A Probabilistic Derivation of LASSO and L12-Norm Feature Selections

Di Ming,[a] Chris Ding,[a] Feiping Nie[b]
[a]Department of Computer Science and Engineering, University of Texas at Arlington, TX 76019, USA
[b]Centre for OPTical Imagery Analysis and Learning, Northwestern Polytechnical University, Xian 710072, China
initialdiming@yahoo.com, chqding@uta.edu, feipingnie@gmail.com

Abstract

LASSO and $\ell_{2,1}$-norm based feature selection had achieved success in many application areas. In this paper, we first derive LASSO and $\ell_{1,2}$-norm feature selection from a probabilistic framework, which provides an independent point of view from the usual sparse coding point of view. From here, we further propose a feature selection approach based on the probability-derived $\ell_{1,2}$-norm. We point out some inflexibility in the standard feature selection that the feature selected for all different classes are enforced to be exactly the same using the widely used $\ell_{2,1}$-norm, which enforces the joint sparsity across all the data instances. Using the probability-derived $\ell_{1,2}$-norm feature selection, allowing certain flexibility that the selected features do not have to be exactly same for all classes, the resulting features lead to better classification on six benchmark datasets.

Introduction

Feature selection is one of important tasks of machine learning. Selecting useful set of features could improve many learning algorithms such as classification, regression, etc. In today's big data environment, many data has high-dimensions, e.g., biology datasets with around 10k features/genes (Bolón-Canedo et al. 2014) are commonplace. Selecting a subset of features reduces the data sizes and, more importantly, simplifies the interpretation of machine learning results, with many applications in gene-expression analysis (Dudoit, Fridlyand, and Speed 2002), proteomic biomarkers discovery (Saeys, Inza, and Larrañaga 2007), molecular cancer prediction (Gao and Church 2005).

Feature selection has been widely investigated in many applications with great assistance to practical performance. The main focus in the literature is on the supervised learning, which evaluates the relevance between features and class labels. The evaluation metric divides feature selection algorithms into three main categories (Guyon and Elisseeff 2003), which are filter, wrapper, embedded methods. Independent of any specific models, filter-type methods such as F-statistic (Ding and Peng 2003) and ReliefF (Robnik-Šikonja and Kononenko 2003) can quickly select features which are most correlated with class labels. However, redundant features are usually present in the subset of selected features via aforementioned algorithms. Thus, mRMR (Peng, Long, and Ding 2005) is proposed to maximize relevance and minimize redundancy simultaneously, which can effectively overcome the shortage of previous methods and further improve the practical performance. On the contrary, wrapper-type methods such as SVM-RFE (Guyon et al. 2002) are dependent on a specific classifier to iteratively search the best feature subset, but which has highly expensive computational cost and potential overfitting risk.

Recently, sparse coding based methods (also called embedded methods) become popular in study of feature selection. This approach combines the advantages of abovementioned two kinds of methods. The sparse model tries to find a compromise between loss and sparsity-induced regularization, e.g., the classic Lasso (Tibshirani 1996) using ℓ_1-norm constraint, also known as sparse coding in dictionary learning. To remove redundant noise features, ℓ_1-SVM (Zhu et al. 2003) is introduced to generate sparse solution for two-class feature selection. On the other hand, in multi-task setting, researchers (Argyriou, Evgeniou, and Pontil 2008), (Liu, Ji, and Ye 2009), (Nie et al. 2010), (Gui et al. 2017) focus on designing a collaborative model to select class-shared features via $\ell_{2,1}$-norm, which is first proposed in (Ding et al. 2006) as rotational invariant ℓ_1-norm for purpose of robust subspace factorization. Similarly, $\ell_{1,\infty}$-norm (Quattoni et al. 2009) is proposed to build a set of jointly sparse models, by means of ℓ_1-ball projection (Duchi et al. 2008). Besides, sparse coding based method is applied to other domains, such as sparse subspace learning (Gui et al. 2012), sparse representation based classification (Lu et al. 2013),etc.

A Probabilistic View of LASSO

The ℓ_1 based LASSO and the closely related $\ell_{1,2}$-norm feature selection are, in some sense, a prescription using sparse coding. In this paper, we show they can be derived from a probability framework, thus provides a strong probabilistic foundation.

In this paper, we propose to use the probability-derived $\ell_{1,2}$-norm feature selection. In this approach, features selected from different classes are not vigorously enforced to be exactly same.

However, most of popular feature selection methods aim at searching features across all the data instances with joint sparsity, which then enforces the selected features to be ex-

actly same for all classes.

Here, we argue that it is better to allowing selected features to have certain flexibility, not exactly same. In applications and real data, different classes could have its own characteristics, e.g., cars and cups have different features. Thus, using vigorously same set of features is not a natural way to pre-process/prescreen the data. Motivated by exclusive feature learning (Zhao, Rocha, and Yu 2009), (Zhou, Jin, and Hoi 2010), (Kong et al. 2014), (Campbell and Allen 2017), in this paper we propose a flexible feature selection method via $\ell_{1,2}$-norm regularization. In previous works, $\ell_{1,2}$-norm is used to either capture the negative correlation which creates competitions between features across all the classes, or eliminate strongly correlated features in two-class setting. Thus, our proposed method has a clear difference from them, that $\ell_{1,2}$-norm is enforced on features of each class to select a subset of features which are most correlated with each class separately. Using the flexible $\ell_{1,2}$-norm feature selection obtains features that generally perform better in many real datasets, including images and bio-microarray data.

The main contributions of this paper include: (1) a probabilistic derivation of LASSO and $\ell_{1,2}$-norm, and illustrating how $\ell_{1,2}$-norm is used to measure the importance of a subset of features for each class; (2) an effective algorithm with rigourous convergence analysis is proposed to compute/select the features using $\ell_{1,2}$-norm regularization, which is a parameter-free method and quickly converges; (3) experimental results on six benchmark datasets, including images and bio-microarray data, show that our proposed flexible feature selection method has an overwhelmed advantage over state-of-the-art algorithms.

Notations and Definitions

In this paper, lower-case letters refer to scalars, boldface lower-case letters refer to vectors, and boldface capital letters refer to matrices. n refers to the number of data instances. d refers to the number of features or data dimensions. k refers to the number of classes. The i-th element of vector $\mathbf{w}$ is presented by w_i. The i-th row and j-th column of matrix $\mathbf{W} = (W_{i,j})$ are denoted as $\mathbf{w}^i$ and $\mathbf{w}_j$, respectively. Given a matrix $\mathbf{W} \in \mathbb{R}^{d\times k}$, the Frobenius-norm of matrix $\mathbf{W}$ is $\|\mathbf{W}\|_F = \sqrt{\sum_{i=1}^{d}\sum_{j=1}^{k} W_{ij}^2}$. In general, the $\ell_{p,q}$ norm of W is defined as $\|W\|_{p,q} = \left(\sum_{j=1}^{k}\left(\sum_{i=1}^{d}|A_{ij}|^p\right)^{q/p}\right)^{1/q}$, with the computational mathematics convention that ℓ_p norm on the first (fastest index) i and ℓ_q norm on the second fast index j. With this convention, the $\ell_{2,1}$-norm based feature selection uses $\|W^T\|_{2,1}$ regularization; the $\ell_{1,2}$-norm based feature selection uses $\|W\|_{1,2}$ regularization; the exclusive LASSO uses $\|W^T\|_{1,2}$ regularization.

A Probabilistic Derivation of LASSO and $\ell_{1,2}$-Norm Feature Selection

First, the variables of the feature selection model are defined as follows. Training data of n labeled feature vectors are denoted as $\mathbf{X} \in \mathbb{R}^{d\times n} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$, where $\mathbf{x}_i \in \mathbb{R}^d$. The corresponding class labels are denoted as $\mathbf{Y} \in \mathbb{R}^{n\times k} = (\mathbf{y}_1, \cdots, \mathbf{y}_k)$, where $\mathbf{y}^i \in \mathbb{R}^k$ represents the class label for $\mathbf{x}_i$ using one-hot vector, i.e., $Y_{ij} = 1$ if $\mathbf{x}_i$ belongs to j-th class, $Y_{ij} = 0$ otherwise. Weights to be learnt are denoted as $\mathbf{W} \in \mathbb{R}^{d\times k} = (\mathbf{w}_1, \cdots, \mathbf{w}_k)$, where $\mathbf{w}_i \in \mathbb{R}^d$ represents the coefficient correlated with i-th class's features.

Our starting point is the LASSO type feature selection formalism using $\ell_{1,2}$-norm:

$$\min_{\widehat{\mathbf{W}}} \quad \left\|\mathbf{X}^T\widehat{\mathbf{W}} - \mathbf{Y}\right\|_F^2 + \lambda\left\|\widehat{\mathbf{W}}\right\|_{1,2}^2 \tag{1}$$

Here we use $\widehat{\mathbf{W}}$ to distinguish it from the following presentation.

We now present a new derivation of Eq.(1) from a probabilistic selection based on ridge regression. We first expand Eq.(1) on $\widehat{\mathbf{W}} = (\widehat{\mathbf{w}}_1, \cdots, \widehat{\mathbf{w}}_k)$

$$\min_{\widehat{\mathbf{W}}} \quad \sum_{j=1}^{k}\left(\left\|\mathbf{X}^T\widehat{\mathbf{w}}_j - \mathbf{y}_j\right\|_2^2 + \lambda\|\widehat{\mathbf{w}}_j\|_1^2\right) \tag{2}$$

Now, we introduce a selection probability vector $\boldsymbol{\theta}_j$ for class j and propose a selection formalism

$$\begin{aligned}\min_{\mathbf{W},\boldsymbol{\Theta}} \quad & \sum_{j=1}^{k}\left(\left\|\mathbf{X}^T(\boldsymbol{\theta}_j^{\frac{1}{2}} \odot \mathbf{w}_j) - \mathbf{y}_j\right\|_2^2 + \lambda\|\mathbf{w}_j\|_2^2\right) \\ s.t. \quad & \boldsymbol{\theta}_j \geq 0, \mathbf{1}^T\boldsymbol{\theta}_j = 1, \quad j = 1, \cdots, k,\end{aligned} \tag{3}$$

where $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_k) \in \mathbb{R}^{d\times k}$, $\mathbf{1}$ is a vector of all 1's with appropriate size, and $\odot$ is a element-wise hadamard product, i.e., $(\mathbf{a} \odot \mathbf{b})_i = a_i b_i$.

Both the optimization problems of Eq.(2) and Eq.(3) are convex and have unique optimal solutions.

Theorem 1. *Optimization problems Eq.(2) and Eq.(3) are equivalent. (A) Once the optimal solution $\{\widehat{\mathbf{w}}_j^*\}$ for Eq.(2) is obtained, the optimal solution for Eq.(3) is given by*

$$\Theta_{ij}^* = \frac{\left|\widehat{W}_{ij}^*\right|}{\left\|\widehat{\mathbf{w}}_j^*\right\|_1}, \mathbf{w}_j^* = (\boldsymbol{\theta}_j^*)^{-\frac{1}{2}} \odot \widehat{\mathbf{w}}_j^* \tag{4}$$

where $i = 1, \cdots, d$ is the feature/dimension index. (B) On the other direction, once $\{\boldsymbol{\theta}_j^*, \mathbf{w}_j^*\}$ for Eq.(3) is obtained, the optimal solution for Eq.(2) are given by $\widehat{\mathbf{w}}_j^* = (\boldsymbol{\theta}_j^*)^{\frac{1}{2}} \odot \mathbf{w}_j^*$.

The proof of this theorem is given in Lemma 2.

LASSO, Nonnegative Garrote and Selective Ridge Regression

Here we discuss LASSO, selective ridge regression and nonnegative Garrote of Breiman (Breiman 1995).

In optimization problems Eq.(2) and Eq.(3), different classes are in fact decoupled. Thus we can optimize them one class at a time. Thus the optimization of Eq.(2) is, in essence, equivalent to the following form (we ignore the index j)

$$\min_{\widehat{\mathbf{w}}} \quad \left\|\mathbf{X}^T\widehat{\mathbf{w}} - \mathbf{y}\right\|_2^2 + \lambda\|\widehat{\mathbf{w}}\|_1^2, \tag{5}$$

This is LASSO, except the ℓ_1 term is squared which does not affect the sparsity of $\widehat{\mathbf{w}}$.

Optimization problem Eq.(3) is in essence what we would call the "selective" ridge regression

$$\begin{array}{ll}\min\limits_{\mathbf{w},\boldsymbol{\theta}} & \left\|\mathbf{X}^T(\boldsymbol{\theta}^{\frac{1}{2}}\odot\mathbf{w})-\mathbf{y}\right\|_2^2+\lambda\|\mathbf{w}\|_2^2 \\ s.t. & \boldsymbol{\theta}\geq 0, \mathbf{1}^T\boldsymbol{\theta}=1.\end{array} \tag{6}$$

This formulation in some sense is close to the nonnegative Garrote:

$$\begin{array}{ll}\min\limits_{\boldsymbol{\theta}} & \left\|\mathbf{X}^T(\boldsymbol{\theta}\odot\mathbf{w}^0)-\mathbf{y}\right\|_2^2 \\ s.t. & \boldsymbol{\theta}\geq 0, \mathbf{1}^T\boldsymbol{\theta}\leq h,\end{array} \tag{7}$$

where $\mathbf{w}^0=\arg\min_{\mathbf{w}}\|\mathbf{X}^T\mathbf{w}-\mathbf{y}\|_2^2$ is the solution to ordinary least squares estimation, and $h\leq 1$ is a constant. In both Eq.(6) and Eq.(7), the selection vector $\boldsymbol{\theta}$ has similar sparsity pattern of the LASSO.

Lemma 2. *Optimization problems Eq.(5) and Eq.(6) are equivalent.*

Proof Starting from Eq.(6), we introduce a new variable $\widehat{\mathbf{w}}=\boldsymbol{\theta}^{\frac{1}{2}}\odot\mathbf{w}$, then $\mathbf{w}=\boldsymbol{\theta}^{-\frac{1}{2}}\odot\widehat{\mathbf{w}}$. Thus, the optimization problem (6) is transformed into

$$\begin{array}{ll}\min\limits_{\widehat{\mathbf{w}},\boldsymbol{\theta}} & \left\|\mathbf{X}^T\widehat{\mathbf{w}}-\mathbf{y}\right\|_2^2+\lambda\sum\limits_{i=1}^{d}\left(\frac{\widehat{w}_i^2}{\theta_i}\right) \\ s.t. & \boldsymbol{\theta}\geq 0, \mathbf{1}^T\boldsymbol{\theta}=1.\end{array} \tag{8}$$

When $\widehat{\mathbf{w}}$ is fixed, solving problem (8) with respect to $\boldsymbol{\theta}$ is

$$\begin{array}{ll}\min\limits_{\boldsymbol{\theta}} & \sum\limits_{i=1}^{d}\left(\frac{\widehat{w}_i^2}{\theta_i}\right) \\ s.t. & \boldsymbol{\theta}>0, \mathbf{1}^T\boldsymbol{\theta}=1,\end{array} \tag{9}$$

which can be solved using Lagrangian multiplier. The optimal solution of $\boldsymbol{\theta}$ is computed as

$$\theta_i=\frac{|\widehat{w}_i|}{\sum_{i'=1}^{d}|\widehat{w}_{i'}|}=\frac{|\widehat{w}_i|}{\|\widehat{\mathbf{w}}\|_1}, \tag{10}$$

where $i=1,\cdots,d$ is the feature/dimension index. With the result of Eq.(10), the objective of Eq.(9) becomes $\sum_{i=1}^{d}\left(\frac{\widehat{w}_i^2}{\theta_i}\right)=\|\widehat{\mathbf{w}}\|_1^2$. Problem Eq.(8) is transformed into a problem identical to problem (5). □

Using Lemma 2, Theorem 1 can be easily proved. Eq.(4) in Theorem 1 comes from Eq.(10).

The above relationships among LASSO, nonnegative Garrote and selective ridge regression provides a probability interpretation of LASSO. To gain further insights, we can easily prove the following

Theorem 3. *The following optimization*

$$\begin{array}{ll}\min\limits_{\mathbf{w},\boldsymbol{\theta}} & \left\|\mathbf{X}^T(\boldsymbol{\theta}^{\frac{1}{2}}\odot\mathbf{w})-\mathbf{y}\right\|_2^2+\lambda\|\mathbf{w}\|_2^2, \\ s.t. & \boldsymbol{\theta}\geq 0, \mathbf{1}^T\boldsymbol{\theta}\leq h.\end{array} \tag{11}$$

where $0<h\leq 1$ is a constant, is identical to

$$\min_{\widehat{\mathbf{w}}} \quad \left\|\mathbf{X}^T\widehat{\mathbf{w}}-\mathbf{y}\right\|_2^2+\tfrac{\lambda}{h}\|\widehat{\mathbf{w}}\|_1^2\,. \tag{12}$$

Once the optimal solution $\widehat{\mathbf{w}}^*$ to problem (12) is found, optimal solution to problem (11) is given by

$$\theta_i^*=\frac{h\,|\widehat{w}_i^*|}{\|\widehat{\mathbf{w}}^*\|_1},\ w_i^*=(\theta_i^*)^{-\frac{1}{2}}\widehat{w}_i^*, \tag{13}$$

where $i=1,\cdots,d$ is the feature/dimension index. When $\widehat{w}_i^*=0, w_i^*=0$. Note that $\mathbf{1}^T\boldsymbol{\theta}^*=h$.

Theorem 3 implies that when we wish to select less number of features using a smaller $h<1$, we need to increase the regularization, see Eq.(12).

A Ranking Method

Strictly speaking, in order to use LASSO to select m features, one has to set λ appropriately to a value λ_m so that exactly m features in optimal solution $\widehat{\mathbf{w}}^*$ are nonzero. The less number of features we desire, the stronger regularization we need to apply — consistent with Theorem 3. We will call this method as strict λ_m method. This strict λ_m method is computationally expensive.

The probability derivation of LASSO of Theorems 1 and 3, as the selection vector $\boldsymbol{\theta}$ from the selective ridge regression, naturally provides a ranking scheme of the features. Once we computed the solution to the LASSO problem Eq.(5), from Eq.(10), the importance of feature i is proportional to $|\widehat{w}_i^*|$. In other words, we rank the importance of features according to $(|\widehat{w}_1^*|,\cdots,|\widehat{w}_d^*|)$, and select the top m ranked features from the sorted order. This ranking selection method is fast in practice.

These two selection methods usually lead to different selected feature sets. In our experiments and from reading many research publications by other researchers, the feature set selected from ranking method generally performs better than the feature set selected via the strict λ_m method. A simple explanation is that the strict λ_m method usually leads to a larger λ_m as compared to the λ used in the ranking method. The larger λ_m used in LASSO usually penalized the regression too severely and thus altered the structural relation among the features. In the ranking method, a smaller λ is used which does not alter the relation among the features. This explanation is further strengthened from the point of view of the selection vector $\boldsymbol{\theta}$ in selective ridge regression.

Beyond The Linear Regression Loss

In formulations Eqs.(1,6,11), the error/loss term uses linear regression. But they can be any other forms of loss $E(\mathbf{W})$. The proofs of Theorems 1 and 3 only depend on the regularization term, and thus hold without any change. In other words, the process from the ℓ_2 regularization to the $\ell_{1,2}$ regularization is purely due the transformation of probabilistic selection.

Feature Selection Using $\ell_{1,2}$-Norm

From here on, we use $\mathbf{W}$ to replace $\widehat{\mathbf{W}}$ in Eq.(1) for notational simplicity.

As explained earlier, flexible feature selection does not enforce rigourously that features selected for every class are exactly same. This is naturally done in the $\ell_{1,2}$ regularization based selection we propose in this paper, written explicitly here for clarity,

$$\|\mathbf{W}\|_{1,2}^2=\sum_{j=1}^{k}\left(\sum_{i=1}^{d}|W_{ij}|\right)^2. \tag{14}$$

As regularization strength parameter λ goes large, different elements in $\sum_{i=1}^{d}|W_{ij}|$ for a fixed class j compete with

each other, and only a few elements (corresponding to different features) will survive (be nonzero), i.e., these features being selected for class j.

To the best of our knowledge, however, flexible feature selection has not been thoroughly investigated so far. The main trend in feature selection is using $\ell_{2,1}$-norm based formalisms (Argyriou, Evgeniou, and Pontil 2008), (Liu, Ji, and Ye 2009), (Nie et al. 2010), (Gui et al. 2017), selecting rows of weight matrix $\mathbf{W}$.

We note that the competition and survival property explained above for $\|\mathbf{W}\|_{1,2}^2$ also happens in exclusive lasso of Zhou et al. (Zhou, Jin, and Hoi 2010). Their formulation is different from our approach here. They use the regularization

$$\left\|\mathbf{W}^T\right\|_{1,2}^2 = \sum_{i=1}^{d}\left(\sum_{j=1}^{k}|W_{ij}|\right)^2. \tag{15}$$

As regularization strength parameter λ goes large, different elements in $\sum_{j=1}^{k}|W_{ij}|$ for a fixed feature i compete with each other, i.e., they are mutually exclusive, and only a few elements (corresponding to different classes) will survive (be nonzero), i.e., feature i being selected for these classes. This competition and survival property is the prominent feature of "exclusive LASSO". In Kong et al (Kong et al. 2014), they use exclusive group norm, $\sum_g \|\mathbf{w}_g\|_1^2$ (where g is the group index) which is very similar to exclusive lasso, except only 2-class case is considered there.

In summary, both our proposed $\ell_{1,2}$-norm based feature selection $\|\mathbf{W}\|_{1,2}^2$ and the exclusive LASSO $\left\|\mathbf{W}^T\right\|_{1,2}^2$ have the competition and survival property (the "exclusive" property), and can be used for flexible feature selection. However, $\ell_{2,1}$-norm based feature selection $\left\|\mathbf{W}^T\right\|_{2,1}$ is not suitable for flexible feature selection.

Efficient Algorithms

We wish to solve the $\ell_{1,2}$-norm based feature selection and the exclusive lasso (eLASSO). They are expressed as

$$E(\mathbf{W}) = \left\|\mathbf{X}^T\mathbf{W}-\mathbf{Y}\right\|_F^2, \tag{16}$$

$$J_{12}(\mathbf{W}) = E(\mathbf{W}) + \lambda\|\mathbf{W}\|_{1,2}^2, \tag{17}$$

$$J_{\mathrm{eLASSO}}(\mathbf{W}) = E(\mathbf{W}) + \lambda\left\|\mathbf{W}^T\right\|_{1,2}^2. \tag{18}$$

We use an iterative algorithm to solve the problem. Let $\mathbf{W}^0, \mathbf{W}^1, \cdots, \mathbf{W}^t, \mathbf{W}^{t+1}, \cdots$ be the solutions at different stages. Our task here is (A) derive an update algorithm $\mathbf{W}^{t+1} = f(\mathbf{W}^t)$, and (B) prove its convergence: $J(\mathbf{W}^{t+1}) \le J(\mathbf{W}^t)$.

We use the auxiliary function approach widely adapted in nonnegative matrix factorization (Lee and Seung 1999), (Lee and Seung 2000), (Ding, Li, and Jordan 2010) to derive an efficient algorithm. A function $G(\mathbf{W}, \widetilde{\mathbf{W}})$ is the auxiliary function of $J(\mathbf{W})$, if it satisfies condition (C1) $J(\mathbf{W}) \le G(\mathbf{W}, \widetilde{\mathbf{W}}), \forall \mathbf{W}, \widetilde{\mathbf{W}}$ and condition (C2) $J(\mathbf{W}) = G(\mathbf{W}, \mathbf{W}), \forall \mathbf{W}$.

The key step is finding the auxiliary function for the objective $J_{12}(\mathbf{W})$ and $J_{\mathrm{eLASSO}}(\mathbf{W})$. We have

Theorem 4. *An auxiliary function for $J_{12}(\mathbf{W})$ is*

$$\begin{aligned} & G_{12}(\mathbf{W}, \mathbf{W}^t) \\ = \; & E(\mathbf{W}) + \lambda\textstyle\sum_{j=1}^{k}\left(\sum_{i=1}^{d}\frac{W_{ij}^2}{|W_{ij}^t|}\right)\|\mathbf{w}_j^t\|_1 \\ = \; & E(\mathbf{W}) + \lambda\textstyle\sum_{j=1}^{k}\mathbf{w}_j^T\mathbf{D}_j\mathbf{w}_j, \end{aligned} \tag{19}$$

where

$$\mathbf{D}_j = \|\mathbf{w}_j^t\|_1\mathrm{diag}(1/|W_{1j}^t|, \cdots, 1/|W_{dj}^t|). \tag{20}$$

An auxiliary function for J_{eLASSO} is

$$\begin{aligned} & G_{\mathrm{eLASSO}}(\mathbf{W}, \mathbf{W}^t) \\ = \; & E(\mathbf{W}) + \lambda\textstyle\sum_{i=1}^{d}\left(\sum_{j=1}^{k}\frac{W_{ij}^2}{|W_{ij}^t|}\right)\|(\mathbf{w}^i)^t\|_1 \\ = \; & E(\mathbf{W}) + \lambda\textstyle\sum_{i=1}^{d}\mathbf{w}^i\mathbf{H}_i(\mathbf{w}^i)^T, \end{aligned} \tag{21}$$

where

$$\mathbf{H}_i = \|(\mathbf{w}^i)^t\|_1\mathrm{diag}(1/|W_{i1}^t|, \cdots, 1/|W_{ik}^t|), \tag{22}$$

and $\mathbf{w}^i$ is a row vector.

The proof of this theorem is given below.

In the following, we focus on deriving the update algorithm of $\ell_{1,2}$-norm based feature selection using $J_{12}(\mathbf{W})$. Algorithm for $J_{\mathrm{eLASSO}}(\mathbf{W})$ can be obtained in identical fashion.

The Update Algorithm

Using Theorem 4, the update algorithm is given by

$$\mathbf{W}^{t+1} = \arg\min_{\mathbf{W}} G_{12}(\mathbf{W}, \mathbf{W}^t). \tag{23}$$

This is solved by setting $\frac{\partial G_{12}(\mathbf{W}, \mathbf{W}^t)}{\partial \mathbf{W}} = 0$. The solution is

$$\mathbf{w}_j^{t+1} = (\mathbf{X}\mathbf{X}^T + \lambda\mathbf{D}_j)^{-1}(\mathbf{X}\mathbf{y}_j), \tag{24}$$

where $j = 1, \cdots, k$ is the class index, and $\mathbf{D}_j$ is defined in Eq.(20). Eq.(24) is the updating equation. Since $G_{12}(\mathbf{W}, \mathbf{W}^t)$ is a strict convex function in $\mathbf{W}$, $\mathbf{w}_j^{t+1}$ obtained is the global optimal solution.

Algorithm 1 Efficient algorithm for solving the $\ell_{1,2}$-norm based feature selection.

Input: Data matrix $\mathbf{X} \in \mathbb{R}^{d\times n}$, labels $\mathbf{Y} \in \mathbb{R}^{n\times k}$.
Output: $\mathbf{W} \in \mathbb{R}^{d\times k}$, $\mathbf{D}_j \in \mathbb{R}^{d\times d}$, $j = 1, \cdots, k$.
Set $t = 0$.
Initialize $\mathbf{W}^t$.
repeat
  for each class $j \in \{1, \cdots, k\}$ **do**
    Compute $\mathbf{D}_j$ via Eq.20.
    Compute $\mathbf{w}_j^{t+1}$ via Eq.24.
  end for
  Set $t = t + 1$.
until Converges

This is a convergent update algorithm, because we have $J_{12}(\mathbf{W}^{t+1}) \le G_{12}(\mathbf{W}^{t+1}, \mathbf{W}^t) \le G_{12}(\mathbf{W}^t, \mathbf{W}^t) = J_{12}(\mathbf{W}^t)$. The first inequality is due to the condition (C1) for the auxiliary function. The second inequality comes from the fact

that $\mathbf{W}^{t+1}$ is the global optimal solution for Eq.(23). The third equality comes from auxiliary function condition (C2).

In summary, we have derived the update algorithm outlined in Algorithm 1 and proved its convergence.

Proof of Theorem 4.

Auxiliary function condition (C1): $J_{12}(\mathbf{W}^{t+1}) \leq G_{12}(\mathbf{W}^{t+1}, \mathbf{W}^t)$. Let the difference between left-hand-side and right-hand-side of above inequality defined as $\Delta = \text{LHS} - \text{RHS}$. We obtain the following

$$\begin{aligned}
\Delta &= \left(\sum_{j=1}^{k} \left\|\mathbf{w}_j^{t+1}\right\|_1^2\right) - \left(\sum_{j=1}^{k} \left(\mathbf{w}_j^{t+1}\right)^T \mathbf{D}_j \left(\mathbf{w}_j^{t+1}\right)\right) \\
&= \sum_{j=1}^{k} \left[\left(\sum_{i=1}^{d} \left|W_{ij}^{t+1}\right|\right)^2 - \left(\sum_{i=1}^{d} \frac{\left|W_{ij}^{t+1}\right|^2}{\left|W_{ij}^{t}\right|}\right)\left(\sum_{i=1}^{d} \left|W_{ij}^{t}\right|\right)\right] \\
&= \sum_{j=1}^{k} \left[\left(\sum_{i=1}^{d} A_{ij}B_{ij}\right)^2 - \left(\sum_{i=1}^{d} A_{ij}^2\right)\left(\sum_{i=1}^{d} B_{ij}^2\right)\right] \leq 0
\end{aligned} \tag{25}$$

where $A_{ij} = \frac{|W_{ij}^{t+1}|}{\sqrt{|W_{ij}^t|}}$, $B_{ij} = \sqrt{|W_{ij}^t|}$. The last inequality in Eq.(25) is obtained according to the Cauchy-Schwarz[1] inequality, which proves condition (C1).

Auxiliary function condition (C2): $G_{12}(\mathbf{W}^t, \mathbf{W}^t) = J_{12}(\mathbf{W}^t)$. From the Eq.19, we obtain the following

$$\begin{aligned}
& G_{12}(\mathbf{W}^t, \mathbf{W}^t) \\
=\ & E(\mathbf{W}^t) + \lambda \textstyle\sum_{j=1}^{k} \left(\sum_{i=1}^{d} \frac{(W_{ij}^t)^2}{|W_{ij}^t|}\right) \|\mathbf{w}_j^t\|_1 \\
=\ & E(\mathbf{W}^t) + \lambda \textstyle\sum_{j=1}^{k} \|\mathbf{w}_j^t\|_1^2 \\
=\ & J_{12}(\mathbf{W}^t),
\end{aligned} \tag{26}$$

which proves condition (C2). Thus, Theorem 4 is proved. □

During the computation, many of the elements W_{ij} become zero due to sparsity. We therefore replace $1/|W_{ij}|$ by $1/(|W_{ij}| + \epsilon)$ where ϵ is a small number $1e{-}7$.

Experiment

For purpose of verifying the effectiveness of our flexible feature selection method via $\ell_{1,2}$-norm, extensive experiments on six benchmark datasets are conducted in comparison with six state-of-the-art algorithms.

Description of Benchmark Datasets

In our experiments, six benchmark datasets including images and bio-microarray data are used to study the performance of feature selection methods on classification. The description of all datasets are given as follows.

Image dataset: there are three image datasets, including MNIST[2] (Lecun et al. 1998), BinAlpha[3], AT&T[4]. Each instance is represented by a vector with all the pixel values in an image. In MNIST, handwritten digits from 0 to 9 are collected as samples. On the other hand, samples in BinAlpha are composed of handwritten letters from A to Z. Both digits and letters have been size-normalized and centered in a fixed-size image. AT&T, also known as the ORL database of faces, is the widely used face recognition dataset, in which images were taken at different times varying the lighting, facial expressions, and facial details.

Microarray dataset: there are three microarray datasets, including Carcinomas (Su et al. 2001), (Yang et al. 2006), Lung (Bhattacharjee et al. 2001), TOX[5] (Kwon et al. 2012). Each instance is represented by a vector with all the genes expression values. Under the first-generation molecular classification scheme, both Carcinomas and Lung are constructed to identify gene subsets whose expression typifies each cancer class, and quantify the extent to which genes are related to specific tumor type. In another hand, TOX focuses on discovering the time-course of changes in adipocyte morphology, adipokines and the global transcriptional landscape in visceral white adipose tissue, during the development of diet-induced obesity.

As compared to image dataset, microarray dataset usually involves a relatively small number of data instances but following with a extremely high dimension of features.

The detail of benchmark datasets is summarized in Table 1.

Dataset	#Classes	#Instances	#Features
MNIST	10	1000	784
BinAlpha	26	1014	320
AT&T	40	400	644
Carcinomas	11	174	9182
Lung	5	203	3312
TOX	4	171	5748

Table 1: Summary descriptions of dataset.

Classification Result and Analysis

Baseline methods: our $\ell_{1,2}$-norm based flexible feature selection method is compared to six state-of-the-art algorithms, including feature selection via $\ell_{2,1}$-norm (Argyriou, Evgeniou, and Pontil 2008), (Liu, Ji, and Ye 2009), (Nie et al. 2010), feature selection via $\ell_{1,\infty}$-norm (Quattoni et al. 2009), exclusive lasso (eLASSO) (Zhou, Jin, and Hoi 2010), mRMR (Peng, Long, and Ding 2005), F-statistic (Ding and Peng 2003), ReliefF (Robnik-Šikonja and Kononenko 2003). Towards a fair comparison, the hyperparameter λ in regression models, e.g., $\ell_{1,2}$ and $\ell_{2,1}$, is adjusted to achieve the same number of nonzero elements in weight matrix $\mathbf{W}$.

Classifiers: k-nearest neighbor (KNN), support vector machine (SVM), and linear regression (LR) with five-fold cross validation are used to evaluate the performance of feature selection on classification. The average of classification performance on different five folds are reported as the final accuracy. The parameter k in KNN is set as 3. LIBSVM

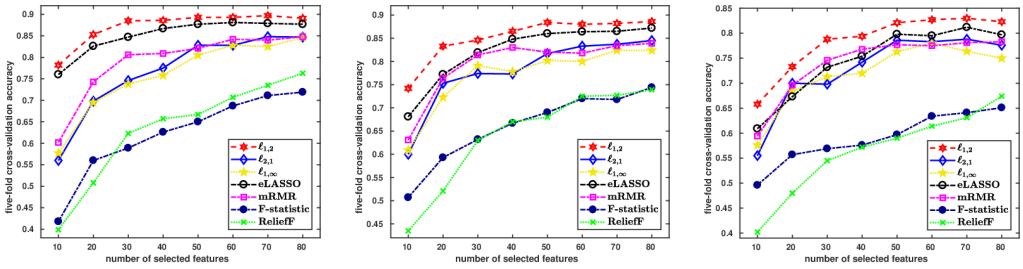
[1]Given any two vectors $\mathbf{x}$ and $\mathbf{y}$, the Cauchy-Schwarz inequality states, in the inner product space, it is always true that $(\sum_i x_i y_i)^2 \leq (\sum_i x_i^2)(\sum_i y_i^2)$.

[2]In MNIST, one hundred samples are randomly chosen out of each class to form a smaller dataset in our experiments.

[3]https://cs.nyu.edu/~roweis/data.html

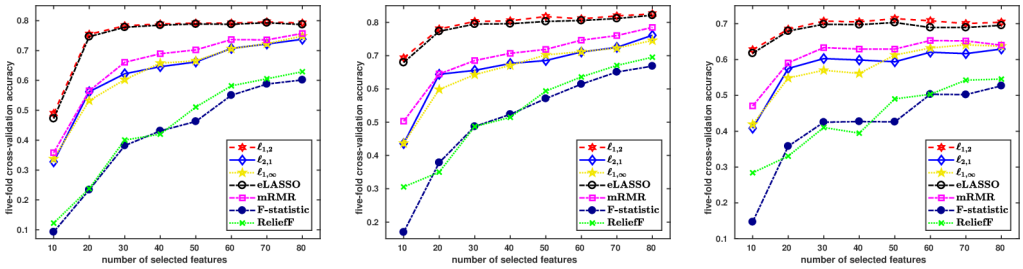
[4]http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

[5]http://featureselection.asu.edu/datasets.php



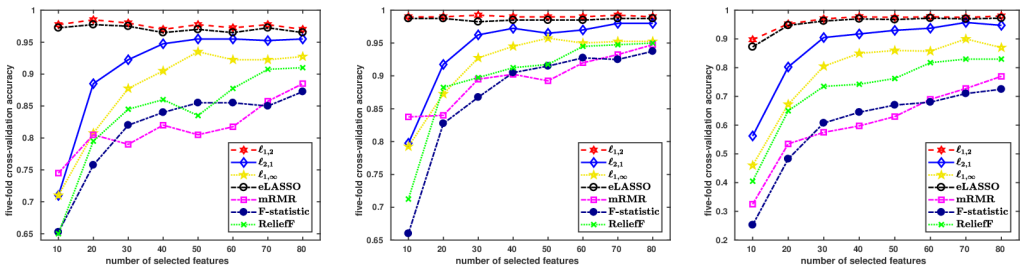
(a) KNN (b) SVM (c) LR

Figure 1: $\ell_{1,2}$ versus state-of-the-arts on MNIST dataset.



(a) KNN (b) SVM (c) LR

Figure 2: $\ell_{1,2}$ versus state-of-the-arts on BinAlpha dataset.



(a) KNN (b) SVM (c) LR

Figure 3: $\ell_{1,2}$ versus state-of-the-arts on AT&T dataset.

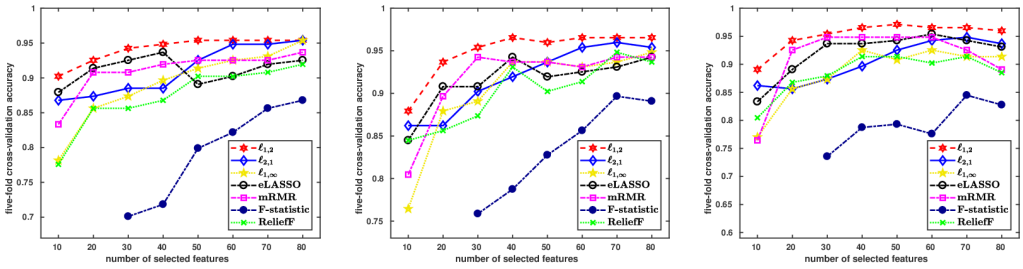
(Chang and Lin 2011) is used as practical implementation of SVM, in which the kernel is set as linear and $C = 1$.

Analysis of experimental results: As it can be seen in Fig. 1−6 that the classification using aforementioned seven feature selection methods is performed on six benchmark datasets. From left to right in each figure, employed classifiers are KNN, SVM, and LR respectively. The number of selected features for each method ranges from 10 to 80, which is marked as the scale of x-axis. The y-axis shows the averaged accuracy of five-fold cross validation.

Among these methods, the simplest F-statistic has the worst performance overall. Compared to F-statistic, another two filter-type methods, such as mRMR and ReliefF, improve the classification accuracy greatly. Moreover, mRMR can even beat sparse coding based methods such as $\ell_{2,1}$-norm or $\ell_{1,\infty}$-norm in some cases.

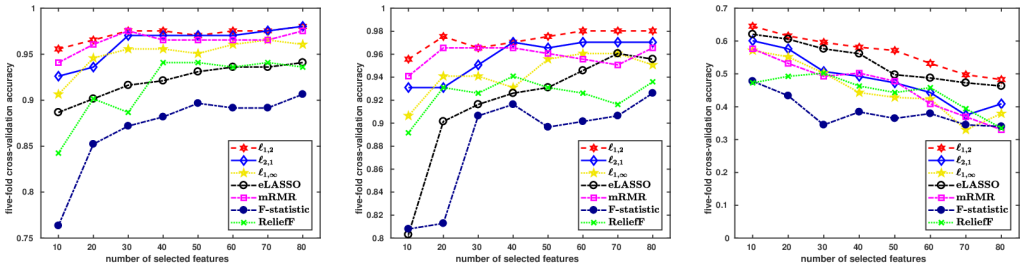
However, filter-type methods are inferior to sparse coding based methods in general. Feature selection via $\ell_{2,1}$-norm performs very close to feature selection via $\ell_{1,\infty}$-norm when classifying not only images but also bio-microarray data, since both methods share the same property that aims at searching a subset of class-shared features across all the data instances. Only the results obtained on AT&T dataset, $\ell_{2,1}$ is obviously better than $\ell_{1,\infty}$ around 5.0%. Among sparse coding based methods, eLASSO is an outstanding one which selects exclusive features as the main purpose, only performing slightly lower than our $\ell_{1,2}$-norm based method around 1.0% on BinAlpha and AT&T. Nevertheless, when the dimension of features becomes very large, eLASSO has a relatively bad results on microarray datasets.

Most importantly, our flexible feature selection method via $\ell_{1,2}$-norm achieves the best results on all six benchmark datasets compared to state-of-the-arts. No matter which classifier is used here, $\ell_{1,2}$ has an overwhelmed advantage over six baseline methods. Besides, $\ell_{1,2}$ has a stable performance without huge degradation, when using any feature subsets. Contrarily, most of baseline methods have a deteriorated performance in different degrees, when the number of selected



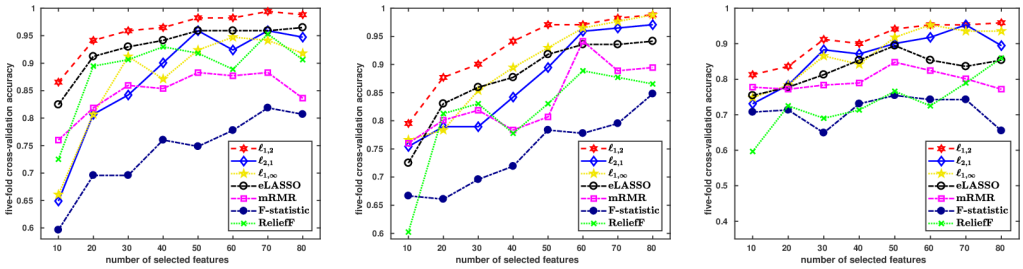
(a) KNN (b) SVM (c) LR

Figure 4: $\ell_{1,2}$ versus state-of-the-arts on Carcinomas dataset. F-statistic using top 10 and 20 features is not plotted in the figure, since the classification accuracy is way below the scale of y-axis.



(a) KNN (b) SVM (c) LR

Figure 5: $\ell_{1,2}$ versus state-of-the-arts on Lung dataset.



(a) KNN (b) SVM (c) LR

Figure 6: $\ell_{1,2}$ versus state-of-the-arts on TOX dataset.

features is relatively small. However, $\ell_{1,2}$ is better than others around 5%-10% using top 10 or 20 features. In summary, experimental results on benchmark datasets verify that $\ell_{1,2}$ based flexible selection is a more nature way to measure the importance of features than class-shared selections.

Conclusion

In this paper, we derive LASSO and $\ell_{1,2}$-norm feature selection from a probabilistic framework. In addition, we further propose a feature selection approach based on $\ell_{1,2}$-norm, allowing flexibility that selected features do not have to be exactly same for all classes. The resulting features lead to significantly better classification than state-of-the-arts algorithms on six benchmark datasets, including images and biomicroarray data.

Acknowledgment. This work is partially supported by a national science foundation grant: NSF-IIS-1633753.

References

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

Bhattacharjee, A.; Richards, W. G.; Staunton, J.; Li, C.; et al. 2001. Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences* 98(24):13790–13795.

Bolón-Canedo, V.; Sánchez-Maroño, N.; Alonso-Betanzos, A.; Benítez, J. M.; and Herrera, F. 2014. A review of microarray datasets and applied feature selection methods. *Information Sciences* 282:111–135.

Breiman, L. 1995. Better subset regression using the nonnegative garrote. *Technometrics* 37(4):373–384.

Campbell, F., and Allen, G. I. 2017. Within group variable

selection through the exclusive lasso. *Electronic Journal of Statistics* 11(2):4220–4257.

Chang, C.-C., and Lin, C.-J. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2:27:1–27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Ding, C., and Peng, H. 2003. Minimum redundancy feature selection from microarray gene expression data. In *Computational Systems Bioinformatics. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003*, 523–528.

Ding, C.; Zhou, D.; He, X.; and Zha, H. 2006. R1-pca: Rotational invariant ℓ_1-norm principal component analysis for robust subspace factorization. In *Proceedings of the 23rd International Conference on Machine Learning*, 281–288.

Ding, C. H. Q.; Li, T.; and Jordan, M. I. 2010. Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32(1):45–55.

Duchi, J.; Shalev-Shwartz, S.; Singer, Y.; and Chandra, T. 2008. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*, 272–279.

Dudoit, S.; Fridlyand, J.; and Speed, T. P. 2002. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* 97(457):77–87.

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

Gui, J.; Sun, Z.; Jia, W.; Hu, R.; Lei, Y.; and Ji, S. 2012. Discriminant sparse neighborhood preserving embedding for face recognition. *Pattern Recognition* 45(8):2884 – 2893.

Gui, J.; Sun, Z.; Ji, S.; Tao, D.; and Tan, T. 2017. Feature selection based on structured sparsity: A comprehensive study. *IEEE Transactions on Neural Networks and Learning Systems* 28(7):1490–1507.

Guyon, I., and Elisseeff, A. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research* 3:1157–1182.

Guyon, I.; Weston, J.; Barnhill, S.; and Vapnik, V. 2002. Gene selection for cancer classification using support vector machines. *Machine Learning* 46(1):389–422.

Kong, D.; Fujimaki, R.; Liu, J.; Nie, F.; and Ding, C. 2014. Exclusive feature learning on arbitrary structures via $\ell_{1,2}$ -norm. In *Advances in Neural Information Processing Systems 27*. 1655–1663.

Kwon, E.-Y.; Shin, S.-K.; Cho, Y.-Y.; Ju Jung, U.; et al. 2012. Time-course microarrays reveal early activation of the immune transcriptome and adipokine dysregulation leads to fibrosis in visceral adipose depots during diet-induced obesity. *BMC genomics* 13:450.

Lecun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11):2278–2324.

Lee, D. D., and Seung, H. S. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401:788–91.

Lee, D. D., and Seung, H. S. 2000. Algorithms for non-negative matrix factorization. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, 535–541.

Liu, J.; Ji, S.; and Ye, J. 2009. Multi-task feature learning via efficient $\ell_{2,1}$-norm minimization. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 339–348.

Lu, C.-Y.; Min, H.; Gui, J.; Zhu, L.; and Lei, Y.-K. 2013. Face recognition via weighted sparse representation. *Journal of Visual Communication and Image Representation* 24(2):111 – 116.

Nie, F.; Huang, H.; Cai, X.; and Ding, C. H. 2010. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *Advances in Neural Information Processing Systems 23*. 1813–1821.

Peng, H.; Long, F.; and Ding, C. 2005. 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.

Quattoni, A.; Carreras, X.; Collins, M.; and Darrell, T. 2009. An efficient projection for $\ell_{1,\infty}$ regularization. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 857–864.

Robnik-Šikonja, M., and Kononenko, I. 2003. Theoretical and empirical analysis of relieff and rrelieff. *Machine Learning* 53(1):23–69.

Saeys, Y.; Inza, I. n.; and Larrañaga, P. 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics* 23(19):2507–2517.

Su, A. I.; Welsh, J. B.; Sapinoso, L. M.; Kern, S. G.; et al. 2001. Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Research* 61(20):7388–7393.

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

Yang, K.; Cai, Z.; Li, J.; and Lin, G. 2006. A stable gene selection in microarray data analysis. *BMC Bioinformatics* 7(1):228.

Zhao, P.; Rocha, G.; and Yu, B. 2009. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics* 37(6A):3468–3497.

Zhou, Y.; Jin, R.; and Hoi, S. 2010. Exclusive lasso for multi-task feature selection. In *Proceedings of International Conference on Artificial Intelligence and Statistics*, 988–995.

Zhu, J.; Rosset, S.; Hastie, T.; and Tibshirani, R. 2003. 1-norm support vector machines. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, 49–56.