Statistics and Its Interface Volume 12 (2019) 599–615

Nonnegative hierarchical lasso with a mixed $(1, \frac{1}{2})$-penalty and a fast solver

Wanling Xie and Hu Yang*

Grouping structures arise naturally in many high dimensional statistical problems. Incorporation of grouping information can efficiently improve the statistical accuracy and model interpretability. In addition, nonnegative constraints are essential to cope with index tracking problems. This paper proposes the nonnegative hierarchical lasso with nonnegative constraints on the coefficients both in low dimensional setting and ultra high dimensional setting, which is capable of simultaneous selection at both the group and within-group levels with overlap, namely the bi-level selection.

In theoretical analysis, we show the nonnegative hierarchical lasso enjoys the oracle properties in group selection when the number of covariates diverges with the sample size under certain regularity conditions. Since there are less works devoted to the theoretical properties of bi-level selection methods in cases where the number of variables or groups is much larger than the sample size, we also derive the oracle inequalities for the prediction and l_1 estimation errors of the estimator under the restricted eigenvalue conditions on the design matrix. It is shown to have group selection consistency and estimation consistency in ultra high-dimensional sparse linear regression models. To get the solution of the nonnegative hierarchical lasso, we propose a fast and efficient iterative half thresholding-based local linear approximation algorithm (IHT-LLA) for solving. Simulations indicate that the nonnegative hierarchical lasso outperforms other nonnegative regularization methods and is robust against possible mis-specified grouping structure. Besides, we further apply our method to the index tracking problems.

Keywords and phrases: Nonnegative hierarchical lasso, Oracle property, Bi-level selection, Index tracking, Oracle inequality.

1. INTRODUCTION

Consider the classical linear regression model

$$y = X\beta + \epsilon, \tag{1}$$

where y is the response and $X = (X_1, \dots, X_p)$ is the design matrix, where $X_j = (x_{1j}, \dots, x_{nj})^T$, $j = 1, \dots, p$ are the predictors and ϵ is the error term. Without loss of generality, we assume the data are centered so that the intercept can be excluded from the regression model.

Suppose the predictors can be naturally divided into J groups and there are p_j variables in the jth group, then it can be rewritten as

$$y_i = \sum_{j=1}^{J} \sum_{k=1}^{p_j} x_{i,jk}\beta_{jk} + \epsilon_i, \quad i = 1, \dots, n. \tag{2}$$

Grouping structures arise from diverse fields of scientific research. For example, in ANOVA, a factor with multiple levels can be referred as a group. Similarly, in gene expression studies, genes which belong to the same biological pathway form a natural group and in genome-association studies, single-nucleotide polymorphisms from the same gene can also be regarded as a group. Furthermore, in nonparametric additive models, each original predictor can be expanded into a set of basis functions. Thus it is desirable for us to incorporate such grouping structures into the model as the prior knowledge. There are several recent literature devoted to address the group variable selection problems. Yuan and Lin [7] proposed the group lasso method which used an l_2 norm of the coefficients from a group and developed a group coordinate descent algorithm to compute the group lasso solutions. Meier et al. [8] extended this approach to the logistic regression model. Wang et al. [9] developed a group SCAD for microarray time course gene expression data. Zhao et al. [10] proposed the composite absolute penalty for grouped and hierarchical variable selection which included the group lasso as a special case. Hu et al. [19] considered the group adaptive elastic-net approach to cope with collinearity. In many applications, however, it is of importance to select both groups and individuals. Huang et al. [13] proposed the group bridge for bi-level selection which yielded sparsity at the group and within group levels. Breheny and Huang [14] introduced a general framework for bi-level selection and derived a local coordinate descent algorithm. Friedman et al. [11] considered a more general penalty which blended the lasso penalty with the group lasso penalty and proposed the sparse group lasso method. The hierarchical lasso proposed by Zhou and Zhu [12] can be regarded as a special case of group bridge with $\gamma = 0.5$. Huang et al. [21] reviewed the several group selection methods involving both group selection and bi-level selection methods.

[*] Corresponding author. ORCID: 0000-0001-6589-8534.

Index tracking, as a popular passive portfolio management strategy, aims at reproducing the performance of a market index. In view of the transaction costs and adjustment of constituent stocks, it is necessary for us to construct a sparse index tracking portfolio against the costly full replication, i.e., purchasing a small amount of assets to replicate the index. There are two kinds of widely used methods of sparse index tracking. The first one is such a two-stage approach, namely stock selection first and then capital allocation. Various stock selection methods have been proposed in the last two decades. For instance, K.J.Oh et al. [23] took advantage of market capitalization to select assets. Dose and Cincotti [24] selected the stocks which are more correlated with the index. Alexander [25] considered a selection method based on the idea if there exists a linear combination of the log-prices of selected assets cointegrated well with the value of index. The above two-stage approach, however, suffers from unclarity on how optimal the resulting portfolio is. To unify these two steps, another approach is to directly penalize the cardinality of the tracking portfolio (i.e., l_0-norm of the coefficients) to simultaneously select assets and allocate the capital to the selected stocks. Note that this problem is highly non-convex due to the l_0-norm term. In general, a common approach is to substitute other non-convex norms for the l_0-norm. In addition, fewer than half the exchanges around the world allow short sales due to the lack of consensus among regulators. Conversely, some countries don't officially prohibit short-selling, yet no short-selling takes place for lack of necessary institutions or request for high fees. Considering the legality and feasibility of short-selling, the nonnegative constraints are commonly used in index tracking management as the short-sale constraints. For this aim, a large amount of literatures have devoted to the nonnegative penalized methods and the application in index tracking. Breiman [1] proposed the nonnegative garrotte and showed its stability when compared to the subset regression and ridge regression. Slawski and Hein [32] considered the nonnegative least squares and gave the rate of convergence. Meinshausen [30] showed the effectiveness of sign-constrained least squares estimation for sparse high-dimensional data. Wu et al. [34] proposed the nonnegative lasso and proved its oracle property. Wu and Yang [33] introduced the nonnegative elastic-net and applied it to the index tracking problem. Yang and Wu [35] further proposed the nonnegative adaptive lasso as an improvement of nonnegative lasso.

Since there has been little discussion about the bi-level selection methods under the nonnegative constraints on the coefficient and notice that $L_{1/2}$ regularizer can be taken as a representative of the L_p regularizer. Compared to the L_0 regularizer and L_1 regularizer, the L_p regularizer can produce more sparse solutions than L_1 regularizer and it is easier to be solved than the L_0 regularizer. In this paper, we propose the nonnegative hierarchical lasso with a group $L_{1/2}$ regularizer which gives nonconvexity on the group level for simultaneous estimation and bi-level selection in sparse high-dimensional linear regression models. This method is an extension of the hierarchical lasso with nonnegative constraints on the coefficients. Under certain appropriate conditions, it is shown to have satisfactory properties both in diverging and ultra high-dimensional settings. The main contribution of this paper is threefold. Firstly, we extend the hierarchical lasso to nonnegative hierarchical lasso and prove it enjoys the oracle properties in group selection when the number of covariates diverges with the sample size and further derive the oracle inequalities in cases where the number of covariates is much larger than the sample size. Secondly, we propose a fast and efficient iterative half thresholding-based local linear approximation algorithm(IHT-LLA) based on the iterative half thresholding algorithm in Xu et.al. [22], group coordinate descent in Wei and Zhu [28] and local linear approximation in Zou and Li [5]. The proposed algorithm is faster than the existing algorithms for solving the group $L_{1/2}$ optimization problem. Thirdly, we show by simulation studies and the real data example that the nonnegative hierarchical lasso outperforms other nonnegative methods for high-dimensional data.

The rest of this paper is organized as follows. In Section 2, we introduce the nonnegative hierarchical lasso. The oracle properties both in the low-dimensional setting and in cases where $p \gg n$ are studied in Section 3. In Section 4, we propose a fast iterative half thresholding-based local linear approximation algorithm(IHT-LLA) for implementation. In Section 5 and Section 6, simulation studies as well as an application are conducted to show the finite sample performance of the proposed method as compared to other nonnegative methods. We conclude with a few remarks in Section 7. All the technical proofs are given in the Appendix.

2. NONNEGATIVE HIERARCHICAL LASSO

In this section, we extend the hierarchical lasso to nonnegative hierarchical lasso to simultaneously select important variables at both the group and within-group levels when the true coefficients are nonnegative.

We denote

$$\beta_{jk} = d_j \alpha_{jk}, \quad j = 1, \ldots, J; k = 1, \ldots, p_j, \tag{3}$$

where $d_j \geq 0$ represents the group information from jth group as the first level of hierarchy to reflect the information that all β_{jk} belong to the jth group by treating each β_{jk} hierarchically, while $\alpha_{jk}{}' s \geq 0$ are at the second level to show the differences within the jth group.

Thus, we consider the following penalized least square question with nonnegative constraints on (d_j, α_{jk})

$$\begin{cases} \min\limits_{d_j, \alpha_{jk}} \ \frac{1}{2} \sum\limits_{i=1}^{n} \left(y_i - \sum\limits_{j=1}^{J} d_j \sum\limits_{k=1}^{p_j} \alpha_{jk} x_{i,jk} \right)^2 \\ + \lambda_1 \sum\limits_{j=1}^{J} d_j + \lambda_2 \sum\limits_{j=1}^{J} \sum\limits_{k=1}^{p_j} \alpha_{jk}, \\ \text{subject to } d_j \geq 0, \alpha_{jk} \geq 0, \end{cases} \tag{4}$$

where $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are the regularization parameters to control the degree of penalization of the groups and individuals, respectively. If $d_j = 0$, then all β_{jk} in the jth group will be equal to 0; if not, choosing proper λ_2 will make some of α_{jk} thus some of β_{jk} shrunken to zero. More precisely, the nonnegative hierarchical lasso can not only select important groups, but also identify important members of these groups, which is referred as bi-level selection.

(4) is an optimization problem w.r.t (d_j, α_{jk}). We show it can be written in an equivalent form w.r.t β_{jk}.

lemma 2.1. *If $(\hat{d}, \hat{\alpha})$ is a local minimizer of* (4)*, then $\hat{\beta}$ is a local minimizer of*

$$\begin{aligned}
&\min_{\beta_{jk} \geq 0} \ \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{J} \sum_{k=1}^{p_j} x_{i,jk} \beta_{jk} \right)^2 \\
(5) \qquad &+ 2\sqrt{\lambda} \sum_{j=1}^{J} \sqrt{\beta_{j1} + \beta_{j2} + \ldots + \beta_{jp_j}},
\end{aligned}$$

where $\lambda = \lambda_1 \cdot \lambda_2$, $\hat{\beta}_{jk} = \hat{d}_j \hat{\alpha}_{jk}$. On the other hand, if $\hat{\beta}$ is a local minimizer of (5)*, then define $(\hat{d}, \hat{\alpha})$ where $\hat{d}_j = 0$, $\hat{\alpha}_j = 0$ if $\hat{\beta}_j = 0$ and $\hat{d}_j = (\lambda_2 \sum_{k=1}^{p_j} \beta_{jk}/\lambda_1)^{1/2}$, $\hat{\alpha}_j = \hat{\beta}_j / \hat{d}_j$ if $\hat{\beta}_j \neq 0$ as a local minimizer of* (4)

Note that the penalty function of (5) is equivalent to a mixed $\left(1, \frac{1}{2}\right)$-penalty, namely the outer bridge penalty with bridge index $\gamma = 0.5$, and the inner lasso penalty so that it characterized by a concave group-level penalty and an individual variable-level 1-norm penalty. Note that both the lasso penalty and bridge penalty with $\gamma \leq 1$ possess the ability of variable selection. Therefore, the nonnegative hierarchical lasso can select not only important groups, but also the individuals.

3. ASYMPTOTIC PROPERTIES

3.1 Oracle properties for a diverging number of variables

In this section, we explore the theoretical properties of the nonnegative hierarchical lasso when the number of covariates is smaller than the sample size and show it has the group selection consistency, namely, under some regularization conditions, correctly selects true groups with asymptotic probability one with appropriate choice of regularization parameters. Meanwhile, we derive the asymptotic distribution of the estimators of the nonzero coefficients.

Without loss of generality, we assume

$$\begin{aligned}
&\beta_{A_j} \neq 0, 1 \leq j \leq J_1, \\
&\beta_{A_j} = 0, J_1 + 1 \leq j \leq J.
\end{aligned}$$

Let $B_1 = \cup_{j=1}^{J_1} A_j$ be the collection of the nonzero groups and $B_2 = B_1^c$ be that of zero groups. Let $\beta_{B_k} = (\beta_j, j \in B_k)^T, k = 1, 2$ and $\beta = (\beta_{B_1}^T, \beta_{B_2}^T)^T$. We assume $\beta^* = (\beta_{B_1}^{*T}, \beta_{B_2}^{*T})^T$ is the true coefficient with $\beta_{B_2}^* = 0$ and $\hat{\beta} = (\hat{\beta}_{B_1}^T, \hat{\beta}_{B_2}^T)^T$ as the nonnegative hierarchical estimates of β^*. Similarly, we denote $X(1)$ and $X(2)$ the submatrices of the design matrix X formed by columns in B_1 and its complement, respectively. Define

$$\begin{aligned}
C^n &= \frac{1}{n} X^T X, C_{11}^n = \frac{1}{n} X^T(1) X(1), \\
C_{22}^n &= \frac{1}{n} X^T(2) X(2), C_{A_j}^n = \frac{1}{n} X_{A_j}^T X_{A_j}.
\end{aligned}$$

Let λ_{max} and λ_{min} be the largest and the smallest eigenvalues of C^n, respectively.

Consider the following conditions:

(A1) The errors $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are uncorrelated with mean zero and finite variance σ^2.

(A2) The maximum multiplicity $C_n^* = \max_k \sum_{j=1}^{J} I_{\{k \in A_j\}}$ is bounded and

$$\begin{aligned}
(6) \qquad &\frac{\lambda}{n\lambda_{min}} \sum_{j=1}^{J_1} \frac{1}{\|\beta_{A_j}^*\|_1} |A_j| \leq \sigma^2 p M_n, \\
&M_n = O(1).
\end{aligned}$$

(A3) $\frac{\lambda(\lambda_{min}/p)^{\frac{3}{2}}}{\sqrt{n}\lambda_{max}^2} \to \infty$.

Condition (A1) is standard in linear regression model. Condition (A2) requests the l_1-norms of the coefficients in the nonzero groups are bounded away from zero. Moreover, both Condition (A2) and Condition (A3) require full rank design, that is, $\text{rank}(X) = p \leq n$. In this case, we still allow the number of features $p = p_n$ to grow with n under the assumption that the design matrix X is column full rank. If $\left\{B_1, \beta_{B_1}^*, J_1\right\}$ are fixed unknowns, then (A2)-(A3) are consequences of

$$\begin{aligned}
&C_n^* = O(1), \lambda_{min} = O(1), \lambda_{max} = O(1), \\
(7) \qquad &\frac{\lambda}{n} \to 0, \frac{\lambda}{p^{\frac{3}{2}}\sqrt{n}} \to \infty.
\end{aligned}$$

lemma 3.1. *Suppose conditions (A1) and (A2) hold, then nonnegative hierarchical lasso possesses the estimation consistency, that is*

$$(8) \qquad \|\hat{\beta} - \beta^*\|_2^2 \leq O_p\left(\frac{\sigma^2 p}{n\lambda_{min}}\right).$$

Theorem 3.1. *Under conditions (A1)-(A3), then the nonnegative hierarchical lasso has group selection consistency. That is*

$$(9) \qquad \mathbb{P}\left\{\hat{\beta}_{B_2} = 0\right\} \to 1.$$

Theorem 3.2. *Suppose $\left\{B_1, \beta_{B_1}^*, J_1\right\}$ are fixed unknowns and* (7) *holds. Suppose*

$$(10) \qquad C_{11}^n \to C_{11}, \frac{1}{\sqrt{n}} X^T(1)\epsilon \to_D W \sim N(0, \sigma^2 C_{11}).$$

Then

$$\hat{u} = \sqrt{n}\big(\hat{\beta}_{B_1} - \beta^*_{B_1}\big) \to_D$$
$$\begin{cases} C_{11}^{-1}W \sim N\big(0, \sigma^2 C_{11}^{-1}\big), \hat{u} \in D^o, \\ M_{j_1,\dots,j_k}W \sim N(0, D), \hat{u} \in D^o_{j_1,\dots,j_k}, \end{cases} \tag{11}$$

where $D = \sigma^2 C_{11}^{-1}(C_{11} - H(H^T C_{11}^{-1} H)^{-1} H^T)C_{11}^{-1}$, $M_{j_1,\dots,j_k} = C_{11}^{-1}[I - H(H^T C_{11}^{-1} H)^{-1} H^T C_{11}^{-1}]$, *$H$ denotes the $|B_1| \times k$ matrix with the main diagonal elements 1 and others 0. The definitions of D^o and $D^o_{j_1,\dots,j_k}$ are given in the appendix.*

remark 3.1. Theorem 3.1 suggests that the nonnegative hierarchical lasso estimates of the zero groups can be all dropped with probability converging to one. Theorem 3.2 shows the estimator of nonzero coefficients is root-n consistent and converges to a block-wise Gaussian distribution. Combine Theorem 3.1 and Theorem 3.2, imply that the nonnegative hierarchical estimator has the oracle property in group selection. The complete proof of Theorem 3.1 and Theorem 3.2 can be found in the Appendix.

3.2 Oracle inequalities and model consistency in high-dimensional linear model

To study the asymptotic properties of nonnegative hierarchical lasso in sparse ultra high-dimensional setting, namely, the number of features as well as the number of groups can be much larger than the sample size, the criterion (5) is rewritten as

$$\min_{\beta_{jk}\geq 0} \frac{1}{2n}\sum_{i=1}^{n}\Big(y_i - \sum_{j=1}^{J}\sum_{k=1}^{p_j} x_{i,jk}\beta_{jk}\Big)^2$$
$$+\lambda_n \sum_{j=1}^{J}\sqrt{\beta_{j1} + \beta_{j2} + \ldots + \beta_{jp_j}}, \tag{12}$$

where $m_1 \leq \max_{1\leq j\leq J} \|\beta_{A_j}\|_1^{\frac{1}{2}} \leq m_2$, the lower bound and upper bound of the square root of l_1 norm of coefficients in each group, respectively.

In what follows, we assume the noise is a vector of $i.i.d$ normal random variables so that it has Gaussian tails. As compared to the Non-Gaussian error, often used in the robust regression, it possesses some nice properties, such as the tail bound on a combination of $i.i.d$ normal random variables. With these properties, we can bound the random part $X_j^T\epsilon$, and further derive the oracle inequalities for the prediction and estimation errors.

Besides the notations used in section 3.1, we need some more notations to state the following results. We denote by $\|A\|_F$ and $|||A|||$ the Frobenius and spectral norms of matrix A, respectively. If A is positive semi-definite and $\lambda_1,\dots,\lambda_k$ are the eigenvalues of A, then we have $\|A\|_F = (\sum_{i=1}^k \lambda_i^2)^{1/2}$ and $|||A||| = \max_{i=1,\dots,k}\lambda_i$. We set $J(\hat{\beta}) = \{j : \hat{\beta}_j \neq 0, 1 \leq j \leq p\}$ with cardinality $M(\hat{\beta}) = |J(\hat{\beta})|$. Let $\Delta = \hat{\beta} - \beta^*, \Delta_{B_1} = \hat{\beta}_{B_1} - \beta^*_{B_1}, \Delta_{A_j} = \hat{\beta}_{A_j} - \beta^*_{A_j}$.

In our theoretical analysis, we make the following regularity conditions throughout.

(B1) For some integer $1 \leq s \leq p$, let

$$\mathcal{B} = \Big\{\Delta \in \mathbb{R}^p : \sum_{j=J_1+1}^{J}\|\Delta_{A_j}\|_1^{\frac{1}{2}} \leq 3\sum_{j=1}^{J_1}\|\Delta_{A_j}\|_1^{\frac{1}{2}}\Big\}. \tag{13}$$

Then the following conditions hold:

$$\kappa(s) = \min_{J_1\leq s}\min_{\substack{\|\Delta\|_2\neq 0\\ \Delta\in\mathcal{B}}}\frac{\|X\Delta\|_2}{\sqrt{n}\|\Delta_{B_1}\|_1} > 0. \tag{14}$$

(B2) The error ϵ is a vector of $i.i.d$ normal random variables with mean 0 and variance σ^2.

(B3) The group number J tends to infinity with n at a rate $\lim_{n\to\infty} log(J)/n = 0$, i.e., $J = e^{n^{a_1}}$ for some $0 < a_1 < 1$

(B4) The maximum multiplicity $C_n^* = \max_k \sum_{j=1}^J I_{\{k\in A_j\}}$ is bounded and there exists a constant $0 \leq a_2 < 1$ such that $\max_{1\leq j\leq J} p_j = O\big(n^{a_2}\big)$.

(B5) There exist a constant $a_3 \in (0, \min\{1-a_1, 1-a_2\})$ such that $\lambda_{max} = O\big(n^{a_3}\big)$.

Condition (B1) is called the restricted eigenvalue assumption on the Gram matrix of X. The integer s plays a role of an upper bound on the group sparsity. This is an extension to our settings of the RE assumption for the usual lasso and Dantzig selector from Bickel et al. [16] and group sparsity from Lounici et al. [17]. RE assumption is widely used in high-dimensional setting to establish non-asymptotic error bounds, which is milder than the incoherence condition. It is equivalent to lower bounding the restricted ℓ_1-eigenvalues of the sample covariance matrix X^TX/n. Since it is a simple form of restricted strong convexity condition for the least-square loss, then follow the discussion in [18], RE assumption holds with high probability for various classes of random design matrix. For example, for a random design matrix X with rows $x_i \in \mathbb{R}^p$ drawn independently and identically distributed from a zero mean sub-Gaussian distribution with covariance Σ, then RE assumption hold with high probability. Condition (B3) implies the number of groups can be much larger than the sample size and allows the dimensionality grows at a exponential rate as the sample size increases. Condition (B4) and (B5) assume the maximal group size and the largest eigenvalue of the sample covariance matrix are bounded at some rate as n grows.

lemma 3.2. *For every $j \in 1, 2, \dots, J$, recall that $C^n_{A_j} = X^T_{A_j}X_{A_j}/n$ and choose*

$$\lambda_n \geq \max_{1\leq j\leq J}\Big\{\frac{2\sigma m_2}{\sqrt{n}}\big(tr\big(C^n_{A_j}\big) + 2|||C^n_{A_j}|||$$
$$\Big(2\gamma log(J) + \sqrt{p_j\gamma log(J)}\Big)^{1/2}\Big\}, \tag{15}$$

where $\gamma > 1$. Then with probability at least $1 - 2J^{1-\gamma}$, for any solution $\hat{\beta}$ of criterion (12) *we have*

$$\sum_{j=J_1+1}^{J} \|\Delta_{A_j}\|_1^{\frac{1}{2}} \le 3 \sum_{j=1}^{J_1} \|\Delta_{A_j}\|_1^{\frac{1}{2}}, \tag{16}$$

$$\frac{\|X\Delta\|_2^2}{n} \le 3\lambda_n \sum_{j=1}^{J_1} \|\Delta_{A_j}\|_1^{\frac{1}{2}}. \tag{17}$$

Theorem 3.3. *Under conditions (B1)-(B3). If $M(\beta^*) \le s$ and the regularization parameter λ_n is chosen such that* (15) *holds with $\gamma > 1$, then with probability at least $1 - 2J^{1-\gamma}$, we have that*

$$\frac{1}{n}\|X\Delta\|_2^2 \le \frac{3^{\frac{4}{3}} J_1^{\frac{2}{3}} C_n^{*\frac{1}{3}} \lambda_n^{\frac{4}{3}}}{\kappa^{\frac{2}{3}}(s)}, \tag{18}$$

$$\|\Delta_{B_1}\|_1 \le \frac{3^{\frac{2}{3}} J_1^{\frac{1}{3}} C_n^{*\frac{2}{3}} \lambda_n^{\frac{2}{3}}}{\kappa^{\frac{4}{3}}(s)}, \tag{19}$$

$$\|\Delta\|_1 \le 16 \frac{3^{\frac{2}{3}} J_1^{\frac{1}{3}} C_n^{*\frac{8}{3}} \lambda_n^{\frac{2}{3}}}{\kappa^{\frac{4}{3}}(s)}, \tag{20}$$

$$M(\hat{\beta}) \le 4m_2^2 \lambda_{max} \frac{3^{\frac{4}{3}} C_n^{*\frac{4}{3}} J_1^{\frac{2}{3}}}{\lambda_n^{\frac{2}{3}} (1 - \frac{\lambda_n}{m_1})^2 \kappa^{\frac{2}{3}}(s)}, \tag{21}$$

where λ_{max} denotes the largest eigenvalue of the Gram matrix of X.

remark 3.2. Inequalities (18)-(21) in Theorem 3.3 are called the oracle inequalities which give non-asymptotic bounds on the prediction and l_1 estimation loss with some probability. The parameter γ controls the probability under which the above inequalities hold.

Theorem 3.4. *Suppose $\{B_1, \beta^*_{B_1}, J_1\}$ are fixed unknowns and conditions (B1)-(B5) hold. Set*

$$\lambda_n = \frac{4\sqrt{2}\hat{\sigma}\sqrt{\lambda_{max}} m_2}{\sqrt{n}} \left(\max_{1 \le j \le J} p_j + Alog(J) \right)^{\frac{1}{2}}, \tag{22}$$

where $A \ge 5/2$ and $\hat{\sigma}$ is some estimator of σ. Then with probability at least $1 - 2J^{1-2A/5}$ we have
(i) group selection consistency:
$\mathbb{P}(\hat{\beta}_{B_2} = 0) \to 1$;
(ii) estimation consistency:
$\|\hat{\beta} - \beta^*\|_1 \le O_p\left(\frac{\sigma^2}{\kappa^3 \kappa(s)} \frac{\max_{1 \le j \le J} p_j + Alog(J)}{n} \right)^{1/3}$.

remark 3.3. Taking the regularization parameter λ of order $((\max_{1 \le j \le J} p_j + Alog(J))/n)^{1/2}$, theorem 3.4 indicates that the nonnegative hierarchical lasso enjoys the group selection consistency when the number of covariates is much larger than the sample size and establishes the upper bound for l_1 estimation errors.

remark 3.4. Since σ is unknown in general, we need a well-chosen estimator $\hat{\sigma}$ of σ. It can be shown that we can take $\hat{\sigma} = y^T y/n$ after centering y, which we can refer to the section 6.2 in [15] for more details.

4. ESTIMATION ALGORITHM

Since the optimization problem (4) w.r.t (d_j, α_{jk}) requires to choose the values of two tuning parameters and it is computationally expensive, we consider the optimization w.r.t β_{jk} as follows

$$\min_{\beta_{jk} \ge 0} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{J} \sum_{k=1}^{p_j} x_{i,jk}\beta_{jk} \right)^2 + \lambda \sum_{j=1}^{J} \sqrt{\beta_{j1} + \beta_{j2} + \ldots + \beta_{jp_j}},$$

multiplying by a positive parameter μ both sides of the above equation gives

$$\min_{\beta \ge 0} \mu\|y - X\beta\|_2^2 + \lambda\mu \sum_{j=1}^{J} \|\beta_{A_j}\|_1^{1/2},$$

where μ is some rescaling factor.

To this end, we propose a fast and efficient iterative half thresholding based local linear approximation algorithm(IHT-LLA) for solving the nonnegative hierarchical lasso problem, which is based on the iterative half thresholding algorithm in Xu et.al. [22], group coordinate descent in Wei and Zhu [28] and local linear approximation in Zou and Li [5].

Following the half threshold algorithm in Xu et.al. [22], for the usual $L_{1/2}$ regularization, the thresholding representation is defined as follows:

$$\hat{\beta}^{(k+1)} = H_{\lambda_k\mu_k, \frac{1}{2}}\left(\hat{\beta}^{(k)} + \mu_k X^T(y - X\hat{\beta}^{(k)}) \right),$$

where $H_{\lambda\mu,1/2}(x) = (h_{\lambda\mu,1/2}(x_1), \ldots, h_{\lambda\mu,1/2}(x_n))$ and

$$h_{\lambda\mu,\frac{1}{2}}(x) = \begin{cases} f_{\lambda\mu,\frac{1}{2}}(x), & |x| > \frac{\sqrt[3]{54}}{4}(\lambda\mu)^{2/3}, \\ 0, & otherwise, \end{cases}$$

with $f_{\lambda\mu,1/2}(x) = 2x(1 + \cos(2\pi/3 - 2\varphi(x)/3))/3$ and $\varphi(x) = \arccos(\lambda\mu(|x|/3)^{-3/2}/8)$.

If $\hat{\beta}_{A_j} = 0, 1 \le j \le J$, the group penalty of the jth group is reduced to the usual $L_{1/2}$ penalty. Therefore, following the idea of iterative half thresholding algorithm, we get the threshold value for $\hat{\beta}_{A_j} = 0, 1 \le j \le J$. If $\hat{\beta}_{A_j} \ne 0, 1 \le j \le J$, then adopting a local linear approximation, the optimization problem with respect to the jth group is closely related to an iteratively reweighted nonnegative adaptive lasso procedure. Suppose $\beta^{(0)} \ge 0$ be the initial estimator, if $\beta^{(0)}_{A_j} = 0$, then set $\hat{\beta}_{A_j} = 0$. Otherwise they can be locally approximated by a linear function as

$$P_\lambda(\beta_{A_j}) \approx P_\lambda(\beta^{(0)}_{A_j}) + \frac{\lambda}{2\|\beta^{(0)}_{A_j}\|_1^{1/2}} \sum_{k=1}^{p_j} (\beta_{jk} - \beta^{(0)}_{jk}). \tag{23}$$

It can be regarded as an nonnegative adaptive lasso problem at each iterative step and can be solved efficiently using a revised coordinate-wise descent algorithms from [6] when suppose the estimator is nonnegative. Assume that the x_{ij} are standardized so that $\sum_i x_{ij}/n = 0, \sum_i x_{ij}^2 = 1$, then one can show that the coordinate-wise update has the form

$$\hat{\beta}_{jk} \leftarrow \left(\beta^{(0)}_{jk} + x_{jk}^T (y - X\beta^{(0)}) - \lambda/4\|\beta^{(0)}_{A_j}\|_1^{1/2} \right)_+ .$$

Obviously, the above algorithm has a blockwise coordinate descent structure. It optimizes a target function with respect to a single group at a time, iterate through all the groups until convergence. Therefore, the IHT-LLA iterates as follows:

1. Center y. Center and normalize x_{jk}.
2. Initialize $\beta^{(0)}_{jk}$.
3. Set $\mu_t = \mu_0 = 1/|||X^T X/n|||^2$, the square of spectral norm of the Gram matrix of X, and λ_t can be chosen from BIC criteria.
4. For the jth group, if $\hat{\beta}^{(t)}_{jk} + \mu_t X^T (y - X\hat{\beta}^{(t)}) \le \sqrt[3]{54}(\lambda_t \mu_t)^{2/3}/4, k = 1, 2 \dots, p_j$, then set $\hat{\beta}^{(t+1)}_{A_j} = 0$. Otherwise, solving the following nonnegative adaptive lasso problem

$$\hat{\beta}^{(t+1)}_{A_j} = \underset{\beta_{A_j} \ge 0}{\arg\min}\, \mu_t \|y^* - X_{A_j}\beta_{A_j}\|_2^2 + \frac{\lambda_t \mu_t}{2\|\beta^{(t)}_{A_j}\|_1^{1/2}} \sum_{k=1}^{p_j} \beta_{jk}, \tag{24}$$

where $y^* = y - \sum_{k \neq j} X_{A_k} \beta^{(t)}_{A_k}$.

5. Repeat step 4 until convergence, e.g., $\|\hat{\beta}^{(t+1)} - \hat{\beta}^{(t)}\|_2 \le \epsilon$, where ϵ is some tolerance level, e.g., $\epsilon = 10^{-5}$.

5. SIMULATION STUDIES

In this section, we use some simulations to demonstrate the finite sample performance of the nonnegative hierarchical lasso both in low-dimensional and high-dimensional settings and compare the results with nonnegative lasso, nonnegative adaptive lasso and nonnegative elastic-net.

5.1 Simulation

Similar to Huang et al. [13], we consider the following settings:

case1: In this case, there are 16 groups, each with 5 predictors. Suppose that the between-group correlation is 0 while the within-group is 0.5, i.e., each group is simulated independently by a multivariate normal random vector $N(0, \Sigma)$ with Σ being the covariance matrix of each group such that all the non-diagonal entries equal to 0.5 and the diagonal elements 1. Set $\epsilon \sim N(0, 1)$. Since there exists bi-level hierarchical structure in our estimator and the group level $d_j \propto \|\beta_{A_j}\|_1^{1/2}$, we let

$$d = (\sqrt{10}, \sqrt{8.5}, \sqrt{2.5}, \underbrace{0, \dots, 0}_{13}),$$

and

$$\beta = (\underbrace{0.5, 2, 2, 2.5, 3}_{5}, \underbrace{1, 1.5, 2, 2, 2}_{5}, \underbrace{1.5, 1, 0, 0, 0}_{5}, \underbrace{0, \dots, 0}_{65}).$$

case2: In this case, there are 100 groups, each with 5 predictors. In each group, the coefficients are either all nonzero or all zero. Like Huang et al [13], we first simulate $R_j, j = 1, \dots, 500$ independently from the standard normal distribution. Then we generate $Z_j, j = 1, \dots, 100$ from the normal distribution with $cov(Z_{j_1}, Z_{j_2}) = 0.5^{|j_1 - j_2|}$ for $j_1, j_2 = 1, \dots, 100, j_1 \neq j_2$. To normalize, the predictors $(X_1, \dots, X_{500})$ are generated by

$$X_{5(j-1)+k} = Z_j/2 + \sqrt{3}R_{5(j-1)+k}/2, \quad j = 1, \dots, 100, k = 1, \dots, 5.$$

We still set $\epsilon \sim N(0, 1)$ and

$$d = (\sqrt{7.5}, \sqrt{7.5}, \sqrt{10}, \sqrt{10}, \sqrt{5}, \underbrace{0, \dots, 0}_{95}),$$

and

$$\beta = (\underbrace{1.5, \dots, 1.5}_{5}, \underbrace{1.5, \dots, 1.5}_{5}, \underbrace{2, \dots, 2}_{5}, \underbrace{2, \dots, 2}_{5}, \underbrace{1, \dots, 1}_{5}, \underbrace{0, \dots, 0}_{475}).$$

case3: Both the predictors and random errors are generated similarly to case2 except that there exists both the important and unimportant variables in some groups. Let

$$d = (\sqrt{5}, \sqrt{10}, 2, \sqrt{3}, 1, \underbrace{0, \dots, 0}_{95}),$$

and

$$\beta = (\underbrace{1, 1, 1, 1, 1}_{5}, \underbrace{2, 2, 2, 2, 2}_{5}, \underbrace{1, 1, 1, 1, 0}_{5}, \underbrace{1.5, 1.5, 0, 0, 0}_{5}, \underbrace{1, 0, 0, 0, 0}_{5}, \underbrace{0, \dots, 0}_{475}).$$

case4: Both the predictors and random errors are simulated in the same way as in case3 except that the sample size is increased from 5 to 20 while the number of groups is decreased to 25. Postulating the number of nonzero coefficients in the 1,3,5 groups is generated independently by a discrete uniform distribution $DU(1, 20)$. To be general, the nonzero components from the first group are simulated independently by a normal $N(1.5, 0.3)$ random variable. Similarly, the nonzero components from the third group are simulated independently by a normal $N(2, 0.2)$ random variable and the fifth by a normal $N(2.5, 0.1)$ random variable. Set $\beta_{61} = \beta_{62} = \cdots = \beta_{500} = 0$.

case5: In this case, the model has groups of different sizes. Set $p = 500, J = 29$ with group sizes $p_{1-4} = 5, p_{5-8} = 15, p_{9-29} = 20$. In each group, the coefficients are either all non-zero or all zero. The data are generated as follows. The predictors $(X_1, \ldots, X_{500})$ are generated by a multivariate normal random vector $N(0, \Sigma)$. Let $\Sigma_{p\times p} = diag(\Sigma_{1,2}, \Sigma_{3,4}, \Sigma_5, \ldots, \Sigma_{29})$, with $\Sigma_{1,2}$ being the covariance matrix for groups 1 and 2 and $\Sigma_{3,4}$ for groups 3 and 4, $\Sigma_j, j = 5, \ldots, 28$ being the covariance matrix for groups $j = 5, \ldots, 28$. Set $(\Sigma_{1,2})_{ij} = 1_{i=j} + 0.5_{1\leq i\neq j\leq 5} + 0.5_{6\leq i\neq j\leq 10} + 0.3_{1\leq i\leq 5, 6\leq j\leq 10} + 0.3_{6\leq i\leq 10, 1\leq j\leq 5}$ such that within-group correlation is 0.5 and between-group correlation is 0.3 for group 1 and 2. Similarly, $(\Sigma_{3,4})_{ij} = 1_{i=j} + 0.5_{1\leq i\neq j\leq 5} + 0.5_{6\leq i\neq j\leq 10} - 0.5_{1\leq i\leq 5, 6\leq j\leq 10} - 0.5_{6\leq i\leq 10, 1\leq j\leq 5}$ such that within-group correlation is 0.5 and between-group correlation is -0.5 for group 3 and 4. For $\Sigma_j, j = 5, \ldots, 8$, we choose a compound symmetry structure with $\rho = 0.5$ while for $\Sigma_j, j = 9, \ldots, 29$ we choose $\rho = 0.2$. The response is calculated based on model (1) with

$$\beta = (\underbrace{1, \ldots, 1}_{5}, \underbrace{0, \ldots, 0}_{5}, \underbrace{2, \ldots, 2}_{5}, \underbrace{0, \ldots, 0}_{485}).$$

case6: Both the predictors and random errors are simulated in the same way as in case3 except that the number of groups is increased to 200 so that there are 1000 predictors. Let

$$d = (\sqrt{7.5}, \sqrt{8}, \underbrace{0, \ldots, 0}_{198}),$$

and

$$\beta = (\underbrace{1.5, 1.5, 1.5, 1.5, 1.5}_{5}, \underbrace{2, 2, 2, 2, 0}_{5}, \underbrace{0, \ldots, 0}_{990}).$$

case7: The same as case6, except that $\epsilon \sim t(8)$, a t distribution with degree of freedom equal to 8.

case8: The same as case6, except that $\epsilon \sim 0.9N(0, 1) + 0.1N(0, 10)$, a mixed normal distribution.

For these cases, all the simulations are repeated 100 times randomly. Within each replication, our simulated data consists of a training set of size 100. To show the performance of nonnegative hierarchical lasso compared to nonnegative lasso, nonnegative adaptive lasso and nonnegative elastic-net, we use BIC criterion to select the tuning parameters and the simulation results are summarized in Table 1. L1 is the l_1-norm of estimation errors in group/individual level, namely, $\|\hat{d} - d^*\|_1$ and $\|\hat{\beta} - \beta^*\|_1$. Similarly, L2 is the l_2-norm of estimation errors in group/individual level, namely, $\|\hat{d} - d^*\|_2^2$ and $\|\hat{\beta} - \beta^*\|_2^2$. ME represents the model error which is computed as $(\hat{\beta} - \beta^*)^T \mathbb{E}(XX^T)(\hat{\beta} - \beta^*)$. FP counts the the number of groups/variables that are false positive. Since the false negative, i.e., the number of nonzero groups or individuals which are not selected is very small, we omit it here. In the parentheses are the corresponding standard deviations.

From Table 1, with any group size and group structure, it can be seen that the nonnegative hierarchical lasso is significantly superior to, and very occasionally inferior to the other methods in group-level both in low-dimensional and high-dimensional set-ups. When the group size is relatively small, this superiority is more obvious. In individual-level, nonnegative hierarchical lasso performs as well as nonnegative adaptive lasso for low dimensionality but attains a lower L1, L2 and VF as the dimensionality grows. As for ME, nonnegative hierarchical lasso also perform the best, followed by nonnegative adaptive lasso, nonnegative elastic-net and nonnegative lasso. Similar results can be obtained under two types of Non-Gaussian error-a t distribution $t(8)$ and a mixed normal distribtion $0.9N(0, 1) + 0.1N(0, 10)$.

5.2 Simulation with group mis-specification

Set $p = 600, J = 60$ with all of the group sizes equal to 10. In each group, the coefficients are either all non-zero or all zero. The predictors $(X_1, \ldots, X_{600})$ are generated independently by a standard normal distribution $N(0, 1)$. The response is calculated based on model (1) with

$$\beta = (\underbrace{0.5, \ldots, 0.5}_{10}, \underbrace{0, \ldots, 0}_{10}, \underbrace{0.5, \ldots, 0.5}_{10}, \underbrace{0, \ldots, 0}_{570}).$$

The group information is mis-specified such that the casual variables $X_9 - X_{10}$ are grouped with the null variables $X_{11} - X_{20}$ and the casual variables $X_{29} - X_{30}$ are grouped with the null variables $X_{31} - X_{40}$. Thus, the group sizes are changed to $p_1 = 8, p_2 = 12, p_3 = 8, p_4 = 12, p_{5-60} = 10$. The percentage of $X_9 - X_{10}$ and $X_{29} - X_{30}$ being selected over the 100 replications and their variation ranges are reported in Table 2. To examine the selection abilities of all these methods, we plot the percentage of the 100 replications when a nonzero coefficient is selected. The results are given in Figure1.

5.3 Computational efficiency

In this section we briefly assess the computational efficiency of our algorithm IHT-LLA. We compare IHT-LLA with the hierarchically iterative algorithm (HI) used in [12]. The simulation data is generated as follows: we fix $n = 100$

Table 1. Simulation results for case1-8

Method	ME	Groups			Variables		
		L1	L2	FP	L1	L2	FP
$n = 100, p = 80, J = 16, \epsilon \sim N(0, 1)$							
NLasso	0.25(0.09)	0.75(0.55)	0.19(0.20)	2.38(2.05)	1.72(0.39)	0.32(0.12)	3.45(2.88)
NEnet	0.23(0.09)	0.61(0.50)	0.15(0.18)	1.99(1.78)	1.75(0.40)	0.34(0.13)	3.08(2.26)
NALasso	0.16(0.07)	0.33(0.36)	0.09(0.13)	0.71(1.03)	1.50(0.36)	0.27(0.11)	1.12(1.36)
NHLasso	0.16(0.06)	0.22(0.35)	0.07(0.18)	0.28(0.68)	1.50(0.38)	0.26(0.12)	1.11(1.27)
$n = 100, p = 500, J = 100, \epsilon \sim N(0, 1)$							
NLasso	0.66(0.17)	3.91(1.97)	1.11(0.68)	13.85(7.21)	4.87(0.88)	1.04(0.30)	15.85(8.86)
NEnet	0.64(0.20)	4.38(2.09)	1.33(0.72)	14.91(8.02)	5.29(0.95)	1.21(0.40)	17.26(9.74)
NALasso	0.44(0.14)	2.18(1.12)	0.63(0.40)	6.79(3.79)	3.92(0.74)	0.80(0.27)	7.26(4.21)
NHLasso	0.35(0.12)	1.09(0.90)	0.56(0.54)	1.61(1.55)	3.22(0.83)	0.58(0.23)	3.27(3.37)
$n = 100, p = 500, J = 100, \epsilon \sim N(0, 1)$							
NLasso	0.60(0.18)	2.70(1.40)	0.69(0.46)	9.28(5.45)	3.41(0.62)	0.70(0.20)	11.48(6.66)
NEnet	0.49(0.11)	3.76(1.48)	1.10(0.65)	13.50(4.37)	3.82(0.97)	0.76(0.29)	16.76(6.01)
NALasso	0.43(0.33)	1.98(2.67)	0.68(1.14)	5.20(7.38)	2.82(1.45)	0.70(0.57)	6.17(8.87)
NHLasso	0.22(0.08)	0.24(0.19)	0.05(0.14)	0.06(0.24)	2.01(0.48)	0.32(0.13)	2.99(1.67)
$n = 100, p = 500, J = 25, \epsilon \sim N(0, 1)$							
NLasso	0.51(0.16)	3.04(3.07)	2.88(5.58)	5.82(2.99)	3.58(0.74)	0.74(0.25)	12.01(6.62)
NEnet	0.52(0.15)	2.96(3.14)	2.86(5.58)	5.33(3.40)	3.65(0.80)	0.78(0.25)	11.16(7.35)
NALasso	0.42(0.12)	2.63(2.54)	2.02(4.46)	5.33(2.44)	3.77(0.67)	0.89(0.31)	11.85(4.64)
NHLasso	0.29(0.10)	1.67(3.54)	2.48(5.84)	0.89(1.96)	2.67(0.59)	0.49(0.18)	5.13(2.89)
$n = 100, p = 500, J = 29, \epsilon \sim N(0, 1)$							
NLasso	0.39(0.15)	0.98(0.69)	0.24(0.25)	3.38(2.51)	1.82(0.37)	0.41(0.17)	4.20(3.59)
NEnet	0.27(0.11)	0.83(0.54)	0.22(0.18)	2.69(2.16)	1.75(0.36)	0.38(0.16)	3.17(2.46)
NALasso	0.23(0.10)	0.79(0.53)	0.22(0.19)	2.40(1.76)	1.73(0.39)	0.40(0.18)	2.86(2.16)
NHLasso	0.11(0.05)	0.16(0.29)	0.07(0.17)	0.23(0.51)	1.19(0.32)	0.22(0.09)	0.51(1.28)
$n = 100, p = 1000, J = 200, \epsilon \sim N(0, 1)$							
NLasso	0.40(0.09)	0.80(0.61)	0.17(0.17)	2.98(2.85)	1.48(0.27)	0.31(0.09)	3.15(3.00)
NEnet	0.17(0.09)	0.67(0.52)	0.16(0.18)	2.68(1.95)	1.23(0.31)	0.23(0.09)	2.86(1.99)
NALasso	0.15(0.07)	0.82(0.81)	0.27(0.30)	2.34(2.53)	1.23(0.38)	0.21(0.10)	2.34(2.54)
NHLasso	0.12(0.06)	0.29(0.54)	0.12(0.30)	0.45(0.99)	1.03(0.38)	0.16(0.09)	1.07(2.14)
$n = 100, p = 1000, J = 200, \epsilon \sim t(8)$							
NLasso	0.59(0.24)	1.21(0.79)	0.29(0.23)	4.41(3.23)	1.84(0.36)	0.42(0.15)	4.68(3.41)
NEnet	0.41(0.20)	1.24(0.75)	0.31(0.24)	4.54(2.91)	1.75(0.39)	0.38(0.15)	5.01(3.13)
NALasso	0.23(0.12)	1.52(1.28)	0.58(0.53)	4.10(3.60)	1.58(0.63)	0.29(0.16)	4.11(3.61)
NHLasso	0.19(0.09)	0.51(0.54)	0.25(0.31)	0.83(1.00)	1.26(0.46)	0.22(0.12)	1.84(1.89)
$n = 100, p = 1000, J = 200, \epsilon \sim 0.9N(0, 1) + 0.1N(0, 10)$							
NLasso	0.95(0.36)	1.82(1.47)	0.48(0.48)	5.87(5.34)	2.66(0.57)	0.81(0.26)	6.06(5.56)
NEnet	0.89(0.34)	1.90(1.41)	0.49(0.49)	5.87(4.98)	2.69(0.56)	0.83(0.25)	6.15(5.41)
NALasso	0.35(0.14)	1.92(1.21)	0.68(0.55)	4.67(3.11)	2.03(0.69)	0.46(0.22)	4.71(3.13)
NHLasso	0.34(0.19)	1.08(1.05)	0.55(0.67)	1.66(1.73)	1.82(0.79)	0.43(0.24)	3.19(3.39)

Table 2. Simulation results for group mis-specification

	Nlasso	Nnet	Nalasso	Nhlasso
pct.X9	0.21	0.20	0.69	1.00
pct.X10	0.18	0.17	0.67	1.00
pct.X29	0.17	0.18	0.63	1.00
pct.X30	0.18	0.16	0.58	0.98
range.X9	0.097-0.553	0.098-0.554	0.060-0.809	0.156-0.853
range.X10	0.010-0.606	0.031-0.605	0.003-0.680	0.220-0.743
range.X29	0.031-0.452	0.031-0.498	0.117-0.261	0.074-0.172
range.X30	0.036-0.559	0.051-0.559	0.046-0.131	0.009-0.397

with the first 50 samples serve as a training set whereas the remaining as the test set, and vary the value of p from 60 to 480 with group size all equal to 3. That is, the number of groups is increased from 20 to 160. The covariates $(x_1, \ldots, x_p)$ are simulated from a multivariate normal distribution with the pairwise correlation between x_i and x_j set to be $corr(i, j) = 0.5^{|i-j|}$. The model built on the training set can be expressed as

$$y_i = \sum_{j=1}^{8} x_{ij}\beta_j^* + \epsilon_i, i = 1, \ldots, 50,$$

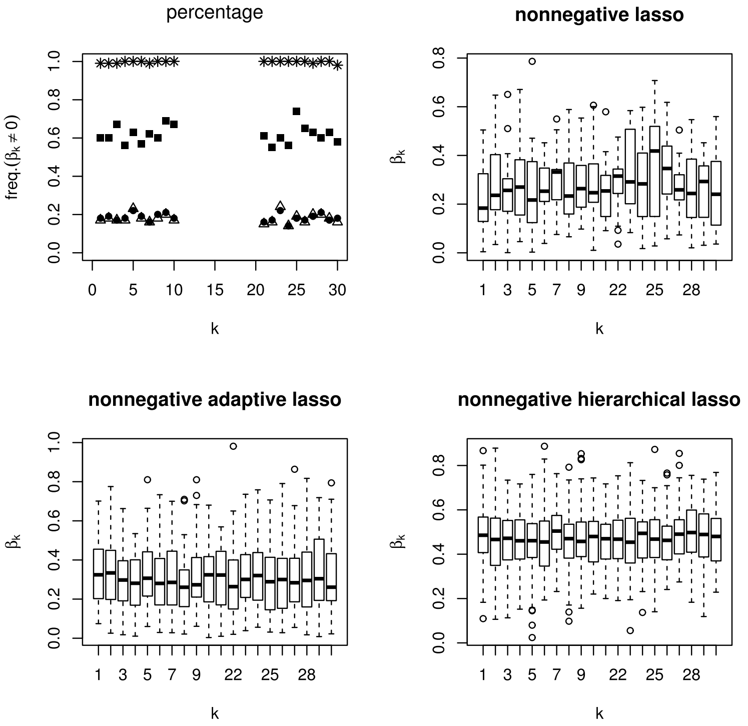

Figure 1. Top left: frequency of X1-X10 and X30-X40 being selected of the nonnegative lasso, nonnegative adaptive lasso and nonnegative hierarchical lasso estimates. Solid dot: nonnegative lasso. Triangle: nonnegative adaptive lasso. Solid square: nonnegative adaptive lasso. Star: nonnegative hierarchical lasso. Top right: box plot of nonnegative lasso. Bottom left: box plot of nonnegative adaptive lasso. Bottom right: box plot of nonnegative hierarchical lasso.

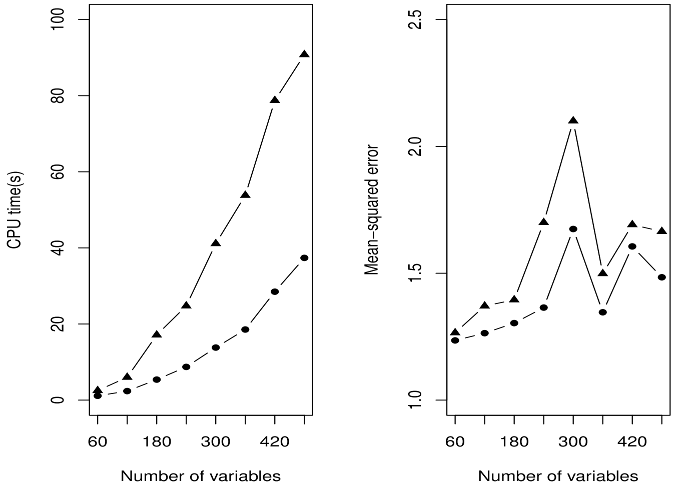

Figure 2. Left: CPU times (in seconds) required to fit the entire coefficients path. Right: mean-squared error on the test set. Solid dot: ITH-LLA. Triangle: HI.

where $\epsilon_i \sim N(0,1)$. In every simulation the nonzero coefficients $\beta_j^*, j = 1, \ldots, 8$ are all equal to 1.5. Both IHT-LLA and HI do their works in R. To show the efficiency of our algorithm, Figure2 presents the CPU times required to fit the entire coefficients path and the mean-squared error (MSE) on the test set, all averaged over 10 runs.

As we shall see, our algorithm IHT-LLA is faster than the hierarchically iterative algorithm (HI), and performs better in terms of prediction accuracy.

6. AN APPLICATION

In this section, we focus on the application of the nonnegative hierarchical lasso in financial market. The performance of the nonnegative hierarchical lasso with grouping information considered is tested to track the S&P 500 index, a market-capitalization-weighted index of the 500 largest U.S. publicly traded companies by market value to represent the largest publicly traded corporations in America.

We apply the nonnegative hierarchical lasso to index tracking mainly for the following:

Firstly, it is costly to select all of the assets. In application, one always construct a sparse index tracking portfolio against the costly full replication. On the other hand, a large quantity of stocks also cause the problem of high-dimensionality. Thus, our method is a fast and effective way to induce sparsity.

Secondly, nonnegative constraints are the conventional restraints especially in financial field as the short-sale constraints. Since the index is compiled via the market capitalization weight which is positive, it is natural to introduce the nonnegative constraints for better replication.

Finally, there exist group structures in stock market. For instance, there are relationships between stocks from the same stock block or industry. Incorporation of such information can improve the predictive accuracy. The nonnegative hierarchical lasso, as a bi-level selection method, gains the ability to achieve this aim.

We use the proposed nonnegative hierarchical lasso to analyze the data from the closing prices of stocks that make up the S&P 500 index, from Feb. 1, 2018 to Sept. 26, 2018. There are 159 observations, and 504 predictors in this data. To demonstrate the predictive accuracy of the proposed approach, the data are divided into two parts, the first 119 observations are regarded as a training set and the remaining are used as a test set. In this work, we let $P_{i,j}$ be the prices of the jth constituent stock and p_i be the S&P 500 index. To eliminate the relativity of observations, we transform the closing price to the daily return, i.e.

$$x_{ij} = \frac{P_{i-1,j}}{P_{i,j}} - 1, \quad i = 2, 3, \ldots, n,$$

$$y_i = \frac{p_{i-1}}{p_i} - 1, i = 2, 3, \quad \ldots, n,$$

where x_{ij} and y_i stand for the return of the jth constituent stock and S&P 500 index, respectively. Therefore, the statistical model built on the training set can be described to a linear regression model:

$$\mathrm{y}_i = \sum_{j=1}^{504} x_{ij}\beta_j^* + \varepsilon_i, \quad i = 1, \ldots, 118, \quad s.t. \quad \beta_j^* \geq 0,$$

where β is assumed to be sparse for sparse replication.

Since the true grouping structure is unknown in practice, we consider a data-driven clustering method which provides an adaptively choice of the number of groups depending on the size of selected model. Note that the hierarchical lasso is a bi-level selection procedure with group-level $d_j \propto \|\beta_{A_j}\|_1^{1/2}, \ j = 1, \ldots, J$ and we conjecture the reasonability of clustering two variables into the same cluster with almost the same size of group-level $d_j, \ j = 1, \ldots, J$. Therefore, our clustering method can be described as follows:

(a) Let each group consists of one variable and use the nonnegative hierarchical lasso to get the solution $\hat{d}_j, \ j = 1, \ldots, J$ for each variable.

(b) Fix the number of groups, e.g., 50,60,70, ...,200, and the membership of groups could be determined by the $\hat{d}_j, \ j = 1, \ldots, J$ from (a).

(c) For a given size of selected model, we determine the number of groups by a CV-like method, e.g., divide the dataset into 5 parts, four of which as the training set and the remaining as the test set. Then we can model the training set by nonnegative hierarchical lasso to compare the mean square error of the test set, respectively. Denote the corresponding number of groups with the smallest MSE by k^*.

(d) Choose the number of groups around k^*, e.g., $k^* - 10, k^* - 9, \ldots, k^* + 9, k^* + 10$, and do (b) and (c) again to find the final optimal number of groups.

For comparison, we also considered the nonnegative lasso, nonnegative adaptive lasso and nonnegative elastic-net from Wu and Yang [34], Yang et al. [35] and Wu et al. [33]. The BIC is also used to select the regularization parameters of these three methods. For evaluation purpose, the Absolute Mean Tracking Error (AMTE), is defined as

$$TrackingError_{Mean} = \frac{\sum(|err_t|)}{T},$$

where $err_t = y_t - \hat{y}_t$ and $\hat{y}_t$ is the fitted or predicted value of y_t, for $t = 1, 2, \ldots, T$.

Our aim is to resort to a small subset of the constituent stocks to replicate the index. Thus, we tune the tuning parameter to select 35 or 50 stocks to demonstrate the fitted and predicted results of our approach as compared to nonnegative lasso, nonnegative adaptive lasso and nonnegative elastic-net. By the aforementioned data-driven clustering method, we set the number of groups is 117(or equivalently,72) when the size of selected model is 35 (or equivalently,50). The results for Absolute Mean Tracking Error (AMTE) are given in Table 3. We only show the predicted results with different number of selected stocks (35 VS 50) in Figure3-4. Since the predicted result of nonnegative elastic-net is thus similar to that of nonnegative lasso, it is omitted for brevity. It can be seen from Table 3 that the nonnegative hierarchical lasso outperforms other methods in terms of AMTE, then the nonnegative adaptive lasso, and worst the nonnegative elastic-net and nonnegative lasso. Also, increasing the number of selected stocks could slightly improve the performance of prediction. Similar conclusions can be drawn from the predicted results in Figure3-4 and the nonnegative hierarchical lasso could be a better choice for long-term prediction, which can be shown in Figure3-4.

Table 3. The fitted and predicted Absolute Mean Tracking Error (AMTE)

	non-lasso	non-elastic net	non-adaptive lasso	non-hierarchical lasso
Fitted AMTE(35)	0.157%	0.155%	0.074%	0.069%
Predicted AMTE(35)	0.135%	0.135%	0.107%	0.086%
Fitted AMTE(50)	0.087%	0.086%	0.043%	0.061%
Predicted AMTE(50)	0.128%	0.128%	0.103%	0.078%

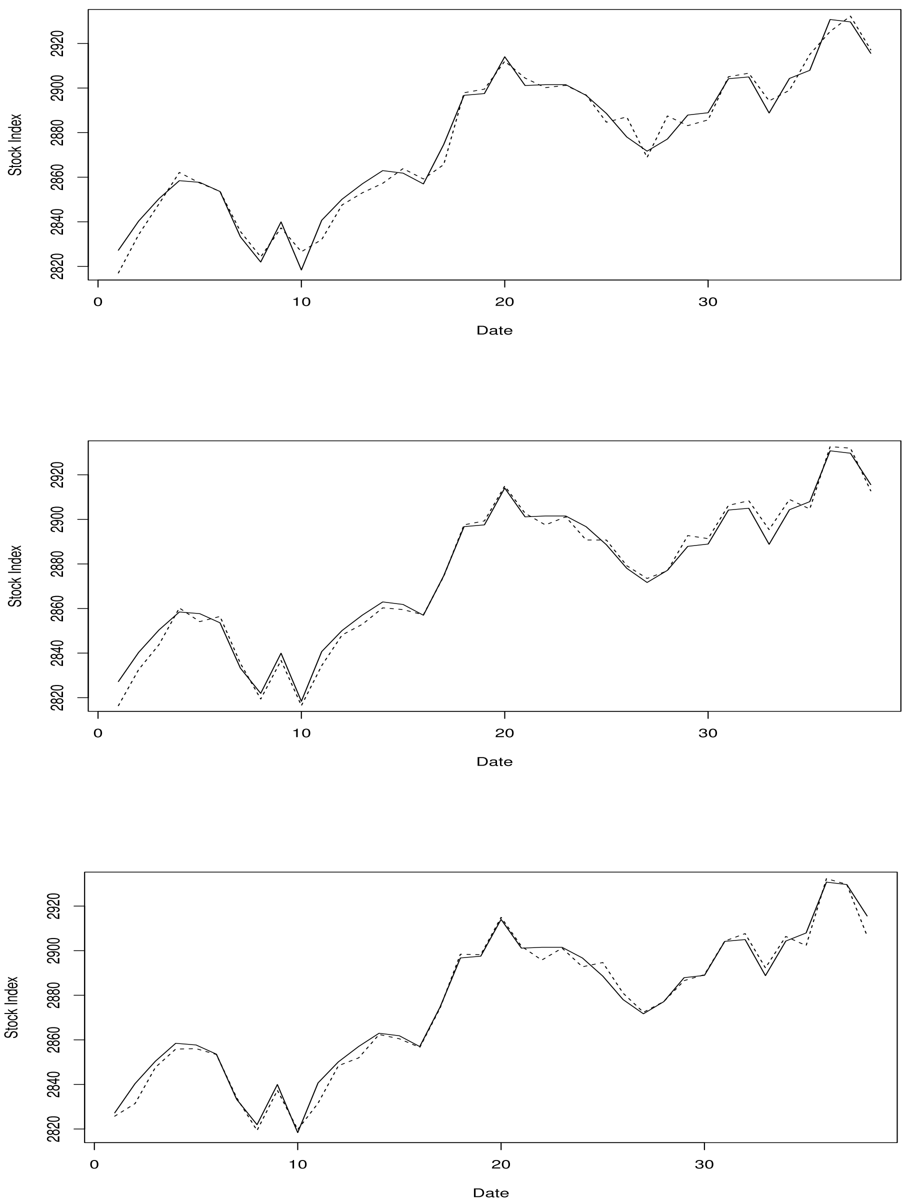

Figure 3. Predicted results for real data with 35 stocks: S&P 500 index (solid line), predicted value (dashed line), nonnegative lasso (Top), nonnegative adaptive lasso (Middle), nonnegative hierarchical lasso (Bottom).

7. CONCLUDING REMARKS

In this paper, we propose the nonnegative hierarchical lasso for bi-level selection both in low-dimensional and ultra high-dimensional linear regression model, and prove its nice statistical properties under certain appropriate conditions. Since its theoretical properties in cases where $p \gg n$, to the best of our knowledge, have not been explored. We also derive the oracle inequalities in cases where the number of covariates is much larger than the sample size.

The nonnegative hierarchical lasso, however, has the lasso penalty as its inner penalty so that it shares some drawbacks with the usual lasso. For example, a single strong predictor could draw other predictors in the same group into the model, which prevents the nonnegative hierarchical lasso from achieving consistency for the selection of individual variables. For further improvements, the lasso penalty could be replaced by adaptive lasso, MCP or SCAD. We could also consider the composite $l_{1/2}$ penalty, i.e. using the $l_{1/2}$ penalty as both the outer and inner penalties. Further work is needed to study the properties of this class of estimators and compare their performance.

Besides, we only focus on the nonnegative hierarchical lasso in the context of linear regression models. The proposed approach can be applied to other regression models

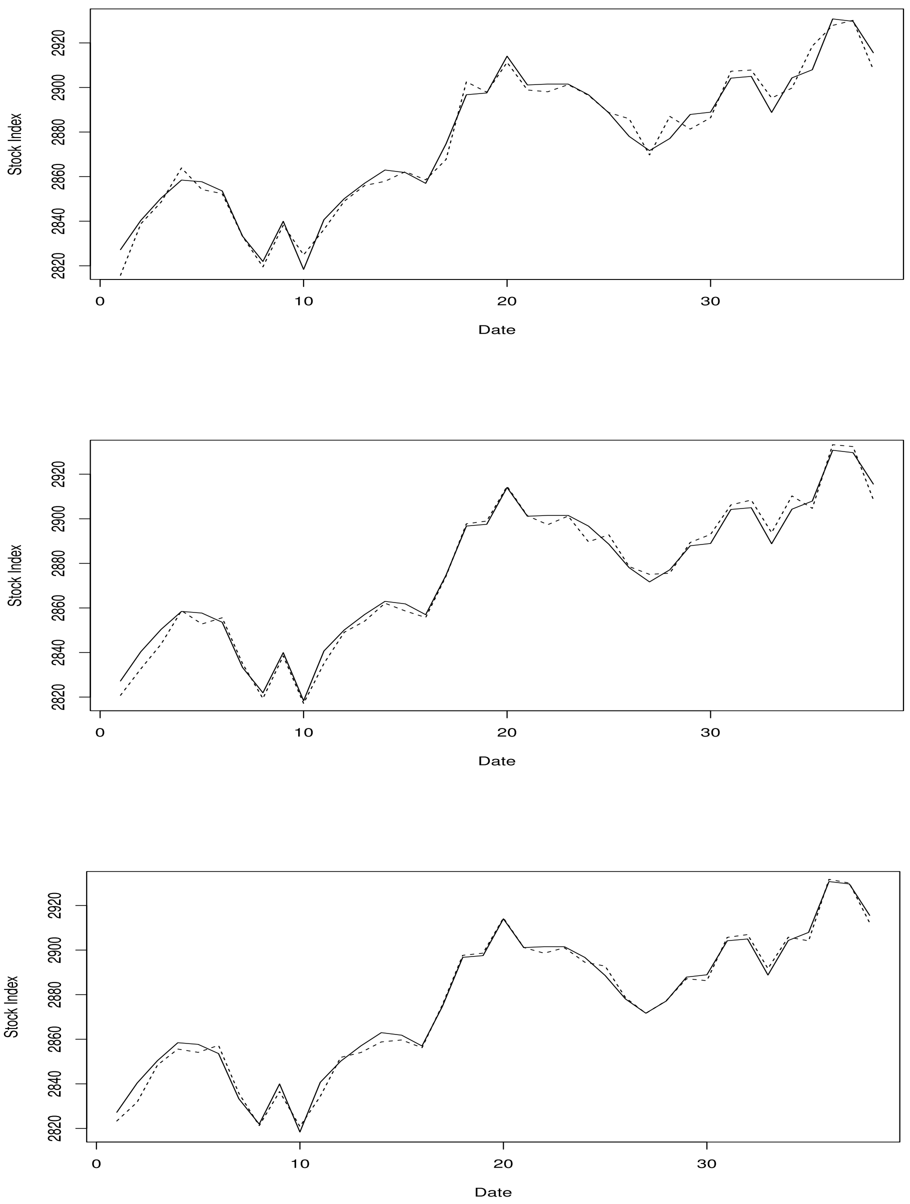

Figure 4. Predicted results for real data with 50 stocks: S&P 500 index (solid line), predicted value (dashed line), nonnegative lasso (Top), nonnegative adaptive lasso (Middle), nonnegative hierarchical lasso (Bottom).

when both the group and individual sparsity are desired. Specifically, it can be naturally extended to the generalized linear models, Cox regression and robust regression. Therefore, it is of interest to further study the theoretical properties and computational algorithms with different loss functions.

ACKNOWLEDGEMENT

This work is supported by the National Natural Science Foundation of China (Grant No. 11671059).

APPENDIX: PROOFS OF THEOREMS

Proof of Lemma 2.1 As stated in Section 2.

Let $Q^*(d, \alpha)$ be the criteria of equation (4) and $Q(\beta)$ be the criteria of equation (5). Suppose that $(\hat{d}, \hat{\alpha})$ is a local minimizer of $Q^*(d, \alpha)$. We will show $\hat{\beta}_{jk} = \hat{d}_j \hat{\alpha}_{jk}, j = 1, \ldots, J, k = 1, \ldots, p_j$ is a local minimizer of $Q(\beta)$ in the following.

Denote $\beta_{jk} = d_j \alpha_{jk}$. Since

$$Q^*(d, \alpha)$$
$$= \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{J} d_j \sum_{k=1}^{p_j} \alpha_{jk} x_{i,jk} \right)^2$$

$$
\begin{aligned}
&+ \lambda_1 \sum_{j=1}^{J} d_j + \lambda_2 \sum_{j=1}^{J}\sum_{k=1}^{p_j} \alpha_{jk} \\
&= \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{J}\sum_{k=1}^{p_j} \beta_{jk} x_{i,jk}\right)^2 \\
&+ \lambda_1 \sum_{j=1}^{J} d_j + \lambda_2 \sum_{j=1}^{J}\sum_{k=1}^{p_j} (d_j)^{-1}\beta_{jk}.
\end{aligned}
$$

Thus for any $\beta \geq 0$, we have $\hat{d}(\beta) = \arg\min_{d\geq 0} Q^*(d,\alpha)$.

Therefore

$$
(25) \qquad \hat{d}_j(\beta) = (\frac{\lambda_1}{\lambda_2})^{-\frac{1}{2}}(\sum_{k=1}^{p_j}\hat{\beta}_{jk})^{\frac{1}{2}}, , j = 1, 2, \ldots, J.
$$

Substitute (25) into $Q^*(d,\alpha)$, it is easy to obtain that

$$
\begin{aligned}
&Q^*(\hat{d},\hat{\alpha}) \\
&= \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{J}\sum_{k=1}^{p_j} \hat{\beta}_{jk} x_{i,jk}\right)^2 \\
&+ \lambda_1 \sum_{j=1}^{J}(\frac{\lambda_1}{\lambda_2})^{-\frac{1}{2}}(\sum_{k=1}^{p_j}\hat{\beta}_{jk})^{\frac{1}{2}} \\
&+ \lambda_2 \sum_{j=1}^{J}\sum_{k=1}^{p_j}(\frac{\lambda_1}{\lambda_2})^{\frac{1}{2}}(\sum_{k=1}^{p_j}\hat{\beta}_{jk})^{-\frac{1}{2}}\hat{\beta}_{jk} \\
&= \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{J}\sum_{k=1}^{p_j} \hat{\beta}_{jk} x_{i,jk}\right)^2 \\
&+ 2\sqrt{\lambda_1\cdot\lambda_2}\sqrt{\sum_{k=1}^{p_j}\hat{\beta}_{jk}},
\end{aligned}
$$

which is equal to $Q(\hat{\beta})$ with $\lambda = \lambda_1\cdot\lambda_2$.

Proof of Lemma 3.1 As stated in Section 3. Let $\hat{\beta}$ be the nonnegative hierarchical lasso estimator for a given λ, then from the definition of $\hat{\beta}$

$$
\begin{aligned}
&\frac{1}{2}\|Y - X\hat{\beta}\|_2^2 + 2\sqrt{\lambda}\sum_{j=1}^{J}\sqrt{\hat{\beta}_{j1} + \ldots + \hat{\beta}_{jp_j}} \\
&\leq \frac{1}{2}\|Y - X\beta^*\|_2^2 + 2\sqrt{\lambda}\sum_{j=1}^{J}\sqrt{\beta^*_{j1} + \ldots + \beta^*_{jp_j}}.
\end{aligned}
$$

By Cauchy-Schwarz inequality

$$
\begin{aligned}
&\sum_{j=1}^{J}\sqrt{\sum_{k=1}^{p_j}\beta^*_{jk}} - \sum_{j=1}^{J}\sqrt{\sum_{k=1}^{p_j}\hat{\beta}_{jk}} \\
&\leq \sum_{j=1}^{J_1}\sqrt{\sum_{k=1}^{p_j}\beta^*_{jk}} - \sum_{j=1}^{J_1}\sqrt{\sum_{k=1}^{p_j}\hat{\beta}_{jk}} \\
&\leq \sum_{j=1}^{J_1}\left[\left(\sum_{k=1}^{p_j}\beta^*_{jk}\right)^{-\frac{1}{2}}\left\|\hat{\beta}_{A_j} - \beta^*_{A_j}\right\|_1\right] \\
&\leq \sum_{j=1}^{J_1}\left[\left\|\beta^*_{A_j}\right\|_1^{-\frac{1}{2}}\left(|A_j|\|\hat{\beta}_{A_j} - \beta^*_{A_j}\|_2^2\right)^{\frac{1}{2}}\right] \\
&\leq \eta_n\left(\sum_{j=1}^{J_1}\|\hat{\beta}_{A_j} - \beta^*_{A_j}\|_2^2\right)^{\frac{1}{2}},
\end{aligned}
$$

where $\eta_n = \sqrt{\sum_{j=1}^{J_1}\frac{|A_j|}{\|\beta^*_{A_j}\|_1}}$.

Since $\sum_{j=1}^{J_1}\|\hat{\beta}_{A_j} - \beta^*_{A_j}\|_2^2 \leq C_n^*\|\hat{\beta} - \beta^*\|_2^2$. Then by the above inequalities we have

$$
\begin{aligned}
&4\sqrt{\lambda}\eta_n\sqrt{C_n^*}|\hat{\beta} - \beta^*\|_2 \\
&\geq \|Y - X\hat{\beta}\|_2^2 - \|Y - X\beta^*\|_2^2 \\
&= \|X\beta^* + \epsilon - X\hat{\beta}\|_2^2 - \|Y - X\beta^*\|_2^2 \\
&= \|X(\hat{\beta} - \beta^*)\|_2^2 + 2\epsilon' X(\beta^* - \hat{\beta}).
\end{aligned}
$$

Let ϵ^* be the projection of ϵ to the span of $\{X_1, X_2, \ldots, X_p\}$ in the sense that $\epsilon^* = X(X^TX)^{-1}X^T\epsilon$, then

$$
\begin{aligned}
2\epsilon^T X(\hat{\beta} - \beta^*) &= 2\epsilon^{*T} X(\hat{\beta} - \beta^*) \\
&\leq 2\|\epsilon^*\|_2\|X(\hat{\beta} - \beta^*)\|_2 \\
&\leq \frac{4\|\epsilon^*\|_2^2 + \|X(\hat{\beta} - \beta^*)\|_2^2}{2}.
\end{aligned}
$$

Thus combine the above inequalities, we have

$$
\begin{aligned}
(26) \qquad &\|X(\hat{\beta} - \beta^*)\|_2^2 \\
&\leq 8\sqrt{\lambda}\eta_n\sqrt{C_n^*}|\hat{\beta} - \beta^*\|_2 + 4\|\epsilon^*\|_2^2.
\end{aligned}
$$

Since $\mathbb{E}\|\epsilon^*\|_2^2 = \sigma^2 tr(X(X^TX)^{-1}X^T) = p\sigma^2$, then

$$
\begin{aligned}
(27) \qquad &\|X(\hat{\beta} - \beta^*)\|_2^2 \\
&\leq 8\sqrt{\lambda}\eta_n\sqrt{C_n^*}|\hat{\beta} - \beta^*\|_2 + 4O_p(p\sigma^2).
\end{aligned}
$$

Note that λ_{min} is the smallest eigenvalue of $\frac{X^TX}{n}$, then we have

$$
\begin{aligned}
&\|\hat{\beta} - \beta^*\|_2^2 \\
&\leq \frac{8\sqrt{\lambda C_n^*}\eta_n\|\hat{\beta} - \beta^*\|_2}{n\lambda_{min}} + \frac{4O_p(p\sigma^2)}{n\lambda_{min}} \\
&\leq \frac{32\lambda\eta_n^2 C_n^*}{n^2\lambda_{min}^2} + \frac{\|\hat{\beta} - \beta^*\|_2^2}{2} + \frac{4O_p(p\sigma^2)}{n\lambda_{min}}.
\end{aligned}
$$

Then condition (A2) gives

$$
(28) \qquad \|\hat{\beta} - \beta^*\|_2^2 \leq \frac{64\lambda\eta_n^2 C_n^*}{n^2\lambda_{min}^2} + \frac{8O_p(p\sigma^2)}{n\lambda_{min}}
$$

$$\leq O_p\left(\frac{p\sigma^2}{n\lambda_{min}}\right).$$

Proof of Theorem 3.1 As stated in Section 3. Define $\widetilde{\beta} = (\widetilde{\beta}_{11}, \widetilde{\beta}_{12}, \ldots, \widetilde{\beta}_{Jp_J})^T$ by

$$\widetilde{\beta}_{jk} = \begin{cases} \hat{\beta}_{jk} \quad 1 \leq j \leq J_1, 1 \leq k \leq p_j, \\ 0 \quad J_1 + 1 \leq j \leq J, 1 \leq k \leq p_j. \end{cases}$$

Then by KKT conditions for any $\hat{\beta}_{jk} \neq 0$ we can obtain that

$$(Y - X\hat{\beta})^T X_{jk} = \sqrt{\lambda} \frac{1}{\sqrt{\hat{\beta}_{j1} + \ldots + \hat{\beta}_{jp_j}}}.$$

By the definition of $\widetilde{\beta}$ and the non-negativity of $\hat{\beta}$, we can validate that

$$\begin{aligned} &(Y - X\hat{\beta})^T X(\hat{\beta} - \widetilde{\beta}) \\ &= \sqrt{\lambda} \sum_{j=1}^{J} \sum_{k=1}^{p_j} \frac{(\hat{\beta}_{jk} - \widetilde{\beta}_{jk})}{\sqrt{\hat{\beta}_{j1} + \ldots + \hat{\beta}_{jp_j}}} \\ &= \sqrt{\lambda} \sum_{j=1}^{J} \|\hat{\beta}_{A_j}\|_1^{-\frac{1}{2}} (\|\hat{\beta}_{A_j}\|_1 - \|\widetilde{\beta}_{A_j}\|_1) \\ &= \sqrt{\lambda} \sum_{j=J_1+1}^{J} \|\hat{\beta}_{A_j}\|_1^{\frac{1}{2}}. \end{aligned}$$

Similarly to the proof of Lemm1, the definition of $\hat{\beta}$ implies

$$\begin{aligned} &\frac{1}{2}\|Y - X\hat{\beta}\|_2^2 + 2\sqrt{\lambda} \sum_{j=1}^{J} \|\hat{\beta}_{A_j}\|_1^{\frac{1}{2}} \\ &\leq \frac{1}{2}\|Y - X\widetilde{\beta}\|_2^2 + 2\sqrt{\lambda} \sum_{j=1}^{J} \|\widetilde{\beta}_{A_j}\|_1^{\frac{1}{2}}. \end{aligned}$$

It is easy to prove that

$$\begin{aligned} &2\sqrt{\lambda} \sum_{j=1}^{J} (\|\hat{\beta}_{A_j}\|_1^{\frac{1}{2}} - \|\widetilde{\beta}_{A_j}\|_1^{\frac{1}{2}}) \\ &\leq \frac{1}{2}\|Y - X\widetilde{\beta}\|_2^2 - \frac{1}{2}\|Y - X\hat{\beta}\|_2^2 \\ &= \frac{1}{2}\|X(\hat{\beta} - \widetilde{\beta})\|_2^2 + (Y - X\hat{\beta})^T X(\hat{\beta} - \widetilde{\beta}). \end{aligned}$$

Note that λ_{max} denotes the largest eigenvalue of the Gram matrix $X^T X/n$ and by Lemma 1, it is easy to get

$$\begin{aligned} \sqrt{\lambda} \sum_{j=J_1+1}^{J} \|\hat{\beta}_{A_j}\|_1^{\frac{1}{2}} &\leq \frac{1}{2} n\lambda_{max} \|\hat{\beta}_{B_2}\|_2^2 \\ &\leq \frac{n\lambda_{max}}{2} \|\hat{\beta} - \beta^*\|_2^2 \leq n\lambda_{max} O_p\left(\frac{p\sigma^2}{n\lambda_{min}}\right). \end{aligned}$$

Now we establish the lower bound of the $\sum_{j=J_1+1}^{J} \|\hat{\beta}_{A_j}\|_1^{\frac{1}{2}}$, it is easy to prove that

$$\begin{aligned} \sum_{j=J_1+1}^{J} \|\hat{\beta}_{A_j}\|_1^{\frac{1}{2}} &\leq \left(\sum_{j=J_1+1}^{J} \|\hat{\beta}_{A_j}\|_1 \right)^{\frac{1}{2}} \\ &\leq \|\hat{\beta}_{B_2}\|_1^{\frac{1}{2}} \leq \|\hat{\beta}_{B_2}\|_2^{\frac{1}{2}}. \end{aligned}$$

Then combine the above two inequalities, we have

$$\sqrt{\lambda} \leq \frac{n\lambda_{max}}{2} \|\hat{\beta}_{B_2}\|_2^{\frac{3}{2}} \leq n\lambda_{max} O_p\left(\frac{p\sigma^2}{n\lambda_{min}}\right)^{\frac{3}{4}}.$$

Thus by condition (A3) we can prove that

$$\begin{aligned} \mathbb{P}\big(\|\hat{\beta}_{B_2}\|_2 > 0\big) &\leq \mathbb{P}\left\{ \frac{\lambda(\lambda_{min}/p)^{\frac{3}{2}}}{\lambda_{max}^2 n^{\frac{1}{2}}} \leq O_p(1) \right\} \\ &\to 0. \end{aligned} \tag{29}$$

Then we complete the proof.

Proof of Theorem 3.2 As stated in Section 3. By Lemma 2, recall that $\|\hat{\beta} - \beta^*\|_2^2 = O_p\big(\frac{\sigma^2 p}{n\lambda_{min}}\big)$. Since $\{B_1, \beta^*_{B_1}, J_1\}$ are fixed, and (7) implies (A2) and (A3), then the proof of lemma 2 still works with the submatrix X_1, thus we have

$$\|\hat{\beta}_{B_1} - \beta^*_{B_1}\|_2^2 = O_p\left(\frac{1}{n}\right). \tag{30}$$

Define

$$V_n(u) = L_n\left(\beta^* + \frac{1}{\sqrt{n}}(u^T, 0^T)^T\right) - L_n\left(\beta^*\right),$$

where $L_n(\beta)$ is the objective function of (5), 0 stands for the zero vector of dimension $|B_2|$. Then by theorem 1, the following hold with large probability

$$\begin{aligned} \hat{\beta} - \beta^* &= \frac{1}{\sqrt{n}}(u^T, 0^T)^T, \\ \hat{u}_n &= \operatorname*{arg\,min}_{\beta^*_{B_1} + u/\sqrt{n} \geq 0} \big\{ V_n(u) : u \in R^{|B_1|} \big\}. \end{aligned}$$

Since the function $V_n(u)$ can be decomposed into two parts

$$\begin{aligned} V_n(u) &= \frac{1}{2}\|\epsilon - X(u^T, 0^T)^T/\sqrt{n}\|_2^2 - \frac{1}{2}\|\epsilon\|_2^2 \\ &+ 2\sqrt{\lambda} \sum_{j=1}^{J_1} \left(\|\beta^*_{A_j} + \frac{1}{\sqrt{n}} u_{A_j}\|_1^{\frac{1}{2}} - \|\beta^*_{A_j}\|_1^{\frac{1}{2}} \right) \\ &= -\frac{1}{\sqrt{n}} u^T X_1^T \epsilon + \frac{1}{2n} u^T X_1^T X_1 u \\ &+ 2\sqrt{\lambda} \sum_{j=1}^{J_1} \left(\|\beta^*_{A_j} + \frac{1}{\sqrt{n}} u_{A_j}\|_1^{\frac{1}{2}} - \|\beta^*_{A_j}\|_1^{\frac{1}{2}} \right) \\ &= V_{1n}(u) + V_{2n}(u). \end{aligned}$$

For the first term, by (10), it is easy to obtain that

$$V_{1n}(u) \to_D -u^T W + \frac{1}{2}u^T C_{11} u. \tag{31}$$

Then for the second term, since

$$\begin{aligned} V_{2n}(u) &= 2\sqrt{\lambda}\sum_{j=1}^{J_1}\left(\|\beta^*_{A_j} + \frac{u_{A_j}}{\sqrt{n}}\|_1^{\frac{1}{2}} - \|\beta^*_{A_j}\|_1^{\frac{1}{2}}\right) \\ &= 2\frac{\sqrt{\lambda}}{\sqrt{n}}\sum_{j=1}^{J_1}\frac{\|u_{A_j}\|_1}{\|\beta^*_{A_j} + \frac{1}{\sqrt{n}}u_{A_j}\|_1^{\frac{1}{2}} + \|\beta^*_{A_j}\|_1^{\frac{1}{2}}}. \end{aligned}$$

Then by (7), we have $V_{2n}(u) \to 0$, therefore

$$V_n(u) \to_D -u^T W + \frac{1}{2}u^T C_{11} u. \tag{32}$$

Then it follows by solving a constrained optimization problem

$$\begin{cases} \min -u^T W + \frac{1}{2}u^T C_{11} u, \\ \text{subject to } \beta^*_{B_1} + \frac{u}{\sqrt{n}} \geq 0. \end{cases} \tag{33}$$

By [3, Theorem 2], it is equivalent to

$$\begin{cases} \min -u^T W + \frac{1}{2}u^T C_{11} u, \\ \text{subject to } u_j \geq 0, j \in \mathcal{I}(\beta^*_{B_1}), \end{cases} \tag{34}$$

where $\mathcal{I}(\beta^*_{B_1})$ denotes the index set of constraints which are active in $\beta^*_{B_1}$, that is, $\mathcal{I}(\beta^*_{B_1}) = \{j \in B_1 : \beta_j = 0\}$. Let D be the feasible region of the above optimization problem and D^o be the relative interior. D_j and $D_{j_1,\ldots,j_k}$ denote the boundary formed by the jth constraint and the intersection of $D_{j_1}, \ldots, D_{j_k}$, respectively. $D^o_{j_1,\ldots,j_k}$ represents the relative interior of $D_{j_1,\ldots,j_k}$. They can be defined as follows

$$\begin{aligned} D^o &= \{u : u_j > 0, j \in \mathcal{I}(\beta^*_{B_1})\}, \\ D_{j_1,\ldots,j_k} &= \{u : u_{j_r} = 0, j_r \in \mathcal{I}(\beta^*_{B_1}), 1 \leq r \leq k; \\ &\qquad u_t \geq 0, t \in \mathcal{I}(\beta^*_{B_1}) \setminus \{j_1, \ldots, j_k\}\}, \\ D^o_{j_1,\ldots,j_k} &= \{u : u_{j_r} = 0, j_r \in \mathcal{I}(\beta^*_{B_1}), 1 \leq r \leq k; \\ &\qquad u_t > 0, t \in \mathcal{I}(\beta^*_{B_1}) \setminus \{j_1, \ldots, j_k\}\}. \end{aligned}$$

If $\hat{u} \in D^o$ then by KKT conditions we have

$$\hat{u} \to_D C_{11}^{-1} W. \tag{35}$$

In contrast, if $\hat{u} \in D^o_{j_1,\ldots,j_k}$, by KKT conditions we obtain

$$\begin{cases} C_{11}u - W - \lambda_{j1,\ldots,j_k} = 0, \\ u_{j_r} = 0, r = 1, \ldots, k, \end{cases}$$

where $\lambda_{j_1,\ldots,j_k}$ is a vector of lagrangian multipliers with $\lambda_j = 0, j \in B_1 \setminus \{j_1, \ldots, j_k\}$.

Let

$$B_{j_1,\ldots,j_k} = \begin{bmatrix} C_{11} & H \\ H^T & 0 \end{bmatrix},$$

where H denotes the $|B_1| \times k$ matrix with the main diagonal elements 1 and others 0.

We can write the inverse of $B_{j_1,\ldots,j_k}$ as a block matrix

$$B^{-1}_{j_1,\ldots,j_k} = \begin{bmatrix} M_{j_1,\ldots,j_k} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}.$$

By the elementary operation of block matrices, it is easy to obtain that

$$M_{j_1,\ldots,j_k} = C_{11}^{-1}[I - H(H^T C_{11}^{-1} H)^{-1} H^T C_{11}^{-1}].$$

Thus we complete the proof.

Proof of Lemma 3.2 As stated in Section 3. By the definition of $\hat{\beta}$, we have

$$\begin{aligned} &\frac{1}{2n}\|Y - X\hat{\beta}\|_2^2 + \lambda_n \sum_{j=1}^{J}\sqrt{\hat{\beta}_{j1} + \ldots + \hat{\beta}_{jp_j}} \\ &\leq \frac{1}{2n}\|Y - X\beta^*\|_2^2 + \lambda_n \sum_{j=1}^{J}\sqrt{\beta^*_{j1} + \ldots + \beta^*_{jp_j}}. \end{aligned}$$

Then by the subadditivity property of the square root, we have that

$$\begin{aligned} &\frac{1}{2n}\|X(\beta^* - \hat{\beta})\|_2^2 + \lambda_n \sum_{j=J_1+1}^{J}\|\hat{\beta}_{A_j}\|_1^{\frac{1}{2}} \\ &\leq \lambda_n \sum_{j=1}^{J_1}\|\hat{\beta}_{A_j} - \beta^*_{A_j}\|_1^{\frac{1}{2}} + \frac{1}{n}\epsilon^T X(\hat{\beta} - \beta^*) \\ &\leq \lambda_n \sum_{j=1}^{J_1}\|\Delta_{A_j}\|_1^{\frac{1}{2}} + \max_{1 \leq k \leq p}\frac{|X_k^T \epsilon|}{n}\|\hat{\beta} - \beta^*\|_1 \\ &\leq \lambda_n \sum_{j=1}^{J_1}\|\Delta_{A_j}\|_1^{\frac{1}{2}} + \sum_{j=1}^{J}\max_{k \in 1,\ldots,p}\frac{|X_k^T \epsilon|}{n}\|\Delta_{A_j}\|_1. \end{aligned}$$

Since $\hat{\beta} \geq 0$, then

$$\|\hat{\beta}_{A_j} - \beta^*_{A_j}\|_1 \leq m_2\|\hat{\beta}_{A_j} - \beta^*_{A_j}\|_1^{\frac{1}{2}}. \tag{36}$$

Consider the random event

$$\mathcal{H} = \left\{\max_{k \in 1,\ldots,p}\frac{|X_k^T \epsilon|}{n} \leq \frac{\lambda_n}{2m_2}\right\},$$

and for every $j \in 1, \ldots, J$, define

$$\begin{aligned} \mathcal{H}_j &= \left\{\max_{k \in A_j}\frac{|X_k^T \epsilon|}{n} \leq \frac{\lambda_n}{2m_2}\right\}, \\ \mathcal{A}_j &= \left\{\frac{\|X_{A_j}^T \epsilon\|_2}{n} \leq \frac{\lambda_n}{2m_2}\right\}. \end{aligned}$$

Obviously,

$$\mathcal{H} = \bigcap_{j=1}^{J}\mathcal{H}_j,$$

$$\mathbb{P}(\mathcal{A}_j) \le \mathbb{P}(\mathcal{H}_j).$$

We note that

$$\mathbb{P}(\mathcal{A}_j) = \mathbb{P}\left(\frac{1}{n^2}\epsilon^T X_{A_j}^T X_{A_j}\epsilon \le \frac{\lambda_n^2}{4m_2^2}\right)$$
$$= \mathbb{P}\left(\frac{\sum_{i=1}^n \nu_{j,i}(\xi_i^2-1)}{\sqrt{2}\|\nu_j\|_2} \le d_j\right),$$

where $\xi_1,\dots,\xi_n$ are *i.i.d* standard normal and $\nu_{j,1},\dots,\nu_{j,n}$ represent the eigenvalues of the matrix $X_{A_j}X_{A_j}^T/n$ which has the same positive ones as $C_{A_j}^n$, and d_j is defined as follows

$$d_j = \frac{n\lambda_n^2/4\sigma^2 m_2^2 - tr(C_{A_j}^n)}{\sqrt{2}\|C_{A_j}^n\|_F}.$$

Applying [16, Lemma B.1] to the event $\mathcal{A}_j$ then we have

$$\mathbb{P}(\mathcal{A}_j^c) \le 2\exp\left\{\frac{-d_j^2/2}{1+\sqrt{2}d_j|||C_{A_j}^n|||/\|C_{A_j}^n\|_F}\right\}.$$

Now choose d_j to make the right-hand side of the above inequality smaller than $2J^{-\gamma}$, after some computation we have

$$d_j \ge \sqrt{2}\gamma log(J)|||C_{A_j}^n|||/\|C_{A_j}^n\|_F + \big((2\gamma log(J)$$
$$|||C_{A_j}^n|||/\|C_{A_j}^n\|_F)^2 + 2\gamma log(J)\big)^{\frac{1}{2}}.$$

The subadditivity property of the square root and inequality $\|C_{A_j}^n\|_F \le \sqrt{p_j}|||C_{A_j}^n|||$ imply inequality (18) holds and then $\mathbb{P}(\mathcal{H}) \ge 1 - 2J^{1-\gamma}$.

Then on the event $\mathcal{H}$ gives inequalities (16) and (17) immediately.

Proof of Theorem 3.3 As stated in Section 3. By assumption (B1) and Lemma 3 we have

$$\kappa^2(s)\|\Delta_{B_1}\|_1^2 \le \frac{1}{n}\|X\Delta\|_2^2 \le 3\lambda_n\sum_{j=1}^{J_1}\|\Delta_{A_j}\|_1^{\frac{1}{2}}$$
$$\le 3C_n^* J_1^{\frac{1}{2}}\lambda_n\|\Delta_{B_1}\|_1^{\frac{1}{2}}.$$

Then we have

$$\|\Delta_{B_1}\|_1 \le \frac{3^{\frac{2}{3}}J_1^{\frac{1}{3}}C_n^{*\frac{2}{3}}\lambda_n^{\frac{2}{3}}}{\kappa^{\frac{4}{3}}(s)}, \tag{37}$$

which coincides with inequality (19). By (16) we obtain

$$\|\Delta\|_1^{\frac{1}{2}} \le \sum_{j=1}^{J}\|\Delta_{A_j}\|_1^{\frac{1}{2}} \le 4C_n^*\|\Delta_{B_1}\|_1^{\frac{1}{2}}. \tag{38}$$

Then we get (20) immediately.

Furthermore, by (17) we have

$$\frac{1}{n}\|X\Delta\|_2^2 \le 3\lambda_n\sum_{j=1}^{J_1}\|\Delta_{A_j}\|_1^{\frac{1}{2}}$$
$$\le 3\lambda_n J_1^{\frac{1}{2}}\|\Delta_{B_1}\|_1^{\frac{1}{2}}$$
$$\le \frac{3^{\frac{4}{3}}J_1^{\frac{2}{3}}C_n^{*\frac{1}{3}}\lambda_n^{\frac{4}{3}}}{\kappa^{\frac{2}{3}}(s)},$$

which is in line with (18).

To prove (21), by the KKT conditions, for any $\hat{\beta}_{jk} \ne 0, j = 1,\dots,J, k = 1,\dots,p_j$ we have

$$\frac{1}{n}X_{jk}^T(Y - X\hat{\beta}) = \frac{\lambda_n}{2\sqrt{\hat{\beta}_{j1}+\cdots+\hat{\beta}_{jp_j}}}. \tag{39}$$

Thus by the definition of events $\mathcal{H}$ yields

$$\frac{1}{n}X_{jk}^T X(\beta^* - \hat{\beta}) - \frac{\lambda_n}{2\|\hat{\beta}_{A_j}\|_1^{\frac{1}{2}}} + \frac{\lambda_n}{\|\beta_{A_j}^*\|_1^{\frac{1}{2}}}$$
$$= \frac{\lambda_n}{\|\beta_{A_j}^*\|_1^{\frac{1}{2}}} - \frac{1}{n}X_{jk}^T\epsilon \ge \frac{\lambda_n}{2m_2}.$$

Then using the Cauchy-Schwarz inequality, on the event $\mathcal{H}$, it holds uniformly over $\{jk{:}\hat{\beta}_{jk} \ne 0,\ j = 1,\dots,J, k = 1,\dots,p_j\}$ that

$$M(\hat{\beta}) \le \frac{2m_2}{\lambda_n}\sum_{jk\in J(\hat{\beta})}\left\{\frac{1}{n}X_{jk}^T X(\beta^* - \hat{\beta})\right.$$
$$\left. - \frac{\lambda_n}{2\|\hat{\beta}_{A_j}\|_1^{\frac{1}{2}}} + \frac{\lambda_n}{\|\beta_{A_j}^*\|_1^{\frac{1}{2}}}\right\}$$
$$\le \frac{2m_2}{\lambda_n}\sum_{jk\in J(\hat{\beta})}\left\{\frac{1}{n}X_{jk}^T X\Delta + \frac{\lambda_n}{\|\beta_{A_j}^*\|_1^{\frac{1}{2}}}\right\}$$
$$\le \frac{2m_2}{\lambda_n}\left\{\sqrt{M(\hat{\beta})}\frac{\|X^TX\Delta\|_2}{n} + \frac{\lambda_n}{m_1}M(\hat{\beta})\right\}.$$

Then we have

$$(1 - \frac{\lambda_n}{m_1})\sqrt{M(\hat{\beta})} \le \frac{2m_2\sqrt{\lambda_{max}}}{\sqrt{n}\lambda_n}\|X\Delta\|_2. \tag{40}$$

After some computation results in (21) immediately.

Proof of Theory 3.4 As stated in Section 3. Since (*ii*) follows from theorem 3, it suffices to prove (*i*). Applying the Lemma 3 and theorem 3 yields

$$\|\Delta_{B_2}\|_1^{\frac{1}{2}} \le \sum_{j=J_1+1}^{J}\|\Delta_{A_j}\|_1^{\frac{1}{2}} \le 3\sum_{j=1}^{J_1}\|\Delta_{A_j}\|_1^{\frac{1}{2}}$$
$$\le 3J_1^{\frac{1}{2}}\|\Delta_{B_1}\|_1^{\frac{1}{2}} \le O_p\left(\frac{C_n^{*\frac{1}{3}}\lambda_n^{\frac{1}{3}}J_1^{\frac{2}{3}}}{\kappa^{\frac{2}{3}}(s)}\right), \tag{41}$$

so that $\mathbb{P}(\hat{\beta}_{B_2} = 0) \to 1$.

Received 14 January 2019

REFERENCES

[1] Breiman, L. Better subset regression using the Nonnegative Garrote, Technometrics 37 (1995) 373–384. MR1365720

[2] Kim, J., Pollard, D. Cube root asymptotics, Ann. Statist. 18 (1990) 191–219. MR1041391

[3] Wang, J. Asymptotics of least squares estimators for constrained nonlinear regression, Ann.Statist. 24(1996) 1316–1326. MR1401852

[4] Fan, J., Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties, J. Amer. Statist. Assoc. 96 (2001) 1348–1360. MR1946581

[5] Zou, H., Li, R. One-step sparse estimates in nonconcave penalized likelihood models, Ann. Statist. 36(2008) 1509–1533. MR2435443

[6] Friedman, J., Hastie, T., Tibshirani, R. Pathwise coordinate optimization, Ann. Appl. Statist. 1 (2007) 302–332. MR2415737

[7] Yuan, M., Lin, Y. Model selection and estimation in regression with grouped variables, Journal of the Royal Statistical Society 68 (2006) 49–67. MR2212574

[8] Meier, L., Sara, V.D.G., Bühlmann, P. The group lasso for logistic regression, Journal of the Royal Statistical Society 70 (2008) 53–71. MR2412631

[9] Wang, L., Chen, G., Li, H. Group SCAD regression analysis for microarray time course gene expression data, Bioinformatics 23 (2007) 1486–1494.

[10] Zhao, P., Rocha, G., Yu, B. The composite absolute penalties family for grouped and hierarchical variable selection, Ann. Statist. 37 (2009) 3468–3497. MR2549566

[11] Friedman, J., Hastie, T., Tibshirani, R. A note on the group lasso and a sparse group lasso, Stat. 2010.

[12] Zhou, N., Zhu, J. Group variable selection via a hierarchical lasso and its oracle property, Statistics and Its Interface 3 (2010) 557–574. MR2754752

[13] Huang, J., Ma, S., Xie, H., et al. A group bridge approach for variable selection, Biometrika 96 (2009) 339–355. MR2507147

[14] Breheny, P., Huang, J. Penalized methods for bi-level variable selection, Statistics and Its Interface 2 (2009) 369. MR2540094

[15] Bühlmann, P., Geer, S.V.D. Statistics for high-dimensional data, Springer Berlin Heidelberg, 2011. MR2807761

[16] Bickel, P.J., Ritov, Y., Tsybakov, A.B. Simultaneous analysis of Lasso and Dantzig selector, Ann. Statist. 37 (2009) 1705–1732. MR2533469

[17] Lounici, K., Pontil, M., Tsybakov, A.B. Oracle inequalities and optimal inference under group sparsity, Ann. Statist. 39 (2011) 2164–2204. MR2893865

[18] Ravikumar, P., Ravikumar, P., Wainwright, M.J., et al. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers, International Conference on Neural Information Processing Systems. Curran Associates Inc. (2009).

[19] Hu, J., Huang, J., Qiu, F. A group adaptive elastic-net approach for variable selection in high-dimensional linear regression, Science China Mathematics 61 (2018) 173–188. MR3744405

[20] Jiang, D., Huang, J. Concave 1-norm group selection, Biostatistics 16 (2015) 252. MR3365427

[21] Huang, J., Breheny, P., Ma, S. A Selective review of group selection in high-dimensional models, Statist. Sci. 27 (2013) 481–499 MR3025130

[22] Xu, Z., Chang, X., Xu, F., et al. L-1/2 regularization: a thresholding representation theory and a fast solver, IEEE Trans. Neural Netw. Learn. Syst. 23(2012) 1013–1027

[23] Oh, K.J., Kim, T.Y., Min, S. Using genetic algorithm to support portfolio optimization for index fund management, Expert Syst. Appl. 28(2005) 371–379.

[24] Dose, C., Cincotti, S. Clustering of financial time series with application to index and enhanced index tracking portfolio, Physica A Stat. Mech. Appl. 355(2005) 145–151. MR2143278

[25] Alexander, C. Optimal hedging using cointegration, Phil. Trans. R. Soc. London A, Math., Phys. Eng. Sci. 357(1999) 2039–2058.

[26] Wang, H., Leng, C. A note on adaptive group lasso, Computational Statistics & Data Analysis 52(2008) 5277–5286. MR2526593

[27] She, Y. An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors, Computational Statistics & Data Analysis, 56(2012) 2976–2990. MR2929353

[28] Wei, F., Zhu, H. Group coordinate descent algorithms for nonconvex penalized regression, Computational Statistics & Data Analysis 56(2012) 316–326. MR2842341

[29] Jobst, N., Horniman, M., Lucas, C., Mitra, G. Computational aspects of alternative portfolio selection models in the presence of discrete asset choice constraints, Quant. Finance 1 (2001) 1–13. MR1863872

[30] Meinshausen, N. Sign-constrained least squares estimation for high-dimensional regression, Electron. J. Statist. 7 (2013) 1607–1631. MR3066380

[31] Meinshausen, N., Bühlmann, P. High-dimensional graphs and variable selection with the lasso, Ann. Statist. 34 (2006) 1436–1462. MR2278363

[32] Slawski, M., Hein, M. Non-negative least squares for high dimensional linear models: Consistency and sparse recovery without regularization, Electron. J. Statist. 7 (2013) 3004–3056. MR3151760

[33] Wu, L., Yang, Y.H. Nonnegative elastic net and application in index tracking, Appl. Math. Comput. 227 (2014) 541–552. MR3146340

[34] Wu, L., Yang, Y.H., Liu, H.Z. Nonnegative-lasso and application in index tracking, Computational Statistics & Data Analysis 70 (2014) 116–126. MR3125482

[35] Yang, Y.H., Wu, L. Nonnegative adaptive lasso for ultra-high dimensionalm regression models and a two-stage method applied in financial modeling, J. Statist. Plann. Inference. 174 (2016) 52–67. MR3477700

[36] Yuan, M., Lin, Y. On the non-negative garrotte estimator, J. R. Statist. Soc. Ser. B 69 (2007) 143–161. MR2325269

Wanling Xie
College of Mathematics and Statistics
Chongqing University
Chongqing 401331, China
E-mail address: 476222791@qq.com

Hu Yang
College of Mathematics and Statistics
Chongqing University
Chongqing 401331, China
E-mail address: yh@cqu.edu.cn