Statistics and Its Interface Volume 15 (2022) 373–382

Multiple penalized regularization for clusters with varying correlation levels*

Wenjun Cao, Lisu Wang†, and Yuehan Yang‡

In this paper, we study the high-dimensional correlated data with multi-level correlations. These data appear frequently in many fields, e.g., genes in gene pathways or stock in industry groups. It motivates us not only to exploit these clusters but also to distinguish the correlation levels. Besides, we analyze the data without pre-specified clustering information to covariates. A two-step method is proposed to address the above problems. The first step focuses on distinguishing the levels and clustering. We aim to divide covariates into sub-vectors, considering both grouping effect and varying correlation. In the second step, we propose a joint estimation and a modified coordinate descent algorithm. The proposed procedure estimates different correlated groups with different penalties. We provide the theoretical guarantees of this method. Numerical comparisons show that the method works effectively on the multi-level correlation structures. We also apply the proposed method to financial data and get interpretable results.

Keywords and phrases: Multi-level correlations, Clustering, Elastic Net, Structured sparsit.

1. INTRODUCTION

Statistical application in the fields of internet and finance can involve extremely large datasets, often coming with complex correlated structures. For example, if we focus on modeling the financial data, S&P 500 indices, it includes ten sectors, e.g., Industrials, Information Technology, Finance, Materials, etc. Each stock is assigned to one of the sectors, providing a grouping structure with 10 groups. Furthermore, each sector has different correlation strength, e.g., stocks in Finance are highly correlated; Material stocks are correlated but not as strong as financial stocks. Generally, in empirical analysis, important and unimportant variables can be correlated; the correlations between covariates can be varying; uncorrelated data and correlated data can be mixed. Statistical and computational methods are required to handle the situations where the data are correlated in various structures [7].

Correlating covariates always cause difficulties in modeling and predicting, such as overfitting and multicollinearity. Penalized methods mitigate these problems by ameliorating the estimation through regularization. The Ridge [12], for instance, balance the bias and variance by shrinking the estimates with an l_2 penalty. The Lasso [24] and the Elastic Net [32], on the other hand, combining with an l_1 penalty, achieve the goal of variable shrinkage while the Ridge cannot. These techniques perform well in general situations [32, 8, 11]. Among, Elastic Net often outperforms the Lasso when there are strongly correlated variables, and this kind of data is common in high-dimensional settings. Many methods are proposed to solve the problems with correlated datasets. Spline-lasso [10] is designed for the data with covariates that can be ordered in some meaningful way. The Combined L-one and Two (CLOT) [1] is introduced for sparse regression and compressed sensing. Yang and Yang [28] proposed an adaptive and reversed penalty to remove the shrinking bias and encourage the group effect. Combining with the nonnegative constraint, the Nonnegative Elastic Net [27] and a two-step method are studied and applied to the constrained index tracking problem in the stock market without short sales. Correlated data with a two-level structure are considered under the Gaussian graphical models [20]. Other extensions include Meier, van de Geer and Bühlmann [15], Tibshirani et al. [25], Hastie, Tibshirani and Wainwright [11], etc.

Recently, researchers have tried far more regularization to exploit groups or clusters information. Group Lasso, proposed by Yuan and Lin [29] encourages the variables within a group to have a shared pattern of sparsity. Motivated by this proposal, the Sparse Group Lasso, proposed by [22], allows sparsity for individual elements within a group; and the Cooperative-lasso Chiquet et al. [3] encourage a shared sign for the nonzero estimates within each group. Further, She [21] proposed the Clustered Lasso method with exact clustering. Witten, Shojaie and Zhang [26] proposed the Cluster Elastic Net which infers clusters of features from the data based on the correlation among covariates as well as association with the response. Tan, Witten and Shojaie [23] proposed the Cluster Graphical Lasso for improved estimation of Gaussian graphical models.

*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, the Program for Innovation Research, the Disciplinary Funding, and the School of Statistics and Mathematics in Central University of Finance and Economics.

†W. Cao and L. Wang contributed equally to this work.
‡Corresponding author.

Above group or cluster penalized procedures exploit external information about the covariates to potentially obtain more accurate results. However, they do not consider the different degrees of correlation levels between groups. Note that practical problems often come with complicated correlations in data. Financial data for instance, often have correlations in varying degrees among hundreds of stocks. Companies involved in the same industry have strong correlations, while some companies owned by one cartel show weak correlations between each other. Similarly, the bank lending data have intricate correlations ranging from the service information, including credit rating and loan record, to personal information and market conditions. These kinds of data usually include strong correlated, weak correlated, and non-correlated groups simultaneously. To address these limitations, this paper focuses on the supervised clustering problem considering multi-level correlations. Besides, motivated by Witten, Shojaie and Zhang [26], we do not assume that the clusters are known a priori. Instead, to obtain the cluster information to more accurately perform the regression, we measure the correlation between covariates as well as association with the response.

We propose a two-step estimation, named Multiple penalized regularization (MPR), to estimate the regression models allowing multiple correlation levels among covariates. We first propose a simple algorithm to cluster the covariates based on their correlations as well as association with the response. According to the correlation levels, the proposed method divides covariates into different clusters including strong correlation clusters, weak correlation clusters, and non-correlation clusters, and then, we estimate each cluster with specific shrinkage and estimate all the coefficients simultaneously. For the penalties on different clusters, we use combinations of the penalties from the Elastic Net, Lasso, and ridge, controlling by two parameters. Single l_2 regularization is germane for the extremely strong correlated variable groups which are seldom sparse, i.e., these groups are usually dense while l_2 penalty encourages grouping effect and reduces multicollinearity. Regularization containing both l_1 and l_2 penalty performs well in correlated variable groups: it simultaneously does variable selection and continuous shrinkage. For the non-correlated groups, since they are non-correlated to the response too, a single l_1 penalty is apposite for covariates in these groups are more likely irrelevant. When processing data with multiple correlations between variables, the proposed method performs well in both low-dimensional or high-dimensional situations compared to various existing methods, such as Lasso, SCAD, MCP [30], etc. Simulations and empirical results show that the MPR is more stable and accurate. Furthermore, we apply the proposed method to financial modeling.

This paper is organized as follows. Section 2 presents the model, method, and algorithm. Section 3 shows the theoretical properties and Section 4 shows the simulations results. The application in Section 5 assesses the performance of the proposed method in financial modeling. We conclude in Section 6. Technical details are provided in the Appendix.

2. METHOD

In this section, we present the details of the proposed method and the related algorithm. Consider the linear regression problem:

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

where y is an n response vector, $X = (X_1, \ldots, X_p)^{\mathrm{T}} \in \mathbb{R}^{n\times p}$, β is the p-dimensional regression coefficient and ϵ is the error vector that $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, n$. We are interested in the case where $n \ll p$ and β has at most q nonzero elements. p and q are allowed to grow with n, and we do not index them with n for notational simplicity. We consider the cases where the predictors are complicatedly correlated. This kind of data appears frequently. Financial data for example, when we consider the index tracking problem, we set the y as market index while X be the returns of p stocks. And the interactions between stocks are intricate.

We consider the data involving different correlated levels among covariates as well as association with the response, i.e., there exist K different correlated levels ($K > 2$). The sample correlations are used to be the correlation measure. A group exists when all the variables in the group are correlated to each other and their sample correlations should be roughly close. We use the Pearson correlation coefficient, a numerical value between -1 and 1, to express the strength of relationship between variables. Let $R = \{r_{jj'}\}_{(p+1)\times(p+1)}$ be the sample correlation coefficient matrix. $|r_{jj'}|$ closes to 1 indicating a strong correlation. The corresponding threshold for different groups are $0 < r^*_{K-1} < \cdots < r^*_1 < 1$. Denote $A_1, \ldots, A_K$ as different clustering groups, sorted high to low levels. The proposed algorithm proceeds as follows:

Algorithm

Clustering. For $k = 1, \ldots, K-1$, initialize $A_k = \{y\}$ and denote $W_k = \{1, ..., p\}/(A_1 \cup \cdots \cup A_{k-1})$. Repeat the following steps:

Step 1. Compute the correlation coefficient matrix $R = \{r_{jj'}\}$ of (X_{W_k}, y).

Step 2. Add j' into A_k when $r^*_k \le |r_{jj'}|$ for $\forall j \in A_k$.

Step 3. When A_k is fixed, remove y from A_k.

Denote $A_K = (\cup_{k=1}^{K-1} A_k)^c$.

Selection and estimation. Obtain the following estimation:q

$$(1) \qquad \hat{\beta} := \arg\min \Big\{ \frac{1}{2}\|y - X\beta\|_2^2 + \sum_{k=1}^{K} P_{A_k}(\beta) \Big\},$$

where $P_{A_k}(\beta) = \lambda_{2,k}\|\beta_{A_k}\|_2^2/2 + \lambda_{1,k}\|\beta_{A_k}\|_1$.

This clustering construction is quite general. In spite of the fact that statistical modeling under the high dimensional

setting always requires sparsity, this assumption is sometimes restrictive. For example, financial returns always depend on the common risk factors, and thus if we consider financial modeling with numerous covariates, there are a considerable proportion of covariates are relevant to the response [6]. It is more natural to consider all of those covariates as important. Our settings fit these data well. Also, it is natural to consider different penalty techniques in different clusters, as they have different grouping effects. Variables in the highly correlated group are more likely toward each other rather than those in the weakly correlated group. The variables in weakly correlated groups are more likely irrelevant to the response. Obtaining these targets requires different strengths of penalties. We will show that the proposed two-step estimation enjoys desirable asymptotic properties with proper choices of tuning parameters.

3. THEORETICAL RESULTS

In this section, we provide the theoretical guarantee for the proposed method. In what follows, we consider the following dimensions, $p = O(\exp(n^{c_1}))$ and $q = O(n^{c_2})$ where $0 \leqslant c_1, c_2 < 1$, p is the number of covariates and q is the number of nonzero elements of *beta*.

We first show results on the error bound of the proposed method. This property requires the following condition:

Condition 1. *Define $C = X^{\mathrm{T}}X/n$. There exists a positive constant κ that*

$$v^{\mathrm{T}}Cv \geqslant \kappa\|v\|_2^2,$$

for all $v \in G(S)$ where $G(S) := \{v \in \mathbb{R}^p : \|v_{S^c}\|_1 \leqslant M\|v_S\|_1\}$ and M is ''a positive constant.

Condition 1 is the Restricted Eigenvalue condition which is usually used to bound the l_2-error between coefficients and estimates [17, 18]. This condition holds with high probability for quite general classes of Gaussian matrices for which the predictors may be highly correlated [19].

Theorem 1. *Suppose Condition 1 holds and assume that there exists positive constants K_1 that $\min(\lambda_1) \propto \max(\lambda_1) \propto K_1\sqrt{n\log p}$ and $\|\beta\|_\infty \leqslant \min(\lambda_1)/4\max(\lambda_2)$ where $\lambda_1 = (\lambda_{1,1}, \ldots, \lambda_{1,K})$ and $\lambda_2 = (\lambda_{2,1}, \ldots, \lambda_{2,K})$. Then with probability as least $1 - o(\exp(-n^{c_1}))$ that*

$$\|\hat{\beta} - \beta\|_2 \leqslant \frac{8\sigma}{\kappa}\sqrt{\frac{q\log p}{n}} \quad \text{and} \quad \|\hat{\beta} - \beta\|_1 \leqslant \frac{24\sigma}{\kappa}\sqrt{q\frac{\log p}{n}}.$$

Remark 1. This result is similar to Negahban et al. [18] which has proved the error bound of the Lasso. More details can be found in the Appendix. Note that correlations among covariates doesn't affect the estimator achieving the theoretical upper bound. This error bound is a general result for penalized regularization. The estimation accuracy of the estimate under correlated data is potentially improved, as shown in simulations and applications.

We then discuss the model selection consistency (sign consistency). Set $S = \{j \in \{1, \ldots, p\} : \beta_j \neq 0\}$ to be the set of nonzero coefficients, $|S| = q$, and $\hat{S} = \{\in \{1, \ldots, p\} : \hat{\beta}_j \neq 0\}$ where $\hat{\beta}$ is the estimate of MPR. Without loss of generality, we assume that $C = X^{\mathrm{T}}X/n$, $C_S = X_S^{\mathrm{T}}X_S/n$ and $C_{S^c} = X_{S^c}^{\mathrm{T}}X_S/n$. Similarly, $\beta_S = \{\beta_j : j \in S\}$. We require the following condition:

Condition 2. *There exists a positive constant $\eta > 0$, with*

$$\|C_{S^c}C_S^{-1}\mathrm{sign}(\beta_S)\|_\infty \leqslant 1 - \eta.$$

Condition 2 is the Irrepresentable condition, which is required for the most convex penalties such as l_1 penalty to achieve model selection consistency [31, 16, 11]. It is stronger than Condition 1, which may hold in the cases where Condition 2 is violated.

The following Theorem shows that under the proper tuning parameters, the estimate of the proposed method is equal in sign with the true model with high probability.

Theorem 2. *Suppose Condition 2 holds and assume that there exists positive constants K_2 and K_3 such that $\min(\lambda_1) \propto \max(\lambda_1) \propto K_2(n\log p)^{1/2}$ and $\min_{j\in S}|\beta_j| > K_3\sqrt{q}\max(\lambda_1)/n$ where $\lambda_1 = (\lambda_{1,1}, \ldots, \lambda_{1,K})$. We have with probability at least $1 - o(\exp(-n^{c_1}))$ that*

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

Remark 2. The above result is similar to Zhao and Yu [31] and Jia and Yu [13] which have proved the sign consistency of the Lasso and the Elastic Net respectively. The proof follows their proof too. The proof of Theorem 2 begins with the Karush-Kuhn-Tucker (KKT) condition of the proposed method, to find a sufficient condition for sign consistency. Because of the multiple correlations between predictors, our KKT condition and tuning parameters are more complex than either the Lasso or the Elastic Net. Then we use the tail probability bound of Gaussian distribution. More details can be found in the Appendix.

Remark 3. The theoretical results of the proposed procedure and those of the elastic net are alike. The difference is that we use different l_1 and l_2 penalties for different groups which have different correlation strength, while the elastic net is using a single l_1 penalty and a single l_2 penalty.

4. SIMULATIONS

4.1 Detailed Algorithm

In this part, we introduce the detailed algorithm of the proposed method, using the three levels estimation of MPR as an example. Many algorithms designed for Lasso can be

used with modifications to solve MPR. The coordinate descent algorithm is refereed since it is fast and simple [9, 8]. Based on the coordinate descent algorithm, we wish to partially optimize with respect to β_j, in other words, the algorithm minimizes one coordinate while keeping others fixed. To obtain the proposed estimation, at each iteration, we calculate all parameters in turn. Let

$$R(\beta) = \frac{1}{2n}\|y - X\beta\|_2^2 + \sum_{k=1}^{K} P_{A_k}(\beta),$$

where $P_{A_k}(\beta) = \lambda_{2,k}\|\beta_{A_k}\|_2^2/2 + \lambda_{1,k}\|\beta_{A_k}\|_1$. Given $\tilde{\beta}_k (k \neq j)$ and calculating the first derivative of $R(\beta)$, if $\tilde{\beta}_j > 0$, we have

$$\frac{\partial R}{\partial \beta_j}|_{\beta=\tilde{\beta}} = -(1/n)\sum_{i=1}^{n} x_{ij}(y_i - x_i^T\tilde{\beta}) + \lambda_{2,k}\tilde{\beta}_j + \lambda_{1,k}, \quad j \in A_k.$$

If $\tilde{\beta}_j < 0$, the expression is similar, and $\tilde{\beta}_j = 0$ is treated separately. Following the same notations of Friedman et al. [9], we use the soft-thresholding operator $S(z,\gamma)$ to express the coordinate-wise update:

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

Let $\frac{\partial R}{\partial \beta_j}|_{\beta=\tilde{\beta}} = 0$, the coordinate-wise update has the following form:

$$\tilde{\beta}_j \leftarrow \frac{S(\frac{1}{n}\sum_{i=1}^{n} x_{ij}(y_i - \tilde{y}_i^{(j)}), \lambda_{1,k})}{1 + \lambda_{2,k}}, \quad j \in A_k, \tag{2}$$

where $\tilde{y}_i^{(j)} = (1/n)\sum_{j \neq k} x_{ik}\tilde{\beta}_k$ is the fitted value excluding the contribution from x_{ij}. Based on the coordinate-wise update (2), we obtain $\hat{\beta}$ until the iteration convergence. Starting with any value of $\tilde{\beta}$, e.g., the zero vector in p dimensions, the update (2) is repeated for $j = 1, 2, \ldots, p, 1, 2, \ldots, p, \ldots$. During each iteration, we calculate all parameters in turn. When p is large, the computational cost is large too, and thus we give the following modification to improve calculation speed. Based on

$$y_i - \tilde{y}_i^{(j)} = (y_i - x_i^T\tilde{\beta}) + x_{ij}\tilde{\beta}_j,$$

we have

$$\frac{1}{n}\sum_{i=1}^{n} x_{ij}(y_i - \tilde{y}_i^{(j)}) = \frac{1}{n}\sum_{i=1}^{n} x_{ij}(y_i - x_i^T\tilde{\beta}) + \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2\tilde{\beta}_j. \tag{3}$$

Through (2) and (3), we obtain an equivalent form of the coordinate-wise update:

$$\tilde{\beta}_j \leftarrow \frac{S(\lambda_{2,k}\tilde{\beta}_j + \lambda_{1,k} + \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2\,\tilde{\beta}_j, \lambda_{1,k})}{1 + \lambda_{2,k}}, \quad j \in A_k, \tag{4}$$

If $\tilde{\beta}_j = 0$ at the last iteration, it is easy to obtain that $\tilde{\beta}_j$ has the following form through the next iteration:

$$\tilde{\beta}_j = S(\lambda_{1,k}, \lambda_{1,k})/(1 + \lambda_{2,k}) = 0, \quad j \in A_k,$$

indicating that $\tilde{\beta}_j$ will remain zero until convergence. In this case, we modify the algorithm by dealing with the active set at each iteration. More specifically, if we obtain $\tilde{\beta}_j = 0$, we will no longer calculate $\tilde{\beta}_j$ during the next iterations and delete the predictors whose coefficients have been shrunk to 0. As the number of irrelevant variables is always much larger than that of relevant variables, this modification effectively saves the computational cost.

4.2 Simulation results

In this section, we use simulation studies to exhibit the performance of the proposed method comparing with the Lasso [24], Elastic-Net [32], MCP [30] and SCAD [5]. We use the R package glmnet to run Lasso and Elastic-Net [8], the results of MCP and SCAD are based on the R package ncvreg [2]. We use cross-validation to select the tuning parameters.

Consider the following linear model

$$y_i = \sum_{ij}^{p} x_{ij}\beta_j + \epsilon_i, \quad i = 1, \ldots, n,$$

where $\epsilon_i \sim N(0,1)$ and $x_{ij} \sim N(0,\Sigma)$. Let $\Sigma = \{\sigma_{jj'}\}_{p\times p}$, $\sigma_{jj'} = 0.9$ denotes the strong correlation and $\sigma_{jj'} = 0.3$ denotes the weak correlation. We consider the following correlation structure: let the first 10 variables strongly correlated; the second 10 variables weakly correlated; and the rest variables non-correlated. We consider two different dimensional setting, $(n,p) = (200, 40), (200, 400)$, and two different coefficients settings: 1) the first 15 coefficients are set as nonzero, equal to 3, and others are set to zero; 2) the first 10 coefficients are set as nonzero, equal to 3, and others are set to zero.

Selection and estimation performance of the five methods is compared in four scenarios. The l_2 norm error ($\|\hat{\beta} - \beta\|_2$), the l_1-norm($\|\hat{\beta} - \beta\|_1$) error, MSE (the mean-squared error) and the estimated number of nonzero coefficients (NZ) are computed. Also, the false positive rate (FPR) and the true positive rate (TPR) are defined as following:

$$\text{FPR} = \frac{|j \in \{1, ..., p\} : \hat{\beta}_j \neq 0 \; and \; \beta_j = 0|}{|j \in \{1, ..., p\} : \beta_j = 0|},$$

$$\text{TPR} = \frac{|j \in \{1, ..., p\} : \hat{\beta}_j \neq 0 \; and \; \beta_j \neq 0|}{|j \in \{1, ..., p\} : \beta_j \neq 0|}.$$

We simulate 100 replicates for every scenario. Results are summarized in Table 1. As one can see, MPR performs better than other methods in both model selection and estimation accuracy. When the relevant covariates have both

Table 1. Performance comparison for four simulation examples.

Method	l_2-error	l_1-error	NZ	FPR	TPR	MSE
		(n, p, q)	=	(200, 40, 10)		
MPR	0.33(0.07)	0.86(0.20)	10.00(0.00)	0.00(0.00)	1.00(0.00)	1.07(0.10)
Lasso	0.70(0.17)	1.81(0.47)	10.00(0.00)	0.00(0.00)	1.00(0.00)	1.08(0.10)
Elastic Net	0.68(0.16)	1.75(0.43)	10.00(0.00)	0.00(0.00)	1.00(0.00)	1.15(0.10)
MCP	2.03(1.54)	4.43(3.31)	9.47(0.66)	0.00(0.00)	0.95(0.06)	1.36(0.49)
SCAD	1.07(0.97)	2.46(1.79)	9.95(0.35)	0.00(0.00)	0.99(0.03)	1.05(0.23)
		(n, p, q)	=	(200, 40, 15)		
MPR	0.50(0.09)	1.61(0.34)	15.60(0.73)	0.02(0.03)	1.00(0.00)	1.00(0.09)
Lasso	0.80(0.17)	2.48(0.52)	16.10(1.03)	0.04(0.04)	1.00(0.00)	1.01(0.09)
Elastic Net	0.78(0.17)	2.43(0.53)	16.77(1.11)	0.07(0.04)	1.00(0.00)	1.05(0.09)
MCP	0.73(0.28)	2.12(0.64)	15.08(0.27)	0.00(0.01)	1.00(0.01)	0.94(0.10)
SCAD	0.71(0.16)	2.12(0.47)	16.15(1.07)	0.05(0.04)	1.00(0.00)	0.93(0.09)
		(n, p, q)	=	(200, 400, 10)		
MPR	0.35(0.03)	0.99(0.07)	10.00(0.00)	0.00(0.00)	1.00(0.00)	1.85(0.10)
Lasso	0.74(0.17)	1.93(0.49)	10.00(0.00)	0.00(0.00)	1.00(0.00)	1.21(0.10)
Elastic Net	0.62(0.14)	1.61(0.39)	10.01(0.09)	0.00(0.00)	1.00(0.00)	1.17(0.10)
MCP	4.08(1.66)	9.92(5.36)	9.50(1.36)	0.00(0.00)	0.84(0.09)	2.30(0.77)
SCAD	10.17(1.09)	31.61(3.22)	6.42(1.67)	0.00(0.00)	0.52(0.09)	9.65(1.57)
		(n, p, q)	=	(200, 400, 15)		
MPR	0.47(0.07)	1.51(0.24)	15.36(0.56)	0.00(0.00)	1.00(0.00)	1.14(0.09)
Lasso	0.94(0.16)	3.02(0.55)	15.24(0.49)	0.00(0.00)	1.00(0.00)	1.37(0.10)
Elastic Net	0.94(0.16)	3.01(0.55)	15.25(0.50)	0.00(0.00)	1.00(0.00)	1.38(0.10)
MCP	2.11(1.56)	5.03(3.64)	15.15(1.11)	0.00(0.00)	0.96(0.05)	1.34(0.50)
SCAD	2.09(1.54)	4.98(3.61)	16.05(2.26)	0.00(0.01)	0.98(0.04)	1.31(0.47)

strong and weak correlations. MCP and SCAD tend to over-select the irrelevant variables. Lasso and Elastic Net identify the relevant covariates and reach a good MSE. However, MPR still outperforms them in l_1-error and l_2-error.

5. EMPIRICAL ANALYSIS

In this section, we apply the proposed method to the index tracking problem in financial modeling. Index tracking is one of the most popular topics in finance. Index tracking, also referred to as passive portfolio management, aims to replicate a specific index to match its performance. Malkiel [14] has shown that the average return of indexed funds, e.g., passive portfolio management, is higher than that of other funds, with an average annual difference of 3 percentage points. Active portfolio management with a high level of scientific management and management standards often cannot beat the market for a long time. Besides, since index tracking does not need to beat the market, this strategy requires taking on smaller market risk than is quired for active portfolio management.

One advantage of applying the proposed method to the index tracking is that this is a high-dimensional data modeling problem as the number of stocks is often hundreds or thousands, whereas the number of observations (days) is tens or hundreds. MPR can select a few rather than many stocks to track the performance of the index. Compared to the full replication or active portfolio management, this method offers a lower-cost route to investing in an entire market.

Second, it is easy to find that the stock data has multiple correlation levels, while MPR is proposed to handling this kind of data. We will show in the following that combining statistical modeling and MPR can track the behavior of the index well. It means that the portfolios obtained by this method replicate the market well and thus stand a good chance of gaining nice investment returns. Other discussions of applying statistical modeling on this problem also can be found in Fan et al. [7], Yang and Yang [28], etc.

We consider S&P500 index from June 2018 till December 2019 and divide the dataset into 18 rolling periods (https://www.wind.com.cn/NewSite/edb.html). Each period includes training (= 100 days) data and testing (= 20 days) data. The training data is used for modeling and the testing data is used for forecasting. Let y_t represents the return of the S&P500 index on day t and x_{jt} represents the return of stock j on day t. Then we can describe the relationship between y_t and x_{jt} by the following linear regression model:

$$y_t = \sum_{j=1}^{500} x_{jt}\beta_j + \varepsilon_t.$$

We divide stocks into three groups, identifying a non-correlated group, a weakly correlated group, and a highly

correlated group. Based on data, we observe that if the correlation coefficient threshold is set above 0.6, the sizes of strongly correlated groups are always less than 20, while when the correlation coefficient threshold is set below 0.3, the sizes of uncorrelated groups are always less than 50. To ensure that all the stocks are divided clearly with each group is large enough to show their difference, we choose 0.3 and 0.6 as the thresholds, and in this case, each group contains about 150 to 200 stocks.

We compare five methods, including MPR, Lasso, Elastic-Net, SCAD, and MCP. We choose the tuning parameter to select around 150 constituent stocks for each method. To measure the performance of above methods, we use the tracking error and tracking difference. Let $\hat{y}_t$ be the predicted value of y_t, $\text{err}_t = y_t - \hat{y}_t$, $R_t = \ln(\hat{y}_t/\hat{y}_{t-1})$ and $R_{tm} = \ln(y_t/y_{t-1})$, tracking error and tracking difference are defined as following:

$$\text{TrackingError}_{year} = \sqrt{252} \times \sqrt{\sum(\text{err}_t - \text{mean}(\text{err}))^2/(T-1)},$$

$$\text{TrackingDifference}_{year} = \frac{252}{T}\sum_{t=1}^{T}(R_t - R_{tm}),$$

where mean(err) is the mean of err_t, $t = 1, ..., T$. The variance is always regarded as a risk measure, e.g., high variance is associated with higher risk. The tracking error, a standard deviation percentage difference, is often used to describe the risk of the obtained portfolios. During the empirical analysis and compared to other methods, the portfolio obtained by the MPR achieves the smallest and stablest tracking error. This result indicates that the MPR offers lower volatility and higher risk-adjusted returns.

The left panel of Figure 1 shows the tracking errors on 18 testing sets. The tracking error produced by MPR is more stable and lower than other mentioned methods. Compared with the other methods, MPR substantially reduces the tracking error by 28%-81%. The right panel of Figure 1 shows the tracking differences on 18 testing sets. As one can see, the tracking difference of MPR is much closer to 0 compared to that of other methods. More specifically, MPR produces the tracking difference in which the absolute values are lower by 30%-90%, showing that it fits the real market yields better. Figure 2 shows the predicted performance of five methods.

Figure 3 discusses the characteristics of the chosen stocks in each industry, comparing the amount and the total market capitalization of the chosen stocks from MPR with those of the constituent stocks in each industry. As shown in the left panel of Figure 3, the percentage of chosen stocks in each industry are alway between 23%to 50%, except utilities and energy industry. Health care and IT are the two most chosen industries from MPR, whose total market capitalization reach 42% and 23%. As shown in the left panel of Figure 3, the proportion of total market capitalization

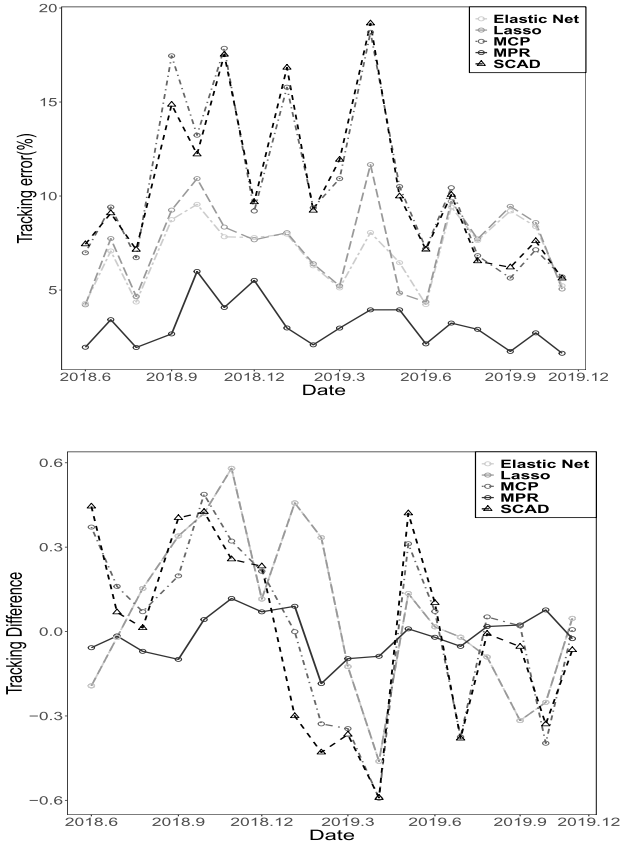

Figure 1. Tracking error and Tracking Difference of MPR, Elastic Net, Lasso, SCAD, and MCP.

between chosen stocks and total stocks are always higher than 20%, except energy industry. The chosen stocks from Utilities and Telecommunications service have the highest proportion in total market capitalization, 77.7% and 65.7% respectively, followed by the real estate, 56.7%, daily consumption, 39.3%.

6. CONCLUSION

In this paper, we propose a two-step method, named Multiple penalized regularization (MPR), for estimating high-dimensional regression models with multi-level correlated predictors. The proposed method is a two-step estimation. The first step serves as a clustering step, in which predictors are classified into different groups, e.g., strong correlation groups, weak correlation groups, and non-correlation groups. The second step is a joint estimation solved by the modified coordinate descent algorithm. We prove the theoretical guarantees of the proposed method. The simulation study shows the advantages of the proposed method compared to other methods. In the empirical part, we apply the MPR to track S&P 500 Index and summarize the tracking errors and tracking differences, all these results are effective and meaningful.

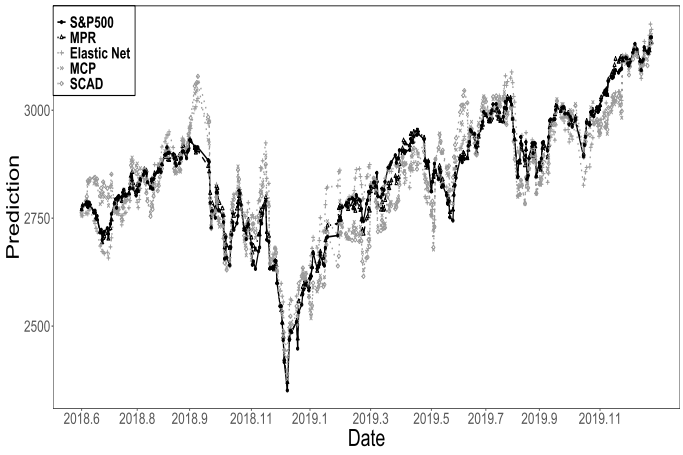

Figure 2. Predicted performance of MPR, Elastic Net, SCAD, and MCP tracking the index.

In the future work, we plan to extend our proposal to more complicated data set with more complicated models. For example, most existing work includes this paper assumes that the models are correctly specified or have fixed dimensionality, yet both features of model misspecification and high dimensionality are prevalent in practice [4]. It would be worthwhile to study how to compare different models when models are unspecified under correlated data.

APPENDIX

We first provide Lemma 1 that shows the solution of MPR satisfies the requirement of Condition 1

Lemma 1. *Assume ϵ_i are i.i.d. Gaussian random variables with mean 0 and variance σ^2, $i = 1, \ldots, n$. We have with a positive constant M that*

$$\|\hat{\beta}_{S^c}\|_1 \leqslant M\|\hat{\beta}_S - \beta_S\|_1.$$

Proof of Lemma 1. We first have the following inequality based on the definition of $\hat{\beta}$:

$$\frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \sum_{k=1}^{K}(\lambda_{2,k}\|\hat{\beta}_{A_k}\|_2^2/2 + \lambda_{1,k}\|\hat{\beta}_{A_k}\|_1)$$
$$\leqslant \frac{1}{2}\|y - X\beta\|_2^2 + \sum_{k=1}^{K}(\lambda_{2,k}\|\beta_{A_k}\|_2^2/2 + \lambda_{1,k}\|\beta_{A_k}\|_1).$$

We first have

$$\frac{1}{2}\|y - X\hat{\beta}\|_2^2 - \frac{1}{2}\|y - X\beta\|_2^2 + \sum_{k=1}^{K}\lambda_{2,k}(\|\hat{\beta}_{A_k}\|_2^2 - \|\beta_{A_k}\|_2^2)/2$$
$$= \frac{1}{2}(\hat{\beta} - \beta)^{\mathrm{T}}(X^{\mathrm{T}}X + \lambda_2 I)(\hat{\beta} - \beta) - \epsilon^{\mathrm{T}}X(\hat{\beta} - \beta)$$
$$+ \sum_{k=1}^{K}\lambda_{2,k}\beta_{A_k}^{\mathrm{T}}(\hat{\beta}_{A_k} - \beta_{A_k}),$$

where $\lambda_2 = (\lambda_{2,1}\mathbf{1}_{A_1}, \ldots, \lambda_{2,K}\mathbf{1}_{A_K})$. Based on $\|\beta\|_\infty \leqslant \min(\lambda_1)/4\max(\lambda_2)$ where $\lambda_1 = (\lambda_{1,1}, \ldots, \lambda_{1,K})$ and $\lambda_2 = (\lambda_{2,1}, \ldots, \lambda_{2,K})$:

$$\beta_{A_k}^{\mathrm{T}}(\hat{\beta}_{A_k} - \beta_{A_k}) \leqslant \|\beta\|_\infty\|\hat{\beta}_{A_k} - \beta_{A_k}\|_1$$
$$\leqslant \frac{1}{4}\min(\lambda_1)\|\hat{\beta}_{S_k} - \beta_{S_k}\|_1 + \frac{1}{4}\min(\lambda_1)\|\hat{\beta}_{S_k^c}\|_1;$$

based on $\{2\|W\|_\infty \leqslant \min(\lambda_1)/\sqrt{n}\}$:

$$2\epsilon^{\mathrm{T}}X(\hat{\beta} - \beta) \leqslant \min(\lambda_1)\|\hat{\beta}_S - \beta_S\|_1 + \min(\lambda_1)\|\hat{\beta}_{S^c}\|_1;$$

and combined with $\|\hat{\beta}_S\|_1 + \|\hat{\beta}_S - \beta_S\|_1 \geqslant \|\beta\|_1$, the following inequality holds:

$$(\hat{\beta} - \beta)^{\mathrm{T}}(X^{\mathrm{T}}X + \lambda_2 I)(\hat{\beta} - \beta) + 2\sum_{k=1}^{K}\lambda_{1,k}\|\hat{\beta}_{S_k^c}\|_1$$
$$\leqslant 2\epsilon^{\mathrm{T}}X(\hat{\beta} - \beta) - 2\sum_{k=1}^{K}\lambda_{2,k}\beta_{A_k}(\hat{\beta}_{A_k} - \beta_{A_k})$$
$$+ 2\sum_{k=1}^{K}\lambda_{1,k}\|\hat{\beta}_{S_k} - \beta_{S_k}\|_1$$
$$\leqslant \sum_{k=1}^{K}\lambda_{1,k}\|\hat{\beta}_{S_k^c}\|_1 + 7/2\sum_{k=1}^{K}\lambda_{1,k}\|\hat{\beta}_{S_k} - \beta_{S_k}\|_1.$$

Since

$$(\hat{\beta} - \beta)^{\mathrm{T}}(X^{\mathrm{T}}X + \lambda_2 I)(\hat{\beta} - \beta) \geqslant 0,$$

assuming that $\min(\lambda_1) \propto \max(\lambda_1)$, we have that with a positive constant M

$$\|\hat{\beta}_{S^c}\|_1 \leqslant M\|\hat{\beta}_S - \beta_S\|_1.$$

□

Proof of Theorem 1. Set

$$F(\beta) = \frac{1}{2}\|y - X\beta\|_2^2 + \sum_{k=1}^{K}\frac{\lambda_{2,k}}{2}\|\beta_{A_k}\|_2^2 + \sum_{k=1}^{K}\lambda_{1,k}\|\beta_{A_k}\|_1$$

and

$$V(u) = F(\hat{\beta}) - F(\beta),$$

where $\hat{u} = \sqrt{n}(\hat{\beta} - \beta)$, and thus we have $\hat{u} := \arg\min V(u)$. Set

$$V(u) = \frac{1}{2}\big(\|y - X\hat{\beta}\|_2^2 - \|y - X\beta\|_2^2\big)$$
$$+ \sum_{k=1}^{K}\frac{\lambda_{2,k}}{2}\big(\|\hat{\beta}_{A_k}\|_2^2 - \|\beta_{A_k}\|_2^2\big)$$

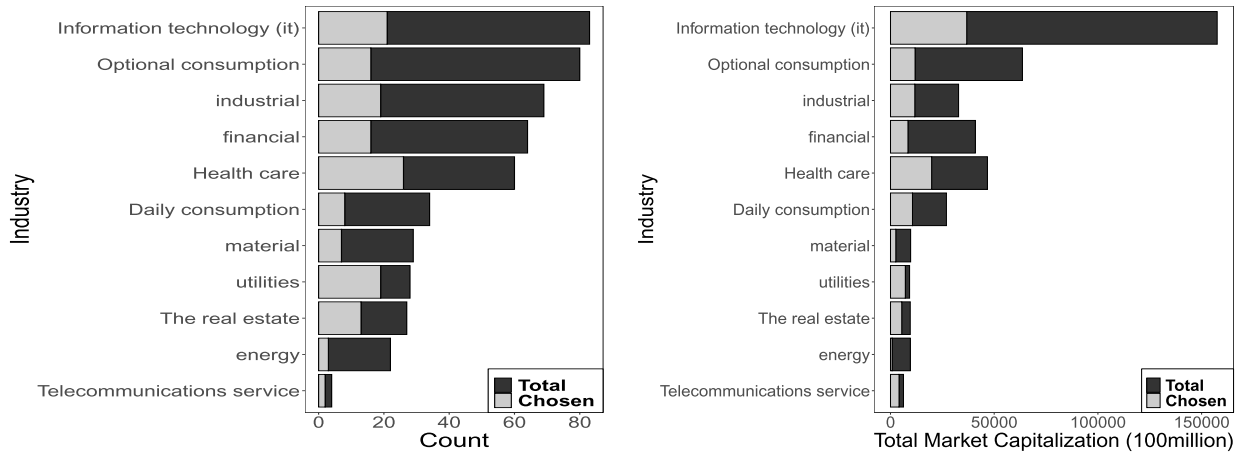

Figure 3. The left panel shows the amount of the chosen stocks comparing the amount of the constituent stocks in each industry. The right panel shows the total market capitalization of the chosen stocks comparing the constituent stocks in each industry. We use the 'Chosen' tag to denote the stocks chosen by MPR in all 18 rolling periods, and the 'Total' tag represents all the stocks that S&P 500 contains.

$$+\sum_{k=1}^{K}\lambda_{1,k}\big(\|\hat{\beta}_{A_k}\|_1-\|\beta_{A_k}\|_1\big)$$
$$\triangleq \frac{1}{2}H_1+\sum_{k=1}^{K}\frac{\lambda_{2,k}}{2}H_{k,2}+\sum_{k=1}^{K}\lambda_{1,k}H_{k,3}.$$

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

$$H_1=u^{\mathrm{T}}Cu-2u^{\mathrm{T}}W,\quad H_2=u_{A_k}^{\mathrm{T}}u_{A_k}/n+2u_{A_k}^{\mathrm{T}}\beta_{A_1}/\sqrt{n},$$
$$H_3\geqslant -\|u_{S_k}\|_1/\sqrt{n}+\|u_{S_k^c}\|_1/\sqrt{n},$$

and thus

$$V(u)\geqslant \frac{1}{2}u^{\mathrm{T}}Cu/n-u^{\mathrm{T}}W/n+\sum_{k=1}^{K}\lambda_{2,k}u_{A_k}^{\mathrm{T}}\beta_{A_k}/\sqrt{n}$$
$$-\sum_{k=1}^{K}\lambda_{1,k}\|u_{S_k}\|_1/\sqrt{n}+\sum_{k=1}^{K}\lambda_{1,k}\|u_{S_k^c}\|_1/\sqrt{n}$$

By RE condition (Condition 1) and Lemma 1, $u^{\mathrm{T}}Cu\geqslant\kappa\|u\|_2^2$. We also have

$$-u^{\mathrm{T}}W\geqslant -\|u_S\|_2\|W_S\|_2-\|W_{S^c}\|_\infty\|u_{S^c}\|_1.$$

Conditional on $\|\beta\|_\infty\leqslant\min(\lambda_1)/4\max(\lambda_2)$, $\{2\|W\|_\infty\leqslant\min(\lambda_1)/\sqrt{n}\}$ where $\lambda_1=(\lambda_{1,1},\ldots,\lambda_{1,K})$ and $\lambda_2=(\lambda_{2,1},\ldots,\lambda_{2,K})$, we have $\{2\|W_S\|_2\leqslant\sqrt{q}\min(\lambda_1)/\sqrt{n}\}$ and

$$\sum_{k=1}^{K}\lambda_{1,k}\|u_{S_k^c}\|_1/\sqrt{n}\geqslant\|W_{S^c}\|_\infty\|u_{S^c}\|_1.$$

The lower bound of $V(u)$ becomes

$$V(u)\geqslant\|u\|_2\Big\{\frac{\kappa}{2}\|u\|_2-\frac{\min(\lambda_1)\sqrt{q}}{\sqrt{n}}\Big\}.$$

Based on $\dfrac{\min(\lambda_1)}{\sqrt{n}}=K_1\sqrt{\log p}$, above inequality implies that

$$\|\hat{u}\|_2\leqslant\frac{8\sigma}{\kappa}\sqrt{q\log p}.$$

Otherwise, we would have $V(\hat{u})>0$, which means the minimum of $V(u)$ does not be attained. □

Proof of Theorem 2. We first use the Karush-Kuhn-Tucker (KKT) conditions for the proposed method. Given the clustering information, i.e., $A_1,\ldots,A_K$, the conditions can be written as

$$\frac{1}{n}X^{\mathrm{T}}(y-X\hat{\beta})=\sum_{k=1}^{K}\lambda_{2,k}\hat{\beta}_{A_k}+\sum_{k=1}^{K}\lambda_{1,k}\gamma_{A_k},\tag{5}$$

$$\gamma_j\in\begin{cases}\{\mathrm{sign}(\hat{\beta}_j)\} & \text{if } \hat{\beta}_j\neq 0\\ [-1,1] & \text{if } \hat{\beta}_j=0\end{cases},\quad \text{for } j=1,\ldots,p.\tag{6}$$

According to KKT conditions, $\hat{\beta}$ is a solution of (1) if and only if $\hat{\beta}$ satisfies (5) and (6). Let $\hat{\beta}_S$, β_S, $\hat{\beta}_{S^c}$, and $\hat{\beta}_{S^c}$ be the S and S^c entries of $\hat{\beta}$ and β, respectively.

Denote $A_1,\ldots,A_K$ as different clustering groups. For notational simplicity and also without loss of generality, we assume that $(1,\ldots,p)=(A_1,\ldots,A_K)$. Then we set $S_k=A_k\cap S$ and $S_k^*=A_k/S_k$. Thus, we have $S=(S_1,\ldots,S_K)$ and $S^c=(S_1^*,\ldots,S_K^*)$. As shown in [31], to prove the sign consistency of MPR, it suffices to prove with high probability that

$$|\hat{\beta}_S-\beta_S|<|\beta_S|,\quad \hat{\beta}_{S^c}=0.$$

Following the same notations of Theorem 1, i.e., denote $\hat{u}=\sqrt{n}(\hat{\beta}-\beta)$, $C=X^{\mathrm{T}}X/n$ and $W=X^{\mathrm{T}}\epsilon/\sqrt{n}$. Among, the S

entries of $\hat{\beta}$ can be described by

$$(C_S + \lambda_{2,S} I/n)\hat{u} = W_S - \lambda_{2,S} I\beta_S/\sqrt{n} - \lambda_{1,S}(\gamma)/\sqrt{n},$$

where $\lambda_{2,S} = (\lambda_{2,1}\mathbf{1}_{S_1}, \ldots, \lambda_{2,K}\mathbf{1}_{S_K})^{\mathrm{T}}$ and $\lambda_{1,S}(\gamma) = (\lambda_{1,1}\gamma_{S_1}, \ldots, \lambda_{1,K}\gamma_{S_K})$. Combined with $|\hat{\beta}_S - \beta_S| < |\beta_S|$,

$$|\sqrt{n}W_S - \lambda_{2,S} I\beta_S - \lambda_{1,S}(\gamma)| < (C_S + \lambda_{2,S} I))|\beta_S|$$

Thus, proving the sign consistency suffices to prove

$$\sqrt{n}|(C_S + \lambda_{2,S} I))^{-1} W_S| < |\beta_S| - C_S^{-1}\lambda_{1,S}(\gamma),$$
$$|C_{S^c}\hat{u} + W_{S^c}| \leqslant \lambda_1^*/\sqrt{n},$$

where $\lambda_1^* = (\lambda_{1,1}\mathbf{1}_{S_1^*}, \ldots, \lambda_{1,K}\mathbf{1}_{S_K^*})^{\mathrm{T}}$. The former inequality holds under the proper choice of $\lambda_1 = (\lambda_1, \ldots, \lambda_K)$, which will be discussed later, and the tail probability bound of Gaussian distribute with $\min_{j\in S}|\beta_j| > K_3\sqrt{q}(\max(\lambda_1))/n$ where K_3 is a positive constant. Combine the above two inequality, to prove the latter inequality suffices to prove

$$\sqrt{n}|C_{S^c}(C_S + \lambda_S \cdot I)^{-1} W_S - W_{S^c}|$$
$$\leqslant \lambda_1^* - |C_{S^c}(C_S + \lambda_S \cdot I)^{-1}\lambda_{1,S}(\gamma)|.$$

To obtain the above inequality, we require the following inequality with a positive constant vector η^*:

$$|C_{S^c} C_S^{-1}\lambda_{1,S}(\gamma)| < \lambda_1^* - \eta^*.$$

When $\min(\lambda_1) \propto \max(\lambda_2)$, above inequality holds when Irrepresentable condition (Condition 2) holds. Now we discuss the convergence rate, which is determined by the tail probability bound of Gaussian distribute. Since ϵ is the error vector that $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, n$, we have

$$P(\|X^{\mathrm{T}}\epsilon/\sqrt{n}\|_\infty > 2\sigma(\log p)^{1/2})$$
$$\leqslant \sum_{j=1}^{p} P(|X_j^{\mathrm{T}}\epsilon/\sqrt{n}| > 2\sigma(\log p)^{1/2})$$
$$< p \cdot \exp(-\frac{4\sigma^2 \log p}{2\sigma^2}) = 1/p.$$

With $\min(\lambda_1) \propto \max(\lambda_1) \propto K_2(n\log p)^{1/2}$ and $\min_{j\in S}|\beta_j| > K_3\sqrt{q}(\max(\lambda_1))/n$, completed the proof. □

Received 1 June 2021

REFERENCES

[1] Ahsen, M. E., Challapalli, N. and Vidyasagar, M. (2017). Two New Approaches to Compressed Sensing Exhibiting Both Robust Sparse Recovery and the Grouping Effect. *Journal of Machine Learning Research* **18** 1745-1768. MR3687597

[2] 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

[3] Chiquet, J., Grandvalet, Y., Charbonnier, C. et al. (2012). Sparsity with sign-coherent groups of variables via the cooperative-lasso. *Annals of Applied Statistics* **6** 795-830. MR2976492

[4] Demirkaya, E., Feng, Y., Basu, P. and Lv, J. (2021). Large-scale model selection in misspecified generalized linear models. *Biometrika* **1** 1-21.

[5] 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

[6] Fan, J., Rigollet, P. and Wang, W. (2015). Estimation of functionals of sparse covariance matrices. *Annals of statistics* **43** 2706-2737. MR3405609

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

[8] 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.

[9] Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics* **1** 302-332. MR2415737

[10] Guo, J. H., Hu, J. C., Jing, B. Y. and Zhang, Z. (2016). Spline-Lasso in High-Dimensional Linear Regression. *Journal of the American Statistical Association* **111** 288-297. MR3494660

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

[14] Malkiel, B. G. (1973). *A Random Walk Down Wall Street: The Time-Tested Strategy for Successful Investing*. W. W. Norton.

[15] Meier, L., van de Geer, S. and Bühlmann, P. (2008). The group LASSO for logistic regression. *Journal of the Royal Statistical Society: Series B* **55** 2183-2202. MR2412631

[16] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Annals of Statistics* **34** 1436-1462. MR2278363

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

[18] 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

[19] Raskutti, G., Wainwright, M. J. and Yu, B. (2010). Restricted Eigenvalue Properties for Correlated Gaussian Designs. *Journal of Machine Learning Research* **11** 2241-2259. MR2719855

[20] Shan, L., Qiao, Z., Cheng, L. and Kim, I. (2020). Joint estimation of the two-level gaussian graphical models across multiple classes. *Journal of Computational and Graphical Statistics* **29** 562-579. MR4153182

[21] She, Y. Y. (2010). Sparse regression with exact clustering. *Electronic Journal of Statistics* **4** 1055-1096. MR2727453

[22] Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics* **22** 231-245. MR3173712

[23] Tan, K. M., Witten, D. and Shojaie, A. (2015). The cluster graphical lasso for improved estimation of Gaussian graphical models. *Computational Statistics & Data Analysis* **85** 23-36. MR3299780

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

[25] Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B* **67** 91-108.

MR2136641
[26] Witten, D. M., Shojaie, A. and Zhang, F. (2014). The cluster elastic net for high-dimensional regression with unknown variable grouping. *Technometrics* **56** 112-122. MR3176577
[27] Wu, L. and Yang, Y. (2014). Nonnegative Elastic Net and Application in Index Tracking. *Applied Mathematics and Computation* **227** 541-552. MR3146340
[28] Yang, Y. and Yang, H. (2021). Adaptive and reversed penalty for analysis of high-dimensional correlated data. *Applied Mathematical Modelling* **92** 63-77. MR4177458
[29] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B* **68** 49-67. MR2212574
[30] Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* **38** 894-942. MR2604701
[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. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B* **67** 301-320. MR2137327

Wenjun Cao
School of Statistics and Mathematics
Central University of Finance and Economics
Bejing, China
E-mail address: 2017310840@email.cufe.edu.cn

Lisu Wang
School of Statistics and Mathematics
Central University of Finance and Economics
Bejing, China
E-mail address: 2018310837@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