The Sparse Group Log Ridge for the Selection of Variable Groups

Song Yao, Lipeng Cui and Sining Ma

Tianjin Light Industry Vocational Technical College, Tianjin, 300350, China

*Corresponding author’s e-mail: yaosong@tjlivtc.edu.cn

Abstract. In recent years, the sparse model is a research hotspot in the field of artificial intelligence. Since the Lasso model ignores the group structure among variables, and can only achieve the selection of scattered variables. Besides, Group Lasso can only select groups of variables. To address this problem, the Sparse Group Log Ridge model is proposed, which can select both groups of variables and variables in one group. Then the MM algorithm combined with the block coordinate descent algorithm can be used for solving. Finally, the advantages of the model in terms of variables selection and prediction are shown through the experiment.

1. Introduction

In fields such as artificial intelligence and machine learning, the space of explanatory variables in many problems of classification and regression tends to be very high-dimensional, or even ultra-highly dimensional. However, high-dimensional data can lead to overfitting in machine learning, thus resulting in poor generalization performance of statistical models. Therefore, variable space dimensionality reduction and variables selection are two problems to be solved. The objective of variable selection is twofold. One is to achieve accurate prediction and classification, and the other is to make the model more interpretable and reduce the complexity of statistical models. The so-called interpretability refers to the simplicity of the model. Obviously, the lower the dimensionality of the variable space is, the better the interpretability of the statistical model is. The Lasso [1] based on the L1-norm penalty proposed by Tibshirani can effectively reduce the dimension of variable space but can only achieve distributed variable selection. In many cases, there is a certain structure among variables. For example, in gene microarray analysis, there are often multiple variants in a gene. When identifying the genetic variant that is associated with the disease, it is more reasonable to divide the variants belonging to the same gene into one group. Given this, some scholars conduct variables selection by taking the structure among variables as a priori information. Group Lasso [2] is a sparse model that uses the group structure of variables as a priori information, which can achieve the selection of variable groups. Sparse models such as SCAD [3], MCP [4], and adaptive Lasso [5] improve the consistency in variables selection of the Lasso. Elastic Net [6] can select groups of highly correlated variables simultaneously. Models such as Lasso have also been applied to the field of support vector machines (SVMs) to constitute sparse SVMs [7-11]. Support vector machine (SVM) has good classification performance, which has been studied extensively in the past ten years. What is more, the standard L2-norm SVM classifier uses a ridge penalty function, so feature selection cannot be implemented. However, feature selection is very important in many situations. For example, selecting important genes is a very important issue in bio-informatics. In many cases, eliminating noise variables can improve the classification accuracy of the classifier. Therefore, how to achieve...
class label classification and feature selection at the same time is crucial, so people naturally think of combining sparse model with support vector machine to achieve classification and feature selection at the same time. Moreover, the sparse model and SVM are similar in form, and both have the form of "loss function + penalty function". Besides, the combination of sparse model and support vector machine produces a series of sparse SVMs that can automatically realize feature selection. For example, the L1-norm SVM in the form of "hinge loss function + L1 norm penalty function" and the SCAD SVM in the form of "hinge loss function + SCAD penalty function", etc. Therefore, it can be seen that by introducing the penalty functions of different regularized sparse models into the SVM, SVMs with different functions of feature selection can be obtained. In addition to applications in statistical variables selection, Lasso and other sparse models have also been applied in many fields such as compressed sensing, signal reconstruction, and image reconstruction, and have an increasingly wide range of applications in fields such as biostatistics, machine learning, data mining, image processing, and signal processing.

This paper proposed Sparse Group Log Ridge (SGLR), a sparse model that can achieve two-layer variables selection based on log penalty [12].

2. SGLR model

2.1. Log Penalty

The log penalty is written in the following form:

\[
\varphi_{\lambda_1,a} (\theta) = \frac{\lambda_1}{\ln(a + 1)} \ln(a|\theta| + 1)
\]

\[
= \begin{cases} 
\frac{\lambda_1}{\ln(a + 1)} \ln(a\theta + 1), \theta \geq 0 \\
\frac{\lambda_1}{\ln(a + 1)} \ln(-a\theta + 1), \theta < 0
\end{cases}
\]  

(1)

where \( \lambda_1 > 0 \) and \( a > 0 \). When \( \lambda_1 = 1 \), the image of the function of log penalty is shown in Fig. 1, where the dashed line and the straight line indicate the two cases when \( a = 1 \) and \( a = 0.1 \), respectively, and the dotted line indicates the image of the L1-norm penalty. The horizontal coordinate theta and the vertical coordinate phi in Fig. 1 are the independent variable \( \theta \) and the function \( \varphi \) in Eq. (1), respectively. From the image of the log penalty, it can be seen that the log penalty tends to the L1-norm penalty when \( a \to 0 \). As the absolute values of the coefficients increase, the log penalty gradually decreases, i.e., the degree to which the variable coefficients are compressed gradually decreases, which is similar to the principle of SCAD penalty and MC penalty. Also, this property can effectively reduce the bias of estimating regression coefficients.

![Figure 1. Image of Log Penalty](image-url)
2.2. Establishment of the SGLR model

When the linear regression model \( y = X\beta + \varepsilon \) is known, where \( \beta \) is the coefficient vector of \( P \times 1 \)-dimension, \( X \) is the design matrix of \( n \times P \) order, and \( y \) is the output vector of \( n \times 1 \)-dimension. The noise \( \varepsilon \) follows a Gaussian distribution. First, \( P \) variables are divided into \( J \) groups. \( \beta_j \) is the coefficient vector of the \( j \)th variable group, \( X_j \) is the sub-design matrix of the \( j \)th variable group, \( d_j \) is the number of variables in the \( j \)th variable group, and \( \{j, 2, \ldots, J\} \). Without loss of generality, we may assume that the data is standardized, and \( \sum_{i=1}^{n} x_{ij} = 0 \), \( \sum_{i=1}^{n} x_{ij}^2 = 1 \), \( \sum_{i=1}^{n} y_j = 0 \), where \( x_{ij} \) is the observation value of the \( p \)th variable in the \( j \)th variable group. Then the SGLR can be written in the following form:

\[
\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^P} \frac{1}{2} \| y - X\beta \|_2^2 + \sum_{j=1}^{J} \phi_{\lambda_j, a}(\| \beta_j \|) + \frac{\lambda_2}{2} \| \beta \|_2^2, \tag{2}
\]

where \( \phi_{\lambda, a}(\cdot) \) is the log penalty, and \( \lambda_1, \lambda_2, \) and \( a \) are adjustable parameters. The integration of the group structure of the penalty function and the L1-norm penalty enables the SGLR to achieve two-layer variables selection, that is, select both the objective variable group and the objective variables within the objective variable groups. The SGLR proposed in this paper is inspired by the hierarchical penalty function, and the penalty function of the hierarchical penalty function can be written

\[
\sum_{j=1}^{J} f_{\text{out}} \left( \sum_{p=1}^{d_j} f_{\text{in}}(\beta_{jp}) \right), \quad \text{where } d_j \text{ is the number of variables contained in the } j \text{th variable group.}
\]

For SGLR, the function \( f_{\text{out}} \) in the outer layer is the log penalty and the function in the inner layer is the L1-norm penalty.

2.3. Solution Algorithm of the SGLR model

In SGLR, solutions cannot be obtained by directly using the block coordinate descent algorithm due to the lack of an explicit solution for a single variable. In this paper, the solution of SGLR is obtained by using MM (Majorize-Minimization) algorithm combined with the block coordinate descent algorithm. Since the log penalty is concave function in the range of \( [0, +\infty) \), the relationship between the log penalty and its first-order Taylor expansion at the current point \( \| \beta_j \|_1 \) is \( \phi(\| \beta_j \|_1) \leq \phi(\| \beta_j \|_1) + \tilde{\Delta}_j(\| \beta_j \|_1 - \| \tilde{\beta}_j \|_1) \), where \( \tilde{\Delta}_j \) is the first-order partial derivative of the log penalty at the current point \( \| \tilde{\beta}_j \|_1 \). And since the equality in the above inequality holds at the current point \( \| \tilde{\beta}_j \|_1 \), \( \phi(\| \tilde{\beta}_j \|_1) + \tilde{\Delta}_j(\| \beta_j \|_1 - \| \tilde{\beta}_j \|_1) \) can be used as the majorizing function with respect to \( \phi(\| \beta_j \|_1) \). Then the optimal solution of the original problem can be obtained through iterative operation, and SGLR can be written as follows.

\[
\begin{align*}
\arg \min_{\beta \in \mathbb{R}^P} & \frac{1}{2} \| y - X\beta \|_2^2 + \phi(\| \tilde{\beta}_j \|_1) + \tilde{\Delta}_j(\| \beta_j \|_1 - \| \tilde{\beta}_j \|_1) + \frac{\lambda_2}{2} \| \beta_j \|_2^2 \\
= & \arg \min_{\beta \in \mathbb{R}^P} \frac{1}{2} \| y - X\beta \|_2^2 + \tilde{\Delta}_j \| \beta_j \|_1 + \frac{\lambda_2}{2} \| \beta_j \|_2^2 \tag{3}
\end{align*}
\]
The reason why the above equality holds is that the elimination of the constant term does not affect the optimal solution to the optimization problem. Then the optimization problem for the $j$th variable groups can be written as follow.

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^n} \frac{1}{2} \left\| y - X\beta \right\|_2^2 + \tilde{\Delta}_j \left\| \beta \right\|_1 + \frac{\lambda_2}{2} \left\| \beta \right\|_2^2$$  \hspace{1cm} (4)

In fact, Eq. (4) is a Lasso problem, which can be solved by means of the block coordinate descent method. Since Eq. (4) is not differentiable at $0$, sub-gradient and subdifferential are used to derive the optimal solution. The subdifferential of $\left\| \beta \right\|_1$ at zero is $\{ e : e \in [-1,1] \}$, where $\tilde{\Delta}_j$ is the first-order partial derivative of the log penalty at the current point $\left\| \beta \right\|_1$. Since the function is not differentiable at zero, it is necessary to use sub-gradient and subdifferential. The following derivation is carried out.

The necessary and sufficient condition for the case that $\beta_{jp} = 0$ is the solution of Eq. (4) as follows.

$$0 \in \{-x_{jp}^T (y - X\beta_{-jp}) + \beta_{jp} \cdot \sum_{i=1}^n x_{ip}^2 + \tilde{\Delta}_j \cdot e + \lambda_2 \beta_{jp}\}$$  \hspace{1cm} (5)

where $e \in [-1,1]$, that is, $e_0 \in [-1,1]$ leads the following equation:

$$- x_{jp}^T (y - X\beta_{-jp}) + (1 + \lambda_2) \beta_{jp} \cdot \sum_{i=1}^n x_{ip}^2 + \tilde{\Delta}_j \cdot e_0 = 0$$  \hspace{1cm} (6)

Set $z_{jp} = x_{jp}^T (y - X\beta_{-jp})$, where

$$\beta_{-jp} = (\beta_1, \ldots, \beta_{(p-1)}, 0, \beta_{(p+1)}, \ldots, \beta_p)$$  \hspace{1cm} (7)

For $\beta_{jp} = 0$ and $\sum_{i=1}^n x_{ip}^2 = 1$, the following can be obtained:

$$\left\| z_{jp} \right\| = \tilde{\Delta}_j \cdot |e_0| \leq \tilde{\Delta}_j$$  \hspace{1cm} (8)

Then, it can be known that $\beta_{jp} = 0$ when $|z_{jp}| \leq \tilde{\Delta}_j$. For $\beta_{jp} \neq 0$, the following can be obtained:

$$- z_{jp} + \beta_{jp} \cdot \sum_{i=1}^n x_{ip}^2 + \tilde{\Delta}_j \cdot \text{sign}(\beta_{jp}) + \lambda_2 \beta_{jp} = 0$$  \hspace{1cm} (9)

For $\sum_{i=1}^n x_{ip}^2 = 1$, the following can be obtained:

$$\beta_{jp} = \frac{z_{jp} - \tilde{\Delta}_j \cdot \text{sign}(\beta_{jp})}{1 + \lambda_2}$$  \hspace{1cm} (10)

Then its explicit solution with respect to a single variable is as follows.

$$\beta_{jp} = \begin{cases} \frac{(z_{jp} - \tilde{\Delta}_j)}{(1 + \lambda_2)}, & z_{jp} > \tilde{\Delta}_j \\ \frac{(z_{jp} + \tilde{\Delta}_j)}{(1 + \lambda_2)}, & z_{jp} < \tilde{\Delta}_j \end{cases}$$  \hspace{1cm} (11)

Once the explicit solution of SGLR with respect to a single variable is obtained, the block coordinate descent algorithm can be solved. The algorithm for solving the SGLR is given below.

(a) Input the response vector $y$, the design matrix $X$, and the initial value $\beta$ of the regression coefficient vector.

(b) Repeat the following steps when $1 \leq j \leq J$.
(i) Set \( \tilde{\lambda}_j = \frac{\lambda a}{(a \| \beta_j \|_1 + 1) \ln(a + 1)} \).

(ii) Repeat the following steps when \( p = 1, 2, \ldots, d_j \) :

Let \( z_{jp} = x_{jp}^T (y - X \beta_{-jp}) \); solve \( \beta_{jp} \) through the Eq. (11); set \( p = p + 1 \); Change \( \beta_{jp} \).

(iii) Set \( j = j + 1 \).

(c) Get the regression coefficient vector \( \beta \) after traversing all the groups once, and determine whether the predefined convergence condition or the number of iterations is satisfied. If not, go to step (b); otherwise, end.

(d) Output the regression coefficient vector \( \beta \).

3. Experiments with Synthetic Datasets

This section generates three different types of datasets: synthetic dataset 1, synthetic dataset 2, and synthetic dataset 3. Among them, synthetic dataset 1 is sparse both in the level of variable groups and variables, while synthetic dataset 2 is sparse only in the level of variable groups and synthetic dataset 3 is sparse both in the level of variable groups and variables, and extremely sparse in the level of variables within group. All three datasets are generated based on the linear regression model \( y = X \beta + \epsilon \).

Synthetic dataset 1: This dataset contains \( 2n=600 \) samples and \( P=500 \) variables, and these 500 variables are divided into 53 variable groups. The grouping of the variables in this dataset and their true coefficient values are \( \beta = (\beta_1, \beta_2, \ldots, \beta_{53})^T \).

Synthetic dataset 2: This dataset contains \( 2n=400 \) samples and \( P=400 \) variables, and these 400 variables are divided into 43 variable groups. The grouping of the variables in this dataset and their true coefficient values are \( \beta = (\beta_1, \beta_2, \ldots, \beta_{43})^T \).

Synthetic dataset 3: This dataset contains \( 2n=300 \) samples and \( P=300 \) variables, and these 300 variables are divided into 32 variable groups. The grouping of the variables in this dataset and their true coefficient values are \( \beta = (\beta_1, \beta_2, \ldots, \beta_{32})^T \).

Table 1. Experiment Results of Synthetic Dataset 1

| Model     | n    | P    | Rel | Noi  | Size | MSE  |
|-----------|------|------|-----|------|------|------|
| Lasso     | 300  | 500  | 10.00 | 0.00 | 10.00 | 0.0452 |
| Group Lasso | 300  | 500  | 9.00 | 6.00 | 15.00 | 0.3973 |
| SGLR      | 300  | 500  | 9.00 | 0.63 | 9.63  | 0.1466 |

Table 2. Experiment Results of Synthetic Dataset 2

| Model     | n    | P    | Rel | Noi  | Size | MSE  |
|-----------|------|------|-----|------|------|------|
| Lasso     | 200  | 400  | 15.00 | 0.56 | 15.56 | 0.0862 |
| Group Lasso | 200  | 400  | 15.00 | 0.00 | 15.00 | 0.5104 |
| SGLR      | 200  | 400  | 15.00 | 0.00 | 15.00 | 0.0769 |

Table 3. Experiment Results of Synthetic Dataset 3

| Model     | n    | P    | Rel | Noi  | Size | MSE  |
|-----------|------|------|-----|------|------|------|
| Lasso     | 150  | 300  | 3.00 | 0.13 | 3.13  | 0.0156 |
| Group Lasso | 150  | 300  | 3.00 | 7.00 | 10.00 | 0.0646 |
| SGLR      | 150  | 300  | 3.00 | 0.00 | 3.00  | 0.0239 |

For the three synthetic datasets, the variables are randomly divided into two sub-datasets containing an equal number of samples, one of which is used as the training dataset and the other as the test dataset. The above division process is repeated for 30 times to obtain 30 experimental results, and the mean value of the 30 experimental results is taken as the final experimental result. The experimental results are shown in Table 1, Table 2, and Table 3. Where, \( n \) denotes the number of training samples, \( P \)
denotes the total number of variables, Size denotes the total number of selected variables, Rel denotes the number of identified objective variables, Noi denotes the number of eliminated redundant variables, and MSE denotes the predicted mean square error. The experimental results of the three synthetic datasets show that SGLR performs well on prediction and variables selection in terms of accuracy.

To verify the capability for two-layer variables selection of SGLR, the first experiment of 30 experiments conducted for the synthetic data set 1 is selected, and the non-zero regression coefficients fitted for that experiment are included in Tables 4, 5, 6, and 7, and the coefficients not listed equal 0. The grouping of the variables in the synthetic dataset 1 and their true coefficient values are

\[ \beta = (\beta_1, \beta_2, \ldots, \beta_{53})^T \]

where each component of the above vectors is as follows:

\[ \beta_1 = (1,2,0,0,0) \]
\[ \beta_2 = (3,2,3,3,0) \]
\[ \beta_3 = (0,0,0,0,0) \]
\[ \beta_4 = (0,0,0,0,0) \]
\[ \beta_5 = (0,0,0,0,0) \]
\[ \beta_6 = (0,0,0,0,0,0) \]
\[ \beta_7 = (0,0,0,0,0,0,0) \]

and \( \beta_{53} = \beta_{52} = \cdots = \beta_3 = 0 \in \mathbb{R}^n \).

By comparing the experimental results of Lasso and Group Lasso, it can be known that Lasso only achieves scattered variables selection, while Group Lasso only achieves the selection of variable groups, where the coefficients of the variables grouped into the same variable group are all zero or all non-zero.

The capability for within-group variables selection of SGLR can be seen by comparing its experimental results with those of Group Lasso. Group Lasso only achieves the selection of variable groups, where the coefficients of variables grouped into the same variable group are all zero or all non-zero, while SGLR achieves within-group variables selection, where the coefficients of variables within the same group are either zero or non-zero.

The capability for variable groups selection of SGLR can be seen by comparing its experimental results with those of Lasso. The true regression coefficient of \( V_{40} \) is 0.3, and \( V_{40} \) is the objective variable expected to be selected, and Lasso selected \( V_{40} \). But the regression coefficient value of \( V_{40} \) obtained in SGLR was 0, and \( V_{40} \) was not selected in SGLR. The reason is that \( V_{40} \) in SGLR was in the same variable group with \( V_{31}-V_{39} \), which are other nine non-significant variables with a true regression coefficient of zero. The existence of a large number of non-significant variables enables the variable group to be a non-significant variable group. \( V_{40} \) is affected by these nine non-significant variables in the same group, resulting in \( V_{40} \) not being selected. In other words, 10 variables in the same group in SGLR are not selected at the same time, that is, the group sparsity and selection of variable groups are achieved. Therefore, the experimental results of Lasso, Group Lasso, and SGLR show that SGLR achieves the selection of both variable groups and within-group variables, that is, SGLR has the two-layer variable selection capability.

Table 4. Coefficients Obtained by Lasso

| Variable | V1 | V2 | V6 | V7 | V8 | V9 | V28 | V29 | V30 | V40 |
|----------|----|----|----|----|----|----|-----|-----|-----|----|
| Coefficient | 0.95 | 1.93 | 2.95 | 1.92 | 2.92 | 2.95 | 0.93 | 0.94 | 0.95 | 0.21 |

Table 5. Coefficients of the First Part Obtained by Group Lasso

| Variable | V1 | V2 | V3 | V4 | V5 | V6 | V7 | V8 | V9 | V10 |
|----------|----|----|----|----|----|----|----|----|----|-----|
| Coefficient | 0.89 | 1.70 | -0.02 | -0.09 | -0.03 | 2.84 | 1.85 | 2.80 | 2.87 | -0.07 |

Table 6. Coefficients of the Second Part Obtained by Group Lasso

| Variable | V26 | V27 | V28 | V29 | V30 |
|----------|-----|-----|-----|-----|-----|
| Coefficient | 0.02 | 0.01 | .80 | .82 | .83 |

Table 7. Coefficients Obtained by SGLR

| Variable | V1 | V2 | V6 | V7 | V8 | V9 | V10 | V28 | V29 | V30 |
|----------|----|----|----|----|----|----|-----|-----|-----|----|
| Coefficient | 0.95 | 1.86 | 2.87 | 1.88 | 2.83 | 2.87 | 0.01 | 0.91 | 0.91 | 0.94 |
Due to the $L_1$-norm penalty being non-derivable at zero, Lasso generates sparse solutions. Also, since Lasso does not use the group structure between variables as a priori information, it only achieves a scattered sparse selection. Both Group Lasso and the Group Berhu model proposed in Section 3 of this paper use the group structure as priori information. Considering the $L_2$-norm penalty being non-derivable at zero, the regression coefficients of variables in the same variable group tend to be zero or non-zero at the same time, thus achieving group sparsity. Based on a hierarchical penalty function and given that the $L_1$-norm penalty is non-derivable at zero, SGLR generates sparse solutions, but due to the prior use of group structure for grouping, the variables in the same group shall interact with each other, thus achieving both group sparsity and within-group sparsity.

4. Experiments with Real Datasets

Based on the birthweight datasets collected by Hosmer and Lemeshow, 189 variables were randomly divided into two sub-datasets containing 100 and 89 samples, respectively, with the sub-dataset containing 100 samples as the training dataset and the other sub-dataset containing 89 samples as the test dataset. The above division process was repeated for 100 times and 100 experimental results were obtained. The mean value of 100 experimental results is taken as the final experimental result. The experimental results are shown in Table 8, from which it can be seen that SGLR has higher accuracy in prediction compared with Lasso and Group Lasso.

| Model          | $n$ | $P$ | Size | MSE   |
|---------------|-----|-----|------|-------|
| Lasso         | 100 | 19  | 3.33 | 0.5119|
| Group Lasso   | 100 | 19  | 10.56| 0.4837|
| Group Log Ridge| 100 | 19  | 10.91| 0.4792|

5. Conclusions and Future Work

Sparse model is a hot research topic in the fields of machine learning and artificial intelligence in recent years. The Lasso based on the $L_1$-norm penalty proposed by Tibshirani can effectively reduce the dimension of variable space but can only achieve distributed variable selection. In many cases, there is a certain structure among variables. For example, in gene microarray analysis, there are often multiple variants in a gene. When identifying the genetic variant that is associated with the disease, it is more reasonable to divide the variants belonging to the same gene into one group. Given this, some scholars conduct variables selection by taking the structure among variables as a priori information. Group Lasso is a sparse model that uses the group structure of variables as a priori information, which can achieve the selection of variable groups.

This paper proposed a new two-layer sparse model SGLR for variables selection. Based on a hierarchical penalty function, the model utilizes the $L_1$-norm penalty that can achieve within-group sparsity internally and log penalty externally. SGLR can identify both significant variable groups in a large number of variables and significant variables within significant groups. Finally, this section verifies its capability of two-layer variables selection through synthetic dataset experiments. Both synthetic dataset experiments and real dataset experiments show that SGLR performs better in terms of prediction accuracy and variables selection consistency compared to sparse models such as Lasso and Group Lasso.

Many sparse models have oracle properties, such as Group SCAD. The oracle property of SGLR proposed in this paper needs to be further explored and validated. A wider selection of support vector machines and other loss functions should be developed. The current sparse support vector machines are mostly researched on the basis of a single hinge loss function. In the future, the sparse support vector machines can be expanded to other types of loss functions such as multi-class loss functions and least squares loss functions to achieve classification in different scenarios. The shortcoming of the Elastic Net, HHSVM and DrSVM is that only highly correlated features can be selected as a group simultaneously. In other words, the Elastic Net, HHSVM and DrSVM can not achieve feature group selection about arbitrary features. How to solve this problem is still an open question. Note that the...
SGLR proposed in this paper can achieve feature group selection about arbitrary features through dividing the features into different groups manually in advance, so can we apply the norm penalty function of the SGLR into the support vector machine classification field to achieve classification and feature group selection about arbitrary features simultaneously? In addition, the group SCAD, group bridge and group MCP are nonconvex unbiased feature group selection methods, can we apply them to the support vector machine classification field to achieve the feature group selection about arbitrary features and classification simultaneously? It is worth exploring these questions in future.

References
[1] Tibshirani R. (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, 58(1): 267-288.
[2] Yuan M, Lin Y. (2006) Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, 68(1): 49-67.
[3] Fan J Q, Li R Z. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96(456): 1348-1360.
[4] Zhang C H. (2010) Nearly unbiased variable selection under minimax concave penalty. Annals of Statistics, 38 (2): 894-942.
[5] Zou H. (2006) The adaptive lasso and its oracle properties. Journal of the American Statistical Association, 101(476): 1418-1429.
[6] Zou H, Hastie T. (2005) Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, 67(2): 301-320.
[7] Zhu J, Rosset S, Hastie T et al. (2004) 1-norm support vector machines. Advances in Neural Information Processing Systems, 16(1): 49-56.
[8] Zhang H H, Ahn J, Lin X et al. (2006) Gene selection using support vector machines with non-convex penalty. Bioinformatics, 22(1): 88-95.
[9] Zuo X, Huang H L, Li H E et al. (2013) P-norm regularized SVM classifier by non-convex conjugate gradient algorithm. In: Proceedings of the 25th Chinese Control and Decision Conference, Guiyang, China: 2685-2690.
[10] Zhang H H, Liu Y, Wu Y et al. (2008) Variable selection for the multicategory SVM via adaptive sup-norm regularization. Electronic Journal of Statistics, 2(8): 149-167.
[11] Zou H. (2007) An improved 1-norm SVM for simultaneous classification and variable selection. In: Proceedings of the 11th International Conference on Artificial Intelligence and Statistics. San Juan, Puerto Rico: 675-681.
[12] Mazumder R, Friedman J H, Hastie T. (2011) Sparse net: coordinate descent with nonconvex penalties. Journal of the American Statistical Association, 106(495): 1125-1138.
[13] Wang L, Chen G, Li H. (2007) Group SCAD regression analysis for microarray time course gene expression data. Bioinformatics, 23(12): 1486-1494.