Statistics and Its Interface Volume 16 (2023) 433–443

An iterative algorithm with adaptive weights and sparse Laplacian shrinkage for regression problems[*]

Xingyu Chen and Yuehan Yang[†]

This paper considers the regression problem with correlation structures among covariates. We propose an iterative algorithm, named Adaptive Sparse Laplacian Shrinkage (AdaSLS). This algorithm bases on a graph-constrained regularization. In each iteration, an adaptive weight is fitted within the feature space obtained from the previous step. Under suitable regularity conditions, AdaSLS obtains the correct feature set and accurate estimation with high probability. Its bias decay at an exponential rate. Numerical comparisons show that AdaSLS improves the accuracy of both variable selection and estimation. We also apply the proposed algorithm on a gene microarray dataset and a chimeric protein dataset, obtaining meaningful results.

Keywords and phrases: High-dimensional data, Correlated effects, Laplacian Matrix, Adaptive weight, Iterative algorithm.

1. INTRODUCTION

With the advent of the data revolution, high-dimensional and large-sample data are analyzed in many fields, such as economics, brain science, environmental science, finance, etc. Numerous studies have focused on high-dimensional problems, in which the number of variables is much larger than that of observations. In recent years, the collected data often contain complicated correlations. The processing of the correlation structures in high-dimensional data is of substantial importance.

Among statistical modeling and analysis, the regression model is a classic model of supervised learning. The model is easy to explain and is less prone to over-fitting. Shrinkage estimation with penalty is a common technique in high-dimensional linear regression models. A great deal of work has focused on variable selection and estimation of penalized methods. For example, Hoerl and Kennard [12] proposed Ridge Regression. Lasso (Least absolute shrinkage operator) was proposed by Tibshirani [27]. To improve the bias of lasso estimation, Fan and Li [7] proposed the smoothly clipped absolute deviation (SCAD) penalty; Zou [32] proposed the adaptive lasso; and then Zhang [29] proposed the minimum concave penalty (MCP). Extensions of the penalized methods have enhanced their applicability and offered theoretical guarantees; for example, see Bühlmann and Van De Geer [4], Hastie, Tibshirani and Wainwright [11], Fan et al. [8].

Yet, when dealing with the correlated data, the above methods do not consider the correlations among predictors, regarding the predictors as independent of each other. As Zhang and Huang [30] pointed out, lasso tends to select only one variable from a group of highly correlated variables. The other methods mentioned above also have such characteristics. To solve this problem, Zou and Hastie [33] proposed the elastic net combining l_1 and l_2 penalty, and then Zou and Zhang [34] proposed the adaptive elastic net. Bondell and Reich [2] proposed the OSCAR (Octagonal Shrinkage and Clustering Algorithm for Regression), and Huang et al. [14] proposed the Mnet which combines MCP and l_2 penalty. These methods deal with some collinearity and group effects among predictors. However, there are still gaps, for these combinations of penalties do not make use of the specific correlation structure among predictors.

When we study the data with correlated structure, ignorance of correlated relationships among predictors may lead to large estimation error. To address this problem, Li and Li [17] proposed a network constrained regularization method, and the variable is the measurement of genomic data on the genetic network. Li and Li [18] focused on the regression analysis assuming that the relationship among predictors is based on an undirected graph and is known. Its penalty is a combination of the l_1 penalty and the Laplacian quadratic penalty that is associated with the graph. But in practical problems, the real graph structure of the variables is usually unknown. Daye and Jeng [6] proposed the weighted fusion method, combining the l_1 penalty and the quadratic form to integrate the information among related variables for estimation and variable selection. Tutz and Ulbricht [28] studied a penalized method based on correlation, which can be considered as a special case of a general quadratic penalty. Pan,

[*] This work was supported by the National Natural Science Foundation of China (Grant No. 12001557); the Youth Talent Development Support Program (QYP202104), the Emerging Interdisciplinary Project, and the Disciplinary Funding of Central University of Finance and Economics.

[†] Corresponding author.

Xie and Shen [23] proposed a grouped penalty based on the l_γ norm of $\gamma > 1$, which can smooth the regression coefficient based on the network structure. Huang et al. [13] proposed the sparse Laplacian shrinkage (SLS) method, which combines the MCP penalty and the Laplace quadratic penalty.

In this paper, we aim to study the data with correlation structures among predictors. When there are complex correlation structures, it is still challenging to keep an accurate estimation. We plan to use the iteration to reduce the bias of estimation, and distinguish the nonzeros from zeros. We propose an iterative algorithm with adaptive sparse Laplacian shrinkage for regression problems, named Adaptive sparse Laplacian shrinkage (AdaSLS). During iterations, AdaSLS uses the penalty term of the adaptive l_1 penalty to ensure the sparsity of the model and the Laplace quadratic penalty to enhance the smoothness among coefficients of correlated predictors. The proposed method constantly updates the weights in the iteration process, reducing the bias of the estimated coefficient caused by mistaken initial weights. The quadratic form of the Laplace penalty can be associated with the Laplace matrix of the undirected weighted graph. Thus, the information of the specific network structure can be added to construct the model. We also apply AdaSLS to empirical data with correlated covariates, including a gene microarray dataset and a chimeric protein dataset. Results show that the AdaSLS has good performance in both two datasets.

This paper is organized as follows. Section 2 presents the method and describes a coordinate descent algorithm for the method, then discusses the ways to construct the adjacency matrix and the relationship between the proposed method and other methods. The theoretical properties of the proposed procedure are shown in Section 3. In Section 4 we show the simulation results. Applications are shown in Section 5. We conclude in Section 6.

2. THE ADAPTIVE SPARSE LAPLACIAN SHRINKAGE ESTIMATOR

2.1 Model and method

We consider the linear regression model with n observations and p predictors:

$$y = X\beta + \epsilon,$$

where $y = (y_1, \cdots, y_n)^{\mathrm{T}}$ is the vector of n responses; X is the $n \times p$ matrix of predictors; $\beta = (\beta_1, \cdots, \beta_p)^{\mathrm{T}}$ is the vector of regression coefficients; $\epsilon = (\epsilon, \cdots, \epsilon_n)^{\mathrm{T}}$ is the vector of random errors. We are interested in the analysis of high-dimensional data with correlation structure. We propose an iterative loss function that combines the weighted lasso and the Laplacian penalty. Among, the weight of l_1 penalty is updated constantly. The Laplacian penalty is used to deal with the correlation structure. The proposed iterative loss function is as following:

$$\begin{aligned} L(\beta; \lambda_1, \lambda_2) = & \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda_1 \sum_{j=1}^{p} w_j|\beta_j| \\ & + \frac{1}{2}\lambda_2 \sum_{1 \le j < j' \le p} |a_{jj'}|(\beta_j - s_{jj'}\beta_{j'})^2, \end{aligned} \tag{1}$$

where w_j is the weight that updates during the coordinate descent process. The first penalty $\lambda_1 \sum\limits_{i=1}^{p} w_j|\beta_j|$ is short for an iterative term and ensures the sparsity of the estimated model. Specifically, during each iteration of the coordinate descent process, $w = (w_1, \cdots, w_p)'$ updates after $\widehat{\beta} = (\widehat{\beta}_1, \cdots, \widehat{\beta}_p)'$ updates. Until convergence, the weights use information from the estimated coefficients constantly to reduce the bias of estimated coefficients that comes from the bias of weight. More details can be found in the Algorithm.

In the second penalty, $a_{jj'}$ is a measure of the correlation between X_j and $X_{j'}$. It describes how related the X_j and $X_{j'}$ are. $s_{jj'} = \mathrm{sign}(a_{jj'})$ denotes the sign of $a_{jj'}$. The second penalty $\frac{1}{2}\lambda_2 \sum\limits_{1 \le j < j' \le p} |a_{jj'}|(\beta_j - s_{jj'}\beta_{j'})^2$ shrinks $\beta_j - s_{jj'}\beta_{j'}$ to zero when $a_{jj'} \neq 0$. The smoothness is affected by the choice of tuning parameter λ_2. Predictors with negative correlations tend to obtain estimates with different signs, and vice versa. Compared to the ridge penalty that all the predictors are shrunk in the same level, the information of correlation structure among the predictors is applied in this function. The second penalty can also be associated with the Laplace matrix L of the undirected graph and can be represented by the Laplace matrix L:

$$\beta^{\mathrm{T}} L \beta = \sum_{1 \le j < j' \le p} |a_{jj'}|(\beta_j - s_{jj'}\beta_{j'})^2, \forall \beta \in \mathbb{R}^p.$$

Based on (1), we then discuss the details of the proposed algorithm in the following. When estimating β_j, other $\beta_{j'}$ for $j' \neq j$ are fixed. At each step, we update $\hat{\beta}_j$ by minimizing the loss function. After all the entries of $\widehat{\beta}$ been updated, each element of the weight $w = (w_1, \cdots, w_p)^{\mathrm{T}}$ are updated by $w_j = 1/\widehat{\beta}_j$ or n respectively for $\widehat{\beta}_j \neq 0$ or $\widehat{\beta}_j = 0$ for $j \in \{1, \cdots, p\}$. Repeat the above steps until convergence. The proposed iterative algorithm is as follows:

Step 1: Given λ_1, λ_2, adjacency matrix $A = (a_{ij})_{p \times p}$ and calculate lasso solution $\beta^{[0]}$.

Step 2: For $m = 1, \ldots,$ and $j \in \{1, \cdots, p\}$, calculate:

$$w_j = \begin{cases} \frac{1}{\hat{\beta}_j^{[m-1]}} & \hat{\beta}_j^{[m-1]} \neq 0 \\ n & \hat{\beta}_j^{[m-1]} = 0 \end{cases},$$

$$S_{1j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}(y_i - \sum_{j' \neq j} x_{ij'}\hat{\beta}_{j'}^{[m-1]}) + \lambda_2 \sum_{j' \neq j} a_{jj'}\hat{\beta}_{j'}^{[m-1]},$$

and

$$S_{2j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 + \lambda_2 \sum_{j' \neq j} |a_{jj'}|.$$

Step 3: Update:

$$\hat{\beta}_j^{(m)} \leftarrow \frac{S(S_{1j}, \lambda_1 w_j)}{S_{2j}}, j \in \{1, \cdots, p\},$$

where

$$S(z,\gamma) = \text{sign}(z)(|z| - \gamma)_+ = \begin{cases} z-\gamma & z>0 \text{ and } |z|>\gamma \\ z+\gamma & z<0 \text{ and } |z|>\gamma \\ 0 & |z|<\gamma \end{cases}.$$

Step 4: If $\sum_j (\hat{\beta}_j^{(m+1)} - \hat{\beta}_j^{[m]})^2 < 10^{-4}$, the iteration stops, otherwise it returns to Step 2 and Step 3.

2.2 Construction of adjacency matrix

The adjacency matrix describes the correlation between variables. We use the following ways to construct the adjacency matrix that describes different correlations.

i. The Pearson's correlation coefficient [24] is a commonly used measure to describe correlation. It describes the linear correlation between predictors well. We define the adjacency matrix according to it as follows:

$$s_{jj'} = \text{sign}(r_{jj'}) \quad \text{and} \quad a_{jj'} = s_{jj'} I(r_{jj'} > c),$$

where $r_{jj'} = X_j' X_{j'} / \|X_j\| \|X_{j'}\|$ is the Pearson's correlation coefficient between jth predictor and j'th predictor, and c is the threshold value which can be determined by the Fisher transformation or other information.

According to Mazumder and Hastie [19], when the population covariance matrix is a block diagonal matrix, the threshold sample covariance is equivalent to the estimated inverse covariance matrix obtained by the Glasso which characterizes the relative independence of predictors. That is, in this case, after standardized, the threshold sample covariance or say Pearson's correlation characterizes the relative independence of predictors. The threshold value c can be determined by specifying the sparsity of the matrix.

ii. The distance correlation coefficient [26] describes both linear and nonlinear association between predictors. This is in contrast to the above correlation coefficient which can only detect linear correlation. We define the adjacency matrix:

$$s_{jj'} = \text{sign}(d_{jj'}) \quad \text{and} \quad a_{jj'} = s_{jj'} I(d_{jj'} > c),$$

where $d_{jj'}$ is the distance correlation coefficient between jth predictor and j'th predictor and c is the threshold value.

iii. The adjacency matrix is also built using the partial residual. That is, during iteration, the adjacency matrix is updated by the last estimate and the partical residual of jth predictor in the mth iteration is

$$e_j^{[m]} = y - \sum_{j' \neq j} X_{j'} \hat{\beta}_{j'}^{[m-1]}.$$

The jth partial residual can be regard as the dependent variable vector corrected for all independent variables except the jth independent variables [16]. According to the partial residual, we define the adjacency matrix in the mth iteration as follows:

$$s_{jj'}^{[m]} = \text{sign}(R_{jj'}^{[m]}) \quad \text{and} \quad a_{jj'}^{[m]} = s_{jj'}^{[m]} I(R_{jj'}^{[m]} > c)$$

where c is the threshold value, and

$$R_{jj'}^{[m]} = (e_j'^{[m]} e_{j'}^{[m]}) / (\|e_j^{[m]}\| \|e_{j'}^{[m]}\|)$$

is the Pearson's correlation coefficient between the partial residual of the jth predictor and that of the j'th predictor.

2.3 Comparison

The proposed procedure is related to but different from the adaptive lasso and the SLS method. The adaptive lasso, for example, leads to an unbiased estimation by introducing the adaptive weights, and thus enjoys the oracle properties. The adaptive weights, on the other hand, are given at the beginning, and if the initial weights do not accurately characterize the model, the estimate will have a substantial bias. The SLS employs the MCP penalty as well as the graph Laplace matrix. This procedure enjoys the oracle properties and deals with the data with correlation structure well. However, it only uses the Pearson correlation coefficient to construct the adjacency matrix, which reflects the linear correlation between predictors, regardless of the nonlinear or other correlation. The AdaSLS has three advantages:

- Different from the adaptive lasso, the proposed method updates the weights constantly throughout iterations based on the coordinate descent process. This strategy reduces the bias caused by the initial weights.
- We consider three types of measures for calculating the correlations between predictors, e.g., Pearson's correlation, distance correlation, and partial residual during iterations, offering additional options for constructing the adjacency matrix. We further discuss that how the threshold sample covariance captures the relative independence of predictors well when the covariance matrix is a block diagonal matrix.
- We provide theoretical guarantees on both the error bound of the estimate and its sign consistency. Compared to [13], we further provide the upper bound of l_2-norm error.

These advantages help enhance the performance of AdaSLS when dealing with the regression problem with correlation structures. Simulation studies and applications show that the proposed method estimates coefficients with small error and low variance and performs well in variable selection.

3. THEORETICAL RESULTS

In this section, we study the theoretical properties of the proposed iterative function with the following dimensionality: $p = O(e^{n^{c_1}})$ and $q = O(n^{c_2})$ where $0 < c_1 + c_2 < 1$. Specifically, we consider the q-sparse model that only q covariates are relevant to the model and q is much smaller than both p and n. In the meantime, denote $S = \{j : \beta_j \neq 0\}$, thus, $|S| = q$. Its estimated value is $\hat{S} = \{j : \hat{\beta}_j \neq 0\}$. We can write X and β in the partitioned form $\beta = (\beta_S^{\mathrm{T}}, \beta_{S^c}^{\mathrm{T}})^{\mathrm{T}}$, $X = (X_S, X_{S^c})$, respectively. Without loss of generality, suppose the covariates are centered and normalized, i.e., $\sum_{i=1}^n x_{ij}^2/n = 1$ for $j = 1, \ldots, p$ and the error vector ϵ are independently distributed from $N(0, \sigma^2)$. We have the following assumption:

Condition 1. *Assume X satisfies the Restricted Eigenvalue (RE) condition: there exists positive constants κ, κ_0 that*

$$v^{\mathrm{T}} X^{\mathrm{T}} X v/n \geqslant \kappa \|v\|_2^2, \quad \textit{for all} \quad v \in G(S),$$

where $G(S) := \{v \in \mathbb{R}^p : \|v_{S^c}\|_1 \leqslant \kappa_0 \|v_S\|_1\}$.

Condition 1 requires a restriction of $X^{\mathrm{T}}X/n$ to the columns in S is invertible. This assumption is widely used to bound the l_2-error between estimates and coefficients [20, 21]. Besides, we need the approval that the proposed estimator satisfies the requirement of Condition 1, i.e., set $n^{1/2}(\hat{\beta} - \beta) \in G(S)$. This result has been proved for the lasso estimator [4]. Compared to the lasso, we require another bound for the Laplace matrix L:

$$\begin{aligned}\beta^{\mathrm{T}} L \beta - \hat{\beta}^{\mathrm{T}} L \hat{\beta} &= \beta^{\mathrm{T}} L \beta - (\hat{\beta} - \beta + \beta)^{\mathrm{T}} L (\hat{\beta} - \beta + \beta)\\ &= -(\hat{\beta} - \beta)^{\mathrm{T}} L (\hat{\beta} - \beta) - 2\beta^{\mathrm{T}} L (\hat{\beta} - \beta)\\ &< 2\|\beta^{\mathrm{T}} L\|_\infty \|\hat{\beta} - \beta\|_1 \leqslant \lambda_1/\lambda_2 \cdot \|\hat{\beta} - \beta\|_1.\end{aligned}$$

Combining above upper bound and Lemma 6.3 of Bühlmann and Van De Geer [4], we conclude $n^{1/2}(\hat{\beta} - \beta) \in G(S)$ for Condition 1 with the same convergence rate, and have the following result:

Theorem 1. *Suppose Condition 1 holds. If $\lambda_1 = 4\sigma\sqrt{\log p/n}$ and $\lambda_2 \Lambda_{\max}(L) \|\beta\|_2 \leqslant \sqrt{q}\lambda_1$. Then there exists a positive constant K that the following holds with probability at least $1 - o(\exp(n^{c_1}))$:*

$$\|\hat{\beta} - \beta\|_2 \leqslant K \sqrt{\frac{q \log p}{n}}.$$

For the sign consistency of the proposed estimation, we require the another condition:

Condition 2. *There exists a positive constant $\eta > 0$ such that*

$$\begin{aligned}&\big\|\big(C_{S^c} + \lambda_2 L_{S^c}\big)\big(C_S + \lambda_2 L_S\big)^{-1}\big[\gamma_S + \frac{\lambda_2}{\lambda_1} L_S \beta_S\big]\\ &\quad - \frac{\lambda_2}{\lambda_1} L_{S^c} \beta_S\|_\infty < 1 - \eta,\end{aligned} \tag{2}$$

where L_S and L_{S^c} denote the matrices with rows and columns of L indexed by $S \times S$ and $S^c \times S$, respectively.

Condition 2 is quite similar to the irrepresentable condition for the lasso [31] and elastic irrepresentable condition for the elastic net [15]. When $L = I$, the above condition equals the elastic irrepresentable condition; when $\lambda_2 = 0$, the above condition equals the irrepresentable condition. There exists λ_2 and L that Condition 2 holds but the irrepresentable condition or the elastic irrepresentable condition cannot. It indicates that Condition 2 is weaker than the both conditions, and thus the proposed method performs better than many other l_1-based methods. The following Theorem shows that with high probability the proposed method is equal in sign with the true model.

Theorem 2. *Suppose Condition 2 holds. If $\lambda_1 = 4\sigma\sqrt{\log p/n}$, $\lambda_2 \Lambda_{\max}(L)\|\beta\|_2 \leqslant \sqrt{q}\lambda_1$ and $\min_{j \in S} |\beta_j| \geqslant 8\sigma \Lambda_{\min}^{-1}(C_S)\sqrt{q \log p/n}$. We have with probability at least $1 - o(\exp(-n^{c_1}))$ that*

$$\mathrm{sign}(\hat{\beta}) = \mathrm{sign}(\beta).$$

4. SIMULATIONS

In this part, we conduct simulation experiments to test the performance of the proposed method in two aspects: estimation accuracy and variable selection. We compare the proposed method with the elastic net [33], MCP [29] and SLS [14]. The R glmnet package [9] is used to run the elastic net, and the results of MCP and SLS are based on the R ncvreg package [3]. We simulate 100 times for each setting.

We set the sample size as 200, the number of predictors as 400. The simulation is based on the following linear regression model:

$$y = X\beta + \epsilon,$$

where $\epsilon \sim N(0, I)$. The predictors are divided into 80 groups, each group has 5 predictors. Among the four hundred predictors, only the first ten predictors have non-zero coefficients. We consider the following three scenarios on different non-zero coefficient settings, i.e., Scenario 1: all the non-zero coefficients are set to be 5; Scenario 2: the first five non-zero coefficients are 5 and the next five are -5; and Scenario 3: the first five non-zero coefficients are 3 and the next five are 5. We consider two kinds of covariance matrices of X:

> Example 1: For the first two groups, predictors within-group are correlated: correlations equal 0.9, i.e., Example 1(a) and $0.9^{|i-j|}$, i.e., Example 1(b). There is no correlation between the two groups, and the rest predictors are independent too.
> Example 2: For each group, predictors within-group are correlated: correlations equal 0.9, i.e., Example 2(a) and $0.9^{|i-j|}$, i.e., Example 2(b). Predictors in different groups are independent.

We use the following four ways to construct the adjacency matrices. Let $a_{jj'}$ be the element of the adjacency matrix in the jth row and j'th column, where $j, j' = 1, \dots, p$:

AdaSLS.1: $a_{jj'} = I(r_{jj'} > c)$, where $c = 0.7$, and $r_{jj'} = cor(X_j, X_{j'})$ denotes the Pearson's correlation coefficient between the jth predictor and the j'th predictor.

AdaSLS.2: $a_{jj'} = I(d_{jj'} > c)$, where $c = 0.7$, and $d_{jj'}$ denotes the distance correlation coefficient between the jth predictor and the j'th predictor.

AdaSLS.3: $a_{jj'} = I(S_{jj'} > c)$ and $S = \mathrm{Cov}(x) = (S_{jj'}, 1 \le j, j' \le p)$ denotes the sample covariance matrix and c is computed by the covariance matrix.

AdaSLS.4: Use the correlation of the partial residual to construct the adjacency matrix and update the adjacency matrix constantly in iterations.

According to our numerical experience, estimation results do not differ much over a range of threshold value c, i.e., valued in $[0.3, 0.8]$, under these scenarios. Thus, we fix $c = 0.7$ in AdaSLS.1 and AdaSLS.2. We choose the parameters λ_1 and λ_2 by 5-fold cross-validation: set the range of λ_1 and λ_2 respectively and then choose the optimal $(\lambda_1^*, \lambda_2^*)$ with minimum prediction error. We compare both the estimation and selection performances of each method. The mean and standard deviation of results are reported in Table 1-3. The l_1-error and l_2-error of the estimating coefficients, the number of nonzero estimating coefficients (NZ), the true positive rate (TPR), and false positive rate (FPR) of variable selection are respectively defined as:

$$l_1\text{-error} = \sum_{j=1}^{p} |\beta_j - \widehat{\beta}_j|, \; l_2\text{-error} = \Big(\sum_{j=1}^{p} (\beta_j - \widehat{\beta}_j)^2\Big)^{1/2},$$

$$\mathrm{NZ} = |j \in \{1, 2, \cdots, p\} : \widehat{\beta}_j \neq 0|,$$

$$\mathrm{TPR} = \frac{|j \in \{1, 2, \cdots, p\} : \widehat{\beta}_j \neq 0 \text{ and } \beta_j \neq 0|}{|j \in \{1, 2, \cdots, p\} : \beta_j \neq 0|},$$

$$\mathrm{FPR} = \frac{|j \in \{1, 2, \cdots, p\} : \widehat{\beta}_j \neq 0 \text{ and } \beta_j = 0|}{|j \in \{1, 2, \cdots, p\} : \beta_j = 0|}.$$

Table 1 shows the comparison between four versions of AdaSLS and Elastic Net, MCP, and SLS in each example under Scenario 1. As one can see, AdaSLS performs well in both estimation and selection. All four versions, AdaSLS estimates the parameters accurately with lower error and lower standard deviation. They select the models with all the true positives and no false positives. In contrast, MCP misses lots of true positives; elastic net and SLS both identify all true positives. However, the compared methods always pick up some false positives. When the coefficients have different signs, as shown in Table 2, AdaSLS performs well. All four versions of AdaSLS obtain lower estimation error. AdaSLS.2 and AdaSLS.4 have the best performance in variable selection. Their adjacency matrices are constructed by the distance correlation of the predictors and the correlation of the partial residual. Both kinds of AdaSLS identify all true positives and false positives. Under Scenario 3, the signal-to-noise becomes smaller, from 5 to 3. As one can see in Table 3, AdaSLS still has a good performance in both estimation and selection for four kinds of methods.

To further show the performance in estimating non-zero coefficients, Figure 1 compares the l_1 estimation error of $\beta : |\beta_j - \widehat{\beta}_j|$ for $\beta_j \neq 0$ $(j = 1, \cdots, 10)$ under three scenarios and two examples. The results for example (b) are similar as those for example (a) and thus omitted. Since the errors of MCP are much higher than the other methods, we compare the performance of the elastic net, SLS, and AdaSLS.1-.4. The red line denotes the mean l_1 error of the elastic net and its $25\% - 75\%$ interval. The orange line denotes the mean l_1 error of the SLS and its $25\% - 75\%$ interval. The rest line denotes the mean l_1 error of AdaSLS.1-.4., respectively. As one can see from Figure 1, the l_1 error of the elastic net and the SLS are always larger than those of AdaSLS. Results of AdaSLS.1-.4. are similar, and all are better than those of the SLS and elastic net. The mean value of AdaSLS is about 0.01, and the range of $25\% - 75\%$ is about $0.003 - 0.02$ under Scenario 1 and 2. Under Scenario 3 which is a smaller signal-to-noise case, the estimation error become a little larger, the mean value of AdaSLS is between 0.01 and 0.02, and the range of $25\% - 75\%$ is about $0.006 - 0.03$ (The above intervals are not shown in the figures.). Those results indicate that AdaSLS performs much better in estimating non-zero coefficients.

5. EMPIRICAL ANALYSIS

In this section, we apply the proposed method to two datasets: gene expression data and protein data. The gene expression data is a rat eye gene microarray and used to find the gene expressions that are related to the gene expression of the TRIM32 gene. The protein data contains the thermostability of proteins and some structural features of the P450 proteins. The AdaSLS method performs well in both two datasets.

5.1 Gene microarray dataset

Firstly, we apply the proposed method to a gene expression problem. In this part, we apply the proposed method to a rat eye gene microarray dataset. The data is from the GEO database[1] and is provided by Scheetz et al. [25]. This data is commonly used in mammalian eye gene expression and related diseases studies, and is also used as an empirical part of some research for gene microarray data. SR/JrHsd (a kind of salt-resistant rat) males were crossed with SHRSP (Spontaneous Hypertensive Rat-Stroke Prone strain) females to generate F1 and F2 generation animals [25]. 120 12-week-old male F2 offspring were selected as samples and there are

[1] Gene Expression Omnibus, www.ncbi.nlm.nih.govgeo, (accession no. GSE5680)

Table 1. Performance comparison in each example under Scenario 1.

Example	Method	l_1-error	l_2-error	NZ	TPR	FPR
Example 1(a)	Elastic net	1.756(0.440)	0.674(0.164)	10.570(1.157)	1	0.001(0.003)
	MCP	53.034(6.902)	17.241(2.091)	4.840(0.788)	0.454(0.072)	0.001(0.002)
	SLS	0.321(0.345)	0.107(0.070)	12.83(5.965)	1	0.007 (0.015)
	AdaSLS.1	0.146(0.081)	0.051(0.027)	10	1	0
	AdaSLS.2	0.137(0.079)	0.047(0.027)	10	1	0
	AdaSLS.3	0.138(0.079)	0.048(0.026)	10	1	0
	AdaSLS.4	0.136(0.066)	0.049(0.022)	10	1	0
Example 1(b)	Elastic net	1.850(0.533)	0.701(0.189)	10.720(1.102)	1	0.002(0.003)
	MCP	45.480(7.347)	15.325(2.338)	5.450(0.770)	0.538(0.079)	0.0002(0.001)
	SLS	0.321(0.391)	0.107(0.088)	12.42(5.819)	1	0.006 (0.015)
	AdaSLS.1	0.128(0.053)	0.047(0.018)	10	1	0
	AdaSLS.2	0.134(0.049)	0.051(0.017)	10	1	0
	AdaSLS.3	0.126(0.061)	0.046(0.021)	10	1	0
	AdaSLS.4	0.161(0.051)	0.061(0.019)	10	1	0
Example 2(a)	Elastic net	1.757(0.439)	0.681(0.166)	10.320(0.649)	1	0.001(0.002)
	MCP	54.438(6.019)	17.661(1.822)	4.670(0.766)	0.440(0.062)	0.001(0.001)
	SLS	0.293(0.077)	0.113(0.027)	10.35(1.149)	1	0.001 (0.003)
	AdaSLS.1	0.122(0.071)	0.044(0.024)	10	1	0
	AdaSLS.2	0.119(0.064)	0.043(0.023)	10	1	0
	AdaSLS.3	0.175(0.080)	0.063(0.027)	10	1	0
	AdaSLS.4	0.137(0.079)	0.048(0.027)	10	1	0
Example 2(b)	Elastic net	1.801(0.534)	0.689(0.197)	10.460(0.797)	1	0.001(0.002)
	MCP	45.834(6.844)	15.457(2.242)	5.420(0.727)	0.535(0.072)	0.0002(0.001)
	SLS	0.290(0.101)	0.111(0.033)	10.46(1.772)	1	0.007 (0.015)
	AdaSLS.1	0.141(0.066)	0.051(0.022)	10	1	0
	AdaSLS.2	0.151(0.073)	0.056(0.025)	10	1	0
	AdaSLS.3	0.188(0.073)	0.069(0.025)	10	1	0
	AdaSLS.4	0.162(0.060)	0.062(0.022)	10	1	0

Table 2. Performance comparison in each example under Scenario 2.

Example	Method	l_1-error	l_2-error	NZ	TPR	FPR
Example 1(a)	Elastic net	1.723(0.388)	0.659(0.164)	10.21(0.518)	1	0.001(0.001)
	MCP	53.636(6.001)	17.447(2.091)	4.67(0.766)	0.448(0.062)	0.0005(0.001)
	SLS	0.224(0.059)	0.084(0.116)	10.25(1.104)	1	0.001 (0.003)
	AdaSLS.1	0.142(0.057)	0.052(0.019)	10	1	0
	AdaSLS.2	0.137(0.059)	0.050(0.020)	10	1	0
	AdaSLS.3	0.129(0.055)	0.046(0.019)	10	1	0
	AdaSLS.4	0.142(0.057)	0.052(0.020)	10	1	0
Example 1(b)	Elastic net	1.780(0.508)	0.683(0.174)	10.38(0.962)	1	0.001(0.002)
	MCP	53.536(6.081)	17.674(1.843)	4.69(0.748)	0.439(0.063)	0.001(0.001)
	SLS	0.355(0.111)	0.136(0.042)	10.29(0.743)	1	0.001 (0.002)
	AdaSLS.1	0.170(0.190)	0.062(0.056)	10.05(0.5)	1	0.0001(0.001)
	AdaSLS.2	0.171(0.062)	0.065(0.022)	10	1	0
	AdaSLS.3	0.152(0.061)	0.057(0.020)	10	1	0
	AdaSLS.4	0.126(0.068)	0.045(0.022)	10	1	0
Example 2(a)	Elastic net	1.791(0.440)	0.683(0.159)	10.2(0.603)	1	0.001(0.002)
	MCP	53.763(6.189)	17.468(1.865)	4.68(0.737)	0.447(0.064)	0.001(0.001)
	SLS	0.224(0.263)	0.087(0.074)	11.87(4.334)	1	0.005 (0.011)
	AdaSLS.1	0.136(0.072)	0.048(0.025)	10	1	0
	AdaSLS.2	0.144(0.081)	0.050(0.028)	10	1	0
	AdaSLS.3	0.154(0.078)	0.055(0.027)	10	1	0
	AdaSLS.4	0.129(0.053)	0.048(0.019)	10	1	0
Example 2(b)	Elastic net	1.694(0.445)	0.657(0.163)	10.3(0.674)	1	0.001(0.002)
	MCP	53.854(6.389)	17.493(1.939)	4.65(0.743)	0.447(0.069)	0.00005(0.001)
	SLS	0.422(0.269)	0.147(0.070)	12.42(1.772)	1	0.006 (0.012)
	AdaSLS.1	0.210(0.509)	0.073(0.168)	10.093(0.914)	1	0.0002(0.002)
	AdaSLS.2	0.130(0.058)	0.046(0.020)	10	1	0
	AdaSLS.3	0.217(0.460)	0.078(0.142)	10.082(0.812)	1	0.0002(0.002)
	AdaSLS.4	0.160(0.069)	0.056(0.024)	10	1	0

Table 3. Performance comparison in each example under Scenario 3.

Example	Method	l_1-error	l_2-error	NZ	TPR	FPR
Example 1(a)	Elastic net	1.787(0.482)	0.678(0.176)	10.91(0.518)	1	0.002(0.004)
	MCP	45.040(6.370)	15.162(2.117)	5.49(0.659)	0.542(0.068)	0.0002(0.001)
	SLS	0.231(0.059)	0.089(0.018)	10.20(1.104)	1	0.0005 (0.001)
	AdaSLS.1	0.138(0.075)	0.046(0.023)	10	1	0
	AdaSLS.2	0.113(0.063)	0.041(0.022)	10	1	0
	AdaSLS.3	0.138(0.073)	0.046(0.022)	10	1	0
	AdaSLS.4	0.133(0.078)	0.044(0.024)	10	1	0
Example 1(b)	Elastic net	1.911(0.494)	0.726(0.178)	10.89(1.034)	1	0.002(0.003)
	MCP	45.277(6.452)	15.255(2.11)	5.49(0.689)	0.541(0.068)	0.0002(0.001)
	SLS	0.343(0.336)	0.114(0.087)	13.39(5.682)	1	0.001 (0.015)
	AdaSLS.1	0.180(0.384)	0.065(0.123)	10.17(1.340)	1	0.0004(0.003)
	AdaSLS.2	0.132(0.063)	0.046(0.020)	10	1	0
	AdaSLS.3	0.179(0.382)	0.065(0.123)	10.18(1.343)	1	0.0005(0.003)
	AdaSLS.4	0.142(0.066)	0.050(0.030)	10	1	0
Example 2(a)	Elastic net	1.848(0.566)	0.695(0.203)	10.81(1.161)	1	0.002(0.003)
	MCP	45.854(7.224)	15.397(2.398)	5.40(0.710)	0.447(0.064)	0.0002(0.001)
	SLS	0.356(0.269)	0.129(0.076)	12.3(4.140)	1	0.005 (0.011)
	AdaSLS.1	0.103(0.046)	0.040(0.019)	10	1	0
	AdaSLS.2	0.252(1.14)	0.078(0.023)	10.128(1.186)	1	0.0003(0.003)
	AdaSLS.3	0.126(0.063)	0.046(0.025)	10	1	0
	AdaSLS.4	0.206(0.083)	0.069(0.029)	10	1	0
Example 2(b)	Elastic net	1.764(0.417)	0.663(0.155)	10.91(1.296)	1	0.002(0.003)
	MCP	46.080(6.389)	15.472(2.430)	5.41(0.753)	0.532(0.080)	0.00002(0.001)
	SLS	0.376(0.380)	0.124(0.089)	14.6(7.531)	1	0.011 (0.019)
	AdaSLS.1	0.171(0.063)	0.064(0.022)	10	1	0
	AdaSLS.2	0.332(0.90)	0.109(0.241)	10.2(1.078)	1	0.001(0.002)
	AdaSLS.3	0.159(0.088)	0.053(0.027)	10	1	0
	AdaSLS.4	0.226(0.097)	0.075(0.029)	10	1	0

more than 31000 different probes in the RNA from the eyes of these F2 animals.

The TRIM32 gene causes Bardet-Biedl syndrome [5], which is a genetically heterogeneous disease of multiple organ systems including the retina. We plan to identify the genes that are associated with TRIM32 gene expression and have the greatest changes in gene expression across samples. Since the number of genes associated with TRIM32 is expected to be small, and we are mainly interested in the genes whose expression values varied greatly in the samples, the data is preprocessed as follows. We calculate the variance of gene expression and select the top 1000 genes with the largest variance, then centralize the TRIM32 gene expression and standardize other gene expressions.

We analyze the data by SLS and the proposed procedure. We use different measures to construct the adjacency matrix, leading to different gene identifications. As expected, the list of genes identified by AdaSLS is shorter than that of SLS, which makes a leaner model. As we don't know the real model as we do in simulation studies, we can't evaluate the true positives and false positives of variable selection. We evaluate the performance through the following measure. The data are divided randomly into 5 subsets of the same size. One of the subsets is chosen as the test data and the rest is used to select parameters by cross-validation and estimate models. Then we use the estimated model to predict the test data. Repeat for all subsets as the test data and calculate the prediction error. The sum of squared predictions' error is also shown in Table 4. As we can see from the result, AdaSLS has a smaller prediction error.

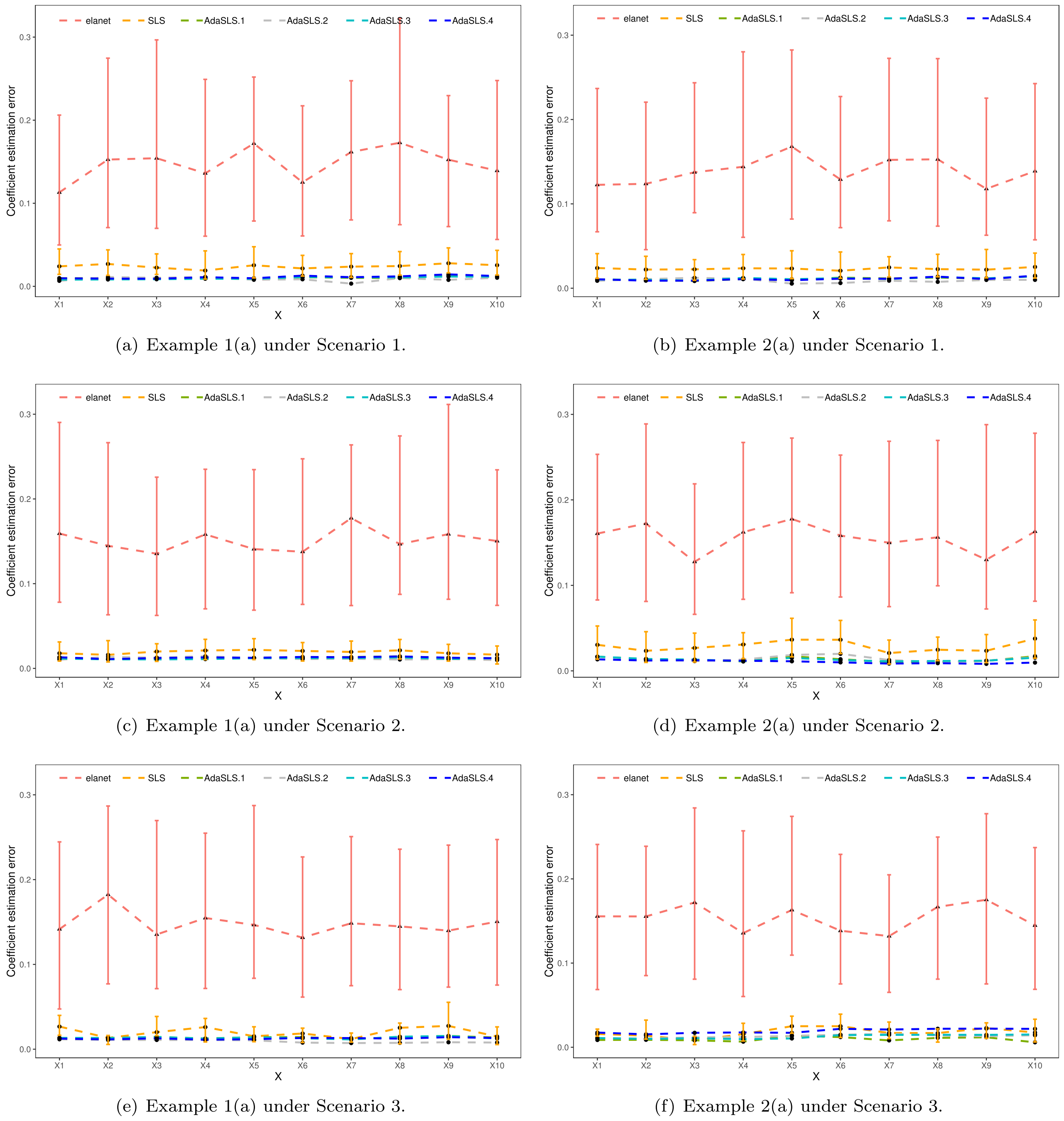
Table 4. Result of Empirical Analysis

Method	Number of selected genes	Prediction errors
SLS	25	2.488
AdaSLS.1	18	2.348
AdaSLS.3	17	2.462
AdaSLS.4	21	2.466

5.2 Chimeric protein dataset

In this part, we apply the proposed method to the protein data. The data is provided by the Romero Lab at UW-Madison[2]. It includes the thermostability and structural features of chimeric P450 proteins with eight block chimeras. Protein analysis is an analytical technique commonly used in biochemistry. The cytochrome P450 proteins are a kind of widely used biological catalysts. It is widely used in pharmaceutical products and other useful compounds for the production of synthetic since it can catalyze many reactions [10]. The thermostability of proteins is of great industrial importance because of their ability to withstand difficult industrial process conditions [22]. The structure features are

[2] Data is in https://github.com/RomeroLab/seq-fcn-data/tree/master/P450_chimeras.



(a) Example 1(a) under Scenario 1.

(b) Example 2(a) under Scenario 1.

(c) Example 1(a) under Scenario 2.

(d) Example 2(a) under Scenario 2.

(e) Example 1(a) under Scenario 3.

(f) Example 2(a) under Scenario 3.

Figure 1. The l_1 estimation error of β: $|\beta_j - \widehat{\beta}_j|$ for $\beta_j \neq 0$ where $j = 1, \ldots, 10$.

simulated via RosettaCommons [1]. The Rosetta biomolecular modeling suite modeled 3-D structures of the chimeric enzymes and then estimated the structural properties of each protein [1]. What we are interested in is how the structural properties of proteins relate to the thermostability of proteins. There are 988 chimeric P450 proteins 50 structure features in the feature data while the thermostability data only includes 242 chimeric P450 proteins. So we match the feature data into the thermostability data by the sequence information of these proteins as the pre-processing. Then we standardized the thermostability and feature data.

Similar to the previous analysis, five subsets are generated. The prediction error of each subset is derived from the model trained by the remaining subsets and the performance is evaluated by the prediction error. The mean of squared prediction errors are SLS:23.44, AdaSLS.1:22.63, AdaSLS.2:21.83, AdaSLS.3:22.27, and AdaSLS.4:22.58. The mean of squared error of AdaSLS is smaller than those of other methods.

6. SUMMARY

In this paper, we consider the problem of estimating the high-dimensional data with correlation structures. To overcome the difficulties in estimation accuracy, we propose the AdaSLS algorithm (Adaptive sparse Laplacian shrinkage) for the variable selection and estimation in the sparse regression models with correlation structure. This algorithm contains two penalties: an adaptive weighted l_1 penalty and the Laplace quadratic penalty. The weighted l_1 penalty term is similar to that in adaptive lasso, but the weight is updated constantly during the iteration process. In this way, we effectively reduce the estimation bias. The Laplace quadratic penalty contains the information of the network structure of the related predictors in the model and enhances the smoothness among coefficients of correlated predictors. Furthermore, different measures are used to construct the adjacency matrix, including the Pearson correlation coefficient, the distance correlation coefficient, and the partial residual. We show that under regular conditions, the proposed method achieves sign consistency and the error bound. We evaluated the method with simulations and two datasets in empirical analysis. Simulations show that the AdaSLS demonstrates the best performance in both estimation and selection. It estimates the model accurately with low error and low standard deviation.

AdaSLS has good performance in correlated data which often appears in reality, e.g., gene data analysis, medical data analysis, and financial data. As we analyzed in the empirical study, gene expressions that are associated with a biological trait are sparse and always highly correlated, in the meantime, the number of samples is usually much smaller than the number of variables since samples are not easy to get. Similarly, in medical data analysis, it is also not easy to obtain samples for people who may not be willing to be studied as a sample and each sample requires a certain amount of examination costs. There are usually various variables that are related to each sample. We have shown that the proposed method obtained a sparser model with smaller prediction error, obtaining meaningful results. The study of our method is of practical significance to deal with this kind of data in reality.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 12001557); Youth Talent Development Support Program (QYP202104), Emerging Interdisciplinary Project, and School of Statistics and Mathematics in Central University of Finance and Economics.

APPENDIX

Proof of Theorem 1. Note that

$$L(\beta;\lambda_1,\lambda_2) = \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda_1\sum_{j=1}^{p} w_j|\beta_j| + \frac{1}{2}\lambda_2\beta^{\mathrm{T}}L\beta,$$

Set $u = \hat{\beta} - \beta$ and $F(u) = L(\hat{\beta};\lambda_1,\lambda_2) - L(\beta;\lambda_1,\lambda_2)$. By the definition of $\hat{\beta}$, we have $F(u) \leqslant 0$ and let

$$F(u) = V_1 + V_2 + V_3,$$

that

$$\begin{aligned}
V_1 &= \frac{1}{2n}(\|y - X\hat{\beta}\|_2^2 - \|y - X\beta\|_2^2),\\
V_2 &= \lambda_1\sum_{j=1}^{p} w_j|\hat{\beta}_j| - \lambda_1\sum_{j=1}^{p} w_j|\beta_j|,\\
V_3 &= \frac{1}{2}(\lambda_2\hat{\beta}^{\mathrm{T}}L\hat{\beta} - \lambda_2\beta^{\mathrm{T}}L\beta).
\end{aligned}$$

Denote $C = X^{\mathrm{T}}X/n$, $W = X^{\mathrm{T}}\epsilon/\sqrt{n}$. We have

$$V_1 = \frac{1}{2}u^{\mathrm{T}}Cu - u^{\mathrm{T}}W/\sqrt{n}$$

Denote u_S, β_S and u_{S^c}, β_{S^c} as the partition of u and β indexed by the set S and S^c. We have

$$V_1 = \frac{1}{2}u^{\mathrm{T}}Cu - u_S^{\mathrm{T}}W_S/\sqrt{n} - u_{S^c}^{\mathrm{T}}W_{S^c}/\sqrt{n},$$

$$V_2 \geqslant -\lambda_1\sum_{j\in S} w_j|u_j| + \lambda_1\sum_{j\in S^c} w_j|u_j|,$$

and

$$\begin{aligned}
V_3 &= \frac{\lambda_2}{2}\{(\hat{\beta} - \beta)^{\mathrm{T}}L(\hat{\beta} - \beta) + 2\hat{\beta}^{\mathrm{T}}L\beta - 2\beta^{\mathrm{T}}L\beta\},\\
&= \frac{\lambda_2}{2}\{u^{\mathrm{T}}Lu + 2u^{\mathrm{T}}L\beta\},\\
&\geqslant \frac{\lambda_2}{2}\Lambda_{\min}(L)\|u\|_2^2 - \lambda_2\Lambda_{\max}(L)\big|u_S^{\mathrm{T}}\beta_S\big|.
\end{aligned}$$

Given w_j, $j = 1, \ldots, p$ and conditional on $\{\|W\|_\infty \leqslant K_0\sqrt{n}\lambda_1\}$ with a positive constant K_0, we have

$$u_{S^c}^{\mathrm{T}}W_{S^c}/\sqrt{n} \leqslant \|u_{S^c}\|_1\|W\|_\infty/\sqrt{n} \leqslant \lambda_1\sum_{j\in S^c} w_j|u_j|,$$

$$u_S^{\mathrm{T}}W_S/\sqrt{n} \leqslant \|u_S\|_2\|W_S\|_2/\sqrt{n} \leqslant \lambda_1\sqrt{q}\|u_S\|_2,$$

and

$$\lambda_1\sum_{j\in S} w_j|u_j| \leqslant \lambda_1\sqrt{q}\|u\|_2.$$

For $\lambda_2\Lambda_{\max}(L)\|\beta\|_2 \leqslant \sqrt{q}\lambda_1$, we have

$$\big|\lambda_2\Lambda_{\max}(L)u_S^{\mathrm{T}}\beta_S\big| \leqslant \lambda_2\Lambda_{\max}(L)\|u\|_2\|\beta\|_2 \leqslant \sqrt{q}\lambda_1\|u\|_2.$$

Combining with RE condition, $F(u)$ thus becomes:

$$\begin{aligned}
F(u) \geqslant{}& \frac{1}{2}\kappa\|u\|_2^2 - u_S^{\mathrm{T}}W_S/\sqrt{n} - \lambda_1\sum_{j\in S} w_j|u_j|\\
&+ \frac{\lambda_2}{2}\Lambda_{\min}(L)(\|u\|_2^2 - 2|u_S^{\mathrm{T}}\beta_S|)
\end{aligned}$$

$$\begin{aligned}&\geqslant \|u\|_2\Big\{\frac{1}{2}\big(\kappa+\lambda_2\Lambda_{\min}(L)\big)\|u\|_2 - 3\sqrt{q}\lambda_1\Big\}\\ &\geqslant 0.\end{aligned}$$

Then by $\lambda_1 = 4\sigma\sqrt{\log p/n}$, we have conditional on $\{\|W\|_\infty \leqslant K\sqrt{n}\lambda_1\}$ that

$$\|\hat\beta-\beta\|_2 = \|u\|_2 \leqslant \frac{6\sqrt{q}\lambda_1}{\kappa+\lambda_2\Lambda_{\min}(L)} = K\sqrt{q\log p/n},$$

where K is a positive constant. We use the tail probability bound of Gaussian distribution for $\{\|W\|_\infty \leqslant K\sqrt{n}\lambda_1\}$:

$$\begin{aligned}P(\|W\|_\infty > K_0\sqrt{n}\lambda_1) &\leqslant P(\|X^{\mathrm T}\epsilon/\sqrt{n}\|_\infty > 4K_0\sigma\sqrt{\log p})\\ &\leqslant p\exp(-\frac{2K_0\sigma^2\log p}{\sigma^2}) = o(e^{-n^{c_1}}).\end{aligned}$$

□

Proof of Theorem 2. By Karush-Kuhn-Tucker conditions, $\hat\beta$ is optimal if and only if

$$\frac{X^{\mathrm T}X\hat\beta}{n} - \frac{X^{\mathrm T}y}{n} + \lambda_2L\hat\beta + \lambda_1\gamma = 0, \tag{3}$$

$$\gamma_j \in \begin{cases}\{\mathrm{sign}(\hat\beta_j)\}, & \hat\beta_j \neq 0,\\ [-1,1], & \text{otherwise.}\end{cases}$$

To prove $\mathrm{sign}(\hat\beta) = \mathrm{sign}(\beta)$ with high probability is suffices to prove $\hat\beta_{S^c} = 0$ and $|u_S| = |\hat\beta_S - \beta_S| < |\beta_S|$. Following the same notation above, i.e., $C = X^{\mathrm T}X/n$, $W = X^{\mathrm T}\epsilon/\sqrt{n}$, further, we set $C_S = X_S^{\mathrm T}X_S/n$ and $C_{S^c} = X_{S^c}^{\mathrm T}X_S/n$. Combine with (3), we need to have

$$\begin{aligned}&\frac{X^{\mathrm T}X\hat\beta}{n} - \frac{X^{\mathrm T}y}{n} + \lambda_2L\hat\beta\\ &= (C_S+\lambda_2L_S)\hat\beta_S - C_S\beta_S - W_S/\sqrt{n}\\ &= (C_S+\lambda_2L_S)u_S + \lambda_2L_S\beta_S - W_S/\sqrt{n} = -\lambda_1\gamma_S \qquad (4)\end{aligned}$$

and

$$\begin{aligned}&\big|(C_{S^c}+\lambda_2L_{S^c})\hat\beta_S - X_{S^c}^{\mathrm T}X_S\beta_S - X_{S^c}^{\mathrm T}\epsilon\big|\\ &= \big|(C_{S^c}+\lambda_2L_{S^c}))u_S + \lambda_2L_S\beta_S - W_S/\sqrt{n}\big| \leqslant |\lambda_1|. \qquad (5)\end{aligned}$$

By $|u_S| < |\beta_S|$, it becomes to prove the following inequality,

$$\begin{aligned}&(C_S+\lambda_2L_S)^{-1}W_S \leqslant \sqrt{n}|\beta_S|\\ &-\sqrt{n}(C_S+\lambda_2L_S)^{-1}\big(\lambda_1\gamma_S+\lambda_2L_S\beta_S\big).\end{aligned}$$

For $j \in S$, with $\lambda_1 = 4\sigma\sqrt{\log p/n}$ and $\lambda_2\Lambda_{\max}(L)\|\beta\|_2 \leqslant \sqrt{q}\lambda_1$, we have

$$\begin{aligned}&\big|(C_S+\lambda_2L_S)^{-1}\big(\lambda_1\gamma_S+\lambda_2L_S\beta_S\big)\big|_j\\ &\leqslant 2\Lambda_{\min}^{-1}(C_S)\sqrt{q}\lambda_1 \leqslant 8\sigma\Lambda_{\min}^{-1}(C_S)\sqrt{q\log p/n}.\end{aligned}$$

Then by $\min_{j\in S}|\beta_j| \geqslant 8\sigma\Lambda_{\min}^{-1}(C_S)\sqrt{q\log p/n}$ and the tail probability bound of Gaussian distribution that:

$$\begin{aligned}&P\Big((C_S+\lambda_2L_S)^{-1}W_S \leqslant \sqrt{n}|\beta_S| -\\ &\sqrt{n}(C_S+\lambda_2L_S)^{-1}\big(\lambda_1\gamma_S+\lambda_2L_S\beta_S\big)\Big)\\ &\geqslant 1 - P(\|W_S\|_\infty > M\sigma\sqrt{q\log p})\\ &\geqslant 1 - q\exp(-\frac{M\sigma^2q\log p}{2\sigma^2})\\ &\geqslant 1 - o(e^{-n^{c_1}}).\end{aligned}$$

Considers the inequality (5), combines with (4), we have

$$\begin{aligned}&\big|C_{S^c}\big(C_S+\lambda_2L_S\big)^{-1}W_S/\sqrt{n} - W_{S^c}/\sqrt{n}\big|\\ &\leqslant \lambda_1\big(1 - \big|\big(C_{S^c}+\lambda_2L_{S^c}\big)\big(C_S+\lambda_2L_S\big)^{-1}\big[\gamma_S + \frac{\lambda_2}{\lambda_1}L_S\beta_S\big]\\ &-\frac{\lambda_2}{\lambda_1}L_{S^c}\beta_S\big|\big).\end{aligned}$$

By Condition 2, above inequality can be written as

$$\big|C_{S^c}\big(C_S+\lambda_2L_S\big)^{-1}W_S - W_{S^c}\big| \leqslant \sqrt{n}\lambda_1\eta.$$

Again, by the tail probability bound of Gaussian distribution, we have

$$\begin{aligned}&P\Big(\big|C_{S^c}\big(C_S+\lambda_2L_S\big)^{-1}W_S - W_{S^c}\big| \leqslant \sqrt{n}\lambda_1\eta\Big)\\ &\geqslant 1 - o(e^{-n^{c_1}}).\end{aligned}$$

□

Received 8 August 2021

REFERENCES

[1] Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O'Meara, M. J., DiMaio, F. P., Park, H., Shapovalov, M. V., Renfrew, P. D., Mulligan, V. K., Kappel, K. et al. (2017). The Rosetta all-atom energy function for macromolecular modeling and design. *Journal of chemical theory and computation* **13** 3031–3048.

[2] Bondell, H. D. and Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics* **64** 115–123. MR2422825

[3] Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics* **5** 232–253. MR2810396

[4] Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media. MR2807761

[5] Chiang, A. P., Beck, J. S., Yen, H.-J., Tayeh, M. K., Scheetz, T. E., Swiderski, R. E., Nishimura, D. Y., Braun, T. A., Kim, K.-Y. A., Huang, J. et al. (2006). Homozygosity mapping with SNP arrays identifies TRIM32, an E3 ubiquitin ligase, as a Bardet–Biedl syndrome gene (BBS11). *Proceedings of the National Academy of Sciences* **103** 6287–6292.

[6] Daye, Z. J. and Jeng, X. J. (2009). Shrinkage and model selection with correlated variables via weighted fusion. *Computational Statistics & Data Analysis* **53** 1284–1298. MR2657091

[7] Fan, J. Q. and Li, R. Z. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360. MR1946581

[8] Fan, J., Li, R., Zhang, C.-H. and Zou, H. (2020). *Statistical foundations of data science*. CRC press.

[9] Friedman, J., Hastie, T. and Tibshirani, R. J. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33** 1–22.

[10] Guengerich, F. P. (2002). Cytochrome P450 enzymes in the generation of commercial products. *Nature Reviews Drug Discovery* **1** 359–366.

[11] Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC. MR3616141

[12] Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.

[13] Huang, J., Ma, S., Li, H. and Zhang, C. H. (2011). The sparse Laplacian shrinkage estimator for high-dimensional regression. *Annals of Statistics* **39** 2021–2046. MR2893860

[14] Huang, J., Breheny, P., Lee, S., Ma, S. and Zhang, C. H. (2016). The Mnet method for variable selection. *Statistica Sinica* **26** 903–923. MR3559936

[15] Jia, J. Z. and Yu, B. (2010). On model selection consistency of elastic net when p≫n. *Statistica Sinica* **20** 595–611. MR2682632

[16] Larsen, W. A. and McCleary, S. J. (1972). The use of partial residual plots in regression analysis. *Technometrics* **14** 781–790.

[17] Li, C. Y. and Li, H. Z. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* **24** 1175–1182.

[18] Li, C. Y. and Li, H. Z. (2010). Variable selection and regression analysis for graph-structured covariates with an application to genomics. *Annals of Applied Statistics* **4** 1498–1516. MR2758338

[19] Mazumder, R. and Hastie, T. (2012). Exact covariance thresholding into connected components for large-scale graphical lasso. *Journal of Machine Learning Research* **13** 781–794. MR2913718

[20] Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics* **37** 246–270. MR2488351

[21] Negahban, S., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science* **27** 1348–1356. MR3025133

[22] Niehaus, F., Bertoldo, C., Kähler, M. and Antranikian, G. (1999). Extremophiles as a source of novel enzymes for industrial application. *Applied Microbiology and Biotechnology* **51** 711–729.

[23] Pan, W., Xie, B. and Shen, X. (2010). Incorporating predictor network in penalized regression with application to microarray data. *Biometrics* **66** 474–484. MR2758827

[24] Pearson, K. (1895). VII. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London* **58** 240–242.

[25] Scheetz, T. E., Kim, K.-Y. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J., Casavant, T. L. et al. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences* **103** 14429–14434.

[26] Székely, G. J., Rizzo, M. L., Bakirov, N. K. et al. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics* **35** 2769–2794. MR2382665

[27] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B* **58** 267–288. MR1379242

[28] Tutz, G. and Ulbricht, J. (2009). Penalized regression with correlation-based penalty. *Statistics and Computing* **19** 239–253. MR2516217

[29] Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* **38** 894–942. MR2604701

[30] Zhang, C. H. and Huang, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Annals of Statistics* **36** 1567–1594. MR2435448

[31] Zhao, P. and Yu, B. (2006). On Model Selection Consistency of Lasso. *Journal of Machine Learning Research* **7** 2541–2563. MR2274449

[32] Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association* **101** 1418–1429. MR2279469

[33] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B* **67** 301–320. MR2137327

[34] Zou, H. and Zhang, H. L. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics* **37** 1733–1751. MR2533470

Xingyu Chen
School of Statistics and Mathematics
Central University of Finance and Economics
Bejing, China
E-mail address: 2020211057@email.cufe.edu.cn

Yuehan Yang
School of Statistics and Mathematics
Central University of Finance and Economics
Bejing, China
E-mail address: yyh@cufe.edu.cn