Consistent Group Selection in High-dimensional Linear Regression

Fengrong Wei  
University of West Georgia  
Carrollton, GA, USA  
Email: fwei@westga.edu

Jian Huang  
University of Iowa  
Iowa City, IA, USA  
Email: jian-huang@uiowa.edu

Abstract

In regression problems where covariates can be naturally grouped, the group Lasso is an attractive method for variable selection, since it respects the grouping structure in the data. We study the selection and estimation properties of the group Lasso in high-dimensional settings when the number of groups exceeds the sample size. We provide sufficient conditions under which the group Lasso selects a model whose dimension is comparable with the underlying model with high probability and is estimation consistent. However, the group Lasso is in general not selection consistent and tends to also select groups that are not important in the model. To improve the selection results, we propose an adaptive group Lasso method, which is a generalization of the adaptive Lasso and requires an initial estimator. We show that the adaptive group Lasso is consistent in group selection under certain conditions, if the group Lasso is used as the initial estimator.


Keywords: group selection, high-dimensional data, penalized regression, rate consistency, selection consistency.

1 Introduction

Consider the linear regression model with \( p \) groups of covariates

\[
Y_i = \sum_{k=1}^{p} X_{ik}' \beta_k + \varepsilon_i, \quad i = 1, \ldots, n,
\]

where \( Y_i \) is the response variable, \( \varepsilon_i \) is the error term, \( X_{ik} \) is a \( d_k \times 1 \) covariate vector representing the \( k \)th group and, \( \beta_k \) is the corresponding \( d_k \times 1 \) regression coefficients vector. For such a model, the group Lasso (Yuan and Lin 2006, Antoniadis and Fan 2001) is an attractive method for variable selection since it respects the grouping structure in the covariates. This method is a natural extension of the Lasso (Tibshirani 1996), in which an \( \ell_2 \) norm of the coefficients associated with a group of variables is used as a component in the penalty function. However, the group Lasso is in general not selection consistent and tends to select more groups than there are in the model. To improve the selection results, we consider an adaptive group Lasso method, which is a generalization of the adaptive Lasso (Zou 2006). We provide sufficient conditions under which the adaptive group Lasso is selection consistent, if the group Lasso is used as the initial estimator.
The need to select groups of variables arises in many statistical modeling and applied problems. For example, in multifactor analysis of variance, a factor with multiple levels can be represented by a group of dummy variables. In nonparametric additive regression, each component can be expressed as a linear combination of a set of basis functions. In both cases, the selection of important factors or nonparametric components amounts to the selection of groups of variables. Several recent papers have considered group selection using penalized methods. In addition to the group Lasso, Yuan and Lin (2006) proposed the group Lars and group nonnegative garroate methods. Kim, Kim and Kim (2006) considered the group Lasso in the context of generalized linear models. Zhao, Rocha and Yu (2008) proposed a composite absolute penalty for group selection, which can be considered a generalization of the group Lasso. Meier, van de Geer and Bühlmann (2008) studied the group Lasso for logistic regression. Huang, Ma, Xie and Zhang (2008) proposed a group bridge method that can be used for simultaneous group and individual variable selection.

There have been much work on the penalized methods for variable selection and estimation with high-dimensional data. Several approaches have been proposed, including the least absolute shrinkage and selection operator (Lasso, Tibshirani 1996), the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li 2001, Fan and Peng 2004), the elastic net (Enet) penalty (Zou and Hastie 2006), and the minimum concave penalty (Zhang 2007). Much progress has been made in understanding the statistical properties of these methods in both fixed $p$ and $p \gg n$ settings. In particular, several recent studies considered the Lasso with regard to its variable selection, estimation and prediction properties, see for example, Knight and Fu (2001); Greenshtein and Ritov (2004); Meinshausen and Buhlmann (2006); Zhao and Yu (2006); Huang, Ma and Zhang (2006); van de Geer (2008); and Zhang and Huang (2008), among others. All these studies are concerned with the Lasso for individual variable selection.

In this article, we study the asymptotic properties of the group Lasso and the adaptive group Lasso in high-dimensional settings when $p \gg n$. We generalize the results about the Lasso obtained in Zhang and Huang (2008) to the group Lasso. We show that, under a generalized sparsity condition and the sparse Riesz condition as well as certain regularity conditions, the group Lasso selects a model whose dimension has the same order as the underlying model, selects all groups whose $\ell_2$ norms are of greater order than the bias of the selected model and is estimation consistent. In addition, under a narrow-sense sparsity condition and using the group Lasso as the initial estimator, the adaptive group Lasso can correctly select important groups with high probability.

Our theoretical and simulation results suggest the following one-step approach to group selection in high-dimensional settings. First, we use the group Lasso to obtain an initial estimator and reduce the dimension of the problem. Then we use the adaptive group Lasso to select the final set of groups of variables. Since the computation of the adaptive group Lasso estimator can be carried using the same algorithm and program for the group Lasso, the computational cost of this one-step approach is approximately twice that of a single group Lasso computation. This approach iteratively using the group Lasso twice follows the idea of the adaptive Lasso (Zou 2006) a proposal by Bühlmann and Meier (2008) in the context of individual variable selection.

The rest of the paper is organized as follows. In Section 2, we state the results on the selection, bias of the selected model and the convergent rate of the group Lasso estimator. In Section 3, we describe the selection and estimation consistency results about the adaptive group Lasso. In Section 4, we use simulation to compare the group Lasso and adaptive group Lasso. Proofs are given in Section 5. Concluding remarks are given in Section 6.
2 The asymptotic properties of the group Lasso

Let $Y = (Y_1, \ldots, Y_n)'$ and $X = (X_1, \cdots, X_p)$, where $X_k$ is the $n \times d_k$ covariate sub-matrix corresponding to the $k$th group. For a given penalty level $\lambda \geq 0$, the group Lasso estimator of $\beta = (\beta_1', \ldots, \beta_p')'$ is

$$
\hat{\beta} = \arg \min_{\beta} \frac{1}{2}(Y - X\beta)^T(Y - X\beta) + \lambda \sum_{k=1}^{p} \sqrt{d_k}||\beta_k||_2, \quad (2.1)
$$

where $\hat{\beta} = (\hat{\beta}_1', \cdots, \hat{\beta}_p')'$.

We consider the model selection and estimation properties of $\hat{\beta}$ under a generalized sparsity condition (GSC) of the model and a sparse Riesz condition (SRC) on the covariate matrix. These two conditions were first formulated in the study of the Lasso estimator (Zhang and Huang 2008). The GSC assumes that, for some $\eta_1 \geq 0$, there exists an $A_0 \subset \{1, \cdots, p\}$ such that $\sum_{k \in A_0} ||\beta_k||_2 \leq \eta_1$, where $||\cdot||_2$ denotes the $\ell_2$ norm. Without loss of generality, let $A_0 = \{q + 1, \cdots, p\}$. Then the GSC is

$$
\sum_{k=q+1}^{p} ||\beta_k||_2 \leq \eta_1. \quad (2.2)
$$

Thus the number of the truly important groups is $q$. A more rigid way to describe sparsity is to assume $\eta_1 = 0$, that is,

$$
||\beta_k||_2 = 0, \quad k = q + 1, \ldots, p. \quad (2.3)
$$

This is a special case of the GSC and we call it the narrow-sense sparsity condition (NSC). In practice, the GSC is a more realistic formulation of a sparse model. However, the NSC can often be considered a reasonable approximation to the GSC, especially when $\eta_1$ is smaller than the noise level associated with model fitting.

The SRC controls the range of eigenvalues of sub-matrix. For $A \subset \{1, \cdots, p\}$, define $X_A = (X_k, k \in A)$, $\Sigma_A = X_A'X_A/n$. Note that $X_A$ is an $n \times \sum_{k \in A} d_k$ matrix. The design matrix $X_A$ satisfies the sparse Riesz condition (SRC) with rank $q^*$ and spectrum bounds $0 < c_s < c^* < \infty$ if

$$
c_s \leq \frac{\|X_A\nu\|^2}{n \|
u\|^2} \leq c^*, \forall A \text{ with } q^* = |A| = \#\{k : k \in A\} \text{ and } \nu \in R^{\sum_{k \in A} d_k}. \quad (2.4)
$$

Let $\hat{A} = \{k : ||\hat{\beta}_k||_2 > 0, 1 \leq k \leq p\}$ which is the set of indices of the groups selected by the group Lasso. An important quantity is the cardinality of $\hat{A}$, defined as

$$
\hat{q} = |\hat{A}| = \#\{k : ||\hat{\beta}_k||_2 > 0, 1 \leq k \leq p\}, \quad (2.5)
$$

which determines the dimension of the selected model. If $\hat{q} = O(q)$, the selected model has comparable dimension as the underlying model. Following Zhang and Huang (2008), we also consider two measures of the selected model. The first measures the error of the selected model,

$$
\tilde{\nu} = \|(I - \hat{P})X\beta\|_2, \quad (2.6)
$$

where $\hat{P}$ is the projection matrix from $R^n$ to the linear span of the set of selected groups and $I \equiv I_n \times n$ is the identity matrix. Thus $\tilde{\nu}^2$ is the sum of squares of the mean vector not
explained by the selected model. To measure the important groups missing in the selected model, define
\[
\zeta_2 = \left( \sum_{k \notin A_0} \|\hat{\beta}_k\|_2^2 I\{\|\hat{\beta}_k\|_2 = 0\} \right)^{1/2}.
\] (2.7)

We now describe several quantities that will be useful in describing the main results. Let \(d_a = \max_{1 \leq k \leq p} d_k\), \(d_b = \min_{1 \leq k \leq p} d_k\), \(d = d_a/d_b\) and \(N_d = \sum_{k=1}^p d_k\). Define
\[
r_1 \equiv r_1(\lambda) = \left( \frac{nc^* \sqrt{d_a} \eta_1}{\lambda d_b q} \right)^{1/2}, \quad r_2 \equiv r_2(\lambda) = \left( \frac{nc^* \eta_2^2}{\lambda^2 d_b q} \right)^{1/2}, \quad \bar{c} = \frac{c^*}{c_s},
\] (2.8)
where \(\eta_2 \equiv \max_{A \subset A_0} \|\sum_{k \in A} X_k \beta_k\|_2\),
\[
M_1 \equiv M_1(\lambda) = 2 + 4r_1^2 + 4\sqrt{d}\bar{c}r_2 + 4d\bar{c}, \quad M_2 \equiv M_2(\lambda) = \frac{2}{3} \left(1 + 4r_1^2 + 2\bar{c} + \frac{4}{3}d\bar{c}\right), \quad M_3 \equiv M_3(\lambda) = \frac{2}{3} \left(1 + 4r_1^2 + 2\bar{c}\right).
\] (2.9), (2.10), (2.11)

Let \(\lambda_{n,p} = 2\sigma \sqrt{8(1 + c_0)d_a d^2 q^* c n c^* \log(N_d \vee a_n)}\), where \(c_0 \geq 0, a_n \geq 0\), satisfying \(pd_a/(N_d \vee a_n)^{1+c_0} \approx 0\) and, \(\lambda_0 = \inf\{\lambda : M_1q + 1 \leq q^*\}, \inf\emptyset = \infty\). We also consider the following constraint
\[
\lambda \geq \max\{\lambda_0, \lambda_{n,p}\}.
\] (2.12)
For large \(p\), the lower bound here is allowed to be \(\lambda_{n,p} = 2\sigma [8(1 + c_0)d_a d^2 q^* c n c^* \log(N_d)]^{1/2}\) with \(a_n = 0\); for fixed \(p\), \(a_n \to \infty\) is required.

We assume the following basic condition.

**Condition (C1)** The errors \(\varepsilon_1, \ldots, \varepsilon_n\) are independent and identically distributed as \(N(0, \sigma^2)\).

**Theorem 2.1** Suppose that \(q \geq 1\) and that (C1), the GSC (2.2) and SRC (2.4) are satisfied. Let \(\hat{\beta}, \hat{\omega}\) and \(\zeta_2\) be defined as in (2.5), (2.6) and (2.7), respectively, for the model \(A\) selected by the group Lasso from (2.1). Let \(M_1, M_2\) and \(M_3\) be defined as in (2.9), (2.10) and (2.11), respectively. Then, if the constraint (2.12) is satisfied, the following assertions hold with probability converging to 1,
\[
\hat{q} \leq \#\{k : \|\hat{\beta}_k\|_2 > 0 \text{ or } k \notin A_0\} \leq M_1(\lambda)q, \quad \hat{\omega}^2 = \|(I - \hat{P})X\beta\|_2^2 \leq M_2(\lambda)B_1^2(\lambda), \quad \zeta_2^2 = \sum_{k \notin A_0} \|\hat{\beta}_k\|_2^2 I\{\|\hat{\beta}_k\|_2 = 0\} \leq \frac{M_3(\lambda)B_1^2(\lambda)}{c_s n},
\]
where \(B_1(\lambda) = ((\lambda d_b q^*)/(nc^*))^{1/2}\).

**Remark 2.1** The condition \(q \geq 1\) is not necessary, since it is only used to express quantities in terms of ratios in (2.8) and Theorem 2.1. If \(q = 0\), we use \(r_1^2 q = nc^* \sqrt{d_a} \eta_1/\lambda d_b\), \(r_2^2 q = nc^* \eta_2^2/\lambda^2 d_b\), to recover \(M_1, M_2\) and \(M_3\) in (2.9), (2.10), (2.11), respectively, resulting \(\hat{q} \leq 4nc^* \sqrt{d_a} \eta_1/\lambda d_b\), \(\hat{\omega}^2 \leq 8\lambda \sqrt{d_a d_b} \eta_1/3\) and \(\zeta_2^2 = 0\).

**Remark 2.2** If \(\eta_1 = 0\) in (2.2), then \(r_1 = r_2 = 0\) and
\[
M_1 = 2 + 4d\bar{c}, \quad M_2 = \frac{2}{3} \left(1 + 2\bar{c} + \frac{16}{3}d\bar{c}\right), \quad M_3 = \frac{2}{3} \left(1 + \frac{2}{3}d\bar{c}(7 + 4\bar{c})\right).
\]
all only depend on $d$ and $\bar{c}$. This suggests that the relative sizes of the groups affect the selection results. Since $d \geq 1$, the most favorable case is $d = 1$, that is, when the groups have equal sizes.

**Remark 2.3** If $d_1 = \cdots = d_p = 1$, the group Lasso simplifies to the Lasso, and Theorem 2.2 is a direct generalization of Theorem 1 on the selection properties of the Lasso obtained by Zhang and Huang (2008). In particular, when $d_1 = \cdots = d_p = 1$, $r_1, r_2, M_1, M_2, M_3$ are the same as the constants in Theorem 1 of Zhang and Huang (2008).

**Remark 2.4** A more general definition of the group Lasso is

$$
\hat{\beta}^* = \arg \min_{\beta} \frac{1}{2}(Y - X\beta)'(Y - X\beta) + \lambda \sum_{k=1}^p (\beta_k'R_k\beta_k)^{1/2},
$$

(2.13)

where $R_k$ is a $d_k \times d_k$ positive definite matrix. This is useful when certain relationships among the coefficients need be specified. By the Cholesky decomposition, there exists a matrix $Q_k$ such that $R_k = d_kQ_kQ_k'$. Let $\beta^* = Q_k\beta$, and $X_k = X_kQ_k^{-1}$. Then (2.13) becomes

$$
\hat{\beta}^* = \arg \min_{\beta} (Y - X^*\beta^*)'(Y - X^*\beta^*) + \lambda \sum_{k=1}^p \sqrt{d_k}\|\beta_k^*\|_2.
$$

The GSC for (2.13) is $\sum_{k=q+1}^p (\beta_k^*Q_k^*Q_k\beta_k^*)^{1/2} \leq \eta_1$. The SRC can be assumed for $X \cdot Q^{-1}$, where $X \cdot Q^{-1} = (X_1Q_1^{-1}, \cdots, X_pQ_p^{-1})$.

Immediately from Theorem 2.1, we have the following corollary.

**Corollary 2.1** Suppose that the conditions of Theorem 2.1 hold and $\lambda$ satisfies the constraint (2.12). Then, with probability converging to one, all groups with $\|\beta_k\|_2^2 > M_3(\lambda)\lambda q^2/(c_*c^*n^2)$ are selected.

From Theorem 2.1 and Corollary 2.1, the group Lasso possesses similar properties to the Lasso in terms of sparsity and bias (Zhang and Huang 2008). In particular, the group Lasso selects a model whose dimension has the same order as the underlying model. Furthermore, all the groups with coefficients whose $\ell_2$ norm are greater than the threshold model given in Corollary 2.1 are selected with high probability.

**Theorem 2.2** Let $\{\bar{c}, \sigma, r_1, r_2, c_0, d\}$ be fixed and $1 \leq q \leq n \leq p \to \infty$. Suppose that the conditions in Theorem 2.1 hold. Then with probability converging to 1,

$$
\|\hat{\beta} - \beta\|_2 \leq \frac{1}{\sqrt{nc_*}} \left(2\sigma \sqrt{M_1 \log(N_d)}q + (r_2 + \sqrt{dM_1\bar{c}})B_1\right) + \sqrt{\frac{c_*r_1^2 + r_2^2}{c_*c^*}} \sqrt{\frac{q\lambda}{n}},
$$

and

$$
\|X\hat{\beta} - X\beta\|_2 \leq 2\sigma \sqrt{M_1 \log(N_d)}q + (2r_2 + \sqrt{dM_1\bar{c}})B_1.
$$

Theorem 2.2 is stated for a general $\lambda$ that satisfies (2.12). The following result is an immediate corollary of Theorem 2.2.

**Corollary 2.2** Let $\lambda = 2\sigma \sqrt{8(1 + c'_0)d_6d^2q^*\bar{c}\bar{c}^*n \log(N_d)}$ with a fixed $c'_0 \geq c_0$. Suppose all the conditions in Theorem 2.2 hold. Then

$$
\|\hat{\beta} - \beta\|_2 = O_p\left(\sqrt{q \log(N_d)/n}\right) \quad \text{and} \quad \|X\hat{\beta} - X\beta\|_2 = O_p\left(\sqrt{q \log(N_d)}\right).
$$

This corollary follows by substituting the given $\lambda$ value into the expressions in the results of Theorem 2.2.
3 Selection consistency of the adaptive group Lasso

As shown in the previous section, the group Lasso has excellent selection and estimation properties. However, there is room to improve, in particular, with regard to selection. Although the group Lasso selects a model whose dimension is comparable to the underlying model, the simulation results reported in Yuan and Lin (2006) and those given below suggest that it tends to select more groups than there are in the underlying model. To correct the tendency of over selection by the group Lasso, we generalize the idea of the adaptive Lasso (Zou 2006) for individual variable selection to the present problem of group selection.

Consider a general group Lasso criterion with a weighted penalty term,\[ \frac{1}{2}(Y - X\beta)'(Y - X\beta) + \sum_{k=1}^{p} w_k \sqrt{d_k} \| \beta_k \|_2, \] (3.1) where \( w_k \) is the weight associated with the \( k \)th group. The \( \lambda_k \equiv \tilde{\lambda} w_k \) can be regarded as the penalty level corresponding to the \( k \)th group. For different groups, the penalty level \( \lambda_k \) can be different. If we can have lower penalty for groups with large coefficients and higher penalty for the groups with small coefficients (in the \( \ell_2 \) sense), we expect to be able to improve variable selection accuracy and reduce estimation bias. One way to obtain the information about whether a group has large or small coefficients is by using a consistent initial estimator. Suppose that an initial estimate \( \tilde{\beta} \) is available. A simple approach to determining the weight is to use the initial estimator. Consider \[ w_k = \frac{1}{\| \tilde{\beta}_k \|_2}, \quad k = 1, \ldots, p. \] (3.2)

Thus for each group, its penalty is proportional to the inverse of the norm of \( \tilde{\beta}_k \). This choice of the penalty levels for each group is a natural generalization of the adaptive Lasso (Zou 2006). In particular, when each group only contains a single variable, (3.2) simplifies to the adaptive Lasso penalty.

Let \( \theta_a = \max_{k \in A_0} \| \beta_k \|_2 \), \( \theta_b = \min_{k \in A_0} \| \beta_k \|_2 \). We say that an initial estimator \( \tilde{\beta} \) is consistent at zero with rate \( r_n \) if \( r_n \max_{k \in A_0} \| \tilde{\beta}_k \|_2 = O_p(1) \) where \( r_n \to \infty \) as \( n \to \infty \) and, there exists a constant \( \xi_b > 0 \) such that, for any \( \varepsilon > 0 \), \( P(\min_{k \in A_0} \| \tilde{\beta}_k \|_2 > \xi_b \theta_b) > 1 - \varepsilon \) for \( n \) sufficiently large.

In addition to (C1), we assume the following conditions.

(C2) The initial estimator \( \tilde{\beta} \) is consistent at zero with rate \( r_n \to \infty \).

(C3) \[ \frac{\sqrt{d_a \log q}}{\sqrt{n} \theta_b} \to 0, \quad \frac{\tilde{\lambda} d_a^{3/2}}{n \theta_b^2} \to 0, \quad \frac{\sqrt{nd \log(p - q)}}{\bar{\lambda} r_n} \to 0, \quad \frac{d_a^{5/2} q^2}{r_n \theta_b \sqrt{d_b}} \to 0. \]

(C4) All the eigenvalues of \( \Sigma_{A^0} \Sigma_{A^0} \) are bounded away from zero and infinity.

Condition (C2) assumes that an initial zero-consistent estimator exists. It is the most critical one and it is general difficult to establish. It assumes that we can consistently differentiate between important and non-important groups. For fixed \( p \) and \( d_k \), the ordinary least squares estimator can be used as the initial estimator. However, when \( p > n \), the least squares estimator is no longer feasible. By Theorems 2.1 and 2.2, the group Lasso
estimator $\hat{\beta}$ is consistent at zero with rate $\sqrt{n/(q \log(Nd))}$. Condition (C3) restricts the numbers of important and non-important groups, as well as variables within the groups. It also places constraints on the penalty parameter and the $\ell_2$ norm of the smallest important group. Condition (C4) assumes that the eigenvalues of $\Sigma_{A^c, A^c}$ are finite and bounded away from zero. This is reasonable since the number of important groups is small in a sparse model. This condition ensures that the true model is identifiable.

Define

$$\hat{\beta}^* = \arg \min \frac{1}{2}(Y - X\beta)'(Y - X\beta) + \tilde{\lambda} \sum_{k=1}^{P} \|\tilde{\beta}_k\|_{2}^{-1} \sqrt{d_k} \|\beta_k\|_2.$$  \hspace{1cm} (3.3)

**Theorem 3.1** Suppose that (C1)-(C4) and NSC (2.3) are satisfied. Then

$$P(\|\hat{\beta}^*_k\|_2 \neq 0, k \notin A_0, \|\hat{\beta}^*_k\|_2 = 0, k \in A_0) \to 1.$$  

Therefore the adaptive group Lasso is selection consistent if the conditions stated in Theorem 3.1 hold.

If we use $\hat{\beta}$ as the initial estimator, (C3) can be changed to

(C3)*

$$\sqrt{\frac{d_a(\log q)}{n\theta_b}} \to 0, \frac{\tilde{\lambda} d_a^{3/2} q}{n\theta_b^2} \to 0, \sqrt{\frac{dq \log(p - q) \log(Nd)}{\lambda}} \to 0, \frac{(d_aq)^{5/2} \sqrt{\log(Nd)}}{\theta_b \sqrt{n\theta_b}} \to 0.$$  

We often have $\tilde{\lambda} = n^\alpha$ for some $0 < \alpha < 1/2$. In this case, the number of non-important groups can be as large as $\exp(n^{2\alpha}/(q \log q))$ with the number of important groups satisfying $q^5 \log q/n \to 0$, assuming that $\theta_b$ and the number of variables within the groups are finite.

**Corollary 3.1** Let the initial estimator $\tilde{\beta} = \hat{\beta}$, where $\hat{\beta}$ is the group Lasso estimator. Supposed that the NSC (2.3) holds and that (C1), (C2), (C3)* and (C4) are satisfied. Then

$$P(\|\hat{\beta}^*_k\|_2 \neq 0, k \notin A_0, \|\hat{\beta}^*_k\|_2 = 0, k \in A_0) \to 1.$$  

This corollary follows directly from Theorem 3.1. It shows that the iterated group Lasso procedure that uses a combination of the group Lasso and the adaptive group Lasso is selection consistent.

**Theorem 3.2** Suppose that the conditions in Theorem 2.2 hold and $\theta_b > t_b$ for some constant $t_b > 0$. If $\tilde{\lambda} \sim O(n^{\alpha})$ for some $0 < \alpha < 1/2$, then

$$\|\hat{\beta}^* - \beta\|_2 = O_p\left(\sqrt{\frac{q}{n} + \frac{\tilde{\lambda}^2}{n^2}}\right) = O_p\left(\sqrt{\frac{q}{n}}\right), \|X\hat{\beta}^* - X\beta\|_2 \sim O\left(\sqrt{\frac{q + \tilde{\lambda}^2}{n}}\right) = O_p\left(\sqrt{q}\right).$$

Theorem 3.2 implies that for the adaptive group Lasso, given a zero-consistent initial estimator, we can reduce a high-dimensional problem to a lower-dimensional one. The convergence rate is improved compared with that of the group Lasso by choosing appropriate penalty parameter $\tilde{\lambda}$. 


4 Simulation studies

In this section, we use simulation to evaluate the finite sample performance of the group Lasso and the adaptive group Lasso. Consider \( \lambda_k = \hat{\lambda}/\|\hat{\beta}_k\|_2 \), if \( \|\hat{\beta}_k\|_2 > 0 \); \( \lambda_k = \infty \), if \( \|\hat{\beta}_k\|_2 = 0 \). Thus we can drop the corresponding covariates \( X_k \) from the model and only consider the groups with \( \|\hat{\beta}_k\|_2 > 0 \). After a scale transformation, we can directly apply the group least angle regression algorithm (Yuan and Lin 2006) to compute the adaptive group Lasso estimator \( \hat{\beta}^* \). The penalty parameters for the group Lasso and the adaptive group Lasso are selected using the BIC criterion (Schwarz 1978).

We consider two scenarios of simulation models. In the first scenario, the group sizes are equal. In the second, the group sizes vary. For every scenario, we consider the cases \( p < n \) and \( p > n \). In all the examples, the sample size is \( n = 200 \).

Example 1. In this example, there are 10 groups and each group consists of 5 covariates. The covariate vector is \( X = (X_1, \cdots, X_{10}) \) where \( X_j = (X_{5(j-1)+1}, \cdots, X_{5(j-1)+5}), 1 \leq j \leq 10 \). To generate \( X \), we first simulate 50 random variables \( R_1, \cdots, R_{50} \) independently from \( N(0, 1) \). Then \( Z_j, j = 1, \cdots, 10 \) are simulated from a multivariate normal distribution with mean zero and \( \text{Cov}(Z_{j1}, Z_{j2}) = 0.6|j_1-j_2| \). The covariates \( X_1, \cdots, X_{50} \) are generated as

\[
X_{5(j-1)+k} = \frac{Z_j + R_{5(j-1)+k}}{\sqrt{2}}, 1 \leq j \leq 10, 1 \leq k \leq 5.
\]

The random error \( \varepsilon \sim N(0, 3^2) \). The response variable \( Y \) is generated from \( Y = \sum_{k=1}^{10} X_k^T \beta_k + \varepsilon \), where \( \beta_1 = (0.5, 1, 1.5, 2, 2.5) \), \( \beta_2 = (2, 2, 2, 2, 2) \), \( \beta_3 = \cdots = \beta_{10} = (0, 0, 0, 0, 0) \).

Example 2. In this example, the number of groups is \( p = 10 \). Each group consists of 5 covariates. The covariates are generated the same way as in Example 1. However, the regression coefficients are \( \beta_1 = (0.5, 1, 1.5, 1, 0.5), \beta_2 = (1, 1, 1, 1, 1), \beta_3 = (-1, 0, 1, 2, 1.5), \beta_4 = (-1.5, 1, 0.5, 0.5, 0.5), \beta_5 = \cdots = \beta_{10} = (0, 0, 0, 0, 0) \).

Example 3. In this example, the number of groups \( p = 210 \) is bigger than the sample size \( n \). Each group consists of 5 covariates. The covariates are generated the same way as in Example 1. However, the regression coefficients are \( \beta_1 = (0.5, 1, 1.5, 1, 0.5), \beta_2 = (1, 1, 1, 1, 1), \beta_3 = (-1, 0, 1, 2, 1.5), \beta_4 = (-1.5, 1, 0.5, 0.5, 0.5), \beta_5 = \cdots = \beta_{210} = (0, 0, 0, 0, 0) \).

Example 4. In this example, the group sizes differ across groups. There are 5 groups with size 5 and 5 groups with size 3. The covariate vector is \( X = (X_1, \cdots, X_{10}) \) where \( X_j = (X_{5(j-1)+1}, \cdots, X_{5(j-1)+5}), 1 \leq j \leq 5 \), and \( X_j = (X_{3(j-6)+26}, \cdots, X_{3(j-6)+28}), 6 \leq j \leq 10 \). In order to generate \( X \), we first simulate 40 random variables \( R_1, \cdots, R_{40} \) independently from \( N(0, 1) \). Then \( Z_j, j = 1, \cdots, 10 \) are simulated with a normal distribution with mean zero and \( \text{Cov}(Z_{j1}, Z_{j2}) = 0.6|j_1-j_2| \). The covariates \( X_1, \cdots, X_{40} \) are generated as

\[
X_{5(j-1)+k} = \frac{Z_j + R_{5(j-1)+k}}{\sqrt{2}}, 1 \leq j \leq 5, 1 \leq k \leq 5,
\]

\[
X_{3(j-6)+25+k} = \frac{Z_j + R_{3(j-6)+25+k}}{\sqrt{2}}, 6 \leq j \leq 10, 1 \leq k \leq 3.
\]

The random error \( \varepsilon \sim N(0, 3^2) \). The response variable \( Y \) is generated from \( Y = \sum_{k=1}^{10} X_k^T \beta_k + \varepsilon \), where \( \beta_1 = (0.5, 1, 1.5, 2, 2.5) \), \( \beta_2 = (2, 0, 0, 2, 2) \), \( \beta_3 = \cdots = \beta_{10} = (0, 0, 0, 0, 0) \), \( \beta_6 = (-1, -2, -3) \), \( \beta_7 = \cdots = \beta_{10} = (0, 0, 0) \).
Example 5. In this example, the number of groups is $p = 10$ and the group sizes differ across groups. The data are generated the same way as in Example 4. However, the regression coefficients are $eta_1 = (0.5, 1, 1.5, 2, 2.5), \beta_2 = (2, 2, 2, 2, 2), \beta_3 = (-1, 0, 1, 2, 3), \beta_4 = (-1.5, 2, 0, 0, 0), \beta_5 = (0, 0, 0, 0, 0), \beta_6 = (2, -2, 1), \beta_7 = (0, -3, 1.5), \beta_8 = (-1.5, 1.5, 2), \beta_9 = (-2, -2, -2), \beta_{10} = (0, 0, 0).

Example 6. In this example, the number of groups $p = 210$ and the group sizes differ across groups. The data are generated the same way as in Example 4. However, the regression coefficients are $eta_1 = (0.5, 1, 1.5, 2, 2.5), \beta_2 = (2, 2, 2, 2, 2), \beta_3 = (-1, 0, 1, 2, 3), \beta_4 = (-1.5, 2, 0, 0, 0), \beta_5 = \cdots = \beta_{100} = (0, 0, 0, 0, 0), \beta_{101} = (2, -2, 1), \beta_{102} = (0, -3, 1.5), \beta_{103} = (-1.5, 1.5, 2), \beta_{104} = (-2, -2, -2), \beta_{105} = \cdots = \beta_{210} = (0, 0, 0).

The results are given in Table 1 based on 400 replications. The columns in the table include the average number of groups selected with standard error in parentheses, the median number (med) of groups selected with the 25% and 75% quantiles of the number of selected groups in parentheses, model error (ME), percentage of occasion on which correct groups are included in the selected model (%incl) and percentage of occasions on which the exactly correct groups are selected (%sel) with standard error in parentheses.

Several observations can be made from Table 1. First, in all six examples, the adaptive group Lasso performs better than the group Lasso in terms of model error and the percentage of correctly selected models. The group Lasso which gives the initial estimator for the adaptive group Lasso includes the correct groups with high probability. And the improvement is considerable for models with different group sizes. Second, the results from models with equal group sizes (Examples 1, 2 and 3) are better than those from models with different group ones (Examples 4, 5 and 6). Finally, when the dimension of the model increases, the performance of both methods becomes worse. This is to be expected since selection in models with a larger number of groups is more difficult.

| $\sigma = 3$ | group Lasso | | | | | adaptive group Lasso | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | mean | med | ME | % incl | % sel | mean | med | ME | % incl | % sel |
| Ex 1, [2] | 2.04 | 2 | 8.79 | 100 | 96.5 | 2.01 | 2 | 8.54 | 100 | 99.5 |
| | (0.18) | (2.2) | (0.94) | (0.18) | | (0.07) | (2.2) | (0.90) | (0.07) | |
| Ex 2, [4] | 4.11 | 4 | 8.52 | 99.5 | 88.5 | 4.00 | 4 | 8.10 | 99.5 | 98.0 |
| | (0.34) | (4.4) | (0.94) | (0.07) | (0.32) | | (0.14) | (4.4) | (0.87) | (0.07) | (0.14) |
| Ex 3, [4] | 4.00 | 4 | 9.48 | 93.0 | 86.5 | 3.94 | 4 | 8.19 | 93.0 | 92.5 |
| | (0.38) | (4.4) | (1.19) | (0.26) | (0.34) | | (0.27) | (4.4) | (0.96) | (0.26) | (0.26) |
| Ex 4, [3] | 3.17 | 3 | 8.78 | 100 | 85.3 | 3 | 3 | 8.36 | 100 | 100 |
| | (0.45) | (3.3) | (1.00) | (0.35) | | | (0) | (3.3) | (0.90) | (0) | (0) |
| Ex 5, [8] | 8.88 | 9 | 7.68 | 100 | 40.0 | 8.03 | 8 | 7.58 | 100 | 97.5 |
| | (0.81) | (8.10) | (0.94) | (0.49) | | | (0.16) | (8.8) | (0.86) | (0) | (0.16) |
| Ex 6, [8] | 12.90 | 9 | 14.61 | 66.5 | 7.0 | 11.49 | 8 | 9.28 | 66.5 | 47.0 |
| | (12.42) | (8.11) | (7.21) | (0.47) | (0.26) | | (12.68) | (7.8) | (5.79) | (0.47) | (0.50) |
5 Concluding remarks

We have studied the asymptotic selection and estimation properties of the group Lasso and adaptive group Lasso in “large $p$, small $n$” linear regression models. For the adaptive group Lasso to be selection consistent, the initial estimator should possess two properties: (a) it does not miss important groups and variables; and (b) it is estimation consistent, although it may not be group-selection or variable-selection consistent. Under the conditions stated in Theorem 2.1, the group Lasso is shown to satisfies these two requirements. Thus the iterated group Lasso procedure, which uses the group Lasso to achieve dimension reduction and generate the initial estimates and then uses the adaptive group Lasso to achieve selection consistency, is an appealing approach to group selection in high-dimensional settings.

Acknowledgements. The authors are grateful to Professor Cun-Hui Zhang for sharing his insights into the problem and related topics. Huang is also a member of the Department of Biostatistics of the University of Iowa. The work of Huang is supported in part by the U.S. NIH grant R01CA120988 and NSF grants DMS 0706108 and 0805670.

6 Proofs

We first define some notations which will be used in proofs. Let $\{k : \| \hat{\beta}_k \|_2 > 0, k \leq p\} \subseteq A_1 \subseteq \{k : X_k^T(Y - X\hat{\beta}) = \lambda \sqrt{d_k}\| \hat{\beta}_k \|_2 \} \cup \{1, \ldots, q\}$. Set $A_2 = \{1, \ldots, p\} \setminus A_1$, $A_3 = A_1 \setminus A_0$, $A_4 = A_1 \cap A_0$, $A_5 = A_2 \setminus A_0$, $A_6 = A_2 \cap A_0$. Thus $A_1 = A_3 \cup A_4$, $A_3 \cap A_4 = \emptyset$, $A_2 = A_5 \cup A_6$, $A_5 \cap A_6 = \emptyset$. Let $|A_i| = \sum_{k \in A_i} d_k$, $N(A_i) = \#\{k : k \in A_i\}$, $i = 1, \ldots, 6$ and $q_1 = N(A_1)$.

Proof of Theorem 2.1. The basic idea used in this proof follows the proof of the rate consistency of the Lasso in Zhang and Huang (2008). However, there are many differences in technical details, e.g., in the characterization of the solution via the KKT conditions, in the constraint needed for the penalty level and in the use of maximal inequalities.

The proof consists of three steps. Step 1 proves some inequalities related to $q_1$, $\hat{\omega}$ and $\zeta_2$. Step 2 translates the results of Step 1 into upper bounds for $\hat{q}$, $\hat{\omega}$ and $\zeta_2$. Step 3 completes the proof by showing the probability of the event in Step 2 converging to one. The details of the complete proof is available from the website www.stat.uiowa.edu/techrep. The sketch of the proof is following:

If $\hat{\beta}$ is a solution of (2.1), by the KKT condition, $X_k'(Y - X\hat{\beta}) = \lambda \sqrt{d_k}\| \hat{\beta}_k \|_2$, $\forall \| \hat{\beta}_k \|_2 > 0$, and $-\lambda \sqrt{d_k} \leq X_k'(Y - X\hat{\beta}) \leq \lambda \sqrt{d_k}, \forall \| \hat{\beta}_k \|_2 = 0$. Then

$$\sigma_{11}^{-1} S_{A_i}/n = (\beta_{A_i} - \hat{\beta}_{A_i}) + \sigma_{11}^{-1} \Sigma_{12} \beta_{A_2} + \sigma_{11}^{-1} X_{A_i}' e/n,$$

$$n \sigma_{22} \beta_{A_2} - n \sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \beta_{A_2} \leq C_{A_2} - X_{A_2}' e - \sigma_{21} \Sigma_{11}^{-1} S_{A_1} + \sigma_{21} \Sigma_{11}^{-1} X_{A_2}' e.$$  \hspace{1cm} (6.1)

where $S_{A_i} = (S_{k_1}', \ldots, S_{k_n}')$, $S_{k_i} = \lambda \sqrt{d_k} s_{k_i}, s_{k_i} = X_{k_i}'(Y - X\hat{\beta})/(\lambda \sqrt{d_k}), C_{A_i} = (C_{k_1}', \ldots, C_{k_n}')$, $C_{k_i} = \lambda \sqrt{d_k} I(\| \hat{\beta}_{k_i} \|_2 = 0) e_{d_{k_i} \times 1}$, all the elements of matrix $e_{d_{k_i} \times 1}$ is 1, $k_i \in A_i$ and $\Sigma_{ij} = X_{A_i} X_{A_j}/n$.

Step 1. Define

$$V_{1j} = \sigma_{11}^{-1/2} Q_{A_{1j}} S_{A_j}/\sqrt{n}, \ j = 1, 3, 4, \ \omega_k = (I - P_{A_1}) X_{A_k} \beta_{A_k}, \ k = 2, \ldots, 6,$$

where $Q_{A_{1j}}$ is the matrix representing the selection of variables in $A_k$ from $A_j$. Define $u = (X_{A_1} \Sigma_{11}^{-1} Q_{A_{11}} S_{A_1}/n - \omega_2)/\| X_{A_1} \Sigma_{11}^{-1} Q_{A_{11}} S_{A_1}/n - \omega_2 \|_2$. From (6.1) and (6.2), $V_{14}(V_{13} +$
\[ V_{14} \leq S'_{A}Q_{A_{1}}\Sigma_{11}^{-1}12\beta_{A_{2}} + S'_{A}Q_{A_{1}}\Sigma_{11}^{-1}X_{A_{2}}'\varepsilon/n + \sqrt{d_{a}}\lambda\sum_{k\in A_{4}}\|\beta_{k}\|_{2}, \|\omega_{2}\|_{2}^{2} \leq \beta'_{A_{2}}(C_{A_{2}} - X_{A_{2}}'\varepsilon - \Sigma_{21}\Sigma_{11}^{-1}S_{A_{1}} + \Sigma_{21}\Sigma_{11}^{-1}X_{A_{2}}'\varepsilon), \] then under GSC,

\[
\|V_{14}\|_{2}^{2} + \|\omega_{2}\|_{2}^{2} \leq (\|V_{14}\|_{2}^{2} + \|\omega_{2}\|_{2}^{2})^{1/2}u' \varepsilon + (\|V_{14}\|_{2} + \|P_{A_{1}}X_{A_{2}}\beta_{A_{2}}\|_{2}) \left( \frac{\lambda^{2}d_{a}N(A_{3})}{nc_{s}(|A_{1}|)} \right)^{1/2} + \sqrt{d_{a}}\lambda\eta_{1} + \lambda\sqrt{d_{a}}\beta_{A_{2}}\|_{2}.
\]  

(6.3)

**Step 2.** Define \( B_{1}^{2} = \lambda^{2}d_{a}q/(nc_{s}(|A_{1}|)) \) and \( B_{2}^{2} = \lambda^{2}d_{a}q/(nc_{s}(|A_{0}| \lor |A_{1}|)). \) In this step, we consider the event \(|u' \varepsilon|^{2} \leq (|A_{1}| \lor d_{b})B_{1}^{2}/(4q_{d_{a}}). \) Consider the set \( A_{1} \) contains all large \( \beta_{k} \neq 0. \) From (6.3), \( \|V_{14}\|_{2}^{2} \leq B_{1}^{2} + 4\sqrt{d_{a}}\lambda\eta_{1} + 4\sqrt{d_{a}}\eta_{2}B_{2} + 4dB_{2}^{2}, \) then

\[
(q_{1} - q)^{+} \leq q + \frac{nc_{s}(|A_{1}|)}{\lambda^{2}d_{b}} \left( \frac{4\sqrt{d_{a}}\lambda\eta_{1} + 4\sqrt{\frac{\lambda^{2}d_{a}q}{nc_{s}(|A_{1}|)}}\eta_{2} + 2\lambda^{2}d_{a}q}{nc_{s}(|A_{1}|)} \right). 
\]  

(6.4)

For general \( A_{1}, \) let \( C_{5} = c^{*}(|A_{5}|)/c_{s}(|A_{1}| \lor |A_{5}|). \) From (6.3),

\[
\|\omega_{2}\|_{2}^{2} \leq 4\left( \frac{B_{1}^{2}}{2} + dB_{2}^{2} + \sqrt{d(1 + \sqrt{C_{5}})\eta_{2}B_{2} + 2\sqrt{d_{a}}\eta_{1}} \right) + \frac{32}{9}dC_{5}B_{2}^{2}.
\]  

(6.5)

From Zhang and Huang (2008), \( \|\omega_{2}\|_{2}^{2} \geq (\|\beta_{A_{1}}\|_{2}(nc_{s}5)^{1/2} - \eta_{2})^{2}, \|X_{A_{2}}\beta_{A_{2}}\|_{2} \leq \eta_{2} + \|X_{A_{5}}\beta_{A_{5}}\|_{2} \leq \eta_{2} + (nc_{s}(|A_{5}|))^{1/2}\|\beta_{A_{2}}\|_{2}, \) then

\[
\|\beta_{A_{5}}\|_{2}^{2}nc_{s,5} \leq \left[ \frac{4}{3}\lambda\sqrt{\frac{d_{a}q}{nc_{s,5}}} \left( 1 + \frac{c^{*}(|A_{5}|)}{c_{s}(|A_{1}|)} \right)^{1/2} + 2\eta_{2} \right]^{2} \]  

+ \[ \frac{8}{3} \left[ \frac{B_{1}^{2}}{4} + \sqrt{d_{a}}\lambda\eta_{1} + \eta_{2} \left( \frac{\lambda^{2}d_{a}q}{nc_{s}(|A_{1}|)} \right)^{1/2} + \frac{\lambda^{2}d_{a}q}{2nc_{s}(|A_{1}|)} - \frac{3}{4}B_{2}^{2} \right].
\]  

(6.6)

where \( c_{s,5} = c_{s}(|A_{1} \lor A_{5}|). \)

**Step 3.** Consider \( c_{s}(|A_{m}|) = c_{s}, c^{*}(|A_{m}|) = c^{*}, \) for \( N(A_{m}) \leq q^{*}, \)

\[
q_{1} \leq N(A_{1} \lor A_{5}) \leq q^{*}, \quad |u' \varepsilon|^{2} \leq \frac{(|A_{1}| \lor d_{b})\lambda^{2}d_{b}}{4d_{a}nc^{*}(|A_{1}|)}.
\]  

(6.7)

We have \( \hat{c} = C_{5} = c^{*}(|A_{5}|)/c_{s}(|A_{1} \lor |A_{5}|) = c^{*}/c_{s} \) and \( c_{s,5} = c_{s}(|A_{1} \lor A_{5}|) = c_{s}. \) From (6.4), (6.5) and (6.6), \( (q_{1} - q)^{+} + q \leq M_{1}q, \|\omega_{2}\|_{2}^{2} \leq M_{2}B_{1}^{2}, \) when (2.12) is satisfied. Define

\[
x_{m}^{*} \equiv \max_{|A|=m} \max_{\|U_{A_{k}}\|_{2}=1, k=1, \ldots, m} \frac{\|X_{A}(X_{A}'X_{A})^{-1}S_{A} - (I - P_{A})X_{A}\|_{2}}{\|X_{A}(X_{A}'X_{A})^{-1}S_{A} - (I - P_{A})X_{A}\|_{2}},
\]  

(6.8)

for \( |A| = q_{1} = m \geq 0, \) \( S_{A_{k}} = (S'_{A_{1}}', \ldots, S'_{A_{m}})' \) where \( S_{A_{k}} = \lambda\sqrt{d_{A_{k}}U_{A_{k}}, \|U_{A_{k}}\|_{2} = 1.} \) Let \( Q_{A} = X_{A}'(X_{A}'X_{A})^{-1} \) where \( X_{k}' = \lambda\sqrt{d_{k}X_{k}} \) for \( k \in A. \) For a given \( A, \) let \( V_{ij} = (0, \ldots, 0, 1, 0, \ldots, 0) \) be the \( |A| \times 1 \) vector with the \( j \)th element in the \( l \)th group to be 1, then by (6.8),

\[
x_{m}^{*} \leq \max_{|A|=m} \max_{l_{j}} \left\{ \left( \frac{\varepsilon'}{\|Q_{A}V_{ij}\|_{2}} + \frac{\|Q_{A}V_{ij}\|_{2}\Sigma_{l_{i} \in A}\sqrt{d_{l}}}{\|Q_{A}U_{A}\|_{2}} + \frac{\varepsilon'(I - P_{A})X_{A}\|_{2}}{|(I - P_{A})X_{A}\|_{2}} \right) \right\}.
\]  


Define \( \Omega_{m_0} = \{(U, \varepsilon) : x_m^* \leq \sqrt{8(1 + c_0)V^2((md_b) \lor d_b) \log(N_d \lor a_n)}, \forall m \geq m_0 \} \). Then \((X, \varepsilon) \in \Omega_{m_0} \Rightarrow \|u\varepsilon\|^2 \leq (x_m^*)^2 \leq (|A| \lor d_b)^2/(4d_an\epsilon^2_c), \) for \( N(A_1) \geq m_0 \geq 0 \). By the definition of \( x_m^* \), it is less than the maximum of \( \binom{p}{m} \sum_{k \in A} d_k \) normal variables with mean 0 and variance \( \sigma^2 V_k^2 \), plus the maximum of \( \binom{p}{m} \) normal variables with mean 0 and variance \( \sigma^2 \). Then \( P\{(X, \varepsilon) \in \Omega_{m_0}\} \rightarrow 1 \) when (6.7) holds. This complete the sketch of the proof of Theorem 2.1.

**Proof of Theorem 2.2.** Consider the case when \( \{c^*, c_*, r_1, r_2, c_0, \sigma, d\} \) are fixed. Then the required configurations in Theorem 2.1 become

\[
M_1 q + 1 < q^*, \quad \eta_1 \leq \frac{r_1^2}{c^*} \frac{q\lambda}{n}, \quad \eta_2 \leq \frac{r_2^2}{c^*} \frac{q\lambda^2}{n}.
\]

Let \( A_1 = \{k : \|\hat{\beta}_k\|_2 > 0 \text{ or } k \notin A_0\} \). Define \( v_1 = X_{A_1}(\hat{\beta}_{A_1} - \beta_{A_1}) \), \( g_1 = X'_{A_1}(Y - X\hat{\beta}) \). Then \( \|v_1\|^2 \geq c_* n \|\hat{\beta}_{A_1} - \beta_{A_1}\|_2^2 \) \((\|\hat{\beta}_{A_1} - \beta_{A_1}\)\()^\prime g_1 = v_1'(X\hat{\beta} - X_{A_1}\hat{\beta}_{A_1} + \varepsilon) - \|v_1\|^2\), and \( \|g_1\|_\infty \leq \max_{k, \|\hat{\beta}_k\|_2 > 0} \lambda \sqrt{d_k} \|k\|_2 \|\hat{\beta}_k\|_2 = \lambda d_a \). Therefore, \( \|v_1\|^2 \leq \eta_2 + \|P_{A_1}\varepsilon\|^2 + \lambda \sqrt{d_a N(A_1)/(nc_*)} \). Since \( \|P_{A_1}\varepsilon\|^2 \leq 2\sigma V N(A_1) \log(N_d) \) with probability converging to one under the normality assumption, \( \|X(\hat{\beta} - \beta)\|_2 \leq 2\eta_2 + \|P_{A_1}\varepsilon\|^2 + \lambda \sqrt{d_a N(A_1)/(nc_*)} \). Then

\[
(\sum_{k \in A_1} \|\hat{\beta}_k - \beta_k\|_2^2) / 2 \leq \frac{\|v_1\|^2}{\sqrt{nc_*}} \leq \frac{1}{\sqrt{nc_*}} (\eta_2 + 2\sigma V N(A_1) \log(N_d) + \sqrt{dM_1 e B_1}).
\]

Since \( A_2 \subset A_0 \), by the second inequality in (6.9), \#\(\{k \in A_0 : \|\hat{\beta}_k\|_2 > \lambda/n\} \leq r_1^2 q/c^* \sim O(q) \). By the SRC and the third inequality in (6.9), \( \sum_{k \in A_0} \|\hat{\beta}_k\|^2 I\{\|\hat{\beta}_k\|_2 > \lambda/n\} \leq \sum_{k \in A_0} \|X_k\hat{\beta}_k I\{\|\hat{\beta}_k\|_2 > \lambda/n\}\|^2/(nc_*) \leq r_2^2 q\lambda^2/(n^2 c_* c^*) \), and \( \sum_{k \in A_0} \|\hat{\beta}_k\|^2 I\{\|\hat{\beta}_k\|_2 \leq \lambda/n\} \leq r_1^2 q\lambda^2/(n^2 c^*) \). Then from (6.10), we have

\[
\|\hat{\beta} - \beta\|_2 \leq \sqrt{\frac{\|v_1\|^2}{\sqrt{nc_*}}} (2\sigma V M_1 \log(N_d) q + (r_2 + \sqrt{dM_1 e B_1}) B_1) + \sqrt{\frac{c_* r_1^2 + r_2^2}{c_e c^*}} \frac{\sqrt{q\lambda}}{n},
\]

\[
\|X\hat{\beta} - X\beta\|_2 \leq 2\sigma V M_1 \log(N_d) q + (r_2 + \sqrt{dM_1 e B_1}) B_1.
\]

This complete the proof of Theorem 2.2.

**Proof of Theorem 3.1.** Let \( \hat{u} = \hat{\beta} - \beta, W = X'\varepsilon / \sqrt{n}, V(u) = \sum_{i=1}^n [((\varepsilon_i - x_i u)^2 - \varepsilon_i^2)] + \sum_{k=1}^p \lambda_k \sqrt{d_k} \|u_k + \beta_k\|_2 \) and \( \hat{u} = \min(\varepsilon - X\hat{\beta})' (\varepsilon - X\hat{\beta}) + \sum_{k=1}^p \lambda_k \sqrt{d_k} \|u_k + \beta_k\|_2 \), where \( \lambda_k = \lambda / \|\hat{\beta}_k\|_2 \). By the KKT conditions, if there exists \( \hat{u} \) such that

\[
\Sigma A^c_0 \hat{A}_0^c (\sqrt{n}\hat{u}_A^c) - W A^c_0 = -S A^c_0 / \sqrt{n}, \quad \|\hat{u}_k\|_2 \leq \|\beta_k\|_2, k \in A^c_0,
\]

\[
-C A_0 / \sqrt{n} \leq \Sigma A_0 A^c_0 (\sqrt{n}\hat{u}_A^c) - W A_0 \leq C A_0 / \sqrt{n},
\]

then \( \|\hat{\beta}_k\|_2 \neq 0 \), for \( k = 1, \ldots, q \), and \( \|\hat{\beta}_k\|_2 = 0 \), for \( k = q + 1, \ldots, p \).

From (6.11) and (6.12), \( (\sqrt{n}\hat{u}_A^c) - \Sigma A^c_0 A^c_0 W A^c_0 = -\frac{1}{\sqrt{n}} \Sigma A^c_0 A^c_0 S A^c_0 \), \( \Sigma A_0 A^c_0 (\sqrt{n}\hat{u}_A^c) - W A_0 = -n^{-1/2} X' A_0 (I - P A_0) \varepsilon \neq -n^{-1/2} \Sigma A_0 A^c_0 A^c_0 S A_0^c \). Define the events

\[
E_1 = \{n^{-1/2}\|(\Sigma A^c_0 A^c_0 X' A_0)\varepsilon\|_2 < \sqrt{n} \|\hat{\beta}_k\|_2 - n^{-1/2}\|(\Sigma A^c_0 A^c_0 S A_0^c)\varepsilon\|_2, k \in A^c_0\},
\]

\[
E_2 = \{n^{-1/2}\|(X' A_0 (I - P A_0)\varepsilon)\|_2 < n^{-1/2}\|C\|_2 - n^{-1/2}\|(\Sigma A_0 A^c_0 A^c_0 S A_0^c)\varepsilon\|_2, k \in A_0\},
\]

where \( \cdot \) means the \( d_k \) dimensional sub-vector of vector \( \cdot \) corresponding to the \( k \)th group. Then \( P(\|\hat{\beta}_k\|_2 \neq 0, k \in A_0, \|\hat{\beta}_k\|_2 = 0, k \notin A_0 \) \( \geq P(E_1 \cap E_2) \), \( P(E_1 \cap E_2) = 1 - P(E_1^c \cup E_2^c) \geq 1 - P(E_1^c) - P(E_2^c) \).
First, we consider $P(E_1^c)$. Define $R = \{\|\hat{\beta}_k\|_2^{-1} \leq c_1\theta_b^{-1}, k \in A_0^c\}$ where $c_1$ is a constant. $P(E_1^c) = P(E_1^c \cap R) + P(E_1^c \cap R^c) \leq P(E_1^c \cap R) + P(R^c)$. By (C2), $P(R^c) \to 0$. Let $N_q = \sum_{k=1}^q d_k$, $\tau_1 \leq \cdots \leq \tau_N_q$ be the eigenvalues of $\Sigma_{A_0^c}^{-1} \Sigma_{A_0^c}^{-1}$ and $\gamma_1, \cdots, \gamma_N_q$ be the associated eigenvectors. The $j$th element in the $l$th group of vector $\Sigma_{A_0^c}^{-1} \Sigma_{A_0^c}^{-1} v$ is $u_{lji} = \sum_{i=1}^{N_q} \tau_{lji}^{-1}(\gamma_j S_{A_0} \gamma_j ij)$. By the Cauchy-Schwarz inequality, $u_{lji}^2 \leq \tau_{lji}^{-2} \sum_{i=1}^{N_q} \|\gamma_i\|_2^2 \|S_{A_0}\|_2^2 \leq \tau_{lji}^{-2} N_q \sum_{k=1}^q \lambda^2_k d_k$. Then, $\|u\|^2 \leq d_k \tau_1^{-2} q^2 \hat{\sigma}_b^2 (\lambda_1 \theta_b^{-1})^2$.

Define $v_{A_0^c} = \sqrt{n}\theta_b - n^{-1/2}c_1 \tau_1^{-1} \sum_{i=1}^{N_q} \|\gamma_i\|_2 \hat{\lambda}_b^{-1}, \eta_{A_0^c} = n^{-1/2}c_1 \sum_{i=1}^{N_q} X_{A_0} \xi_{A_0} = n^{-1/2}X_{A_0} (I - P_{A_0}) \xi_{A_0}$, $C_{A_0} = \{\max_{k \in A_0^c} \|\eta_k\|_2 \geq v_{A_0^c}\}$, then $P(E_1^c) \leq P(C_{A_0})$. By Lemmas 1 and 2 of Huang, Ma and Zhang (2008), $P(C_{A_0}) \leq K(d_a \log n)^{1/2}/v_{A_0^c}$ where $K$ is a constant. $K(d_a \log n)^{1/2}/v_{A_0^c} \to 0$ from (C3). Then, $P(E_1^c \cap R) \to 0$, $P(E_1^c) \to 0$.

Next, we consider $P(E_2^c)$. Similarly as above, define $D = \{\|\hat{\beta}_k\|_2^{-1} > r_n, k \in A_0\} \cap R$. $P(E_2^c) \leq P(E_2^c \cap D) + P(D^c)$. By (C2), $P(D^c) \to 0$. $|\sum_{i=1}^{N_q} \sum_{j=1}^{N_q} (X_{A_0}\xi_{A_0})_{ij} x_{1i}| \leq \sum_{i=1}^{N_q} u_i/n \leq \tau_1^{-1} q^2 \sum_{i=1}^{N_q} \|\gamma_i\|_2 \hat{\lambda}_b^{-1}$, where $u_i$ is the $l$th element in the $l$th group of vector $\Sigma_{A_0^c}^{-1} \Sigma_{A_0^c}^{-1}$. Define $v_{A_0} = n^{-1/2} \hat{\lambda}_b n/\sqrt{\delta_0} - n^{-1/2} \tau_1^{-1} q^2 \hat{\sigma}_b^2 \hat{\gamma}_b \lambda_1 \theta_b^{-1}, C_{A_0} = \{\max_{k \in A_0} \|\eta_k\|_2 \geq v_{A_0}\}$, then $P(E_2^c) \leq P(C_{A_0})$. By Lemmas 1 and 2 of Huang, Ma and Zhang (2008), $P(C_{A_0}) \leq K(d_a \log (p - q))^{1/2}/v_{A_0}$. $K(d_a \log (p - q))^{1/2}/v_{A_0} \to 0$ from (C3). Then $P(E_2^c \cap D) \to 0$, $P(E_2^c) \to 0$. This completes the proof the Theorem 3.1. □

Proof of Theorem 3.2. Let $\hat{A} = \{k: \|\hat{\beta}_k\|_2 > 0, k = 1, \cdots, p\}$, then $\sum_{k \in \hat{A}} \|\hat{\beta}_k\|_2 = 0$, the dimension of problem 3.1 is reduced to $\hat{q}, \hat{q} \leq q^*$ and $\hat{A}_c \subset A_0$. By the definition of $\hat{\beta}$,

$$
\frac{1}{2} \|Y - X_{\hat{A}} \hat{\beta}_A\|_2^2 + \lambda \sum_{k \in \hat{A}} \sqrt{d_k} \|\hat{\beta}_k\|_2 \leq \frac{1}{2} \|Y - X_{\hat{A}} \hat{\beta}_A\|_2^2 + \lambda \sum_{k \in \hat{A}} \sqrt{d_k} \|\beta_k\|_2 \|\beta_k\|_2. \tag{6.13}
$$

$$
\eta^* = \lambda \sum_{k \in \hat{A}} \sqrt{d_k} \|\beta_k\|_2 - \|\hat{\beta}_k\|_2 \leq \lambda \sum_{k \in \hat{A}} \sqrt{d_k} \|\hat{\beta}_k - \beta_k\|_2. \tag{6.14}
$$

Let $\delta \hat{A} = \sum_{k \in \hat{A}} \left(\|\hat{\beta}_k\| - \|\beta_k\|_2\right), D = \sum_{k \in \hat{A}} \sqrt{d_k} \hat{\beta}_k - \|Y - X_{\hat{A}} \hat{\beta}_A\|_2 = \delta \hat{A} - \delta \hat{A} - 2(D \hat{\epsilon}) \hat{A}$. By (6.13) and (6.14), $\delta \hat{A} - \delta \hat{A} - 2(D \hat{\epsilon}) \hat{A} \leq 0$, then $\|\hat{\beta}_A - \beta_A\|_2 \leq \|D \hat{\epsilon}\|_2 + \|D \hat{\epsilon}\|_2$. Thus $\|\hat{\beta}_A - \beta_A\|_2 \leq 6\|D \hat{\epsilon}\|_2 + 6\eta^*$. Let $D_1$ be the $i$th column of $D$, $E(\|D_1\|_2^2) = \sigma^2 \|D D^T\| = \sigma^2 \|D_1\|_2^2$. Then with probability converging to one, $\|\hat{\beta}_A - \beta_A\|_2 \leq 6\sigma^2 M^2 \xi_0 (c_n) (\lambda \sqrt{\eta^*} (c_n (\xi_0 \xi_0 c_n))/n)^2 /2 + \|\hat{\beta}_A - \beta_A\|_2^2 /2$. Thus for $\lambda = n^a$ for some $0 < a < 1/2$, with probability converging to one,

$$
\|\hat{\beta}_A - \beta_A\|_2 \leq \sqrt{6\sigma^2 M^2 \xi_0 (c_n) (\lambda \sqrt{\eta}^* (c_n (\xi_0 \xi_0 c_n))/n)^2 /2} \sim O(\sqrt{\eta^*/n}),
$$

and $\|X_{\hat{A}} \hat{\beta}_A - X_{\hat{A}} \beta_A\|_2 \leq \sqrt{\eta^*} \|\hat{\beta}_A - \beta_A\|_2 \sim O(\sqrt{\eta^*/n})$. This completes the proof of Theorem 3.2. □

References


