Certifiably Polynomial Algorithm for Best Group Subset Selection

Yanhang Zhang†, Junxian Zhu†, Jin Zhu†, Xueqin Wang‡

April 27, 2021

Abstract

Best group subset selection aims to choose a small part of non-overlapping groups to achieve the best interpretability on the response variable. It is practically attractive for group variable selection; however, due to the computational intractability in high dimensionality setting, it doesn’t catch enough attention. To fill the blank of efficient algorithms for best group subset selection, in this paper, we propose a group-splicing algorithm that iteratively detects effective groups and excludes the helpless ones. Moreover, coupled with a novel Bayesian group information criterion, an adaptive algorithm is developed to determine the true group subset size. It is certifiable that our algorithms enable identifying the optimal group subset in polynomial time under mild conditions. We demonstrate the efficiency and accuracy of our proposal by comparing state-of-the-art algorithms on both synthetic and real-world datasets.

1 Introduction

Consider a linear model with $J$ non-overlapping groups:

$$y = \sum_{j=1}^{J} X_{G_j} \beta_{G_j} + \varepsilon,$$

where $y \in \mathbb{R}^n$ is the response variable, $X_{G_j} \in \mathbb{R}^{n \times p_j}$ is the design matrix of the $j$th group, $\beta_{G_j} \in \mathbb{R}^{p_j}$ is the regression coefficients of $j$th group and $\varepsilon$ is the random error term. Here $\{G_j\}_{j=1}^{J}$ are group indices of $p$ predictors such that $\bigcup_{j=1}^{J} G_j = \{1, \ldots, p\}$ and $G_i \cap G_j = \emptyset$ when $i \neq j$, and we suppose the group size of the $j$th group is $p_j$. We name the linear model as group linear model, and it simplifies to an ordinary linear model when $p_1 = \cdots = p_J = 1$. Group linear model is practically useful for the analysis of ubiquitousy existing variables with certain group structures. For instance, a categorical variable with several levels is often represented by a group of dummy variables. Besides, in a nonparametric additive model, a continuous component can be represented by a set of basis functions (e.g., a linear combination of spline basis functions). Finally, specific prior knowledge can impose group structures on variables. A typical example is that the

---

†School of Statistics, Renmin University of China, zhangyh98@ruc.edu.cn
‡Southern China Center for Statistical Science, Department of Statistical Science, School of Mathematics, Sun Yat-Sen University, zhangyh93@mail3.sysu.edu.cn, zhujx1@mail2.sysu.edu.cn, zhuj37@mail2.sysu.edu.cn
‡Department of Statistics and Finance/International Institute of Finance, School of Management, University of Science and Technology of China, wangxq20@ustc.edu.cn
genes belonging to the same biological pathway can be considered as a group in the genomic data analysis.

In recent decades, high dimensional group selection plays an essential role in various applications and has attracted considerable attention of many researchers [Bach, 2008; Zhao et al., 2009; Obozinski et al., 2011; Huang et al., 2012; Won et al., 2020]. One of the most natural formulations for this research is the best group subset selection (BGSS):

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2n} \|y - X\beta\|^2_2, \quad \text{s.t. } \|\beta\|_{0,r} \leq T, \quad (1)$$

where $$\|\beta\|_{0,r} = \sum_{j=1}^{J} I(\|\beta_{G_j}\|_r \neq 0)$$ is the $$\ell_r$$ norm for some $$r \geq 0$$ and model size $$T$$ is a positive integer to be determined from data. BGSS is generally NP-hard problem [Natarajan, 1995]. A naive approach to solve problem (1) is exhaustively searching all possible combination of $$T$$ groups, and then information criterion like AIC [Akaike, 1974], BIC [Schwarz, 1978] or EBIC [Chen and Chen, 2008], is used as a criterion for the choice of model size $$T$$. Unfortunately, even though with novel branch-and-bound and advanced implementation, the computational complexity of the naive approach exponentially increases as $$J$$, and thus, it is computationally infeasible on a modern computer when $$J$$ is larger than 30 [Bertsimas et al., 2016].

To circumvent the computational intractability of the BGSS in the context of high dimensionality, many regularization methods have been developed for group variable selection and estimation. One popular approach of group variable selection is group Lasso (GLasso) [Yuan and Lin, 2006], a natural extension of Lasso estimator [Tibshirani, 1996]. It applies $$\ell_2$$ penalty, the ridge penalty, within a group and applies $$\ell_1$$ penalty, the Lasso penalty, across the groups. Thus GLasso only encourages sparsity at the group level, equivalently that within a group, coefficients will either all be zero or all nonzero. [Huang and Zhang, 2010] provided the concept of strong group sparsity and showed that for strongly group-sparse signals, GLasso performs much better than standard Lasso. [Lounici et al., 2011] established the oracle inequalities of the prediction and $$\ell_2$$ error bounds for GLasso. To remedy over-shrinkage of GLasso method, researchers develop adaptive Lasso [Zou, 2006] or extend nonconvex penalties such as SCAD [Fan and Li, 2001] and MCP [Zhang, 2010] to group variable selection. Specifically, [Wang and Leng, 2008] proposed adaptive GLasso and proved its oracle property. [Wei and Huang, 2010] studied the asymptotic selection and estimation properties of adaptive GLasso when the number of predictors $$p$$ is larger than sample size $$n$$ under sparse Riesz condition [Zhang and Huang, 2008]. As for nonconvex group penalties, [Wang et al., 2007] extended the SCAD penalty to group selection and proved its oracle property. [Huang et al., 2012] proposed group MCP, and proved its oracle property for the high dimensional scenario, in which the number of groups $$J$$ exceeds the sample size $$n$$. Another two interesting non-convex penalties are the $$\ell_{q,r}$$ penalty ($$0 \leq q \leq 1, r \geq 1$$) and Capped-$$\ell_1$$ penalty, which are investigated by [Hu et al., 2017] and [Phan and Le Thi, 2019], respectively.

A special and notable regularization method for group selection considers the Lagrangian form of problem (1):

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2n} \|y - X\beta\|^2_2 + \lambda \|\beta\|_{0,r}, \quad (2)$$

where $$\lambda > 0$$ is the tuning parameter. Some remarkable methods solving problem (2) have been obtained in recent years. [Jiao et al., 2016] recruited a primal-dual active set strategy to group selection with $$\ell_{0,2}$$ penalty and conducted a theoretical analysis of the algorithm, such as a provable
finite-step convergence and support recovery. Hu et al. [2017] proposed a proximal gradient method to solve (2) and showed the proposed algorithm converged to a local minimum. However, owing to the nonconvexity of $\ell_{0,r}$ penalty, (1) and (2) are not equivalent, and there is no clear correspondence between $\lambda$ and $T$ such that $\ell_{0,r}$ regularization approaches do not provide how to select an exact model size. In fact, tuning the values of $\lambda$ in problem (2) or other group penalties such as GLasso are quite time-consuming [Huang et al., 2018]. In comparison, problem (1) directly controls the exact level of model size via the choice of model size $T$. Besides, Shen et al. [2013] pointed out that (1) is more preferable over (2) in terms of statistical properties of the solution.

In this paper, our primary aim is to design polynomial algorithms to obtain an exact solution of problem (1). To the best of our knowledge, our work is the first study for BGSS problem under the high dimensional setting, and our contributions are summarized as follows:

- A novel group-splicing (GSplicing) algorithm is proposed for iteratively solving BGSS. The iteration of GSplicing must stop in finite steps. Under mild conditions, GSplicing’s convergence rate is provably linear with a high probability, and its computational complexity is polynomially increases. Specifically, the complexity is proportional to sample size $n$ and the number of groups $J$.

- In terms of group subset selection, GSplicing perfectly recovers the true group subset with high probability given an exact group subset size. When model size exceeds the true group subset size, the output of GSplicing includes the true group subset with a high probability.

- A novel and efficient Bayesian group information criterion for selecting an optimal group subset size. By integrating the information criterion and GSplicing, an adaptive algorithm is designed to determine both true group subset and model size simultaneously. Theoretically, the adaptive algorithm can identify the true group subset in polynomial time with a high probability even when group subset size is unknown. Furthermore, a heuristic strategy is equipped to accelerate the adaptive algorithm.

- A generalization of our method is deployed to logistic regression, poisson regression [Nelder and Wedderburn, 1972] and Cox proportional hazard model [Cox, 1972] to perform binary classification and counting/censored response modeling, which are frequently encountered in data analysis. The helpful extension exhibit our proposal is valuable for more general problems.

1.1 Organization

The rest of this paper is organized as follows. In Section 2, we detail the proposed methods for the selection of the best group subset and develop a novel information criterion for the choice of optimal group size. Section 3 shows the theoretical properties of our algorithms. Numerical experiments follow in Section 4, in which we apply our methods and several state-of-the-art methods in both synthetic datasets and two real-world datasets. The complete technical proofs and generalization of our method are provided in the appendix.

1.2 Notations

Let $S = \{1, \ldots, J\}$. For a given group subset $A \subseteq S$ with size $|A|$, denote $\#A = \sum_{j \in A} p_j$ as the total number of variables included in $\bigcup_{j \in A} G_j$, and $\beta_A = (\beta_{G_j}, j \in A) \in \mathbb{R}^{\#A}$ as the
sub-vector of coefficients $\beta$ constrained in $\cup_{j \in A} G_j$, $X_A = (X_{G_j}, j \in A) \in \mathbb{R}^{n \times \#(A)}$ as the sub-matrix of design matrix $X$ constrained in $\cup_{j \in A} G_j$. We denote $I_p_j$ as a $p_j \times p_j$ identity matrix.

We define $[\cdot]$ as a function that returns the nearest integer, $I(\cdot)$ as an indicator function and $L(\beta) = \frac{1}{2n} \|y - X\beta\|_2^2$ as the loss function. Throughout this paper, we assume to apply a groupwise orthonormalization, i.e., by a QR decomposition, to obtain $X_{G_j}^\top X_{G_j}/n = I_{p_j}$ for all $j \in S$.

2 Methodology

We first introduce an algorithm for solving BGSS in Subsection 2.1. Furthermore, we develop an algorithm for jointly learning an optimal model size and BGSS’s solution in Subsection 2.2.

2.1 Group-Splicing algorithm

The augmented Lagrangian form of problem (1) with $\ell_{0,2}$ constraint is

$$
\begin{align*}
\min_{\beta \in \mathbb{R}^p} & \quad \frac{1}{2n} \|y - X\beta\|_2^2 + d^\top (\beta - Ku) + \kappa \frac{1}{2}\|\beta - Ku\|_2^2, \\
\text{s.t.} & \quad \|v\|_{0,2} \leq T, \\
& \quad \beta = Ku,
\end{align*}
$$

where $\kappa$ is a positive constant, $v \in \mathbb{R}^p$ is an instrumental variable, $d \in \mathbb{R}^p$ is the dual variable and $K$ is a block diagonal matrix with the $j$th sub-matrix $K_j$ equals to $\sqrt{p_j}I_{p_j}$. Here $K$ is used to eliminate the impact of group sizes. Especially, when groups contain the same group size, the block diagonal matrix $K \propto I_p$.

Without loss of generality, the response variable and predictors are centered around the mean and let $\|v\|_{0,2} = T$. From the Karush-Kuhn-Tucker (KKT) conditions of (3), we derive the selected set and unselected set as

$$
\begin{align*}
\mathcal{A}^o &= \{j \in S : \sum_{i=1}^J I\left(\frac{1}{p_j} \|\beta_{G_j}^o\|_2^2 + \frac{1}{\kappa}d_{G_j}^o\|_2^2 \leq \frac{1}{p_i} \|\beta_{G_i}^o\|_2^2 + \frac{1}{\kappa}d_{G_i}^o\|_2^2 \right) \leq T\}, \\
\mathcal{I}^o &= S \setminus \mathcal{A}^o = (\mathcal{A}^o)^c,
\end{align*}
$$

where primal variable $\beta^o$ and dual variable $d^o$ are defined as

$$
\begin{align*}
\beta^o_{\mathcal{A}^o} &= (X_{\mathcal{A}^o}^\top X_{\mathcal{A}^o})^{-1}X_{\mathcal{A}^o}^\top y, \quad \beta^o_{\mathcal{I}^o} = 0, \\
d^o_{\mathcal{A}^o} = 0, \quad d^o_{\mathcal{I}^o} = X_{\mathcal{I}^o}^\top (y - X\beta^o)/n.
\end{align*}
$$

We can approximate the KKT conditions iteratively. Let $\{\mathcal{A}^k, \mathcal{I}^k, \beta^k, d^k\}$ be the solution in the $k$th iteration. We update $\{\mathcal{A}^{k+1}, \mathcal{I}^{k+1}\}$ by

$$
\begin{align*}
\mathcal{A}^{k+1} &= \{j \in S : \sum_{i=1}^J I\left(\frac{1}{p_j} \|\beta_{G_j}^k\|_2^2 + \frac{1}{\kappa}d_{G_j}^k\|_2^2 \leq \frac{1}{p_i} \|\beta_{G_i}^k\|_2^2 + \frac{1}{\kappa}d_{G_i}^k\|_2^2 \right) \leq T\}, \\
\mathcal{I}^{k+1} &= S \setminus \mathcal{A}^{k+1} = (\mathcal{A}^{k+1})^c.
\end{align*}
$$

Then, update $\{\beta^{k+1}, d^{k+1}\}$ by
\[ \beta^{k+1}_{A_k} = (X_{A_k}^T X_{A_k+1})^{-1} X_{A_k+1}^T y, \quad \beta^{k+1}_{I_k} = 0, \]
\[ d^{k+1}_{A_k} = 0, \quad d^{k+1}_{I_k} = X_{I_k}^T (y - X \beta^{k+1})/n. \]

In [4], \( \kappa \) weighs the importance of the dual variable \( d^k \) in the iterations. It is worthy to note that a large value of \( \kappa \) (e.g., \( \kappa = +\infty \)) makes a small update on the selected set, which implies only a small number of groups change between selected set and unselected set in the next iteration, while a small \( \kappa \) (e.g., \( \kappa = 0 \)) might completely change selected set. Motivated by this observation, we consider the update [4] as an exchange between selected set and unselected set, which we call “splicing” procedure [Zhu et al. 2020]. Thus we can select \( \kappa \) by determining the number of elements to be exchanged. We precisely characterize the idea in the next paragraph.

Suppose the size of the exchanged group subset is a constant positive integer \( C(\leq |A^k|) \). And the smallest \( C \) groups in \( A^k \) and the largest \( C \) groups in \( I^k \) is defined as:

\[
S_{C,1}^k = \{ j \in A^k : \sum_{i \in A^k} I \left( \frac{1}{p_j} \| \beta_{G_i}^k \|_2^2 \geq \frac{1}{p_i} \| \beta_{G_i} \|_2^2 \right) \leq C \} \\
= \{ j \in A^k : \sum_{i \in A^k} I \left( \frac{1}{p_j} \| \beta_{G_i}^k \|_2 \geq \frac{1}{p_i} \| \beta_{G_i} \|_2 \right) \leq C \} 
\]

and

\[
S_{C,2}^k = \{ j \in I^k : \sum_{i \in I^k} I \left( \frac{1}{p_j} \| \beta_{G_i}^k \|_2^2 \leq \frac{1}{p_i} \| \beta_{G_i} \|_2^2 \right) \leq C \} \\
= \{ j \in I^k : \sum_{i \in I^k} I \left( \frac{1}{p_j} \| \beta_{G_i}^k \|_2 \leq \frac{1}{p_i} \| \beta_{G_i} \|_2 \right) \leq C \},
\]

where the last equation in (5) follows from \( d^{k}_{A^k} = 0 \) and the last equation in (6) follows from \( \beta^{k}_{I_k} = 0 \). The groups in \( S_{C,1}^k \) and \( S_{C,2}^k \) have intuitive explanations. In fact, according to the following lemma, \( S_{C,1}^k \) (\( S_{C,2}^k \)) can be interpreted as the groups in \( A^k \) (\( I^k \)) with smallest (largest) averaged contributions to loss function:

**Lemma 1**

(i) For any \( j \in A^k \), the contributions to the decreases of \( L(\beta^k) \) by discarding \( j \)th group is

\[ L(\beta^{A^k\setminus \{j\}}) - L(\beta^k) = \frac{1}{2} \| \beta_{G_i}^k \|_2^2, \]

where \( \beta^{A^k\setminus \{j\}} \) is the estimator assigning the \( j \)th group of \( \beta^k \) to be zero.

(ii) For any \( j \in I^k \), the contributions to the decreases of \( L(\beta^k) \) by adding \( j \)th group is

\[ L(\beta^k) - L(\beta^k + t^k_j) = \frac{1}{2} \| d_{G_j}^k \|_2^2, \]

where \( t^k_j = \arg \min_{t_G, j \neq 0} L(\beta^k + t), d_{G_j}^k = X_{G_j}^T (y - X \beta^k)/n. \)

More importantly, the concepts of \( S_{C,1}^k \) and \( S_{C,2}^k \) connect the exchanged group subset size \( C \) and the weight of dual parameter \( \kappa \):
Lemma 2  Given exchanged subset size $C$, the corresponding range of $\kappa$ in the $k$th iteration is

$$
\kappa \in \left( \frac{\min_{j \in \mathcal{S}_k} \|d_{G_j}^k\|^2 / p_j}{\max_{i \in \mathcal{S}_k} \|\beta_{G_i}^k\|^2 / p_i} \right)^{\frac{1}{2}}, \left( \frac{\min_{j \in \mathcal{S}_{k,2}} \|d_{G_j}^k\|^2 / p_j}{\max_{i \in \mathcal{S}_{k,2}} \|\beta_{G_i}^k\|^2 / p_i} \right)^{\frac{1}{2}}.
$$

Lemma 2 recommends that we can choose a subset size $C$ instead of adaptively selecting $\kappa$. To minimize the loss function in (1), a natural way is choosing $C$ such that loss function can decrease after updating selected set. The procedure determining $C$ to find a better selected set (in the sense of loss value) refers to "group-splicing". We summarise group-splicing algorithm as Algorithm 1.

**Algorithm 1 Group-Splicing (GSplcing) algorithm**

**Require:** $X$, $y$, $\{G_j\}_{j=1}^T$, $T$, $\mathcal{A}^0$, $\pi_T$, $C_{\text{max}}$.
1: Initial primal variable $\beta^0$ and dual variable $d^0$

$$
\beta^0_{\mathcal{A}^0} = (X_{\mathcal{A}^0}^T X_{\mathcal{A}^0})^{-1} X_{\mathcal{A}^0}^T y, \quad \beta^0_{\mathcal{T}^0} = 0,
$$

$$
d^0_{\mathcal{T}^0} = X_{\mathcal{T}^0}^T (y - X \beta^0_{\mathcal{T}^0}) / n, \quad d^0_{\mathcal{A}^0} = 0.
$$

2: for $k = 0, 1, \ldots, k_{\text{max}}$ do
3: Compute $L = \frac{1}{2n} \|y - X \beta^k\|^2_2$ and update $\mathcal{S}_1^k, \mathcal{S}_2^k$

$$
\mathcal{S}_1^k = \{ j \in \mathcal{A}^k : \sum_{i \in \mathcal{A}^k} I(\frac{1}{p_j} \|\beta_{G_j}^k\|^2_2 \geq \frac{1}{p_i} \|\beta_{G_i}^k\|^2_2) \leq C_{\text{max}} \},
$$

$$
\mathcal{S}_2^k = \{ j \in \mathcal{T}^k : \sum_{i \in \mathcal{T}^k} I(\frac{1}{p_j} \|d_{G_j}^k\|^2_2 \leq \frac{1}{p_i} \|d_{G_i}^k\|^2_2) \leq C_{\text{max}} \}.
$$

4: for $C = C_{\text{max}}, \ldots, 1$, do
5: Let $\tilde{\mathcal{A}}_C^k = (\mathcal{A}^k \setminus \mathcal{S}_1^k) \cup \mathcal{S}_2^k$ and $\tilde{\mathcal{T}}_C^k = (\mathcal{T}^k \setminus \mathcal{S}_2^k) \cup \mathcal{S}_1^k$.
6: Update primal variable $\tilde{\beta}$ and dual variable $\tilde{d}$

$$
\tilde{\beta}_{\tilde{\mathcal{A}}_C^k} = (X_{\tilde{\mathcal{A}}_C^k}^T X_{\tilde{\mathcal{A}}_C^k})^{-1} X_{\tilde{\mathcal{A}}_C^k}^T y, \quad \tilde{\beta}_{\tilde{\mathcal{T}}_C^k} = 0,
$$

$$
\tilde{d}_{\tilde{\mathcal{A}}_C^k} = X_{\tilde{\mathcal{A}}_C^k}^T (y - X \tilde{\beta}) / n, \quad \tilde{d}_{\tilde{\mathcal{T}}_C^k} = 0.
$$

7: Compute $\tilde{L} = \frac{1}{2n} \|y - X \tilde{\beta}\|^2_2$.
8: if $L - \tilde{L} > \pi_T$, then
9: Denote $(\tilde{\mathcal{A}}_C^k, \tilde{\mathcal{T}}_C^k, \tilde{\beta}, \tilde{d})$ as $(\mathcal{A}^{k+1}, \mathcal{T}^{k+1}, \beta^{k+1}, d^{k+1})$ and break.
10: else
11: Update $\mathcal{S}_1^k, \mathcal{S}_2^k$

$$
\mathcal{S}_1^k = \mathcal{S}_1^k \setminus \arg \max \{ \frac{1}{p_i} \|\beta_{G_i}^k\|^2_2 \},
$$

$$
\mathcal{S}_2^k = \mathcal{S}_2^k \setminus \arg \min \{ \frac{1}{p_i} \|d_{G_i}^k\|^2_2 \}.
$$

12: end if
13: end for
14: if $\mathcal{A}^{k+1} = \mathcal{A}^k$, then
15: Stop and denote the solution at the $k$th iteration as $(\tilde{\mathcal{A}}, \tilde{\beta})$.
16: end if
17: end for

**Ensure:** $(\tilde{\mathcal{A}}, \tilde{\beta})$. 


We detail about a part of input parameters in Algorithm 1. The first one is the model size $T$. We fix the size of selected set during iterations to $T$ groups so that we can select an exact model size. The second one is the initial selected set $A^0$. Typically, we choose the $T$ largest elements of set $\{\|X^\top_G y\|_2^2/p_j, j = 1, \ldots, J\}$ as $A^0$. $\pi_T$ in Algorithm 1 is a threshold related to the given model size $T$. Once the algorithm recovers the true group subset, we may splice some irrelevant groups, and then the loss function may decrease slightly. The threshold $\pi_T$ can reduce these unnecessary iterations. Typically, $\pi_T$ is relatively small, i.e., $\pi_T = 0.1T \log(nJ)/n$. The last input parameter $C_{\text{max}}$ is an important parameter, which controls the maximum exchanged size in GSplicing algorithm. With a large value of $C_{\text{max}}$, the more time in each iteration but the fewer iterations are required in GSplicing algorithm. On the contrary, more iterations but less time are required in each iteration in GSplicing algorithm with a smaller value. Typically, we set $C_{\text{max}}$ with a small size, i.e., $C_{\text{max}} = 2$.

It is worth noting that GSplicing algorithm outputs a solution in finite steps since the loss function in each iteration always decreases at least $\pi_T$ and the choice of the selected set with fixed model size $T$ is finite. It is also easy to see that, by replacing the loss function in Lemma 1 as the negative log (partial-)likelihood function, we can extend group-splicing technique to solve the best group subset selection for generalized linear model [Nelder and Wedderburn, 1972] and Cox proportional hazard model [Cox 1972]. We defer detail derivation in the supplementary material.

### 2.2 Adaptive group-splicing algorithm

It is crucial to decide the optimal model size that is usually unknown in practice. Our idea is to take model size $T$ as a tuning parameter and run GSplicing algorithm along a search sequence of $T$. Indeed, we can simply set the search sequence from $T = 1$ to $T = T_{\text{max}}$, in which $T_{\text{max}}$ is the upper bound of potential model size. Then, we can use some model selection techniques, such as cross-validation and information criterion, to determine the optimal model size.

In high dimensional scenario, the Bayes information criterion (BIC) may be too liberal [Chen and Chen, 2008]. To better adapt to high dimensionality, we propose a novel information criterion for group selection, which is named as Bayesian group information criterion (BGIC). The BGIC supported on $\bigcup_{j \in \hat{A}} G_j$ is defined as

$$\text{BGIC}_{\gamma}(\hat{A}) = n \log \hat{L}(\hat{\beta}) + (\gamma \log J + \log n)\#\{\hat{A}\},$$

where $\hat{\beta}$ is the least squares estimator on support $\hat{A}$, and $\gamma$ is a pre-specific non-negative value. To identify the true group structures in the high dimensional scenario, BGIC considers the complexity penalty is log $J$, controlling the diverging rate by constant $\gamma$, and preserves the same penalty of sample size, log $n$, as BIC. In particular, BGIC simplifies to ordinary BIC when $\gamma = 0$, which eliminates the effect of model complexity on group selection. Employing BGIC, we design a sequential group-splicing algorithm, which is summarized as follows.
Algorithm 2 Sequential Group-Splicing (SGSplicing) algorithm

Require: $X, y, \{G_j\}_{j=1}^J, T_{\text{max}}, C_{\text{max}}, \gamma$

1: for $T = 1, \ldots, T_{\text{max}}$ do
2: \hspace{0.5cm} $A_T^0 \leftarrow A_{T-1}^0 \cup \arg \max_j \{\|X_{G_j}^T(y - X\hat{\beta}_{T-1})\|^2 / p_j\}$.
3: \hspace{0.5cm} $(\hat{A}_T, \hat{\beta}_T) \leftarrow \text{GSplicing}(X, y, \{G_j\}_{j=1}^J, T, A_T^0, \pi_T, C_{\text{max}})$.
4: \hspace{0.5cm} $\text{BGIC}_T \leftarrow \text{BGIC}_{\gamma}(\hat{A}_T)$.
5: end for
6: $T^* \leftarrow \arg \min_T \{\text{BGIC}_T\}$.

Ensure: $(\hat{A}_{T^*}, \hat{\beta}_{T^*})$.

We discuss the parameters $T_{\text{max}}$ and $\gamma$ in Algorithm 2. Algorithm 2 would miss the true model size if $T_{\text{max}}$ is too small (e.g., a small constant); while the runtime would visibly increase if $T_{\text{max}}$ is proportional to $J$. From our theoretical result (see Theorem 2), we suggest $T_{\text{max}} = \left\lceil \frac{n}{p_{\text{max}}} \log(J) \right\rceil$. Analogous to EBIC [Chen and Chen, 2008], a large $\gamma$ achieves low positive selection rate (PSR) and low false discovery rate (FDR) while a small one enlarges both of them. Our empirical studies suggest that, generally, $\gamma = \frac{1}{2}$ or $\gamma = \frac{1}{4}$ well balances the PSR and FDR.

Figure 1: A typical “model size v.s. BGIC” plot. The $x$-axis is model size, and the $y$-axis is BGIC’s value recorded in SGSplicing algorithm. The entries of design matrix $X$ are i.i.d. sampled from $\mathcal{N}(0, 1)$, and the matrix shape is $100 \times 200$. The error term $\varepsilon$ are i.i.d. $\mathcal{N}(0, \frac{1}{2})$. Take the two adjacent variables as one group, and set the true coefficients $\beta = (1, 1, 1, 1, 1, -1, -1, -1, -1, 0, \ldots, 0)$. Two tuning parameters in SGSplicing are: $T_{\text{max}} = 15$ and $C_{\text{max}} = 2$. The orange vertical dash line indicates the true group subset size.

Two typical BGIC values of Algorithm 2 are presented in Figure 1, from which we see that the BGIC decreases from $T = 1$ to $T = 5$, but it increases as $T$ larger than 5. In other words, the BGIC path of SGSplicing algorithm is a strictly unimodal function achieving minimum at the true group subset size $T = 5$. Motivated by this observation, we propose a heuristic search algorithm based on the golden-section search technique [Kiefer, 1953], an efficient method for finding the extremum of a unimodal function, to determine $T$ that minimizing BGIC. We summarize the golden-section group-splicing algorithm in Algorithm 3. Algorithm 3 only needs to run GSplicing again.
algorithm $O(\log_{0.618-1} T_{\text{max}})$ times, which reduce the $O(T_{\text{max}})$ times runs of GSplicing algorithm in Algorithm 2.

Algorithm 3 Golden-section Group-Splicing (GGSplicing) algorithm

Require: $X, y, \{G_j\}_{j=1}^J, T_{\text{min}}, T_{\text{max}}, C_{\text{max}}, \gamma$.

1: Initial the value of $T_1$ and $T_2$

   $T_1 \leftarrow [0.618 \times T_{\text{min}} + 0.382 \times T_{\text{max}}],$

   $T_2 \leftarrow [0.382 \times T_{\text{min}} + 0.618 \times T_{\text{max}}].$

2: Run GSplicing algorithm given model size $T_1$ and $T_2$.
3: Compute the information criterion and denote as $\text{BGIC}_{T_1}$ and $\text{BGIC}_{T_2}$.
4: while $T_1 \neq T_2$, do
5:   if $\text{BGIC}_{T_1} \leq \text{BGIC}_{T_2}$, then
6:     $T_{\text{max}} \leftarrow T_2, T_2 \leftarrow T_1, T_1 \leftarrow [0.618 \times T_{\text{min}} + 0.382 \times T_{\text{max}}].$
7:     $\text{BGIC}_{T_2} \leftarrow \text{BGIC}_{T_1}.$
8:   else
9:     $T_{\text{min}} \leftarrow T_1, T_1 \leftarrow T_2, T_2 \leftarrow [0.382 \times T_{\text{min}} + 0.618 \times T_{\text{max}}].$
10:    $\text{BGIC}_{T_1} \leftarrow \text{BGIC}_{T_2}.$
11:    Run GSplicing algorithm given model size $T_2$ and compute $\text{BGIC}_{T_2}$.
12:  end if
13: end while
14: $(\hat{A}, \hat{\beta}) \leftarrow \text{GSplicing}(X, y, T_1, A_0^T, \{G_j\}_{j=1}^J, \pi T_1, C_{\text{max}}).$

Ensure: $(\hat{A}, \hat{\beta})$.

3 Theoretical properties

In this section, we study the theoretical properties of our algorithms. We first discuss the exact support recovery property of the algorithms in Subsection 3.1. Next, we show the selection consistency of SGSplicing algorithm under BGIC in Subsection 3.2. In Subsection 3.3, we establish the linear convergence rate of GSplicing algorithm. Based on this result, we provide the convergence steps of GSplicing algorithm and polynomial time complexity of SGSplicing algorithm. Finally, we provide the $\ell_2$ error bounds of the estimated coefficients in each iteration in Subsection 3.4.

Before formally present theoretical properties, we discuss some technical conditions. The first condition constrains the behavior of the error terms:

(C1) $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d with mean zero and sub-Gaussian tails, i.e., there exists a $\sigma > 0$ such that $P(|\varepsilon_i| > z) \leq 2 \exp(-z^2/\sigma^2)$, for all $z \geq 0$.

Condition (C1) assumes the error terms are all sub-Gaussian random variables. As one of the most common conditions in high-dimensional variable selection, Condition (C1) frequently appears in statistical literature for high-dimensional data analysis [Huang et al., 2018, Zhang, 2010].

Our second condition is a constraint for the design matrix, called the sparse Riesz condition (SRC). The SRC, as a relatively mild identifiability condition for a linear regression model, is first
proposed by Zhang and Huang [2008] and has been recruited to study the theoretical property of Lasso [Zhang and Huang] 2008, MCP [Zhang] 2010. It also has been extended to the group selection case, serves as an indispensable ingredient for the analysis of the selection consistency of adaptive GLasso and group MCP [Wei and Huang 2010 Huang et al.] 2012. We say a design matrix $X$ satisfies SRC with order $\tau$ and spectrum bound $0 < c_s \leq c^* < \infty$ if

$$c_s \|u\|^2 \leq \frac{\|X_A u\|^2}{n} \leq c^* \|u\|^2, \; \forall u \neq 0, u \in \mathbb{R}^{|A|}$$ with $|A| \leq \tau$.

Mathematically, the SRC guarantees the spectrum of the diagonal sub-matrices of matrix $X^\top X/n$ can be bounded. Our second condition is formally stated as:

(C2) $X$ satisfies SRC with order $2T$ and spectrum bound $\{c_s, c^*\}$.

Moreover, if Condition (C2) holds, we can derive a conclusion bounding the spectrum of the off-diagonal sub-matrices: As for the spectrum of the off-diagonal sub-matrices, we can be bounded by constant $\omega$ with order $\tau$, equivalently that for two non-overlapping index sets $A$ and $B$

$$\omega \|u\|^2 \geq \frac{\|X^\top_A X_B u\|^2}{n}, \; \forall u \neq 0, u \in \mathbb{R}^{|B|}$$ with $|A| \leq \tau, |B| \leq \tau$.

The bound has been used in Huang et al. [2018] to analyze the theoretical properties of SDAR, which is also easy to be extended to the group case. We consider $\omega$ for the group case, where $\omega$ imposes the approximate orthogonality of two disjoint groups. It can be shown that constant $\omega$ can be bounded by $(1 - c_s) \lor (c^* - 1)$ when $X$ satisfies SRC with order $2\tau$. Therefore, $\omega$ exists if $X$ satisfies SRC with order $2\tau$, thus we do not require more assumptions on the correlation of inter-groups with the satisfaction of SRC.

Here we give some useful notations. Denote $A^*$ as the true group subset and $\beta^*$ as the true coefficients. Let the true group subset size $|A^*| = s$, maximum group size $p_{\text{max}} = \max \{p_j : 1 \leq j \leq J\}$, minimum group size $p_{\text{min}} = \min \{p_j : 1 \leq j \leq J\}$, relative group size $t = \sqrt{p_{\text{max}}/p_{\text{min}}}$ and the magnitude of minimum group signal $\min\{\|\beta_{\text{g}}\|_2^2\} = \vartheta$. Denote $\delta_1 = O(Jp_{\text{max}} \exp\{-nC_1\vartheta/s\text{p}_{\text{max}}\})$ and $\delta_2 = O(Jp_{\text{max}} \exp\{-nC_2\vartheta/Tp_{\text{max}}\})$ for some positive constants $C_1$ and $C_2$ depending on $T$. Here we assume the $p_{\text{max}}$ is quite smaller than the sample size $n$ or the number of groups $J$, which is appropriate in practice.

The technical condition is given in Condition (C3):

(C3) $0 < \mu_T < 1$, where $\mu_T = \frac{2c^*[(1+\eta)(1+t)\omega]^2}{(1-\eta)(c_s - \frac{\omega^2}{c^*})^3}$ is a constant depending on $T$ for some constant $0 < \eta < 1$.

To make sure $0 < \mu_T < 1$ holds, let $\delta = (1 - c_s) \lor (c^* - 1)$, which is closely related to the RIP (restricted isometry property) constant $\delta_T$ for $X$ [Candes and Tao] 2005. In can be verified that a sufficient condition for (C3) holds when $t = 1$, which implies groups contain the same group size, is $\delta \leq 0.2055$, i.e., $c_s \geq 0.7945$, $c^* \leq 1.2055$, which is weaker than the condition $\delta \leq 0.1599$ in Huang et al. [2018]. When $t = \sqrt{2}$, which means the largest group size is twice than the smallest one, the sufficient condition is $\delta \leq 0.1837$, i.e., $c_s \geq 0.8163$, $c^* \leq 1.1837$. In addition, it is easy to verify that $\mu_T > 0$ if $c_s \geq \delta$.

Finally, we provide some conditions about threshold $\pi_T$ and the magnitude of $\beta^*$:
(C4) The threshold in Algorithm 1 satisfies \( \pi_T = O\left( \frac{T p_{\text{max}} \log(n J)}{n} \right) \).

(C5) The magnitude of minimum group signal satisfies \( \frac{1}{\vartheta} = o\left( \frac{n}{T p_{\text{max}} \log(n J)} \right) \).

3.1 Support recovery

We begin by considering the support recovery when given model size \( T \geq s \).

**Theorem 1** Denote \((\hat{A}, \hat{I}, \hat{\beta}, \hat{d})\) is the solution of Algorithm 1. If Conditions (C1)-(C5) hold, when \( T \geq s \), we have

\[
P(\hat{A} \supseteq A^*) \geq 1 - \delta_1 - \delta_2,
\]

and especially, if \( T = s \), we have

\[
P(\hat{A} = A^*) \geq 1 - \delta_1 - \delta_2.
\]

Theorem 1 shows that, with a high probability, the output of Algorithm 1 is a no-false-exclusion estimator. The constraint in Condition (C3) implies that \( t = 1 \), an even group size structure, is the most favorable case, and the satisfaction of Condition (C2) is much easier. In addition, we can derive the smallest convergence steps with an even group size structure as Subsection 3.3 shows in this case.

**Corollary 1** Suppose both the conditions in Theorem 1 hold. When \( T \geq s \), we have

\[
\lim_{n \to \infty} P(\hat{A} \supseteq A^*) = 1,
\]

and especially, if \( T = s \), we have

\[
\lim_{n \to \infty} P(\hat{A} = A^*) = 1.
\]

Corollary 1 guarantees that, under mild conditions, the estimator output by the GSplicing certainly include all important groups if \( T \geq s \) or exactly recover the support set if \( T = s \) with probability converging to one. To guarantee the asymptotic convergence of \( \delta_1 \) and \( \delta_2 \), we only need a quite weak assumption, that is, \( p_{\text{max}} = o(n) \). It requires that the magnitude of \( p_{\text{max}} \) should be controlled by sample size \( n \), which implies with a large sample size, we allow large \( p_{\text{max}} \) in the data. Similar variable selection consistent properties are also held for the adaptive GLasso [Wei and Huang, 2010]. Moreover, due to the characteristics of Algorithm 1 if the conditions in Corollary 1 hold and \( T = s \), then Algorithm 1 provably provides an oracle estimator.

3.2 Selection consistency

We show that SGSplcing algorithm can identify the true group subsets under proposed BGIC with an appropriate choice of \( \gamma \). Here we require some mild assumptions on the number of groups \( J \) and the true group subset size \( |A^*| \). We summarize them as Condition (C6) below.

(C6) The number of groups \( J \) satisfies \( J = O(n^k) \) for some positive constant \( k \), and the true group subset size satisfies \( |A^*| = o\left( \frac{n}{p_{\text{max}} \gamma \log(J) + \log(n)} \right) \).

In Condition (C6), we require the assumption on the true group subset size \( |A^*| \) according to the form of BGIC, where \( \gamma \) here is defined in BGIC.
Theorem 2 Assume both the conditions in Theorem 1 and Condition (C6) hold. Under the BGIC, we run Algorithm 2 with maximum model size $T_{\text{max}} = o\left(\frac{n}{p_{\text{max}} \log J}\right)$. If $\gamma > C(1 + \alpha) - \frac{1}{k}$, we have

$$P\left\{ \min_{\hat{A} \neq A^*, \hat{A} \subseteq S} \text{BGIC}_\gamma(\hat{A}) > \text{BGIC}_\gamma(A^*) \right\} = 1 - O(J^{-\alpha} p_{\text{max}}),$$

for some constants $0 < \alpha < 1$ and $C > 0$ depending on model size $T$ in Algorithm 1.

Theorem 2 shows that SGSplicing algorithm under BGIC can identify the true group subset when the number of groups $J$ grows polynomially with sample size $n$ with a high probability. When $p_{\text{max}} = o(J^\alpha)$, a weak assumption as the discussion in Corollary 1, the probability in Theorem 2 converges to one asymptotically. Consequently, we will derive the oracle property of SGSplicing algorithm as Corollary 2 shows.

Corollary 2 Assume the conditions in Theorem 2 hold. Then the solution of SGSplicing algorithm is an oracle estimator with a high probability,

$$P(\hat{\beta} = \hat{\beta}^o) = 1 - O(J^{-\alpha} p_{\text{max}}),$$

for some constant $0 < \alpha < 1$, where $\hat{\beta}^o$ is the least-squares estimator given the true group subset $A^*$.

3.3 Convergence steps and polynomial complexity

We first show that the linear convergence rate of Algorithm 1 and the maximum number of iterations until the selected set covers the true group subset.

Theorem 3 Denote $(A^k, T^k, \beta^k, d^k)$ as the results of Algorithm 1 in the $k$th iteration. Assume Conditions (C1)-(C5) hold. If $T \geq s$, then we have

(i) \[ |2nL(\beta^k) - 2nL(\beta^*)| \leq \mu_k^{\frac{1}{2}} y^2 \]

(ii) \[ A^k \supseteq A^*, \text{ if } k > \log \frac{1}{\mu_T} \frac{\|y\|_2^2}{(1 - \frac{\eta}{2})n(c_* - \frac{\omega^2}{c_*})\theta}, \]

with probability $1 - \delta_1 - \delta_2$.

Part (i) in Theorem 3 establishes the linear convergence rate of Algorithm 1 which shows that the estimation error of loss function in the $k$th iteration can be bounded. These bounds decay geometrically until converge, and the constant $\mu_T < 1$ controls the convergence rate of the error bounds of loss function. Part (ii) show that the output of Algorithm 1 covers the true group subset $A^*$ at most $O(\log \frac{1}{\mu_T} \frac{\|y\|_2^2}{(1 - \frac{\eta}{2})n(c_* - \frac{\omega^2}{c_*})\theta})$ iterations. When $T = s$, the maximum convergence steps are a direct consequence as below.

Corollary 3 Assume the conditions in Theorem 3 hold. If $T = s$, then we have

$$A^k = A^*, \text{ if } k > \log \frac{1}{\mu_T} \frac{\|y\|_2^2}{(1 - \frac{\eta}{2})n(c_* - \frac{\omega^2}{c_*})\theta},$$
with probability $1 - \delta_1 - \delta_2$.

In the case when $T = s$, Theorem 1 has shown us that the output of Algorithm 1 is the same as the oracle least-squares estimator with a high probability. More precisely, under the same conditions, Corollary 3 provides the maximum number of iterations of Algorithm 1 until converges.

Next, we provide the maximum convergence steps of Algorithm 1 when $T > s$ in the following Corollary.

**Corollary 4** Assume the conditions in Theorem 3 hold. If $T > s$, then Algorithm 1 will stop at most $O(\log \frac{1}{\mu_T} \frac{\|y\|_2^2}{n(c_\star - \frac{\omega}{\sqrt{c_\star}})})$ with probability $1 - O(\frac{J - \alpha}{p_{\max}})$ for some constant $0 < \alpha < 1$.

In the case when $T > s$, we can show that, once $A^k$ covers true group subset $A^\star$, the decreases of loss function in the next iteration are no smaller than the threshold $\pi_T$ in Algorithm 1. Based on the results of convergence steps, we further show that the SGSplicing algorithm only requires polynomial computational time with a high probability, which highlights the computational efficiency of our method.

**Theorem 4** Assume Conditions (C1)-(C5) hold. The computational complexity of Algorithm 2 for a given $T_{\max}$ with probability $1 - O(\frac{J - \alpha}{p_{\max}})$ is

$$O(T_{\max} \log \frac{\|y\|_2^2}{p_{\max} \log(nJ)} + \frac{n\|y\|_2^2}{p_{\max} \log(nJ)}) (np + nT_{\max}p_{\max} + JC_{\max})$$

for some constant $0 < \alpha < 1$.

The computational complexity given in Theorem 4 is proportional to sample size $n$, the number of groups $J$, the upper bound $T_{\max}$, the maximum group size $p_{\max}$ and some of the logarithmic terms of them, which illustrates Algorithm 2 terminates in polynomial time.

### 3.4 $\ell_2$ error bounds

We consider the $\ell_2$ error bounds of the estimator during iterations of Algorithm 1 which is substantially different from the error bounds of the penalized regression in literature such as Wei and Huang [2010], Cai et al. [2011] and Wainwright [2019]. We analyze these error bounds for two reasons: firstly, analyzing the error bounds finally output by GSplicing algorithm becomes quite trivial due to Corollary 1 the nature of Algorithm 1; secondly, the analysis of the $\ell_2$ error bounds during Algorithm 1 running helps understand how the estimator gradually approaches the ground truth. Our main result for the $\ell_2$ error bounds is given by:

**Theorem 5** Assume Conditions (C1)-(C5) hold, when $T \geq s$, we have

$$P\left(\|\beta^k - \beta^\star\|_2^2 \leq \frac{1 + \eta + \frac{\omega}{c_\star}}{(1 - \frac{\eta}{2})n(c_\star - \frac{\omega}{c_\star})\mu_T^k\|y\|_2^2}\right) \geq 1 - \delta_1 - \delta_2.$$

Theorem 5 shows that, with a high probability, the estimation error between the estimated coefficients in the $k$th iteration and true coefficients can be bounded by a term which is proportional to the $\|y\|_2^2$. Furthermore, an interesting fact unveiled by Theorem 5 is that the error bounds decay geometrically during iterations, and the constant $\mu_T$ controls the linear convergence rate of
the $\ell_2$ error bounds. Furthermore, we can establish a relationship between the prediction error during iterations.

**Corollary 5** Assume the conditions in Theorem 5 hold, the prediction error in the $k$th iteration satisfies:

$$P \left( \|X(\beta^k - \beta^*)\|_2^2 \leq \frac{1 + \eta + \frac{\omega}{c_2} c^* \mu_T^k \|y\|_2^2}{(1 - \frac{\eta}{2})(c_s - \frac{\omega^2}{c_2})} \right) \geq 1 - \delta_1 - \delta_2.$$ 

4 Numerical experiments

This section is devoted to illustrating the empirical performance of SGSplicing and GGSplicing on both synthetic and real-world datasets. In comparison, we use other state-of-the-art group selection methods: GLasso and group MCP (GMCP), where the latter one generally outperforms group SCAD [Breheny and Huang, 2015]. For GLasso and GMCP, we use R package grpreg [Breheny and Huang, 2015] to implement the procedures.

4.1 Simulation studies

For the simulation setting, we consider two main scenarios, that are, small-$n$-large-$p$ scenario and small-$n$-large-$J$ scenario, which frequently occur in group selection problem.

We generate the synthetic datasets from the linear model:

$$y = X\beta + \sigma_1 \varepsilon,$$

where error term $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ are i.i.d drawn from $\mathcal{N}(0, 1)$, and each row of $X$ are independently sampled from multivariate normal matrix $\tilde{X} \sim \mathcal{MVN}(0, \Sigma)$. Here we consider two structures of correlation matrix $\Sigma$:

- Constant correlation structure: $\Sigma_{ij} = \rho^{i(j \neq j)}$.
- Exponential correlation structure: $\Sigma_{ij} = \rho^{|i-j|}$.

For coefficients $\beta$, we generate coefficients for each group in a more general way as Meier et al. [2008] shows. To generate $\beta_{G_j}$, we generate $p_j + 1$ numbers from $\mathcal{N}(0, \sigma_2^2)$ independently, named that $R_1^j, \ldots, R_{p_j+1}^j$, and define the $i$th $(1 \leq i \leq p_j)$ element of $\beta_{G_j}$ as

$$\beta_{G_j,i} = R_i^j - \frac{1}{p_j + 1} \sum_{i=1}^{p_j+1} R_i^j.$$ 

A large $\sigma_2$ implies a large magnitude of $\beta$, leading to a model with a large signal to noise ratio (SNR); on the contrary, a large $\sigma_1$ enhances the impact of noise on $Y$, leading to a model with low SNR.

To assess the group selection and parameter estimation, we use the following criteria:

- True Positives (TP) : The number of intersection between $\hat{A}$ and $A^*$.
- False Positives (FP) : The number of intersection between $\hat{A}$ and $T^*$.
Mathews Correlation Coefficient (MCC):
\[
MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}},
\]
where TN is the number of intersection between $\hat{I}$ and $I^*$ and FN is the number of intersection between $\hat{I}$ and $A^*$. A large value of MCC illustrates a better group selection performance.

Relative Prediction Error (RePE):
\[
RePE = \|X\hat{\beta} - X\beta^*\|_2/\|X\beta^*\|_2.
\]

Especially, we will give the averaged CPU time (ms) of each algorithm in the rightmost column of the result table, and the standard deviations of these criteria are shown in the parentheses. All simulation results are based on 200 replications.

4.1.1 Small-\(n\)-large-\(p\)

In the small-\(n\)-large-\(p\) scenario, we set the sample size \(n\) and dimensionality \(p'\) of $\hat{X}$ are 500 and 250, respectively, and generate $\hat{X}$ by both constant and exponential correaltional structures. Next we generate 1000 random variables $R_1, \ldots, R_{1000}$ independently from $\mathcal{N}(0, 1)$. Then we generate the design matrix $X$ and group structures as
\[
X_{4(j-i)+k} = \frac{\hat{X}_j + R_{4(j-i)+k}}{\sqrt{2}}, \quad 1 \leq j \leq 250, 1 \leq k \leq 4,
\]
in which the group selection problem has 250 non-overlapping groups with same group sizes, that is, \(p_1 = \cdots = p_{250} = 4\). For the true coefficients, we set \(\sigma_2 = 1\) and the first 10 groups with non-zero coefficients. For the error term, we set \(\sigma_1 = 3\). For SGSplicing and GGSplicing algorithm, we set the maximum model size $T_{\text{max}} = \left\lceil \frac{n}{p_{\text{max} \log J}} \right\rceil$ and $C_{\text{max}} = 2$. For GLasso and GMCP, we set the number of tuning parameter $\lambda$ is 100. All the group selection methods determine the optimal tuning parameter by BIC for fairness. We set the correlation coefficient $\rho = 0.2 : 0.2 : 0.8$ in both simulation.

Both Tables 1 and 2 show the performance of the group selection methods become worse when $\rho$ increases. As we can see, although GLasso has the largest TP among these methods, due to its much larger FP, the MCC of GLasso is quite smaller than GMCP, SGSplicing and GGSplicing. Compared with GMCP, SGSplicing and GGSplicing have similar TP but the FP is significantly smaller than GMCP, which clearly illustrates that our methods can control the FP at reasonably low levels but achieve high TP simultaneously. Furthermore, although the group selection performance of GMCP and our methods are similar, the RePE of our methods are much smaller than GMCP, which demonstrates that our methods are more accurate and efficient on estimation. In terms of speed, GGSplicing is the fastest, significantly accelerating the selection procedures of SGSplicing but still preserving great performances. GLasso is slightly computationally faster than SGSplicing while GMCP is the slowest, and the computational time is much larger than the other three methods.
4.1.2 Small-\(n\)-large-\(J\)

We consider the high dimensional scenario. Different from above settings, we consider a 
discretized settings here. We set sample size \(n = 600\) and dimensionality \(p' = 800\), and generate \(\tilde{X}\) 
by both constant and exponential correlation structures. In this setting, the number of groups \(J\) is 
800, exceeding the sample size \(n = 600\), which is quite a high dimensional group selection problem. 
We discretize the \(i\)th column of \(\tilde{X}\) by 
\[
\tilde{X}_{ij} = \begin{cases} 
0 , & \tilde{X}_{ij} < \Phi^{-1}(\frac{1}{4}), \\
1 , & \Phi^{-1}(\frac{1}{4}) \leq \tilde{X}_{ij} < \Phi^{-1}(\frac{1}{2}), \\
2 , & \Phi^{-1}(\frac{1}{2}) \leq \tilde{X}_{ij} < \Phi^{-1}(\frac{3}{4}), \\
3 , & \tilde{X}_{ij} \geq \Phi^{-1}(\frac{3}{4}), 
\end{cases}
\]
where \(\tilde{X}_{ij}\) is the \(j\)th elements of \(\tilde{X}_i\), and \(\Phi^{-1}(x)\) is the \(x\) quantile of the standard normal distribution. 
Then we take \((I(\tilde{X}_i = 0), I(\tilde{X}_i = 1), I(\tilde{X}_i = 2))\) as the \(i\)th group. By doing so we obtain group 
structures with 800 groups and the same group sizes, that is, \(p_1 = \cdots = p_{800} = 3\). For the true 
coefficients, we set \(\sigma_2 = 2\) and the first 15 groups with non-zero coefficients. For the error term, 
we set \(\sigma_1 = 3\). The settings of input parameters for all the methods are the same as Subsection 
4.1.1. Slightly different from Subsection 4.1.1, here we run SGSplicing and GGSplicing algorithm 
under BGIC with \(\gamma = \frac{1}{4}\) while GLasso and GMCP still use BIC to determine the optimal tuning 
parameter. We set the correlation coefficient \(\rho = 0.2 : 0.2 : 0.8\) in both simulation.

In high dimensional scenario, GLasso and GMCP become inaccurate and unstable. From 
Tables 3-4, we see that GLasso has a similar performance as in the small-\(n\)-large-\(p\) scenario, a 
better TP but a much worse FP than the other methods. SGSplicing and GGSplicing preserve 
a competitive TP but a much smaller FP than GMCP, which leads to better performances on 
MCC and RePE. In terms of speed, GLasso and SGSplicing have similar computational time, and 
GGSplicing is significantly faster while GMCP cost the most running time.

In summary, simulation studies demonstrate that SGSplicing and GGSplicing outperform 
Glasso and GMCP on both group selection and estimation and are generally more efficient, more 
stable and more accurate in the high dimensional scenario. Furthermore, our proposed BGIC has 
an efficient performance on group selection in high dimensional settings.
Table 1: Simulation results of small-$n$-large-$p$ with exponential correlation structure.

| $\rho$ | Method   | TP   | FP   | MCC   | RePE   | time (ms) |
|-------|----------|------|------|-------|--------|-----------|
| 0.2   | GLasso   | 9.83 (0.39) | 15.77 (2.52) | 0.59 (0.04) | 0.33 (0.05) | 298.75    |
|       | GMCP     | 9.19 (1.05) | 0.20 (0.50)  | 0.95 (0.06)  | 0.26 (0.07)  | 1384.20   |
|       | SGSplicing | 9.17 (0.86) | 0.04 (0.18)  | 0.95 (0.05)  | 0.22 (0.05)  | 405.85    |
|       | GGSplicing | 9.17 (0.86) | 0.04 (0.18)  | 0.95 (0.05)  | 0.22 (0.05)  | 113.35    |
| 0.4   | GLasso   | 9.78 (0.45) | 15.76 (2.77) | 0.59 (0.04) | 0.34 (0.05) | 311.90    |
|       | GMCP     | 9.20 (0.92) | 0.32 (0.62)  | 0.94 (0.05)  | 0.26 (0.06)  | 1418.05   |
|       | SGSplicing | 9.14 (0.93) | 0.06 (0.25)  | 0.95 (0.05)  | 0.23 (0.05)  | 419.70    |
|       | GGSplicing | 9.14 (0.93) | 0.05 (0.23)  | 0.95 (0.05)  | 0.23 (0.05)  | 111.65    |
| 0.6   | GLasso   | 9.81 (0.43) | 15.44 (2.82) | 0.60 (0.04)  | 0.33 (0.05)  | 323.45    |
|       | GMCP     | 9.05 (1.07) | 0.17 (0.40)  | 0.94 (0.06)  | 0.27 (0.08)  | 1455.25   |
|       | SGSplicing | 8.98 (1.03) | 0.05 (0.22)  | 0.94 (0.06)  | 0.23 (0.06)  | 432.25    |
|       | GGSplicing | 8.99 (1.01) | 0.05 (0.24)  | 0.94 (0.06)  | 0.23 (0.06)  | 108.25    |
| 0.8   | GLasso   | 9.78 (0.45) | 14.97 (2.91) | 0.60 (0.04)  | 0.34 (0.06)  | 343.95    |
|       | GMCP     | 8.92 (1.19) | 0.24 (0.54)  | 0.93 (0.07)  | 0.27 (0.08)  | 1516.35   |
|       | SGSplicing | 9.02 (0.89) | 0.03 (0.17)  | 0.94 (0.05)  | 0.23 (0.05)  | 461.90    |
|       | GGSplicing | 9.02 (0.86) | 0.03 (0.17)  | 0.95 (0.05)  | 0.23 (0.05)  | 114.55    |

Table 2: Simulation results of small-$n$-large-$p$ with constant correlation structure.

| $\rho$ | Method   | TP   | FP   | MCC   | RePE   | time (ms) |
|-------|----------|------|------|-------|--------|-----------|
| 0.2   | GLasso   | 9.82 (0.40) | 15.46 (2.75) | 0.60 (0.04) | 0.33 (0.05) | 353.00    |
|       | GMCP     | 9.21 (1.05) | 0.16 (0.43)  | 0.95 (0.06)  | 0.26 (0.07)  | 1554.00   |
|       | SGSplicing | 9.13 (0.88) | 0.04 (0.21)  | 0.95 (0.05)  | 0.22 (0.05)  | 493.90    |
|       | GGSplicing | 9.13 (0.89) | 0.04 (0.21)  | 0.95 (0.05)  | 0.22 (0.05)  | 119.60    |
| 0.4   | GLasso   | 9.74 (0.48) | 14.94 (2.76) | 0.60 (0.04)  | 0.34 (0.05)  | 352.95    |
|       | GMCP     | 9.14 (1.12) | 0.27 (0.55)  | 0.94 (0.07)  | 0.26 (0.08)  | 1619.45   |
|       | SGSplicing | 9.12 (0.97) | 0.05 (0.21)  | 0.95 (0.06)  | 0.22 (0.05)  | 499.60    |
|       | GGSplicing | 9.12 (0.97) | 0.06 (0.24)  | 0.95 (0.05)  | 0.23 (0.05)  | 119.35    |
| 0.6   | GLasso   | 9.80 (0.44) | 13.99 (2.84) | 0.62 (0.04)  | 0.34 (0.06)  | 373.60    |
|       | GMCP     | 9.04 (1.15) | 0.17 (0.43)  | 0.94 (0.07)  | 0.26 (0.08)  | 1628.90   |
|       | SGSplicing | 8.99 (1.01) | 0.06 (0.24)  | 0.94 (0.06)  | 0.23 (0.06)  | 513.35    |
|       | GGSplicing | 9.00 (1.01) | 0.04 (0.24)  | 0.94 (0.06)  | 0.23 (0.06)  | 114.50    |
| 0.8   | GLasso   | 9.68 (0.57) | 13.18 (2.63) | 0.62 (0.05)  | 0.34 (0.06)  | 385.75    |
|       | GMCP     | 8.87 (1.24) | 0.22 (0.46)  | 0.93 (0.07)  | 0.27 (0.08)  | 1619.15   |
|       | SGSplicing | 8.95 (0.99) | 0.05 (0.22)  | 0.94 (0.05)  | 0.23 (0.05)  | 515.95    |
|       | GGSplicing | 8.96 (0.94) | 0.04 (0.21)  | 0.94 (0.06)  | 0.23 (0.05)  | 114.80    |
Table 3: Simulation results of small-n-large-J with exponential correlation structure.

| $\rho$ | Method    | TP       | FP       | MCC       | RePE      | time (ms) |
|--------|-----------|----------|----------|-----------|-----------|-----------|
| 0.2    | GLasso    | 14.31 (0.79) | 41.89 (4.32) | 0.48 (0.03) | 0.27 (0.04) | 764.57    |
|        | GMCP      | 13.16 (1.49) | 0.90 (7.15)  | 0.92 (0.07) | 0.22 (0.06) | 2150.77   |
|        | SGSplicing| 12.91 (1.38) | 0.04 (0.21)  | 0.92 (0.05) | 0.18 (0.05) | 732.17    |
|        | GGSplicing| 12.90 (1.38) | 0.04 (0.21)  | 0.92 (0.05) | 0.18 (0.05) | 221.93    |
| 0.4    | GLasso    | 14.27 (0.83) | 42.15 (4.79) | 0.48 (0.03) | 0.27 (0.05) | 788.97    |
|        | GMCP      | 12.91 (1.44) | 0.45 (0.79)  | 0.91 (0.05) | 0.22 (0.06) | 2186.50   |
|        | SGSplicing| 12.76 (1.46) | 0.03 (0.17)  | 0.92 (0.05) | 0.18 (0.05) | 748.07    |
|        | GGSplicing| 12.76 (1.45) | 0.03 (0.17)  | 0.92 (0.06) | 0.18 (0.05) | 223.63    |
| 0.6    | GLasso    | 14.21 (0.80) | 41.25 (4.94) | 0.48 (0.03) | 0.27 (0.05) | 790.97    |
|        | GMCP      | 12.82 (1.43) | 0.54 (0.88)  | 0.90 (0.05) | 0.23 (0.06) | 2171.30   |
|        | SGSplicing| 12.67 (1.39) | 0.02 (0.14)  | 0.92 (0.05) | 0.19 (0.05) | 750.90    |
|        | GGSplicing| 12.65 (1.42) | 0.02 (0.16)  | 0.91 (0.06) | 0.19 (0.05) | 220.50    |
| 0.8    | GLasso    | 13.87 (0.94) | 39.52 (5.52) | 0.48 (0.04) | 0.28 (0.05) | 793.37    |
|        | GMCP      | 12.08 (1.65) | 0.75 (1.01)  | 0.87 (0.07) | 0.23 (0.06) | 2206.20   |
|        | SGSplicing| 12.12 (1.55) | 0.06 (0.24)  | 0.89 (0.06) | 0.19 (0.05) | 755.00    |
|        | GGSplicing| 12.18 (1.57) | 0.06 (0.27)  | 0.90 (0.06) | 0.19 (0.05) | 226.40    |

Table 4: Simulation results of small-n-large-J with constant correlation structure.

| $\rho$ | Method    | TP       | FP       | MCC       | RePE      | time (ms) |
|--------|-----------|----------|----------|-----------|-----------|-----------|
| 0.2    | GLasso    | 14.31 (0.81) | 38.99 (4.61) | 0.49 (0.03) | 0.27 (0.04) | 745.30    |
|        | GMCP      | 13.12 (1.42) | 0.43 (0.82)  | 0.92 (0.05) | 0.21 (0.07) | 2012.80   |
|        | SGSplicing| 12.90 (1.37) | 0.03 (0.17)  | 0.92 (0.05) | 0.18 (0.04) | 754.45    |
|        | GGSplicing| 12.90 (1.37) | 0.04 (0.18)  | 0.92 (0.05) | 0.18 (0.04) | 198.30    |
| 0.4    | GLasso    | 14.14 (0.94) | 34.98 (4.33) | 0.51 (0.03) | 0.27 (0.05) | 741.95    |
|        | GMCP      | 12.89 (1.58) | 0.45 (0.74)  | 0.91 (0.05) | 0.21 (0.05) | 2057.35   |
|        | SGSplicing| 12.68 (1.47) | 0.07 (0.26)  | 0.91 (0.06) | 0.19 (0.05) | 744.00    |
|        | GGSplicing| 12.68 (1.46) | 0.07 (0.26)  | 0.91 (0.05) | 0.19 (0.05) | 201.75    |
| 0.6    | GLasso    | 13.95 (0.97) | 30.55 (4.50) | 0.53 (0.04) | 0.28 (0.05) | 749.35    |
|        | GMCP      | 12.64 (1.78) | 0.53 (1.07)  | 0.90 (0.07) | 0.22 (0.06) | 2109.55   |
|        | SGSplicing| 12.44 (1.70) | 0.07 (0.28)  | 0.90 (0.07) | 0.19 (0.05) | 780.15    |
|        | GGSplicing| 12.44 (1.71) | 0.07 (0.28)  | 0.90 (0.07) | 0.19 (0.05) | 210.40    |
| 0.8    | GLasso    | 13.50 (1.19) | 25.59 (4.61) | 0.55 (0.05) | 0.28 (0.06) | 749.40    |
|        | GMCP      | 12.17 (1.76) | 2.36 (12.84) | 0.86 (0.10) | 0.23 (0.09) | 2065.80   |
|        | SGSplicing| 11.86 (1.66) | 0.04 (0.18)  | 0.88 (0.07) | 0.19 (0.05) | 782.70    |
|        | GGSplicing| 11.81 (1.64) | 0.06 (0.26)  | 0.88 (0.06) | 0.20 (0.05) | 199.45    |
4.1.3 Computational complexity analysis

We study the computational complexity for SGSplicing and GGSplicing algorithms, and we are particularly interested in the effects of group numbers \( J \) and group size on the computational complexity. We adopt the experiments setting in subsection 4.1.1 with \( \sigma_1 = \sigma_2 = 1 \) and true group subset size \( s = 10 \). For the effect of group numbers \( J \), we fix the group size \( p_1 = \cdots = p_J = 4 \) and sample size \( n = 300 \) when \( J \) varies from 300 to 500. For the effect of group size, we fix the number of groups \( J = 300 \) and sample size \( n = 300 \) when the equal group size varies from 2 to 10. Figure 2A shows the runtime of both our methods linearly grows as the increase of group numbers.

![Figure 2](image)

Figure 2: (A) Number of groups (x-axis) versus runtime (y-axis) scatterplot. (B) Group size (x-axis) versus runtime (y-axis) scatterplot. In both (A) and (B), the green and red straight lines are characterized by equation \( y = a + bx \) with different slopes and intercepts.

Furthermore, the slopes of the two straight lines demonstrate that the computational complexity of SGSplicing has a stronger growth trend than GGSplicing when group numbers increase. From Figure 2B, group sizes show a similar impact on runtime like group numbers. From the two Figures, we conclude that the empirical runtime analysis matches the theoretical result presented in Theorem 4.

4.2 Real data analysis

We illustrate the competitive performance of our method by two real-world datasets.

4.2.1 Gene expression data analysis

We consider a genomic dataset in a rat eye disease study by Scheetz et al. [2006]. The dataset consists of 120 twelve-week-old male rats, which collects the expression of TRIM32, a gene that has been shown to cause Bardet-Biedl syndrome, and other 18975 related genes that potentially influence the expression of TRIM32. Although there are numerous potential genes, we expect only
a small subset of genes to be related to TRIM32, which inspires us to focus our attention on a small subset of these 18975 genes [Huang et al., 2010, Breheny and Huang, 2015].

Owing to ultra-high dimensionality, we first select 1000 genes by a sure independence screening procedure using ball correlation by R package Ball [Pan et al., 2019, Zhu et al., 2021], and consider a five-term natural cubic spline basis expansion of these genes. By doing so, the objective problem is a group selection problem with sample size \( n = 120 \) and dimensionality \( p = 5000 \), which is a sparse, high-dimensional group selection problem. For GLasso and GMCP, we choose the optimal tuning parameter by BIC and set the number of tuning parameter \( \lambda \) is 100. For SGSplicing and GGSplicing, we run both the algorithms with \( C_{\text{max}} = 2 \) under our proposed BGIC with \( \gamma = \frac{1}{2} \). To assess the performance of these methods, the 120 rats are randomly split into a training set with 100 samples and a test set with the remaining 20 samples, and then compute the mean square error (MSE) in the test set. We replicate these random-splitting procedures 200 times and compute the average of the numbers of genes selected and MSE in the test set. The results are displayed in Table 5.

Table 5: Analysis result for TRIM32 dataset based on 200 replications.

| Method     | Number of groups | 10×MSE  |
|------------|------------------|---------|
| GLasso     | 27.43 (3.53)     | 0.96 (0.39) |
| GMCP       | 1.18 (0.67)      | 0.99 (0.33) |
| SGSplicing | 1.01 (0.10)      | 0.94 (0.26) |
| GGSplicing | 1.01 (0.10)      | 0.94 (0.26) |

From Table 5, we see that GLasso selects far more groups than the other two methods, but this does not lead to the best prediction performance. GMCP selects a much fewer number of groups than GLasso, and has the largest averaged MSE among these methods. Compared with GLasso and GMCP, although SGSplicing and GGSplicing select the sparsest model, they have competitive, even better, performance on prediction.

4.2.2 Asthma data analysis

We analyze the asthma dataset that can be found in R package SNPassoc [González et al., 2007]. This dataset consists of 1578 subjects with 51 genetic factors, i.e., single-nucleotide polymorphism (SNP) markers, and five epidemiological predictors: country, gender, age, body mass index (BMI), and smoke status. Our primary interest is to design a model to automatically classified asthma patients (cases) and healthy subjects (controls). Among these 1578 individuals, there are 1238 cases and 340 controls.

Next, we introduce the pre-processing detail for predictors. We discretize these two continuous variables, age and BMI, into categorical variables. For the age variable, it ranges from 26.4 to 56.5. So we consider dividing the samples into young group and middle-aged group. For the BMI variable, we divide it into four categories: underweight (Below 18.5), normal (18.5 - 24.9), overweight (25.0 - 29.9) and obese (30.0 and Above). We code each SNP markers as 0, 1, 2 for homozygous, heterozygous, and mutation homozygous genotypes, respectively, and retain SNP markers with acceptably low rates of missing data (< 10%) and high minor allele frequency (> 10%). For these SNP markers, we impute the missing data based on Wright’s equilibrium law.
We expand the SNP markers and remaining categorical variables with a group of dummy variables, and then we obtain the design matrix with dimensionality $p = 112$.

We divide the dataset in a stratified manner by 60% for training and 40% for testing, and fit a grouped logistic regression model to predict the status in the test set. All the methods conducted model selection by BIC. We compute the accuracy in the test set for all the methods and replicate the procedures 200 times. The results are presented in Table 6.

Table 6: Analysis result for asthma dataset based on 200 replications.

| Method     | Number of groups | Number of variables | Accuracy   |
|------------|------------------|---------------------|------------|
| GLasso     | 5.47 (1.72)      | 16.89 (3.51)        | 0.80 (0.01)|
| GMCP       | 2.48 (1.32)      | 11.26 (2.34)        | 0.80 (0.01)|
| SGSplcing  | 1.26 (0.59)      | 9.38 (0.91)         | 0.80 (0.01)|
| GGSplicing | 1.41 (0.85)      | 9.66 (1.47)         | 0.80 (0.01)|

From Table 6, we see that these three methods have similar classification performance in the test set. Among these methods, SGSplcing and GGSplicing still select the sparsest model while GLasso selects the largest number of groups and variables. This result illustrates that our method achieves a competitive prediction accuracy with fewer groups, demonstrating our method’s efficiency.

To better understand the genetic factors that influence asthma, we calculate their group selection frequencies via stability selection [Meinshausen and Bühlmann, 2010]. In specific, we randomly pick out a subsample of size $\lfloor n/2 \rfloor$ to train all methods and select groups. Repeating the procedure 400 times, we can determine the selection frequency by the frequency of being selected into the model. We report the groups with selection frequency at least 90%, a threshold recommended by [Meinshausen and Bühlmann, 2010]. Note that, due to the discontinuous solution path of GGSplicing, we do not consider it here.

Table 7: Selection frequencies of groups selected by each method. An empty value implies a selection frequency less than 90%.

| Variable    | GLasso | GMCP | SGSplcing |
|-------------|--------|------|-----------|
| country     | 1.00   | 1.00 | 1.00      |
| gender      | 1.00   | 0.98 | 0.95      |
| BMI         | 1.00   | 0.98 | 1.00      |
| rs1422993   | 1.00   | 0.94 | 0.99      |
| rs2303063   | 0.99   |     | 0.97      |
| age         | 0.98   |     |           |
| rs324381    | 0.96   |     |           |
| rs1430094   | 0.96   |     |           |
| rs2274276   | 0.95   |     |           |
| rs324396    | 0.95   |     |           |
| rs765023    | 0.93   | 0.93|           |
| rs1430090   | 0.92   |     |           |

The stability selection results are summarized in Table 7, from which we see that SGSplcing
selects 5 groups with a high frequency while GMCP selects a similar model size but with a lower frequency on genetic factors, and GLasso selects much more than 5 groups. Interestingly, rs1422993 is frequently selected by all the methods, indicating its strong association with the response.

5 Conclusion

GSplicing algorithm is an efficient group selection method for non-overlapping group structures that can select an exact model size. It is derived based on the augmented Lagrangian form of the BGSS problem and requires the decreases of loss function to exceed a predetermined threshold. Thus GSplicing algorithm can guarantee finite-step convergence without any assumptions. To better adapt to the high dimensional scenario, we propose a novel information criterion called BGIC, of which the ordinary BIC is a special case. Coupled with BGIC, we develop an adaptive algorithm called SGSplicing to determine the optimal model size. Furthermore, we show that SGSplicing solves the BGSS and recovers the true group subset in polynomial time under mild conditions. Motivated by the unimodal-like solution path of SGSplicing algorithm, we also develop an efficient heuristic algorithm, GGSplicing, which significantly accelerates the selection procedures of SGSplicing. More importantly, we conduct a complete theoretical analysis of the output of GSplicing under some appropriate assumptions, i.e., support recovery, the linear convergence rate, the convergence steps and the $\ell_2$ error bounds of the estimated coefficients. Finally, the numerical experiments illustrate that our method is competitive with or outperforms state-of-the-art group selection methods.

Moreover, group selection for overlapping group structures is also an attractive and practical research [Zhao et al. 2009, Jacob et al. 2009, Jenatton et al. 2011]. Extending BGSS for overlapping structures is a constructive approach to solve the overlapping group selection problem but is not analyzed in this paper. Therefore, it is of interest to study the extension of GSplicing algorithm for overlapping group structures in the future.

A Proof

A.1 Proof of Lemma 1

Proof 1 For part (i), note that

$$L(\beta^{A^k,j}) - L(\beta^k) = \frac{1}{2n} \| y - X \beta^{A^k,j} \|_2^2 - \frac{1}{2n} \| y - X \beta^k \|_2^2$$

$$= \frac{1}{2n} (\beta^{A^k,j} - \beta^k)^\top X^\top (\beta^{A^k,j} - \beta^k)$$

$$= \frac{1}{2n} (\beta^k_{G_j})^\top X_{G_j}^\top X_{G_j} \beta^k_{G_j}$$

$$= \frac{1}{2} \| \beta^k_{G_j} \|_2^2,$$

where the third equality follows from the definition of $\beta^{A^k,j}$ and the last equality uses the assumption of $X_{G_j}^\top X_{G_j}/n = I_{p_j}$.

Next, we prove part (ii). From the property of least-squares estimator, we can write $t^k_j = (X_{G_j}^\top X_{G_j})^{-1} X_{G_j}(y - X \beta^k) = (X_{G_j}^\top X_{G_j}/n)^{-1} d^k_{G_j}$. Following similar derivation in the proof of part
(i), we have
\[
L(\beta^k) - L(\beta^k + t_j^k) = \frac{1}{2n} \|y - X\beta^k\|^2 - \frac{1}{2n} \|y - X(\beta^k + t_j^k)\|^2 \\
= \frac{1}{2n} (t_j^k)^\top X_{G_j}^\top X_{G_j} t_j^k \\
= \frac{1}{2} (d_j^k)^\top (X_{G_j}^\top X_{G_j}/n)^{-1} d_j^k \\
= \frac{1}{2} \|d_j^k\|^2, 
\]
where the third equality uses the expression of \(t_j^k\).

\[\Box\]

A.2 Proof of Lemma 2

**Proof 2** Denote \(i \in S_{C,1}^k\) and \(j \in S_{C,2}^k\). Since \(i \in A^k\) and \(j \in I^k\), we have
\[
\frac{1}{\kappa^2 p_j} \|d_j^k\|^2_2 = \frac{1}{p_j} \|\beta^k_{G_j}\|_2 + \frac{1}{\kappa} \|d_j^k\|^2_2 \\
\leq \frac{1}{p_i} \|\beta^k_{C_{i+1,1}}\|_2 + \frac{1}{\kappa} \|d_j^k\|^2_2 = \frac{1}{p_i} \|\beta^k_{C_{i+1,1}}\|_2. 
\]

By some simple algebra,
\[
\kappa^2 \leq \frac{\|d_j^k\|^2_2/p_j}{\|\beta^k_{C_{i+1,1}}\|_2/p_i}. 
\]

Similarly, given the exchange subset size as \(C + 1\), we have
\[
\kappa^2 \leq \frac{\|d_j^k\|^2_2/p_j}{\|\beta^k_{C_{i+1,1}}\|_2/p_i} \leq \frac{\min_{j \in S_{C,2}^k} \|d_j^k\|^2_2/p_j}{\max_{i \in S_{C,1}^k} \|\beta^k_{C_{i+1,1}}\|_2/p_i}. 
\]

Combine (7) and (8),
\[
\frac{\min_{j \in S_{C,2}^k} \|d_j^k\|^2_2/p_j}{\max_{i \in S_{C,1}^k} \|\beta^k_{C_{i+1,1}}\|_2/p_i} \leq \kappa^2 \leq \frac{\min_{j \in S_{C,2}^k} \|d_j^k\|^2_2/p_j}{\max_{i \in S_{C,1}^k} \|\beta^k_{C_{i+1,1}}\|_2/p_i}. 
\]

Therefore,
\[
\kappa \in \left( \frac{\min_{j \in S_{C,2}^k} \|d_j^k\|^2_2/p_j}{\max_{i \in S_{C,1}^k} \|\beta^k_{C_{i+1,1}}\|_2/p_i} \right)^{\frac{1}{2}}, \left( \frac{\min_{j \in S_{C,2}^k} \|d_j^k\|^2_2/p_j}{\max_{i \in S_{C,1}^k} \|\beta^k_{C_{i+1,1}}\|_2/p_i} \right)^{\frac{1}{2}}. 
\]

\[\Box\]

Before detailing the proofs of the main theorems, we first provide three useful Lemmas as follows. To simplify the proof, we provide some useful inequalities occurred frequently in following proofs in Lemma 3. In Lemma 4, we provide the upper bound of \(\|\beta^*_{A_{12}}\|_2\) and \(\|\beta^*_{I_{12}}\|_2\) in terms of \(\|\beta^*_{I_1}\|_2\) with a high probability. In Lemma 5, we show that the components related to error term \(\epsilon\) can be controlled by \(n\) and \(\|\beta^*_{I_1}\|_2\) with a high probability. In both Lemma 4 and Lemma 5, we
use Conditions (C1) and (C2) and Hoeffding inequality frequently.

Here we give some useful notations. Denote

\[ A_1 = \hat{A} \cap A^*, \quad A_2 = \hat{A} \cap I^*, \]
\[ I_1 = \hat{I} \cap A^*, \quad I_2 = \hat{I} \cap I^*. \]

Let the size of exchange subset \( C = |I_1| \) and the exchange subsets be

\[ S_1 = \{ j \in \hat{A} : \sum_{i \in A} I \left( \frac{1}{p_j} \| \hat{\beta}_{G_j} \|_2^2 \geq \frac{1}{p_i} \| \hat{\beta}_{G_i} \|_2^2 \right) \leq C \} \]

and

\[ S_2 = \{ j \in \hat{I} : \sum_{i \in I} I \left( \frac{1}{p_j} \| \hat{d}_{G_j} \|_2^2 \leq \frac{1}{p_i} \| \hat{d}_{G_i} \|_2^2 \right) \leq C \}, \]

where \( S_1 \) represents the subset exchanging from \( \hat{A} \) to \( \hat{I} \), and \( S_2 \) represents the subset exchanging from \( \hat{I} \) to \( \hat{A} \). Denote

\[ A_{11} = A_1 \cap (S_1)^c, \quad A_{12} = A_1 \cap S_1, \]
\[ A_{21} = A_2 \cap (S_1)^c, \quad A_{22} = A_2 \cap S_1, \]

and

\[ I_{11} = I_1 \cap S_2, \quad I_{12} = I_1 \cap (S_2)^c, \]
\[ I_{21} = I_2 \cap S_2, \quad I_{22} = I_2 \cap (S_2)^c, \]

where each subset defines the subset exchanging or preserving in the subset denoted at the beginning of this section.

Let \( \hat{A} = (\hat{A} \setminus S_1) \cup S_2 \) and \( \hat{I} = (\hat{I} \setminus S_2) \cup S_1 \) be the selected set and unselected set after exchanging, respectively. \( H_A = X_A (X_A^T X_A)^{-1} X_A^T \) is the hat matrix and \( X_{G_j}^{(i)} \) be the ith column of sub-matrix \( X_{G_j} \). In addition, we denote the least-squares estimator constrained in \( \cup_{j \in A} G_j \) by \( \hat{\beta}_A \), and denote \( \beta_A^* \) as the true coefficients constrained in \( \cup_{j \in A} G_j \). For simplicity, given \( A \subseteq B \), we denote \( e_A u_B \) as the vector \( u_B \) constrained in \( \cup_{j \in A} G_j \), where \( e_A \), a vector with the same length as \( u_B \), supports on \( \cup_{j \in A} G_j \) and all nonzero components are one.

### A.3 Proofs of Lemma 3

**Lemma 3** Let \( A \) and \( B \) be disjoint subsets of \( S \), with both \( |A| \) and \( |B| \) is smaller than \( \tau \). Assume \( X \) satisfies SRC with order \( 2\tau \). Then for any \( u \in \mathbb{R}^{|A|} \), we have

\[ \sqrt{nc} \| u \|_2 \leq \| X_A u \|_2 \leq \sqrt{nc} \| u \|_2 \]  \hspace{1cm} (9)

\[ nc \| u \|_2 \leq \| X_A^T X_A u \|_2 \leq nc \| u \|_2 \]  \hspace{1cm} (10)
\[
\frac{\|u\|_2}{nc^*} \leq \|(X_A^T X_A)^{-1}u\|_2 \leq \frac{\|u\|_2}{nc^*}, \quad (11)
\]

\[
n(c_* - \frac{\omega^2}{c_*})\|u\|_2 \leq \|X_A^T (I - H_B) X_A u\|_2 \leq nc^*\|u\|_2, \quad (12)
\]

\[
\frac{\|u\|_2}{nc^*} \leq \|(X_A^T (I - H_B) X_A)^{-1}u\|_2 \leq \frac{\|u\|_2}{n(c_* - \frac{\omega^2}{c_*})}, \quad (13)
\]

where \(c_*, c^*\) and \(\omega\) are defined in Condition (C2). In addition, \(c_*\) decreases while \(c^*\) increases as \(\tau\) increases, \(\omega\) is bounded by \((1 - c_*) \lor (c^* - 1)\) and is an increasing function of \(\tau\).

**Proof 3** The assumption of SRC implies the spectrum of \(X_A^T X_A/n\) can be controlled in \([c_*, c^*]\), which obtains (9)-(11). When \(\tau\) increases, \(c_*\) decreases and \(c^*\) increases obviously. Moreover, the spectrum of \(X_A^T X_B/n\) can be bounded by \((1 - c_*) \lor (c^* - 1)\) since \(X_A^T X_B/n\) is a sub-matrix of \(X_{AUB}^T X_{AUB}/n - I\), where \(I\) is an identity matrix with rank \#\{A \cup B\}.

For (12), because \((I - H_B)\) is an idempotent matrix, so its spectrums are either 0 or 1. Therefore, we have \(\|X_A^T (I - H_B) X_A u\|_2 \leq \|X_A^T X_A u\|_2 \leq nc^*\|u\|_2\), which proves the right side of (12). For the left side of (12), note that

\[
X_A^T (I - H_B) X_A = X_A^T X_A - X_A^T X_B (X_B^T X_B)^{-1} X_B X_A.
\]

Then we have

\[
\|X_A^T (I - H_B) X_A u\|_2 \geq \|X_A^T X_A u\|_2 - \|X_A^T X_B (X_B^T X_B)^{-1} X_B X_A u\|_2
\]

\[
\geq nc_*\|u\|_2 - n\frac{\omega^2}{c_*}\|u\|_2
\]

\[
=n(c_* - \frac{\omega^2}{c_*})\|u\|_2,
\]

where the first inequality follows from the triangle inequality, and the second inequality follows from (10), (11) and the definition of \(\omega\).

This proves (12), and (13) is a direct consequence of (12).

\(\Box\)

**A.4 Proofs of Lemma 4**

**Lemma 4** Assume the conditions in Theorem 1 hold, then with probability \(1 - \delta_1\), we have

\[
\|\beta_{A12}^*\|_2 \leq (1 + \eta)\frac{(1 + t)\omega}{c_*}\|\beta_{I1}^*\|_2 \quad (14)
\]

and

\[
\|\beta_{I12}^*\|_2 \leq (1 + \eta)\frac{(1 + t)\omega}{c_* - \frac{\omega^2}{c_*}}\|\beta_{I1}^*\|_2. \quad (15)
\]
Proof 4 By the definition of $S_1$, we have
\[
\frac{1}{|A_{12}|} \sum_{j \in A_{12}} \frac{1}{p_j} \| \hat{\beta}_{G_j} \|_2^2 \leq \frac{1}{|A_{21}|} \sum_{j \in A_{21}} \frac{1}{p_j} \| \hat{\beta}_{G_j} \|_2^2.
\]

Then,
\[
\| \hat{\beta}_{A_{12}} \|_2 \leq t \sqrt{\frac{|A_{12}|}{|A_{21}|}} \| \hat{\beta}_{A_{21}} \|_2.
\]

To obtain the expressions of $\hat{\beta}_{A_1}$ and $\hat{\beta}_{A_2}$, we first constrain $X_{A_1}$ and $X_{A_2}$ in the orthogonal complement of $X_{A_2}$ and $X_{A_1}$, respectively, and then use the least-squares estimation method. By doing these procedures, we have
\[
\hat{\beta}_{A_1} = (X_{A_1}^\top (I - H_{A_2})X_{A_1})^{-1}X_{A_1}^\top (I - H_{A_2})y = \beta_{A_1}^* + (X_{A_1}^\top (I - H_{A_2})X_{A_1})^{-1}X_{A_1}^\top (I - H_{A_2})X_{I_1}^* \beta_{I_1}^* + \varepsilon
\]

and
\[
\hat{\beta}_{A_2} = (X_{A_2}^\top (I - H_{A_1})X_{A_2})^{-1}X_{A_2}^\top (I - H_{A_1})y = \beta_{A_2}^* + (X_{A_2}^\top (I - H_{A_1})X_{A_2})^{-1}X_{A_2}^\top (I - H_{A_1})X_{I_2}^* \beta_{I_2}^* + \varepsilon
\]

where the second equalities in both (17) and (18) follows from the projection property of $I - H_{A_2}$ and $I - H_{A_1}$, respectively, and the last equality in (18) follows from the definition of $A_2$. By (17), we have
\[
\| \hat{\beta}_{A_1} \|_2 \geq \| \beta_{A_1}^* \|_2 - \| e_{A_{12}}^\top (X_{A_1}^\top (I - H_{A_2})X_{A_1})^{-1}X_{A_1}^\top (I - H_{A_2})X_{I_1}^* \beta_{I_1}^* \|_2 - \| e_{A_{12}} (X_{A_1}^\top (I - H_{A_2})X_{A_1})^{-1}X_{A_1}^\top (I - H_{A_2})\varepsilon \|_2
\] \[
\geq \| \beta_{A_1}^* \|_2 - \frac{\omega}{c_\varepsilon} \| \beta_{I_1}^* \|_2 - \| e_{A_{12}} (X_{A_1}^\top (I - H_{A_2})X_{A_1})^{-1}X_{A_1}^\top (I - H_{A_2})\varepsilon \|_2.
\]

where the first inequality follows from the triangle inequality, and the second inequality follows from (13) and the definition of $\omega$. By (18), we have
\[
\| \hat{\beta}_{A_2} \|_2 \leq \| e_{A_{21}}^\top (X_{A_2}^\top (I - H_{A_1})X_{A_2})^{-1}X_{A_2}^\top (I - H_{A_1})X_{I_2}^* \beta_{I_2}^* \|_2 + \| e_{A_{21}} (X_{A_2}^\top (I - H_{A_1})X_{A_2})^{-1}X_{A_2}^\top (I - H_{A_1})\varepsilon \|_2
\] \[
\leq \frac{\omega}{c_\varepsilon} \| \beta_{I_2}^* \|_2 + \| e_{A_{21}} (X_{A_2}^\top (I - H_{A_1})X_{A_2})^{-1}X_{A_2}^\top (I - H_{A_1})\varepsilon \|_2,
\]

26
where uses the same technique as the proof of (19) shows. Combine (16), (19) and (20),

$$
\|\beta^*_{A_{12}}\|_2 \leq (1 + \frac{t}{c_s} \omega) \|\beta^*_{Z_{11}}\|_2 + \\
\|e^T_{A_{12}} (X^T_{A_1} (I - H_{A_2}) X_{A_1})^{-1} X^T_{A_1} (I - H_{A_2}) e\|_2 + \\
t \|e^T_{A_{21}} (X^T_{A_2} (I - H_{A_1}) X_{A_2})^{-1} X^T_{A_2} (I - H_{A_1}) e\|_2.
$$

(21)

Next we bound the components related to error term $e$ in terms of $\|\beta^*_{Z_{11}}\|_2$,

$$
P(\|e^T_{A_{12}} (X^T_{A_1} (I - H_{A_2}) X_{A_1})^{-1} X^T_{A_1} (I - H_{A_2}) e\|_2 > \frac{n \omega (c_s - \frac{\omega^2}{c_s})}{c_s} \|\beta^*_{Z_{11}}\|_2)
\leq P(\|X^T_{A_1} (I - H_{A_2}) e\|_2 > \frac{n \omega (c_s - \frac{\omega^2}{c_s})}{c_s} \|\beta^*_{Z_{11}}\|_2)
\leq \sum_{j \in A_1} \sum_{i=1}^{p_j} P(\|X^{(i)}_{G_j} e\|_2 > \frac{n \omega (c_s - \frac{\omega^2}{c_s})}{c_s \sqrt{\#\{A_1\}}} \|\beta^*_{Z_{11}}\|_2)
\leq 2 J p_{\text{max}} \exp \{-n C_1 \vartheta / sp_{\text{max}}\}
\leq \frac{\delta_1}{2},
$$

(23)

where the first inequality follows from the right side of (13), the second inequality follows from the idempotency of $I - H_{A_1}$ and simple algebra, and the third inequality follows from the Hoeffding’s inequality and $\#\{A_1\} \leq sp_{\text{max}}$ in which $C_1$ is some positive constant depending on $T$. Use the same technique in (23),

$$
P(\|e^T_{A_{21}} (X^T_{A_2} (I - H_{A_1}) X_{A_2})^{-1} X^T_{A_2} (I - H_{A_1}) e\|_2 > \frac{n \omega (c_s - \frac{\omega^2}{c_s})}{c_s} \|\beta^*_{Z_{11}}\|_2)
\leq 2 J p_{\text{max}} \exp \{-n C_1 \vartheta / sp_{\text{max}}\}
\leq \frac{\delta_1}{2}.
$$

(24)

Combine (22), (23) and (24),

$$
P(\|\beta^*_{A_{12}}\|_2 \leq (1 + \frac{t}{c_s} \omega) \|\beta^*_{Z_{11}}\|_2) \geq 1 - \delta_1,
$$

27
which proves (14).
Next we give the proof of (15). By the definition of $S_2$, we have
\[
\frac{1}{|I_{12}|} \sum_{j \in I_{12}} \frac{1}{p_j} \| \hat{d}_{G_j} \|_2^2 \leq \frac{1}{|I_{21}|} \sum_{j \in I_{21}} \frac{1}{p_j} \| \hat{d}_{G_j} \|_2^2.
\]
Since $|I_{11}| + |I_{21}| = |S_2| = |I_1| = |I_{11}| + |I_{12}| = C$, we have $|I_{21}| = |I_{12}|$. Then we obtain,
\[
\| \hat{d}_{I_{12}} \|_2 \leq t \| \hat{d}_{I_{21}} \|_2. \tag{25}
\]
Note that
\[
n \| \hat{d}_{I_{12}} \|_2 = \| X_{I_{12}}^T (I - H_{\bar{A}}) y \|_2 \\
= \| X_{I_{12}}^T (I - H_{\bar{A}}) (X_{I_{12}} \beta_{I_{12}}^* + \varepsilon) \|_2 \\
\geq \| X_{I_{12}}^T (I - H_{\bar{A}}) X_{I_{12}} \beta_{I_{12}}^* \|_2 - \| X_{I_{12}}^T (I - H_{\bar{A}}) X_{I_{11}} \beta_{I_{11}}^* \|_2 - \| X_{I_{12}}^T (I - H_{\bar{A}}) \varepsilon \|_2 \\
\geq n (c_s - \frac{\omega^2}{c_s}) \| \beta_{I_{12}}^* \|_2 - n \omega \| \beta_{I_{11}}^* \|_2 - \| X_{I_{12}}^T (I - H_{\bar{A}}) \varepsilon \|_2, \tag{26}
\]
where the second equality follows from the projection property of $I - H_{\bar{A}}$, the first inequality follows from the triangle inequality, and the second inequality follows from (12) and the definition of $\omega$. Similar to (26), we have
\[
n \| \hat{d}_{I_{21}} \|_2 = \| X_{I_{21}}^T (I - H_{\bar{A}}) y \|_2 \\
= \| X_{I_{21}}^T (I - H_{\bar{A}}) (X_{I_{21}} \beta_{I_{21}}^* + \varepsilon) \|_2 \\
\leq \| X_{I_{21}}^T (I - H_{\bar{A}}) X_{I_{21}} \beta_{I_{21}}^* \|_2 + \| X_{I_{12}}^T (I - H_{\bar{A}}) \varepsilon \|_2 \\
\leq n \omega \| \beta_{I_{21}}^* \|_2 + \| X_{I_{21}}^T (I - H_{\bar{A}}) \varepsilon \|_2. \tag{27}
\]
Combine (25), (26) and (27),
\[
(1 + t) n \omega \| \beta_{I_{12}}^* \|_2 + t \| X_{I_{21}}^T (I - H_{\bar{A}}) \varepsilon \|_2 \\
\geq n (c_s - \frac{\omega^2}{c_s}) \| \beta_{I_{12}}^* \|_2 - \| X_{I_{12}}^T (I - H_{\bar{A}}) \varepsilon \|_2. \tag{28}
\]
Note that
\[
P(\| X_{I_{21}}^T (I - H_{\bar{A}}) \varepsilon \|_2 \geq n \omega \| \beta_{I_{21}}^* \|_2) \\
\leq \sum_{j \in I_{21}} \sum_{i=1}^{p_j} P(\| (X_{G_j}^{(i)})^\top \varepsilon \|_2 > \frac{n \omega c_s}{\sqrt{\# I_{21}}}) \| \beta_{I_{21}}^* \|_2 \\
\leq 2 J_{p_{\text{max}}} \exp \{-n C_1 \theta / sp_{\text{max}}\} \\
\leq \frac{\delta_1}{2}, \tag{29}
\]
where the first inequality follows from the idempotency of $I - H_{\hat{A}}$ and simple algebra, and the second inequality follows from $\#\{I_{21}\} \leq \text{sp}_{\text{max}}$. Here positive constant $C_1$ is related to $T$. Similarly, we can bound $\|X_{I_{12}}^\top (I - H_{\hat{A}})\varepsilon\|_2$ as (29) shows,

$$P(\|X_{I_{12}}^\top (I - H_{\hat{A}})\varepsilon\|_2 \geq \eta c_* \|\beta_{I_1}^*\|_2)$$

$$\leq 2Jp_{\text{max}} \exp\{-nC_1 \vartheta / \text{sp}_{\text{max}}\}$$  \hspace{1cm} (30)

Combine (28), (29) and (30),

$$P(\|\beta_{I_{12}}^*\|_2 \leq (1 + \eta) \frac{(1 + t) \omega}{c_* - \frac{\omega^2}{c_*}} \|\beta_{I_1}^*\|_2) \geq 1 - \delta_1.$$

This proves (15), which completes the proof of Lemma 4.

\[\square\]

A.5 Proofs of Lemma 5

**Lemma 5** Assume the conditions in Theorem 1 hold, then we have

$$|(X_{I_{12}} \beta_{I_1}^*)^\top (I - H_{\hat{A}})\varepsilon| \leq \frac{\eta}{8} n(c_* - \frac{\omega^2}{c_*}) \|\beta_{I_1}^*\|_2^2,$$  \hspace{1cm} (31)

$$|(X_{A_{12} \cup I_{12}} \beta_{A_{12} \cup I_{12}}^*)^\top (I - H_{\hat{A}})\varepsilon| \leq \frac{\eta}{8} n(c_* - \frac{\omega^2}{c_*}) \|\beta_{I_1}^*\|_2^2,$$  \hspace{1cm} (32)

with probability $1 - \delta_1$, and

$$\|H_{\hat{A}}\varepsilon\|_2 \leq \sqrt{\frac{\eta}{4} n(c_* - \frac{\omega^2}{c_*}) \|\beta_{I_1}^*\|_2^2}$$  \hspace{1cm} (33)

$$\|H_{\hat{A}}\varepsilon\|_2 \leq \sqrt{\frac{\eta}{4} n(c_* - \frac{\omega^2}{c_*}) \|\beta_{I_1}^*\|_2^2}$$  \hspace{1cm} (34)

with probability $1 - \delta_2$. 

29
Proof 5 Firstly, we show \[31\]:

\[
P(\|X_{12}\beta_{12}^*\|^2(I - H_{\tilde{A}})\varepsilon | > \frac{\eta}{8} n(c_* - \frac{\omega^2}{c_*}) \|\beta_{12}^*\|_2^2)
\]
\[
\leq P(\|X_{12}^T\varepsilon\|_2 \|\beta_{12}^*\|_2 > \frac{\eta}{8} n(c_* - \frac{\omega^2}{c_*}) \|\beta_{12}^*\|_2^2
\]
\[
\leq \sum_{j \in I_1} \sum_{i=1}^{p_j} P(\|X^{(i)}_{G_i}\varepsilon | > \frac{\eta}{8 \sqrt{\# I_1}} n(c_* - \frac{\omega^2}{c_*}) \|\beta_{12}^*\|_2^2
\]
\[
\leq 2J_{\max} \exp\{-nC_1\vartheta / sp_{\max}\}
\]
\[
\leq 6
\]

where the first inequality follows from the idempotency of \(I - H_{\tilde{A}}\) and Cauchy inequality, and the third inequality follows from \(\# I_1 \leq sp_{\max}\) in which \(C_1\) is some positive constant depending on \(T\).

Next we prove \[32\]. With probability \(1 - \delta_1\), we have

\[
\|\beta_{12\cup I_{12}}^*\|_2^2 = \|\beta_{A_{12}\cup I_{12}}^*\|_2^2 + \|\beta_{I_{12}}^*\|_2^2
\]
\[
\leq [(1 + \eta)(1 + t)\omega]^2 \|\beta_{A_{12}}^*\|_2^2 + (1 + \eta)(1 + t)\omega \|\beta_{I_{12}}^*\|_2^2
\]
\[
\leq 2[(1 + \eta)(1 + t)\omega \|\beta_{A_{12}}^*\|_2^2 + \|\beta_{I_{12}}^*\|_2^2
\]

where the first inequality follows from \[14\] and \[13\], and the second inequality uses the fact that \(c_* > c_* - \frac{\omega^2}{c_*} > 0\). Thus we have

\[
P(\|X_{12\cup I_{12}}\beta_{12\cup I_{12}}^*\|^2(I - H_{\tilde{A}})\varepsilon | > \frac{\eta}{8} n(c_* - \frac{\omega^2}{c_*}) \|\beta_{12\cup I_{12}}^*\|_2^2
\]
\[
\leq P(\|X_{A_{12}\cup I_{12}}^T\varepsilon\|_2 \|\beta_{A_{12}\cup I_{12}}^*\|_2 > \frac{\eta}{8} n(c_* - \frac{\omega^2}{c_*}) \|\beta_{A_{12}\cup I_{12}}^*\|_2^2
\]
\[
\leq \sum_{j \in A_{12}\cup I_{12}} \sum_{i=1}^{p_j} P(\|X^{(i)}_{G_i}\varepsilon | > \frac{\eta}{8 \sqrt{T}(1 + \eta)(1 + t)\omega \sqrt{\# I_{12}} \|\beta_{A_{12}\cup I_{12}}^*\|_2^2
\]
\[
\leq 2J_{\max} \exp\{-nC_1\vartheta / sp_{\max}\}
\]
\[
\leq 6
\]

where the second inequality follows from \[35\], and the fourth inequality follows \(\# I_{12} \leq sp_{\max}\).
Next we show (33).

\[ P(\|H_A\|_2 > \sqrt{\frac{n}{4}}(c_* - \frac{\omega^2}{c_*} \|\beta_{\tilde{I}}\|_2^2)) \]

\[ = P(\|X_A(X_A^TX_A)^{-1}X_A^T\|_2 > \sqrt{\frac{n}{4}}(c_* - \frac{\omega^2}{c_*} \|\beta_{\tilde{I}}\|_2^2)) \]

\[ \leq P\left(\frac{\sqrt{nc^*}}{nc^*} \|X_A\|_2 > \sqrt{\frac{n}{4}}(c_* - \frac{\omega^2}{c_*} \|\beta_{\tilde{I}}\|_2^2)\right) \]

\[ \leq \sum_{j \in A} \sum_{i=1}^{p_j} P(\{(X_{G_i}^{(i)})^T \|\varepsilon\| > n \sqrt{\frac{n}{4\#\{A\}}(c_* - \frac{\omega^2}{c_*} \|\beta_{\tilde{I}}\|_2^2)}) \]

\[ \leq 2Jp_{\text{max}} \exp\{-nC_2\theta/Tp_{\text{max}}\} \]

\[ \leq \delta_2, \]

where the first inequality follows from (10) and (11), and the third inequality follows from \#\{A\} \leq Tp_{\text{max}}. Here positive constant \(C_2\) is related to \(T\).

This proves (33), and (34) can be proved as (33).  

\[ \square \]

By Lemma 33, we can detail the proofs of the theoretical properties. We first give the proof of Theorem 1, the support recovery, and then based on it, give the proof of selection consistency of Algorithm 2 under BGIC. Next based on the conclusion of Theorem 1, we prove Theorem 3 and Corollary 4, the convergence step. With the convergence step, we then show the proof of computational complexity. At last, we show the proof of Theorem 5, in which we show the \(\ell_2\) error bounds in each iteration.

### A.6 Proof of Theorem 1

**Proof** Assume \(\tilde{I} \neq \emptyset\) and show that it will lead to a contradiction. Note that \(A^* \cap \tilde{I} = A_{12} \cup \tilde{I}_{12}\). Assume \(\hat{\beta}\) is the least-squares estimator supported on \(\cup_{j \in \tilde{I}} G_j\).

The loss function of \((\hat{A}, \tilde{I})\) is

\[ 2nL(\hat{\beta}) = y^T(I - H_{\hat{A}})y \]

\[ = (X_{A_{12} \cup \tilde{I}_{12}} \beta_{A_{12} \cup \tilde{I}_{12}}^* + \varepsilon)^T(I - H_{\hat{A}})(X_{A_{12} \cup \tilde{I}_{12}} \beta_{A_{12} \cup \tilde{I}_{12}}^* + \varepsilon) \]

\[ = (X_{A_{12} \cup \tilde{I}_{12}} \beta_{A_{12} \cup \tilde{I}_{12}}^*)^T(I - H_{\hat{A}})X_{A_{12} \cup \tilde{I}_{12}} \beta_{A_{12} \cup \tilde{I}_{12}}^* + \varepsilon^T(I - H_{\hat{A}})\varepsilon + \]

\[ 2(X_{A_{12} \cup \tilde{I}_{12}} \beta_{A_{12} \cup \tilde{I}_{12}}^*)^T(I - H_{\hat{A}})\varepsilon \]

\[ \leq nc^* \|\beta_{A_{12} \cup \tilde{I}_{12}}\|_2^2 + \|\beta_{\tilde{I}_{12}}\|_2^2 + \varepsilon^T \varepsilon + |\varepsilon^T H_{\hat{A}}\varepsilon| \]

\[ \leq 2nc^* \frac{(1 + \eta)(1 + t)\omega}{c_* - \frac{\omega^2}{c_*}} \|\beta_{\tilde{I}}\|_2^2 + \varepsilon^T \varepsilon + |\varepsilon^T H_{\hat{A}}\varepsilon| + \]

\[ 2\|X_{A_{12} \cup \tilde{I}_{12}} \beta_{A_{12} \cup \tilde{I}_{12}}^*\|_2^2 + \|\beta_{A_{12} \cup \tilde{I}_{12}}\|_2^2 \]

\[ \leq 2nc^* \frac{(1 + \eta)(1 + t)\omega}{c_* - \frac{\omega^2}{c_*}} \|\beta_{\tilde{I}}\|_2^2 + \varepsilon^T \varepsilon + |\varepsilon^T H_{\hat{A}}\varepsilon| + \]

\[ 2\|X_{A_{12} \cup \tilde{I}_{12}} \beta_{A_{12} \cup \tilde{I}_{12}}^*\|_2^2 + \|\beta_{A_{12} \cup \tilde{I}_{12}}\|_2^2, \]

31
Thus Algorithm 1 continues iterations and does not output a solution in current iteration, which attains minimum when model size and the fourth inequality follows from Conditions (C4) and (C5).

Similarly, the following equality holds for the loss function of $(\hat{A}, \hat{I})$ is

$$
2nL(\hat{\beta}) = y^\top (I - H_{\hat{A}})y = (X_{I_1} \beta_{I_1} + \epsilon)^\top (I - H_{\hat{A}})(X_{I_1} \beta_{I_1} + \epsilon) \geq n(c_* - \frac{\omega^2}{c_*}) \|\beta_{I_1}^*\|_2^2 + \epsilon^\top \epsilon - 2\|X_{I_1} \beta_{I_1}^*\|^2 (I - H_{\hat{A}})\epsilon| - |\epsilon^\top H_{\hat{A}}\epsilon|.
$$

Therefore, with probability $1 - \delta_1 - \delta_2$, we have

$$
L(\hat{\beta}) - L(\hat{\beta}) \geq \frac{1}{2n} [n(c_* - \frac{\omega^2}{c_*}) \|\beta_{I_1}^*\|_2^2 - 2nc^* \left( \frac{(1 + \eta)(1 + t) \omega^2}{c_* - \frac{\omega^2}{c_*}} \right) \|\beta_{I_1}^*\|_2^2 - 2\|X_{I_1} \beta_{I_1}^*\|^2 (I - H_{\hat{A}})\epsilon| - |\epsilon^\top H_{\hat{A}}\epsilon|]
\geq \frac{1}{2n} [(1 - \eta)n(c_* - \frac{\omega^2}{c_*}) \|\beta_{I_1}^*\|_2^2 - 2nc^* \left( \frac{(1 + \eta)(1 + t) \omega^2}{c_* - \frac{\omega^2}{c_*}} \right) \|\beta_{I_1}^*\|_2^2]
\geq \frac{1}{2n} \cdot \frac{2}{2}
\geq \frac{1}{2n} \cdot \frac{2}{2}
\geq \pi_T,
$$

where the second inequality follows from (31)-(34), the third inequality follows from Condition (C3) and the fourth inequality follows from Conditions (C4) and (C5).

Consequently,

$$
P(L(\hat{\beta}) - L(\hat{\beta}) > \pi_T) \geq 1 - \delta_1 - \delta_2.
$$

Thus Algorithm 1 continues iterations and does not output a solution in current iteration, which leads to a contradiction with $I_1 \neq \emptyset$. Furthermore, we can lead to a similar contradiction with other assumptions in terms of $A_{12}$ and $I_{12}$.

\[ \square \]

### A.7 Proof of Theorem 2

**Proof** To prove Theorem 2, we need to analyze the gap between $\log L(\beta_1)$ and $\log L(\beta_2)$. By using the inequality that $1 - \frac{1}{x} \leq \log x \leq x - 1$, for any $x > 0$, we have

$$
\frac{L(\beta_1) - L(\beta_2)}{L(\beta_1)} \leq \log \frac{L(\beta_1)}{L(\beta_2)} \leq \frac{L(\beta_1) - L(\beta_2)}{L(\beta_2)}.
$$

Let $\hat{\beta}^* = \arg \min_{\beta_{I^*} \neq 0} L(\beta)$ be the least-squares estimator supported on $\cup j \in A^* G_j$. Next we show BGIC attains minimum when model size $T = s$. Note that from the conclusion of Theorem 1, when $T \geq s$,
we have
\[ P(\hat{A} \supseteq A^*) \geq 1 - \delta_1 - \delta_2. \]

Therefore, from the projection property of \( I - H_{\hat{A}} \), when \( T \geq s \), we have
\[ L(\hat{\beta}) = \frac{1}{2n} y^\top (I - H_{\hat{A}}) y = \frac{1}{2n} \varepsilon^\top (I - H_{\hat{A}}) \varepsilon. \]

Especially, when \( T = s \), we have
\[ L(\hat{\beta}) = L(\hat{\beta}^*) = \frac{1}{2n} \varepsilon^\top (I - H_{A^*}) \varepsilon. \]

Firstly, we consider the case when given model size \( T < s \). With probability \( 1 - \delta_1 - \delta_2 \), we have
\[ 2n L(\hat{\beta}) - 2n L(\hat{\beta}^*) = y^\top (I - H_{\hat{A}}) y - \varepsilon^\top (I - H_{A^*}) \varepsilon \]
\[ = (X_{I_1} \beta_{I_1}^* + \varepsilon)^\top (I - H_{\hat{A}}) (X_{I_1} \beta_{I_1}^* + \varepsilon) - \varepsilon^\top (I - H_{A^*}) \varepsilon \]
\[ \geq n \left( c_* - \frac{\omega^2}{c_*} \right) \| \beta_{I_1}^* \|_2^2 - 2 \| (X_{I_1} \beta_{I_1}^*)^\top (I - H_{\hat{A}}) \varepsilon \| - \| \varepsilon^\top H_{\hat{A}} \varepsilon \| - \| \varepsilon^\top H_{A^*} \varepsilon \|, \] (39)

where the last inequality follows from (31), (33) and (37). Similar to the proof of (33) in Lemma 5, we have
\[ P(\| \varepsilon^\top H_{A^*} \varepsilon \| > \frac{m \eta}{2} (c_* - \frac{\omega^2}{c_*}) \| \beta_{I_1}^* \|_2^2) \leq \delta_2. \] (40)

Combine (39) and (40), with probability \( 1 - \delta_1 - \delta_2 \),
\[ 2n L(\hat{\beta}) - 2n L(\hat{\beta}^*) \geq (1 - \eta) n \left( c_* - \frac{\omega^2}{c_*} \right) \| \beta_{I_1}^* \|_2^2. \] (41)

For \( 2n L(\hat{\beta}^*) \), we have
\[ 2n L(\hat{\beta}^*) = \varepsilon^\top (I - H_{A^*}) \varepsilon \]
\[ \leq \| \varepsilon \|_2^2 + \| \varepsilon^\top H_{A^*} \varepsilon \| \]
\[ \leq 2n L(\beta^*) + \frac{m \eta}{2} (c_* - \frac{\omega^2}{c_*}) \| \beta_{I_1}^* \|_2^2, \] (42)

with probability \( 1 - \delta_2 \). Combine (38), (41) and (42), with probability \( 1 - \delta_1 - \delta_2 \),
\[ \log \frac{L(\hat{\beta})}{L(\hat{\beta}^*)} \geq \frac{(1 - \eta)(c_* - \frac{\omega^2}{c_*}) \| \beta_{I_1}^* \|_2^2}{2 L(\beta^*) + \frac{m \eta}{2} (c_* - \frac{\omega^2}{c_*}) \| \beta_{I_1}^* \|_2^2} = O(1). \]

Consequently,
\[
\text{BGIC}_\gamma(\hat{A}) - \text{BGIC}_\gamma(A^*) = n \log \frac{L(\hat{\beta})}{L(\beta^*)} - (\#\{A^*\} - \#\{\hat{A}\})(\gamma \log J + \log n) \\
\geqslant O(n) - \#\{A^*\}(\gamma \log J + \log n) \\
\geqslant O(n) - |A^*|p_{\max}(\gamma \log J + \log n) \\
\geqslant O(n) - o(n) \\
> 0
\]

for sufficiently large \(n\) where the third inequality follows from Condition (C6).

On the other hand, consider the case when \(T > s\). Let \(\hat{A} = A^* \cup B\),

\[
L(\hat{\beta}^*) - L(\hat{\beta}) = \frac{1}{2n} \varepsilon^T (H_{\hat{A}} - H_{A^*}) \varepsilon \\
= \frac{1}{2n} \varepsilon^T (I - H_{A^*}) H_{\hat{A}} (I - H_{A^*}) \varepsilon \\
= \frac{1}{2n} \varepsilon^T (I - H_{A^*}) X_B (X_B^T (I - H_{A^*}) X_B)^{-1} X_B^T (I - H_{A^*}) \varepsilon \\
= \frac{1}{2n} \| (X_B^T (I - H_{A^*}) X_B)^{-\frac{1}{2}} \|_2^2.
\]

Note that

\[
P(\frac{1}{\sqrt{2n}} \| (X_B^T (I - H_{A^*}) X_B)^{-\frac{1}{2}} X_B^T (I - H_{A^*}) \varepsilon \|_2 \geqslant t) \\
\leqslant P(\| X_B^T (I - H_{A^*}) \varepsilon \|_2 \geqslant \sqrt{\frac{2}{\#\{B\}} (c_* - \frac{\omega^2}{c_*}) nt}) \\
\leqslant \sum_{j \in B} \sum_{i=1}^{p_i} P(\| (X_B^{(i)})^T \varepsilon \| > \sqrt{\frac{2}{\#\{B\}} (c_* - \frac{\omega^2}{c_*}) nt}) \\
\leqslant 2Jp_{\max} \exp\{-\frac{C_1 nt^2}{\#\{B\}}\} = \delta
\]

for some positive constant \(C_1\) depending on \(T\). Given any probability \(\delta\), with probability \(1 - \delta\), we can calculate the corresponding \(t\) as follows

\[
\frac{1}{2n} \| (X_B^T (I - H_{A^*}) X_B)^{-\frac{1}{2}} X_B^T (I - H_{A^*}) \varepsilon \|_2^2 \leqslant \#\{B\} \log \frac{2Jp_{\max}}{\delta}.
\]

Let \(\delta = O(J^{-\alpha} p_{\max})\) for some constant \(0 < \alpha < 1\),

\[
L(\hat{\beta}^*) - L(\hat{\beta}) = \frac{1}{2n} \| (X_B^T (I - H_{A^*}) X_B)^{-\frac{1}{2}} X_B^T (I - H_{A^*}) \varepsilon \|_2^2 \\
\leqslant \#\{B\} \log(J^{1+\alpha})
\]

(44)

Define \(\hat{\sigma}\) as the standard deviation of random variable \(\varepsilon_i\). Similar to (44), with probability
1 − O(J−αp_{max}), we have

\[
L(\hat{\beta}) = \frac{1}{2n}\epsilon^\top(I - H_{\hat{A}})\epsilon \\
\geq \frac{1}{2n}\norm{\epsilon}_2^2 - \frac{1}{2n}\epsilon^\top H_{\hat{A}}\epsilon \\
\geq \frac{\hat{\sigma}^2}{2} - \frac{\#\{\hat{A}\} \log(J^{1+\alpha})}{nC_2},
\]

for some positive constant C_2 depending on T, where the second inequality follows from the law of large number. From the assumption T_{max} = O(n_{p_{max}} \log J), we have

\[
\frac{\#\{\hat{A}\} \log(J^{1+\alpha})}{nC_2} \leq \frac{T_{max}p_{max} \log(J^{1+\alpha})}{nC_2} \rightarrow 0
\]

for sufficiently large n and L(\hat{\beta}) > 0. Combine (38), (44), (45) and (46),

\[
n \log \frac{L(\hat{\beta}^*)}{L(\hat{\beta})} \leq \frac{\#\{B\} \log(J^{1+\alpha})}{\hat{\sigma}^2} - \frac{\#\{A\} \log(J^{1+\alpha})}{nC_2} \rightarrow \frac{2\#\{B\} \log(J^{1+\alpha})}{\hat{\sigma}^2C_1}
\]

for sufficiently large n. For simplicity, let C = \frac{2}{\hat{\sigma}^2C_1}. Consequently,

\[
BGIC_\gamma(A^*) - BGIC_\gamma(\hat{A}) = n \log \frac{L(\hat{\beta}^*)}{L(\hat{\beta})} - \#\{B\} (\gamma \log J + \log n) \\
\leq C \#\{B\} \log(J^{1+\alpha}) - \#\{B\} \log(n^{\gamma k + 1}) \\
\leq \#\{B\} \log(n^{kC(1+\alpha)}) - \#\{B\} \log(n^{\gamma k + 1}) \\
< 0
\]

for sufficient large n when kC(1 + \alpha) < \gamma k + 1.

Therefore, when \gamma > C(1 + \alpha) - \frac{1}{k}, Algorithm 2 under BGIC identifies the true group subset A^* with probability 1 − O(J−αp_{max}).

\[\square\]

A.8 Proof of Theorem 3

We will give the proof of convergence step and ℓ_2 error bound in each iteration of Algorithm 1. Denote S_1^k and S_2^k as the exchange subsets in the kth iteration and

\[
A_1^k = A^k \cap A^*, \quad I_1^k = I^k \cap A^*,
\]

and

\[
A_{12}^k = A_1^k \cap S_1^k, \quad I_{12}^k = I_1^k \cap (S_2^k)^c.
\]

Proof 8 Note that I_{12}^{k+1} = A_{12}^k \cup I_{12}^k. Similar to (36), the error of loss function in the (k + 1)th
iteration is
\[
|2nL(\beta^{k+1}) - 2nL(\beta^*)| \\
\leq |(X_{T_1}^{k+1}\beta_{T_1}^{k+1} + \epsilon)^\top (I - H_{A_{k+1}})(X_{T_1}^{k+1}\beta_{T_1}^{k+1} + \epsilon) - \epsilon^\top \epsilon| \\
\leq |(X_{T_1}^{k+1}\beta_{T_1}^{k+1})^\top (I - H_{A_{k+1}})(X_{T_1}^{k+1}\beta_{T_1}^{k+1})| + \\
2|(X_{T_1}^{k+1}\beta_{T_1}^{k+1})^\top (I - H_{A_{k+1}})\epsilon| + |\epsilon^\top H_{A_{k+1}}\epsilon| \tag{47}
\]

where \( h_1(\epsilon) = 2|(X_{A_{k+1}'} \beta_{A_{k+1}'} + \epsilon)(I - H_{A_{k+1}'} + \epsilon)^\top | + |\epsilon^\top H_{A_{k+1}'}\epsilon| \).

Similar to (37), the error of loss function in the \( k \)th iteration is
\[
|2nL(\beta^k) - 2nL(\beta^*)| = |(X_{T_1}^k\beta_{T_1}^k + \epsilon)^\top (I - H_{A_k})(X_{T_1}^k\beta_{T_1}^k + \epsilon) - \epsilon^\top \epsilon| \\
\geq |(X_{T_1}^k\beta_{T_1}^k)^\top (I - H_{A_k})(X_{T_1}^k\beta_{T_1}^k) - \\
2|(X_{T_1}^k\beta_{T_1}^k)^\top (I - H_{A_k})\epsilon| - |\epsilon^\top H_{A_k}\epsilon| \tag{48}
\]

where \( h_2(\epsilon) = 2|(X_{T_1}^k\beta_{T_1}^k)^\top (I - H_{A_k})\epsilon| + |\epsilon^\top H_{A_k}\epsilon| \). Use the technique in the proof of Lemma 5.

\[
h_1(\epsilon) \leq \frac{\eta}{2} \mu_T n(c^* - \frac{\omega^2}{c^*})^2 \beta_{T_1}^k \| \beta_{T_1}^k \|_2^2 \tag{49}
\]

and
\[
h_2(\epsilon) \leq \frac{\eta}{2} n(c^* - \frac{\omega^2}{c^*})^2 \beta_{T_1}^k \| \beta_{T_1}^k \|_2^2 \tag{50}
\]

with probability \( 1 - \delta_1 - \delta_2 \). Combine (47)-(50),
\[
|2nL(\beta^{k+1}) - 2nL(\beta^*)| \leq 2nc^* \frac{(1 + \eta)(1 + t)\omega}{c_s - \frac{\omega^2}{c_s}}^2 \beta_{T_1}^k \| \beta_{T_1}^k \|_2^2 + h_1(\epsilon) \\
\leq \mu_T (1 - \eta)n(c^* - \frac{\omega^2}{c^*})^2 \beta_{T_1}^k \| \beta_{T_1}^k \|_2^2 + h_1(\epsilon) \tag{51}
\]

Let \( A^0 = \emptyset \) and use (51) repeatedly,
\[
|2nL(\beta^{k+1}) - 2nL(\beta^*)| \leq \mu_T |2nL(\beta^{k+1}) - 2nL(\beta^*)| \\
\leq \mu_T |2nL(\beta^0) - 2nL(\beta^*)| \\
\leq \mu_T |\beta_T^{k+1} - \beta_T^0|_2 \tag{52}
\]
This completes the proof of part (i) in Theorem 3, which shows the error bounds of loss function decay geometrically.

Next we provide the lower bound of the error of loss function when \( I_1 \neq \emptyset \). From (48) and (50), with probability \( 1 - \delta_1 - \delta_2 \), we have

\[
|2nL(\beta^k) - 2nL(\beta^*)| \geq n(c_* - \frac{\omega^2}{c_*})\|\beta^*_I\|_2^2 - h_2(\varepsilon) \\
\geq (1 - \frac{\eta}{2})n(c_* - \frac{\omega^2}{c_*})\theta \\
> 0.
\]  

(53)

When the upper bounds in (52) is smaller than the lower bounds in (53), we conclude that \( A^k \supseteq A^* \).

Therefore, we have

\[
|2nL(\beta^k) - 2nL(\beta^*)| \leq \mu_T^k \|y\|_2^2 \leq (1 - \frac{\eta}{2})n(c_* - \frac{\omega^2}{c_*})\theta
\]

holds, equivalently,

\[
A^k \supseteq A^*, \quad k > \log \frac{1}{\nu_T} \frac{\|y\|_2^2}{(1 - \frac{\eta}{2})n(c_* - \frac{\omega^2}{c_*})\theta},
\]

which completes the proof of Theorem 3.

Especially, when \( T = s \), we have

\[
A^k = A^*, \quad k > \log \frac{1}{\nu_T} \frac{\|y\|_2^2}{(1 - \frac{\eta}{2})n(c_* - \frac{\omega^2}{c_*})\theta}.
\]

\[\square\]

A.9 Proof of Corollary 4

Proof 9 Assume \( A^k \supseteq A^* \), then we have \( A^{k+1} \supseteq A^* \) with probability \( 1 - \delta_1 - \delta_2 \) from part (ii) of Theorem 3. Therefore, we have \( L(\beta^{k+1}) = \frac{1}{2n}\varepsilon^T(1 - H_{A^{k+1}})\varepsilon \) and \( L(\beta^k) = \frac{1}{2n}\varepsilon^T(1 - H_{A^k})\varepsilon \).

Following similar derivation in (45), with probability \( 1 - O(J^{-c}p_{\text{max}}) \), we have

\[
L(\beta^k) \leq \frac{\sigma^2}{2} + \frac{\# \{A^k\} \log(J^{1+\alpha})}{nC}
\]

and

\[
L(\beta^{k+1}) \geq \frac{\sigma^2}{2} - \frac{\# \{A^{k+1}\} \log(J^{1+\alpha})}{nC},
\]

where \( C \) is some positive constant depending on \( T \). Therefore,

\[
L(\beta^k) - L(\beta^{k+1}) \leq \frac{2\# \{A^{k+1}\} \log(J^{1+\alpha})}{nC} \leq \frac{2T p_{\text{max}} \log(J^{1+\alpha})}{nC} \leq \pi_T.
\]

The gap of loss function is smaller than the threshold \( \pi_T \) in Algorithm 1 after \( k \)th iteration. Combining part (ii) of Theorem 3, we conclude that Algorithm 1 stops after \( O(\log \frac{1}{\nu_T} \frac{\|y\|_2^2}{(1 - \frac{\eta}{2})n(c_* - \frac{\omega^2}{c_*})\theta}) \)
iterations with probability $1 - O(J^{-\alpha}p_{max})$ when $T > s$.

□

A.10 Proof of Theorem 4

Proof 10 First of all, consider $0 < T < s$. Since the loss function decreases at least $\pi_T$ in each iteration, thus Algorithm 1 stops after $O(\|y\|_2^2/\pi_T)$ iterations for each fixed $T$.

Next consider $s \leq T \leq T_{max}$. For a fixed $T$, by Corollary 4 and Condition (C5), Algorithm 1 stops after $O(\log T_{max}/\log(nJ))$ iterations.

Now we analyze the computational complexity of Algorithm 1 for a given model size $T$. Firstly, computing the primal variable and dual variable takes a total $O(T_{max}p + np)$ steps and then computing the $\ell_2$ norm of each group takes $O(p)$ step. Next, finding the smallest or largest $T$ contributions takes $O(J)$ steps via Hoare’s selection algorithm [Hoare, 1961]. Because the procedure repeats at most $C_{max}$ times, $O(JC_{max})$ steps at most are demanded. Therefore, the total computational complexity of Algorithm 1 is

$$O((\log \frac{\|y\|_2^2}{T_{max}})I(s \leq T) + \frac{\|y\|_2^2}{\pi_T}I(s > T)) \times O(T(np + T_{max}p + JC_{max})).$$

Since $T$ varies from 1 to $T_{max}$ in Algorithm 2, thus the total computational complexity is

$$O(\log \frac{\|y\|_2^2}{T_{max}}p_{max}I + \frac{\|y\|_2^2}{T_{max}}I(nT_{max}p_{max} + J2_{max}C_{max})) +$$

$$\frac{n\|y\|_2^2}{T_{max}}(np_{max} + np_{max} + JsC_{max}))$$

$$\leq O(T_{max}(T_{max} \log \frac{\|y\|_2^2}{p_{max}log(nJ)}) + \frac{n\|y\|_2^2}{p_{max}}(np_{max} + T_{max}p_{max} + JC_{max})).$$

□

A.11 Proof of Theorem 5

Proof 11 Note that

$$P(\|\hat{X}_{I_{Ak}}^TX_{A_k}^{-1}X_{A_k}^T\hat{\epsilon}\|_2 \leq \sqrt{\eta}\|\beta^*_I\|_2) \geq 1 - \delta_2.$$

we have

$$\|\beta^k - \beta^*\|_2^2 = \|\beta_{A_k}^k - \beta_{A_k}^*\|_2^2 + \|\beta_{\hat{I}_k}^*\|_2^2$$

$$= \|\hat{X}_{A_k}^T\hat{X}_{A_k}^{-1}X_{A_k}^Ty - \beta_{A_k}^*\|_2^2 + \|\beta_{\hat{I}_k}^*\|_2^2$$

$$= \|\hat{X}_{A_k}^T(X_{A_k}^T\hat{X}_{A_k}^{-1}X_{A_k}^T\hat{\epsilon} + \hat{\beta}_{A_k}^* + \hat{\beta}_{\hat{I}_k}^*) - \beta_{A_k}^*\|_2^2 + \|\beta_{\hat{I}_k}^*\|_2^2$$

$$= \|\hat{X}_{A_k}^T(X_{A_k}^T\hat{X}_{A_k}^{-1}X_{A_k}^T\beta_{\hat{I}_k}^* + \epsilon)\|_2^2 - \|\beta_{A_k}^*\|_2^2 + \|\beta_{\hat{I}_k}^*\|_2^2$$

$$\leq (1 + \frac{\omega}{\epsilon_\hat{\beta}})\|\beta_{\hat{I}_k}^*\|_2^2 - \|\beta_{A_k}^*\|_2^2$$ (54)
Combine (52), (53) and (54),
\[
\|\beta^k - \beta^*\|_2^2 \leq \frac{1 + \frac{\omega}{c_\ast}}{(1 - \frac{n}{2})n(c_\ast - \frac{\omega^2}{c_\ast})} [2nL(\beta^k) - 2nL(\beta^*)] \\
\leq \frac{1 + \frac{\omega}{c_\ast}}{(1 - \frac{n}{2})n(c_\ast - \frac{\omega^2}{c_\ast})} \mu_k \|g\|^2_2.
\]

This completes the proof of Theorem 5. □

B Generalization of GSplicing algorithm

In this section, we present the generalized form of GSplicing algorithm, in which we extend GSplicing algorithm to generalized linear model and Cox proportional hazard model. Define \( L(\beta) \) as the negative log-likelihood function. Denote \( g_G^j = [\nabla L(\beta)]_{Gj} \) as the \( j \)th group gradient of \( l(\beta) \), and \( h_G^j = [\nabla^2 L(\beta)]_{Gj} \) as the \( j \)th group diagonal sub-matrix of hessian matrix of \( L(\beta) \). Let dual variable \( d_G^j = -g_G^j \) and \( \Psi_G^j = (h_G^j)^{\frac{1}{2}} \).

We first define the contributions to the decreases of \( L(\beta^k) \) in the backward and forward direction in the \( k \)th iteration.

**Lemma 6**  
(i) For any \( j \in A^k \), the contributions to the decreases of \( L(\beta^k) \) by discarding \( j \)th group is,
\[
L(\beta^{A^k \setminus j}) - L(\beta^k) = \frac{1}{2} (\beta_{G^j}^k)^T h_G^j \beta_{G^j}^k = \frac{1}{2} \|\bar{\beta}_{G^j}^k\|^2_2,
\]
where \( \beta^{A^k \setminus j} \) is the estimator assigning the \( j \)th group of \( \beta^k \) to be zero and \( \bar{\beta}_{G^j}^k = \Psi_G^j \beta_{G^j}^k \).

(ii) For any \( j \in I^k \), the contributions to the decreases of \( L(\beta^k) \) by adding \( j \)th group is,
\[
L(\beta^k) - L(\beta^k + t^k_j) = \frac{1}{2} (d_{G^j}^k)^T (h_{G^j}^k)^{-1} d_{G^j}^k = \frac{1}{2} \|\bar{d}_{G^j}^k\|^2_2,
\]
where \( \bar{d}_{G^j}^k = (\Psi_{G^j}^k)^{-1} d_{G^j}^k \), and \( t^k_j = (\Psi_{G^j}^k)^{-1} d_{G^j}^k \).
Proof 12 We first prove part (i). The local quadratic approximation of $L(\beta^k)$ is

$$L(\beta^{A_k\setminus j} = L(\beta^k) + \frac{\partial L(\beta)}{\partial \beta_{A_k\setminus j}} |_{\beta^{A_k\setminus j} = \beta^{A_k\setminus j}}, \frac{\partial L(\beta)}{\partial \beta_{j}} |_{\beta_{j}^k} + \frac{1}{2} (\beta^{A_k\setminus j} - \beta^{A_k\setminus j})^T \frac{\partial^2 L(\beta)}{\partial A_k A_k} |_{\beta^{A_k\setminus j} = \beta^{A_k\setminus j}}, \frac{\partial^2 L(\beta)}{\partial \beta_{j}^2} |_{\beta_{j}^k}$$

Next we prove (ii). The local quadratic approximation of $L(\beta^k)$ is

$$L(\beta^{k} + t_j^k = L(\beta^k) + \frac{\partial L(\beta)}{\partial \beta_{A_k}} |_{\beta^{k} = \beta^k}, \frac{\partial L(\beta)}{\partial \beta_{j}} |_{\beta_{j}^k} + \frac{1}{2} (t_j^k)^T \frac{\partial^2 L(\beta)}{\partial \beta_{j}^2} |_{\beta_{j}^k} + \frac{1}{2} (\beta^k - \beta^k)^T \frac{\partial^2 L(\beta)}{\partial A_k \beta_{j}} |_{\beta^k} t_j^k$$

Denote $\tilde{t}_j^k = \Psi_{G_j} t_j^k$ and $\tilde{d}_{G_j} = (\Psi_{G_j})^{-1} d_{G_j}$. Consequently, we have

$$L(\beta^k) - L(\beta^k + t_j^k) = -\frac{1}{2} (t_j^k)^T h_{G_j} t_j^k + (\gamma_{G_j})^T t_j^k$$

$$= -\frac{1}{2} ||\tilde{t}_j^k - \tilde{d}_{G_j}||_2^2 + \frac{1}{2} ||\tilde{d}_{G_j}||_2^2$$

When $\tilde{t}_j^k = \tilde{d}_{G_j}$, $L(\beta^k) - L(\beta^k + t_j^k)$ attains the maximum.}

Next we provide the expressions for four important statistical models.

Case 1: Group linear model.

In group linear model, the loss function is

$$L(\beta) = \frac{1}{2} ||y - X\beta||_2^2.$$

We have

$$d_{G_j} = X_{G_j}^T (y - X\beta)/n, \; \Psi_{G_j} = (X_{G_j} X_{G_j}/n)^{\frac{1}{2}}, \; j = 1, \ldots, J.$$
Under the assumption of orthonormalization, we have $\Psi_{G_j} = I_{p_j}$. Thus for linear regression model, we do not need to update $\Psi$ during iteration procedures.

**Case 2 : Group logistic model.**
Given the data $\{(X_i, y_i)\}_{i=1}^n$ with $y_i \in \{0, 1\}, X_i \in \mathbb{R}^p$, and denote $X_i = (X_{i,G_1}, \ldots, X_{i,G_J})^\top$. Consider the logistic model $\log\{\pi/(1-\pi)\} = \beta_0 + x^\top \beta$ with $x \in \mathbb{R}^p$ and $\pi = P(y = 1|x)$.
Thus the negative log-likelihood function is

$$L(\beta_0, \beta) = \sum_{i=1}^n \{\log(1 + \exp(\beta_0 + X_i^\top \beta)) - y_i(\beta_0 + X_i^\top \beta)\}.$$ 

We have

$$d_{G_j} = X_{G_j}^\top (y - \pi), \quad \Psi_{G_j} = (X_{G_j}^\top WX_{G_j})^{\frac{1}{2}}, \quad j = 1, \ldots, J,$$
where $\pi = (\pi_1, \ldots, \pi_n)$ with $\pi_i = \exp(X_i^\top \beta)/(1 + \exp(X_i^\top \beta))$, and $W$ is a diagonal matrix with $i$th diagonal entry equal to $\pi_i(1 - \pi_i)$.

**Case 3 : Group poisson model.**
Given the data $\{(X_i, y_i)\}_{i=1}^n$ with $y_i \in \mathbb{N}, X_i \in \mathbb{R}^p$, and denote $X_i = (X_{i,G_1}, \ldots, X_{i,G_J})^\top$. Consider the poisson model $\log(\mathbb{E}(y|x)) = \beta_0 + x^\top \beta$ with $x \in \mathbb{R}^p$.
Thus the negative log-likelihood function is

$$L(\beta_0, \beta) = \sum_{i=1}^n \{\exp(\beta_0 + X_i^\top \beta) + \log(y_i!) - y_i(\beta_0 + X_i^\top \beta)\}.$$ 

We have

$$d_{G_j} = X_{G_j}^\top (y - \eta), \quad \Psi_{G_j} = (X_{G_j}^\top WX_{G_j})^{\frac{1}{2}}, \quad j = 1, \ldots, J,$$
where $\eta = (\eta_1, \ldots, \eta_n)$ with $\eta_i = \exp(\beta_0 + X_i^\top \beta)$, and $W$ is a diagonal matrix with $i$th diagonal entry equal to $\eta_i$.

**Case 4 : Group Cox proportional hazard model.**
Given the survival data $\{(T_i, \delta_i, x_i)\}_{i=1}^n$ with observation of survival time $T_i$ an censoring indicator $\delta_i$. Consider the Cox proportional hazard model $\lambda(x|t) = \lambda_0(t) \exp(x^\top \beta)$ with a baseline hazard $\lambda_0(t)$ and $x \in \mathbb{R}^p$. By the method of partial likelihood, we can write the negative log-likelihood function as

$$L(\beta) = \log \{ \sum_{i:T_i > t_i} \exp(X_{i,G}^\top \beta) \} - \sum_{i: \delta_i = 1} X_{i,G}^\top \beta.$$ 

We have
\[ d_{G_j} = \sum_{i : \delta_i = 1} (X_{i,G_j} - \sum_{i' : T_{i'} > T_i} X_{i',G_j} \omega_{i,i'}), \]

\[ \Psi_{G_j} = \left\{ \sum_{i : \delta_i = 1} \left( \sum_{i' : T_{i'} > T_i} \omega_{i,i'} X_{i',G_j} \right) \right\}^\top - \sum_{i' : T_{i'} > T_i} \omega_{i,i'} X_{i',G_j} X_{i',G_j}^\top \right\}^{\frac{1}{2}}, \]

where \( \omega_{i,i'} = \exp(X_{i'}^\top \beta) / \sum_{i' : T_{i'} > T_i} \exp(X_{i'}^\top \beta). \)

For generalized linear model and Cox proportional hazard model, GSplicing algorithm is slightly different from linear model. Firstly, we need to update \( \Psi_{G_j} \) in each iteration, which is always an identity matrix in linear model. Secondly, due to the lack of explicit solutions for generalized linear model and Cox proportional hazard model, we use Newton-Raphson method to obtain an iterative solution. Except for these two extra operations, GSplicing algorithm for generalized linear model and Cox proportional hazard model is the same as for linear model.

References

Francis R. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9(40):1179–1225, 2008. URL [http://jmlr.org/papers/v9/bach08b.html](http://jmlr.org/papers/v9/bach08b.html).

Peng Zhao, Guilherme Rocha, and Bin Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A):3468–3497, 12 2009. doi:10.1214/07-AOS584 URL [https://doi.org/10.1214/07-AOS584](https://doi.org/10.1214/07-AOS584).

Guillaume Obozinski, Martin J. Wainwright, and Michael I. Jordan. Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics*, 39(1):1–47, 02 2011. doi:10.1214/09-AOS776 URL [https://doi.org/10.1214/09-AOS776](https://doi.org/10.1214/09-AOS776).

Jian Huang, Patrick Breheny, and Shuangge Ma. A selective review of group selection in high-dimensional models. *Statistical Science*, 27(4):481–499, 2012.

Daehan Won, Hasan Manzour, and Wanpracha Chaowalitwongse. Convex optimization for group feature selection in networked data. *INFORMS Journal on Computing*, 32(1):182–198, 2020. doi:10.1287/ijoc.2018.0868.

Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.

Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 03 1978. doi:10.1214/aos/1176344136 URL [https://doi.org/10.1214/aos/1176344136](https://doi.org/10.1214/aos/1176344136).

Jiahua Chen and Zehua Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 09 2008. ISSN 0006-3444. doi:10.1093/biomet/asn034.
Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44(2):813–852, 04 2016. doi:10.1214/15-AOS1388. URL https://doi.org/10.1214/15-AOS1388.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. doi:https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Junzhou Huang and Tong Zhang. The benefit of group sparsity. *The Annals of Statistics*, 38(4):1978–2004, 08 2010. doi:10.1214/09-AOS778 URL https://doi.org/10.1214/09-AOS778.

Karim Lounici, Massimiliano Pontil, Sara van de Geer, and Alexandre B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39(4):2164–2204, 08 2011. doi:10.1214/11-AOS896.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–924, 04 2010. doi:10.1214/09-AOS729 URL https://doi.org/10.1214/09-AOS729.

Hansheng Wang and Chenlei Leng. A note on adaptive group lasso. *Computational statistics & data analysis*, 52(12):5277–5286, 2008.

Fengrong Wei and Jian Huang. Consistent group selection in high-dimensional linear regression. *Bernoulli*, 16(4):1369–1384, 11 2010. doi:10.3150/10-BEJ252 URL https://doi.org/10.3150/10-BEJ252.

Cun-Hui Zhang and Jian Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567–1594, 08 2008. doi:10.1214/07-AOS520 URL https://doi.org/10.1214/07-AOS520.

Lifeng Wang, Guang Chen, and Hongzhe Li. Group scad regression analysis for microarray time course gene expression data. *Bioinformatics*, 23(12):1486–1494, 2007.

Yaohua Hu, Chong Li, Kaiken Meng, Jing Qin, and Xiaoqi Yang. Group sparse optimization via $\ell_{p,q}$ regularization. *Journal of Machine Learning Research*, 18(30):1–52, 2017. URL http://jmlr.org/papers/v18/15-651.html.

Duy Nhat Phan and Hoai An Le Thi. Group variable selection via $\ell_{p,0}$ regularization and application to optimal scoring. *Neural Networks*, 118:220–234, 2019.

Yuling Jiao, Bangti Jin, and Xiliang Lu. Group sparse recovery via the $\ell_0(\ell_2)$ penalty: Theory and algorithm. *IEEE Transactions on Signal Processing*, 65(4):998–1012, 2016.
Jian Huang, Yuling Jiao, Yanyan Liu, and Xiliang Lu. A constructive approach to \( l_0 \) penalized regression. *The Journal of Machine Learning Research*, 19(1):403–439, 2018.

Xiaotong Shen, Wei Pan, Yunzhang Zhu, and Hui Zhou. On constrained and regularized high-dimensional regression. *Annals of the Institute of Statistical Mathematics*, 65(5):807–832, 2013.

J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972. doi:https://doi.org/10.2307/2344614.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972. doi:https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

Junxian Zhu, Canhong Wen, Jin Zhu, Heping Zhang, and Xueqin Wang. A polynomial algorithm for best-subset selection problem. *Proceedings of the National Academy of Sciences*, 2020. ISSN 0027-8424. doi:10.1073/pnas.2014241117

Jack Kiefer. Sequential minimax search for a maximum. *Proceedings of the American mathematical society*, 4(3):502–506, 1953.

Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.

Tony Cai, Weidong Liu, and Xi Luo. A constrained \( \ell_1 \) minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Patrick Breheny and Jian Huang. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*, 25(2):173–187, 2015.

Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

Todd E Scheetz, Kwangyoun Kim, Ruth E Swiderski, A R Philp, Terry A Braun, Kevin L Knudtson, Anne M Dorrance, Gerald F Dibona, Jian Huang, Thomas L Casavant, et al. Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences of the United States of America*, 103(39):14429–14434, 2006.

Jian Huang, Joel L. Horowitz, and Fengrong Wei. Variable selection in nonparametric additive models. *The Annals of Statistics*, 38(4):2282–2313, 08 2010. doi:10.1214/09-AOS781 URL https://doi.org/10.1214/09-AOS781

Wenliang Pan, Xueqin Wang, Weinan Xiao, and Hongtu Zhu. A generic sure independence screening procedure. *Journal of the American Statistical Association*, 114(526):928–937, 2019.

Jin Zhu, Wenliang Pan, Wei Zheng, and Xueqin Wang. Ball: An r package for detecting distribution difference and association in metric spaces. *Journal of Statistical Software, Articles*, 97(6):1–31, 2021. ISSN 1548-7660. doi:10.18637/jss.v097.i06 URL https://www.jstatsoft.org/v097/i06
Juan R. González, Lluís Armengol, Xavier Solé, Elisabet Guinó, Josep M. Mercader, Xavier Estivill, and Víctor Moreno. SNPassoc: an R package to perform whole genome association studies. *Bioinformatics*, 23(5):654–655, 01 2007. ISSN 1367-4803. doi:10.1093/bioinformatics/btm025. URL https://doi.org/10.1093/bioinformatics/btm025.

Sewall Wright. Evolution in mendelian populations. *Genetics*, 16(2):97, 1931.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010. doi:https://doi.org/10.1111/j.1467-9868.2010.00740.x.

Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th annual international conference on machine learning*, pages 433–440, 2009.

Rodolphe Jenatton, Jean-Yves Audibert, and Francis Bach. Structured variable selection with sparsity-inducing norms. *The Journal of Machine Learning Research*, 12:2777–2824, 2011.

C. A. R. Hoare. Algorithm 65: Find. *Commun. ACM*, 4(7):321–322, July 1961. ISSN 0001-0782. doi:10.1145/366622.366647