Consistent Group Selection in High-Dimensional Linear Regression

Fengrong Wei
Department of Mathematics
University of Iowa
Iowa City, IA 52242
USA
Email: fengrong-wei@uiowa.edu

Jian Huang
Department of Statistics and Actuarial Science
University of Iowa
Iowa City, IA 52242
USA
Email: jian-huang@uiowa.edu

The University of Iowa
Department of Statistics and Actuarial Science
Technical Report No. 387

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 improved 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.

AMS 2000 subject classification: Primary 62J05, 62J07; secondary 62H25.

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

1 Introduction

Consider the linear regression model with p groups of covariates

$$Y_i = \sum_{k=1}^{p} X'_{ik}\beta_k + \varepsilon_i,\ i = 1, \cdots, n,$$

where Y_i is the response variable, ε_i is the error term, X_{ik} is a $d_k \times 1$ covariate vector representing the kth group and, β_k is the $d_k \times 1$ vector of corresponding regression coefficients of the kth group. For such a model, the group Lasso (Yuan and Lin 2005, 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 ℓ_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 of the group Lasso, 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 lasso, group Lars and group nonnegative garrote 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. They showed that the group Lasso is consistent under certain conditions and proposed a block coordinate descent algorithm that can handle high-dimensional data. 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); Meinshausen and Yu (2008); 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 (GSC) and the sparse Riesz condition introduced in Zhang and Huang (2008) 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 ℓ_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 (NSC) 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 that iteratively uses the group Lasso twice follows the idea of 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 $Y = (Y_1, \dots, Y_n)'$ and $X = (X_1, \cdots, X_p)$, where X_k is the $n \times d_k$ covariate submatrix corresponding to the kth group. For a given penalty level $\lambda \geq 0$, the group Lasso estimator of $\beta = (\beta_1', \dots, \beta_p')'$ is

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

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 $A_0 \in \{1, \cdots, p\}$ such that $\sum_{k \in A_0} \|\beta_k\|_2 \leq \eta_1$, where $\| \cdot \|_2$ denotes the ℓ_2 norm. Without loss of generality, let $A_0 = \{q+1, \cdots, p\}$. Then the GSC is

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

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

$$\|\beta_k\|_2 = 0, k = q+1, \dots, p. \tag{2.3}$$

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 η_1 is smaller than the noise level associated with model fitting.

The SRC controls the range of eigenvalues of sub-matrix consisted of a fixed number of design vectors x_j. For $A \subset \{1, \cdots, p\}$, define

$$X_A = (X_k, k \in A),\ \Sigma_{AA} = X_A' X_A / n.$$

Note that X_A is an $n \times \sum_{k \in A} d_k$ matrix. The design matrix X satisfies the sparse Riesz condition (SRC) with rank q^* and spectrum bounds $0 < c_* < c^* < \infty$ if

$$c_* \leq \frac{\|X_A \nu\|_2^2}{n\|\nu\|_2^2} \leq c^*, \forall A \text{ with } q^* = |A| = \#\{k : k \in A\} \text{ and } \nu \in R^{\sum_{k \in A} d_k}. \tag{2.4}$$

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

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

which determines the dimension of the selected model. If $\hat{q} = O(q_n)$, 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,

$$\tilde{\omega} = \|(I - \hat{P})X\beta\|_2 \tag{2.6}$$

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 $\tilde{\omega}^2$ 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

$$\zeta_2 = (\sum_{k\notin A_0} \|\beta_k\|_2^2 I\{\|\hat{\beta}_k\|_2 = 0\})^{1/2}. \tag{2.7}$$

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

$$r_1 \equiv r_1(\lambda) = \left(\frac{nc^*\sqrt{d_a}\eta_1}{\lambda d_b q}\right)^{1/2}, \; r_2 \equiv r_2(\lambda) = \left(\frac{nc^*\eta_2^2}{\lambda^2 d_b q}\right)^{1/2}, \; \bar{c} = \frac{c^*}{c_*} \tag{2.8}$$

where $\eta_2 \equiv \max_{A\subset A_0} \|\sum_{k\in A} X_k\beta_k\|_2$,

$$M_1 \equiv M_1(\lambda) = 2 + 4r_1^2 + 4\sqrt{d\bar{c}}\, r_2 + 4d\bar{c}, \tag{2.9}$$

$$M_2 \equiv M_2(\lambda) = \frac{2}{3}\Big(1 + 4r_1^2 + 2d\bar{c} + 4\sqrt{2d}(1+\sqrt{\bar{c}})\sqrt{\bar{c}}\, r_2 + \frac{16}{3}d\bar{c}^2\Big), \tag{2.10}$$

$$M_3 \equiv M_3(\lambda) = \frac{2}{3}\Big(1 + 4r_1^2 + 4\sqrt{d\bar{c}}(1 + 2\sqrt{1+\bar{c}}\,)r_2 + 3r_2^2 + \frac{2}{3}d\bar{c}(7+4\bar{c})\Big). \tag{2.11}$$

Let

$$\lambda_{n,p} = 2\sigma\sqrt{8(1+c_0)d_a d^2 q^* \bar{c} n c^* \log(N_d \vee a_n)},$$

where $c_0 \geq 0$, $a_n \geq 0$, satisfying $pd_a/(N_d \vee a_n)^{1+c_0} \approx 0$. We also consider the following constraints

$$\lambda_0 = \inf\{\lambda : M_1 q + 1 \leq q^*\}, \; \inf\emptyset = \infty, \tag{2.12}$$

$$\lambda \geq \max\{\lambda_0, \lambda_{n,p}\}. \tag{2.13}$$

For large p, the lower bound here is allowed to be $\lambda_{n,p} = 2\sigma[8(1+c_0)d_a d^2 q^* \bar{c} n c^* \log(N_d)]^{1/2}$ with $a_n = 0$; for fixed p, $a_n \to \infty$ is required.

We assume the following basic condition.

(C1) The errors $\varepsilon_1, \cdots, \varepsilon_n$ 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 $\hat{q}, \tilde{\omega}$ and ζ_2 be defined as in* (2.5), (2.6) *and* (2.7), *respectively, for the model $\hat{A}$ 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 constraints* (2.12), (2.13) *are satisfied, the following assertions hold with probability converging to 1,*

$$\begin{aligned}
&\hat{q} \leq \#\{k : \|\hat{\beta}_k\|_2 > 0 \;\; or \;\; k \notin A_0\} \leq M_1(\lambda)q, \\
&\tilde{\omega}^2 = \|(I - \hat{P})X\beta\|_2^2 \leq M_2(\lambda)B_1^2(\lambda), \\
&\zeta_2^2 = \sum_{k \notin A_0} \|\beta_k\|_2^2 I\{\|\hat{\beta}_k\|_2 = 0\} \leq \frac{M_3(\lambda)B_1^2(\lambda)}{c_* n},
\end{aligned}$$

where $B_1(\lambda) = ((\lambda^2 d_b^2 q)/(nc^*))^{1/2}$.

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*

$$r_1^2 q = \frac{nc^*\sqrt{d_a}\eta_1}{\lambda d_b}, \; r_2^2 q = \frac{nc^*\eta_2^2}{\lambda^2 d_b},$$

to recover M_1, M_2 and M_2 in (2.9), (2.10), (2.11), respectively, resulting

$$\hat{q} \leq \frac{4nc^*\sqrt{d_a}\eta_1}{\lambda d_b}, \; \tilde{\omega}^2 \leq \frac{8}{3}\lambda\sqrt{d_a}d_b\eta_1, \; \zeta_2^2 = 0.$$

Remark 2.2 *If $\eta_1 = 0$ in (2.2), then $r_1 = r_2 = 0$ and*

$$\begin{aligned}
M_1 &= 2 + 4d\bar{c}, \\
M_2 &= \frac{2}{3}\Big(1 + 2d\bar{c} + \frac{16}{3}d\bar{c}^2\Big), \\
M_3 &= \frac{2}{3}\Big(1 + \frac{2}{3}d\bar{c}(7 + 4\bar{c})\Big),
\end{aligned}$$

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 2.1 is a direct generalization of Theorem 1 on the selection properties of the Lasso obtained by Zhang and Huang (2008). In particular, when $d_1 = \cdots = d_p = 1$,*

$$\begin{aligned}
r_1 &= \left(\frac{nc^*\eta_1}{\lambda q}\right)^{1/2}, \; r_2 = \left(\frac{nc^*\eta_2^2}{\lambda^2 q}\right)^{1/2}, \\
M_1 &= 2 + 4r_1^2 + 4\sqrt{\bar{c}}\,r_2 + 4\bar{c}, \\
M_2 &= \frac{2}{3}\Big(1 + 4r_1^2 + 2\bar{c} + 4\sqrt{2}(1 + \sqrt{\bar{c}})\sqrt{\bar{c}}\,r_2 + \frac{16}{3}\bar{c}^2\Big), \\
M_3 &= \frac{2}{3}\Big(1 + 4r_1^2 + 4\sqrt{\bar{c}}(1 + 2\sqrt{1 + \bar{c}}\,)r_2 + 3r_2^2 + \frac{2}{3}\bar{c}(7 + 4\bar{c})\Big),
\end{aligned}$$

which 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*

$$\widehat{\beta^*} = \arg\min_{\beta} \frac{1}{2}(Y - X\beta)^T(Y - X\beta) + \lambda \sum_{k=1}^{p} (\beta_k' R_k \beta_k)^{1/2}, \tag{2.14}$$

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 $R_k = d_k Q_k^T Q_k$. Let $\beta^* = Q_k \beta$, and $X_k^* = X_k Q_k^{-1}$. Then (2.14) becomes

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

The generalized sparsity condition for (2.14) 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 $X \cdot Q^{-1} = (X_1 Q_1^{-1}, \cdots, X_p Q_p^{-1})$.

Immediately from Theorem 2.1, we have the following corollary.

Corollary 2.1 *Suppose that the conditions of Theorem 2.1 hold and λ satisfies the constraint (2.13). Then, with probability converging to one, all groups with $\|\beta_k\|_2^2 > M_3(\lambda) q \lambda^2/(c_* c^* n^2)$ 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 ℓ_2 norm are greater than the threshold 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 the following assertions hold with probability converging to 1,*

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

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 λ that satisfies (2.12) and (2.13). The following result is an immediate corollary of Theorem 2.2 .

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

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

This corollary follows by substituting the given λ 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 (2005) 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,

$$\frac{1}{2}(Y - X\beta)'(Y - X\beta) + \tilde{\lambda}\sum_{k=1}^{p} w_k \sqrt{d_k}\|\beta_k\|_2, \tag{3.1}$$

where w_k is the weight associated with the kth group. The $\lambda_k \equiv \tilde{\lambda} w_k$ can be regarded as the penalty level corresponding to the kth group. For different groups, the penalty level λ_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 ℓ_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

$$w_k = \frac{1}{\|\tilde{\beta}_k\|_2}, \ k = 1, \ldots, p. \tag{3.2}$$

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. Below, we first give a general result concerning the selection and estimation properties of the adaptive group Lasso.

Let $\theta_a = \max_{k\in A_0^c} \|\beta_k\|_2$, $\theta_b = \min_{k\in A_0^c} \|\beta_k\|_2$. 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 $\varepsilon > 0$, $P(\min_{k\in A_0^c} \|\tilde{\beta}_k\|_2 > \xi_b\theta_b) > 1 - \varepsilon$ 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)

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

(C4) All the eigenvalues of $\Sigma_{A_0^c A_0^c}$ are bounded away from zero and infinity.

Define

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

Theorem 3.1 *Suppose that (C1)-(C4) and the narrow-sense sparsity condition (2.3) are satisfied. Then*

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

Therefore, the adaptive group Lasso is selection consistent if an initial estimation consistent estimator is available and the conditions stated in Theorem 3.1 hold. 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 $\sqrt{n/(q\log(N_d))}$. If we use $\hat{\beta}$ as the initial estimator, (C3) can be changed to

(C3)*

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

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*

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

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 adaptive group Lasso is selection consistent.

Theorem 3.2 *Suppose that the conditions in Theorem 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*

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

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

$$\lambda_k = \begin{cases} \frac{\tilde{\lambda}}{\|\hat{\beta}_k\|_2}, & \text{if } \|\hat{\beta}_k\|_2 > 0, \\ \infty, & \text{if } \|\hat{\beta}_k\|_2 = 0. \end{cases}$$

If $\lambda_k = \infty$, then $\hat{\beta}_k^* = 0$. 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 $n = 200$.

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 $X_j = (X_{5(j-1)+1}, \cdots, X_{5(j-1)+5}), 1 \le j \le 10$. 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 with mean zero and $Cov(Z_{j_1}, Z_{j_2}) = 0.6^{|j_1 - j_2|}$. The covarites $X_1, \cdots, X_{50}$ are generated as

$$X_{5(j-1)+k} = \frac{Z_j + R_{5(j-1)+k}}{\sqrt{2}}, 1 \le j \le 10, 1 \le k \le 5,$$

The random error $\varepsilon \sim N(0, 3^2)$. The response variable Y is generated from $Y = \sum_{k=1}^{10} X_k' \beta_k + \varepsilon$, where

$$\begin{aligned} &\beta_1 = (0.5, 1, 1.5, 2, 2.5), \beta_2 = (2, 2, 2, 2, 2), \\ &\beta_3 = \cdots = \beta_{10} = (0, 0, 0, 0, 0). \end{aligned}$$

Example 2. In this example, the number of groups is $p = 10$. Each group consists of 5 covariates. The covaraites are generated the same way as in Example 1. However, the regression coefficients

$$\begin{aligned} &\beta_1 = (0.5, 1, 1.5, 1, 0.5), \beta_2 = (1, 1, 1, 1, 1), \\ &\beta_3 = (-1, 0, 1, 2, 1.5), \beta_4 = (-1.5, 1, 0.5, 0.5, 0.5), \\ &\beta_5 = \cdots = \beta_{10} = (0, 0, 0, 0, 0). \end{aligned}$$

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 covaraites are generated the same way as in Example 1. However, the regression coefficients

$$\begin{aligned} &\beta_1 = (0.5, 1, 1.5, 1, 0.5), \beta_2 = (1, 1, 1, 1, 1), \\ &\beta_3 = (-1, 0, 1, 2, 1.5), \beta_4 = (-1.5, 1, 0.5, 0.5, 0.5), \\ &\beta_5 = \cdots = \beta_{210} = (0, 0, 0, 0, 0). \end{aligned}$$

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 $X_j = (X_{5(j-1)+1}, \cdots, X_{5(j-1)+5}), 1 \le j \le 5$; and $X_j = (X_{3(j-6)+26}, \cdots, X_{3(j-6)+28}), 6 \le j \le 10$. 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 $Cov(Z_{j_1}, Z_{j_2}) = 0.6^{|j_1 - j_2|}$. The covarites $X_1, \cdots, X_{40}$ are generated as

$$X_{5(j-1)+k} = \frac{Z_j + R_{5(j-1)+k}}{\sqrt{2}}, 1 \le j \le 5, 1 \le k \le 5,$$

$$X_{3(j-6)+25+k} = \frac{Z_j + R_{3(j-6)+25+k}}{\sqrt{2}}, 6 \le j \le 10, 1 \le k \le 3,$$

The random error $\varepsilon \sim N(0, 3^2)$. The response variable Y is generated from $Y = \sum_{k=1}^{10} X_k\beta_k + \varepsilon$, where

$$\begin{aligned}
\beta_1 &= (0.5, 1, 1.5, 2, 2.5), \beta_2 = (2, 0, 0, 2, 2),\\
\beta_3 &= \cdots = \beta_5 = (0, 0, 0, 0, 0),\\
\beta_6 &= (-1, -2, -3),\\
\beta_7 &= \cdots = \beta_{10} = (0, 0, 0).
\end{aligned}$$

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

$$\begin{aligned}
\beta_1 &= (0.5, 1, 1.5, 2, 2.5), \beta_2 = (2, 2, 2, 2, 2),\\
\beta_3 &= (-1, 0, 1, 2, 3), \beta_4 = (-1.5, 2, 0, 0, 0),\\
\beta_5 &= (0, 0, 0, 0, 0),\\
\beta_6 &= (2, -2, 1), \beta_7 = (0, -3, 1.5),\\
\beta_8 &= (-1.5, 1.5, 2), \beta_9 = (-2, -2, -2),\\
\beta_{10} &= (0, 0, 0).
\end{aligned}$$

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

$$\begin{aligned}
\beta_1 &= (0.5, 1, 1.5, 2, 2.5), \beta_2 = (2, 2, 2, 2, 2),\\
\beta_3 &= (-1, 0, 1, 2, 3), \beta_4 = (-1.5, 2, 0, 0, 0),\\
\beta_5 &= \cdots = \beta_{100} = (0, 0, 0, 0, 0),\\
\beta_{101} &= (2, -2, 1), \beta_{102} = (0, -3, 1.5),\\
\beta_{103} &= (-1.5, 1.5, 2), \beta_{104} = (-2, -2, -2),\\
\beta_{105} &= \cdots = \beta_{210} = (0, 0, 0).
\end{aligned}$$

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 of groups selected with the 25% and 75% quantiles of the number of selected groups in parentheses, model error, percentage of occasion on which correct groups are included in the selected model and percentage of occasions on which the exactly correct groups are selected 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.

Table 1: Simulation study. Mean number of groups selected, median number of variables selected (med), model error (ME), percentage of occasions on which correct groups are included in the selected model (% incl) and percentage of occasions on which exactly correct groups are selected (% sel), averaged over 400 replications with estimated standard errors in parentheses, by the group Lasso and adaptive group Lasso, for Examples 1-6. The true numbers of groups are included in [ ] in the first column.

	group Lasso					adaptive group Lasso				
$\sigma = 3$	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)	(0.18)	(0.07)	(2,2)	(0.90)	(0)	(0.07)
Ex 2, [4]	4.11	4	8.52	99.5%	88.5%	4.00	4	8.10	99.5%	98%
	(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%	86.5%	3.94	4	8.19	93%	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)	(0.35)	(0)	(3,3)	(0.90)	(0)	(0)
Ex 5, [8]	8.88	9	7.68	100%	40%	8.03	8	7.58	100%	97.5%
	(0.81)	(8,10)	(0.94)	(0)	(0.49)	(0.16)	(8,8)	(0.86)	(0)	(0.16)
Ex 6, [8]	12.90	9	14.61	66.5%	7%	11.49	8	9.28	66.5%	47%
	(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 NIH grant R01CA120988 and NSF grants DMS 0706108 and 0805670.

6 Proofs

We first define some notations which will be used in proofs. Let

$$\{k : \|\hat{\beta}_k\|_2 > 0, k \leq p\} \subseteq A_1 \subseteq \{k : X_k^T(Y - X\hat{\beta}) = \frac{\lambda\sqrt{d_k}\hat{\beta}_k}{\|\hat{\beta}_k\|_2}\} \cup \{1, \cdots, q\}.$$

Set $A_2 = \{1, \cdots, p\} \setminus A_1$, $A_3 = A_1 \setminus A_0$, $A_4 = A_1 \cap A_0$, $A_5 = A_2 \setminus A_0$, $A_6 = A_2 \cap A_0$. Thus we have $A_1 = A_3 \cup A_4$, $A_3 \cap A_4 = \emptyset$, $A_2 = A_5 \cup A_6$, $A_5 \cap A_6 = \emptyset$. Let $|A_i| = \sum_{k \in A_i} d_k$ which means the number of columns in the sub-matrix X_{A_i} of covariate matrix X and $N(A_i)$ be the cardinality of set A_i, $i = 1, \cdots, 6$. Assume $q_1 = N(A_1)$.

Lemma 1 *Let $A_k \subset \{1, \cdots, p\}$, $X_{A_k} = (X_k, k \in A_k)$ and $\Sigma_{1k} = X_{A_1}^T X_{A_k}/n$. Then*

$$\frac{\|\nu\|_2^2}{c^*(|A_1|)} \leq \|\Sigma_{11}^{-1/2}\nu\|_2^2 \leq \frac{\|\nu\|_2^2}{c_*(|A_1|)}, \qquad \|\beta_{A_k}\|_1^2 \leq \frac{\|X_{A_k}\beta_{A_k}\|_2^2 N(A_k)}{nc_*(|A_k|)}$$

for all ν of proper dimension. Further more, if $A_k \cap A_1 = \emptyset$, then

$$\|\beta_{A_k}\|_2^2 + \|\Sigma_{11}^{-1}\Sigma_{1k}\beta_{A_k}\|_2^2 \leq \frac{\|(I - P_1)X_{A_k}\beta_{A_k}\|_2^2}{nc_*(|A_1 \cup A_k|)},$$

where P_1 is the projection to the span of $\{x_{kl}, l = 1, \cdots, d_k, k \in A_1\}$.

Lemma 1 is Lemma 1 in Zhang and Huang (2008).

Proof of Theorem 2.1. The basic idea used in the proof of Theorem 2.1 follows that in the proof of the rate consistency of the Lasso in Zhang and Huang (2008). However, there are many difference 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. Thus we write out the details of the proof.

The proof consists of three steps. Step 1 proves some inequalities related to q_1, $\tilde{\omega}$ and ζ_2. Step 2 translates the results of Step 1 into upper bounds for $\hat{q}$, $\tilde{\omega}$ and ζ_2. Step 3 completes the proof by showing the probability of the event in Step 2 converging to one.

Since $\hat{\beta}$ is a solution of (2.1), by the KKT condition,

$$\begin{cases} X_k'(Y - X\hat{\beta}) = \frac{\lambda\sqrt{d_k}\hat{\beta}_k}{\|\hat{\beta}_k\|_2}, & \forall \|\hat{\beta}_k\|_2 > 0, \\ -\lambda\sqrt{d_k} \leq X_k'(Y - X\hat{\beta}) \leq \lambda\sqrt{d_k}, & \forall \|\hat{\beta}_k\|_2 = 0, \end{cases} \tag{6.1}$$

where the second inequality is componentwise.

Define

$$s_k = \frac{X_k'(Y - X\hat{\beta})}{\lambda\sqrt{L_k}}.$$

From (6.1), we have

$$\|s_k\|_2 \begin{cases} = 1, & \text{if } \|\hat{\beta}_k\|_2 > 0, \\ \leq 1, & \text{if } \|\hat{\beta}_k\|_2 = 0, \end{cases}$$

and

$$X_{A_1}'(Y - X_{A_1}\hat{\beta}_{A_1}) = S_{A_1}, \tag{6.2}$$

$$-C_{A_2} \leq X_{A_2}'(Y - X_{A_1}\hat{\beta}_{A_1}) \leq C_{A_2}, \tag{6.3}$$

where $S_{A_i} = (S_{k_1}', \cdots, S_{k_{q_i}}')'$ is an $|A_i| \times 1$ vector, $S_{k_i} = \lambda\sqrt{d_{k_i}}s_{k_i}$; $C_{A_i} = (C_{k_1}', \cdots, C_{k_{q_i}}')'$ is an $|A_i| \times 1$ vector, $C_{k_i} = \lambda\sqrt{d_{k_i}}I(\|\hat{\beta}_{k_i}\|_2 = 0)e_{d_{k_i}\times 1}$, where all the elements of matrix $e_{d_{k_i}\times 1}$ is 1 and $k_i \in A_i$. From (6.2) and (6.3),

$$X_{A_1}'(X_{A_1}\beta_{A_1} + X_{A_2}\beta_{A_2} + \varepsilon - X_{A_1}\hat{\beta}_{A_1}) = S_{A_1},$$

namely

$$n\Sigma_{11}\beta_{A_1} + n\Sigma_{12}\beta_{A_2} + X_{A_1}'\varepsilon - n\Sigma_{11}\hat{\beta}_{A_1} = S_{A_1},$$

and

$$-C_{A_2} \leq n\Sigma_{21}(\beta_{A_1} - \hat{\beta}_{A_1}) + n\Sigma_{22}\beta_{A_2} + X_{A_2}^{'}\varepsilon \leq C_{A_2}.$$

Since all the eigenvalues of Σ_{11} are bounded below by $c_*(|A_1|)$, we assume without loss of generality that Σ_{11} is of full rank. By the definition of Σ,

$$\frac{\Sigma_{11}^{-1}}{n}S_{A_1} = (\beta_{A_1} - \hat{\beta}_{A_1}) + \Sigma_{11}^{-1}\Sigma_{12}\beta_{A_2} + \frac{\Sigma_{11}^{-1}}{n}X_{A_1}^{'}\varepsilon, \tag{6.4}$$

and

$$n\Sigma_{22}\beta_{A_2} - n\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\beta_{A_2} \leq C_{A_2} - X_{A_2}^{'}\varepsilon - \Sigma_{21}\Sigma_{11}^{-1}S_{A_1} + \Sigma_{21}\Sigma_{11}^{-1}X_{A_1}^{'}\varepsilon. \tag{6.5}$$

where $\Sigma_{ij} = X_{A_i}^{'}X_{A_j}/n$.

Step 1. Define

$$V_{1j} = \frac{1}{\sqrt{n}}\Sigma_{11}^{-1/2}Q_{A_j1}^{'}S_{A_j},\ j = 1, 3, 4, \tag{6.6}$$

and

$$\omega_k = (I - P_1)X_{A_k}\beta_{A_k}, k = 2, \cdots, 6.$$

where Q_{A_kj} is the matrix representing the selection of variables in A_k from A_j.

Consider $j = 4$. For any $k \in A_4$, by the definition of A_4, we have $\|\hat{\beta}_k\|_2 > 0$, then $\|s_k\|_2^2 = 1$, so $\|s_{A_4}\|_2^2 = \sum_{k\in A_4}\|s_k\|_2^2 = N(A_4)$. Since $q_1 = N(A_3) + N(A_4) \leq q + N(A_4)$, then $\|s_{A_4}\|_2^2 \geq q_1 - q$.

By the definition of V_{1j} in (6.6) and Lemma 1,

$$\|V_{14}\|_2^2 = \frac{1}{n}\|\Sigma_{11}^{-1/2}Q_{A_41}^{'}S_{A_4}\|_2^2 \geq \frac{\|Q_{A_41}^{'}S_{A_4}\|_2^2}{nc^*(|A_1|)} = \frac{\sum_{k\in A_4}\|\lambda\sqrt{d_k}s_k\|_2^2}{nc^*(|A_1|)} \geq \frac{\lambda^2 d_b(q_1 - q)}{nc^*(|A_1|)}.$$

That is,

$$\|V_{14}\|_2^2 \geq \frac{\lambda^2 d_b(q_1 - q)}{nc^*(|A_1|)}. \tag{6.7}$$

By (6.4),

$$\begin{aligned} V_{14}^{'}(V_{13} + V_{14}) &= S_{A_4}^{'}Q_{A_41}\frac{\Sigma_{11}^{-1}}{n}S_{A_1} \\ &= S_{A_4}^{'}Q_{A_41}\left[(\beta_{A_1} - \hat{\beta}_{A_1}) + \Sigma_{11}^{-1}\Sigma_{12}\beta_{A_2} + \frac{\Sigma_{11}^{-1}}{n}X_{A_1}^{T}\varepsilon\right] \\ &= S_{A_4}^{'}Q_{A_41}\Sigma_{11}^{-1}\Sigma_{12}\beta_{A_2} + \frac{S_{A_4}^{'}Q_{A_41}\Sigma_{11}^{-1}X_{A_1}^{T}\varepsilon}{n} + S_{A_4}^{'}(\beta_{A_4} - \hat{\beta}_{A_4}). \end{aligned}$$

For any $k \in A_4$, $A_4 = A_1 \cap A_0$, $s_k = \hat{\beta}_k/\|\hat{\beta}_k\|_2$, we have

$$S_{A_4}^{'}\beta_{A_4} = \sum_{k\in A_4}\frac{\lambda\sqrt{d_k}}{\|\hat{\beta}_k\|_2}\hat{\beta}_k^{'}\beta_k = \sum_{k\in A_4}\lambda\sqrt{d_k}\|\beta_k\|_2 \leq \lambda\sqrt{d_a}\sum_{k\in A_4}\|\beta_k\|_2.$$

Thus, since $\|\hat{\beta}_k\|_2 > 0$ for any $k \in A_4$, we have

$$(\lambda\sqrt{d_k}\frac{\hat{\beta}_k}{\|\hat{\beta}_k\|_2})^T\hat{\beta}_k = \lambda\sqrt{d_k}\|\hat{\beta}_k\|_2 > 0.$$

Thus $S^{'}_{A_4}\hat{\beta}_{A_4} > 0$. Therefore,

$$V^{'}_{14}(V_{13}+V_{14}) \le S^{'}_{A_4}Q_{A_4 1}\Sigma_{11}^{-1}\Sigma_{12}\beta_{A_2} + \frac{S^{'}_{A_4}Q_{A_4 1}\Sigma_{11}^{-1}X^{'}_{A_1}\varepsilon}{n} + \sqrt{d_a}\lambda \sum_{k\in A_4}\|\beta_k\|_2.$$

By (6.5),

$$\begin{aligned}&\|\omega_2\|_2^2 = (X_{A_2}\beta_{A_2})^{'}(I-P_1)X_{A_2}\beta_{A_2} = \beta^{'}_{A_2}(n\Sigma_{22}\beta_{A_2} - n\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\beta_{A_2})\\ &\le \beta^{'}_{A_2}(C_{A_2} - X^{'}_{A_2}\varepsilon - \Sigma_{21}\Sigma_{11}^{-1}S_{A_1} + \Sigma_{21}\Sigma_{11}^{-1}X^{'}_{A_1}\varepsilon)\\ &= -\omega^{'}_2\varepsilon - S^{'}_{A_1}\Sigma_{11}^{-1}\Sigma_{12}\beta_{A_2} + \beta^{'}_{A_2}C_{A_2}.\end{aligned}$$

It follows that

$$\begin{aligned}&V^{'}_{14}(V_{13}+V_{14}) + \|\omega_2\|_2^2\\ &\le (\frac{S^{'}_{A_4}Q_{A_4 1}\Sigma_{11}^{-1}X^{'}_{A_1}}{n} - \omega^{'}_2)\varepsilon - S^{'}_{A_3}Q_{A_3 1}\Sigma_{11}^{-1}\Sigma_{12}\beta_{A_2}\\ &+ \beta^{'}_{A_2}C_{A_2} + \sqrt{d_a}\lambda\sum_{k\in A_4}\|\beta_k\|_2.\end{aligned}$$

Define

$$u = \frac{X_{A_1}\Sigma_{11}^{-1}Q^{'}_{A_4 1}S_{A_4}/n - \omega_2}{\|X_{A_1}\Sigma_{11}^{-1}Q^{'}_{A_4 1}S_{A_4}/n - \omega_2\|_2}.$$

Since X_{A_1} and ω_2 are orthogonal,

$$\|X_{A_1}\Sigma_{11}^{-1}Q^{'}_{A_4 1}S_{A_4}/n - \omega_2\|_2^2 = \|V_{14}\|_2^2 + \|\omega_2\|_2^2.$$

Thus

$$\begin{aligned}&\|V_{14}\|_2^2 + \|\omega_2\|_2^2 + V^{'}_{14}V_{13} \le (\|V_{14}\|_2^2 + \|\omega_2\|_2^2)^{1/2}|u^{'}\varepsilon|\\ &+ \|V_{13}\|_2\|\Sigma_{11}^{-1/2}\Sigma_{12}\beta_{A_2}\|_2\sqrt{n} + \lambda\sqrt{d_a}\|\beta_{A_2}\|_2 + \sqrt{d_a}\lambda\sum_{k\in A_4}\|\beta_k\|_2.\end{aligned} \tag{6.8}$$

By the definition of V_{1j} in (6.6) and Lemma 1,

$$\|V_{13}\|_2^2 = \frac{1}{n}\|\Sigma_{11}^{-1/2}Q^{'}_{A_3 1}S_{A_3}\|_2^2 \le \frac{\|Q^{'}_{A_3 1}S_{A_3}\|_2^2}{nc_*(|A_1|)} = \frac{\sum_{k\in A_3}\|\lambda\sqrt{d_k}s_k\|_2^2}{nc_*(|A_1|)} \le \frac{\lambda^2 d_a N(A_3)}{nc_*(|A_1|)}.$$

That is,

$$\|V_{13}\|_2^2 \le \frac{\lambda^2 d_a N(A_3)}{nc_*(|A_1|)}.$$

Since $A_2 = A_5 \cup A_6$, $A_5 \cap A_6 = \emptyset$ under the generalized sparsity condition, we have $\|\beta_{A_2}\|_2 \le \|\beta_{A_5}\|_2 + \|\beta_{A_6}\|_2$.

By (6.8), we have

$$\begin{aligned}&\|V_{14}\|_2^2 + \|\omega_2\|_2^2\\ &\le (\|V_{14}\|_2^2 + \|\omega_2\|_2^2)^{1/2}u^{'}\varepsilon\\ &+ \|V_{14}\|_2\|V_{13}\|_2 + \|V_{13}\|_2\|P_1X_{A_2}\beta_{A_2}\|_2 + \sqrt{d_a}\lambda\sum_{k\in A_4\cup A_6}\|\beta_k\|_2 + \lambda\sqrt{d_a}\|\beta_{A_5}\|_2\\ &\le (\|V_{14}\|_2^2 + \|\omega_2\|_2^2)^{1/2}u^{'}\varepsilon + \frac{\lambda\sqrt{d_aN(A_3)}}{\sqrt{nc_*(|A_1|)}}\|V_{14}\|_2\\ &+ \|P_1X_{A_2}\beta_{A_2}\|_2\left(\frac{\lambda^2 d_a N(A_3)}{nc_*(|A_1|)}\right)^{1/2} + \sqrt{d_a}\lambda\eta_1 + \lambda\sqrt{d_a}\|\beta_{A_5}\|_2.\end{aligned} \tag{6.9}$$

Step 2. Define

$$B_1 = \left(\frac{\lambda^2 d_b q}{nc^*(|A_1|)}\right)^{1/2}, \ B_2 = \left(\frac{\lambda^2 d_b q}{nc_*(|A_0| \vee |A_1|)}\right)^{1/2}.$$

In this step, let's consider the event

$$|u^{'}\varepsilon|^2 \leq \frac{(|A_1| \vee d_b)\lambda^2 d_b}{4d_a nc^*(|A_1|)} = (|A_1| \vee d_b)\frac{B_1^2}{4qd_a}.$$

we will prove this event has probability going to 1 later in *Step* 3.

By (6.7),

$$\|V_{14}\|_2^2 \geq \frac{q_1 B_1^2}{q} - B_1^2.$$

This implies

$$|u^{'}\varepsilon|^2 \leq \frac{q_1 d_a B_1^2}{4qd_a} \leq \frac{1}{4}(\|V_{14}\|_2^2 + B_1^2),$$

and

$$\begin{aligned}(\|V_{14}\|_2^2 + \|\omega_2\|_2^2)^{1/2}|u^{'}\varepsilon| &\leq \frac{1}{4}(\|V_{14}\|_2^2 + \|\omega_2\|_2^2) + |u^{'}\varepsilon|^2 \\ &\leq \frac{1}{2}(\|V_{14}\|_2^2 + \frac{\|\omega_2\|_2^2 + B_1^2}{2}).\end{aligned}$$

By (6.9),

$$\begin{aligned}\|V_{14}\|_2^2 + \|\omega_2\|_2^2 \leq \frac{1}{2}\|V_{14}\|_2^2 + \frac{1}{4}\|\omega_2\|_2^2 + \frac{1}{4}B_1^2 + \sqrt{d_a}\lambda\eta_1 + \sqrt{d_a}\lambda\|\beta_{A_5}\|_2 \\ + \frac{\lambda\sqrt{d_a q}}{\sqrt{nc_*(|A_1|)}}\|V_{14}\|_2 + \|X_{A_2}\beta_{A_2}\|_2\left(\frac{\lambda^2 d_a q}{nc_*(|A_1|)}\right)^{1/2}.\end{aligned}$$

It follows that

$$\begin{aligned}\|V_{14}\|_2^2 + \frac{3}{2}\|\omega_2\|_2^2 &\leq \frac{B_1^2}{2} + 2\sqrt{d_a}\lambda\eta_1 + 2\sqrt{d_a}\lambda\|\beta_{A_5}\|_2 \\ &+ 2(\|V_{14}\|_2 + \|X_{A_2}\beta_{A_2}\|_2)\left(\frac{\lambda^2 d_a q}{nc_*(|A_1|)}\right)^{1/2}.\end{aligned} \tag{6.10}$$

Consider the set A_1 contains all large $\beta_k \neq 0$. We have $q_1 \geq q$ and

$$\{k : \|\hat{\beta}_k\|_2 > 0 \ or \ k \notin A_0\} \subseteq A_1 \subseteq \{k : X_k^{'}(Y - X\hat{\beta}) = \frac{\lambda\sqrt{d_k}\hat{\beta}_k}{\|\hat{\beta}_k\|_2} \ or \ k \notin A_0\}.$$

Thus $A_5 = A_2 \setminus A_0 = \emptyset$, $\|\beta_{A_5}\|_2 = 0$, $N(A_3) = q \leq q_1$ and $\|\Sigma_{11}^{-1/2}\Sigma_{12}\beta_{A_2}\|_2\sqrt{n} = \|P_1 X_{A_2}\beta_{A_2}\|_2 = \|P_1 X_{A_6}\beta_{A_6}\|_2 \leq \eta_2$.

By (6.10),

$$\|V_{14}\|_2^2 + \frac{3}{2}\|\omega_2\|_2^2 \leq \frac{B_1^2}{2} + 2\sqrt{d_a}\lambda\eta_1 + 2\sqrt{d}\eta_2 B_2 + 2\sqrt{d}B_2\|V_{14}\|_2.$$

If $x^2 \leq c + 2bx$, then $x^2 \leq 2c + 4b^2$. Therefore,

$$\|V_{14}\|_2^2 \leq B_1^2 + 4\sqrt{d_a}\lambda\eta_1 + 4\sqrt{d}\eta_2 B_2 + 4dB_2^2.$$

It follows from (6.7) that

$$\begin{aligned}(q_1 - q)^+ &\leq \frac{nc^*(|A_1|)\|V_{14}\|_2^2}{\lambda^2 d_b} = \frac{q\|V_{14}\|_2}{B_1^2} \\ &\leq q + \frac{nc^*(|A_1|)}{\lambda^2 d_b}\left(4\sqrt{d_a}\lambda\eta_1 + 4\sqrt{\frac{\lambda^2 d_a q}{nc_*(|A_1|)}}\eta_2 + \frac{4\lambda^2 d_a q}{nc_*(|A_1|)}\right).\end{aligned} \tag{6.11}$$

For general A_1, A_5 may be no longer empty, $N(A_3) + N(A_5) \leq q$. Then

$$\begin{aligned}&\|\beta_{A_5}\|_2\sqrt{d_a}\lambda + \left(\frac{\lambda^2 d_a q}{nc_*(|A_1|)}\right)^{1/2}\|P_1 X_{A_2}\beta_{A_2}\|_2 \\ &\leq \left(\frac{\lambda^2 d_a N(A_5)}{nc_*(|A_5|)}\right)^{1/2}\|X_{A_5}\beta_{A_5}\|_2 + \left(\frac{\lambda^2 d_a q}{nc_*(|A_1|)}\right)^{1/2}\|P_1 X_{A_2}\beta_{A_2}\|_2 \\ &\leq \left(\frac{2\lambda^2 d_a q}{nc_*(|A_1| \vee |A_0|)}\right)^{1/2}\max(\|P_1 X_{A_2}\beta_{A_2}\|_2, \|X_{A_5}\beta_{A_5}\|_2) \\ &\leq \left(\frac{2\lambda^2 d_a q}{nc_*(|A_1| \vee |A_0|)}\right)^{1/2}\left(\sqrt{nc^*(|A_5|)}\|\beta_{A_5}\|_2 + \|X_{A_6}\beta_{A_6}\|_2\right) \\ &\leq \left(\frac{2\lambda^2 d_a q}{nc_*(|A_1| \vee |A_0|)}\right)^{1/2}\left(\sqrt{C_5}\|\omega_2\|_2 + (1 + \sqrt{C_5})\eta_2\right),\end{aligned}$$

where $C_5 = c^*(|A_5|)/c_*(|A_1| \cup |A_5|)$.

By (6.10),

$$\begin{aligned}&\|V_{14}\|_2^2 + \frac{3}{2}\|\omega_2\|_2^2 \leq \frac{B_1^2}{2} + 2\|V_{14}\|_2\left(\frac{\lambda^2 d_a q}{nc_*(|A_1|)}\right)^{1/2} \\ &+ 2\left(\frac{2\lambda^2 d_a q}{nc_*(|A_1| \vee |A_0|)}\right)^{1/2}\left(\sqrt{C_5}\|\omega_2\|_2 + (1 + \sqrt{C_5})\eta_2\right) + 2\sqrt{d_a}\lambda\eta_1 \\ &\leq \frac{B_1^2}{2} + \|V_{14}\|_2^2 + dB_2^2 + \sqrt{8d}(1 + \sqrt{C_5})\eta_2 B_2 + 2\sqrt{d_a}\lambda\eta_1 + \sqrt{8dC_5}B_2\|\omega_2\|_2.\end{aligned}$$

If $x^2 \leq c + bx$, then $x^2 \leq 2c + b^2$. Thus

$$\|\omega_2\|_2^2 \leq \frac{4}{3}\left(\frac{B_1^2}{2} + dB_2^2 + \sqrt{d}(1 + \sqrt{C_5})\eta_2 B_2 + 2\sqrt{d_a}\eta_1\right) + \frac{32}{9}dC_5 B_2^2. \tag{6.12}$$

From Zhang and Huang (2008), we have $\|\omega_2\|_2^2 \geq (\|\beta_{A_5}\|_2(nc_{*,5})^{1/2} - \eta_2)^2$, $\|X_{A_2}\beta_{A_2}\|_2 \leq \eta_2 + \|X_{A_5}\beta_{A_5}\|_2 \leq \eta_2 + (nc^*(|A_5|))^{1/2}\|\beta_{A_5}\|_2$. It follows that

$$\begin{aligned}&\frac{3}{2}(\|\beta_{A_5}\|_2\sqrt{nc_{*,5}} - \eta_2)^2 \leq \frac{B_1^2}{2} + 2\sqrt{d_a}\lambda\eta_1 + 2\sqrt{d_a}\lambda\|\beta_{A_5}\|_2 + \\ &\frac{\lambda^2 d_a q}{nc_*(|A_1|)} + 2\left(\eta_2 + \|\beta_{A_5}\|_2\sqrt{nc^*(|A_5|)}\right)\left(\frac{\lambda^2 d_a q}{nc_*(|A_1|)}\right)^{1/2}.\end{aligned}$$

Since $N(A_3) + N(A_5) = q$, by the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\sqrt{d_a}\lambda\|\beta_{A_5}\|_2 + \sqrt{nc^*(|A_5|)}\|\beta_{A_5}\|_2 \left(\frac{\lambda^2 d_a q}{nc_*(|A_1|)}\right)^{1/2} \\
&\le \|\beta_{A_5}\|_2 \left(\lambda\sqrt{d_a} + \lambda\sqrt{\frac{c^*(|A_5|)d_a q}{c_*(|A_1|)}}\right) \\
&\le \|\beta_{A_5}\|_2\lambda\sqrt{d_a q}\left(1+\frac{c^*(|A_5|)}{c_*(|A_1|)}\right)^{1/2}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\|\beta_{A_5}\|_2^2 nc_{*,5} \le \frac{2}{3}[\frac{B_1^2}{2} + 2\sqrt{d_a}\lambda\eta_1 + 2\eta_2\left(\frac{\lambda^2 d_a q}{nc_*(|A_1|)}\right)^{1/2} \\
&+ 2\|\beta_{A_5}\|_2\lambda\sqrt{d_a q}\left(1+\frac{c^*(|A_5|)}{c_*(|A_1|)}\right)^{1/2} + \frac{\lambda^2 d_a q}{nc_*(|A_1|)}] + 2\eta_2\|\beta_{A_5}\|_2\sqrt{nc_{*,5}} - \eta_2^2 \\
&\le \frac{4}{3}[\frac{B_1^2}{4} + \sqrt{d_a}\lambda\eta_1 + \eta_2\left(\frac{\lambda^2 d_a q}{nc_*(|A_1|)}\right)^{1/2} + \frac{\lambda^2 d_a q}{2nc_*(|A_1|)} - \frac{3}{4}\eta_2^2] \\
&+ \|\beta_{A_5}\|_2\sqrt{nc_{*,5}}[\frac{4}{3}\lambda\sqrt{\frac{d_a q}{nc_{*,5}}}\left(1++\frac{c^*(|A_5|)}{c_*(|A_1|)}\right)^{1/2} + 2\eta_2],
\end{aligned}$$

where $c_{*,5} = c_*(|A_1 \cup A_5|)$.

Since $x^2 \le c + 2bx$ implies $x^2 \le 4b^2 + 2c$ for $b^2 + c \ge 0$, we have

$$\begin{aligned}
\|\beta_{A_5}\|_2^2 nc_{*,5} \le &\left[\frac{4}{3}\lambda\sqrt{\frac{d_a q}{nc_{*,5}}}\left(1+\frac{c^*(|A_5|)}{c_*(|A_1|)}\right)^{1/2} + 2\eta_2\right]^2 \\
&+ \frac{8}{3}\left[\frac{B_1^2}{4} + \sqrt{d_a}\lambda\eta_1 + \eta_2\left(\frac{\lambda^2 d_a q}{nc_*(|A_1|)}\right)^{1/2} + \frac{\lambda^2 d_a q}{2nc_*(|A_1|)} - \frac{3}{4}\eta_2^2\right].
\end{aligned} \tag{6.13}$$

Step 3. Consider $c_*(|A_m|) = c_*, c^*(|A_m|) = c^*$, for $N(A_m) \le q^*$ and,

$$q_1 \le N(A_1 \cup A_5) \le q^*,\ |u^{'}\varepsilon|^2 \le \frac{(|A_1| \vee d_b)\lambda^2 d_b}{4d_a nc^*(|A_1|)}. \tag{6.14}$$

Then we have $\bar{c} = C_5 = c^*(|A_5|)/c_*(|A_1| \vee |A_5|) = c^*/c_*$, $c_{*,5} = c_*(|A_1 \cup A_5|) = c_*$, $B_2^2 = \bar{c}B_1^2$, $r_1^2 = \sqrt{d_a}\lambda\eta_1/B_1^2$ and $r_2^2 = \eta_2^2/B_1^2$. From (6.11), (6.12) and (6.13), we have:

$$(q_1 - q)^+ + q \le (2 + 4r_1^2 + 4\sqrt{d\bar{c}}r_2 + 4d\bar{c})q, \tag{6.15}$$

$$\|\omega_2\|_2^2 \le \frac{1}{3}\left[2 + 4d\bar{c} + 4\sqrt{8d}(1+\sqrt{\bar{c}})\sqrt{\bar{c}}r_2 + 8r_1^2 + \frac{32}{3}d\bar{c}^2\right]B_1^2, \tag{6.16}$$

and

$$nc_*\|\beta_{A_5}\|_2^2 \le \frac{1}{3}\left[2 + 8r_1^2 + 8\sqrt{d\bar{c}}(1+2\sqrt{1+\bar{c}})r_2 + 4d\bar{c}\left(1+\frac{4}{3}(1+\bar{c})\right) + 6r_2^2\right]B_1^2. \tag{6.17}$$

By the definition of M_1, M_2, M_3, (6.15), (6.16) and (6.17) become

$$\begin{aligned} (q_1 - q)^+ + q &\leq M_1 q, \\ \|\omega_2\|_2^2 &\leq M_2 B_1^2, \\ nc_* \|\tilde{\gamma}_{A_5}\|_2^2 &\leq M_3 B_1^2. \end{aligned}$$

when (2.12) and (2.13) are satisfied.

Define

$$x_m^* \equiv \max_{|A|=m} \max_{\|U_{A_k}\|_2=1, k=1,\cdots,m} \left| \varepsilon' \frac{X_A (X_A' X_A)^{-1} \bar{S}_A - (I - P_A) X\beta}{\|X_A (X_A' X_A)^{-1} \bar{S}_A - (I - P_A) X\beta\|_2} \right|, \tag{6.18}$$

for $|A| = q_1 = m \geq 0$, $\bar{S}_A = (\bar{S}_{A_1}', \cdots, \bar{S}_{A_m}')'$ where $\bar{S}_{A_k} = \lambda\sqrt{d_{A_k}} U_{A_k}$, $\|U_{A_k}\|_2 = 1$. Let $Q_A = X_A^* (X_A' X_A)^{-1}$ where $X_k^* = \lambda\sqrt{d_k} X_k$ for $k \in A$. By (6.18),

$$x_m^* \equiv \max_{|A|=m} \max_{\|U_{A_k}\|_2=1, k=1,\cdots,m} \left| \varepsilon' \frac{Q_A U_A - (I - P_A) X\beta}{\|Q_A U_A - (I - P_A) X\beta\|_2} \right|. \tag{6.19}$$

For a given A, let $V_{lj} = (0, \cdots, 0, 1, 0, \cdots, 0)$ be the $|A| \times 1$ vector with the jth element in the lth group to be 1. Then

$$U_A = \sum_{l \in A} \sum_{j=1}^{d_l} \alpha_{lj} V_{lj},$$

and $\sum_{j=1}^{d_l} \alpha_{lj}^2 = 1$. We have

$$\begin{aligned} &|\varepsilon' (Q_A U_A - (I - P_A) X\beta)| = |\sum_{l \in A} \sum_{j=1}^{d_l} \alpha_{lj} \varepsilon' Q_A V_{lj} - \varepsilon' (I - P_A) X\beta| \\ &\leq \max_{l,j} |\varepsilon' Q_A V_{lj}| \sum_{l \in A} \sum_{j=1}^{d_l} |\alpha_{lj}| + |\varepsilon' (I - P_A) X\beta| \\ &\leq \max_{l,j} |\varepsilon' Q_A V_{lj}| \sum_{l \in A} \sqrt{d_l} + |\varepsilon' (I - P_A) X\beta|. \end{aligned}$$

By the definition of Q_A,

$$\|Q_A U_A - (I - P_A) X\beta\|_2^2 = (\|Q_A U_A\|_2^2 + \|(I - P_A) X\beta\|_2)^{1/2},$$

By the SRC,

$$\|Q_A U_A\|_2^2 = U_A^T Q_A^T Q_A U_A \geq \frac{\lambda^2 d_b}{nc^*(|A|)} U_A^T U_A = \frac{\lambda^2 d_b}{nc^*(|A|)} m.$$

Define

$$V = \frac{\max_{l,j} \|Q_A V_{lj}\|_2 \sum_{k \in A} \sqrt{d_k}}{\|Q_A U_A\|_2}.$$

By the definition of Q_A and SRC,

$$V \leq \frac{(\lambda\sqrt{d_a} \sum_{k \in A} \sqrt{d_k})/(nc_*(|A|))^{1/2}}{(\lambda\sqrt{d_b m})/(nc^*(|A|))^{1/2}} = \frac{d_a \sqrt{\bar{c} m}}{\sqrt{d_b}} \leq \frac{d_a \sqrt{\bar{c} M_1 q}}{\sqrt{d_b}}.$$

Thus by (6.19),

$$x_m^* \le \max_{|A|=m} \max_{l,j} \left\{ \left| \varepsilon^{'} \frac{Q_A V_{lj}}{\|Q_A V_{lj}\|_2} \right| \frac{\|Q_A V_{lj}\|_2 \sum_{l \in A} \sqrt{d_l}}{\|Q_A U_A\|_2} + \left| \frac{\varepsilon^{'}(I-P_A)X\beta}{\|(I-P_A)X\beta\|_2} \right| \right\}$$

Next, we show that (6.14) holds for the penalty level vector λ satisfying (2.12) and (2.13) with probability going to 1. Define

$$\Omega_{m_0} = \{(U,\varepsilon) : x_m^* \le \sigma\sqrt{8(1+c_0)V^2((md_b)\vee d_b)\log(N_d \vee a_n)}, \forall m \ge m_0\}.$$

By the definition of $\lambda_{n,p}$ and x_m^*, we have

$$(X,\varepsilon) \in \Omega_{m_0} \Rightarrow |u^{'}\varepsilon|^2 \le (x_m^*)^2 \le \frac{(|A_1| \vee d_b)\lambda^2 d_b}{4d_a n c^*}, \text{ for } N(A_1) \ge m_0 \ge 0.$$

Therefore (6.15) implies that for all $\lambda_k, k = 1, \cdots, p$ satisfying (2.12) and (2.13),

$$(X,\varepsilon) \in \Omega_q \Rightarrow q_1 \le M_1 q.$$

Since $\varepsilon_1, \cdots, \varepsilon_n$ are iid normal with $E\varepsilon_i = 0$ and $Var(\varepsilon_i) = \sigma^2$, 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 $\sigma^2 V_\varepsilon^2$, plus the maximum of $\binom{p}{m}$ normal variables with mean 0 and variance σ^2,

$$\begin{aligned}
&P\{(X,\varepsilon) \in \Omega_{m_0}\} \\
&\ge [1 - \sum_{m=0}^{\infty} \binom{p}{m} md_a \exp(-(md_b \vee d_b)(1+c_0)\log(N_d \vee a_n))] \cdot \\
&[1 - \sum_{m=0}^{\infty} \binom{p}{m} \exp(-(md_b \vee d_b)(1+c_0)V_\varepsilon \log(N_d \vee a_n))] \\
&\ge \left(1 - \sum_{m=0}^{\infty} \frac{p^m d_a^m}{m!} \frac{1}{(N_d \vee a_n)^{(1+c_0)(md_b \vee d_b)}}\right) \cdot \left(1 - \sum_{m=0}^{\infty} \frac{p^m}{m!} \frac{1}{(N_d \vee a_n)^{(1+c_0)(md_b \vee d_b)}}\right) \\
&\ge \left\{2 - \frac{(pd_a)^{d_b}}{(N_d \vee a_n)^{(1+c_0)d_b}} - \exp(\frac{pd_a}{(N_d \vee a_n)^{(1+c_0)}})\right\} \cdot \\
&\left\{2 - \frac{(p)^{d_b}}{(N_d \vee a_n)^{(1+c_0)d_b}} - \exp(\frac{p}{(N_d \vee a_n)^{(1+c_0)}})\right\} \\
&\to 1,
\end{aligned}$$

when (6.14) holds, since $pd_a/(N_d \vee a_n)^{(1+c_0)} \approx 0$. This complete the proof of Theorem 2.1. □

Proof of Theorem 2.2. For simplicity, we only consider the case when $\{c^*, c_*, r_1, r_2, c_0, \sigma, d\}$ are fixed and $\lambda/\sqrt{n} \ge 2\sigma\sqrt{8(1+c_0)d_a d^2 q^* \bar{c} c^* \log(N_d)} \to \infty$. In this case, M_1, M_2 and M_3 are fixed constant and the required configurations for GSC, SRC, (2.12) and (2.13) in Theorem 2.1 become

$$M_1 q + 1 < q^*,\ \eta_1 \le \frac{r_1^2}{c^*}\frac{q\lambda}{n},\ \eta_2^2 \le \frac{r_2^2}{c^*}\frac{q\lambda^2}{n}. \tag{6.20}$$

Consider $p \gg n > q^* > q \to \infty$, let $A_1 = \{k : \|\hat{\beta}_k\|_2 > 0 \ or\ k \notin A_0\}$. Define $v_1 = X_{A_1}(\hat{\beta}_{A_1} - \beta_{A_1})$ and $g_1 = X_{A_1}^T(Y - X\hat{\beta})$. By Lemma 1,

$$\|v_1\|_2^2 = \|X_{A_1}(\hat{\beta}_{A_1} - \beta_{A_1})\|_2^2 = n\|\Sigma_{11}^{1/2}(\hat{\beta}_{A_1} - \beta_{A_1})\|_2^2 \ge c_* n\|\hat{\beta}_{A_1} - \beta_{A_1}\|_2^2,$$

and

$$\begin{aligned}&(\hat{\beta}_{A_1}-\beta_{A_1})^{'}g_1 = v_1^{'}(Y - X_{A_1}\hat{\beta}_{A_1}) = v_1^{'}(X\beta + \varepsilon - X_{A_1}\hat{\beta}_{A_1} - X_{A_1}\beta_{A_1} + X_{A_1}\beta_{A_1})\\ &= v_1^{'}(X\beta - X_{A_1}\beta_{A_1} + \varepsilon) - \|v_1\|_2^2.\end{aligned}$$

By the Karush-Kuhn-Tucker conditions,

$$\|g_1\|_\infty \le \max_{k, \|\hat{\beta}_k\|_2 > 0} \|\frac{\lambda\sqrt{d_k}\hat{\beta}_k}{\|\hat{\beta}_k\|_2}\|_\infty = \lambda d_a,$$

and

$$\|X\beta - X_{A_1}\beta_{A_1}\|_2 = \|X_{A_2}\beta_{A_2}\|_2 \le \eta_2.$$

Therefore,

$$\begin{aligned}\|v_1\|_2 &\le \|X\beta - X_{A_1}\beta_{A_1} + P_1\varepsilon\|_2 + n^{-1/2}\|\Sigma_{11}^{-1/2}g_1\|_2\\ &\le \eta_2 + \|P_1\varepsilon\|_2 + \frac{\|g_1\|_2}{\sqrt{nc_*}} \le \eta_2 + \|P_1\varepsilon\|_2 + \lambda\sqrt{\frac{d_a N(A_1)}{nc_*}}.\end{aligned} \tag{6.21}$$

Since $\|P_1\varepsilon\|_2 \le 2\sigma\sqrt{N(A_1)\log(N_d)}$ with probability converging to one under the normality assumption, we have

$$\begin{aligned}&\|X(\hat{\beta}-\beta)\|_2 \le \|X_{A_1}(\hat{\beta}_{A_1}-\beta_{A_1})\|_2 + \|X_{A_2}\beta_{A_2}\|_2\\ &\le 2\eta_2 + \|P_1\varepsilon\|_2 + \lambda\sqrt{\frac{d_a N(A_1)}{nc_*}}\\ &\le 2r_2B_1 + 2\sigma\sqrt{N(A_1)\log(N_d)} + \lambda\sqrt{\frac{d_a q_1}{nc_*}}\\ &\le 2r_2B_1 + 2\sigma\sqrt{N(A_1)\log(N_d)} + \lambda\sqrt{\frac{d_a M_1 q}{nc_*}}\\ &\le (2r_2 + \sqrt{4M_1 d\bar{c}})B_1 + 2\sigma\sqrt{N(A_1)\log(N_d)}.\end{aligned}$$

It follows that

$$(\sum_{k\in A_1}\|\hat{\beta}_k - \beta_k\|_2^2)^{1/2} \le \frac{\|v_1\|_2}{\sqrt{nc_*}} \le \frac{1}{\sqrt{nc_*}}(\eta_2 + 2\sigma\sqrt{N(A_1)\log(N_d)} + \sqrt{dM_1\bar{c}}B_1). \tag{6.22}$$

Now since $A_2 \subset A_0$, by the second inequality in (6.20),

$$\#\{k \in A_0 : \|\beta_k\|_2 > \lambda/n\} \le \frac{r_1^2}{c^*}q \sim O(q).$$

Thus by the SRC and the third inequality in (6.20),

$$\sum_{k\in A_0}\|\beta_k\|_2^2 I\{\|\beta_k\|_2 > \lambda/n\} \le \frac{1}{nc_*}\sum_{k\in A_0}\|X_k\beta_k I\{\|\beta_k\|_2 > \lambda/n\}\|_2^2 \le \frac{\eta_2^2}{nc_*} \le \frac{r_2^2 q\lambda^2}{n^2 c_* c^*}, \tag{6.23}$$

and

$$\sum_{k\in A_0}\|\beta_k\|_2^2 I\{\|\beta_k\|_2 \le \lambda/n\} \le \frac{r_1^2 q}{c^*}\frac{\lambda^2}{n^2}. \tag{6.24}$$

By (6.22),(6.23) and (6.24), we have

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

and

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

This complete the proof of Theorem 2.2. □

Proof of Theorem 3.1. Let $\hat{u} = \hat{\beta} - \beta$ and $W = X^T\varepsilon/\sqrt{n}$ and let

$$V(u) = \sum_{i=1}^{n}[(\varepsilon_i - x_i u)^2 - \varepsilon_i^2)] + \sum_{k=1}^{p}\lambda_k\sqrt{d_k}\|u_k + \beta_k\|_2,$$

$$\hat{u} = \min_u(\varepsilon - Xu)^{'}(\varepsilon - Xu) + \sum_{k=1}^{p}\lambda_k\sqrt{d_k}\|u_k + \beta_k\|_2,$$

where $\lambda_k = \tilde{\lambda}/\|\tilde{\beta}_k\|_2$. By the KKT conditions,

$$\begin{aligned} X_k^{'}(Y - X\hat{\beta}) &= \lambda_k\sqrt{d_k}\frac{\hat{\beta}_k}{\|\hat{\beta}_k\|_2}, && \text{if } \|\hat{\beta}_k\|_2 \neq 0, \\ -\lambda_k\sqrt{d_k}e_{d_k\times 1} &\leq X_k^{'}(Y - X\hat{\beta}) \leq \lambda_k\sqrt{d_k}e_{d_k\times 1}, && \text{if } \|\hat{\beta}_k\|_2 = 0. \end{aligned}$$

If there exists $\hat{u}$ such that

$$\Sigma_{A_0^c A_0^c}(\sqrt{n}\hat{u}_{A_0^c}) - W_{A_0^c} = -\frac{1}{\sqrt{n}}S_{A_0^c}, \tag{6.25}$$

$$\|\hat{u}_k\|_2 \leq \|\beta_k\|_2, \text{for } k \in A_0^c,$$

and

$$-\frac{C_{A_0}}{\sqrt{n}} \leq \Sigma_{A_0 A_0^c}(\sqrt{n}\hat{u}_{A_0^c}) - W_{A_0} \leq \frac{C_{A_0}}{\sqrt{n}}, \tag{6.26}$$

then we have $\|\hat{\beta}_k\|_2 \neq 0$, for $k = 1, \cdots, q$, and $\|\hat{\beta}_k\|_2 = 0$, for $k = q+1, \cdots, p$. By (6.25),

$$(\sqrt{n}\hat{u}_{A_0^c}) - \Sigma_{A_0^c A_0^c}^{-1}W_{A_0^c} = -\frac{1}{\sqrt{n}}\Sigma_{A_0^c A_0^c}^{-1}S_{A_0^c}.$$

By (6.26),

$$\begin{aligned} \Sigma_{A_0 A_0^c}(\sqrt{n}\hat{u}_{A_0^c}) - W_{A_0} &= \Sigma_{A_0 A_0^c}\Sigma_{A_0^c A_0^c}^{-1}W_{A_0^c} - W_{A_0} - \frac{1}{\sqrt{n}}\Sigma_{A_0 A_0^c}\Sigma_{A_0^c A_0^c}^{-1}S_{A_0^c} \\ &= -n^{-1/2}X_{A_0}^{'}(I - P_{A_0^c})\varepsilon - n^{-1/2}\Sigma_{A_0 A_0^c}\Sigma_{A_0^c A_0^c}^{-1}S_{A_0^c}. \end{aligned}$$

Define the events

$$E_1 = \{n^{-1/2}\|(\Sigma_{A_0^c A_0^c}^{-1}X_{A_0^c}^{'}\varepsilon)_k\|_2 < \sqrt{n}\|\beta_k\|_2 - n^{-1/2}\|(\Sigma_{A_0^c A_0^c}^{-1}S_{A_0^c})_k\|_2,\ k \in A_0^c\},$$

and

$$E_2 = \{n^{-/2}\|(X_{A_0}^{'}(I - P_{A_0^c})\varepsilon)_k\|_2 < n^{-1/2}\|C_k\|_2 - n^{-1/2}\|(\Sigma_{A_0 A_0^c}\Sigma_{A_0^c A_0^c}^{-1}S_{A_0^c})_k\|_2,\ k \in A_0\}.$$

where $(\cdot)_k$ means the d_k dimensional sub-vector of vector $(\cdot)$ corresponding to the kth group for $k \in A_0$ or $k \in A_0^c$. Then we have, $P(\|\hat{\beta}_k\|_2 \neq 0, k \in A_0$, and $\|\hat{\beta}_k\|_2 = 0, k \notin A_0) \geq P(E_1 \cap E_2)$.

Since $P(E_1 \cap E_2) = 1 - P(E_1^c \cup E_2^c) \geq 1 - P(E_1^c) - P(E_2^c)$, to show $P(\|\hat{\beta}_k\|_2 \neq 0, k \in A_0$ and $\|\hat{\beta}_k\|_2 = 0, k \notin A_0) \to 1$, it suffices to show $P(E_1^c) \to 0$ and $P(E_2^c) \to 0$.

First, we consider $P(E_1^c)$. Define $R = \{\|\tilde{\beta}_k\|_2^{-1} \leq c_1\theta_b^{-1}, k \in A_0^c\}$ where c_1 is a constant. Then

$$P(E_1^c) = P(E_1^c \bigcap R) + P(E_1^c \bigcap R^c) \leq P(E_1^c \bigcap R) + P(R^c).$$

By condition $(C2)$, $P(R^c) \to 0$. So it suffices to show that $P(E_1^c \bigcap R) \to 0$.

Let $N_q = \sum_{k=1}^q d_k$, let $\tau_1 \leq \cdots \leq \tau_{N_q}$ be the eigenvalues of $\Sigma_{A_0^c A_0^c}$ and let $\gamma_1, \cdots, \gamma_{N_q}$ be the associated eigenvectors. We have $\Sigma_{A_0^c A_0^c}^{-1} = \sum_{l=1}^{N_q} \tau_l^{-1}\gamma_l\gamma_l^T$. The jth element in the lth group of vector $\Sigma_{A_0^c A_0^c}^{-1} S_{A_0^c}$ is

$$u_{lj} = \sum_{l'=1}^{N_q} \tau_{l'}^{-1}(\gamma_{l'}' S_{A_0^c})\gamma_{lj}.$$

By the Cauchy-Schwartz inequality,

$$u_{lj}^2 \leq \tau_1^{-2} \sum_{l=1}^{N_q} \|\gamma_l\|_2^2 \|S_{A_0^c}\|_2^2 = \tau_1^{-2} N_q \|S_{A_0^c}\|_2^2 \leq \tau_1^{-2} N_q (\sum_{k=1}^q \lambda_k^2 d_k). \tag{6.27}$$

Therefore,

$$\|u_k\|_2^2 \leq d_k \tau_1^{-2} q^2 d_a^2 (\tilde{\lambda} c_1 \theta_b^{-1})^2.$$

Define $\upsilon_{A_0^c} = \sqrt{n}\theta_b - n^{-1/2}c_1\tau_1^{-1}qd_a^{3/2}\tilde{\lambda}\theta_b^{-1}$, $\eta_{A_0^c} = n^{-1/2}\Sigma_{A_0^c A_0^c}^{-1} X_{A_0^c}^T \varepsilon$, $\xi_{A_0} = n^{-1/2}X_{A_0}^T(I - P_{A_0^c})\varepsilon$ and

$$C_{A_0^c} = \{\max_{k \in A_0^c} \|\eta_k\|_2 \geq \upsilon_{A_0^c}\}.$$

Then $P(E_1^c) \leq P(C_{A_0^c})$. By Lemmas 1 and 2 of Huang, Ma and Zhang (2008), we have,

$$P(C_{A_0}^c) \leq \frac{K(d_a \log q)^{1/2}}{\upsilon_{A_0^c}}$$

where K is a constant. Under condition $(C3)$,

$$\frac{k(d_a \log q)^{1/2}}{\upsilon_{A_0^c}} = \frac{K(d_a \log q)^{1/2}}{\sqrt{n}(\theta_b - (\tilde{\lambda} c_1 d_a^{3/2} q)/(\theta_b n))} \to 0$$

Namely, $P(E_1^c \bigcap R) \to 0$. Thus $P(E_1^c) \to 0$.

Next, we consider $P(E_2^c)$. Similarly as above, define $D = \{\|\tilde{\beta}_k\|_2^{-1} > r_n, k \in A_0\} \bigcap R$. Then

$$P(E_2^c) = P(E_2^c \bigcap D) + P(E_2^c \bigcap D^c) \leq P(E_2^c \bigcap D) + P(D^c).$$

By condition $(C2)$, we know $P(D^c) \to 0$. Thus it suffices to show that $P(E_2^c \bigcap D) \to 0$. By (6.27),

$$|\frac{1}{n} \sum_{l=1}^{N_q} \sum_{i=1}^{n} (X_{A_0})_{ij} (X_{A_0^c})_{il} u_l| \leq \sum_{l=1}^{N_q} |u_l| \leq \tau_1^{-1} q^2 d_a^2 \tilde{\lambda} c_1 \theta_b^{-1},$$

where u_l is the lth element of vector $\Sigma_{A_0^c A_0^c}^{-1} S_{A_0^c}$.

Define $\upsilon_{A_0} = n^{-1/2}\tilde{\lambda} r_n \sqrt{d_b} - n^{-1/2}\tau_1^{-1} q^2 d_a^{5/2} \tilde{\lambda} c_1 \theta_b^{-1}$,

$$C_{A_0} = \{\max_{k \in A_0} \|\xi_k\|_2 > \upsilon_{A_0}\}.$$

Then $P(Q^c) \leq P(C_{A_0})$. By Lemmas 1 and 2 of Huang et al. (2008), we have,

$$P(C_{A_0}) \leq \frac{K(d_a \log(p-q))^{1/2}}{\upsilon_{A_0}}$$

Under $(C3)$,

$$\frac{K(d_a \log(p-q))^{1/2}}{\upsilon_{A_0}} = \frac{K\sqrt{n d_a}(\log(p-q))^{1/2}}{\tilde{\lambda} r_n (\sqrt{d_b} - (d_a^{5/2} q^2 c_1)/(\theta_b r_n \tau_1))} \to 0.$$

Namely $P(E_2^c \bigcap D) \to 0$. Thus $P(E_2^c) \to 0$. This completes the proof the Theorem 3.1. □

Proof of Theorem 3.2. Let $\hat{A} = \{k : \|\tilde{\beta}_k\|_2 > 0, k = 1, \cdots, p\}$, $\hat{q} = \#\{k : k \in \hat{A}\}$. By the definition of λ_k, when $\|\tilde{\beta}_k\|_2 = 0$, we have $\|\hat{\beta}_k^*\|_2 = 0$. So $\sum_{k \in \hat{A}^c} \|\hat{\beta}_k^*\|_2 = 0$.

Given the initial estimator from the group Lasso, the dimension of our problem 3.1 is reduced to $\hat{q}$ and $\hat{q} \leq M_1 q \leq q^*$ and $\hat{A}_c \subset A_0$ by Theorem 2.1 with probability converging to one. By the definition of $\hat{\beta}^*$,

$$\frac{1}{2}\|Y - X_{\hat{A}}\hat{\beta}^*_{\hat{A}}\|_2^2 + \tilde{\lambda}\sum_{k \in \hat{A}} \frac{\sqrt{d_k}}{\|\tilde{\beta}_k\|_2}\|\hat{\beta}^*_k\|_2 \leq \frac{1}{2}\|Y - X_{\hat{A}}\beta_{\hat{A}}\|_2^2 + \tilde{\lambda}\sum_{k \in \hat{A}} \frac{\sqrt{d_k}}{\|\tilde{\beta}_k\|_2}\|\beta_k\|_2. \qquad (6.28)$$

$$\eta^* = \tilde{\lambda}\sum_{k \in \hat{A}} \frac{\sqrt{d_k}}{\|\tilde{\beta}_k\|_2}(\|\beta_k\|_2 - \|\hat{\beta}^*_k\|_2) \leq \tilde{\lambda}\sum_{k \in \hat{A}} \frac{\sqrt{d_k}}{\|\tilde{\beta}_k\|_2}\|\hat{\beta}^*_k - \beta_k\|_2. \qquad (6.29)$$

Let $\delta_{\hat{A}} = \Sigma^{1/2}_{\hat{A}\hat{A}}(\hat{\beta}^*_{\hat{A}} - \beta_{\hat{A}})$, $D = \Sigma^{-1/2}_{\hat{A}\hat{A}} X'_{\hat{A}}$. We have

$$\begin{aligned}
&\frac{1}{2}\|Y - X_{\hat{A}}\hat{\beta}^*_{\hat{A}}\|_2^2 - \frac{1}{2}\|Y - X_{\hat{A}}\beta_{\hat{A}}\|_2^2 \\
&= \frac{1}{2}\|X_{\hat{A}}(\hat{\beta}^*_{\hat{A}} - \beta_{\hat{A}})\|_2^2 + \varepsilon' X_{\hat{A}}(\beta_{\hat{A}} - \hat{\beta}_{\hat{A}}) \\
&= \frac{1}{2}\delta'_{\hat{A}}\delta_{\hat{A}} - (D\varepsilon)'\delta_{\hat{A}}.
\end{aligned}$$

By (6.28) and (6.29),

$$\frac{1}{2}\delta'_{\hat{A}}\delta_{\hat{A}} - (D\varepsilon)'\delta_{\hat{A}} - \eta^* \leq 0.$$

Therefore,

$$\|\delta_{\hat{A}} - D\varepsilon\|_2^2 - \|D\varepsilon\|_2^2 - 2\eta^* \leq 0.$$

It follows that $\|\delta_{\hat{A}} - D\varepsilon\|_2 < \|D\varepsilon\|_2 + (2\eta^*)^{1/2}$. By the triangle inequality,

$$\|\delta_{\hat{A}}\|_2 \leq \|\delta_{\hat{A}} - D\varepsilon\|_2 + \|D\varepsilon\|_2 \leq \|D\varepsilon\|_2 + (2\eta^*)^{1/2}.$$

Thus

$$\|\delta_{\hat{A}}\|_2^2 \leq 6\|D\varepsilon\|_2^2 + 6\eta^*.$$

Let D_i be the ith column of D. Then $D\varepsilon = \sum_{i=1}^n D_i\varepsilon_i$. Since $\varepsilon_i, i = 1, \cdots, p$ are iid $N(0, \sigma^2)$,

$$E(\|D\varepsilon\|_2^2) = \sum_{i=1}^n \|D_i\|_2^2 E\varepsilon_i^2 = \sigma^2 tr(D'D) = \sigma^2\hat{q}.$$

Then by the SRC and consistent of the group Lasso estimator, with probability converging to one,

$$nc_*\|\hat{\beta}_{\hat{A}} - \beta_{\hat{A}}\|_2^2 \leq 6\sigma^2 M_1 q + \tilde{\lambda}\frac{\sqrt{d_a}}{\xi_b\theta_b}\|\hat{\beta}_{\hat{A}} - \beta_{\hat{A}}\|_2,$$

namely

$$\|\hat{\beta}_{\hat{A}} - \beta_{\hat{A}}\|_2^2 \leq \frac{6\sigma^2 M_1 q}{nc_*} + \left(\tilde{\lambda}\frac{\sqrt{d_a}}{\xi_b\theta_b nc_*}\right)^2/2 + \|\hat{\beta}_{\hat{A}} - \beta_{\hat{A}}\|_2^2/2.$$

Thus for $\tilde{\lambda} = n^\alpha$ for some $0 < \alpha < 1/2$, with probability converging to one,

$$\|\hat{\beta}_{\hat{A}} - \beta_{\hat{A}}\|_2 \leq \sqrt{\frac{6\sigma^2 M_1}{c_*}\frac{q}{n} + \frac{d_a}{(\xi_b\theta_b c_*)^2}(\frac{\tilde{\lambda}}{n})^2} \sim O(\sqrt{q/n}),$$

and $\|X_{\hat{A}}\hat{\beta}_{\hat{A}} - X_{\hat{A}}\beta_{\hat{A}}\|_2 \leq \sqrt{nc^*}\|\hat{\beta}_{\hat{A}} - \beta_{\hat{A}}\|_2$, then $\|X_{\hat{A}}\hat{\beta}_{\hat{A}} - X_{\hat{A}}\beta_{\hat{A}}\|_2 \sim O(\sqrt{q})$. This completes the proof of Theorem 3.2. □

References

[1] Antoniadis, A. and Fan, J. (2001). Regularization of wavelet approximation (with Discussion). *J. Am. Statist. Assoc.* **96** 939-967.

[2] Bühlmann, P. and Meier, L. (2008). Discussion of "One-step sparse estimates in nonconcave penalized likelihood models" by H. Zou and R. Li. To appear in the *Ann. Statist.* **36** 1534-1541.

[3] Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n (with Discussion). *Ann. Statist.* **35**, 2313-2351.

[4] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression (with Discussion). *Ann. Statist.* **32** 407-499.

[5] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96** 1348-1360.

[6] Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32** 928-961.

[7] Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10** 971-988.

[8] Huang, J., Horowitz, J. L. and Ma, S. G. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.* **36** 587-613.

[9] Huang, J., Ma, S., Xie, H.L. and Zhang, C.H. (2007). A group bridge approach for variable selection. Technical Report No.376. The University of Iowa. Accepted for publication by *Biometrika*.

[10] Huang, J., Ma, S. and Zhang, C.H. (2006). Adaptive lasso for sparse high-dimensional regression models. Technical Report No. 374. The University of Iowa. To appear in *Statist. Sinica.*

[11] Knight, K. and Fu, W. J. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28** 1356-1378.

[12] Kim, Y., Kim, J. and Kim, Y. (2006). The blockwise sparse regression. *Statist. Sinica* **16** 375-390.

[13] Meier, L., van de Geer, S. and Bühlmann, P. (2008). Group Lasso for logisitc regression. *J. R. Statist. Soc. Ser. B* **70** 53-71.

[14] Meinshausen, N. and Buhlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34** 1436-1462.

[15] Meinshausen, N. and Yu, B. (2006). Lasso-type recovery of sparse representations for high-dimensional data. Technical report, Department of Statistics, University of California, Berkeley. To appear in *Ann. Statist.*

[16] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461-464.

[17] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Statist. Soc. Ser. B* **58** 267-288.

[18] van de Geer, S. (2006). High-dimensional generalized linear models and the Lasso. *Ann. Statist.* **36** 614-645.

[19] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. Ser. B* **68** 49-67.

[20] Zhang, C.-H. (2007). Penalized linear unbiased selection. Technical report No. 2007-003. Department of Statistics, Rutgers University. www.stat.rutgers.edu/resources/technical_reports07.html.

[21] Zhang, C.-H. and Huang, J. (2008). Model-selection consistency of the LASSO in high-dimensional linear regression. *Ann. Statist.* **36** 1567-1594.

[22] Zhao, P., Rocha, G. and Yu, B. (2008). Grouped and hierarchical model selection through composite absolute penalties. To appear in the *Ann. Statist.*

[23] Zhao, P. and Yu, B. (2006). On model selection consistency of LASSO. *Journal of Machine Learning Research.* **7** 2541 - 2563

[24] Zou, H. (2006). The adaptive Lasso and its oracle properties. *J. Am. Statist. Assoc.* **101** 1418-1429.

[25] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B* **67** 301-320.

[26] Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with Discussion). *Ann. Statist.* **36** 1509-1533.