Simultaneous grouping pursuit and feature selection over an undirected graph [*]

Yunzhang Zhu, Xiaotong Shen and Wei Pan

Summary

In high-dimensional regression, grouping pursuit and feature selection have their own merits while complementing each other in battling the curse of dimensionality. To seek a parsimonious model, we perform simultaneous grouping pursuit and feature selection over an arbitrary undirected graph with each node corresponding to one predictor. When the corresponding nodes are reachable from each other over the graph, regression coefficients can be grouped, whose absolute values are the same or close. This is motivated from gene network analysis, where genes tend to work in groups according to their biological functionalities. Through a nonconvex penalty, we develop a computational strategy and analyze the proposed method. Theoretical analysis indicates that the proposed method reconstructs the oracle estimator, that is, the unbiased least squares estimator given the true grouping, leading to consistent reconstruction of grouping structures and informative features, as well as to optimal parameter estimation. Simulation studies suggest that the method combines the benefit of grouping pursuit with that of feature selection, and compares favorably against its competitors in selection accuracy and predictive performance. An application to eQTL data is used to illustrate the methodology, where a network is incorporated into analysis through an undirected graph.

Key words: Network analysis, nonconvex minimization, prediction, structured data.

1 Introduction and background

For high-dimensional structured data, the dimension of parameters of interest is usually high. This occurs, for instance, in a study of identifying disease-causing genes for Parkinson's disease, where expression profiles of 22283 genes are collected from 105 patients with 55 disease versus 50 control cases; see [12] for more details. In such a situation, the number of candidate genes $p = 22283$ is much higher than the sample size $n = 105$. To battle

[*] Yunzhang Zhu is a Ph.D. candidate, School of Statistics, University of Minnesota, Minneapolis, MN 55455 (Email: zhuxx351@umn.edu). Xiaotong Shen is Professor, School of Statistics, University of Minnesota, Minneapolis, MN 55455 (Email: xshen@stat.umn.edu). Wei Pan is Professor, Division of Biostatistics, University of Minnesota, Minneapolis, MN 55455 (Email: weip@biostat.umn.edu). The authors would like to thank the editor, the associate editor and anonymous referees for helpful comments and suggestions.

the “curse of dimensionality”, one must exploit additional dependency structures from gene interactions, grouping and causal relationships. In other words, low-dimensional structures must be identified and integrated with present biological knowledge for data analysis. The central issue this article addresses is simultaneous estimation of grouping and sparseness structures, called simultaneous grouping pursuit and feature selection, for structured data over a given undirected graph.

In linear regression, we consider structured data, where dependencies among predictors are loosely modeled by connectivity of an undirected graph. Grouping is only possible when predictors are connected through paths over the graph, representing prior biological information. In this setting, we identify homogeneous subgroups of regression coefficients in absolute values, including the zero-coefficient group (feature selection). This investigation is motivated from the foregoing study, where simultaneous grouping pursuit and feature selection becomes essential over a network describing biological functionalities of genes.

Grouping pursuit has not received much attention in the literature. There is a paucity of literature for guiding practice. Two types of grouping have been investigated so far, identifying coefficients of the same values and absolute values, called Types I and II, respectively. For Type I grouping, the Fused Lasso of [17] introduces a L_1-regularization method for estimating homogeneous subgroups in a certain serial order; [15] proposes a nonconvex method for all possible homogeneous subgroups; [11] studies parameter estimation of the Fused Lasso. For Type II grouping, the OSCAR [2] suggests pairwise L_∞-penalties, and [10] employs a weighted L_γ-regularization over a graph, and [7] uses a Type I grouping method involving the pairwise sample correlations. It is Type II grouping that we shall study here. Yet, simultaneous grouping pursuit and feature selection over an arbitrary undirected graph remains under-studied. In particular, neither the interrelation between grouping pursuit and feature selection nor the impact of graph on grouping is known.

One major issue in feature selection is that highly correlated predictors impose a challenge, that is, if some predictors are included in a model then predictors that are highly correlated with them tend to be excluded in the model. This results in inaccurate feature selection. To resolve this issue, several attempts have been made. Adaptive model selection corrects the selection bias through data-driven penalty [13], and Elastic Net [26] encourages highly correlated predictors to stay together by imposing an additional ridge penalty. Relevant works can be founded in [8, 19, 21]. Despite progress, this issue remains unsettled.

Embedding feature selection into the framework of grouping pursuit, we study simultaneous grouping pursuit and feature selection through a nonconvex method. As to be seen, the method, combining the benefit of grouping pursuit with that of feature selection, outperforms either alone in predictive performance as well as accuracy of both grouping pursuit and feature selection.

This article establishes three main results. First, grouping pursuit and feature selection are complementary through the proposed method. On one hand, grouping pursuit guides feature selection to yield more accurate selection than that without it. This resolves the aforementioned issue of feature selection, because highly correlated predictors can be set to be informative as an entire group when they are grouped together through grouping pursuit. On the other hand, accuracy of grouping pursuit is enhanced through feature selection by removing the group of redundant predictors. Second, simultaneous grouping pursuit and feature selection is an integrated process, improving a model's predictive performance by reducing estimation variance while maintaining roughly the same amount of bias. Third, a graph plays a critical role in the process of grouping pursuit and feature selection. A "sufficiently precise" graph, to be defined in Definition 2, enables the proposed method to handle the least favorable situation in which informative or non-informative predictors are perfectly correlated, which is impossible for other feature selection methods.

Technically, we derive a finite-sample error bound for accuracy of grouping pursuit and feature selection of the proposed method, based on which we prove that the method consistently reconstructs the unbiased least squares estimator given the true grouping, called the *oracle estimator* in what follows, as $n, p \to \infty$. This permits roughly exponentially many predictors in $p = \exp\left(n\frac{C_{\min}}{20\sigma^2 p_0}\right)$, for grouping pursuit consistency and feature selection consistency, where σ^2 is the noise variance and C_{min} a quantity to be introduced later in (6). In addition, the optimal performance of the oracle estimator is recovered by the proposed method in parameter estimation. Most strikingly, if the graph provides a sufficient amount of information regarding grouping, then the proposed method continues to do so even when informative or non-informative predictors are perfectly correlated, whereas feature selection alone is inconsistent without grouping pursuit [14].

To demonstrate utility of the proposed method, we analyze a dataset consisting of 210 unrelated individuals in [18], where the DNA single nucleotide polymorphisms (SNPs) data are obtained from the International HapMap Project, together with the expression data from lymphoblastoid cell lines with the Illumina Sentrix Human-6 Expression BeadChip. Then we identify some SNP locations that map *cis*-acting DNA variants for a representative gene, GLT1D1.

The article is organized in six sections. Section 2 introduces the proposed method, followed by computational developments in Section 3. Section 4 is devoted to a theoretical analysis of the proposed method for oracle properties. Section 5 performs some simulations and demonstrates, in simulations, that the proposed method compares favorably against some competitors. An application to analysis of SNPs data is presented as well. Section 6 contains technical proofs.

2 Proposed method

Consider a linear model in which responses Y_i depends on a vector of p predictors:

$$\boldsymbol{Y} = (Y_1, \cdots, Y_n)^T = \boldsymbol{X}\boldsymbol{\beta}^0 + \boldsymbol{\varepsilon} = \sum_{i=1}^{p} \beta_i^0 \boldsymbol{x}_i + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_{p\times p}), \tag{1}$$

where $\boldsymbol{\beta}^0 = (\beta_1^0, \cdots, \beta_p^0)^T$ is a vector of regression coefficients, and $\boldsymbol{X}$ is independent of random error $\boldsymbol{\varepsilon}$. In (1), our goal is to estimate homogeneous subgroups of components of $\boldsymbol{\beta}$ in sizes, including the zero-coefficient group of $\boldsymbol{\beta}$, particularly when p greatly exceeds n.

In (1), each predictor corresponds to one node over a given undirected graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, describing prior knowledge concerning grouping, where $\mathcal{N} = \{1, \cdots, p\}$ is a set of nodes, and $\mathcal{E}$ consists of edges connecting nodes. If nodes i and j are reachable from each other, then predictors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ can be grouped; otherwise, they are impossible.

For simultaneous grouping pursuit and feature selection, we propose a nonconvex regularization cost function to minimize through pairwise comparisons over $\mathcal{G}$:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} g(\boldsymbol{\beta}) \equiv \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \Big(\frac{1}{2} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda_1 \boldsymbol{p}_1(\boldsymbol{\beta}) + \lambda_2 \boldsymbol{p}_2(\boldsymbol{\beta}) \Big),$$

$$\text{where } \boldsymbol{p}_1(\boldsymbol{\beta}) = \sum_{j=1}^{p} J_\tau(|\beta_j|), \quad \boldsymbol{p}_2(\boldsymbol{\beta}) = \sum_{(j,j') \in \mathcal{E}} J_\tau\big(\big||\beta_j| - |\beta_{j'}|\big|\big), \tag{2}$$

where $J_\tau(x) = \min(\frac{x}{\tau}, 1)$ is a surrogate of the L_0-function [16]; and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ and τ are nonnegative tuning parameters. In (2), grouping penalty $\boldsymbol{p}_2(\boldsymbol{\beta})$ controls only magnitudes of differences or sums of coefficients ignoring their signs over $\mathcal{G}$. Through $\boldsymbol{p}_j(\boldsymbol{\beta})$; $j = 1, 2$, simultaneous grouping pursuit and feature selection is performed by adaptive shrinkage toward unknown locations and the origin jointly, where only large coefficients and pairwise differences are shrunken.

In (2), the proposed method is designed to outperform grouping pursuit alone and feature selection alone, through tuning two regularizers. Moreover, the method is positively impacted by the prior information specified by the given graph. These aspects will be confirmed by our theoretical analysis in Section 5.

To understand the role that $\boldsymbol{p}_2(\boldsymbol{\beta})$ plays, we now examine alternative forms of penalties for grouping. Five forms of $\boldsymbol{p}_2(\boldsymbol{\beta})$ have been proposed, including Elastic Net with $\boldsymbol{p}_2(\boldsymbol{\beta}) = \sum_{j=1}^{p} \beta_j^2 = \frac{1}{2(p-1)} \sum_{j<j'} \big((\beta_j - \beta_{j'})^2 + (\beta_j + \beta_{j'})^2 \big)$, a graph version of Elastic Net [8] with $\boldsymbol{p}_2(\boldsymbol{\beta}) = \sum_{(j,j') \in \mathcal{E}} \big(\frac{\beta_j}{\sqrt{d_j}} - \frac{\beta_{j'}}{\sqrt{d_{j'}}} \big)^2$ with d_i being the number of direct neighbors of node x_i

in $\mathcal{G}$, the OSCAR with $\boldsymbol{p}_2(\boldsymbol{\beta}) = \sum_{j<j'} \max(|\beta_j|, |\beta_{j'}|)$, and a weighted penalty [10] with $\boldsymbol{p}_2(\boldsymbol{\beta}) = \sum_{(j,j')\in\mathcal{E}} 2^{1/\gamma'} \big(\frac{|\beta_j|^\gamma}{w_j} + \frac{|\beta_{j'}|^\gamma}{w_{j'}}\big)^{1/\gamma}$, $\frac{1}{\gamma} + \frac{1}{\gamma'} = 1$ and weight factor $\boldsymbol{w}$, and [7] proposes $\boldsymbol{p}_2(\boldsymbol{\beta}) = \sum_{(j,j')\in\mathcal{E}} |\beta_j - sign(\hat{\rho}_{jj'})\beta_{j'}|)$, where $sign(\hat{\rho}_{jj'})$ is the sign of the sample correlation between predictors $\boldsymbol{x}_j$ and $\boldsymbol{x}_{j'}$. Although these grouping penalties and their variants can improve accuracy of feature selection, additional estimation bias may occur due to strict convexity of $\boldsymbol{p}_2(\boldsymbol{\beta})$ as in the Lasso case [23] or due to possible graph misspecification. For instance, additional bias may be introduced by the grouping penalty in [7], when $\hat{\rho}_{jj'}$ wrongly estimates the sign of $\hat{\beta}_j\hat{\beta}_{j'}$. Despite good empirical performance, statistical properties of these methods have not been studied, regarding grouping pursuit as well as its impact on feature selection.

The proposed nonconvex grouping penalty resolves aforementioned issues of convex grouping penalties through adaptive shrinkage, because it shrinks small differences in absolute values, as opposed to large ones. As a result, estimation bias is reduced as compared to a convex penalty. This phenomenon has been noted in feature selection, where there is a trade-off between estimation bias and feature selection consistency [25]. Most critically, as to be shown later by both theoretical results and numerical examples, the nonconvex method continues to perform well even when the graph is wrongly specified, which is unlike a convex method.

3 Computation

This section develops a computational method for nonconvex minimization in (2) through difference convex (DC) programming [1]. One key idea to DC programming is decomposing the objective $g(\boldsymbol{\beta})$ into a difference of two convex functions $g(\boldsymbol{\beta}) = g_1(\boldsymbol{\beta}) - g_2(\boldsymbol{\beta})$, where

$$
\begin{aligned}
g_1(\boldsymbol{\beta}) &= \frac{1}{2}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|^2 + \frac{\lambda_1}{\tau}\sum_{j=1}^{p}|\beta_j| + \frac{\lambda_2}{\tau}\sum_{(j,j')\in\mathcal{E}}\big(|\beta_j+\beta_{j'}|+|\beta_j-\beta_{j'}|\big),\\
g_2(\boldsymbol{\beta}) &= \frac{\lambda_1}{\tau}\sum_{j=1}^{p}\max(|\beta_j|-\tau,0) + \frac{\lambda_2}{\tau}\sum_{(j,j')\in\mathcal{E}}\max(2|\beta_j|-\tau,2|\beta_{j'}|-\tau,|\beta_j|+|\beta_{j'}|).
\end{aligned}
$$

Our unconstrained DC method is then summarized as follows.

Algorithm 1:

Step 1. (Initialization) Supply an initial estimate $\hat{\boldsymbol{\beta}}^{(0)}$, for instance, $\hat{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$. Specify precision tolerance level $\epsilon > 0$.

Step 2. (Iteration) At iteration $k+1$, compute $\hat{\boldsymbol{\beta}}^{(k+1)}$ by solving subproblem

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\Big(g_1(\boldsymbol{\beta}) - \langle\boldsymbol{\beta},\nabla g_2(\hat{\boldsymbol{\beta}}^{(k)}),\rangle\Big) \tag{3}$$

where $\nabla g_2(\hat{\boldsymbol{\beta}}^{(k)})$ is a gradient vector of $g_2(\boldsymbol{\beta})$ at $\hat{\boldsymbol{\beta}}^{(k)}$ and $\langle\cdot,\cdot\rangle$ is the inner product. (Perturbation) For each j, if $|\beta_j| = \tau$ or there exists j' such that $(j,j')\in\mathcal{E}$ and $\big||\beta_j|-|\beta_{j'}|\big| = \tau$, we perturb β_j by $\beta_j \pm \epsilon^*$ to strictly decrease the cost function.

Step 3. (Stopping rule) Terminate when $g(\hat{\boldsymbol{\beta}}^{(k+1)}) - g(\hat{\boldsymbol{\beta}}^{(k)}) \le \epsilon$.

Next we present some computational properties of **Algorithm 1**.

Theorem 1 *For any $\boldsymbol{\beta}$, if $|\beta_j| = \tau$ for some j; $1 \le j \le p$, or $\big||\beta_j| - |\beta_{j'}|\big| = \tau$ for some (j,j'); $j' \ne j$, then we can perturb the β_j to strictly decrease the value of $g(\boldsymbol{\beta})$ in (2). Moreover,* **Algorithm 1** *converges exactly in finite iteration steps from any initial value.*

The finite convergence property of **Algorithm 1** is unique, due primarily to piecewise linearity of $\boldsymbol{p}_j(\boldsymbol{\beta})$; $j = 1,2$. However, other smooth non convex (differentiable) penalties may not possess this computationally attractive feature.

4 Theory

This section considers a constrained L_0-version of (2) for theoretical investigation:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} S(\boldsymbol{\beta}) = \frac{1}{2}\big\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\big\|^2, \text{ subject to}$$
$$\sum_{j=1}^{p} \mathbb{I}(|\beta_j| \neq 0) \leq C_1, \sum_{(j,j')\in\mathcal{E}} \mathbb{I}\big(\big||\beta_j| - |\beta_{j'}|\big| \neq 0\big) \leq C_2. \tag{4}$$

Moreover, we study a constrained computational surrogate of the L_0-version (4):

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} S(\boldsymbol{\beta}) = \frac{1}{2}\big\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\big\|^2, \text{ subject to}$$
$$\sum_{j=1}^{p} J_\tau(|\beta_j|) \leq C_1, \sum_{(j,j')\in\mathcal{E}} J_\tau\big(\big||\beta_j| - |\beta_{j'}|\big|\big) \leq C_2, \tag{5}$$

where the three non-negative tuning parameters (C_1, C_2, τ) control two-level adaptive shrinkage toward unknown locations and the origin. As discussed in Section 3, the DC method described in **Algorithm 1** targets at a local minimizer of (2), which can be viewed a convex relaxation of (4) or (5).

With regard to simultaneous grouping pursuit and feature selection, we will prove that global minimizers of (4) and (5) reconstruct the ideal "oracle estimator" as if the true grouping were available in advance. As a result of the reconstruction, key properties of the oracle estimator are simultaneously achieved by the proposed method.

4.1 The oracle estimator

Throughout this section, we write the $n \times p$ design matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_p)$, where $\boldsymbol{x}_i$ is the ith column of $\boldsymbol{X}$. Denote by $\lambda_{min}\big(\boldsymbol{A}\big)$ the smallest eigenvalue of a square matrix $\boldsymbol{A}$. For any vector $\boldsymbol{\beta} \in \mathbb{R}^p$, rewrite $\boldsymbol{\beta}$ as $(\boldsymbol{\beta}_{\mathcal{I}_0}, \boldsymbol{\beta}_{\mathcal{I}_1}, \cdots, \boldsymbol{\beta}_{\mathcal{I}_K})$, where $\boldsymbol{\beta}_{\mathcal{I}_0} = \boldsymbol{0}$, $\boldsymbol{\beta}_{\mathcal{I}_j} = (\alpha_j \boldsymbol{1}_{\mathcal{I}_{j1}}, -\alpha_j \boldsymbol{1}_{\mathcal{I}_{j2}})^T$; $j = 1, \cdots, K$, is a vector of length $|\mathcal{I}_j|$, with $\mathcal{I}_j = \mathcal{I}_{j1} \cup \mathcal{I}_{j2}$ and $\mathcal{I}_{j1} \cap \mathcal{I}_{j2} = \emptyset$, consisting of two disjoint subgroups with coefficients being opposite signs, where $|\mathcal{I}_{j1}| = 0$ or $|\mathcal{I}_{j2}| = 0$ is permitted. Given $\boldsymbol{\beta}$, let $\mathcal{G} = \big(\mathcal{I}_0, \mathcal{I}_1, \cdots, \mathcal{I}_K\big)$ with $\mathcal{I}_j = \mathcal{I}_{j_1} \cup \mathcal{I}_{j_2}$, which partitions $\mathcal{I} = \{1, \cdots, p\}$. Given $\mathcal{G}$, define $\boldsymbol{X}_{\mathcal{G}}$ as $\Big(\sum_{k\in\mathcal{I}_{11}} \boldsymbol{x}_k - \sum_{k\in\mathcal{I}_{12}} \boldsymbol{x}_k, \cdots, \sum_{k\in\mathcal{I}_{K1}} \boldsymbol{x}_k - \sum_{k\in\mathcal{I}_{K2}} \boldsymbol{x}_k\Big)$ to be a collapsed matrix by collapsing columns of $\boldsymbol{X}$ according to $\mathcal{G}$. Given $B = \{i_1, \cdots, i_{|B|}\} \in \mathcal{I}$, where $i_1 < \cdots < i_{|B|}$, define $\boldsymbol{X}_B$ as $(\boldsymbol{x}_{i_1}, \cdots, \boldsymbol{x}_{i_{|B|}})$ to be a submatrix of $\boldsymbol{X}$; and $\boldsymbol{\beta}_B$ to be vector $(\beta_{i_1}, \cdots, \beta_{i_{|B|}})$ for any $\boldsymbol{\beta} \in \mathbb{R}^p$.

Definition 1 (Oracle estimator) *Given the true grouping* $\mathcal{G}^0 = \big(\mathcal{I}_0^0, \mathcal{I}_1^0, \cdots, \mathcal{I}_{K_0}^0\big)$ *with* $\mathcal{I}_j^0 = \mathcal{I}_{j1}^0 \cup \mathcal{I}_{j2}^0, j = 1, \cdots, K_0$*, the oracle estimator* $\hat{\boldsymbol{\beta}}^{ol} = (\hat{\beta}_1^{ol}, \cdots, \hat{\beta}_p^{ol})^T$ *is* $\hat{\beta}_k^{ol} = \hat{\alpha}_j$ *if* $k \in \mathcal{I}_{j1}^0$, $\hat{\beta}_k^{ol} = -\hat{\alpha}_j$ *if* $k \in \mathcal{I}_{j2}^0$*;* $j = 1, \cdots, K_0$*, and* $\hat{\beta}_k^{ol} = 0$ *if* $k \in \mathcal{I}_0^0$*, where* $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \cdots, \hat{\alpha}_{K_0}) = \operatorname{argmin}_{\boldsymbol{\alpha} \in \mathbb{R}^{K_0}} \frac{1}{2} \big\| \boldsymbol{Y} - \boldsymbol{X}_{\mathcal{G}^0} \boldsymbol{\alpha} \big\|^2$.

The *oracle estimator* is the unbiased least squares estimate given the true grouping $\mathcal{G}^0$. It reduces to the oracle estimator for feature selection alone when no homogeneous groups exist for informative predictors.

4.2 Non-asymptotic probability error bounds

This section derives a non-asymptotic probability error bound for simultaneous grouping pursuit and feature selection, based on which we prove that (4) and (5) reconstruct the oracle estimator. This implies grouping pursuit consistency as well as feature selection consistency, under one simple assumption, what we call the degree-of-separation condition.

Let $\mathcal{S} = \big\{\mathcal{G} \neq \mathcal{G}^0 : C_1(\mathcal{G}) \leq p_0; C_2(\mathcal{G}, \mathcal{E}) \leq c_0\big\}$ be a constrained set defined in (4), with $C_1(\mathcal{G}) = \sum_{j=1}^p \mathbb{I}(|\beta_j| \neq 0) = |\mathcal{I} \setminus \mathcal{I}_0|$ and $C_2(\mathcal{G}, \mathcal{E}) = \sum_{(j,j') \in \mathcal{E}} \mathbb{I}\big(\big||\beta_j| - |\beta_{j'}|\big| \neq 0\big) = \sum_{0 \leq i < i' \leq K} \sum_{j \in \mathcal{I}_i, j' \in \mathcal{I}_{i'}} \mathbb{I}\big((j, j') \in \mathcal{E}\big)$, $p_0 = C_1(\mathcal{G}^0)$ and $c_0 = C_2(\mathcal{G}^0, \mathcal{E})$.

Let $A \subset \{1, \cdots, p\}$, and $A_0 = \mathcal{I} \setminus \mathcal{I}_0$ whose size $|A_0| \equiv p_0$. Define $\mathcal{S}_A = \big\{\mathcal{G} \in \mathcal{S} : \mathcal{I} \setminus \mathcal{I}_0 = A\big\}$ to be a set of groupings indexed by set A of nonzero coefficients. Let $S_i^* \equiv \max_{A:|A_0 \setminus A| = i} |\mathcal{S}_A|$ be the maximal of $\mathcal{S}_A$ satisfying $|A_0 \setminus A| = i$ and further let $S^* = \exp\big(\max_{1 \leq i \leq p_0} \frac{\log S_i^*}{i}\big)$. Finally, let $K_i^* \equiv \max_{\mathcal{G} \in \mathcal{S}: |A_0 \setminus A| = i} K(\mathcal{G})$, with $K^* = \max_{1 \leq i \leq p_0} \frac{K_i^*}{i}$.

The degree-of-separation condition is stated as follows.

$$C_{\min} \geq d_0 \frac{2 \log p + K^* + 2 \log S^*}{n} \sigma^2, \tag{6}$$

where $d_0 > 10$ is a constant, $C_{\min} \equiv \min_{\mathcal{G} \in \mathcal{S}} \frac{\|(I - \boldsymbol{P}_{\mathcal{G}}) \boldsymbol{X}_{A_0} \boldsymbol{\beta}_{A_0}^0\|^2}{|A_0 \setminus A| n}$, and $\boldsymbol{P}_{\mathcal{G}}$ is a projection onto the linear space spanned by columns of the collapsed design matrix $\boldsymbol{X}_{\mathcal{G}}$. Here $C_{\min}$ describes the least favorable situation for simultaneous grouping pursuit and feature selection, and characterizes the level of difficulty of the underlying problem.

In (6), the graph specification may have an impact on C_{min}. We introduce the notion of "consistent" graph in Definition 2. A "consistent" graph is a minimal requirement for reconstruction of the oracle estimator, where there exists a path in $\mathcal{E}$ connecting any two predictors in the same true group.

Definition 2 (*"Consistent"* graph) *An undirected graph $\big(\mathcal{N}, \mathcal{E}\big)$ is consistent with respect to the true grouping $\mathcal{G}^0 = \big(\mathcal{I}_0^0, \cdots, \mathcal{I}_{K_0}^0\big)$, if for any $j = 1, \cdots, K_0$, $\mathcal{E}|_{\mathcal{I}_j^0}$, the subgraph restricted on node set $\mathcal{I}_j^0$, is connected.*

We now present our non-asymptotic probability error bounds for global minimizers of (4) and (5) in terms of $(C_{min}, n, p, p_0, \sigma^2)$, where p_0, p may depend on n.

Theorem 2 *(L_0 method) If $\mathcal{E}$ is consistent with respect to $\mathcal{G}^0$, then for a global minimizer of (4) $\hat{\boldsymbol{\beta}}^{l_0}$ with estimated grouping $\hat{\mathcal{G}}^{l_0}$ at $(C_1, C_2) = (p_0, c_0)$,*

$$\mathbb{P}\Big(\hat{\boldsymbol{\beta}}^{l_0} \neq \hat{\boldsymbol{\beta}}^{ol}\Big) \;\; \leq \;\; \exp\Big(-\frac{n}{10\sigma^2}\big(C_{min} - 20\sigma^2\frac{\log p}{n} - 10\sigma^2\frac{K^*}{n} - 10\sigma^2\frac{\log S^*}{n}\big)\Big). \qquad (7)$$

Under (6), $\mathbb{P}\Big(\hat{\mathcal{G}}^{l_0} \neq \mathcal{G}^0\Big) \leq \mathbb{P}\Big(\hat{\boldsymbol{\beta}}^{l_0} \neq \hat{\boldsymbol{\beta}}^{ol}\Big) \to 0$, and $\frac{1}{n}\mathbb{E}\|\hat{\boldsymbol{\beta}}^{l_0} - \hat{\boldsymbol{\beta}}^0\|^2 = \big(1 + o(1)\big)\frac{1}{n}\mathbb{E}\|\hat{\boldsymbol{\beta}}^{ol} - \hat{\boldsymbol{\beta}}^0\|^2 = \big(1 + o(1)\big)\frac{K_0}{n}$, as $n, p \to \infty$.

Theorem 3 *(Surrogate method) If $\mathcal{E}$ is consistent with respect to $\mathcal{G}^0$, then for a global minimizer of (5) $\hat{\boldsymbol{\beta}}^g$ with estimated grouping $\hat{\mathcal{G}}^g$ when $(C_1, C_2) = (p_0, c_0)$; $\tau \leq 2\sigma\sqrt{\frac{\log p}{2np^3\lambda_{max}\big(\boldsymbol{X}^T\boldsymbol{X}\big)}}$,*

$$\mathbb{P}\Big(\hat{\boldsymbol{\beta}}^{g} \neq \hat{\boldsymbol{\beta}}^{ol}\Big) \;\; \leq \;\; \exp\Big(-\frac{n}{10\sigma^2}\big(C_{min} - 20\sigma^2\frac{\log p}{n} - 10\sigma^2\frac{K^*}{n} - 20\sigma^2\frac{\log S^*}{n}\big)\Big). \qquad (8)$$

Under (6), $\mathbb{P}\Big(\hat{\mathcal{G}}^{g} \neq \mathcal{G}^0\Big) \leq \mathbb{P}\Big(\hat{\boldsymbol{\beta}}^{g} \neq \hat{\boldsymbol{\beta}}^{ol}\Big) \to 0$, and $\frac{1}{n}\mathbb{E}\|\hat{\boldsymbol{\beta}}^{g} - \hat{\boldsymbol{\beta}}^0\|^2 = \big(1 + o(1)\big)\frac{1}{n}\mathbb{E}\|\hat{\boldsymbol{\beta}}^{ol} - \hat{\boldsymbol{\beta}}^0\|^2 = \big(1 + o(1)\big)\frac{K_0}{n}$, as $n, p \to \infty$.

In Theorems 2 and 3, K^* and S^* need to be computed. Next we present some bounds for (K^*, S^*).

Corollary 1 *If $\mathcal{E}$ is a fused graph, that is $\mathcal{E} = \{(i, i+1) : i = 1, \cdots, p-1\}$, then*

$$S^* \leq \sum_{i=1}^{K_0} \binom{p_0}{i} \leq p_0^{K_0+1}, \text{ and } K^* \leq K_i^* \leq K_0; i = 1, \cdots, K_0 - 1. \tag{9}$$

As a result, (2) and (3) reduce to

$$\begin{aligned} \mathbb{P}\Big(\hat{\boldsymbol{\beta}}^{l_0} \neq \hat{\boldsymbol{\beta}}^{ol}\Big) &\leq \exp\Big(-\frac{n}{10\sigma^2}\big(C_{min} - 20\sigma^2\frac{\log p}{n} - 10K_0\sigma^2\frac{\log p_0}{n})\Big), & (10) \\ \mathbb{P}\Big(\hat{\boldsymbol{\beta}}^{g} \neq \hat{\boldsymbol{\beta}}^{ol}\Big) &\leq \exp\Big(-\frac{n}{10\sigma^2}\big(C_{min} - 20\sigma^2\frac{\log p}{n} - 20K_0\sigma^2\frac{\log p_0}{n})\Big). & (11) \end{aligned}$$

For the purpose of comparing simultaneous grouping pursuit and feature selection with feature selection alone without grouping pursuit, we present (7) and (8) in a parallel manner as that in [14] for feature selection alone, where the degree of separation for feature selection alone is $C_{min}^T = \inf_{A \neq A_0, |A| \leq p_0} \Big(\big(|A_0 \setminus A| n\big)^{-1} \|(I - \boldsymbol{P}_A)\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0}^0\|^2\Big)$, which is in contrast to C_{min} in (6). Specifically, the feature selection estimators in [14] correspond to that in (4) and (5) with $(C_1, C_2) = (p_0, +\infty)$. By the necessary condition in Theorem 1 of [14], the necessary condition for feature selection alone requires that

$$C_{\min}^T \geq d_1 \frac{\log p}{n}\sigma^2, \text{ as } n, p \to +\infty \tag{12}$$

for some $d_1 > 0$. Note that the lower bound of C_{min} in (6) can be larger than that of C_{min}^T in (12). This generally means that, in terms of complexity, the problem of recovering *oracle estimator* in the sense of simultaneous grouping pursuit and feature selection is more difficult than that of feature selection alone.

To study the impact of a graph on simultaneous grouping pursuit and feature selection, we introduce another notion "sufficiently preciseness" in Definitions 3. A sufficiently precise graph is consistent, and the number of correctly connected edges for each true group is two times higher than that of wrongly connected ones, where within group connections refer to correct connections whereas between group connections are defined to be wrongly corrected.

Definition 3 (*"Sufficiently precise"* graph) *For any index sets $\mathcal{I}_j$; $j = 1, 2$, $\mathcal{I}_1 \cap \mathcal{I}_2 = \emptyset$, we define $d_{\mathcal{E}}(\mathcal{I}_1, \mathcal{I}_2) = \sum_{i \in \mathcal{I}_1; j \in \mathcal{I}_2} \mathbb{I}\big((i, j) \in \mathcal{E}\big)$ to be the number of connections between* *them over $\mathcal{E}$. A graph is sufficiently precise with respect to $\mathcal{G}^0$, if it is a consistent graph and satisfies: for any $j = 0, \cdots, K_0$, the number of within-group connections exceeds two times that of between-group connections for $\mathcal{I}_j^0$, that is, $d_{\mathcal{E}}(E, \mathcal{I}_j^0 \setminus E) > 2d_{\mathcal{E}}(E, \cup_{i \neq j}\mathcal{I}_i^0)$, for any $E \subset \mathcal{I}_j^0$.*

Lemma 1 below establishes a connection between C_{min} and C_{min}^T, and describes their behaviors in presence of perfectly correlated predictors.

Lemma 1 *(Level of difficulty) For any consistent graph,*

$$C_{min} \geq \eta^2 c_{min}, \quad C_{min}^T \geq \gamma^2 c_{min}, \text{ and } \gamma \geq \eta, \tag{13}$$

where $c_{min} = \min_{|B| \leq 2|\mathcal{I} \setminus \mathcal{I}_0^0|, \mathcal{I} \setminus \mathcal{I}_0^0 \subseteq B} \lambda_{min}\Big(n^{-1}\boldsymbol{X}_B^T \boldsymbol{X}_B\Big)$, $\eta^2 = \min\big(\min_{(j,j'):j \sim j', |\beta_j^0| \neq |\beta_{j'}^0|} \frac{1}{2}(|\beta_j^0| - |\beta_{j'}^0|)^2, \gamma^2\big)$, and $\gamma = \min_{j \in A_0} |\beta_j^0|$. If the graph is sufficiently precise, and $\mathcal{I}_i^0$ can be further partitioned into perfectly correlated subgroups $\mathcal{I}_i^0 = \{A_{i1}, \cdots, A_{in_i}\}$; $i = 1, \cdots, K_0$, then

$$C_{min} \geq c_{min}^G \min_{\boldsymbol{\alpha}, \boldsymbol{A}} \|\boldsymbol{\gamma} - \boldsymbol{A}\boldsymbol{\alpha}\| > 0, \text{ and } C_{min}^T = 0, \tag{14}$$

where $\boldsymbol{A} = (a_{ns})$ is a $N_0 \times (K_0 - 1)$ matrix with $a_{ns} \in \mathbb{Z}$, $N_0 = \sum_{i=1}^{K_0} n_i$, $\sum_{s=1}^{K_0-1} |a_{ns}| \leq |A_{im}|$, $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_{N_0})$ with $\gamma_i = |A_{im}|\beta_i^0$; $n = \sum_{j=1}^{i-1} n_k + m$; $m = 1, \cdots, n_i, i = 1, \cdots, K_0$, and

$$c_{\min}^G = \min_{B:|B \cap (\mathcal{I} \setminus \mathcal{I}_0^0)| \leq p_0, |B \cap \mathcal{I}_0^0| \leq p_0, |B \cap A_{im}| \leq 1, i=1,\cdots,K_0, m=1,\cdots,n_i} \lambda_{min}\Big(n^{-1}\boldsymbol{X}_B^T \boldsymbol{X}_B\Big).$$

Here $c_{\min}^G = c_{\min}$ in absence of perfectly correlated predictors, and $c_{\min}^G \geq c_{\min}$ otherwise.

Lemma 1 says that simultaneous grouping pursuit and feature selection is generally more difficult than feature selection alone, as described by the degree-of-separation condition for $C_{\min}$ and $C_{\min}^T$ in (6) and (12). Importantly, the impact of grouping pursuit on feature selection is evident in situations where some informative features are perfectly correlated. When a graph is sufficiently precise, simultaneous grouping and feature selection continues to work when $C_{\min} > 0$ by Lemma 1. However, any feature selection method breaks down because of non-identifiable models when $C_{\min}^T = 0$, leading to inconsistent selection in view of the necessary condition in Theorem 1 of [14]. In other words, simultaneous grouping and feature selection overcomes the difficulty of highly correlated features in feature selection.

Lemma 2 *The results in Theorems 2 and 3 continue to hold for fixed* p *with* $n \to +\infty$ *with (6) replaced by* $\lim_{n\to+\infty} nC_{min} = +\infty$.

5 Numerical examples

5.1 Simulations

This section examines operating characteristics of the proposed method and compares it against some competitors, through simulations, with regard to accuracy of grouping pursuit as well as feature selection, in addition to accuracy of parameter estimation. The competitors are OSCAR [2], *GFlasso* [7] and aGrace [8].

To measure accuracy of grouping pursuit and feature selection, we introduce four separate metrics. For the accuracy of feature selection, we use false and negative positives for feature selection, denoted by $VFP = \frac{\sum_{j=1}^{p} \mathbb{I}(\hat{\beta}_j \neq 0, \beta_j^0 = 0)}{p - p_0} \mathbb{I}(p_0 \neq p)$ and $VFN = \frac{\sum_{j=1}^{p} \mathbb{I}(\hat{\beta}_j = 0, \beta_j^0 \neq 0)}{p_0} \mathbb{I}(p_0 \neq 0)$. For grouping pursuit, we consider false and negative positives for feature selection, that is, $GFP = \frac{\sum_{(j,j') \in \mathcal{E}^0} \mathbb{I}(\hat{\beta}_j sign(\beta_j^0) \neq \hat{\beta}_{j'} sign(\beta_{j'}^0))}{|\mathcal{E}^0|} \mathbb{I}(|\mathcal{E}^0| > 0)$ and $GFN = \frac{\sum_{(j,j') \notin \mathcal{E}^0} \mathbb{I}(|\hat{\beta}_j| = |\hat{\beta}_{j'}|)}{p(p-1)/2 - |\mathcal{E}^0|} \mathbb{I}\big(|\mathcal{E}^0| < p(p-1)/2\big)$. Clearly, *VFP*, *VFN*, *GFP* and *GFN* are between $[0,1]$, with a small value indicating high accuracy for variable selection and grouping pursuit.

To measure the performance of parameter estimation for $\hat{\boldsymbol{\beta}}$, we use predictive mean squared error $PMSE(\hat{\boldsymbol{\beta}}) = \frac{\|\boldsymbol{Y}^{\text{test}} - \boldsymbol{X}^{\text{test}} \hat{\boldsymbol{\beta}}\|^2}{n^{\text{test}}}$, where $\boldsymbol{Y}^{\text{test}}, \boldsymbol{X}^{\text{test}}$ are test data and and n^{test} is the sample size of the test data. In simulations, the values of PMSE are reported, as well as values of (*VFP*, *VFN*, *GFP*, *GFN*).

Example 1 (Gene network: Large p **but small** n**).** Consider a regulatory gene network example in [8], where an entire network consists of 200 subnetworks, each with one transcription factor (TF) and its 10 regulatory target genes; see [8] for a display of the network. For this network, each predictor is generated according to $\mathcal{N}(0,1)$. To mimic a regulatory relationship, the predictor of each target gene and the TF had a bivariate normal distribution with correlation $\rho = .2, .5, .9$; conditional on the TF, the target genes are independent. In addition, $\varepsilon_i \sim \mathcal{N}(0, \sigma_e^2)$ with $\sigma_e^2 = \frac{\sum_j^p (\beta_j^0)^2}{4}$. The true regression coefficients are:

$$\boldsymbol{\beta}^0 = \big(2, \underbrace{2/\sqrt{10}, \ldots, 2/\sqrt{10}}_{10}, -2, \underbrace{-2/\sqrt{10}, \ldots, -2/\sqrt{10}}_{10},$$

$$4, \underbrace{4/\sqrt{10}, \ldots, 4/\sqrt{10}}_{10}, -4, \underbrace{-4/\sqrt{10}, \ldots, -4/\sqrt{10}}_{10}, \underbrace{0, \ldots, 0}_{p-44}\big)^T, \quad p = 2200$$

Tables 1 and 2 about here

As suggested by Table 1, the proposed method compares favorably against its competitors across all the situations, in terms of parameter estimation and accuracy of grouping pursuit and feature selection. Interestingly, *GFlasso* and aGrace perform similarly. Furthermore, all the graph-based methods performs reasonably well except Elastic Net where it does not exploit the informative graph information.

To see the impact of grouping pursuit on feature selection and vice versa, we compare the proposed method with (λ_1, λ_2) jointly against feature selection alone with $(\lambda_1, \lambda_2 = 0)$, and grouping pursuit alone with $(\lambda_1 = 0, \lambda_2)$. As indicated in Table 2, simultaneous grouping pursuit and feature selection outperforms either, as expected. The improvement in accuracy of feature selection is large, as measured by VFP, VFN, where nearly perfect reconstruction is evident. This is in contrast to accuracy of feature selection alone, where the false negative rate is high for either, in the presence of highly correlated predictors with the TF-gene correlation .9. This confirms our foregoing discussion about the impact of grouping pursuit on feature selection. Meanwhile, feature selection also enhances grouping pursuit as evident from an improvement over grouping pursuit alone.

Example 2 (Impact of erroneous edges) To understand the impact of specification of prior knowledge on a method's performance, we consider the network in **Example 1** with a varying fraction of erroneous edges adding into the network, involving different correlation structures among predictors. In set-up 1, we set the TF-gene correlation to be .9 with independent TF's. For set-up 2, the TF-TF correlation is set to be .5 so that the correlation between the informative and noisy TF's is .5. For both the set-ups, we randomly add $k = 0, 10, 100$ edges between each active and other inactive nodes. As a result, the network has $p_0 k$ more edges than that in the previous example, where p_0 is the number of active nodes. In this case, the true regression coefficients are

$$\boldsymbol{\beta}^0 = \big(\underbrace{2, \ldots, 2}_{11}, \underbrace{-2, \ldots, -2}_{11}, \underbrace{4, \ldots, 4}_{11}, \underbrace{-4, \ldots, -4}_{11}, \underbrace{0, \ldots, 0}_{p-44}\big)^T, \quad p = 2200,$$

with $\sigma_e^2 = 1$. Moreover, we use the "oracle recovery rate", defined as the percentage of times that the oracle estimator is reconstructed over 100 simulation replications. The total number of erroneous edges is 0, 440 and 4400. Results of Example 1 in presence of erroneous edges are also reported in Table 4 with correlation .9 and the average number of erroneous edges $0, 2, 10$.

Tables 3 and 4 about here

As suggested by Table 3, the proposed method performs best in terms of parameter estimation and reconstruction of the oracle estimator across all the set-ups. As a result, it yields accurate identification of grouping structures, as evident by nearly zero false positives and negatives for grouping and feature selection VFP, VFN, GFP and GFN. Interestingly, our algorithm gives a high percentage of reconstructing the oracle estimator across all the situations, indicating that it has a high chance to produce a global minimizer that is the oracle estimator with a high probability as suggested by Theorem 3. In fact, our method has a recovery rate between 100% and 85% in set-up 1, whereas it has a rate from 78% to 73% in set-up 2. Note that the recovery percentage depends on the design matrix. Overall, the level of difficulty for set-up 2 is higher, because of stronger correlations between informative and noisy predictors.

Compared to other methods, $GFlasso$ and aGrace perform slightly worse in parameter estimation but much worse in terms of oracle reconstruction. These methods seem sensitive to erroneous edges in the graph, especially in setup 2 where correlation between informative and noise variables incur bias to $GFlasso$. Finally, neither OSCAR nor Elastic Net performs well, because OSCAR is heavily biased and Elastic Net has not utilized the informative knowledge specified by the graph.

Next we investigate sensitivity of erroneous edges of the specified graphs on performance of a method. As suggested by Table 4, the oracle recovery rate dips from 67% to 35% as the average number of erroneous edges increases from 0 to 10 in Example 1. However, in Example 2, the proposed method does not seem sensitive, giving nearly unchanged PMSEs and small differences in the oracle recovery rate, where the error variance is much smaller with $\sigma_e^2 = 1$ compared to $\sigma_e^2 = 20$ in Example 1. The performance of aGrace and $GFlasso$ deteriorate significantly, as the number of erroneous edges increases from 10 to 100 for each informative node. For aGrace, it has an elevated PMSE value from 1.11 to 1.45 and 1.55 in set-ups 1 and 2. This is expected because aGrace incurs additional bias through erroneous edges. For $GFlasso$, its PMSE values increase from 1.12 to 1.16 for $k = 100$ in set-up 1, but from 1.12 to 1.36 for $k = 10$ and to 1.54 for $k = 100$. This is also expected because $GFlasso$ uses the correlations among variables as weights to alleviate bias, which can be affected by erroneous edges between correlated predictors.

Finally, based on Theorem 2, Corollary 1 and our numerical experience, in addition to the graph specification, the oracle recovery probability depends on error variance σ^2, the level of difficulty η^2, sample size n and the number of predictors p. Our numerical results suggest that our "sufficiently precise" condition for oracle recovery may be a bit conservative but is still qualitatively correct in that given the rest are the same, the less erroneous edges one have in the graph, the better chance one can recover the oracle.

Example 3 (Illustration of Corollary 1) The error bound in Corollary 1 suggests that the recovery rate depends on the number of groups K_0 and the level of difficulty η^2. We now perform a simulation study to confirm. Consider two scenarios

$$\beta^0 = \big(\underbrace{1,\ldots,1}_{p_0/K_0},\underbrace{2,\ldots,2}_{p_0/K_0},\cdots,\underbrace{K_0,\ldots,K_0}_{p_0/K_0},\underbrace{0,\ldots,0}_{p-p_0}\big)^T, \quad \eta^2 = 1/2.$$

$$\beta^0 = \big(\underbrace{3,\ldots,3}_{p_0/K_0},\underbrace{6,\ldots,6}_{p_0/K_0},\cdots,\underbrace{3K_0,\ldots,3K_0}_{p_0/K_0},\underbrace{0,\ldots,0}_{p-p_0}\big)^T, \quad \eta^2 = 9/2$$

with $p_0 = 100$, $p = 1000$ and $K_0 = 2, 5, 10, 20$. The correlation structure remains the same as in Example 1 but has within-group correlation .9 with $n = 200$.

Tables 5 about here

As suggested by Table 5, the oracle recovery rate deteriorates dramatically in both scenarios as K_0 increases from 2 to 20, as well as PMSE. Moreover, the recovery rate in the second scenario is higher with a smaller PMSE. This is in agreement with Corollary 1.

In conclusion, the proposed method performs well against its competitors in terms of parameter estimation and identifying grouping structures. In addition, it is less sensitive to the imprecise graph knowledge.

5.2 Data analysis: eQTL data

To study genetic variation, one important approach is identifying DNA sequence elements controlling gene expressions. By treating a gene's expression as a quantitative trait, one can identify DNA loci regulating the gene expression, called eQTL, which bridges the gap between genetic variants and clinical outcomes, providing biological insights into molecular mechanisms underlying complex disease missed by genome-wide association studies. Furthermore, there is increasing evidence showing that eQTLs are more likely to be disease risk loci, or can be used to boost statistical power to detect disease loci [9, 24]. Such a genomescale study utilizes DNA single nucleotide polymorphisms (SNPs) and gene expression data. The current practice of eQTL analysis is limited to simple single gene-single SNP analysis, which ignores joint effects of multiple SNPs. Here we apply the proposed method for a single gene-multiple SNP analysis.

Our focus here is mapping *cis*-acting DNA variants for a representative gene, GLT1D1. As in [18], we pre-process the data, and select SNPs lying within 500kb upstream of the transcription start site (TSS) and 500kb downstream of the transcription end site (TES) of gene GLT1D1. After monomorphic SNPs are removed, 1782 SNPs remain. As discussed in [18], the standard approach uses a univariate (or marginal) least squares (U-OLS) by regressing the expression level of GLT1D1 on each of the SNPs, coded as 0, 1 and 2, representing the count of the minor allele for the SNP. It is known that the standard approach has some potential drawbacks for data of this type. First, physically nearby SNPs tend to be correlated due to linkage disequilibrium. As a result, a true causal SNP may introduce spurious associations of its nearby SNPs with gene expressions, leading to false positives. Second, most of the genes are regulated by multiple factors or loci. This means that a univariate analysis considering only one SNP a time can be inefficient. To overcome these issues, we consider high-dimensional linear regression with the expression of gene GLT1D1 as our response and 1782 SNPs as our predictors, where simultaneous grouping pursuit and feature selection is performed, and a graph is constructed based on pairwise sample correlations exceeding a cut-off 0.6; see Figure 1 for display a subnetwork. Although this cut-off is somewhat arbitrary, it has been used to construct co-expression networks [27].

For our SNPs data, the number of SNPs $p = 1782$ is much larger than n, but biologically only a few SNPs are expected to be relevant and the correlation structure of physically nearby SNPs needs to be considered. This makes a compelling case for simultaneous grouping pursuit and feature selection to build a simpler model with higher predictive accuracy. To capture the correlation structure induced by physical locations of SNPs, a graph is constructed based on pairwise sample correlations, with a correlation stronger than 0.6 being connected; see Figure 1 for a display of the graph. Also considered is a fused type of graph, defined by a consecutive series order as in the Fused Lasso. For a comparison, we also examine the Lasso, TLP and OSCAR, where the first two perform feature selection alone and the last one does grouping pursuit and feature selection. For each method, the tuning parameter selection is achieved by randomly dividing the samples into two subset, one training set consisting of 140 samples, one tuning set consisting of 70 samples. Then, by applying the cross-validated model to the whole data set, the prediction error (PE)'s are computed, as well as the numbers of nonzero regression coefficients and homogeneous groups, based on the tuning set for the expressions of GLT1D1.

As suggested in Table 6, the proposed method not only yields a parsimonious model with the smallest mean PE but also includes one pair of physically nearby SNPs. To confirm our analysis, note that the proposed method and TLP, the proposed method with $\lambda_2 = 0$, both tend to include a subset of those SNPs having significant p-values in the marginal analysis of [18]. In contrast, the Lasso and TLP identify no grouping structure, and OSCAR is less parsimonious, including many more SNPs with less significant marginal p-values.

Our final model contains one pair of physically nearby SNPs, locations 787 and 790; see Table 7. Interestingly, adjacent locations 788 and 789 are not included in the model, because of their small pairwise sample correlations with the other nearby locations. By comparison, the fused type of graph does not seem promising, and other methods include more isolated locations. Our statistical result can be cross-validated biologically through a confirmative experiment focusing on the SNP regions near locations 787-790.

Figure 1, Tables 6 and 7 about here

6 Discussion

This article proposes a method for high-dimensional least square regression, performing simultaneous grouping pursuit and feature selection over an undirected graph describing grouping information *a priori*. Our theoretical analysis indicates that the proposed method as well as its computational surrogate reconstructs the *oracle estimator* even in difficult situations involving highly-correlated predictors when the graph is precise enough. Our numerical analysis suggests that the proposed method outperforms its competitors in accuracy of selection in addition to estimation. In particular, we have illustrated the application of our method to a single gene-multilocus eQTL analysis; its natural extension is to multiple gene-multilocus eQTL analysis, as advocated by [22, 3], though our method differs from the former two in that ours is built in a general framework of penalized regression.

In order for the proposed method to be useful, further investigation is necessary to understand the interplay between grouping pursuit and feature selection.

7 Appendix

Proof of Lemma 1: Before proceeding, we introduce some notations. Let $\widetilde{\boldsymbol{X}}$ be a matrix with column vectors $(\tilde{\boldsymbol{x}}_1, \cdots, \tilde{\boldsymbol{x}}_p)$, where $\tilde{\boldsymbol{x}}_k = \boldsymbol{x}_k$ if $k \in \big(\cup_{j=1}^{K_0} \mathcal{I}_{j1}^0\big) \cup \mathcal{I}_0^0$; $\tilde{\boldsymbol{x}}_k = -\boldsymbol{x}_k$ otherwise. In other words, $\widetilde{\boldsymbol{X}}$ is generated by flipping signs of columns of $\boldsymbol{X}$ when their indices are in $\cup_{j=1}^{K_0} \mathcal{I}_{j2}^0$. For any partition $\mathcal{G} = (\mathcal{I}_0, \mathcal{I}_1, \cdots, \mathcal{I}_K)$ with $\mathcal{I}_i = \mathcal{I}_{i1} \cup \mathcal{I}_{i2}, i = 1, \cdots, K$, let $S_{\mathcal{G}}(k) = 1$ if $k \in \big(\cup_{i=1}^K \mathcal{I}_{i1}\big) \cup \mathcal{I}_0$ and $S_{\mathcal{G}}(k) = -1$ otherwise. For $\mathcal{G} \in \mathcal{S}$, let $A = \mathcal{I} \setminus \mathcal{I}_0$, and $A_0 = \mathcal{I} \setminus \mathcal{I}_0^0$. Denote by $s_k = S_{\mathcal{G}^0}(k) S_{\mathcal{G}}(k)$; $k = 1, \cdots, p$.

To lower bound C_{min}, note that $\tilde{c}_{min} = \min_{|B| \leq 2|\mathcal{I} \setminus \mathcal{I}_0^0|, \mathcal{I} \setminus \mathcal{I}_0^0 \subseteq B} \lambda_{min}\Big(n^{-1} \widetilde{\boldsymbol{X}}_B^T \widetilde{\boldsymbol{X}}_B\Big) = c_{min}$, because $\widetilde{\boldsymbol{X}}_B^T \widetilde{\boldsymbol{X}}_B = \boldsymbol{X}_B^T \boldsymbol{X}_B$ for any B by definition. For $\mathcal{G} \in \mathcal{S}$, write $\boldsymbol{X}_{A_0} \boldsymbol{\beta}_{A_0}^0 - \boldsymbol{X}_{\mathcal{G}} \boldsymbol{\alpha}$ as

$$\sum_{i=1}^{K_0} \sum_{j=1}^{K} \sum_{k \in \mathcal{I}_i^0 \cap \mathcal{I}_j} (S_{\mathcal{G}}^0(k) \beta_k^0 - s_k \alpha_j) \tilde{\boldsymbol{x}}_k + \sum_{i=1}^{K_0} \sum_{k \in \mathcal{I}_i^0 \setminus A} S_{\mathcal{G}}^0(k) \beta_k^0 \tilde{\boldsymbol{x}}_k + \sum_{j=1}^{K} \sum_{k \in \mathcal{I}_j \setminus (\mathcal{I} \setminus \mathcal{I}_0^0)} s_k \alpha_j \tilde{\boldsymbol{x}}_k.$$

Then $\|(I-\boldsymbol{P}_{\mathcal{G}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\|^2 = \min_{\boldsymbol{\alpha}\in\mathbb{R}^K} \|\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0} - \boldsymbol{X}_{\mathcal{G}}\boldsymbol{\alpha}\|^2$ is lower bounded by

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^K}\Big(\sum_{i=1}^{K_0}\sum_{j=1}^{K}\sum_{k\in\mathcal{I}_i^0\cap\mathcal{I}_j}(S_{\mathcal{G}^0}(k)\beta_k^0 - s_k\alpha_j)^2 + \sum_{i=1}^{K_0}\sum_{k\in\mathcal{I}_i^0\setminus A}(\beta_k^0)^2 + \sum_{j=1}^K|\mathcal{I}_j\setminus A_0|\alpha_j^2\Big)c_{min}n \equiv I.$$

If $\mathcal{I}_i^0\setminus A\neq\emptyset$ for some i; $1\leq i\leq K_0$, then $I\geq nc_{min}\sum_{k\in\mathcal{I}_i^0\setminus A}(\beta_k^0)^2\geq nc_{min}\eta^2$. Otherwise, $\mathcal{I}_i^0\setminus A=\emptyset$; $i=1,\cdots,K_0$, implying that $A_0\subseteq A$. Note further that $|A|\leq|A_0|$ for $\mathcal{G}\in\mathcal{S}$ by assumption. Then $A_0=A$. Hence $I=\min_{\boldsymbol{\alpha}\in\mathbb{R}^K}\Big(\sum_{i=1}^{K_0}\sum_{j=1}^{K}\sum_{k\in\mathcal{I}_i^0\cap\mathcal{I}_j}(S_{\mathcal{G}^0}(k)\beta_k^0 - s_k\alpha_j)^2\Big)c_{\min}n$. Next two cases are examined.

For each j; $1\leq j\leq K$, (a) if there exist two indices i',i'' with $1\leq i'\neq i''\leq K_0$ such that $\mathcal{I}_{i'}^0\cap\mathcal{I}_j\neq\emptyset$ and $\mathcal{I}_{i''}^0\cap\mathcal{I}_j\neq\emptyset$, then

$$\begin{aligned} I &\geq nc_{min}\min_{\boldsymbol{\alpha}\in\mathbb{R}^K}\Big(\sum_{k\in\mathcal{I}_{i'}^0\cap\mathcal{I}_j}(S_{\mathcal{G}^0}(k)\beta_k^0 - s_k\alpha_j)^2 + \sum_{k\in\mathcal{I}_{i''}^0\cap\mathcal{I}_j}(S_{\mathcal{G}^0}(k)\beta_k^0 - s_k\alpha_j)^2\Big) \\ &\geq nc_{min}\min_{(j,j'):|\beta_j^0|\neq|\beta_{j'}^0|}\frac{1}{2}(|\beta_j^0|-|\beta_{j'}^0|)^2\geq nc_{min}\eta^2; \end{aligned}$$

otherwise, (b) there exists at most one index i^* with $1\leq i^*\leq K_0$ such that $\mathcal{I}_j\subseteq\mathcal{I}_{i^*}^0$, or $\mathcal{G}^0$ is coarser than $\mathcal{G}$. This implies that $C_2(\mathcal{G},\mathcal{E})\geq C_2(\mathcal{G}^0,\mathcal{E})=c_0$, which in turn yields that $C_2(\mathcal{G},\mathcal{E})>c_0$ when $\mathcal{G}\neq\mathcal{G}^0$ by graph consistency. This contradicts to the tuning assumption that $C_2(\mathcal{G},\mathcal{E})\leq c_0$. The bound of I in (a) thus establishes (13).

For (14), two cases are considered for any $\mathcal{G}\in\mathcal{S}$: (c) if there exists an index subset of length l^* $\{i_1,\cdots,i_{l^*}\}\subseteq\{1,\cdots,K_0\}$ and that of length (l^*-1) $\{j_1,\cdots,j_{l^*-1}\}\subseteq\{1,\cdots,K\}$ such that $\mathcal{I}_{i_1}^0\cup\cdots\cup\mathcal{I}_{i_{l^*}}^0\subseteq\mathcal{I}_{j_1}\cup\cdots\cup\mathcal{I}_{j_{l^*-1}}$ for some l^* with $1\leq l^*\leq K$; otherwise, (d) for any l with $1\leq l\leq K$, $\{i_1,\cdots,i_l\}$, $(\mathcal{I}_{i_1}^0\cup\cdots\cup\mathcal{I}_{i_l}^0)\nsubseteq(\mathcal{I}_{j_1}\cup\cdots\cup\mathcal{I}_{j_k})\neq\emptyset$ for $k<l$.

For (c), let $\mathcal{J}=(A\cup A_0)\setminus(\mathcal{I}_{i_1}^0\cup\cdots\cup\mathcal{I}_{i_{l^*}}^0)$, $L(\boldsymbol{X}_{\mathcal{J}})=\boldsymbol{X}_{\mathcal{J}}\boldsymbol{\beta}^0_{\mathcal{J}}-\sum_{k\in\mathcal{J}}\big(\sum_{j=1}^K\alpha_j\mathbb{I}(k\in\mathcal{I}_j)\big)\boldsymbol{x}_k$, $\boldsymbol{\alpha}=(\alpha_{j_1},\cdots,\alpha_{j_{l^*-1}})\in\mathbb{R}^{l^*-1}$ and $a_{ts}^{(m)}=\sum_{k\in A_{i_tm}}\pm\mathbb{I}(k\in\mathcal{I}_{j_s})$; $t=1,\cdots,l^*$, $s=1,\cdots,l^*-1$, $m=1,\cdots,n_t$. For any $\mathcal{G}\in\mathcal{S}$, $\|(I-\boldsymbol{P}_{\mathcal{G}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\|^2$ is lower bounded by

$$
\begin{aligned}
& \min_{\boldsymbol{\alpha}} \Big\| \sum_{t=1}^{l^*} \beta_{i_t}^0 \sum_{k \in \mathcal{I}_{i_t}^0} \tilde{\boldsymbol{x}}_k - \sum_{k \in \mathcal{I}_{i_1}^0 \cup \cdots \cup \mathcal{I}_{i_{l^*}}^0} \tilde{\boldsymbol{x}}_k \sum_{s=1}^{l^*-1} (\pm \alpha_{j_s}) \mathbb{I}(k \in \mathcal{I}_{j_s}) + L(\boldsymbol{X}_{\mathcal{J}}) \Big\|^2 \\
\geq \;& \min_{\boldsymbol{\alpha}} \Big\| \sum_{t=1}^{l^*} \sum_{m=1}^{n_t} |A_{i_t m}| \beta_{i_t}^0 \boldsymbol{z}_{i_t m} - \sum_{t=1}^{l} \sum_{m=1}^{n_t} \Big(\boldsymbol{z}_{i_t m} \big(\sum_{s=1}^{l^*-1} \alpha_{j_s} \sum_{k \in A_{i_t m}} \pm \mathbb{I}(k \in \mathcal{I}_{j_s}) \big) \Big) + L(\boldsymbol{X}_{\mathcal{J}}) \Big\|^2 \\
\geq \;& \min_{\boldsymbol{\alpha}, a_{ts}^{(m)}} \Big\| \sum_{t=1}^{l^*} \sum_{m=1}^{n_t} (|A_{i_t m}| \beta_{i_t}^0 - \sum_{s=1}^{l^*-1} \alpha_{j_s} a_{ts}^{(m)}) \boldsymbol{z}_{i_t m} + L(\boldsymbol{X}_{\mathcal{J}}) \Big\|^2 \\
\geq \;& n c_{min}^G \min_{\boldsymbol{\alpha}, a_{ts}^{(m)}} \sum_{t=1}^{l^*} \sum_{m=1}^{n_t} (|A_{i_t m}| \beta_{i_t}^0 - \sum_{s=1}^{l^*-1} \alpha_{j_s} a_{ts}^{(m)})^2 \geq n c_{min}^G \min_{\boldsymbol{\alpha}, \boldsymbol{A}} \| \boldsymbol{\gamma} - \boldsymbol{A} \boldsymbol{\alpha} \|^2,
\end{aligned}
$$

implying (14).

For (d), we will show that it does not occur under sufficient preciseness. Suppose that (d) does. By Hall's Theorem [4], there exists a matching of $\{\mathcal{I}_1^0 \cup \cdots \cup \mathcal{I}_{K_0}^0\}$ into $\{\mathcal{I}_1 \cup \cdots \cup \mathcal{I}_K\}$. Without loss of generality, we may assume $\mathcal{I}_1 \cap \mathcal{I}_1^0 \neq \emptyset, \cdots, \mathcal{I}_{K_0} \cap \mathcal{I}_{K_0}^0 \neq \emptyset$. For $D \subseteq \mathcal{I} = \{1, \cdots, p\}$, let $d_{\mathcal{E}}(D) = \sum_{i,i' \in D; i < i'} \mathbb{I}\big((i,i') \in \mathcal{E}\big)$, and $\mathcal{I}_{ij} = \mathcal{I}_i^0 \cap \mathcal{I}_j$. Then

$$
\begin{aligned}
2\big(C_2(\mathcal{G}, \mathcal{E}) - C_2(\mathcal{G}^0, \mathcal{E})\big) &= 2\big(d_{\mathcal{E}}(\mathcal{I}) - \sum_{j=0}^{K} d_{\mathcal{E}}(\mathcal{I}_j)\big) - 2\big(d_{\mathcal{E}}(\mathcal{I}) - \sum_{i=0}^{K_0} d_{\mathcal{E}}(\mathcal{I}_i^0)\big) \\
= \;& \Big(\sum_{i=0}^{K_0} \sum_{j=0}^{K} d_{\mathcal{E}}(\mathcal{I}_{ij}, \mathcal{I}_i^0 \setminus \mathcal{I}_{ij}) \Big) - \Big(\sum_{j=0}^{K} \sum_{i=0}^{K_0} d_{\mathcal{E}}(\mathcal{I}_{ij}, \mathcal{I}_j \setminus \mathcal{I}_{ij}) \Big). \qquad (15)
\end{aligned}
$$

To simplify (15), consider two cases: (e) if $\mathcal{I}_i^0 \not\subseteq \mathcal{I}_i$ thus $\mathcal{I}_i^0 \backslash \mathcal{I}_{ii} \neq \emptyset$ for any i; $0 \leq i \leq K_0$; otherwise (f) the set $\mathcal{I}_* \equiv \{i : \mathcal{I}_i^0 \subseteq \mathcal{I}_i\}$ is nonempty.

For (e), note that $\mathcal{I}_{ii} \neq \emptyset$, hence that $\mathcal{I}_i^0 \backslash \mathcal{I}_{ij} \neq \emptyset$ for any $i \neq j$; $0 \leq i \leq K_0, 0 \leq j \leq K$. By sufficiently preciseness, $d_{\mathcal{E}}(\mathcal{I}_{ij}, \mathcal{I}_i^0 \backslash \mathcal{I}_{ij}) > 2 d_{\mathcal{E}}(\mathcal{I}_{ij}, \mathcal{I}_j \backslash \mathcal{I}_{ij}) > d_{\mathcal{E}}(\mathcal{I}_{ij}, \mathcal{I}_j \backslash \mathcal{I}_{ij})$; $i = 0, \cdots, K_0, j = 0, \cdots, K$, implying that $C_2(\mathcal{G}, \mathcal{E}) > C_2(\mathcal{G}^0, \mathcal{E}) = c_0$ in (15), which contradicts to the tuning assumption that $C_2(\mathcal{G}, \mathcal{E}) \leq c_0$.

For (f), let $\mathcal{I}_*^1 = \{0, 1, \cdots, K_0\} \setminus \mathcal{I}_*$ and $\mathcal{I}_*^2 = \{0, 1, \cdots, K\} \setminus \mathcal{I}_*$. Now, $\mathcal{I}_i^0 \subseteq \mathcal{I}_i, i \in \mathcal{I}_*$. Since $|\cup_{i=1}^{K_0} \mathcal{I}_i^0| \geq |\cup_{j=1}^{K} \mathcal{I}_j|$, $1 \leq |\mathcal{I}_*| < K_0$. Hence $\mathcal{I}_{ij} = \emptyset, i \in \mathcal{I}_*, j \neq i$ and $\mathcal{I}_i^0 \backslash \mathcal{I}_{ii} = \emptyset, i \in \mathcal{I}_*$. Now (15) becomes

$$
\sum_{i \in \mathcal{I}_*^1} \sum_{j=0}^{K} d_{\mathcal{E}}(\mathcal{I}_{ij}, \mathcal{I}_i^0 \setminus \mathcal{I}_{ij}) - \sum_{j \in \mathcal{I}_*} d_{\mathcal{E}}(\mathcal{I}_j \setminus \mathcal{I}_j^0, \mathcal{I}_j^0) - \sum_{i \in \mathcal{I}_*^1} \sum_{j=0}^{K} d_{\mathcal{E}}(\mathcal{I}_{ij}, \mathcal{I}_j \setminus \mathcal{I}_{ij}). \qquad (16)
$$

By sufficiently preciseness, $\sum_{i \in \mathcal{I}_*^1} \sum_{j=0}^K d_{\mathcal{E}}(\mathcal{I}_{ij}, \mathcal{I}_i^0 \setminus \mathcal{I}_{ij}) > 2 \sum_{i \in \mathcal{I}_*^1} \sum_{j=0}^K d_{\mathcal{E}}(\mathcal{I}_{ij}, \mathcal{I}_j \setminus \mathcal{I}_{ij})$. This together with $\sum_{j \in \mathcal{I}_*} d_{\mathcal{E}}(\mathcal{I}_j \setminus \mathcal{I}_j^0, \mathcal{I}_j^0) \leq \sum_{j \in \mathcal{I}_*} \sum_{i \in \mathcal{I}_*^1} d_{\mathcal{E}}(\mathcal{I}_{ij}, \mathcal{I}_j^0) \leq \sum_{i \in \mathcal{I}_*^1} \sum_{j=0}^K d_{\mathcal{E}}(\mathcal{I}_{ij}, \mathcal{I}_j \setminus \mathcal{I}_{ij})$ yields that $C_2(\mathcal{G}, \mathcal{E}) > C_2(\mathcal{G}^0, \mathcal{E}) = c_0$ in (16), which is impossible as before. Consequently (f) does not occur under sufficiently preciseness. This completes the proof.

Proof of Theorem 1: The proof is similar to the convergence proof in [14]. Hence it will be omitted.

Proof of Theorem 3: Before proceeding, we introduce some notations. Define $\widehat{\mathcal{G}} = \big(\hat{\mathcal{I}}_0, \hat{\mathcal{I}}_1, \cdots, \hat{\mathcal{I}}_K\big)$ with $\hat{\mathcal{I}}_i = \hat{\mathcal{I}}_{i1} \cup \hat{\mathcal{I}}_{i2}$; $i = 1, \cdots, K$ as follows. First, $|\hat{\beta}_j^g|$'s are ordered by their values. Second, check any two consecutive ordered values of $|\hat{\beta}_j^g|$, and set j_1 and j_2 to be in one group if $||\hat{\beta}_{j_1}^g| - |\hat{\beta}_{j_2}^g|| \leq \tau$. Third, let $\hat{\mathcal{I}}_0$ be the group whose range contains zero, and $\hat{\mathcal{I}}_0 = \emptyset$ otherwise. Finally, for each $1 \leq i \leq K$, partition $\hat{\mathcal{I}}_i$ into $\hat{\mathcal{I}}_{i1}$ and $\hat{\mathcal{I}}_{i2}$ by grouping components $\hat{\beta}_j$'s of the same sign together. Consequently, (i) $\max_{j \in \hat{\mathcal{I}}_0} |\hat{\boldsymbol{\beta}}_j^g| \leq \tau$; (ii) $\big||\hat{\beta}_{j_1}^g| - |\hat{\beta}_{j_2}^g|\big| \leq \tau$ for any $1 \leq j_1, j_2 \leq K$; (iii) $\hat{\beta}_{j_1}^g \hat{\beta}_{j_2}^g < 0$ for any $j_1 \in \mathcal{I}_{i1}^1, j_2 \in \hat{\mathcal{I}}_{i2}$; $i = 1, \cdots, K$.

Next we show that $\hat{\boldsymbol{\beta}}^g = \hat{\boldsymbol{\beta}}^{ol}$ when $\widehat{\mathcal{G}} = \mathcal{G}^0$. Now $p_1 = \mathcal{I} \setminus \hat{\mathcal{I}}_0^0 = p_0$. By (5), $\frac{1}{\tau} \sum_{j \in \hat{\mathcal{I}}_0} |\hat{\beta}_j^g| + p_1 \leq p_0$, with $p_0 = p_1$, yields that $\hat{\beta}_j^g = 0$; $j \in \mathcal{I}_0^1$. In addition, the second constraint of (5) implies $\sum_{i=1}^K \sum_{j, j' \in \mathcal{I}_i, (j,j') \in \mathcal{E}} \frac{\big||\hat{\beta}_j^g| - |\hat{\beta}_{j'}^g|\big|}{\tau} \leq 0$, yielding that $\hat{\beta}_j^g = -\hat{\beta}_{j'}^g$; $j \in \hat{\mathcal{I}}_{i1}$, $j' \in \hat{\mathcal{I}}_{i2}, (j, j') \in \mathcal{E}$ and $\hat{\beta}_{j_1}^g = \hat{\beta}_{j_1}^g$; $j_1, j_2 \in \hat{\mathcal{I}}_{i1}$ or $j_1, j_2 \in \hat{\mathcal{I}}_{i2}, (j_1, j_2) \in \mathcal{E}$. By graph consistency of $\mathcal{E}$, $\mathcal{E}|_{\hat{\mathcal{I}}_i}$ is connected, implying that $\hat{\beta}_j^g = -\hat{\beta}_{j'}^g$; $j \in \hat{\mathcal{I}}_{i1}$, $j' \in \hat{\mathcal{I}}_{i2}$ and $\hat{\beta}_{j_1}^g = \hat{\beta}_{j_1}^g$; $j_1, j_2 \in \hat{\mathcal{I}}_{i1} \cup \hat{\mathcal{I}}_{i2}$. This further implies that $\hat{\boldsymbol{\beta}}^g = \hat{\boldsymbol{\beta}}^{ol}$, hence that $\{\widehat{\mathcal{G}} = \mathcal{G}^0\} \subseteq \{\hat{\boldsymbol{\beta}}^g = \hat{\boldsymbol{\beta}}^{ol}\}$. Thus

$$\mathbb{P}(\hat{\boldsymbol{\beta}}^g \neq \hat{\boldsymbol{\beta}}^{ol}, \widehat{\mathcal{G}} \neq \mathcal{G}^0) \leq \mathbb{P}\Big(S(\hat{\boldsymbol{\beta}}^g) - S(\hat{\boldsymbol{\beta}}^{ol}) \leq 0, \widehat{\mathcal{G}} \neq \mathcal{G}^0\Big) \equiv I, \tag{17}$$

To bound I, we first obtain lower bounds of $S(\hat{\boldsymbol{\beta}}^g) - S(\hat{\boldsymbol{\beta}}^{ol})$. Let $\bar{\boldsymbol{\beta}} = (\bar{\beta}_1, \cdots, \bar{\beta}_p)$, with $\bar{\beta}_j = sign(\hat{\beta}_j^g) \frac{\sum_{j' \in \hat{\mathcal{I}}_i} |\hat{\beta}_{j'}^g|}{|\hat{\mathcal{I}}_i|}$; $j \in \hat{\mathcal{I}}_i$, $i = 1, \cdots, K$ and $\bar{\beta}_j = 0$; $j \in \hat{\mathcal{I}}_0$. Then $|\bar{\beta}_j - \hat{\beta}_j^g| \leq (|\hat{\mathcal{I}}_i| - 1)\tau$ for $j \in \hat{\mathcal{I}}_i$; $i = 0, \cdots, K$. Note that

$$
\begin{aligned}
&\|\boldsymbol{Y}-\boldsymbol{X}\bar{\boldsymbol{\beta}}\|^2 \ge \|(\boldsymbol{I}-\boldsymbol{P}_{\widehat{\mathcal{G}}})\boldsymbol{Y}\|^2 = \|(\boldsymbol{I}-\boldsymbol{P}_{\widehat{\mathcal{G}}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0} + (\boldsymbol{I}-\boldsymbol{P}_{\widehat{\mathcal{G}}^g})\boldsymbol{\epsilon}\|^2,\\
&\|\boldsymbol{X}\bar{\boldsymbol{\beta}}-\boldsymbol{X}\hat{\boldsymbol{\beta}}^g\|^2 \le \lambda_{max}\big(\boldsymbol{X}^T\boldsymbol{X}\big)\|\bar{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}^g\|^2 \le \lambda_{max}\big(\boldsymbol{X}^T\boldsymbol{X}\big)\tau^2\sum_{i=0}^{K}(|\hat{\mathcal{I}}_i|-1)^2|\hat{\mathcal{I}}_i|\\
&\le \lambda_{max}\big(\boldsymbol{X}^T\boldsymbol{X}\big)p^3\tau^2.
\end{aligned}
$$

Using the inequality $\|U+V\|^2 \ge \frac{a-1}{a}\|U\|^2-(a-1)\|V\|^2$ for any real vectors $U,V\in\mathbb{R}^p$ and $a>0$, we have

$$
\begin{aligned}
&S(\hat{\boldsymbol{\beta}}^g)=\frac{1}{2}\big\|\boldsymbol{Y}-\boldsymbol{X}\bar{\boldsymbol{\beta}}+\boldsymbol{X}\bar{\boldsymbol{\beta}}-\boldsymbol{X}\hat{\boldsymbol{\beta}}^g\big\|^2 \ge \frac{a-1}{2a}\|\boldsymbol{Y}-\boldsymbol{X}\bar{\boldsymbol{\beta}}\|^2-\frac{a-1}{2}\|\boldsymbol{X}\bar{\boldsymbol{\beta}}-\boldsymbol{X}\hat{\boldsymbol{\beta}}^g\|\\
&\ge \frac{a-1}{2a}\|(\boldsymbol{I}-\boldsymbol{P}_{\widehat{\mathcal{G}}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}+(\boldsymbol{I}-\boldsymbol{P}_{\widehat{\mathcal{G}}})\boldsymbol{\epsilon}\|^2-\frac{(a-1)\lambda_{max}\big(\boldsymbol{X}^T\boldsymbol{X}\big)p^3\tau^2}{2}\\
&\ge \frac{a-1}{2a}\Big(\|(\boldsymbol{I}-\boldsymbol{P}_{\widehat{\mathcal{G}}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\|^2+\|(\boldsymbol{I}-\boldsymbol{P}_{\widehat{\mathcal{G}}})\boldsymbol{\epsilon}\|^2+2\boldsymbol{\epsilon}^T(\boldsymbol{I}-\boldsymbol{P}_{\widehat{\mathcal{G}}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}-\frac{\lambda}{a-1}\Big),
\end{aligned}
$$

where $\lambda=a(a-1)\lambda_{max}\big(\boldsymbol{X}^T\boldsymbol{X}\big)p^3\tau^2$. This yields that

$$
\begin{aligned}
&2a\Big(S(\hat{\boldsymbol{\beta}}^g)-S(\hat{\boldsymbol{\beta}}^{ol})\Big)=2a\Big(S(\hat{\boldsymbol{\beta}}^g)-\frac{1}{2}\|(I-\boldsymbol{P}_{\mathcal{G}^0})\boldsymbol{\epsilon}\|^2\Big)\\
&\ge 2(a-1)\boldsymbol{\epsilon}^T(\boldsymbol{I}-\boldsymbol{P}_{\widehat{\mathcal{G}}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}+(a-1)\|(\boldsymbol{I}-\boldsymbol{P}_{\widehat{\mathcal{G}}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\|^2-\boldsymbol{\epsilon}^T\Big(\boldsymbol{I}+(a-1)\boldsymbol{P}_{\widehat{\mathcal{G}}}\Big)\boldsymbol{\epsilon}\\
&-\lambda\equiv -L(\widehat{\mathcal{G}})+b(\widehat{\mathcal{G}}),
\end{aligned}
$$

where $L(\mathcal{G})\equiv\big(\boldsymbol{\epsilon}-(a-1)(\boldsymbol{I}-\boldsymbol{P}_{\mathcal{G}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\big)^T\Big(\boldsymbol{I}+(a-1)\boldsymbol{P}_{\mathcal{G}}\Big)\big(\boldsymbol{\epsilon}-(a-1)(\boldsymbol{I}-\boldsymbol{P}_{\mathcal{G}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\big)$, $b(\mathcal{G})=a(a-1)\|(\boldsymbol{I}-\boldsymbol{P}_{\mathcal{G}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\|^2-\lambda$. Note that $L(\mathcal{G})=L_1(\mathcal{G})+L_2(\mathcal{G})$, where $L_1(\mathcal{G})=\big(\boldsymbol{\epsilon}-(a-1)(\boldsymbol{I}-\boldsymbol{P}_{\mathcal{G}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\big)^T\Big(\boldsymbol{I}-\boldsymbol{P}_{\mathcal{G}}\Big)\big(\boldsymbol{\epsilon}-(a-1)(\boldsymbol{I}-\boldsymbol{P}_{\mathcal{G}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\big)$, which follows $\chi^2_{k,\Lambda}$ of freedom $n-K$ and non-central parameter $\Lambda=(a-1)^2\sigma^{-2}\|(\boldsymbol{I}-\boldsymbol{P}_{\mathcal{G}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\|^2\ge(a-1)^2nC_{min}/\sigma^2$ and $L_2(\mathcal{G})=a\boldsymbol{\epsilon}^T\boldsymbol{P}_{\mathcal{G}}\boldsymbol{\epsilon}$ is independent of $L_1(\mathcal{G})$.

Recall that $\mathcal{S}=\big\{\mathcal{G}\neq\mathcal{G}^0: C_1(\mathcal{G})\le p_0; C_2(\mathcal{G},\mathcal{E})\le c_0\big\}$. Let $\widehat{A}=\mathcal{I}\setminus\widehat{\mathcal{I}}_0$. By Markov's inequality with any $t<\frac{1}{2a}$, it follows from (17) that

$$\begin{aligned}
I &\leq \sum_{A:|A_0\backslash A|=i}\sum_{\mathcal{G}\in\mathcal{S}_A}\mathbb{P}\Big(L(\mathcal{G})\geq b(\mathcal{G}),\widehat{\mathcal{G}}=\mathcal{G},\widehat{A}=A\Big)\\
&\leq \sum_{A:|A_0\backslash A|=i}\sum_{\mathcal{G}\in\mathcal{S}_A}\mathbb{E}\exp\Big(\frac{t}{\sigma^2}L_1(\mathcal{G})\Big)\mathbb{E}\exp\Big(\frac{t}{\sigma^2}L_2(\mathcal{G})\Big)\exp\Big(-\frac{t}{\sigma^2}b(\mathcal{G})\Big)\\
&= \sum_{i=1}^{p_0}\sum_{A:|A_0\backslash A|=i}S_i^*\frac{\exp\Big(\frac{t(a-1)^2niC_{\min}}{(1-2t)\sigma^2}\Big)\exp\Big(-\frac{t}{\sigma^2}(-\lambda+a(a-1)niC_{\min})\Big)}{(1-2at)^{\frac{K_i^*}{2}}(1-2t)^{\frac{n-K_i^*}{2}}}\\
&\leq \sum_{i=1}^{p_0}\binom{p_0}{p_0-i}\sum_{j=0}^{i}\binom{p-p_0}{j}\frac{S_i^*}{(1-2t)^{n/2}}\exp\Big(-n\frac{t(a-1)iC_{min}}{\sigma^2}\frac{1-2at}{1-2t}\Big)\big(\frac{1-2t}{1-2at}\big)^{K_i^*/2}
\end{aligned}$$

where $S_i^*\equiv\max_{A\in\mathcal{A},|A_0\backslash A|=i}|\mathcal{S}_A|$ and $K_i^*\equiv\max_{\mathcal{G}\in\mathcal{S}_A,|A_0\backslash A|=i}K(\mathcal{G})$, as defined. This, together with the fact that $\binom{p_0}{p_0-i}\leq p_0^i$, $\sum_{j=1}^{i}\binom{p-p_0}{j}\leq(p-p_0)^i$ and $(p-p_0)p_0\leq\frac{p^2}{4}$, yields

$$I \leq \sum_{i=1}^{p_0}\frac{p^2}{4}S_i^*\exp\Big(-n\frac{t(a-1)iC_{min}}{\sigma^2}\frac{1-2at}{1-2t}\Big)\big(\frac{1-2t}{1-2at}\big)^{K_i^*/2}\frac{1}{(1-2t)^{n/2}} \qquad (18)$$

provided that $\frac{t}{\sigma^2}\lambda\leq 1$. Let $K^*=\max_{1\leq i\leq p_0}K_i^*/i,\log(S^*)=\max_{1\leq i\leq p_0}\log(S_i^*)/i$. For simplification, choose $t=\frac{1}{4(a-1)}$, $c=\frac{2a-3}{a-2}>2$, and a to satisfy $2\frac{n}{\log S^*}>a>4+\frac{n}{4\log S^*}$. Then (18) becomes:

$$\begin{aligned}
I &\leq \sum_{i=1}^{p_0}\frac{p^2}{4}S_i^*\exp\Big(-n\frac{1}{4c\sigma^2}iC_{min}\Big)c^{K_i^*/2}\frac{1}{(1-2t)^{n/2}}\\
&\leq \exp\Big(-\frac{n}{10\sigma^2}\big(C_{min}-20\sigma^2\frac{\log p}{n}-10\sigma^2\frac{K^*}{n}-20\sigma^2\frac{\log|\mathcal{S}|}{n}\big)\Big),
\end{aligned}$$

provided that $\tau\leq\frac{2\sigma}{p}\sqrt{\frac{\log p}{2np\lambda_{max}\big(\boldsymbol{X}^T\boldsymbol{X}\big)}}$. This leads to (8).

For the risk property, let $D=25\sigma^2$ and $G=\{\frac{1}{n}\|\boldsymbol{X}\hat{\boldsymbol{\beta}}^{tl}-\boldsymbol{X}\boldsymbol{\beta}^0\|^2\geq D\}$. Then

$$\frac{1}{n}\mathbb{E}\|\boldsymbol{X}\hat{\boldsymbol{\beta}}^g-\boldsymbol{X}\boldsymbol{\beta}^0\|^2=\frac{1}{n}\mathbb{E}\|\boldsymbol{X}\hat{\boldsymbol{\beta}}^g-\boldsymbol{X}\boldsymbol{\beta}^0\|^2(\mathbb{I}(G)+\mathbb{I}(G^c))\equiv T_1+T_2.$$

For T_1, note that $\frac{1}{4n}\|\boldsymbol{X}\hat{\boldsymbol{\beta}}^g-\boldsymbol{X}\boldsymbol{\beta}^0\|^2-\frac{1}{2n}\|\boldsymbol{\epsilon}\|^2\leq\frac{1}{2n}\|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}^g\|^2\leq\frac{1}{2n}\|\boldsymbol{\epsilon}\|^2$. By Markov's inequality with $t=\frac{1}{3}$, $T_1=\int_D^\infty\mathbb{P}\big(\frac{1}{n}\|\boldsymbol{X}\hat{\boldsymbol{\beta}}^{tl}-\boldsymbol{X}\boldsymbol{\beta}^0\|^2\geq x\big)dx$ is upper bounded by

$$\begin{aligned}
&\textstyle\int_D^\infty\mathbb{P}\Big(\frac{1}{n}\|\boldsymbol{\epsilon}\|^2\geq\frac{x}{4}\Big)dx\leq\int_D^\infty\mathbb{E}\exp\big(\frac{t\|\boldsymbol{\epsilon}\|^2}{\sigma^2}\big)\exp\big(-nt\frac{x}{4\sigma^2}\big)dx\\
&\textstyle\leq\int_D^\infty\exp\big(-\frac{n}{12\sigma^2}(x-24\sigma^2)\big)dx=\frac{12\sigma^2}{n}\exp\big(-\frac{n}{12}\big),
\end{aligned}$$

implying that $T_1=o(\frac{p_0}{n}\sigma^2)$. For T_2, then,

$$
\begin{aligned}
T_2 &\leq D\mathbb{P}(\hat{\boldsymbol{\beta}}^g \neq \hat{\boldsymbol{\beta}}^{ol}) + \frac{1}{n}\mathbb{E}\|\boldsymbol{X}\hat{\boldsymbol{\beta}}^{ol} - \boldsymbol{X}\boldsymbol{\beta}^0\|^2 \\
&= 25\sigma^2\mathbb{P}(\hat{\boldsymbol{\beta}}^g \neq \hat{\boldsymbol{\beta}}^{ol}) + \frac{K_0}{n}\sigma^2 = (o(1)+1)\frac{K_0}{n}\sigma^2.
\end{aligned}
$$

The desired result then follows. This completes the proof.

Proof of Theorem 2: The proof is similar to that of Theorem 3 with some minor modifications. In the present case, let $\widehat{\mathcal{G}}^{l_0}$ be a grouping associated with $\hat{\boldsymbol{\beta}}^{l_0}$. Then $\hat{\boldsymbol{\beta}}^{l_0} = \hat{\boldsymbol{\beta}}^{ol}$ if $\widehat{\mathcal{G}}^{l_0} = \mathcal{G}^0$. This means $\{\hat{\boldsymbol{\beta}}^{l_0} \neq \hat{\boldsymbol{\beta}}^{ol}\} = \{\widehat{\mathcal{G}}^{l_0} \neq \mathcal{G}^0\}$. Then

$$
\mathbb{P}(\hat{\boldsymbol{\beta}}^{l_0} \neq \hat{\boldsymbol{\beta}}^{ol}) \leq \sum_{i=0}^{p_0} \mathbb{P}\Big(S(\hat{\boldsymbol{\beta}}^{l_0}) - S(\hat{\boldsymbol{\beta}}^{ol}) \leq 0, \widehat{\mathcal{G}}^{l_0} \neq \mathcal{G}^0\Big) \equiv I
$$

Note that $S(\hat{\boldsymbol{\beta}}^{l_0}) \equiv \frac{1}{2}\big\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{l_0}\big\|^2 \geq \frac{1}{2}\big\|(\boldsymbol{I} - \boldsymbol{P}_{\widehat{\mathcal{G}}^{l_0}})(\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0} + \boldsymbol{\epsilon})\big\|^2$. Then

$$
\begin{aligned}
&2\Big(S(\hat{\boldsymbol{\beta}}^{l_0}) - S(\hat{\boldsymbol{\beta}}^{ol})\Big) \geq \big\|(\boldsymbol{I} - \boldsymbol{P}_{\widehat{\mathcal{G}}^{l_0}})(\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0} + \boldsymbol{\epsilon})\big\|^2 - \|(\boldsymbol{I} - \boldsymbol{P}_{\mathcal{G}^0})\boldsymbol{\epsilon}\|^2 \\
&= 2\boldsymbol{\epsilon}^T(\boldsymbol{I} - \boldsymbol{P}_{\widehat{\mathcal{G}}^{l_0}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0} + \|(\boldsymbol{I} - \boldsymbol{P}_{\widehat{\mathcal{G}}^{l_0}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\|^2 + \|(\boldsymbol{I} - \boldsymbol{P}_{\widehat{\mathcal{G}}^{l_0}})\boldsymbol{\epsilon}\|^2 - \|(\boldsymbol{I} - \boldsymbol{P}_{\mathcal{G}^0})\boldsymbol{\epsilon}\| \\
&\geq 2\boldsymbol{\epsilon}^T(\boldsymbol{I} - \boldsymbol{P}_{\widehat{\mathcal{G}}^{l_0}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0} + \|(\boldsymbol{I} - \boldsymbol{P}_{\widehat{\mathcal{G}}^{l_0}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\|^2 - \boldsymbol{\epsilon}^T\boldsymbol{P}_{\widehat{\mathcal{G}}^{l_0}}\boldsymbol{\epsilon} \equiv -L(\widehat{\mathcal{G}}^{l_0}) + b(\widehat{\mathcal{G}}^{l_0}), \quad (19)
\end{aligned}
$$

where $L(\mathcal{G}) \equiv L_1(\mathcal{G}) + L_2(\mathcal{G}) = 2\boldsymbol{\epsilon}^T(\boldsymbol{I} - \boldsymbol{P}_{\mathcal{G}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0} + \boldsymbol{\epsilon}^T\boldsymbol{P}_{\mathcal{G}}\boldsymbol{\epsilon}$, $b(\mathcal{G}) = \|(\boldsymbol{I} - \boldsymbol{P}_{\mathcal{G}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\|^2$, and $L_1(\mathcal{G}) \equiv -2\boldsymbol{\epsilon}^T(\boldsymbol{I} - \boldsymbol{P}_{\mathcal{G}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}$ and $L_2(\mathcal{G}) \equiv \boldsymbol{\epsilon}^T\boldsymbol{P}_{\mathcal{G}}\boldsymbol{\epsilon}$, and $L_1(\mathcal{G})$ are $L_2(\mathcal{G})$ are independent. Recall that $\mathcal{S} = \big\{\mathcal{G} : \mathcal{G} \neq \mathcal{G}^0; C_1(\mathcal{G}) \leq p_0; C_2(\mathcal{G}, \mathcal{E}) \leq c_0\big\}$. Let $\widehat{A} = \mathcal{I} \setminus \widehat{\mathcal{I}}_0$, Then, for any $0 < t < 1/2$ by Markov's inequality,

$$
\begin{aligned}
I &\leq \sum_{A\in\mathcal{A}}\sum_{\mathcal{G}\in\mathcal{S}_A} \mathbb{P}\Big(L(\mathcal{G}) \geq b(\mathcal{G}), \widehat{\mathcal{G}} = \mathcal{G}, \widehat{A} = A\Big) \\
&\leq \sum_{A\in\mathcal{A}}\sum_{\mathcal{G}\in\mathcal{S}_A} \mathbb{E}\exp\Big(\frac{t}{\sigma^2}L_1(\mathcal{G})\Big)\mathbb{E}\exp\Big(\frac{t}{\sigma^2}L_2(\mathcal{G})\Big)\exp\Big(-\frac{t}{\sigma^2}b(\mathcal{G})\Big) \\
&= \sum_{i=1}^{p_0}\sum_{A\in\mathcal{A}, |A_0\setminus A| = i} S_i^*\exp\Big(-\frac{t-t^2}{2\sigma^2}niC_{min}\Big)\frac{1}{(1-2t)^{K_i^*}} \\
&\leq \sum_{i=1}^{p_0}\binom{p_0}{p_0 - i}\sum_{j=0}^{i}\binom{p-p_0}{j}S_i^*\exp\Big(-\frac{t-t^2}{2\sigma^2}niC_{min}\Big)\frac{1}{(1-2t)^{K_i^*}}
\end{aligned}
$$

where $S_i^* \equiv \max_{A\in\mathcal{A}, |A_0\setminus A| = i}|\mathcal{S}_A|, K(\mathcal{G}), K_i^* \equiv \max_{\mathcal{G}\in\mathcal{S}_A, |A_0\setminus A| = i}K(\mathcal{G})$, as defined. This, together with the fact that $\binom{p_0}{p_0-i} \leq p_0^i$, $\sum_{j=1}^{i}\binom{p-p_0}{j} \leq (p-p_0)^i$ and $(p-p_0)p_0 \leq \frac{p^2}{4}$, yields

$$I \;\le\; \sum_{i=1}^{p_0} \frac{p^2}{4} S_i^* \;\exp\Big(-\frac{t-t^2}{2\sigma^2} niC_{min}\Big)\frac{1}{(1-2t)^{K_i^*}}$$

Let $K^* = \max_{1\le i\le p_0} \frac{K_i^*}{i}$ and $\log S^* = \max_{1\le i\le p_0} \frac{\log S_i^*}{i}$. To simplify the bound we choose $t = \frac{e-1}{2e} > \frac{3}{10}$, where $\frac{t-t^2}{2} > \frac{1}{10}$

$$I \;\le\; \exp\Big(-\frac{n}{10\sigma^2}\big(C_{min} - 20\sigma^2\frac{\log p}{n} - 10\sigma^2\frac{K^*}{n} - 10\sigma^2\frac{\log S^*}{n}\big)\Big)$$

This leads to (7).

The proof for the risk property is the same and is omitted. This completes the proof.

Proof of Corollary 1: Easily, $K^* \le K_i^* \le K_0$. Note that for any $A \subset \mathcal{I}$ with $|A| \neq p_0$, $|S_A| \le \sum_{i=0}^{K_0-1}\binom{|A|}{i} \le \sum_{i=0}^{K_0-1}\binom{p_0}{i}$. Thus, $S_i^* = \max_{A\in\mathcal{A},|A_0\setminus A|=i} |\mathcal{S}_A| \le \sum_{i=1}^{K_0-1}\binom{p_0}{i} \le p_0^{K_0}$ and $S^* = \exp\big(\max_{1\le i\le p_0}\frac{\log S_i^*}{i}\big) \le \max_{1\le i\le p_0} S_i^* \le p_0^{K_0}$. Using the bounds derived in Theorem 2 and 3, we obtain the desired results.

References

[1] An, L.T.H. and Tao, P.D. (2005). The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. *Annals of Operations Research*, Vol. **133**, 23-46.

[2] Bondell, H. D. and Reich, B. J. (2008). Simultaneous regression shrinkage, feature selection, and supervised clustering of predictors with OSCAR. *Biometrics*, **64**, 115-23.

[3] Bottolo, L., Petretto, E., Blankenberg, S., Cambien, F., Cook, S. A., Tiret, L. and Richardson, S. (2011). Bayesian Detection of Expression Quantitative Trait Loci Hot Spots. *Genetics*, **189**, 1449-1459.

[4] Chartrand, G. (1985). *Introductory graph theory*, Prindle, Weber and Schmidt.

[5] Friedman, J., Hastie, T., Hofling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, **1**, 302-332.

[6] Fan, J., and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348-1360.

[7] Kim, S.,and Xing, E. (2009). Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS Genetics*, **5**, e1000587. doi:10.1371/journal.pgen.1000587.

[8] Li, C., and Li, H. (2010). Variable selection and regression analysis for graph-structured covariates with an application to genomics. *The Annals of Applied Statistics*, **4**, 1498-1516.

[9] Nicolae, D. L., Gamazon E, Zhang, W., Duan, S., Dolan, M. E., et al. (2010) Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. *PLoS Genet*, **6(4)**, e1000888.

[10] Pan, W., Xie, B. and Shen, X. (2010). Incorporating predictor network in penalized regression with application to microarray data. *Biometrics*. **66**, 474-484.

[11] Rinaldo, A. (2009). Properties and refinements of the Fused Lasso. *Ann. Statist.*, **37**, 2922-2952.

[12] Scherzer, C.R., Eklund, A.C., Morse, L.J., Liao, Z., et. al (2007). Molecular markers of early Parkinson's disease based on gene expression in blood. *Proceedings of the National Academy of Sciences*, **104**, 955-960.

[13] Shen, X., and Ye, J. (2002) Adaptive model selection. *Journal of the American Statistical Association*, **97**, 210-221.

[14] Shen, X., Pan, W., Zhu, Y. and Zhou, H. (2013). On constrained and regularized high-dimensional regression. *Annals of the Institute of Statistical Mathematics*, **1**, 1-26.

[15] Shen, X., and Huang, H. (2010). Grouping pursuit through a regularization solution surface. *Journal of the American Statistical Association*, **105**, 727-739.

[16] Shen, X., Pan, W., and Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, **107**, 223-232.

[17] Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B*, **67**, 91-108.

[18] Veyriera, J.B., et al. (2008). High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet*, **4**, e1000214.

[19] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, **68**, 49-67.

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

[21] Zhao, P., Rocha, G. and Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical feature selection. *The Annals of Statistics*, **37**, 3468-3497.

[22] Zhang, W., Zhu, J., Schadt, E. E., Liu, J. S. (2010) A Bayesian partition method for detecting pleiotropic and epistatic eQTL modules. *PLoS Computional Biology*, **6(1)**, e1000642.

[23] Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, **38**, 894-942.

[24] Zhong, H., Yang, X., Kaplan, L. M., Molony, C. and Schadt, E. E. (2010). Integrating Pathway Analysis and Genetics of Gene Expression for Genome-wide Association Studies. *The American Journal of Human Genetics*, **86**, 581-591.

[25] Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association.*, **101**, 1418-1429.

[26] Zou, H. and Trevor, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, **67(2)**, 301-320.

[27] Zhou, X., Kao, M.J., and Wong, W. W. (2002) Transitive functional annotation by shortest-path analysis of gene expression data. *PNAS*, **99**, 12783-12788.

Table 1: Sample means (SD in parentheses) of prediction mean squared error (PMSE), accuracy of feature selection VFP and VFN, accuracy of grouping pursuit GFP and GFN, based on 100 simulation replications in Example 1, for our proposed method (Grouping), adaptive Grace (aGrace) [8], *GFlasso* [7], Elastic-Net (Enet) [26] and Oscar [2]

Correlation	Method	PMSE	VFP	VFN	GFP	GFN
$Cor = .9$	Ours	20.6(2.2)	.09%(.21%)	.00%(.00%)	.16%(.40%)	.00%(.00%)
	GFlasso	22.6(2.4)	11.3%(4.93%)	.15%(.66%)	20.8%(8.50%)	.14%(.59%)
	Oscar	22.7(2.5)	63.2%(10.6%)	.00%(.00%)	83.6%(8.20%)	.00%(.00%)
	Enet	45.7(4.9)	18.2%(22.6%)	6.29%(5.84%)	22.4%(3.51%)	6.95%(6.30%)
	aGrace	22.5(2.4)	39.1%(43.0%)	.00%(.00%)	43.1%(37.5%)	.00%(.00%)
$Cor = .5$	Ours	20.5(2.1)	.17%(.56%)	.25%(.84%)	.33%(1.07%)	.24%(.84%)
	GFlasso	22.6(2.5)	15.10%(6.30%)	.00%(.00%)	27.15%(10.4%)	.00%(.00%)
	Oscar	24.3(2.8)	72.7%(6.23%)	.00%(.00%)	89.9%(3.90%)	.00%(.00%)
	Enet	40.8(4.7)	40.6%(42.8%)	3.43%(3.80%)	2.12%(8.85%)	6.35%(5.15%)
	aGrace	22.4(2.5)	36.2%(41.4%)	.00%(.00%)	41.2%(36.8%)	.00%(.00%)
$Cor = .2$	Ours	20.8(2.1)	.04%(.18%)	.84%(3.38%)	.09%(.36%)	.83%(3.35%)
	GFlasso	22.6(2.5)	19.7%(7.73%)	.00%(.00%)	34.7%(12.3%)	.00%(.00%)
	Oscar	26.7(3.3)	68.3%(9.80%)	.00%(.00%)	86.7%(8.49%)	.00%(.00%)
	Enet	47.1(6.7)	10.7%(12.8%)	16.1%(7.45%)	17.5%(4.83%)	14.8%(6.54%)
	aGrace	23.9(3.3)	35.7%(34.7%)	.13%(.54%)	45.8%(30.3%)	.10%(.42%)

Table 2: Sample means (SD in parentheses) of prediction mean squared error (PMSE), accuracy of feature selection VFP and VFN, accuracy of grouping pursuit GFP and GFN, based on 100 simulation replications in Example 1, for feature selection alone with $\lambda_2 = 0$ in (2) (TLP), grouping pursuit alone with $\lambda_1 = 0$ in (2) (Grouping), and simultaneous grouping pursuit and feature selection (Both).

Correlation	Method	PMSE	VFP	VFN	GFP	GFN
$Cor = .9$	Both	20.55(2.23)	.09%(.21%)	.00%(.00%)	.16%(.40%)	.00%(.00%)
	TLP	24.54(2.43)	.01%(.03%)	45.7%(9.44%)	.02%(.06%)	45.5%(9.44%)
	Grouping	372(218)	100%(.00%)	.00%(.00%)	82.8%(22.8%)	18.9%(24.6%)
$Cor = .5$	Both	20.54(2.12)	.17%(.56%)	.25%(.84%)	.33%(1.07%)	.24%(.84%)
	TLP	31.86(3.49)	.09%(.13%)	42.8%(9.01%)	.19%(.26%)	42.6%(9.02%)
	Grouping	462(47.8)	100%(.00%)	.00%(.00%)	49.3%(.96%)	59.4%(11.9%)
$Cor = .2$	Both	20.75(2.12)	.04%(.18%)	.84%(3.38%)	.08%(.36%)	.83%(3.35%)
	TLP	41.66(5.57)	.42%(.59%)	50.1%(13.1%)	.84%(1.17%)	49.7%(13.2%)
	Grouping	287(29.1)	100%(.00%)	.00%(.00%)	50.6%(.71%)	69.2%(12.4%)

Table 3: Sample means (SD in parentheses) of prediction mean squared error (PMSE), accuracy of feature selection VFP and VFN, accuracy of grouping pursuit GFP and GFN, as well as %Oracle, the percentage of time that our method reconstructs the oracle estimator, based on 100 simulation replications in Example 2, for our proposed method (Our), adaptive Grace (aGrace) [8], *GFlasso* [7], Elastic-Net (Enet) [26] and Oscar [2]. Setups have the TF-TF correlation of 0 and .5; k is the average number of erroneous edges.

Setup 1	Method	PMSE	VFP	VFN	GFP	GFN	%Oracle
$k = 0$	Ours	1.02(.02)	.00%(.00%)	.00%(.00%)	.00%(.00%)	.00%(.00%)	100%
	GFlasso	1.12(.05)	8.41%(1.66%)	.00%(.00%)	16.0%(3.03%)	.00%(.00%)	0%
	Oscar	1.20(.07)	85.0%(5.78%)	.00%(.00%)	96.4%(2.60%)	.00%(.00%)	0%
	Enet	1.51(.13)	1.42%(.44%)	.00%(.00%)	2.84%(.87%)	.00%(.00%)	0%
	aGrace	1.11(.05)	9.37%(3.31%)	.00%(.00%)	17.7%(5.93)	.00%(.00%)	0%
$k = 10$	Ours	1.02(.02)	.00%(.01%)	.00%(.00%)	.00%(.01%)	.00%(.00%)	90%
	GFlasso	1.12(.05)	11.3%(3.31%)	.00%(.00%)	20.7%(5.73%)	.00%(.00%)	0%
	Oscar	1.49(.14)	100%(.00%)	.00%(.00%)	88.7%(3.29%)	.00%(.00%)	0%
	Enet	1.53(.13)	1.39%(.45%)	.00%(.00%)	2.77%(.90%)	.00%(.00%)	0%
	aGrace	1.45(.10)	100%(.00%)	.00%(.00%)	96.7%(.61%)	.00%(.00%)	0%
$k = 100$	Ours	1.02(.02)	.00%(.01%)	.00%(.00%)	.00%(.01%)	.00%(.00%)	85%
	GFlasso	1.16(.06)	100%(.00%)	.00%(.00%)	89.2%(2.71%)	.00%(.00%)	0%
	Oscar	1.49(.12)	100%(.00%)	.00%(.00%)	90.6%(2.81%)	.00%(.00%)	0%
	Enet	1.52(.13)	1.38%(.45%)	.00%(.00%)	2.75%(.88%)	.00%(.00%)	0%
	aGrace	1.45(.11)	100%(.00%)	.00%(.00%)	96.0%(.81%)	.00%(.01%)	0%
Setup 2	Method	PMSE	VFP	VFN	GFP	GFN	%Oracle
$k = 0$	Ours	1.02(.02)	.18%(.54%)	.00%(.00%)	.36%(1.07%)	.00%(.00%)	%75
	GFlasso	1.12(.05)	12.4%(4.47%)	.00%(.00%)	23.1%(7.60%)	.00%(.00%)	0%
	Oscar	1.25(.08)	38.9%(7.93%)	.00%(.00%)	61.3%(9.11%)	.00%(.00%)	0%
	Enet	1.59(.15)	1.92%(.29%)	.00%(.00%)	3.81%(.57%)	.00%(.00%)	0%
	aGrace	1.11(.05)	11.5%(4.16%)	.00%(.00%)	21.6%(7.32%)	.00%(.00%)	0%
$k = 10$	Ours	1.02(.02)	.01%(.01%)	.00%(.00%)	.01%(.02%)	.00%(.00%)	78%
	GFlasso	1.36(.11)	3.57%(1.74%)	.00%(.00%)	6.71%(3.23%)	.00%(.00%)	0%
	Oscar	1.54(.15)	100%(.00%)	.00%(.00%)	83.9(4.96%)	.00%(.00%)	0%
	Enet	1.61(.16)	1.45%(.46%)	.00%(.23%)	2.9%(.91%)	.02%(.22%)	0%
	aGrace	1.51(.12)	100%(.00%)	.00%(.00%)	94.3%(1.3%)	.00%(.00%)	0%
$k = 100$	Ours	1.02(.02)	.01%(.02%)	.00%(.00%)	.01%(.03%)	.00%(.00%)	73%
	GFlasso	1.54(.14)	100%(.00%)	.00%(.00%)	87.6%(3.93%)	.00%(.00%)	0%
	Oscar	1.54(.14)	100%(.00%)	.00%(.00%)	87.8%(3.78%)	.00%(.00%)	0%
	Enet	1.61(.16)	1.45%(.46%)	.00%(.23%)	2.9%(.91%)	.02%(.22%)	0%
	aGrace	1.51(.13)	100%(.00%)	.00%(.00%)	95.1%(.84%)	.00%(.02%)	0%

Table 4: Performance of our methods after adding k ($k = 0, 2, 10$) erroneous edges for each informative predictors in Example 2.

Eroneous edges	PMSE	VFP	VFN	GFP	GFN	%Oracle
0	20.55(2.23)	.09%(.21%)	.00%(.00%)	.16%(.40%)	.00%(.00%)	67%
2	20.63(2.16)	.02%(.05%)	.93%(2.29%)	.03%(.10%)	.93%(2.27%)	58%
10	20.64(2.17)	.01%(.05%)	3.00%(4.78%)	.03%(.11%)	2.98%(4.74%)	35%

Table 5: Performance of our methods with different numbers of groups and different levels of difficulty in Example 3.

	$\gamma_{min} = 1$		$\gamma_{min} = 3$	
# groups	PMSE	%Oracle	PMSE	%Oracle
2	1.01(.02)	97%	1.01(.02)	91%
5	1.04(.03)	30%	1.03(.02)	76%
10	1.10(.05)	4%	1.06(.03)	79%
20	1.28(.09)	0%	1.15(.09)	32%

Table 6: Mean prediction error (PE), number of non-zero regression coefficient estimates, percentage of grouping s, for four competing methods, in the eQTL analysis for gene GLT1D1 in Section 5.2.

	Tuning			Final Model	
Method	PE	# non-zeros	% grouping s	# non-zeros	% grouping s
Lasso	0.93(0.07)	6.67(2.08)	0(0)	3	0
OSCAR	0.90 (0.07)	42.67(17.90)	0.26(0.17)	16	0.01
TLP	0.87(0.01)	1.33(0.58)	0(0)	1	0
Fuse	0.87(0.01)	1.33(0.58)	0(0)	1	0
Ours	0.85(0.04)	1.66(0.58)	0.67(0.58)	2	1

Table 7: Parameter estimation for the final model in Section 5.2, where only nonzero coefficients are displayed.

	Estimates		
Method	$\hat{\beta}_{787}$	$\hat{\beta}_{790}$	$\hat{\beta}_{1667}$
Lasso	0.064	2.451	-0.101
OSCAR	1.439	1.439	-0.347
TLP	0	5.090	0
Our-Fuse	0	5.090	0
Ours	2.874	2.874	0

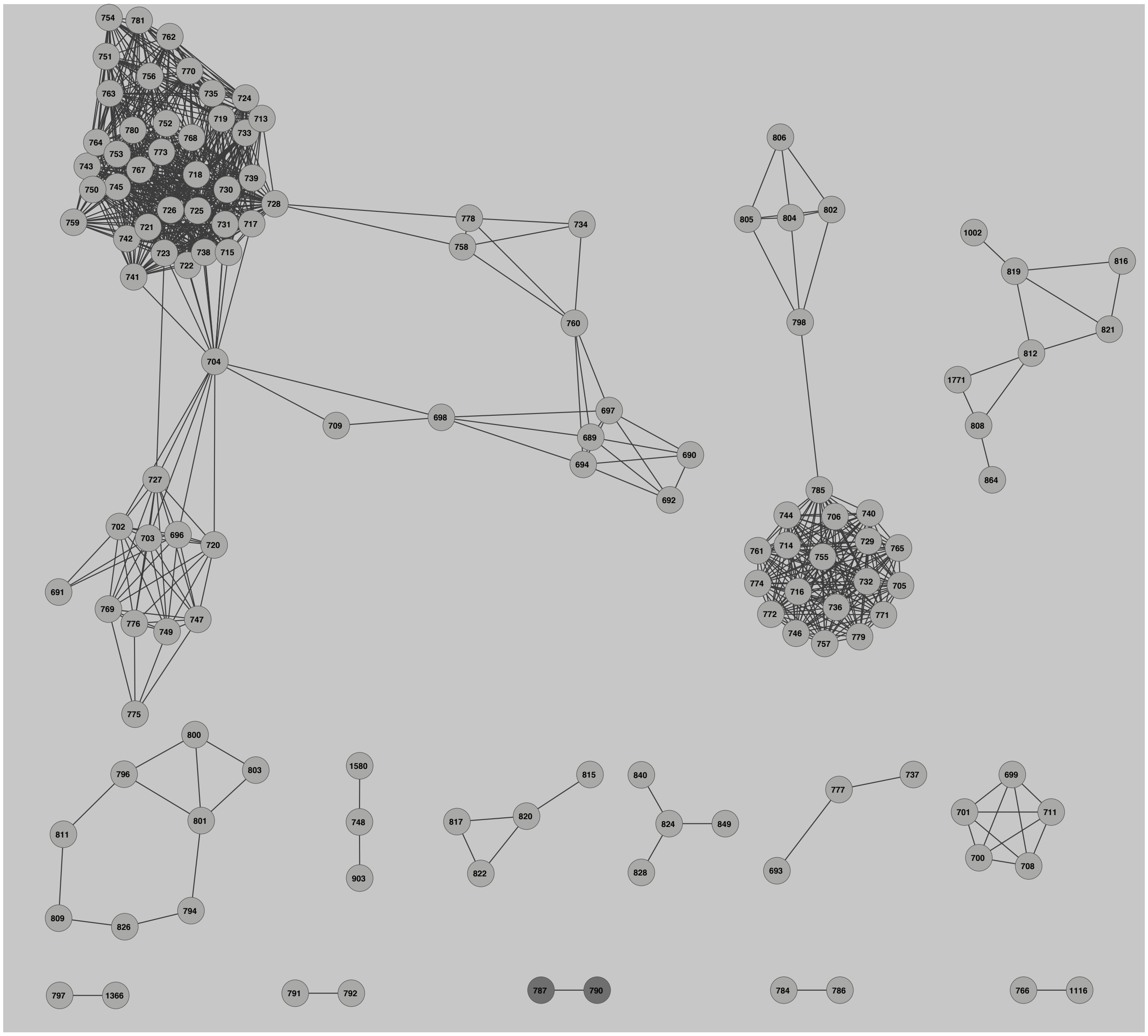

Figure 1: Subnetwork consisting of SNPs around informative locations, defined by correlation stronger than .6. Here SNP's locations are numbered with adjacent numbers indicating nearby locations.