tuning its parameter, and the $\ell_{2,p}$-norm $(0 < p \leq 1)$ based regularization function is able to boost the sparsity for feature selection. To solve the optimization problem with multiple non-smooth and non-convex functions when $r, p < 1$, we develop an efficient solver under the general umbrella of Iterative Reweighted Least Square (IRLS) algorithms. Our algorithm has been proved to converge with a theoretical convergence order of at least $\min(2-r, 2-p)$. The experimental results have demonstrated that our method could achieve competitive feature selection performance on publicly available datasets compared with state-of-the-art feature selection methods, with reduced computational cost.

1. Introduction

Feature selection plays an important role in high-dimensional data analysis for selecting informative features and removing irrelevant or redundant ones (Cawley et al., 2006; Guyon & Elisseeff, 2003; Kira & Rendell, 1992; Lewis, 1992; Peng et al., 2005). Among existing feature selection methods, sparsity regularization based methods are appealing for their excellent performance (Argyriou & Evgeniou, 2007; Bradley & Mangasarian, 1998; Liu et al., 2009; Nie et al., 2010; Obozinski et al., 2006; Tibshirani, 1996; Wang et al., 2008; Xiang et al., 2012). In particular, ℓ_1-norm has been widely adopted in feature selection algorithms, such as Lasso (Tibshirani, 1996) and sparse SVM (Bradley & Mangasarian, 1998; Wang, et al., 2008). Built upon ℓ_1-norm based regularization models, $\ell_{2,1}$-norm has been used for feature selection in problems with multiple tasks or multiple classes (Argyriou & Evgeniou, 2007; Liu, et al., 2009; Nie, et al., 2010; Obozinski, et al., 2006; Xiang, et al., 2012). More recently, ℓ_p-norm and $\ell_{2,\mathrm{p}}$-norm $(0 < p < 1)$ based regularization models have gained increasing attention (Bolon-Canedo, et al., 2013; Chartrand & Staneva, 2008; Kong & Ding, 2014; Liu et al., 2007; Peng & Fan, 2016; Zhang et al., 2014) since they can yield sparser solutions than ℓ_1-norm and $\ell_{2,1}$-norm based models (Chartrand, 2007; Zeng et al., 2014).

Although a variety of sparsity regularization based feature selection methods with different sparse regularization functions have been developed, most of them adopt a least square loss function. The least square loss function has good data-fitting performance. However, it is sensitive to outliers. A robust feature selection (RFS) method with joint $\ell_{2,1}$-norm minimization on both the loss function and regularization function was proposed (Nie, et al., 2010; Xiang, et al., 2012) and has been extended (Wang & Chen, 2013) with joint $\ell_{2,p}$-norm $(0 < p \leq 1)$. However, it is not necessary to use the same norm for both the loss function and sparse regularization function. To make the sparsity regularized feature selection method more flexible, we propose a general sparsity regularized feature selection (GSR-FS) algorithm that optimizes a $\ell_{2,r}$-norm $(0 < r \leq 2)$ based loss function and a $\ell_{2,p}$-norm $(0 < p \leq 1)$ sparse regularization function. Particularly, the $\ell_{2,r}$-norm $(0 < r \leq 2)$ based loss function can balance the data fitting and robustness to outliners and the $\ell_{2,p}$-norm ($0 < p < 1$) based regularization function is able to boost the model sparsity for feature selection.

The optimization algorithms used in the existing sparsity regularized methods typically handle optimization problems with one non-smooth term[1] and are not suitable for our optimization problem with 2 non-smooth terms when $r, p \leq 1$. Iteratively reweighted least squares (IRLS) based methods have been widely used to solve sparse optimization problems in many fields (Candes et al., 2008; Chartrand & Yin, 2008; Gorodnitsky & Rao, 1997; Lu et al., 2014). However, the existing IRLS based algorithms only handle optimization problems with no more than one non-smooth function (Lu et al., 2015). To optimize our problem, we develop a novel algorithm based on IRLS with a convergence order of at least $\min(2-r, 2-p)$.

Our method has been validated based on 6 publicly available datasets and achieved competitive feature selection performance with respect to both classification accuracy and computational cost compared with 6 state-of-the-art feature selection algorithms, including Minimum-Redundancy Maximum-Relevance (mRMR) (Peng, et al., 2005), ReliefF (Kira & Rendell, 1992), Multi-Task Feature Selection (MTFS) (Argyriou & Evgeniou, 2007; Liu, et al., 2009; Obozinski, et al., 2006), Robust Feature Selection (RFS) (Nie, et al., 2010; Xiang, et al., 2012), an extended RFS(E-RFS) (Wang & Chen, 2013), and Rank One Update Algorithm (RK1U) (Zhang, et al., 2014).

2. A unified sparse feature selection algorithm

Given a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, its $\ell_{2,p}$-norm($p > 0$) is defined as:

$$\|\boldsymbol{A}\|_{2,p} = \left(\sum_{i=1}^{m}\left(\sum_{j=1}^{n}|\boldsymbol{a}_{i,j}|^{2}\right)^{\frac{p}{2}}\right)^{\frac{1}{p}} = \left(\sum_{i=1}^{m}(\|\boldsymbol{a}_i\|_2^2)^{\frac{p}{2}}\right)^{\frac{1}{p}}, \quad (1)$$

where $\|\boldsymbol{a}_i\|_2$ denotes ℓ_2-norm of the i-th row vector of $\boldsymbol{A}$.

Given m training samples $\boldsymbol{X} = \{\boldsymbol{x}^i\}_{i=1}^{m}$, $\boldsymbol{x}^i \in \mathbb{R}^n$, belonging to c(c $\geq$ 2) classes, and their class labels $\boldsymbol{Y} = \{\boldsymbol{f}^i\}_{i=1}^{m}$, $\boldsymbol{f}^i = [-1, \dots, 1, \dots, -1] \in \mathbb{R}^{\mathrm{c}}$ (the j-th element is 1 and others are -1 for the i^{th} data point belonging to the j^{th} class). In this paper, we adopt a $\ell_{2,r}$-norm ($0 < r \leq 2$) based loss function and a $\ell_{2,p}$-norm ($0 < p \leq 1$) based regularization function for feature selection, *i.e.*,

$$\min_{\boldsymbol{W}} \mathcal{J}(\boldsymbol{W}) = \|\boldsymbol{X}\boldsymbol{W} - \boldsymbol{Y}\|_{2,r}^{r} + \lambda\|\boldsymbol{W}\|_{2,p}^{p}, \quad (2)$$

where $\boldsymbol{W} \in \mathbb{R}^{\mathrm{n}\times\mathrm{c}}$ is the weight matrix to be learned, and non-zero rows of $\boldsymbol{W}$ indicate the selected features.

We choose $\|\boldsymbol{X}\boldsymbol{W} - \boldsymbol{Y}\|_{2,r}^{r}$ ($0 < r \leq 2$) instead of the traditional $\|\boldsymbol{X}\boldsymbol{W} - \boldsymbol{Y}\|_F^2$ as the loss function for following reasons. In general, a loss function with smaller r is more robust to outliers, whereas with larger r has better data-fitting performance. As indicated by the plots shown in Figure 1 (a), a small $r \leq 1$ for ℓ_r-norm could reduce the impact of an outlier on the loss function compared with a larger r. The impact of outliers in classification is also illustrated by 2D linear classification models with different settings of the ℓ_r-norm based loss function. In particular, as shown in Figure 1 (b), the classification models remain the same for different values of r if no outlier sample is present in the training data. However, the classification model with a ℓ_r-norm based loss function could change dramatically with different values of r if the training data contain outlier samples, and the classification models with a smaller r are more robust to outlier samples, as illustrated by Figure 1 (c). The regularization function $\|\boldsymbol{W}\|_{2,p}^{p}$ has a direct impact on the solution's sparsity, and small $p \leq 1$ is able to boost sparsity. The proposed method in Eqn. (2) is a generalization of existing sparsity regularization based feature selection methods, and many of them are special cases of the proposed method. Differences between our method and the existing methods under comparison are summarized in Table 1.

3. A novel IRLS method

The optimization problem of Eqn. (2) is a difficult problem with 2 non-convex, non-smooth functions when $0 < r < 1$ and $0 < p < 1$. To solve this problem, we propose an iterative algorithm, and at each iteration step we remodel the optimization problem of Eqn. (2) as a re-weighted least square minimization problem with analytical solutions.

When $\|\boldsymbol{x}_i\boldsymbol{W} - \boldsymbol{y}_i\| \neq 0$ and $\|\boldsymbol{w}_i\| \neq 0$ ($\boldsymbol{x}_i\boldsymbol{W} - \boldsymbol{y}_i$ and $\boldsymbol{w}_i$ are the i^{th} row vector of $\boldsymbol{X}\boldsymbol{W} - \boldsymbol{Y}$ and $\boldsymbol{W}$, respectively), the gradient of $\mathcal{J}(\boldsymbol{W})$ in Eqn. (2) with respect to $\boldsymbol{W}$ is

$$\frac{\partial\mathcal{J}(\boldsymbol{W})}{\partial\boldsymbol{W}} = 2\boldsymbol{X}^T\boldsymbol{S}_1(\boldsymbol{X}\boldsymbol{W} - \boldsymbol{Y}) + 2\lambda\,\boldsymbol{S}_2\boldsymbol{W}, \quad (3)$$

where $\boldsymbol{S}_1 \in \mathbb{R}^{m\times m}$ and $\boldsymbol{S}_2 \in \mathbb{R}^{n\times n}$ are diagonal matrices with diagonal elements $\boldsymbol{s}_1^{ii} = r/(2\|\boldsymbol{x}_i\boldsymbol{W} - \boldsymbol{y}_i\|_2^{2-r})$ and $\boldsymbol{s}_2^{ii} = p/2\|\boldsymbol{w}_i\|_2^{2-p}$. When $\|\boldsymbol{x}_i\boldsymbol{W} - \boldsymbol{y}_i\| = 0$ and $\|\boldsymbol{w}_i\| = 0$, we adopt the same strategy as (Nie, et al., 2010). Setting $\partial\mathcal{J}(\boldsymbol{W})/\partial\boldsymbol{W}$ to be 0, we have a solution of Eqn. (2), i.e.,

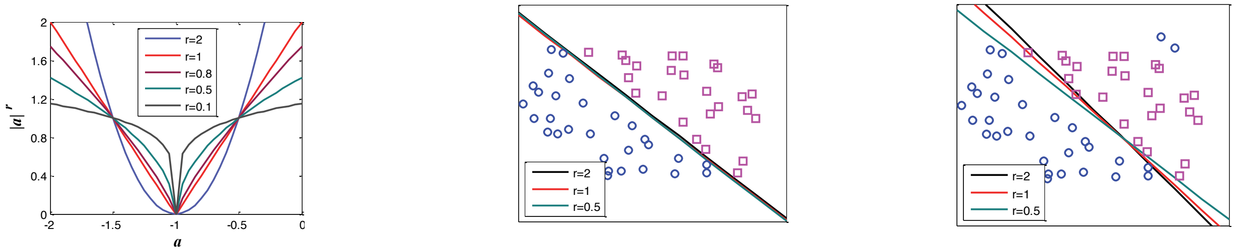
$$\boldsymbol{W} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{S}_1\boldsymbol{X} + \lambda\,\boldsymbol{S}_2)^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{S}_1\boldsymbol{Y}. \quad (4)$$

If $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ are fixed, we construct an auxiliary objective function $\mathcal{J}_1(\boldsymbol{W})$ to have the same gradient as $\mathcal{J}(\boldsymbol{W})$ in Eqn. (2),

$$\mathcal{J}_1(\boldsymbol{W}) = \|\boldsymbol{\Sigma}_1(\boldsymbol{X}\boldsymbol{W} - \boldsymbol{Y})\|_F^2 + \lambda\|\boldsymbol{\Sigma}_2\boldsymbol{W}\|_F^2, \quad (5)$$

where $\boldsymbol{\Sigma}_1 = (\boldsymbol{S}_1)^{1/2}$ and $\boldsymbol{\Sigma}_2 = (\boldsymbol{S}_2)^{1/2}$, and their i^{th} diagonal elements are $\boldsymbol{\varsigma}_1^{ii} = \sqrt{\mathrm{r}/2}/\|\boldsymbol{x}_i\boldsymbol{W} - \boldsymbol{y}_i\|_2^{1-r/2}$ and $\boldsymbol{\varsigma}_2^{ii} = \sqrt{p/2}/\|\boldsymbol{w}_i\|_2^{1-p/2}$, respectively.

[1] RFS and the extended RFS reformulated the optimization objective function with 2 non-smooth terms as a problem with one non-smooth term since both the loss function and the regularization function adopt the same $\ell_{2,p}$-norm($0 < p \leq 1$) (Nie, et al., 2010; Wang & Chen, 2013).



(a) (b) (c)

Figure 1： (a) $|a|^r$ with different r. (b) Optimal classification lines for ℓ_r-norm loss functions without any outlier. (c) Optimal classification lines for ℓ_r-norm loss functions with outliers

Table1: Comparisons of MTFS, RK1U, RFS, Extended RFS and our method

Method name	Objective function	r	p	Same r and p?
MTFS (Obozinski, et al., 2006)	$\|\boldsymbol{XW}-\boldsymbol{Y}\|_F^2+\lambda\|\boldsymbol{W}\|_{2,1}$	$r=2$	$p=1$	NO
RK1U (Zhang, et al., 2014)	$\|\boldsymbol{XW}-\boldsymbol{Y}\|_F^2+\lambda\|\boldsymbol{W}\|_{2,p}^p$	$r=2$	$0<p\le 1$	NO
RFS(Nie, et al., 2010)	$\|\boldsymbol{XW}-\boldsymbol{Y}\|_{2,1}+\lambda\|\boldsymbol{W}\|_{2,1}$	$r=1$	$p=1$	YES
E-RFS (Wang & Chen, 2013)	$\|\boldsymbol{XW}-\boldsymbol{Y}\|_{2,p}^p+\lambda\|\boldsymbol{W}\|_{2,p}^p$	$0<r\le 1$	$0<p\le 1$	YES
Ours	$\|\boldsymbol{XW}-\boldsymbol{Y}\|_{2,r}^r+\lambda\|\boldsymbol{W}\|_{2,p}^p$	$0<r\le 2$	$0<p\le 1$	NO

Then, we can obtain a solution by solving the re-weighted least square minimization problem, i.e.,

$$\min_{\boldsymbol{W}} \|\boldsymbol{\Sigma}_1(\boldsymbol{XW}-\boldsymbol{Y})\|_F^2+\lambda\|\boldsymbol{\Sigma}_2\boldsymbol{W}\|_F^2. \quad (6)$$

Since $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ (or $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$) are functions of $\boldsymbol{W}$, we use an iterative algorithm to compute the solution. At each iterative step, $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ are fixed first, then $\boldsymbol{W}$ is obtained according to Eqn. (4), and finally we update $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ based on $\boldsymbol{W}$, as summarized in Algorithm 1. Its convergence is proved in the following subsection.

Algorithm 1. A unified sparse feature selection algorithm

```
1.  Input: data points $\{x^i\}_{i=1}^m (x^i \in \mathbb{R}^n)$ and their corresponding label $\{y^i\}_{i=1}^m$; loss function norm order $r$; sparse regularization norm order $p$; regularization parameter $\lambda$; number of features $d$ to be selected.
2.  Construct $\boldsymbol{X}$ and $\boldsymbol{Y}$
3.  Set $k=1$ and initialize $\boldsymbol{S}_{1_0} \in \mathbb{R}^{m\times m}$ and $\boldsymbol{S}_{2_0} \in \mathbb{R}^{n\times n}$ as identity matrices
4.  Repeat
5.  Calculate $\boldsymbol{W}_k = \left(\boldsymbol{X}^T\boldsymbol{S}_{1_{k-1}}\boldsymbol{X}+\lambda\,\boldsymbol{S}_{2_{k-1}}\right)^{-1}\boldsymbol{X}^T\boldsymbol{S}_{1_{k-1}}\boldsymbol{Y}$
6.  Calculate $\boldsymbol{E}_k = \boldsymbol{XW_k}-\boldsymbol{Y}$
7.    Update $\boldsymbol{S}_{1_k}$, where $i$-th diagonal elements is $\frac{r/2}{\left\|\boldsymbol{e}_{i_k}\right\|_2^{2-r}}$
8.    Update $\boldsymbol{S}_{2_k}$,where $i$-th diagonal elements is $\frac{p/2}{\left\|\boldsymbol{w}_{i_k}\right\|_2^{2-p}}$
9.    Update $k=k+1$
10. Until convergence
11. Output: Sort all features according to $\|\boldsymbol{w}_i\|_2$ and select the top largest $d$ features.
```

4. Convergence analysis and convergence rate

The objective function $\mathcal{J}(\mathbf{W})$ monotonically decreases at every iteration step and Algorithm 1 finally converges.

Lemma 1. *Given any nonzero vectors* $\boldsymbol{a}$ *and* $\boldsymbol{b}$, *we have*

$$\|\boldsymbol{b}\|_2^{2\theta}-\theta\cdot\frac{\|\boldsymbol{b}\|_2^2}{\|\boldsymbol{a}\|_2^{2-2\theta}}\le(1-\theta)\|\boldsymbol{a}\|_2^{2\theta}, \quad (7)$$

where $0<\theta<1$ *and the equality holds if and only if* $\boldsymbol{a}=\boldsymbol{b}$.

Proof please see the supplementary material.

Based on Lemma 1, we have Lemma 2.

Lemma 2. *Given an optimization problem:*

$$\min_{\boldsymbol{Z}} f(\boldsymbol{Z})+\|\boldsymbol{\Sigma\Phi}(\boldsymbol{Z})\|_F^2, \quad s.t.\ \boldsymbol{Z}\in\mathcal{F}, \quad (8)$$

where $f(\boldsymbol{Z})$ *and* $\boldsymbol{\Phi}(\boldsymbol{Z})$ *are matrix functions of* $\boldsymbol{Z}$, $\mathcal{F}$ *is the feasible region*, and $\boldsymbol{\Sigma}$ *is a diagonal matrix with its* i^{th} *diagonal element equal to* $\sqrt{q/2}/\|\boldsymbol{\Phi}(\boldsymbol{Z}_0)_i\|_2^{1-q/2}$ *(*$\boldsymbol{Z}_0$ *is one element in* $\mathcal{F}$, $\boldsymbol{\Phi}(\boldsymbol{Z}_0)_i$ *is the* i^{th} *row vector of* $\boldsymbol{\Phi}(\boldsymbol{Z}_0)$ *and* $0<q\le 2$*). If* $\boldsymbol{Z}^*$ *is the optimal solution of the above optimization problem Eqn. (8), we have*

$$f(\boldsymbol{Z}^*)+\|\boldsymbol{\Phi}(\boldsymbol{Z}^*)\|_{2,q}^q\le f(\boldsymbol{Z}_0)+\|\boldsymbol{\Phi}(\boldsymbol{Z}_0)\|_{2,q}^q, \quad (9)$$

Proof. Since $\boldsymbol{Z}^*$ is the optimal solution of Eqn.(8), we have

$$f(\boldsymbol{Z}^*)+\|\ \boldsymbol{\Sigma\Phi}(\boldsymbol{Z}^*)\ \|_F^2\le f(\boldsymbol{Z}_0)+\|\ \boldsymbol{\Sigma\Phi}(\boldsymbol{Z}_0)\ \|_F^2. \quad (10)$$

Therefore

$$f(\boldsymbol{Z}^*)+\sum_i\frac{q}{2}\frac{\|\boldsymbol{\Phi}(\boldsymbol{Z}^*)_i\|_2^2}{\|\boldsymbol{\Phi}(\boldsymbol{Z}_0)_i\|_2^{2-q}} \le f(\boldsymbol{Z}_0)+\sum_i\frac{q}{2}\|\boldsymbol{\Phi}(\boldsymbol{Z}_0)_i\|_2^q. \quad (11)$$

When $0<q<2$, according to Lemma 1, we have

$$\sum_i\left(\|\boldsymbol{\Phi}(\boldsymbol{Z}^*)_i\|_2^q-\frac{q}{2}\frac{\|\boldsymbol{\Phi}(\boldsymbol{Z}^*)_i\|_2^2}{\|\boldsymbol{\Phi}(\boldsymbol{Z}_0)_i\|_2^{2-q}}\right) \le \sum_i\left(1-\frac{q}{2}\right)\|\boldsymbol{\Phi}(\boldsymbol{Z}_0)_i\|_2^q\ . \quad (12)$$

Summing Eqn. (11) and Eqn. (12), we obtain

$$f(\boldsymbol{Z}^*) + \sum_i \|\boldsymbol{\Phi}(\boldsymbol{Z}^*)_i\|_2^q \leq \ f(\boldsymbol{Z}_0) + \sum_i \|\boldsymbol{\Phi}(\boldsymbol{Z}_0)_i\|_2^q. \quad (13)$$

Finally, we obtain

$$f(\boldsymbol{Z}^*) + \|\boldsymbol{\Phi}(\boldsymbol{Z}^*)\|_{2,q}^q \ \leq f(\boldsymbol{Z}_0) + \|\boldsymbol{\Phi}(\boldsymbol{Z}_0)\|_{2,q}^q, \quad (14)$$

where the equality holds if and only if $\boldsymbol{\Phi}(\boldsymbol{Z}^*) = \boldsymbol{\Phi}(\boldsymbol{Z_0})$.

When $q = 2$, $\boldsymbol{\Sigma}$ becomes an identity matrix, the equality in Eqn. (14) still holds. □

Theorem 1. *The objective function of Eqn. (2) monotonically decreases at every iteration step, i.e.,*

$$\mathcal{J}(\boldsymbol{W}_k) \leq \mathcal{J}(\boldsymbol{W}_{k-1}), \quad (15)$$

and it converges to a limit point.

Proof. According to Eqn. (6), we have

$$\left\|\boldsymbol{\Sigma}_{1_{k-1}}(\boldsymbol{XW}_k - \boldsymbol{Y})\right\|_F^2 + \lambda\left\|\boldsymbol{\Sigma}_{2_{k-1}}\boldsymbol{W}_k\right\|_F^2$$
$$\leq \left\|\boldsymbol{\Sigma}_{1_{k-1}}(\boldsymbol{XW}_{k-1} - \boldsymbol{Y})\right\|_F^2 + \lambda\left\|\boldsymbol{\Sigma}_{2_{k-1}}\boldsymbol{W}_{k-1}\right\|_F^2, (16)$$

where the i th diagonal elements of $\boldsymbol{\Sigma}_{1_{k-1}}$ and $\boldsymbol{\Sigma}_{2_{k-1}}$ are $\boldsymbol{\varsigma}_{1_{k-1}}^{ii} = \sqrt{\mathrm{r}/2}/\|\boldsymbol{x}_i\boldsymbol{W}_{k-1} - \boldsymbol{y}_i\|_2^{1-r/2}$ and $\boldsymbol{\varsigma}_{2_{k-1}}^{ii} = \sqrt{p/2}/\|\boldsymbol{w}_{i_{k-1}}\|_2^{1-p/2}$, respectively.

According to Lemma 2, let $f(\boldsymbol{W}) = \left\|\boldsymbol{\Sigma}_{1_{k-1}}(\boldsymbol{XW} - \boldsymbol{Y})\right\|_F^2$ and $\boldsymbol{\Phi}(\boldsymbol{W}) = \lambda\boldsymbol{W}$, we have

$$\left\|\boldsymbol{\Sigma}_{1_{k-1}}(\boldsymbol{XW}_k - \boldsymbol{Y})\right\|_F^2 + \lambda\|\boldsymbol{W}_k\|_{2,p}^p$$
$$\leq \left\|\boldsymbol{\Sigma}_{1_{k-1}}(\boldsymbol{XW}_{k-1} - \boldsymbol{Y})\right\|_F^2 + \lambda\|\boldsymbol{W}_{k-1}\|_{2,p}^p. \quad (17)$$

Setting $f(\boldsymbol{W}) = \lambda\|\boldsymbol{W}\|_{2,p}^p$ and $\boldsymbol{\Phi}(\boldsymbol{W}) = \boldsymbol{XW} - \boldsymbol{Y}$, according to Lemma 2 we have

$$\|\boldsymbol{XW}_k - \boldsymbol{Y}\|_{2,r}^r + \lambda\|\boldsymbol{W}_k\|_{2,p}^p$$
$$\leq \|\boldsymbol{XW}_{k-1} - \boldsymbol{Y}\|_{2,r}^r + \lambda\|\boldsymbol{W}_{k-1}\|_{2,p}^p. \quad (18)$$

So, $\mathcal{J}(\boldsymbol{W}_\mathrm{k}) \leq \mathcal{J}(\boldsymbol{W}_{\mathrm{k}-1})$, and the equality holds if and only if $\boldsymbol{W}_\mathrm{k} = \boldsymbol{W}_{\mathrm{k}-1}$. Since the lower bound of $\mathcal{J}(\boldsymbol{W}_\mathrm{k})$ is limited, $\mathcal{J}(\boldsymbol{W}_\mathrm{k})$ converges to a limit point. □

Theorem 2. *Sequence* $\{\boldsymbol{W}_k\}$ *produced in Algorithm 1 converges, and the limit point is a stationary point of Eqn. (2).*

Proof can be found in the supplementary material. When $p = 1$ and $r \geq 1$, Eqn. (2) is a convex optimization problem, hence its solution obtained by Algorithm 1 is the globally optimal. When $0 < p < 1$ or $0 < r < 1$, it may converge to a local optimum.

The convergence rate of Algorithm 1 is derived as following. If $\boldsymbol{W}^*$ is the optimal solution of $\min_{\boldsymbol{W}}\|\boldsymbol{XW} - \boldsymbol{Y}\|_{2,r}^r + \lambda\|\boldsymbol{W}\|_{2,p}^p$, then the optimal residual $\boldsymbol{E}^* = \boldsymbol{XW}^* - \boldsymbol{Y}$. When $\boldsymbol{W}^*$ is sparse, the rows of $\boldsymbol{W}^*$ can be split into two parts: $\boldsymbol{W}_1^*$ and $\boldsymbol{W}_2^*$, where $\boldsymbol{W}_2^* = \boldsymbol{0}$ and $\boldsymbol{W}_1^*$ is the remainder. In the same way as partitioning $\boldsymbol{W}^*$ into $\boldsymbol{W}_1^*$ and $\boldsymbol{W}_2^*$, the rows of $\boldsymbol{W}_k$ and the columns of $\boldsymbol{X}$ are partitioned into $\boldsymbol{W}_{1_k}$and $\boldsymbol{W}_{2_k}$, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, respectively. Similarly, $\boldsymbol{E}^*$ can be split into $\boldsymbol{E}_1^*$ and $\boldsymbol{E}_2^*$, where $\boldsymbol{E}_2^* = \boldsymbol{0}$ and $\boldsymbol{E}_1^*$ is the remainder, and the rows of $\boldsymbol{E}_k$ ($\boldsymbol{E}_k = \boldsymbol{XW}_k - \boldsymbol{Y}$) and the columns of $\boldsymbol{I}$ ($\in \mathbb{R}^{\mathrm{m}\times\mathrm{m}}$) are partitioned into $\boldsymbol{E}_{1_k}$ and $\boldsymbol{E}_{2_k}$, $\boldsymbol{L}_1$ and $\boldsymbol{L}_2$, accordingly. We define $\boldsymbol{A}_1 = [\boldsymbol{X}_1,\ \boldsymbol{L}_1]$, $\boldsymbol{A}_2 = [\boldsymbol{X}_2,\ \boldsymbol{L}_2]$, $\boldsymbol{U}_{1_k}^T = [\boldsymbol{W}_{1_k}, -\boldsymbol{E}_{1_k}]^T$, $\boldsymbol{U}_{2_k}^T = [\boldsymbol{W}_{2_k}, -\boldsymbol{E}_{2_k}]^T$ and $(\boldsymbol{U}_1^*)^T = [\boldsymbol{W}_1^*, -\boldsymbol{E}_1^*]^T$. Then we have Lemma 3.

Lemma 3. *The following inequalities hold in the successive iteration steps of Algorithm 1.*

$$\|\boldsymbol{U}_{1_k} - \boldsymbol{U}_1^*\| \leq \frac{\frac{2}{p}\left\|\boldsymbol{W}_{2_{k-1}}\right\|_{2,2-p}^{2-p} + \frac{2\lambda}{r}\left\|\boldsymbol{E}_{2_{k-1}}\right\|_{2,2-r}^{2-r}}{\frac{2}{p}\left\|\boldsymbol{W}_{1_{k-1}}\right\|_{2,2-p}^{2-p} + \frac{2\lambda}{r}\left\|\boldsymbol{E}_{1_{k-1}}\right\|_{2,2-p}^{2-p}} \cdot$$
$$\sqrt{s_0}\,\|\boldsymbol{I} - \boldsymbol{B}^+\boldsymbol{B}\|^2\|\boldsymbol{A}_1^+\boldsymbol{A}_2\|^2\,\|\boldsymbol{U}_1^*\|, \quad (19)$$

$$\|\boldsymbol{U}_{2_k}\| \leq \frac{\frac{2}{p}\left\|\boldsymbol{W}_{2_{k-1}}\right\|_{2,2-p}^{2-p} + \frac{2\lambda}{r}\left\|\boldsymbol{E}_{2_{k-1}}\right\|_{2,2-r}^{2-r}}{\frac{2}{p}\left\|\boldsymbol{W}_{1_{k-1}}\right\|_{2,2-p}^{2-p} + \frac{2\lambda}{r}\left\|\boldsymbol{E}_{1_{k-1}}\right\|_{2,2-p}^{2-p}} \cdot$$
$$\sqrt{s_0}\,\|\boldsymbol{I} - \boldsymbol{B}^+\boldsymbol{B}\|\|\boldsymbol{A}_1^+\boldsymbol{A}_2\|\|\boldsymbol{U}_1^*\|, \quad (20)$$

where s_0 *is the number of columns of* $\boldsymbol{A}_1$, *and* $\boldsymbol{B} = (\boldsymbol{I} - \boldsymbol{A}_1\boldsymbol{A}_1^+)\boldsymbol{A}_2$.

Proof please see the supplementary material. According to Lemma 3, we can obtain the convergence order of Algorithm 1.

Theorem 3. *The convergence order of Algorithm 1 is at least* $min(2 - p, 2 - r)$.

Proof can be found in the supplementary material.

5. Experiments

5.1 Results based on a synthetic dataset

To investigate how the loss function's parameter r in our method affects the feature selection performance, we generated a synthetic dataset using following procedure. First, we generated n samples with features $\boldsymbol{X}_1,\ \boldsymbol{X}_2 \in \mathbb{R}^{n\times d_1}$, where elements of $\boldsymbol{X}_1,\ \boldsymbol{X}_2$ were randomly generated according to Gaussian distribution $\mathcal{N}(0,1)$. Second, we introduced redundant features $\boldsymbol{X}_3 = 0.5(\boldsymbol{X}_1, +\boldsymbol{X}_2) + \boldsymbol{\epsilon} \in \mathbb{R}^{n\times d_1}$ to the samples, where elements of $\boldsymbol{\epsilon}$ were randomly generated according to $\mathcal{N}(0,0.1)$. Third, irrelevant features $\boldsymbol{X}_4 \in \mathbb{R}^{n\times d_2}$ were injected into the samples, where elements of $\boldsymbol{X}_4$ were randomly generated according to uniform distribution $\mathcal{U}(-1,1)$. So, we obtained samples with features $\boldsymbol{X} = [\boldsymbol{X}_1,\ \boldsymbol{X}_2,\ \boldsymbol{X}_3,\ \boldsymbol{X}_4] \in \mathbb{R}^{n\times d}$, $d = 3d_1 + d_2$. Then, we generated multi-tasks labels for these samples as $\boldsymbol{Y}_0 = \boldsymbol{XW} + \boldsymbol{\varsigma} \in \mathbb{R}^{n\times c}$, where $\boldsymbol{W} = [\mathbf{W}_1;\ \mathbf{W}_2; \mathbf{0}] \in \mathbb{R}^{d\times c}$, $\boldsymbol{W}_1,\ \boldsymbol{W}_2 \in \mathbb{R}^{d_1\times c}$ and their elements were randomly generated according to uniform distribution $\mathcal{U}(0,1)$, and $\boldsymbol{\varsigma}$ was randomly generated according to $\mathcal{N}(0, 0.5)$. Finally, to simulate outlier samples, we randomly picked a subset of $\boldsymbol{Y}_0$ with a percentage of a, and reversed their positive or negative signs, yielding new labels $\boldsymbol{Y}_1$. Setting n =2000, d_1 =100, d_2 =700, c = 5, and a = 0, 0.01, and 0.1, we obtained 3 simulated data sets, each of them having 1000 features, among which 200 features were informative.

We evaluated our method using 10-fold cross validation based on the simulated dataset with respect to different $r = 0.5$, 1, 2 by setting $p = 1$. The performance was gauged with root mean square error (RMSE) between actual values and predicted values base on top 200 selected features. As shown in Figure 2, the least square loss function had the best data-fitting performance for samples without outliers. However, it was sensitive to outliers as reflected by relatively larger RMSE when the samples contained 1% and 10% outliers. Not surprisingly, the feature selection with $\ell_{2,r}$-norm based loss functions ($r \leq 1$) was robust to outliers but might sacrifice data-fitting accuracy. All these results indicated that the loss function's parameter should be adaptive to the problem under study.

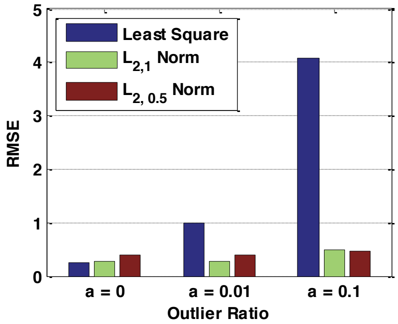


Figure 2: Average RMSE of 10-fold cross-validation for least square loss, $\ell_{2,1}$-norm loss, and $\ell_{2,0.5}$-norm loss, respectively.

5.2 Classification experiments on real-world datasets

We also evaluated our algorithm, referred to as general sparsity regularized feature selection (GSR-FS) based on 6 publicly available real-world datasets. In particular, 2 datasets were obtained from UCI, including ISOLET and SEMEION. Particularly, ISOLET is a speech recognition data set with 7797 samples in 26 classes, and each sample has 617 features. SEMEION contains 1593 handwritten images from ~80 persons, stretched in a rectangular box of 16 × 16. Three face image datasets were obtained from AR, ORL, and the frontal pose sub-dataset (09) of CMU-PIE. Particularly, AR has 1680 samples with 2000 features, ORL contains 400 samples with 92 × 112 pixels as features, and the CMU-PIE subset contains images of 64 persons with different illuminations. The 6th dataset contains confusable hand writing images 4 and 9, obtained from MNIST.

We compared our method with 4 sparsity regularized feature selection algorithms, including MTFS (Argyriou & Evgeniou, 2007; Liu, et al., 2009; Obozinski, et al., 2006), RFS (Nie, et al., 2010; Xiang, et al., 2012), an extended RFS (E-RFS) (Wang & Chen, 2013), and RK1U (Zhang, et al., 2014). We also compared our method with two filter feature selection methods, namely ReliefF (Kira & Rendell, 1992) and mRMR (Peng, et al., 2005).

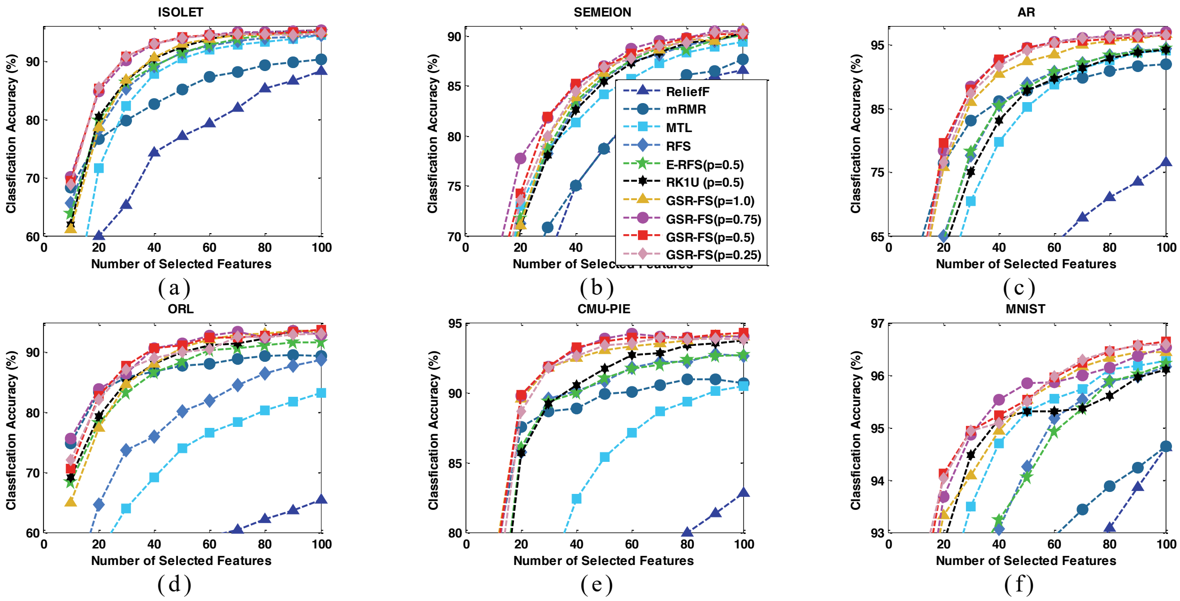
In our experiments, we first normalized all the features to have zero mean and unit standard deviation. Then, 10 trials were carried out on each dataset for feature selection. In each trial, each dataset was randomly spilt into training and testing subsets with a ratio of 6:4. Classification accuracy was used to evaluate the feature selection methods. Particularly, linear SVM (Chang & Lin, 2011) was used to build classifiers based on the selected features. The parameter C of linear SVM classifiers was tuned using a cross-validation strategy by searching a candidate set of [10^{-3}, 10^{-2}, 10^{-1}, 1, 10^{1}, 10^{2}]. The regularized parameter λ in our algorithm, MTFS, RFS and RK1U was tuned using the same cross-validation strategy by searching a candidate set of [10^{-3}, 10^{-2}, 10^{-1}, 1, 10^{1}, 10^{2}].

Our algorithm has 2 hyper parameters r and p. For evaluating the impact of p on the sparsity and directly comparing our method with MTFS, RFS and RK1U, we evaluated our algorithm by setting $p = 1$, 0.75, 0.5 and 0.25. Since the loss function with a smaller r is more robust to outliers but a larger r of the loss function may yield better data-fitting performance, in our experiments r was tuned by cross-validation with a candidate set of [0.5, 1, 2]. For the E-RFS and RK1U, $p = 0.5$ since it had better classification performance than other values (Wang & Chen, 2013; Zhang, et al., 2014).

Table 2 summarizes mean and standard deviation of the classification rates in 10 trails for classifiers built on the top 50 features. The average classification accuracy rates with top [10, 20, …, 100] features are shown in Figure 3. These results demonstrated that our method with different p

Table 2： Mean and standard deviation of the classification accuracy (%, mean±std) of Linear-SVM classifiers built on the top 50 features selected by different algorithms on different datasets.

Algorithm	ReliefF	mRMR	MTFS	RFS	Extended RFS (p=0.5)	RK1U (p=0.5)	GSR-FS (p=1.0)	GSR-FS (p=0.75)	GSR-FS (p=0.5)	GSR-FS (p=0.25)
ISOLET	77.02±0.82	85.10±0.62	90.40±0.68	91.38±0.73	91.36±0.56	92.50±0.79	92.98±0.44	93.92±0.46	**94.10±0.29**	93.93±0.40
SEMEION	78.70±1.39	78.75±1.82	84.11±1.12	85.82±1.25	85.89±1.61	85.45±1.01	86.24±1.45	**86.97±1.13**	86.87±0.89	86.58±1.17
AR	57.89±4.58	87.77±1.36	85.24±2.99	88.90±1.41	88.36±1.82	87.90±1.88	92.38±1.24	94.49±0.56	**94.51±0.96**	94.06±0.89
ORL	56.50±6.12	87.81±3.23	74.00±3.60	80.25±2.74	88.44±2.11	90.13±1.81	90.88±2.83	**91.50±2.27**	91.13±2.93	90.06±2.33
CMU-PIE	75.35±1.50	89.91±1.04	85.41±1.39	90.83±0.79	90.98±1.01	91.70±0.97	93.01±0.87	**93.85±0.83**	93.67±1.08	93.39±0.79
MNIST	90.62±0.35	92.55±0.19	95.30±0.21	94.25±0.38	94.04±0.55	95.31±0.20	95.53±0.21	**95.85±0.24**	95.53±0.18	95.49±0.35



(a) (b) (c)

(d) (e) (f)

Figure 3：Average classification accuracy of 10 trials for classifiers built on the selected features by different algorithms. The results shown were obtained on (a) ISOLET, (b)SEMEION, (c) AR, (d) ORL, (e) CMU-PIE, and (f) MNIST.

Table 3： Running Time (unit: second) by different Algorithms

	ReliefF	mRMR	RFS	Extended RFS(p=0.5)	RK1U (p=0.5)	GSR-FS (p=1.0)	GSR-FS (p=0.75)	GSR-FS (p=0.5)	GSR-FS (p=0.25)
ISOLET	319.48	52.85	2954.07	2100.20	6467.20	26.70	24.70	21.47	**17.01**
SEMEION	6.18	**4.70**	56.38	32.93	101.69	11.11	5.52	5.58	5.60
AR	37.13	24.67	67.44	48.92	2070.54	62.08	30.19	26.36	**20.49**
ORL	23.04	157.69	12.01	7.81	5750.64	12.11	11.97	4.60	**3.21**
CMU-PIE	89.35	62.24	66.47	67.31	4563.38	49.38	53.22	35.64	**19.43**
MNIST	676.21	48.83	9275.84	6199.82	1050.98	26.87	22.56	**18.81**	18.87

achieved overall the best classification accuracy on most of the datasets, especially when $p = 0.75, 0.5$. When $p = 0.5$, our method performed better than E-RFS and RK1U. Not surprisingly, the sparsity regularization methods had better performance than filter methods.

5.3 Computational cost

We also compared our algorithm with other methods with respect to their computation cost[2]. The convergence of all the sparse feature selection algorithms was determined based on the same criterion: the change of objective function value is less than 10^{-4} between 2 successive iteration steps with the regularized parameter λ=1. And the filter algorithms ran until the top 100 features were selected. We set $r = 2$ in our algorithm. We ran different methods on a desktop with an Intel i7-4470 CPU, 3.4GHz and 8G RAM. The computational costs of different algorithms are summarized in Table 3. As shown in Table 3, our algorithm was faster than other sparse feature selection algorithms on most of the datasets, and had similar costs as mRMR and ReliefF. Particularly, RK1U and our method achieved similar classification performance on several datasets under study, but the computational time of RK1U was more than 50 times longer than ours on average.

6. Discussions and Conclusion

We have presented a general framework for sparsity regularization based feature selection and a novel iterative reweighted least square minimization optimization algorithm. Several existing sparsity regularized feature selection methods could be treated as its special cases. The objective function of our method consists of a $\ell_{2,\mathrm{r}}$-norm $(0 < r \leq 2)$ based loss function and a $\ell_{2,\mathrm{p}}$-norm $(0 < p \leq 1)$ sparse regularization function, yielding an adaptive solution for handling outliers by turning its parameters. Such flexibility could improve feature selection performance as demonstrated by the experimental results. The novel IRLS algorithm is capable of solving problems with multiple non-smooth functions, and could find its applications in other fields.

[2] MTFS was implemented in C, and other algorithms were implemented in Matlab. So, we did not directly compare our algorithm with MTFS. However, RK1U was faster than MTFS (M. Zhang, et al, 2014), and our method was faster than RK1U.

We will extend our method to constrained optimization problems and investigate how to choose optimal parameters r and p in addition to the cross-validation strategy adopted in the present study.

Acknowledgments

This study was supported in part by National Key Basic Research and Development Program of China (2015CB856404), National Natural Science Foundation of China (81271514, 61473296), and NIH grants (EB022573, CA189523, and MH070365).

References

Argyriou, A. and Evgeniou, T., Multi-task feature learning. *Advances in Neural Information Processing Systems*, 2007.

Bradley, P. and Mangasarian, O., Feature selection via concave minimization and support vector machines. *International Conference on Machine Learning*, 1998.

Candes, E.J., et al., Enhancing Sparsity by Reweighted L_1 Minimization. *Journal of Fourier Analysis and Applications*. 14: 877-905, 2008.

Cawley, G.C., et al., Sparse multinomial ogistic regression via Bayesian L_1 regularisation. *Advances in Neural Information Processing Systems,*, 2006.

Chang, C.C. and Lin, C.J., LIBSVM: A Library for Support Vector Machines. *Acm Transactions on Intelligent Systems and Technology*. 2, 2011.

Chartrand, R., Exact reconstruction of sparse signals via nonconvex minimization. *Ieee Signal Processing Letters*. 14: 707-710, 2007.

Chartrand, R. and Staneva, V., Restricted isometry properties and nonconvex compressive sensing. *Inverse Problems*. 24, 2008.

Chartrand, R. and Yin, W.T., Iteratively reweighted algorithms for compressive sensing. *2008 Ieee International Conference on Acoustics, Speech and Signal Processing, Vols 1-12*: 3869-3872, 2008.

Gorodnitsky, I.F. and Rao, B.D., Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm. *Ieee Transactions on Signal Processing*. 45: 600-616, 1997.

Guyon, I. and Elisseeff, A., An introduction to variable and feature selection. *Journal of Machine Learning Research*, 2003.

Kira, K. and Rendell, L.A., A Practical Approach to Feature-Selection. *Machine Learning*: 249-256, 1992.

Kong, D. and Ding, C., Non-Convex Feature Learning via L_{p,inf} Operator. *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

Lewis, D.D., Feature-Selection and Feature-Extraction for Text Categorization. *Speech and Natural Language*: 212-217, 1992.

Liu, J., et al., Multi-Task Feature Learning Via Efficient L2,1-Norm Minimization. *Uncertainty in Artificial Intelligence*, 2009.

Liu, Y.F., et al., Support vector machines with adaptive L-q penalty. *Computational Statistics & Data Analysis*. 51: 6380-6394, 2007.

Lu, C., et al., Proximal iteratively reweighted algorithm with multiple splitting for nonconvex sparsity optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2014.

Lu, C.Y., et al., Smoothed Low Rank and Sparse Matrix Recovery by Iteratively Reweighted Least Squares Minimization. *Ieee Transactions on Image Processing*. 24, 2015.

Nie, F.P., et al., Efficient and Robust Feature Selection via Joint L2,1-Norms Minimization. *Advances in Neural Information Processing Systems*, 2010.

Obozinski, G., et al., Multi-task feature selection. *Technical report, Department of Statistics,University of California, Berkeley,*, 2006.

Peng, H.C., et al., Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *Ieee Transactions on Pattern Analysis and Machine Intelligence*. 27: 1226-1238, 2005.

Peng, H. and Fan, Y. Direct Sparsity Optimization Based Feature Selection for Multi-Class Classification. the 25th International Joint Conference on Artificial Intelligence, 2016.

Tibshirani, R., Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B-Methodological*. 58: 267-288, 1996.

Wang, L. and Chen, S., L_{2,p} Matrix Norm and Its Application in Feature Selection. *arXiv preprint arXiv:1303.3987*, 2013.

Wang, L., et al., Hybrid huberized support vector machines for microarray classification and gene selection. *Bioinformatics*. 24: 412-419, 2008.

Xiang, S.M., et al., Discriminative Least Squares Regression for Multiclass Classification and Feature Selection. *Ieee Transactions on Neural Networks and Learning Systems*. 23: 1738-1754, 2012.

Zeng, J.S., et al., L-1/2 Regularization: Convergence of Iterative Half Thresholding Algorithm. *Ieee Transactions on Signal Processing*. 62: 2317-2329, 2014.

Zhang, M., et al., Feature Selection at the Discrete Limit. *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.