Bernoulli **16**(4), 2010, 1369–1384
DOI: 10.3150/10-BEJ252

Consistent group selection in high-dimensional linear regression

FENGRONG WEI[1] and JIAN HUANG[2]

[1]*Department of Mathematics, University of West Georgia, 1601 Maple Street, Carrollton, GA 30118, USA. E-mail: fwei@westga.edu*
[2]*Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA 52242, USA. E-mail: jian-huang@uiowa.edu*

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 also tends to 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

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

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 corresponding $d_k \times 1$ vector of regression coefficients. For such a model, the group Lasso (Antoniadis and Fan (2001), Yuan and Lin (2006)) 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, 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 problems and applications. 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) have proposed the 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. Huang, Ma, Xie and Zhang (2008) proposed a group bridge method that can be used for simultaneous group and individual variable selection.

There has 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 of 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 concerning 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 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 (see page 1371) 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. We then 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 out 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)) and 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 convergent rate of the group Lasso estimator. In Section 3, we

describe the selection and estimation consistency results concerning 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 $X = (X_1, \ldots, X_p)$, 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 $\beta = (\beta_1', \ldots, \beta_p')'$ is

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

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

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 $A_0 \subset \{1, \ldots, p\}$ 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 $A_0 = \{q + 1, \ldots, p\}$. The GSC is then

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

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

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

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 the submatrix. For $A \subset \{1, \ldots, p\}$, we define $X_A = (X_k, k \in A)$ and $\Sigma_{AA} = X_A' X_A / n$. Note that X_A is an $n \times \sum_{k \in A} d_k$ matrix. The design matrix X_A satisfies the sparse Riesz condition (SRC) with rank q^* and spectrum bounds $0 < c_* < c^* < \infty$ if

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

Let $\hat{A} = \{k\colon\ \|\hat{\beta}_k\|_2 > 0, 1 \leq k \leq p\}$, 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

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

which determines the dimension of the selected model. If $\hat{q} = \mathrm{O}(q)$, then the selected model has dimension comparable to 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 accounted for by the selected model. To measure the important groups missing in the selected model, we define

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

We now describe several quantities that will be useful in describing the main results. Let $d_a = \max_{1\le k\le p} d_k$, $d_b = \min_{1\le k\le p} d_k$, $d = d_a/d_b$ and $N_d = \sum_{k=1}^p d_k$. Define

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

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

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

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

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

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 $c_0 \ge 0$ and $a_n \ge 0$, satisfying $pd_a/(N_d \vee a_n)^{1+c_0} \approx 0$, and $\lambda_0 = \inf\{\lambda\colon\ M_1 q + 1 \le q^*\}$, where $\inf\varnothing = \infty$. We also consider the constraint

$$\lambda \ge \max\{\lambda_0, \lambda_{n,p}\}. \tag{2.12}$$

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

We assume the following basic condition.

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

Theorem 2.1. *Suppose that $q \ge 1$ 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. If the constraint* (2.12) *is satisfied, then the following assertions hold with probability*

converging to 1:

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

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

Remark 2.1. The condition $q \ge 1$ 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 $r_1^2 q = nc^*\sqrt{d_a}\eta_1/(\lambda d_b)$ and $r_2^2 q = nc^*\eta_2^2/(\lambda^2 d_b)$ to recover M_1, M_2 and M_3 in (2.9), (2.10), (2.11), respectively, giving the results $\hat{q} \le 4nc^*\sqrt{d_a}\eta_1/\lambda d_b$, $\tilde{\omega}^2 \le 8\lambda\sqrt{d_a}d_b\eta_1/3$ and $\zeta_2^2 = 0$.

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

$$M_1 = 2 + 4d\bar{c}, \qquad M_2 = \tfrac{2}{3}\big(1 + 2d\bar{c} + \tfrac{16}{3}d\bar{c}^2\big), \qquad M_3 = \tfrac{2}{3}\big(1 + \tfrac{2}{3}d\bar{c}(7 + 4\bar{c})\big),$$

all of which depend only on d and $\bar{c}$. This suggests that the relative sizes of the groups affect the selection results. Since $d \ge 1$, 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 $d_1 = \cdots = d_p = 1$, 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

$$\hat{\beta}^* = \arg\min_{\beta} \frac{1}{2}(Y - X\beta)'(Y - X\beta) + \lambda \sum_{k=1}^{p} (\beta_k' R_k \beta_k)^{1/2}, \tag{2.13}$$

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 $R_k = d_k Q_k' Q_k$. Let $\beta^* = Q_k\beta$, and $X_k^* = X_k Q_k^{-1}$. 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 $\sum_{k=q+1}^{p} (\beta_k' Q_k' Q_k \beta_k)^{1/2} \le \eta_1$. The SRC can be assumed for $X \cdot Q^{-1}$, where $X \cdot Q^{-1} = (X_1 Q_1^{-1}, \ldots, X_p Q_p^{-1})$.

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* (2.12). *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 of the groups with coefficients whose ℓ_2-norms 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* $1 \le q \le n \le p \to \infty$. *Suppose that the conditions in Theorem 2.1 hold. Then, with probability converging to* 1, *we have*

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

and

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

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

Corollary 2.2. *Let* $\lambda = 2\sigma\sqrt{8(1 + c_0')d_ad^2q^*\bar{c}c^*n\log(N_d)}$ *with a fixed* $c_0' \ge c_0$. *Suppose that all of the conditions in Theorem 2.2 hold. We then have*

$$\|\hat{\beta} - \beta\|_2 = \mathrm{O}_p\bigl(\sqrt{q\log(N_d)/n}\bigr) \quad \textit{and} \quad \|X\hat{\beta} - X\beta\|_2 = \mathrm{O}_p\bigl(\sqrt{q\log(N_d)}\bigr).$$

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 for improvement, particularly with regard to selection. Although the group Lasso selects a model whose dimension is comparable to that of the underlying model, the simulation results reported in Yuan and Lin (2006) and those reported below suggest that it tends to select more groups than there are in the underlying model. To correct the tendency of overselection 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 groups with small coefficients (in the ℓ_2 sense), then 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

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

Thus, for each group, its penalty is proportional to the inverse of the norm of $\tilde{\beta}_k$. This choice of the penalty level 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 $\theta_a = \max_{k\in A_0^c}\|\beta_k\|_2$ and $\theta_b = \min_{k\in A_0^c}\|\beta_k\|_2$. We say that an initial estimator $\tilde{\beta}$ is consistent at zero with rate r_n if $r_n \max_{k\in A_0}\|\tilde{\beta}_k\|_2 = \mathrm{O}_p(1)$, 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)
$$\frac{\sqrt{d_a(\log q)}}{\sqrt{n}\theta_b} \to 0, \qquad \frac{\tilde{\lambda}d_a^{3/2}q}{n\theta_b^2} \to 0, \qquad \frac{\sqrt{nd\log(p-q)}}{\tilde{\lambda}r_n} \to 0, \qquad \frac{d_a^{5/2}q^2}{r_n\theta_b\sqrt{d_b}} \to 0;$$

(C4) all of the eigenvalues of $\Sigma_{A_0^cA_0^c}$ are bounded away from zero and infinity.

Condition (C2) assumes that an initial zero-consistent estimator exists. It is the most critical one and is generally 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 $\sqrt{n/(q\log(N_d))}$. 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 ℓ_2-norm of the smallest important group. Condition (C4) assumes that the eigenvalues of $\Sigma_{A_0^cA_0^c}$ 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

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

Theorem 3.1. *If* (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 2.1 hold.

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

(C3)*

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

$$\frac{(d_a q)^{5/2}\sqrt{\log(N_d)}}{\theta_b\sqrt{nd_b}} \to 0.$$

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 θ_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. Suppose that the NSC* (2.3) *holds and that* (C1), (C2), (C3)* *and* (C4) *are satisfied. We then have*

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

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 that $\theta_b > t_b$ for some constant $t_b > 0$. If $\tilde{\lambda} \sim \mathrm{O}(n^{\alpha})$ for some $0 < \alpha < 1/2$, then*

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

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 an 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. Let $\lambda_k = \tilde{\lambda}/\|\hat{\beta}_k\|_2$, if $\|\hat{\beta}_k\|_2 > 0$; if $\|\hat{\beta}_k\|_2 = 0$, then $\lambda_k = \infty$, $\hat{\beta}_k^* = 0$. We can thus 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 of the examples, the sample size is $n = 200$.

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

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

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

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

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

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 $X = (X_1, \ldots, X_{10})$, where $X_j = (X_{5(j-1)+1}, \ldots, X_{5(j-1)+5})$, $1 \le j \le 5$, and $X_j = (X_{3(j-6)+26}, \ldots, X_{3(j-6)+28})$, $6 \le j \le 10$. In order to generate X, we first simulate 40 random variables $R_1, \ldots, R_{40}$, independently from $N(0, 1)$. Then, Z_j, $j = 1, \ldots, 10$ are simulated with a normal distribution with mean zero and

$\text{cov}(Z_{j_1}, Z_{j_2}) = 0.6^{|j_1 - j_2|}$. The covariates $X_1, \ldots, X_{40}$ are generated as

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

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

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

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

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

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

Table 1. Simulation study 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	98.00
	(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	3.00	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	40.0	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	7.0	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)

('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 sizes (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.

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; (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 satisfy 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.

6. Proofs

We first introduce some notation which will be used in proofs. Let $\{k\colon\ \|\hat{\beta}_k\|_2 > 0, k \le p\} \subseteq A_1 \subseteq \{k\colon\ X_k'(Y - X\hat{\beta}) = \lambda\sqrt{d_k}\hat{\beta}_k/\|\hat{\beta}_k\|_2\} \cup \{1, \ldots, q\}$. Set $A_2 = \{1, \ldots, p\} \setminus A_1$, $A_3 = A_1 \setminus A_0$, $A_4 = A_1 \cap A_0$, $A_5 = A_2 \setminus A_0$ and $A_6 = A_2 \cap A_0$. Thus, we have $A_1 = A_3 \cup A_4$, $A_3 \cap A_4 = \varnothing$, $A_2 = A_5 \cup A_6$ and $A_5 \cap A_6 = \varnothing$. Let $|A_i| = \sum_{k \in A_i} d_k$, $N(A_i) = \#\{k\colon\ k \in A_i\}$, $i = 1, \ldots, 6$ 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, for example, in the characterization of the solution via the Karush–Kuhn–Tucker (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, $\tilde{\omega}$ 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 1. The details of the complete proof are available from the website www.stat.uiowa.edu/techrep. We will sketch the proof in the following.

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

$$\Sigma_{11}^{-1}S_{A_1}/n = (\beta_{A_1} - \hat{\beta}_{A_1}) + \Sigma_{11}^{-1}\Sigma_{12}\beta_{A_2} + \Sigma_{11}^{-1}X_{A_1}'\varepsilon/n, \tag{6.1}$$

$$n\Sigma_{22}\beta_{A_2} - n\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\beta_{A_2} \le C_{A_2} - X_{A_2}'\varepsilon - \Sigma_{21}\Sigma_{11}^{-1}S_{A_1} + \Sigma_{21}\Sigma_{11}^{-1}X_{A_1}'\varepsilon, \tag{6.2}$$

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

Step 1. Define

$$V_{1j} = \Sigma_{11}^{-1/2}Q_{A_j1}'S_{A_j}/\sqrt{n}, \qquad j = 1, 3, 4, \qquad \omega_k = (I - P_{A_1})X_{A_k}\beta_{A_k}, \qquad k = 2, \ldots, 6,$$

where Q_{A_kj} is the matrix representing the selection of variables in A_k from A_j. Define $u = X_{A_1}\Sigma_{11}^{-1}Q_{A_41}'S_{A_4}/n - \omega_2/\|X_{A_1}\Sigma_{11}^{-1}Q_{A_41}'S_{A_4}/n - \omega_2\|_2$. From (6.1) and (6.2), we have $V_{14}'(V_{13} + V_{14}) \le S_{A_4}'Q_{A41}\Sigma_{11}^{-1}\Sigma_{12}\beta_{A_2} + S_{A_4}'Q_{A41}\Sigma_{11}^{-1}X_{A_1}'\varepsilon/n + \sqrt{d_a}\lambda\sum_{k\in A_4}\|\beta_k\|_2$ and $\|\omega_2\|_2^2 \le \beta_{A_2}'(C_{A_2} - X_{A_2}'\varepsilon - \Sigma_{21}\Sigma_{11}^{-1}S_{A_1} + \Sigma_{21}\Sigma_{11}^{-1}X_{A_1}'\varepsilon)$. Then, under GSC,

$$\begin{aligned}\|V_{14}\|_2^2 + \|\omega_2\|_2^2 \le{}& (\|V_{14}\|_2^2 + \|\omega_2\|_2^2)^{1/2}u'\varepsilon + (\|V_{14}\|_2 + \|P_1X_{A_2}\beta_{A_2}\|_2)\left(\frac{\lambda^2 d_a N(A_3)}{nc_*(|A_1|)}\right)^{1/2}\\ &+ \sqrt{d_a}\lambda\eta_1 + \lambda\sqrt{d_a}\|\beta_{A_5}\|_2.\end{aligned} \tag{6.3}$$

Step 2. Define $B_1^2 = \lambda^2 d_b q/(nc^*(|A_1|))$ and $B_2^2 = \lambda^2 d_b q/(nc_*(|A_0| \vee |A_1|))$. In this step, we consider the event $|u'\varepsilon|^2 \le (|A_1| \vee d_b)B_1^2/(4qd_a)$. Suppose that the set A_1 contains all large $\beta_k \ne 0$. From (6.3), $\|V_{14}\|_2^2 \le B_1^2 + 4\sqrt{d_a}\lambda\eta_1 + 4\sqrt{d}\eta_2 B_2 + 4dB_2^2$, so we have

$$(q_1 - q)^+ \le q + \frac{nc^*(|A_1|)}{\lambda^2 d_b}\left(4\sqrt{d_a}\lambda\eta_1 + 4\sqrt{\frac{\lambda^2 d_a q}{nc_*(|A_1|)}}\eta_2 + \frac{4\lambda^2 d_a q}{nc_*(|A_1|)}\right). \tag{6.4}$$

For general A_1, let $C_5 = c^*(|A_5|)/c_*(|A_1| \cup |A_5|)$. From (6.3),

$$\|\omega_2\|_2^2 \le \frac{4}{3}\left(\frac{B_1^2}{2} + dB_2^2 + \sqrt{d}\big(1 + \sqrt{C_5}\big)\eta_2 B_2 + 2\sqrt{d_a}\eta_1\right) + \frac{32}{9}dC_5B_2^2. \tag{6.5}$$

From Zhang and Huang (2008), $\|\omega_2\|_2^2 \ge (\|\beta_{A_5}\|_2(nc_{*,5})^{1/2} - \eta_2)^2$ and $\|X_{A_2}\beta_{A_2}\|_2 \le \eta_2 + \|X_{A_5}\beta_{A_5}\|_2 \le \eta_2 + (nc^*(|A_5|))^{1/2}\|\beta_{A_5}\|_2$. By the Cauchy–Schwarz inequality, then, we have

$$\begin{aligned}\|\beta_{A_5}\|_2^2 nc_{*,5} \le{}& \left[\frac{4}{3}\lambda\sqrt{\frac{d_a q}{nc_{*,5}}}\left(1 + \frac{c^*(|A_5|)}{c_*(|A_1|)}\right)^{1/2} + 2\eta_2\right]^2\\ &+ \frac{8}{3}\left[\frac{B_1^2}{4} + \sqrt{d_a}\lambda\eta_1 + \eta_2\left(\frac{\lambda^2 d_a q}{nc_*(|A_1|)}\right)^{1/2} + \frac{\lambda^2 d_a q}{2nc_*(|A_1|)} - \frac{3}{4}\eta_2^2\right],\end{aligned} \tag{6.6}$$

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

Step 3. Letting $c_*(|A_m|) = c_*$, $c^*(|A_m|) = c^*$ for $N(A_m) \le q^*$, we have

$$q_1 \le N(A_1 \cup A_5) \le q^*, \qquad |u'\varepsilon|^2 \le \frac{(|A_1| \vee d_b)\lambda^2 d_b}{4d_a n c^*(|A_1|)}. \tag{6.7}$$

We have $\bar{c} = C_5 = c^*(|A_5|)/c_*(|A_1| \vee |A_5|) = c^*/c_*$ and $c_{*,5} = c_*(|A_1 \cup A_5|) = c_*$. From (6.4), (6.5) and (6.6), $(q_1 - q)^+ + q \le M_1 q$, $\|\omega_2\|_2^2 \le M_2 B_1^2$, $nc_*\|\tilde{\gamma}_{A_5}\|_2^2 \le M_3 B_1^2$ when (2.12) is satisfied. Define

$$x_m^* \equiv \max_{|A|=m} \max_{\|U_{A_k}\|_2=1, k=1,\dots,m} \left| \varepsilon' \frac{X_A (X_A' X_A)^{-1} \bar{S}_A - (I - P_A) X\beta}{\|X_A (X_A' X_A)^{-1} \bar{S}_A - (I - P_A) X\beta\|_2} \right| \tag{6.8}$$

for $|A| = q_1 = m \ge 0$, $\bar{S}_A = (\bar{S}_{A_1}', \dots, \bar{S}_{A_m}')'$, 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 $k \in A$. For a given A, let $V_{lj} = (0, \dots, 0, 1, 0, \dots, 0)$ be the $|A| \times 1$ vector with the jth element in the lth group being 1. Then, by (6.8),

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

If we define $\Omega_{m_0} = \{(U, \varepsilon)\colon x_m^* \le \sigma\sqrt{8(1 + c_0) V^2 ((md_b) \vee d_b) \log(N_d \vee a_n)}\ \forall m \ge m_0\}$, then $(X, \varepsilon) \in \Omega_{m_0} \Rightarrow |u'\varepsilon|^2 \le (x_m^*)^2 \le (|A_1| \vee d_b)\lambda^2 d_b / (4d_a n c^*)$ for $N(A_1) \ge m_0 \ge 0$. 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 $\sigma^2 V_\varepsilon^2$, plus the maximum of $\binom{p}{m}$ normal variables with mean 0 and variance σ^2. It follows that $P\{(X, \varepsilon) \in \Omega_{m_0}\} \to 1$ when (6.7) holds. This completes 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. The required configurations in Theorem 2.1 then become

$$M_1 q + 1 < q^*, \qquad \eta_1 \le \frac{r_1^2}{c^*} \frac{q\lambda}{n}, \qquad \eta_2^2 \le \frac{r_2^2}{c^*} \frac{q\lambda^2}{n}. \tag{6.9}$$

Let $A_1 = \{k\colon \|\hat{\beta}_k\|_2 > 0 \text{ or } k \notin A_0\}$. Define $v_1 = X_{A_1}(\hat{\beta}_{A_1} - \beta_{A_1})$ and $g_1 = X_{A_1}'(Y - X\hat{\beta})$. We then have $\|v_1\|_2^2 \ge c_* n \|\hat{\beta}_{A_1} - \beta_{A_1}\|_2^2$, $(\hat{\beta}_{A_1} - \beta_{A_1})' g_1 = v_1'(X\beta - X_{A_1}\beta_{A_1} + \varepsilon) - \|v_1\|_2^2$ and $\|g_1\|_\infty \le \max_{k, \|\hat{\beta}_k\|_2 > 0} \|\lambda\sqrt{d_k}\hat{\beta}_k / \|\hat{\beta}_k\|_2\|_\infty = \lambda d_a$. Therefore, $\|v_1\|_2 \le \eta_2 + \|P_{A_1}\varepsilon\|_2 + \lambda\sqrt{d_a N(A_1)/(nc_*)}$. Since $\|P_{A_1}\varepsilon\|_2 \le 2\sigma\sqrt{N(A_1)\log(N_d)}$ with probability converging to 1 under the normality assumption, $\|X(\hat{\beta} - \beta)\|_2 \le 2\eta_2 + \|P_{A_1}\varepsilon\|_2 + \lambda\sqrt{d_a N(A_1)/(nc_*)}$. We then have

$$\left(\sum_{k \in A_1} \|\hat{\beta}_k - \beta_k\|_2^2 \right)^{1/2} \le \frac{\|v_1\|_2}{\sqrt{nc_*}} \le \frac{1}{\sqrt{nc_*}} \left(\eta_2 + 2\sigma\sqrt{N(A_1)\log(N_d)} + \sqrt{dM_1}\bar{c}B_1 \right). \tag{6.10}$$

Since $A_2 \subset A_0$, by the second inequality in (6.9), $\#\{k \in A_0\colon \|\beta_k\|_2 > \lambda/n\} \le r_1^2 q/c^* \sim \mathrm{O}(q)$. By the SRC and the third inequality in (6.9), $\sum_{k\in A_0} \|\beta_k\|_2^2 I\{\|\beta_k\|_2 > \lambda/n\} \le \sum_{k\in A_0} \|X_k\beta_k \times I\{\|\beta_k\|_2 > \lambda/n\}\|_2^2/(nc_*) \le r_2^2 q\lambda^2/(n^2 c_* c^*)$ and $\sum_{k\in A_0} \|\beta_k\|_2^2 I\{\|\beta_k\|_2 \le \lambda/n\} \le r_1^2 q\lambda^2/(c^* n^2)$. From (6.10), we then have

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

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

This completes 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 $\hat u = \min_u(\varepsilon - Xu)'(\varepsilon - Xu) + \sum_{k=1}^p \lambda_k\sqrt{d_k}\|u_k + \beta_k\|_2$, where $\lambda_k = \tilde\lambda/\|\tilde\beta_k\|_2$. By the KKT conditions, if there exists $\hat u$ such that

$$\Sigma_{A_0^c A_0^c}\bigl(\sqrt{n}\hat u_{A_0^c}\bigr) - W_{A_0^c} = -S_{A_0^c}/\sqrt{n}, \qquad \|\hat u_k\|_2 \le \|\beta_k\|_2 \qquad \text{for } k \in A_0^c, \tag{6.11}$$

$$-C_{A_0}/\sqrt{n} \le \Sigma_{A_0 A_0^c}\bigl(\sqrt{n}\hat u_{A_0^c}\bigr) - W_{A_0} \le C_{A_0}/\sqrt{n}, \tag{6.12}$$

then $\|\hat\beta_k\|_2 \ne 0$ for $k = 1,\dots,q$ and $\|\hat\beta_k\|_2 = 0$ for $k = q+1,\dots,p$.

From (6.11) and (6.12), $(\sqrt{n}\hat u_{A_0^c}) - \Sigma^{-1}_{A_0^c A_0^c} W_{A_0^c} = -\frac{1}{\sqrt{n}}\Sigma^{-1}_{A_0^c A_0^c} S_{A_0^c}$ and $\Sigma_{A_0 A_0^c}(\sqrt{n}\hat u_{A_0^c}) - W_{A_0} = -n^{-1/2}X'_{A_0}(I - P_{A_0^c})\varepsilon - n^{-1/2}\Sigma_{A_0 A_0^c}\Sigma^{-1}_{A_0^c A_0^c} S_{A_0^c}$. Define the events

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

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

where $(\cdot)_k$ denotes the d_k-dimensional subvector of the vector $(\cdot)$ corresponding to the kth group. We then have $P(\|\hat\beta_k\|_2 \ne 0, k \in A_0,\ \text{and } \|\hat\beta_k\|_2 = 0, k \notin A_0) \ge P(E_1 \cap E_2)$ and $P(E_1 \cap E_2) = 1 - P(E_1^c \cup E_2^c) \ge 1 - P(E_1^c) - P(E_2^c)$.

First, we consider $P(E_1^c)$. Define $R = \{\|\tilde\beta_k\|_2^{-1} \le c_1\theta_b^{-1}, k \in A_0^c\}$, where c_1 is a constant. $P(E_1^c) = P(E_1^c \cap R) + P(E_1^c \cap R^c) \le P(E_1^c \cap R) + P(R^c)$. By (C2), $P(R^c) \to 0$. Let $N_q = \sum_{k=1}^q d_k$, $\tau_1 \le \cdots \le \tau_{N_q}$ be the eigenvalues of $\Sigma_{A_0^c A_0^c}$ and $\gamma_1,\dots,\gamma_{N_q}$ be the associated eigenvectors. The jth element in the lth group of vector $\Sigma^{-1}_{A_0^c A_0^c} S_{A_0^c}$ is $u_{lj} = \sum_{l'=1}^{N_q} \tau_{l'}^{-1}(\gamma_{l'}' S_{A_0^c})\gamma_{lj}$. By the Cauchy–Schwarz inequality, $u_{lj}^2 \le \tau_1^{-2}\sum_{l=1}^{N_q}\|\gamma_l\|_2^2\|S_{A_0^c}\|_2^2 = \tau_1^{-2}N_q\|S_{A_0^c}\|_2^2 \le \tau_1^{-2}N_q(\sum_{k=1}^q \lambda_k^2 d_k)$. Therefore, $\|u_k\|_2^2 \le d_k\tau_1^{-2}q^2 d_a^2(\tilde\lambda c_1\theta_b^{-1})^2$.

If we define $\upsilon_{A_0^c} = \sqrt{n}\theta_b - n^{-1/2}c_1\tau_1^{-1}q d_a^{3/2}\tilde\lambda\theta_b^{-1}$, $\eta_{A_0^c} = n^{-1/2}\Sigma^{-1}_{A_0^c A_0^c} X'_{A_0^c}\varepsilon$, $\xi_{A_0} = n^{-1/2} \times X'_{A_0}(I - P_{A_0^c})\varepsilon$, $C_{A_0^c} = \{\max_{k\in A_0^c}\|\eta_k\|_2 \ge \upsilon_{A_0^c}\}$, then $P(E_1^c) \le P(C_{A_0^c})$. By Lemmas 1 and 2 of Huang, Ma and Zhang (2008), $P(C^c_{A_0}) \le K(d_a\log q)^{1/2}/\upsilon_{A_0^c}$, where K is a constant, $k(d_a\log q)^{1/2}/\upsilon_{A_0^c} \to 0$ from (C3). We then have $P(E_1^c \cap R) \to 0$, $P(E_1^c) \to 0$.

Next, we consider $P(E_2^c)$. Similarly as above, define $D = \{\|\tilde{\beta}_k\|_2^{-1} > r_n, k \in A_0\} \cap R$. $P(E_2^c) \le P(E_2^c \cap D) + P(D^c)$. By (C2), $P(D^c) \to 0$. $|\sum_{l=1}^{N_q} \sum_{i=1}^{n} (X_{A_0})_{ij} (X_{A_0^c})_{il} u_l| \le \sum_{l=1}^{N_q} |u_l/n| \le \tau_1^{-1} q^2 d_a^2 \tilde{\lambda} c_1 \theta_b^{-1}$, where u_l is the lth element of vector $\Sigma_{A_0^c A_0^c}^{-1} S_{A_0^c}$. If we 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 $P(Q^c) \le P(C_{A_0})$, $P(C_{A_0}) \le K(d_a \log(p-q))^{1/2}/\upsilon_{A_0}$. $K(d_a \log(p-q))^{1/2}/\upsilon_{A_0} \to 0$ from (C3). We then have $P(E_2^c \cap D) \to 0$, $P(E_2^c) \to 0$. This completes the proof of Theorem 3.1. □

Proof of Theorem 3.2. If we let $\hat{A} = \{k\colon \|\tilde{\beta}_k\|_2 > 0, k = 1, \ldots, p\}$, then $\sum_{k \in \hat{A}^c} \|\hat{\beta}_k^*\|_2 = 0$, the dimension of our problem (3.1) is reduced to $\hat{q}$, $\hat{q} \le q^*$ and $\hat{A}_c \subset A_0$. By the definition of $\hat{\beta}^*$, we have

$$\frac{1}{2}\|Y - X_{\hat{A}}\hat{\beta}_{\hat{A}}^*\|_2^2 + \tilde{\lambda} \sum_{k \in \hat{A}} \frac{\sqrt{d_k}}{\|\tilde{\beta}_k\|_2} \|\hat{\beta}_k^*\|_2 \le \frac{1}{2}\|Y - X_{\hat{A}}\beta_{\hat{A}}\|_2^2 + \tilde{\lambda} \sum_{k \in \hat{A}} \frac{\sqrt{d_k}}{\|\tilde{\beta}_k\|_2} \|\beta_k\|_2, \tag{6.13}$$

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

If we let $\delta_{\hat{A}} = \Sigma_{\hat{A}\hat{A}}^{1/2}(\hat{\beta}_{\hat{A}}^* - \beta_{\hat{A}})$ and $D = \Sigma_{\hat{A}\hat{A}}^{-1/2} X'_{\hat{A}}$, then $\|Y - X_{\hat{A}}\hat{\beta}_{\hat{A}}^*\|_2^2/2 - \|Y - X_{\hat{A}}\beta_{\hat{A}}\|_2^2/2 = \delta'_{\hat{A}}\delta_{\hat{A}}/2 - (D\varepsilon)'\delta_{\hat{A}}$. By (6.13) and (6.14), $\delta'_{\hat{A}}\delta_{\hat{A}}/2 - (D\varepsilon)'\delta_{\hat{A}} - \eta^* \le 0$, so $\|\delta_{\hat{A}} - D\varepsilon\|_2^2 - \|D\varepsilon\|_2^2 - 2\eta^* \le 0$. By the triangle inequality, $\|\delta_{\hat{A}}\|_2 \le \|\delta_{\hat{A}} - D\varepsilon\|_2 + \|D\varepsilon\|_2$. Thus, $\|\delta_{\hat{A}}\|_2^2 \le 6\|D\varepsilon\|_2^2 + 6\eta^*$.

Let D_i be the ith column of D. $E(\|D\varepsilon\|_2^2) = \sigma^2 \operatorname{tr}(D'D) = \sigma^2 \hat{q}$. Then, with probability converging to 1, $\|\hat{\beta}_{\hat{A}} - \beta_{\hat{A}}\|_2^2 \le 6\sigma^2 M_1 q/(nc_*) + (\tilde{\lambda}\sqrt{d_a}/(\xi_b \theta_b n c_*))^2/2 + \|\hat{\beta}_{\hat{A}} - \beta_{\hat{A}}\|_2^2/2$. Thus, for $\tilde{\lambda} = n^{\alpha}$ for some $0 < \alpha < 1/2$, with probability converging to 1,

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

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

Acknowledgements

The authors are grateful to Professor Cun-Hui Zhang for sharing his insights into the problem and related topics. The work of Jian Huang is supported in part by NIH Grant R01CA120988 and NSF Grants DMS-07-06108 and 0805670.

References

Antoniadis, A. and Fan, J. (2001). Regularization of wavelet approximation (with discussion). *J. Amer. Statist. Assoc.* **96** 939–967. MR1946364

Bühlmann, P. and Meier, L. (2008). Discussion of "One-step sparse estimates in nonconcave penalized likelihood models," by H. Zou and R. Li. *Ann. Statist.* **36** 1534–1541. MR2435444

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

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32** 928–961. MR2065194

Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10** 971–988. MR2108039

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. MR2396808

Huang, J., Ma, S. and Zhang, C.H. (2006). Adaptive lasso for sparse high-dimensional regression models. *Statist. Sinica* **18** 1603–1618. MR2469326

Kim, Y., Kim, J. and Kim, Y. (2006). The blockwise sparse regression. *Statist. Sinica* **16** 375–390. MR2267240

Knight, K. and Fu, W.J. (2001). Asymptotics for lasso-type estimators. *Ann. Statist.* **28** 1356–1378. MR1805787

Meier, L., van de Geer, S. and Bühlmann, P. (2008). Group Lasso for logisitc regression. *J. R. Stat. Soc. Ser. B* **70** 53–71. MR2412631

Meinshausen, N. and Buhlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B* **58** 267–288. MR1379242

van de Geer, S. (2008). High-dimensional generalized linear models and the Lasso. *Ann. Statist.* **36** 614–645. MR2396809

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B* **68** 49–67. MR2212574

Zhang, C.H. (2007). Penalized linear unbiased selection. Technical Report 2007-003, Dept. Statistics, Rutgers Univ.

Zhang, C.H. and Huang, J. (2008). Model-selection consistency of the LASSO in high-dimensional linear regression. *Ann. Statist.* **36** 1567–1594. MR2435448

Zhao, P., Rocha, G. and Yu, B. (2008). Grouped and hierarchical model selection through composite absolute penalties. *Ann. Statist.* **36** 1567–1594. MR2549566

Zhao, P. and Yu, B. (2006). On model selection consistency of LASSO. *J. Mach. Learn. Res*. **7** 2541–2563. MR2274449

Zou, H. (2006). The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

Zou, H. and Hastie, T. (2006). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B* **67** 301–320. MR2137327

Received November 2008 and revised September 2009