Submitted to the Annals of Statistics

THE BENEFIT OF GROUP SPARSITY

By Junzhou Huang

Computer Science Department, Rutgers University

By Tong Zhang

Statistics Department, Rutgers University

This paper develops a theory for group Lasso using a concept called *strong group sparsity*. Our result shows that group Lasso is superior to standard Lasso for strongly group-sparse signals. This provides a convincing theoretical justification for using group sparse regularization when the underlying group structure is consistent with the data. Moreover, the theory predicts some limitations of the group Lasso formulation that are confirmed by simulation studies.

1. Introduction. We are interested in the sparse learning problem for least squares regression. Consider a set of p basis vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$ where $\mathbf{x}_j \in \mathbb{R}^n$ for each j. Here, n is the sample size.

Denote by X the $n \times p$ data matrix, with column j of X being $\mathbf{x}_j$. Given an observation $\mathbf{y} = [y_1, \ldots, y_n] \in \mathbb{R}^n$ that is generated from a sparse linear combination of the basis vectors plus a stochastic noise vector $\epsilon \in \mathbb{R}^n$:

$$\mathbf{y} = X\bar{\beta} + \epsilon = \sum_{j=1}^{d} \bar{\beta}_j \mathbf{x}_j + \epsilon,$$

where we assume that the target coefficient $\bar{\beta}$ is sparse. Throughout the paper, we consider fixed design only. That is, we assume X is fixed, and randomization is with respect to the noise ϵ (and thus the observation $\mathbf{y}$). Note that we do not assume that the noise ϵ is zero-mean.

Define the support of a sparse vector $\beta \in \mathbb{R}^p$ as

$$\mathrm{supp}(\beta) = \{j : \beta_j \neq 0\},$$

and $\|\beta\|_0 = |\mathrm{supp}(\beta)|$. A natural method for sparse learning is L_0 regularization:

$$\hat{\beta}_{L0} = \arg\min_{\beta \in \mathbb{R}^p} \|X\beta - \mathbf{y}\|_2^2 \quad \text{subject to } \|\beta\|_0 \leq k,$$

where k is the sparsity. Since this optimization problem is generally NP-hard, in practice, one often consider the following L_1 regularization problem, which is the standard convex relaxation of L_0:

$$\hat{\beta}_{L1} = \arg\min_{\beta \in \mathbb{R}^p} \left[\frac{1}{n} \|X\beta - \mathbf{y}\|_2^2 + \lambda \|\beta\|_1 \right],$$

where λ is an appropriately chosen regularization parameter. This method is often referred to as Lasso in the statistical literature.

In practical applications, one often knows a group structure on the coefficient vector $\bar{\beta}$ so that variables in the same group tend to be zeros or nonzeros simultaneously. The purpose of this paper is to show that if such a structure exists, then better results can be obtained.

2. Strong Group Sparsity. For simplicity, we shall only consider non-overlapping groups in this paper, although our analysis can be adapted to handle moderately overlapping groups (that is, each feature is only covered by a constant number of groups, and the resulting analysis depends on this constant).

Assume that $\{1,\ldots,p\} = \cup_{j=1}^m G_j$ is partitioned into m disjoint groups $G_1, G_2, \ldots, G_m$: $G_i \cap G_j = \emptyset$ when $i \neq j$. Moreover, throughout the paper, we let $k_j = |G_j|$, and $k_0 = \max_{j\in\{1,\ldots,m\}} k_j$. Given $S \subset \{1,\ldots,m\}$ that denotes a set of groups, we define $G_S = \cup_{j\in S} G_j$.

Given a subset of variables $F \subset \{1,\ldots,p\}$ and a coefficient vector $\beta \in \mathbb{R}^p$, let β_F be the vector in $\mathbb{R}^{|F|}$ which is identical to β in F. Similar, X_F is the $n \times |F|$ matrix with columns identical to X in F.

The following method, often referred to as group Lasso, has been proposed to take advantage of the group structure:

$$\hat{\beta} = \arg\min_{\beta} \left[\frac{1}{n} \|X\beta - \mathbf{y}\|_2^2 + \sum_{j=1}^m \lambda_j \|\beta_{G_j}\|_2 \right]. \tag{1}$$

The purpose of this paper is to develop a theory that characterizes the performance of (1). We are interested in conditions under which group Lasso yields better estimate of $\bar{\beta}$ than the standard Lasso.

Instead of the standard sparsity assumption, where the complexity is measured by the number of nonzero coefficients k, we introduce the strong group sparsity concept below. The idea is to measure the complexity of a sparse signal using group sparsity in addition to coefficient sparsity.

DEFINITION 2.1. *A coefficient vector $\bar{\beta} \in \mathbb{R}^p$ is (g,k) strongly group-sparse if there exists a set S of groups such that*

$$\mathrm{supp}(\bar{\beta}) \subset G_S, \quad |G_S| \leq k, \quad |S| \leq g.$$

The new concept is referred to as strong group-sparsity because k is used to measure the sparsity of $\bar{\beta}$ instead of $\|\bar{\beta}\|_0$. If this notion is beneficial, then $k/\|\bar{\beta}\|_0$ should be small, which means that the signal has to be efficiently covered by the groups. In fact, the group Lasso method does not work well when $k/\|\bar{\beta}\|_0$ is large. In that case, the signal is only *weak group sparse*, and one needs to use $\|\bar{\beta}\|_0$ to precisely measure the real sparsity of the signal. Unfortunately, such information is not included in the group Lasso formulation, and there is no simple fix of this problem using variations of group Lasso. This is because our theory requires that the group Lasso regularization term is strong enough to dominate the noise, and the strong regularization causes a bias of the order $O(k)$ which cannot be removed. This is one fundamental drawback which is inherent to the group Lasso formulation.

We shall mention that this paper focuses on the scenario that each group is finite dimensional, and our analysis relies on the overall sparsity k. For some applications, each group may be an infinite dimensional Hilbert space, and the group Lasso can be used to learn combinations of kernels (see [1, 5] for analysis and references). For such problems, our analysis does not apply because the sparsity k may be infinity. Also in such case, Lasso cannot be run and thus group Lasso will be the only natural formulation.

3. Related Work. The idea of using group structure to achieve better sparse recovery performance has received much attention. For example, group sparsity has been considered for simultaneous sparse approximation [12] and multi-task compressive sensing [4] from the Bayesian

hierarchical modeling point of view. Under the Bayesian hierarchical model framework, data from all sources contribute to the estimation of hyper-parameters in the sparse prior model. The shared prior can then be inferred from multiple sources. Although the idea can be justified using standard Bayesian intuition, there are no theoretical results showing how much better (and under what kind of conditions) the resulting algorithms perform.

In [11], the authors attempted to derive a bound on the number of samples needed to recover block sparse signals, where the coefficients in each block are either all zero or all nonzero. In our terminology, this corresponds to the case of group sparsity with equal size groups. The algorithm considered there is a special case of (1) with $\lambda_j \to 0^+$. However, their result is very loose, and does not demonstrate the advantage of group Lasso over standard Lasso.

In the statistical literature, the group Lasso (1) has been studied by a number of authors [1, 5, 7, 8, 13]. There were no theoretical results in [13]. Although some theoretical results were developed in [1, 7], neither showed that group Lasso is superior to the standard Lasso. In particular, although [7] is related to our work (in the sense that it also studies parameter estimation error), the analysis does not try to show the advantage of group Lasso over standard Lasso.

The authors of [5] showed that group Lasso can be superior to standard Lasso when each group is an infinite dimensional kernel, by using an argument completely different from ours (they relied on the fact that meaningful analysis can be obtained for kernel methods in infinite dimension). Their idea cannot be adapted to show the advantage of group Lasso in finite dimensional scenarios of interests such as in the standard compressive sensing setting. Therefore our analysis, which focuses on the latter, is complementary to their work.

Another related work is [8], where the authors considered a special case of group Lasso in the multi-task learning scenario, and showed that the number of samples required for recovering the exact support set may be smaller for group Lasso under appropriate conditions. The analysis is quite tight but with different assumptions than what we make in this paper. That is, there are major differences between our analysis and their analysis. For example, the group formulation we consider here is more general and includes the multi-task scenario as a special case. Moreover, we study signal recovery performance in 2-norm instead of the exact recovery of support set in their analysis. The sparse eigenvalue condition employed in this work is different from the irrepresentable type condition in their analysis (which is required for exact support set recovery). Under our assumptions, either Lasso nor group Lasso may be able to recover the exact support set.

In the above context, the main contribution of this work is the introduction of the strong group sparsity concept, under which a satisfactory theory of group Lasso is developed. Our result shows that strongly group sparse signals can be estimated more reliably using group Lasso, in that it requires fewer number of samples in the compressive sensing setting, and is more robust to noise in the statistical estimation setting.

Finally, we shall mention that independent of the authors, results similar to those presented in this paper have also been obtained in [6] with a similar technical analysis. However, while our paper studies the general group Lasso formulation, only the special case of multi-task learning is considered in [6].

4. Assumptions. The following assumption on the noise is important in our analysis. It captures an important advantage of group Lasso over standard Lasso under the strong group sparsity assumption.

ASSUMPTION 4.1 (Group noise condition). *There exist non-negative constants a, b such that for any fixed group $j \in \{1, \ldots, m\}$, and $\eta \in (0, 1)$: with probability larger than $1-\eta$, the noise projection*

to the j-th group is bounded by:

$$\|(X_{G_j}^\top X_{G_j})^{-0.5} X_{G_j}^\top (\epsilon - \mathbb{E}\epsilon)\|_2 \leq a\sqrt{k_j} + b\sqrt{-\ln \eta}.$$

The importance of the assumption is that the concentration term $\sqrt{-\ln \eta}$ does not depend on k. This reveals a significant benefit of group Lasso over standard Lasso: that is, the concentration term does not increase when the group size increases. This implies that if we can correctly guess the group sparsity structure, the group Lasso estimator is more stable with respect to stochastic noise than the standard Lasso.

We shall point out that this assumption holds for independent sub-Gaussian noise vectors, where $e^{t(\epsilon_i - \mathbb{E}\epsilon_i)} \leq e^{t^2\sigma^2/2}$ for all t and $i = 1, \ldots, n$. It can be shown that one may choose $a = 2.8$ and $b = 2.4$ when $\eta \in (0, 0.5)$. Since a complete treatment of sub-Gaussian noise is not important for the purpose of this paper, we only prove this assumption under independent Gaussian noise, which can be directly calculated.

PROPOSITION 4.1. *Assume the noise vector ϵ are independent Gaussians: $\epsilon_i - \mathbb{E}\epsilon_i \sim N(0, \sigma_i^2)$, where each $\sigma_i \leq \sigma$ $(i = 1, \ldots, n)$. Then Assumption 4.1 holds with $a = \sigma$ and $b = \sqrt{2}\sigma$.*

The next assumption handles the case that true target is not exactly sparse. That is, we only assume that $X\bar{\beta} \approx \mathbb{E}\mathbf{y}$.

ASSUMPTION 4.2 (Group approximation error condition). *There exist $\delta a, \delta b \geq 0$ such that for all group $j \in \{1, \ldots, m\}$: the projection of error mean $\mathbb{E}\epsilon$ to the j-th group is bounded by:*

$$\|(X_{G_j}^\top X_{G_j})^{-0.5} X_{G_j}^\top \mathbb{E}\epsilon\|_2/\sqrt{n} \leq \sqrt{k_j}\delta a + \delta b.$$

As mentioned earlier, we do not assume that the noise is zero-mean. Hence $\mathbb{E}\epsilon$ may not equal zero. In other words, this condition considers the situation that the true target is not exactly sparse. It resembles algebraic noise in [15] but takes the group structure into account. Similar to [15], we have the following result.

PROPOSITION 4.2. *Consider a (g, k) strongly group sparse coefficient vector $\bar{\beta}$ such that*

$$\frac{1}{n}\|X\bar{\beta} - \mathbb{E}\mathbf{y}\|_2^2 \leq \Delta^2,$$

and $a_0, b_0 \geq 0$. Then there exists (g', k') strongly group sparse $\bar{\beta}'$ such that $k'a_0^2 + g'b_0^2 \leq 2(ka_0^2 + gb_0^2)$, $\|X\bar{\beta}' - \mathbb{E}\mathbf{y}\|_2 \leq \|X\bar{\beta} - \mathbb{E}\mathbf{y}\|_2$, $\mathrm{supp}(\bar{\beta}) \subset \mathrm{supp}(\bar{\beta}')$, and for all group j:

$$\|(X_{G_j}^\top X_{G_j})^{-0.5} X_{G_j}^\top (X\bar{\beta}' - \mathbb{E}\mathbf{y})\|_2/\sqrt{n} \leq (a_0\sqrt{k_j} + b_0)\Delta/\sqrt{ka_0^2 + b_0^2}.$$

The proposition shows that if the approximation error of $\bar{\beta}$ is $\Delta = \|X\bar{\beta} - \mathbb{E}\mathbf{y}\|_2/\sqrt{n}$, then we may find an alternative target $\bar{\beta}'$ with similar sparsity for which we can take $\delta a = a_0\Delta/\sqrt{ka_0^2 + b_0^2}$ and $\delta b = b_0\Delta/\sqrt{ka_0^2 + b_0^2}$ in Assumption 4.2. This means that in Theorem 5.1 below, by choosing $a_0 = a$ and $b_0 = b\sqrt{\ln(m/\eta)}$, the contribution of the approximation error to the reconstruction error $\|\hat{\beta} - \bar{\beta}\|_2$ is $O(\Delta)$. Note that this assumption does not show the benefit of group Lasso over standard Lasso. Therefore in order to compare our results to that of the standard Lasso, one may consider

the simple situation where $\delta a = \delta b = 0$. That is, the target is exactly sparse. The only reason to include Assumption 4.2 is to illustrate that our analysis can handle approximate sparsity.

The last assumption is a sparse eigenvalue condition, used in the modern analysis of Lasso (e.g., [2, 15]). It is also closely related to (and slightly weaker than) the RIP (restricted isometry property) assumption [3] in the compressive sensing literature. This assumption takes advantage of group structure, and can be considered as (a weaker version of) group RIP. We introduce a definition before stating the assumption.

DEFINITION 4.1. *For all $F \subset \{1, \ldots, p\}$, define*

$$\rho_-(F) = \inf\left\{\frac{1}{n}\|X\beta\|_2^2/\|\beta\|_2^2 : \mathrm{supp}(\beta) \subset F\right\},$$
$$\rho_+(F) = \sup\left\{\frac{1}{n}\|X\beta\|_2^2/\|\beta\|_2^2 : \mathrm{supp}(\beta) \subset F\right\}.$$

Moreover, for all $1 \le s \le p$, define

$$\rho_-(s) = \inf\{\rho_-(G_S) : S \subset \{1, \ldots, m\}, |G_S| \le s\},$$
$$\rho_+(s) = \sup\{\rho_+(G_S) : S \subset \{1, \ldots, m\}, |G_S| \le s\}.$$

ASSUMPTION 4.3 (Group sparse eigenvalue condition). *There exist $s, c > 0$ such that*

$$\frac{\rho_+(s) - \rho_-(2s)}{\rho_-(s)} \le c.$$

Assumption 4.3 illustrates another advantage of group Lasso over standard Lasso. Since we only consider eigenvalues for sub-matrices consistent with the group structure $\{G_j\}$, the ratio $\rho_+(s)/\rho_-(s)$ can be significantly smaller than the corresponding ratio for Lasso (which considers all subsets of $\{1, \ldots, p\}$ up to size s). For example, assume that all group sizes are identical $k_1 = \ldots = k_m = k_0$, and s is a multiple of k_0. For random projections used in compressive sensing applications, only $n = O(s + (s/k_0)\ln m)$ projections are needed for Assumption 4.3 to hold. In comparison, for standard Lasso, we need $n = O(s \ln p)$ projections. The difference can be significant when p and k_0 are large. More precisely, we have the following random projection sample complexity bound for the group sparse eigenvalue condition. Although we assume Gaussian random matrix in order to state explicit constants, it is clear that similar results hold for other sub-Gaussian random matrices.

PROPOSITION 4.3 (Group-RIP). *Suppose that elements in X are iid standard Gaussian random variables $N(0,1)$. For any $t > 0$ and $\delta \in (0,1)$, let*

$$n \ge \frac{8}{\delta^2}[\ln 3 + t + k\ln(1 + 8/\delta) + g\ln(em/g)].$$

Then with probability at least $1 - e^{-t}$, the random matrix $X \in \mathbb{R}^{n\times p}$ satisfies the following group-RIP inequality for all (g,k) strongly group-sparse vector $\bar\beta \in \mathbb{R}^p$,

$$(1-\delta)\|\bar\beta\|_2 \le \frac{1}{\sqrt{n}}\|X\bar\beta\|_2 \le (1+\delta)\|\bar\beta\|_2. \tag{2}$$

5. Main Results. Our main result is the following signal recovery (2-norm parameter estimation error) bound for group Lasso.

THEOREM 5.1. *Suppose that Assumption 4.1, Assumption 4.2, and Assumption 4.3 are valid. Take* $\lambda_j = (A\sqrt{k_j} + B)/\sqrt{n}$, *where both* A *and* B *can depend on data* $\mathbf{y}$. *Given* $\eta \in (0,1)$, *with probability larger than* $1-\eta$, *if the following conditions hold:*

- $A \geq 4\max_j \rho_+(G_j)^{1/2}(a + \delta a\sqrt{n})$,
- $B \geq 4\max_j \rho_+(G_j)^{1/2}(b\sqrt{\ln(m/\eta)} + \delta b\sqrt{n})$,
- $\bar{\beta}$ *is a* (g,k) *strongly group-sparse coefficient vector,*
- $s \geq k + k_0$,
- *Let* $\ell = s - (k-k_0) + 1$, *and* $g_\ell = \min\{|S| : |G_S| \geq \ell, S \subset \{1,\ldots,m\}\}$, *we have*

$$c^2 \leq \frac{\ell A^2 + g_\ell B^2}{72(kA^2 + gB^2)},$$

then the solution of (1) satisfies:

$$\|\hat{\beta} - \bar{\beta}\|_2 \leq \frac{\sqrt{4.5}}{\rho_-(s)\sqrt{n}}(1 + 0.25c^{-1})\sqrt{A^2k + gB^2}.$$

The first four conditions of the theorem are not critical, as they are just definitions and choices for λ_j. The fifth assumption is critical, which means that the group sparse eigenvalue condition has to be satisfied with some c that is not too large. In order to satisfy the condition, ℓ should be chosen relatively large as the right hand side is linear in ℓ. However, this implies that s also grow linearly. It is possible to find s so that the condition is satisfied when c^2 in Assumption 4.3 grows sub-linearly in s. Consider the situation that $\delta a = \delta b = 0$. If the conditions of Theorem 5.1 is satisfied, then

$$\|\hat{\beta} - \bar{\beta}\|_2^2 = O((k + g\ln(m/\eta))/n).$$

In comparison, The Lasso estimator can only achieve the bound

$$\|\hat{\beta}_{L1} - \bar{\beta}\|_2^2 = O((\|\bar{\beta}\|_0 \ln(p/\eta))/n).$$

If $k/\|\bar{\beta}\|_0 \ll \ln(p/\eta)$ (which means that the group structure is useful) and $g \ll \|\bar{\beta}\|_0$, then the group Lasso is superior. This is consistent with intuition. However, if $k \gg \|\bar{\beta}\|_0 \ln(p/\eta)$, then group Lasso is inferior. This happens when the signal is not strongly group sparse.

Theorem 5.1 also suggests that if the group sizes are not even, then group Lasso may not work well when the signal is contained in small sized groups. This is because in such case g_ℓ can be significantly smaller than g even with relatively large ℓ, which means we have to choose a large s and small c, implying a poor bound. This prediction is confirmed in Section 7.2 using simulated data. Intuitively, group Lasso favors large sized groups because the 2-norm regularization for large group size is weaker. Adjusting regularization parameters λ_j not only fails to work in theory, but also impractical since it is unrealistic to tune many parameters. This unstable behavior with respect to uneven group size may be regarded as another drawback of the group Lasso formulation.

In the following, we present two simplifications of Theorem 5.1 that are easier to interpret. The first is the compressive sensing case, which does not consider stochastic noise.

COROLLARY 5.1 (Compressive sensing). *Suppose that Assumption 4.1 and Assumption 4.2 are valid with* $a = b = \delta b = 0$. *Take* $\lambda_j = 4\sqrt{k_j}\max_j \rho_+(G_j)^{1/2}\delta a$. *Let* $\bar{\beta}$ *be a* (k,g) *strongly group-sparse*

signal, $\ell = k$, and $s = 2k + k_0 - 1$. If $(\rho_+(s) - \rho_-(2s))/\rho_-(s) \leq 1/\sqrt{72}$, then the solution of (1) satisfies:

$$\|\hat{\beta} - \bar{\beta}\|_2 \leq \frac{6\sqrt{2} + 18}{\rho_-(s)} \max_j \rho_+(G_j)^{1/2} \delta a \sqrt{k}.$$

If $\delta a = 0$, then we can achieve exact recovery. Moreover, Proposition 4.2 implies that we may choose a target with similar sparsity such that $\delta a\sqrt{k} = O(\|X\bar{\beta} - \mathbb{E}\mathbf{y}\|_2/\sqrt{n})$. This implies a bound

$$\|\hat{\beta} - \bar{\beta}\|_2 = O(\|X\bar{\beta} - \mathbb{E}\mathbf{y}\|_2/\sqrt{n}).$$

If we have even sized groups, the number of samples n required for Corollary 5.1 to hold (that is, $(\rho_+(s) - \rho_-(2s))/\rho_-(s) \leq 1/\sqrt{72}$) is $O(k + g\ln(m/g))$, where $g = k/k_0$. In comparison, although a similar result holds for Lasso, it requires sample size of order $\|\bar{\beta}\|_0 \ln(p/\|\bar{\beta}\|_0)$. Again, group Lasso has a significant advantage if $k/\|\bar{\beta}\|_0 \ll \ln(p/\|\bar{\beta}\|_0)$, $g \ll \|\bar{\beta}\|_0$, and p is large.

The following corollary is for even sized groups, and the result is simpler to interpret. For standard Lasso, $B = O(\sqrt{\ln p})$, and for group Lasso, $B = O(\sqrt{\ln m})$. The benefit of group Lasso is the division of B^2 by k_0 in the bound, which is a significant improvement when the dimensionality p is large. The disadvantage of group Lasso is that the signal sparsity $\|\bar{\beta}\|_0$ is replaced by the group sparsity k. This is not an artifact of our analysis, but rather a fundamental drawback inherent to the group Lasso formulation. The effect is observable, as shown in our simulation studies.

COROLLARY 5.2 (Even group size). *Suppose that Assumption 4.1 and Assumption 4.2 are valid. Assume also that all groups are of equal sizes: $k_0 = k_j$ for $j = 1, \ldots, m$. Given $\eta \in (0, 1)$, let*

$$\lambda_j = (A\sqrt{k_0} + B)/\sqrt{n},$$

where $A \geq 4\max_j \rho_+(G_j)^{1/2}(a + \delta a\sqrt{n})$ and $B \geq 4\max_j \rho_+(G_j)^{1/2}(b\sqrt{\ln(m/\eta)} + \delta b\sqrt{n})$. Let $\bar{\beta}$ be a $(k, k/k_0)$ strongly group-sparse signal. With probability larger than $1 - \eta$, if

$$6\sqrt{2}(\rho_+(k + \ell) - \rho_-(2k + 2\ell))/\rho_-(k + \ell) < \sqrt{\ell/k}$$

for some $\ell > 0$ that is a multiple of k_0, then the solution of (1) satisfies:

$$\|\hat{\beta} - \bar{\beta}\|_2 \leq \rho_-(k + \ell)^{-1}(\sqrt{4.5} + 4.5\ell/k)\sqrt{A^2 + B^2/k_0}\sqrt{k/n}.$$

6. Parameter Estimation Lower Bound. The following parameter estimation lower bound applies to all statistical estimators. In order to simplify the proof, we intentionally exclude the $\Omega(k/n)$ term from the lower bound (see comments in the proof), as this is a well-known term from the classical parametric statistics.

THEOREM 6.1. *Given an $n \times p$ design matrix X, we define $\forall \bar{\beta} \in \mathbb{R}^p$ the following probability density for $\mathbf{y} \in \mathbb{R}^n$:*

$$p_{\bar{\beta}}(\mathbf{y}) = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\|\mathbf{y} - X\bar{\beta}\|_2^2/(2\sigma^2)}.$$

Let $H(g, k)$ be the family of (g, k) strongly group-sparse signals in $\mathbb{R}^p$ with respect to a set of m pre-defined groups with even group size $k_0 = p/m$, where $k = gk_0$. Let $\hat{\beta}(\mathbf{y}) \in \mathbb{R}^p$ be an arbitrary statistical estimator of $\bar{\beta}$ based on $\mathbf{y} \sim p_{\bar{\beta}}$. If $g < m/2$, then we have

$$\sup_{\bar{\beta} \in H(g,k)} \mathbb{E}_{\mathbf{y} \sim p_{\bar{\beta}}} \|X(\hat{\beta}(\mathbf{y}) - \bar{\beta})\|_2^2 \geq \sigma^2 \frac{\rho_-(2g)}{32\rho_+(2g)} [g\ln((m - g)/g) - (g + 2)\ln 4].$$

It implies the following lower bound on the 2-norm parameter estimation error:

$$\sup_{\bar{\beta}\in H(g,k)} \mathbb{E}_{\mathbf{y}\sim p_{\bar{\beta}}} \|\hat{\beta}(\mathbf{y}) - \bar{\beta}\|_2^2 \geq \sigma^2 \frac{\rho_-(2g)}{96n\rho_+(2g)^2}[g\ln((m-g)/g) - (g+2)\ln 4].$$

The theorem shows that under the sparse eigenvalue conditions, the advantage of group Lasso over standard Lasso is real. For standard sparsity, we take $k_0 = 1$, and the parameter estimation lower bound is $\Omega(k\ln(p/k)/n)$. Since Lasso does not take advantage of group structure, it follows that there exists a k-sparse signal for which Lasso can only achieve parameter estimation error of $\Omega(k\ln(p/k)/n)$, independent of the signal's group structure. In comparison, if this signal is (g,k) strongly group-sparse with respect to a pre-defined group structure, then the lower bound is $\Omega(g\ln(m/g)/n)$. Since the classical parametric statistics implies that the lower bound for any statistical estimator cannot be better than $\Omega(k/n)$ with k features, we obtain a lower bound of $\Omega((k+g\ln(m/g))/n)$ under strong group-sparsity (with even group size), which matches our upper bound obtained for group Lasso. This means that group Lasso achieves the optimal minimax rate for 2-norm parameter estimation up to a constant factor that depends on $\rho_+(\cdot)$ and $\rho_-(\cdot)$.

Moreover, we note that in the setting of compressive sensing, the RIP condition at sparsity k requires $\Omega(k\ln(p/k))$ random projections. In general, $\Omega(k\ln(p/k))$ random projections are also needed in order to reconstruct a k-sparse signal. This claim follows from some classical n-width results in approximation theory. However, similar results for group-sparsity is not simple to derive. Therefore we shall not include such results here.

7. Simulation Studies. We want to verify our theory by comparing group Lasso to Lasso on simulation data. For quantitative evaluation, the recovery error is defined as the relative difference in 2-norm between the estimated sparse coefficient vector β_{est} and the ground-truth sparse coefficient $\bar{\beta}$: $\|\beta_{est} - \bar{\beta}\|_2/\|\bar{\beta}\|_2$.

The regularization parameter λ in Lasso is chosen with five-fold cross validation. In group Lasso, we simply suppose the regularization parameter $\lambda_j = (\lambda\sqrt{k_j})/\sqrt{n}$ for $j = 1, 2, ..., m$. The regularization parameter λ is then chosen with five-fold cross validation. Here we set $B = 0$ in the formula $\lambda_j = O(A\sqrt{k_j} + B)$. Since the relative performance of group Lasso versus standard Lasso is similar with other values of B, in order to avoid redundancy, we do not include results with $B \neq 0$.

7.1. *Even group size.* In this set of experiments, the projection matrix X is generated by creating an $n \times p$ matrix with i.i.d. draws from a standard Gaussian distribution $N(0,1)$. For simplicity, the rows of X are normalized to unit magnitude. Zero-mean Gaussian noise with standard deviation $\sigma = 0.01$ is added to the measurements. Our task is to compare the recovery performance of Lasso and Group Lasso for these (g,k) strongly group sparse signals.

7.1.1. *With correct group structure.* In this experiment, we randomly generate (g,k) strongly group sparse coefficients with values ± 1, where $p = 512$, $k = 64$ and $g = 16$. There are 128 groups with even group size of $k_0 = 4$. Here the group structure coincides with the signal sparsity: $k = \|\bar{\beta}\|_0$.

Figure 1 shows an instance of generated sparse coefficient vector and the recovered results by Lasso and group Lasso respectively when $n = 3k = 192$. Since the sample size n is only three times the signal sparsity k, the standard Lasso does not achieve good recovery results, whereas the group Lasso achieves near perfect recovery of the original signal.

Figure 2(a) shows the effect of sample size n, where we report the averaged recover error over 100 random runs for each sample size. Group Lasso is clearly superior in this case. These results show

that the the group Lasso can achieve better recovery performance for (g, k) strongly group sparse signals with fewer measurements, which is consistent with our theory.

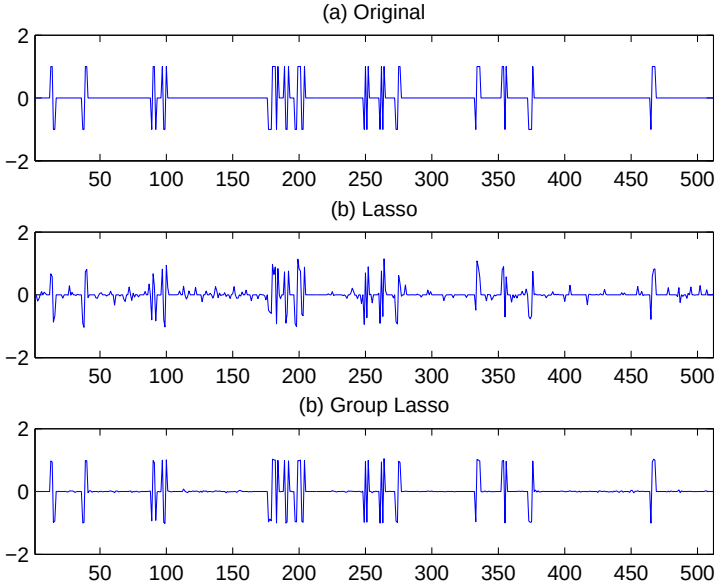


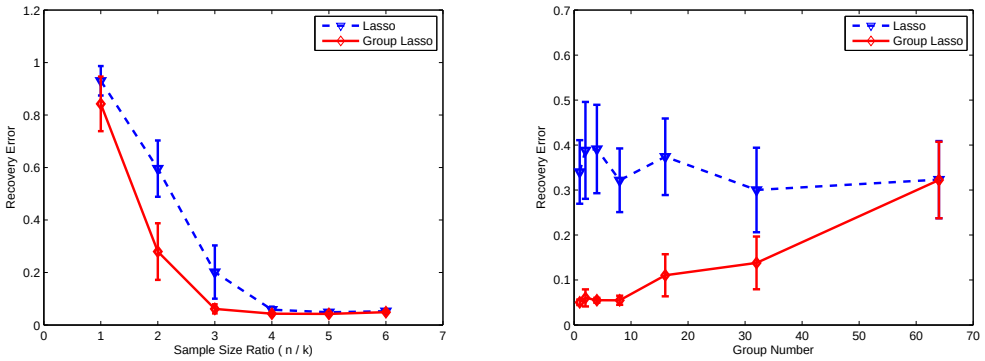
FIG 1. *Recovery results when the assumed group structure is correct. (a) Original data; (b) results with Lasso (recovery error is 0.3444); (c) results with Group Lasso (recovery error is 0.0419)*

To study the effect of the group number g (with k fixed), we set the sample size $n = 160$ and then change the group number while keeping other parameters unchanged. Figure 2(b) shows the recovery performance of the two algorithms, averaged over 100 random runs for each sample size. As expected, the recovery performance for Lasso is independent to the group number within statistical error. Moreover, the recovery results for group Lasso are significantly better when the group number g is much smaller than the sparsity $k = 64$. When $g = k$, the group Lasso becomes identical to Lasso, which is expected. This shows that the recovery performance of group Lasso degrades when g/k increases, which confirms our theory.

7.1.2. *With incorrect group structure.* In this experiment, we assume that the known group structure is not exactly the same as the sparsity of the signal (that is, $k > \|\bar{\beta}\|_0$). We randomly generate strongly group sparse coefficients with values ± 1, where $p = 512$, $\|\bar{\beta}\|_0 = 64$ and $g = 16$. In the first experiment, we let $k = 4\|\bar{\beta}\|_0$, and use $m = 32$ groups with even group size of $k_0 = 16$.

Figure 3 shows one instance of the generated sparse signal and the recovered results by Lasso and group Lasso respectively when $n = 3\|\bar{\beta}\|_0 = 192$. In this case, the standard Lasso obtains better recovery results than the group Lasso. Figure 2(a) shows the effect of sample size n, where we report the averaged recover error over 100 random runs for each sample size. The group Lasso recovery performance is clearly inferior to that of the Lasso. This shows that group Lasso fails when $k/\|\bar{\beta}\|_0$ is relatively large, which is consistent with our theory.

To study the effect of $k/\|\bar{\beta}\|_0$ on the group Lasso performance, we keep $\|\bar{\beta}\|_0$ fixed, and simply vary the group size as $k_0 = 1, 2, 4, 8, 16, 32, 64$ with $k/\|\bar{\beta}\|_0 = 1, 1, 1, 2, 4, 8, 16$. Figure 4(b) shows the performance of the two algorithms with different group sizes k_0 in terms of recovery error. It



(a) (b)

FIG 2. *Recovery performance: (a) recovery error vs. sample size ratio n/k; (b) recovery error vs. group number g*

shows that the performance of group Lasso is better when $k/\|\bar{\beta}\|_0 = 1$. However, when $k/\|\bar{\beta}\|_0 > 1$, the performance of group Lasso deteriorates.

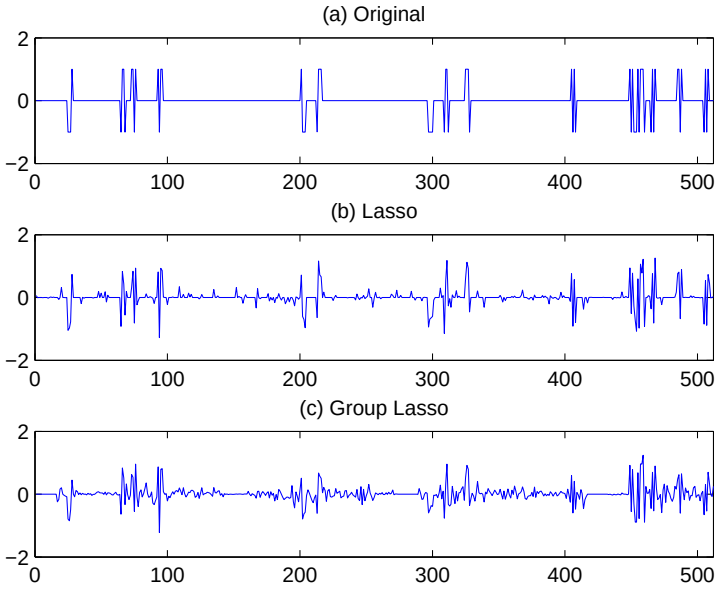
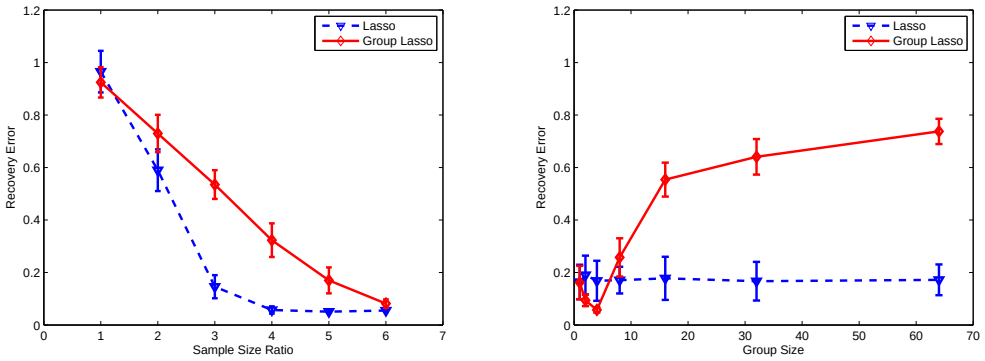


FIG 3. *Recovery results when the assumed group structure is incorrect. (a) Original data; (b) results with Lasso (recovery error is 0.3616); (c) results with Group Lasso (recovery error is 0.6688)*

7.2. *Uneven group size.* In this set of experiments, we randomly generate (g, k) strongly sparse coefficients with values ± 1, where $p = 512$, and $g = 4$. There are 64 uneven sized groups. The projection matrix X and noises are generated as in the even group size case. Our task is to compare the recovery performance of Lasso and Group Lasso for (g, k) strongly sparse signals with $\|\bar{\beta}\|_0 = k$. To reduce the variance, we run each experiment 100 times and report the average performance.

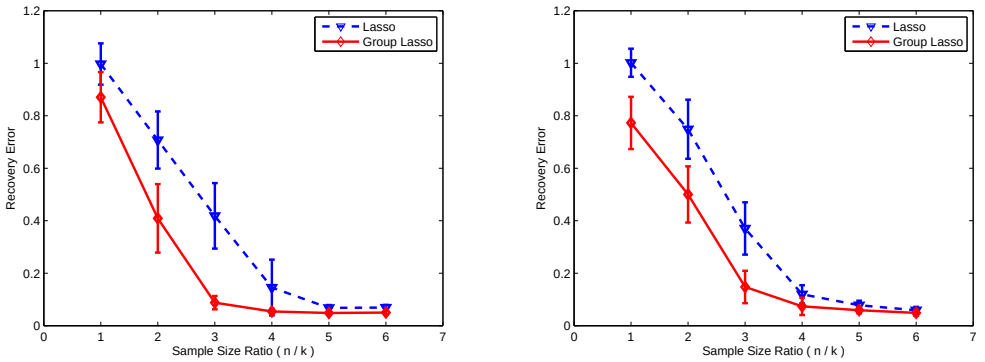
In the first experiment, the group sizes of 64 groups are randomly generated and the $g = 4$ active



(a) (b)

FIG 4. *Recovery performance: (a) recovery error vs. sample size ratio n/k; (b) recovery error vs. group size k_0*

groups are randomly extracted from these 64 groups. Figure 5(a) shows the recovery performance of Lasso and group Lasso with increasing sample size (measurements) in terms of recovery error. Similar to the case of even group size, the group Lasso obtains better recovery results than those with Lasso. It shows that the group Lasso is superior when the group sizes are randomly uneven.



(a) (b)

FIG 5. *Recovery performance: (a) g active groups have randomly uneven group sizes; (b) half of g active groups are single element groups and another half of g active groups have large group size*

As discussed after Theorem 5.1, because group Lasso favors large sized groups, if the signal is contained in small sized groups, then the performance of group Lasso can be relatively poor. In order to confirm this claim of Theorem 5.1, we consider the special case where 32 groups have large group sizes and each of the remaining 32 groups has only one element. First, we consider the case where half of $g = 4$ active groups are extracted from the single element groups and the other half of $g = 4$ active groups are extracted from the groups with large size. Figure 5(b) shows the signal recovery performance of Lasso and group Lasso. It is clear that the group Lasso performs better, but the results are not as good as those of Figure 5(a).

Moreover, Figure 6(a) shows the recovery performance of Lasso and group Lasso when all of the $g = 4$ active groups are extracted from large sized groups. We observe that the relative performance of group Lasso improves. Finally, Figure 6(b) shows the recovery performance of Lasso and group

Lasso when all of the $g = 4$ active groups are extracted from single element groups. It is obvious that the group Lasso is inferior to Lasso in this case. This confirms the prediction of Theorem 5.1 that suggests that group Lasso favors large sized groups.

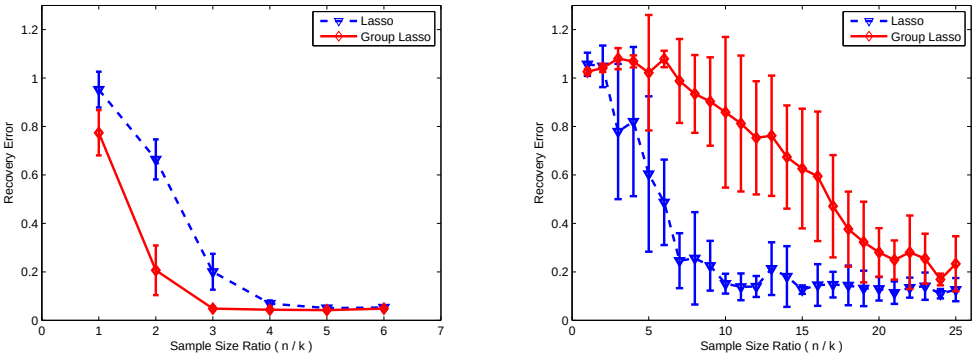


FIG 6. *Recovery performance: (a) all g active groups have large group size; (b) all g active groups are single element groups*

8. Conclusion. In this paper we introduced a concept called strong group sparsity that characterizes the signal recovery performance of group Lasso. In particular, we showed that group Lasso is superior to standard Lasso when the underlying signal is strongly group-sparse:

- Group Lasso is more robust to noise due to the stability associated with group structure.
- Group Lasso requires a smaller sample size to satisfy the sparse eigenvalue condition required in the modern sparsity analysis.

However, group Lasso can be inferior if the signal is only weakly group-sparse, or covered by groups with small sizes. Moreover, group Lasso does not perform well with overlapping groups (which is not analyzed in this paper). Better learning algorithms are needed to overcome these limitations.

References.

[1] Francis R. Bach. Consistency of the group lasso and multiple kernel learning. *JMLR*, 9:1179–1225, 2008.

[2] Peter Bickel, Yaacov Ritov, and Alexandre Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

[3] Emmanuel J. Candes and Terence Tao. Decoding by linear programming. *IEEE Trans. on Information Theory*, 51:4203–4215, 2005.

[4] S. Ji, D. Dunson, and L. Carin. Multi-task compressive sensing. *IEEE Transactions on Signal Processing*, 2008. Accepted.

[5] Vladimir Koltchinskii and Ming Yuan. Sparse recovery in large ensembles of kernel machines. In *COLT'08*, 2008.

[6] Karim Lounici, Massimiliano Pontil, Alexandre B. Tsybakov, and Sara A. van de Geer. Taking advantage of sparsity in multi-task learnng. In *COLT'09*, 2009.

[7] Yuval Nardi and Alessandro Rinaldo. On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics*, 2:605–633, 2008.

[8] G. Obozinski, M. J. Wainwright, and M. I. Jordan. Union support recovery in high-dimensional multivariate regression. Technical Report 761, UC Berkeley, 2008.

[9] G. Pisier. The volume of convex bodies and Banach space geometry. 1989. Cambridge University Press.

[10] Holger Rauhut, Karin Schnass, and Pierre Vandergheynst. Compressed sensing and redundant dictionaries. *IEEE Transactions on Information Theory*, 54(5), 2008.

[11] M. Stojnic, F. Parvaresh, and B. Hassibi. On the reconstruction of block-sparse signals with an optimal number of measurements. 2008. Preprint.

[12] D. Wipf and B. Rao. An empirical Bayesian strategy for solving the simultaneous sparse approximation problem. *IEEE Transactions on Signal Processing*, 55(7):3704–3716, 2007.
[13] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of The Royal Statistical Society Series B*, 68(1):49–67, 2006.
[14] Tong Zhang. Information theoretical upper and lower bounds for statistical estimation. *IEEE Trans. Info. Theory*, 52:1307–1321, 2006.
[15] Tong Zhang. Some sharp performance bounds for least squares regression with l_1 regularization. *Ann. Statist.*, 37(5A):2109–2144, 2009.

APPENDIX A: PROOF OF PROPOSITION 4.1

Without loss of generality, we may assume $\sigma_i > 0$ for all i (otherwise, we can still let $\sigma_i > 0$ and then just take the limit $\sigma_i \to 0$ for some i).

For notation simplicity, we remove the subscript j from the group index, and consider group G with k variables.

Let Σ be the diagonal matrix with σ_i as its diagonal elements. We can find an $n \times k$ matrix $Z = X_G(X_G^\top \Sigma X_G)^{-0.5}$, such that $Z^\top \Sigma Z = I_{k\times k}$. Let $\xi = Z^\top(\epsilon - \mathbb{E}\epsilon) \in \mathbb{R}^k$. Since $\forall v \in \mathbb{R}^n$,

$$\|(X_G^\top X_G)^{-0.5} X_G^\top v\|_2 = \|(Z^\top Z)^{-0.5} Z^\top v\|_2,$$

we have

$$\begin{aligned}\frac{\|(X_G^\top X_G)^{-0.5} X_G^\top(\epsilon - \mathbb{E}\epsilon)\|_2^2}{\xi^\top \xi} &\le \sup_{v\in\mathbb{R}^n} \frac{v^\top Z(Z^\top Z)^{-1} Z^\top v}{v^\top Z Z^\top v} \\ &= \sup_{u\in\mathbb{R}^k} \frac{u^\top (Z^\top Z)^{-1} u}{u^\top u} = \sup_{u\in\mathbb{R}^k} \frac{u^\top Z^\top \Sigma Z u}{u^\top (Z^\top Z) u} \\ &\le \sup_{v\in\mathbb{R}^n} \frac{v^\top \Sigma v}{v^\top v} \le \sigma^2.\end{aligned}$$

Therefore, we only need to show that with probability at least $1-\eta$ for all $\eta \in (0,1)$:

$$\|\xi\|_2 \le a\sqrt{k} + b\sqrt{-\ln \eta} \tag{3}$$

with $a = 1$ and $b = \sqrt{2}$.

To prove this inequality, we note that the condition $Z^\top \Sigma Z = I_{k\times k}$ means that the covariance matrix of ξ is $I_{k\times,k}$. Therefore the components of ξ are k iid Gaussians $N(0,1)$, and the distribution of $\|\xi\|_2^2$ is χ^2. Many methods have been suggested to approximate the tail probability of χ^2 distribution. For example, a well-known approximation of $\|\xi\|_2$ is the normal $N(\sqrt{k-0.5}, 0.5)$, which would imply $a = b = 1$ in (3). The weaker bound with $a = 1$ and $b = \sqrt{2}$ can be obtained through direct integration.

APPENDIX B: PROOF OF PROPOSITION 4.2

We consider the following group-greedy procedure starting with $\bar\beta^{(0)} = \bar\beta$, and form $(k^{(\ell)}, g^{(\ell)})$ strongly group sparse $\bar\beta^{(\ell)}$ as follows for $\ell = 1, 2, \ldots$

- let $r^{(\ell-1)} = X\bar\beta^{(\ell-1)} - \mathbb{E}\mathbf{y}$,
- let $j^{(\ell)} = \arg\max_j [\|(X_{G_j}^\top X_{G_j})^{-0.5} X_{G_j}^\top r^{(\ell-1)}\|_2 / \sqrt{k_j a_0^2 + b_0^2}]$,
- let $\bar\beta^{(\ell)} = \bar\beta^{(\ell-1)}$; and then reset its coefficients in group G_j as $\bar\beta_{G_j}^{(\ell)} = \bar\beta_{G_j}^{(\ell)} - (X_{G_j}^\top X_{G_j})^{-1} X_{G_j}^\top r^{(\ell-1)}$, where $j = j^{(\ell)}$.

It is not difficult to check that

$$\|r^{(\ell-1)}\|_2^2 - \|r^{(\ell)}\|_2^2 = \|(X_{G_j}^\top X_{G_j})^{-0.5} X_{G_j}^\top r^{(\ell-1)}\|_2^2,$$

$k^{(\ell)} - k^{(\ell-1)} \leq k_j$, $g^{(\ell)} - g^{(\ell-1)} \leq 1$, with $j = j^{(\ell)}$. Therefore if for all $0 \leq \ell \leq t$, we have

$$\arg\max_j \left[\|(X_{G_j}^\top X_{G_j})^{-0.5} X_{G_j}^\top r^{(\ell)}\|_2 / \sqrt{k_j a_0^2 + b_0^2} \right] \geq \sqrt{n}\Delta / \sqrt{k a_0^2 + b_0^2},$$

then by summing over $\ell = 1, \ldots, t, t+1$, we obtain

$$\begin{aligned} n\Delta^2 =& \|r^{(0)}\|_2^2 \geq \sum_{\ell=1}^{t+1} [\|r^{(\ell-1)}\|_2^2 - \|r^{(\ell)}\|_2^2] \\ \geq& n \sum_{\ell=1}^{t+1} [(k^{(\ell)} - k^{(\ell-1)}) a_0^2 + (g^{(\ell)} - g^{(\ell-1)}) b_0^2] \Delta^2 / (k a_0^2 + b_0^2) \\ \geq& n[(k^{(t+1)} - k) a_0^2 + (g^{(t+1)} - g) b_0^2] \Delta^2 / (k a_0^2 + b_0^2). \end{aligned}$$

This implies that

$$k^{(t+1)} a_0^2 + g^{(t+1)} b_0^2 \leq 2(k a_0^2 + g b_0^2).$$

Therefore if we let t be the first time $k^{(t+1)} a_0^2 + g^{(t+1)} b_0^2 > 2(k a_0^2 + g b_0^2)$, then there exists $\ell \leq t$, such that $\bar{\beta}' = \beta^{(\ell)}$ satisfies the requirement.

APPENDIX C: PROOF OF PROPOSITION 4.3

The following lemma is taken from [9].

LEMMA C.1. *Consider the unit sphere $S^{k-1} = \{x : \|x\|_2 = 1\}$ in $\mathbb{R}^k$ ($k \geq 1$). Given any $\varepsilon > 0$, there exists an ε-cover $Q \subset S^{k-1}$ such that $\min_{q \in Q} \|x - q\|_2 \leq \varepsilon$ for all $\|x\|_2 = 1$, with $|Q| \leq (1 + 2/\varepsilon)^k$.*

The following concentration result for χ^2 distribution is similar to Proposition 4.1, and can be obtained from direct integration. We skip the detailed calculation. This is where the Gaussian assumption is used in the proof. A similar result holds for sub-Gaussian random variables.

LEMMA C.2. *Let $\xi \in \mathbb{R}^n$ be a vector of n iid standard Gaussian variables: $\xi_i \sim N(0,1)$. Then $\forall \epsilon \geq 0$:*

$$\Pr\left[|\|\xi\|_2 - \sqrt{n}| \geq \epsilon \right] \leq 3e^{-\epsilon^2/2}.$$

The derivation of the following estimate employs a standard proof technique (for example, see [10]).

LEMMA C.3. *Suppose X is generated according to Proposition 4.3. For any fixed set $S \subset \{1, \ldots, p\}$ with $|S| = k$ and $0 < \delta < 1$, we have with probability exceeding $1 - 3(1 + 8/\delta)^k e^{-n\delta^2/8}$:*

$$(1 - \delta)\|\beta\|_2 \leq \frac{1}{\sqrt{n}} \|X_S \beta\|_2 \leq (1 + \delta)\|\beta\|_2 \tag{4}$$

for all $\beta \in \mathbb{R}^k$.

PROOF. It is enough to prove the conclusion in the case of $\|\beta\|_2 = 1$. According to Lemma C.1, given $\epsilon_1 > 0$, there exists a finite set $Q = \{q_i\}$ with $|Q| \leq (1 + 2/\epsilon_1)^k$ such that $\|q_i\|_2 = 1$ for all i, and $\min_i \|\beta - q_i\|_2 \leq \epsilon_1$ for all $\|\beta\|_2 = 1$.

For each i, Since elements of $\xi = X_S q_i$ are iid Gaussians $N(0, 1)$, Lemma C.2 implies that $\forall \epsilon_2 > 0$:

$$\Pr\left[|\|X_S q_i\|_2 - \sqrt{n}\|q_i\|_2| \geq \sqrt{n}\epsilon_2\right] \leq 3e^{-n\epsilon_2^2/2}.$$

Taking union bound for all $q_i \in Q$, we obtain with probability exceeding $1 - 3(1 + 2/\epsilon_1)^k e^{-n\epsilon_2^2/2}$: for all $q_i \in Q$,

$$(1 - \epsilon_2) \leq \frac{1}{\sqrt{n}}\|X_S q_i\|_2 \leq (1 + \epsilon_2).$$

Now, we define ρ as the smallest nonnegative number such that

$$\frac{1}{\sqrt{n}}\|X_S \beta\|_2 \leq (1 + \rho) \tag{5}$$

for all $\beta \in \mathbb{R}^k$ with $\|\beta\|_2 = 1$. Since for all $\|\beta\|_2 = 1$, we can find $q_i \in Q$ such that $\|\beta - q_i\|_2 \leq \epsilon_1$, we have

$$\|X_S \beta\|_2 \leq \|X_S q_i\|_2 + \|X_S(\beta - q_i)\|_2 \leq \sqrt{n}(1 + \epsilon_2 + (1 + \rho)\epsilon_1),$$

where we used (5) in the derivation. Since ρ is the smallest non-negative constant for which (5) holds, we have

$$\sqrt{n}(1 + \rho) \leq \sqrt{n}(1 + \epsilon_2 + (1 + \rho)\epsilon_1),$$

which implies that

$$\rho \leq (\epsilon_1 + \epsilon_2)/(1 - \epsilon_1).$$

Now we choose $\epsilon_1 = \delta/4$ and $\epsilon_2 = \delta/2$. Since $0 < \delta < 1$, it is easy to see that $\rho \leq \delta$. This proves the upper bound. For the lower bound, we note that for all $\|\beta\|_2 = 1$ with $\|\beta - q_i\|_2 \leq \epsilon_1$, we have

$$\|X_S \beta\|_2 \geq \|X_S q_i\|_2 - \|X_S(\beta - q_i)\|_2 \geq \sqrt{n}(1 - \epsilon_2 - (1 + \rho)\epsilon_1),$$

which leads to the desired result. □

Proof of Proposition 4.3. For each subset $S \subset \{1, \ldots, m\}$ of groups with $|S| \leq g$ and $|G_S| \leq k$, we know from C.3 that for all β such that $\text{supp}(\beta) \subset G_S$:

$$(1 - \delta)\|\beta\|_2 \leq \frac{1}{\sqrt{n}}\|X\beta\|_2 \leq (1 + \delta)\|\beta\|_2$$

with probability exceeding $1 - 3(1 + 8/\delta)^k e^{-n\delta^2/8}$.

Since the number of such groups S can be no more than $C_m^g \leq (em/g)^g$, by taking the union bound, we know that the group RIP in Equation (2) fails with probability less than

$$3(em/g)^g(1 + 8/\delta)^k e^{-n\delta^2/8} \leq e^{-t}.$$

APPENDIX D: TECHNICAL LEMMAS

The following lemmas are adapted from [15] to handle group sparsity structure. Similar techniques can be found in [2]. The first lemma is in [15].

LEMMA D.1. *Let $A = X^\top X/n$, and let I and J be non-overlapping indices in $\{1,\ldots,p\}$. We have*

$$\|A_{I,J}\|_2 \leq \sqrt{(\rho_+(I) - \rho_-(I \cup J))(\rho_+(J) - \rho_-(I \cup J))},$$

where the matrix 2-norm is defined as $\|A_{I,J}\|_2 = \sup_{\|u\|_2=\|v\|_2=1} |u^\top A_{I,J} v|$.

The next lemma uses the previous result to control the contribution of the non-signal part G^c of an error vector u to the product $u_G^\top A_{G,G^c} u_{G^c}$.

LEMMA D.2. *Given $u \in \mathbb{R}^p$ and $S \subset \{1,\ldots,m\}$. Consider $\ell \geq 1$ and define*

$$\lambda_-^2 = \min\left\{\sum_{j\in S'} \lambda_j^2 : |G_{S'}| \geq \ell\right\}.$$

Let $S_0 \subset \{1,\ldots,m\} - S$ contain indices j of largest values of $\|u_{G_j}\|_2/\lambda_j$ ($j \notin S$), and satisfies the condition $\ell \leq |G_{S_0}| < \ell + k_0$. Let $G = G_S \cup G_{S_0}$. Then

$$\sqrt{\sum_{j\notin S\cup S_0} \|u_{G_j}\|_2^2} \leq (2\lambda_-)^{-1} \sum_{j\notin S} \lambda_j \|u_{G_j}\|_2$$

and

$$\frac{1}{n}\left|\sum_{j\notin S\cup S_0} u_G^\top X_G^\top X_{G_j} u_{G_j}\right| \leq \lambda_-^{-1} \tilde{\rho}_+(|G|, \ell + k_0 - 1)\|u_G\|_2 \sum_{j\notin S} \lambda_j \|u_{G_j}\|_2,$$

where $\tilde{\rho}_+(|G|,\ell+k_0-1) = \sqrt{(\rho_+(|G|) - \rho_-(|G| + \ell + k_0 - 1))(\rho_+(\ell + k_0 - 1) - \rho_-(|G| + \ell + k_0 - 1))}$.

PROOF. Without loss of generality, we assume that $S = \{1,\ldots,g\}$, and we assume that $j > g$ is in descending order of $\|u_{G_j}\|_2/\lambda_j$. Let $S_0, S_1, \ldots$ be the first, second, etc, consecutive blocks of $j > g$, such that $\ell \leq |G_{S_k}| < \ell + k_0$ (except for the last S_k). If we let $G^k = G_{S_k}$, then:

$$\begin{aligned}
\sum_{j\notin S\cup S_0} \|u_{G_j}\|_2^2 &\leq \left[\sum_{j\notin S\cup S_0} \lambda_j \|u_{G_j}\|_2\right]\left[\max_{j\notin S\cup S_0} \|u_{G_j}\|_2/\lambda_j\right] \\
&\leq \left[\sum_{j\notin S\cup S_0} \lambda_j \|u_{G_j}\|_2\right]\left[\min_{j\in S_0} \|u_{G_j}\|_2/\lambda_j\right] \\
&\leq \left[\sum_{j\notin S\cup S_0} \lambda_j \|u_{G_j}\|_2\right]\left[\sum_{j\in S_0} \lambda_j \|u_{G_j}\|_2 / \sum_{j\in S_0} \lambda_j^2\right] \\
&\leq \frac{[\sum_{j\notin S} \lambda_j \|u_{G_j}\|_2]^2}{4\lambda_-^2}.
\end{aligned}$$

This proves the first inequality of the lemma. Note that the second inequality follows from the descending order of $\|u_{G_j}\|_2/\lambda_j$ for $j > g$. Similarly, we have

$$\begin{aligned}
\sum_{k\geq 1}\|u_{G^k}\|_2 =& \sum_{k\geq 1}\sqrt{\sum_{j\in S_k}\|u_{G_j}\|_2^2}\\
\leq& \sum_{k\geq 1}\sqrt{\sum_{j\in S_k}\lambda_j\|u_{G_j}\|_2}\sqrt{\max_{j\in S_k}\|u_{G_j}\|_2/\lambda_j}\\
\leq& \sum_{k\geq 1}\sqrt{\sum_{j\in S_k}\lambda_j\|u_{G_j}\|_2}\sqrt{\min_{j\in S_{k-1}}\|u_{G_j}\|_2/\lambda_j}\\
\leq& \sum_{k\geq 1}\sqrt{\sum_{j\in S_k}\lambda_j\|u_{G_j}\|_2}\sqrt{\sum_{j\in S_{k-1}}\lambda_j|u_{G_j}\|_2/\sum_{j\in S_{k-1}}\lambda_j^2}\\
\leq& \lambda_-^{-1}\sum_{k\geq 1}\sqrt{\sum_{j\in S_k}\lambda_j\|u_{G_j}\|_2}\sqrt{\sum_{j\in S_{k-1}}\lambda_j|u_{G_j}\|_2}\\
\leq& \lambda_-^{-1}\sum_{k\geq 1}\frac{1}{2}\left[\sum_{j\in S_k}\lambda_j\|u_{G_j}\|_2+\sum_{j\in S_{k-1}}\lambda_j|u_{G_j}\|_2\right]\\
\leq& \lambda_-^{-1}\sum_{k\geq 0}\sum_{j\in S_k}\lambda_j\|u_{G_j}\|_2 = \lambda_-^{-1}\sum_{j\notin S}\lambda_j\|u_{G_j}\|_2.
\end{aligned}$$

Therefore

$$\begin{aligned}
n^{-1}\left|\sum_{j\notin S\cup S_0}u_G^\top X_G^\top X_{G_j}u_{G_j}\right| \leq& n^{-1}\sum_{k\geq 1}|u_G^\top X_G^\top X_{G^k}u_{G^k}|\\
\leq& n^{-1}\sum_{k\geq 1}\|X_G^\top X_{G^k}\|_2\|u_{G^k}\|_2\|u_G\|_2\\
\leq& \tilde{\rho}_+(|G|,\ell+k_0-1)\|u_G\|_2\sum_{k\geq 1}\|u_{G^k}\|_2\\
\leq& \tilde{\rho}_+(|G|,\ell+k_0-1)\lambda_-^{-1}\|u_G\|_2\sum_{j\notin S}\lambda_j\|u_{G_j}\|_2.
\end{aligned}$$

Note that Lemma D.1 is used to bound $\|X_G^\top X_{G^k}\|_2$. This proves the second inequality of the lemma. □

The following lemma shows that the group L_1-norm of the group Lasso estimator's non-signal part is small (compared to the group L_1-norm of the parameter estimation error in the signal part).

LEMMA D.3. *Let* $\mathrm{supp}(\bar{\beta}) \in G_S$ *for some* $S \subset \{1,\ldots,m\}$. *Assume that for all* j:

$$\lambda_j \geq 4\rho_+(G_j)^{1/2}\|(X_{G_j}^\top X_{G_j})^{-1/2}X_{G_j}^\top\epsilon\|_2/\sqrt{n}.$$

Then the solution of (1) satisfies:

$$\sum_{j\notin S}\lambda_j\left\|\hat{\beta}_{G_j}\right\|_2 \leq 3\sum_{j\in S}\lambda_j\|\bar{\beta}_{G_j}-\hat{\beta}_{G_j}\|_2.$$

PROOF. The first order condition is:

$$2X^\top X(\hat{\beta}-\bar{\beta}) - 2X^\top \epsilon + \sum_{j=1}^m \lambda_j n v_j = 0, \tag{6}$$

where $v_j = \hat{\beta}_{G_j} / \left\|\hat{\beta}_{G_j}\right\|_2$ when $\hat{\beta}_{G_j} \neq 0$; $\|v_j\|_2 \leq 1$ and $\mathrm{supp}(v_j) \subset G_j$ when $\hat{\beta}_{G_j} = 0$. It implies that

$$\hat{\beta}^\top v_j = \|\hat{\beta}_{G_j}\|_2, \qquad |(\hat{\beta}-\bar{\beta})^\top v_j| \leq \|(\hat{\beta}-\bar{\beta})_{G_j}\|_2.$$

By multiplying both sides by $(\hat{\beta}-\bar{\beta})^\top$, we obtain

$$0 \geq -2(\hat{\beta}-\bar{\beta})^\top X^\top X(\hat{\beta}-\bar{\beta}) = -2(\hat{\beta}-\bar{\beta})^\top X^\top \epsilon + \sum_{j=1}^m \lambda_j n (\hat{\beta}-\bar{\beta})^\top v_j.$$

Therefore

$$\begin{aligned}
&\sum_{j\notin S} \lambda_j \left\|\hat{\beta}_{G_j}\right\|_2 \\
\leq &\sum_{j\in S} \lambda_j \|\bar{\beta}_{G_j} - \hat{\beta}_{G_j}\|_2 + 2(\hat{\beta}-\bar{\beta})^\top X^\top \epsilon/n \\
\leq &\sum_{j\in S} \lambda_j \|\bar{\beta}_{G_j} - \hat{\beta}_{G_j}\|_2 + 2\sum_{j=1}^m \rho_+(G_j)^{1/2} \|(\hat{\beta}-\bar{\beta})_{G_j}\|_2 \|(X_{G_j}^\top X_{G_j})^{-1/2} X_{G_j}^\top \epsilon\|_2/\sqrt{n} \\
\leq &\sum_{j\in S} \lambda_j \|\bar{\beta}_{G_j} - \hat{\beta}_{G_j}\|_2 + 0.5\sum_{j=1}^m \lambda_j \|(\hat{\beta}-\bar{\beta})_{G_j}\|_2.
\end{aligned}$$

Note that the last inequality follows from the assumption of the lemma. By simplifying the above inequality, we obtain the desired bound. □

The following lemma bounds parameter estimation error by combining the previous two lemmas.

LEMMA D.4. *Let* $\mathrm{supp}(\bar{\beta}) \in G_S$ *for some* $S \subset \{1,\ldots,m\}$. *Consider* $\ell \geq 1$ *and let* $s = |G_S| + \ell + k_0 - 1$. *Define*

$$\begin{aligned}
\lambda_-^2 =& \min\left\{\sum_{j\in S'} \lambda_j^2 : |G_{S'}| \geq \ell\right\}, \\
\tilde{\rho}_+(s, s-|G_s|) =& \sqrt{(\rho_+(s) - \rho_-(2s-|G_S|))(\rho_+(s-|G_S|) - \rho_-(2s-|G_S|))}.
\end{aligned}$$

If for all j:

$$\lambda_j \geq 4\rho_+(G_j)^{1/2} \|(X_{G_j}^\top X_{G_j})^{-1/2} X_{G_j}^\top \epsilon\|_2/\sqrt{n},$$

and

$$6\frac{\tilde{\rho}_+(s, s-|G_s|)}{\rho_-(s)} \leq \frac{\lambda_-}{\sqrt{\sum_{j\in S}\lambda_j^2}},$$

then the solution of (1) satisfies:

$$\|(\hat{\beta}-\bar{\beta})\|_2 \leq \frac{1.5}{\rho_-(s)}\left(1 + 1.5\lambda_-^{-1}\sqrt{\sum_{j\in S}\lambda_j^2}\right)\sqrt{\sum_{j\in S}\lambda_j^2}.$$

PROOF. Define S_0 as in Lemma D.2. Let $G = \cup_{j\in S\cup S_0} G_j$. By multiplying both sides of (6) by $(\hat\beta - \bar\beta)_G^\top$, we obtain

$$2(\hat\beta - \bar\beta)_G^\top X_G^\top X(\hat\beta - \bar\beta) - 2(\hat\beta - \bar\beta)_G^\top X^\top \epsilon + \sum_{j\in S\cup S_0} \lambda_j n(\hat\beta - \bar\beta)_{G_j}^\top v_j = 0.$$

Similar to the proof in Lemma D.3, we use the assumptions on λ_j to obtain:

$$4n^{-1}(\hat\beta - \bar\beta)_G^\top X_G^\top X(\hat\beta - \bar\beta) + \sum_{j\in S_0} \lambda_j \left\|\hat\beta_{G_j}\right\|_2 \le 3\sum_{j\in S} \lambda_j \|\hat\beta_{G_j} - \bar\beta_{G_j}\|_2. \tag{7}$$

Now, Lemma D.2 implies that

$$\begin{aligned}
&(\hat\beta - \bar\beta)_G^\top X_G^\top X(\hat\beta - \bar\beta)\\
\ge&(\hat\beta - \bar\beta)_G^\top X_G^\top X_G(\hat\beta - \bar\beta)_G - \tilde\rho_+(s, s - |G_S|)\lambda_-^{-1} n\|(\hat\beta - \bar\beta)_G\|_2 \sum_{j\notin S} \lambda_j \|(\hat\beta - \bar\beta)_{G_j}\|_2.
\end{aligned}$$

By applying Lemma D.3, we have

$$\begin{aligned}
&n^{-1}(\hat\beta - \bar\beta)_G^\top X_G^\top X(\hat\beta - \bar\beta)\\
\ge&\rho_-(G)\|(\hat\beta - \bar\beta)_G\|_2^2 - 3\tilde\rho_+(s, s - |G_S|)\lambda_-^{-1}\|(\hat\beta - \bar\beta)_G\|_2 \sum_{j\in S} \lambda_j \|(\hat\beta - \bar\beta)_{G_j}\|_2\\
\ge&\rho_-(G)\|(\hat\beta - \bar\beta)_G\|_2^2 - 3\tilde\rho_+(s, s - |G_S|)\lambda_-^{-1} \sqrt{\sum_{j\in S} \lambda_j^2}\|(\hat\beta - \bar\beta)_G\|_2^2\\
\ge&0.5\rho_-(G)\|(\hat\beta - \bar\beta)_G\|_2^2.
\end{aligned}$$

The assumption of the lemma is used to derive the last inequality. Now plug this inequality into (7), we have

$$\|(\hat\beta - \bar\beta)_G\|_2^2 \le 1.5\rho_-(G)^{-1} \sum_{j\in S} \lambda_j \|\hat\beta_{G_j} - \bar\beta_{G_j}\|_2 \le 1.5\rho_-(G)^{-1} \sqrt{\sum_{j\in S} \lambda_j^2}\|(\hat\beta - \bar\beta)_G\|_2.$$

This implies

$$\|(\hat\beta - \bar\beta)_G\|_2^2 \le 2.25\rho_-(G)^{-2} \sum_{j\in S} \lambda_j^2.$$

Now Lemma D.2 and Lemma D.3 imply that

$$\begin{aligned}
\|(\hat\beta - \bar\beta)\|_2^2 - \|(\hat\beta - \bar\beta)_G\|_2^2 \le&0.25\lambda_-^{-2} \left[\sum_{j\notin S} \lambda_j \|(\hat\beta - \bar\beta)_{G_j}\|_2\right]^2\\
\le&2.25\lambda_-^{-2} \left[\sum_{j\in S} \lambda_j \|(\hat\beta - \bar\beta)_{G_j}\|_2\right]^2\\
\le&2.25\lambda_-^{-2} \sum_{j\in S} \lambda_j^2 \|(\hat\beta - \bar\beta)_G\|_2^2.
\end{aligned}$$

By combining the previous two displayed inequalities, we obtain the lemma. □

APPENDIX E: PROOF OF THEOREM 5.1

Assumption 4.1 implies that with probability larger than $1-\eta$, uniformly for all groups j, we have

$$\|(X_{G_j}^\top X_{G_j})^{-0.5} X_{G_j}^\top (\epsilon - \mathbb{E}\epsilon)\|_2 \leq a\sqrt{k_j} + b\sqrt{\ln(m/\eta)}.$$

It follows that with the choice of A, B, and λ_j, $\lambda_j \geq 4\rho_+(G_j)^{1/2}\|(X_{G_j}^\top X_{G_j})^{-1/2} X_{G_j}^\top \epsilon\|_2/\sqrt{n}$ for all j. Moreover, assumptions of the theorem also imply that $\tilde{\rho}_+(s, s-|G_S|) \leq \rho_+(s) - \rho_-(2s)$, and

$$\frac{\tilde{\rho}_+(s, s-|G_S|)}{\rho_-(s)} \leq \frac{\rho_+(s) - \rho_-(2s)}{\rho_-(s)} \leq c \leq \frac{\sqrt{\ell A^2 + g_\ell B^2}}{6\sqrt{2(kA^2+gB^2)}} \leq \frac{\lambda_-}{6\sqrt{\sum_{j\in S}\lambda_j^2}}.$$

Note that we have used $\sum_{j\in S'}[A^2 k_j + B^2] \leq n\sum_{j\in S'}\lambda_j^2 \leq 2\sum_{j\in S'}[A^2 k_j + B^2]$.

Therefore the conditions of Lemma D.4 are satisfied. Its conclusion implies that

$$\begin{aligned}
\|(\hat{\beta}-\bar{\beta})\|_2 \leq & \frac{1.5}{\rho_-(s)}\left(1 + 1.5\lambda_-^{-1}\sqrt{\sum_{j\in S}\lambda_j^2}\right)\sqrt{\sum_{j\in S}\lambda_j^2} \\
\leq & \frac{1.5}{\rho_-(s)}\left(1+\frac{1}{4c}\right)\sqrt{\sum_{j\in S}\lambda_j^2} \\
\leq & \frac{1.5}{\rho_-(s)}\left(1+\frac{1}{4c}\right)\sqrt{2(A^2k + B^2 g)/n}.
\end{aligned}$$

This proves the theorem.

APPENDIX F: PROOF OF THEOREM 6.1

First, we recall the standard definition of KL divergence:

$$D_{KL}(p_{\bar{\beta}}||p_{\hat{\beta}}) = \int_{\mathbf{y}} p_{\bar{\beta}}(\mathbf{y}) \ln(p_{\bar{\beta}}(\mathbf{y})/p_{\hat{\beta}}(\mathbf{y}))d\mathbf{y}.$$

Our proof relies on the following lower bound result, with an appropriately chosen $B \subset H(g,k)$ to be determined later. Although the bound is related to other standard lower-bound techniques such as Fano's inequality, it is easier to apply for our purpose. The lemma itself is a special case of a more general lower bound theorem in [14] with uniform prior on B; it is a direct translation using our notations.

LEMMA F.1. *Consider an arbitrary finite set $B \subset \mathbb{R}^p$ and let $N = |B|$. For an arbitrary estimator $\hat{\beta}(\mathbf{y}) \in \mathbb{R}^p$ of $\bar{\beta}$ from $\mathbf{y} \sim p_{\bar{\beta}}$, we have*

$$\frac{1}{N}\sum_{\bar{\beta}\in B}\mathbb{E}_{\mathbf{y}\sim p_{\bar{\beta}}}\|X(\bar{\beta}-\hat{\beta}(\mathbf{y}))\|_2^2 \geq 0.5\sup\left\{\epsilon : \inf_{\bar{\beta}'\in\mathbb{R}^p} \ln\frac{N}{|\{\bar{\beta}\in\mathbb{R}^p : \|X(\bar{\beta}-\bar{\beta}')\|_2^2 < \epsilon\}|} \geq 2\Delta_B + \ln 4\right\},$$

where $\Delta_B = N^{-2}\sum_{\bar{\beta},\bar{\beta}'\in B} D_{KL}(p_{\bar{\beta}}||p_{\bar{\beta}'})$.

The following result relates KL-divergence and in-sample prediction error.

LEMMA F.2. *We have*

$$D_{KL}(p_{\bar{\beta}}||p_{\hat{\beta}}) = \frac{\|X(\bar{\beta}-\hat{\beta})\|_2^2}{2\sigma^2}.$$

PROOF. By definition, we have

$$\begin{aligned}D_{KL}(p_{\bar\beta}||p_{\hat\beta}) =& \int_{\mathbf{y}\in\mathbb{R}^n} p_{\bar\beta}(\mathbf{y})\ln(p_{\bar\beta}(\mathbf{y})/p_{\hat\beta}(\mathbf{y}))d\mathbf{y}\\ =& \int_{\mathbf{y}\in\mathbb{R}^n} \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\|\mathbf{y}-X\bar\beta\|_2^2/(2\sigma^2)} \times \frac{\|\mathbf{y}-X\hat\beta\|_2^2 - \|\mathbf{y}-X\bar\beta\|_2^2}{2\sigma^2} d\mathbf{y}\\ =& \frac{\|X(\hat\beta-\bar\beta)\|_2^2}{2\sigma^2},\end{aligned}$$

which implies the lemma. □

The following result is used to define a set B in order to apply Lemma F.1.

LEMMA F.3. *Given positive integer $g < m/2$. Let N be the largest number such that there exist subsets $S_1,\ldots,S_N \subset \{1,\ldots,m\}$: $|S_j| = g$ and $|S_i - S_j| \geq g$ for $i \neq j$. Then we have*

$$\ln N \geq 0.5g\ln((m-g+1)/(4g)).$$

PROOF. Let S_0 be a subset of $\{1,\ldots,m\}$ of cardinality g, chosen uniformly at random without replacement. Then for each $j = 1,\ldots,N$:

$$\begin{aligned}P[|S_0 - S_j| < g] =& \frac{\sum_{\ell>g/2} C_g^\ell C_{m-g}^{g-\ell}}{C_m^g} = \sum_{\ell>g/2} C_g^\ell \frac{g!}{(g-\ell)!}\frac{(m-g)!^2}{m!(m+\ell-2g)!}\\ \leq& \sum_{\ell>g/2} C_g^\ell g^\ell \frac{(m-g)^{g-\ell}}{(m-g+1)^g} \leq \sum_{\ell>g/2} C_g^\ell (g/(m-g+1))^{g/2}\\ \leq& 2^g (g/(m-g+1))^{g/2}.\end{aligned}$$

Since N is the largest, for any S_0, there exists j such that $|S_0 - S_j| < g$. It follows that

$$1 = P[\exists j : |S_0 - S_j| < g] \leq \sum_{j=1}^N P[|S_0 - S_j| < g] \leq N(4g/(m-g+1))^{g/2}.$$

This implies the desired bound. □

Now, we can apply Lemma F.1 with the following B. Let $\delta = \sigma\sqrt{(2n\rho_+(2g))^{-1}\ln(N/4)}$, where $N = |B|$. We choose $B \subset H(g,k)$ such that each $\bar\beta \in B$ has components $\bar\beta_j \in \{0, \delta/\sqrt{k}\}$. Moreover, we assume that any two different elements $\bar\beta, \bar\beta' \in B$ satisfy the separation condition $\|\bar\beta - \bar\beta'\|_2 \geq \delta$. Lemma F.3 implies that we can find such a set B (for each j in Lemma F.3, we define a corresponding $\bar\beta \in B$ with $\mathrm{supp}(\bar\beta) = G_{S_j}$) so that $N = |B| \geq (m-g+1)^{0.5g}/(4g)^{0.5g}$.

We observe that B has the property that for any two different elements $\bar\beta, \bar\beta' \in B$:

$$\rho_-(2g)\frac{n\delta^2}{2\sigma^2} \leq n\rho_-(2g)\frac{\|\bar\beta-\bar\beta'\|_2^2}{2\sigma^2} \leq \frac{\|X(\bar\beta-\bar\beta')\|_2^2}{2\sigma^2} \leq n\rho_+(2g)\frac{\|\bar\beta-\bar\beta'\|_2^2}{2\sigma^2} \leq \rho_+(2g)\frac{n\delta^2}{\sigma^2}.$$

Therefore, in Lemma F.1, we have

$$\Delta_B \leq \sup_{\bar\beta,\bar\beta'\in B} D_{KL}(p_{\bar\beta}||p_{\bar\beta'}) = \sup_{\bar\beta,\bar\beta'\in B} \frac{\|X(\bar\beta-\bar\beta')\|_2^2}{2\sigma^2} \leq \rho_+(2g)\frac{n\delta^2}{\sigma^2} \leq 0.5\ln(N/4).$$

This means if we pick $\epsilon = n\rho_-(2g)\delta^2/4$ in Lemma F.1, then $\forall \bar{\beta}' \in \mathbb{R}^p$: $|\{\bar{\beta} \in B : \|X(\bar{\beta} - \bar{\beta}')\|_2^2 < \epsilon\}| \leq 1$ and thus

$$\frac{1}{N}\sum_{\bar{\beta}\in B} \mathbb{E}_{\mathbf{y}\sim p_{\bar{\beta}}} \|X(\bar{\beta} - \hat{\beta}(\mathbf{y}))\|_2^2 \geq 0.5\epsilon = \sigma^2 \frac{\rho_-(2g)}{16\rho_+(2g)} \ln(N/4),$$

which proves the first lower-bound of the theorem.

Note that the estimator $\hat{\beta}(\mathbf{y})$ does not have to be in $H(g,k)$. In order to see that the first lower bound implies the second lower bound, let $\hat{\beta}'(\mathbf{y})$ be the best 2-norm approximation of $\hat{\beta}(\mathbf{y})$ in $H(g,k)$ (i.e., keeping the g groups of $\hat{\beta}(\mathbf{y})$ with largest values). Then simple algebra implies that $\|\hat{\beta}(\mathbf{y}) - \bar{\beta}\|_2^2 \geq \|\hat{\beta}'(\mathbf{y}) - \bar{\beta}\|_2^2/3 \geq (3n\rho_+(2g))^{-1}\|X(\hat{\beta}'(\mathbf{y}) - \bar{\beta})\|_2^2$. Now the first lower-bound of the theorem, applied to $\|X(\hat{\beta}'(\mathbf{y}) - \bar{\beta})\|_2^2$, implies the desired lower bound for $\|\hat{\beta}(\mathbf{y}) - \bar{\beta}\|_2^2$.

Finally, we observe that the above definition of B only considers the effect of choosing g out of m groups. We intentionally skipped the effect of estimating coefficients within any selected k features to simplify the calculation. From the proof of Lemma F.3, it is not hard to see that we can incorporate this effect and increase $\ln N$ to $\Omega(k + g\ln(m/g))$. This will give an improved lower bound.

Computer Science Department
Rutgers University
Piscataway,NJ 08854, USA
jzhuang@eden.rutgers.edu

Statistics Department
Rutgers University
Piscataway,NJ 08854, USA
tzhang@stat.rutgers.edu