4-29-2012

Optimality of Graphlet Screening in High Dimensional Variable Selection

Jiashun Jin
Carnegie Mellon University, jiashun@stat.cmu.edu

Cun-Hui Zhang
Rutgers University

Qi Zhang
University of Pittsburgh

Optimality of Graphlet Screening in High Dimensional Variable Selection

Jiashun Jin[1], Cun-Hui Zhang[2] and Qi Zhang[3]
[1]Carnegie Mellon University, [2]Rutgers University, [3] University of Pittsburgh

April 29, 2012

Abstract

Consider a linear model $Y = X\beta + \sigma z$, where X has n rows and p columns and $z \sim N(0, I_n)$. We assume both p and n are large, including the case of $p \gg n$. The unknown signal vector β is assumed to be sparse in the sense that only a small fraction of its components is nonzero. The goal is to identify such nonzero coordinates (i.e., variable selection).

We are primarily interested in the regime where signals are both *rare and weak* so that successful variable selection is challenging but is still possible. Researches on rare and weak signals to date have been focused on the unstructured case, where the Gram matrix $G = X'X$ is nearly orthogonal. In this paper, G is only assumed to be sparse in the sense that each row of G has relatively few large coordinates (diagonals of G are normalized to 1). The sparsity of G naturally induces the sparsity of the so-called *graph of strong dependence* (GOSD). The key insight is that there is an interesting interplay between the signal sparsity and graph sparsity: in a broad context, the signals decompose into many small-size components of GOSD that are disconnected to each other.

We propose *Graphlet Screening* (GS) for variable selection. This is a two-step Screen and Clean procedure, where in the first step, we screen subgraphs of GOSD with sequential χ^2-tests, and in the second step, we clean with penalized MLE. The main methodological innovation is to use GOSD to guide both the screening and cleaning processes.

For any variable selection procedure $\hat{\beta}$, we measure its performance with the Hamming distance between the sign vectors of $\hat{\beta}$ and β, and assess the optimality by the convergence rate of the Hamming distance. Compared with more stringent criterions such as exact support recovery or oracle property, which demand strong signals, the Hamming distance criterion is more appropriate for weak signals since it naturally allows a small fraction of errors.

We show that in a broad class of situations, Graphlet Screening achieves the optimal rate of convergence in terms of the Hamming distance. Well-known procedures such as the L^0-penalization method and the L^1-penalization methods do not utilize graph structure for variable selection, so they generally do not achieve the optimal rate of convergence, even in very simple settings and even when the tuning parameters are ideally set.

Keywords: Asymptotic minimaxity, graph of strong dependence (GOSD), graph of least favorables (GOLF), Hamming distance, Graphlet Screening, phase diagram, Rare and Weak signal model, Screen and Clean, sparsity.

AMS 2000 subject classifications: Primary 62G05, 62G10; secondary 62G20.

Acknowledgments: We thank Pengsheng Ji and Tracy Ke for help and pointers. JJ and QZ were partially supported by NSF CAREER award DMS-0908613. QZ was also partially supported by the NIH Grant P50-MH080215-02. CHZ was partially supported by the NSF Grants DMS-0906420, DMS-1106753 and NSA Grant H98230-11-1-0205. The research was supported in part by Computational Resources on PittGrid.

1 Introduction

Consider a linear regression model

$$Y = X\beta + \sigma z, \qquad z \sim N(0, I_n), \tag{1.1}$$

where the design matrix $X = X_{n,p}$ has n rows and p columns. Throughout this paper, we assume the diagonals of the Gram matrix

$$G = X'X$$

are normalized to 1 (and approximately 1 in the random design model). Motivated by the recent trend of 'Big Data' where massive datasets consisting of millions or billions of observations and variables are mined for associations and patterns (e.g. genomics, compressive sensing), we are primarily interested in the case where both p and n are large with $p \geq n$ (though this should not be taken as a restriction). The signal vector β is unknown to us, but is presumably *sparse* in the sense that only a small proportion of its coordinates is nonzero. The main interest of this paper is to identify such nonzero coordinates (i.e., variable selection).

1.1 The paradigm of rare and weak signals

We are primarily interested in the regime where the signals are both *rare* and *weak*: Whether we are talking about clickstreams in web browsing or genome scans or tick-by-tick financial data, most of what we see is noise; the signals, mostly very subtle, are hard to find, and it's easy to be fooled.

While rarity (or sparsity) of the signal is a well-accepted concept in high dimensional data analysis, the weakness of the signal is a much neglected notion. Many contemporary studies of variable selection have focused on rare and strong signals, where the so-called *oracle property* or *probability of exact support recovery* are used as the measure of optimality. Typically, these works assume the signals are sufficiently strong, so that the variable selection problem does not involve the subtle tradeoff between signal sparsity and signal strength. However, such a tradeoff is of great interest from both scientific and practical perspectives.

In this paper, we focus on the regime where the signals are so rare and weak that they are barely separable from the noise. We are interested in the exact demarcation that separates the region of impossibility from the region of possibility. In the region of impossibility, the signal is so rare and weak that successful variable selection is impossible. In the region of possibility, the signals are strong enough so that successful variable selection is possible. in the sense of committing a much smaller number of selection errors than the number of signals. This is a very delicate situation, where it is of major interest to develop methods that yields successful variable selection.

When signals are rare and weak, exact recovery is usually impossible, and oracle property or probability of exact support recovery is no longer an appropriate criterion for assessing optimality. In this paper, we use the minimax Hamming distance as a measure of optimality. Hamming distance is the expected number of components for which the estimated signs and true signs of the regression coefficients disagree. Our primary goal is to study the rate of minimax Hamming distance and to develop procedures that achieve the minimax rate.

1.2 Exploiting the sparsity of the graph of strong dependence

Most of the work to date on "rare and weak" effects consider the completely unstructured case where no two features interact with each other in a significant way [11, 12, 23, 29]. However, in numerous applications, there are relationships between predictors which are important to consider.

In this paper, we are primarily interested in the class of linear models where the Gram matrix G is 'sparse', in the sense that each row of G only has relatively few large coordinates. Linear models where G are sparse can be found in the following application areas.

- *Compressive sensing.* We are interested in a very high dimensional sparse vector β. The plan is to store or transmit n linear functionals of β and then reconstruct it. For $1 \leq i \leq n$, we choose a p-dimensional coefficient vector X_i and observe $Y_i = X_i'\beta + z_i$ with an error z_i. The so-called Gaussian design is often considered [9, 10, 2], where $X_i \overset{iid}{\sim} N(0, \Omega/n)$ with a sparse covariance matrix Ω. In this example, the sparsity of Ω induces that of $G = X'X$.

- *Genetic Regulatory Network (GRN).* For $1 \leq i \leq n$, $W_i = (W_i(1), \ldots, W_i(p))'$ represents the expression level of p different genes corresponding to the i-th patient. Approximately, $W_i \overset{iid}{\sim} N(\alpha, \Sigma)$, where the contrast mean vector α is sparse reflecting that only few genes are differentially expressed between a normal patient and a diseased one [30]. Frequently, the concentration matrix $\Omega = \Sigma^{-1}$ is believed to be sparse, and can be effectively estimated in some cases (e.g. [3, 4]), or can be assumed as known in others, with the so-called "data about data" available [26]. To estimate α, one may consider $Y = n^{-1/2}\sum_{i=1}^n W_i \sim N(\sqrt{n}\alpha, \Sigma)$ and use brute-force thresholding. However, such an approach is inefficient as it neglects the correlation structure. Alternatively, let $\hat{\Omega}$ be a positive-definite estimate of Ω, the problem can be re-formulated as the following linear model: $(\hat{\Omega})^{1/2}Y \approx \Omega^{1/2}Y \sim N(\Omega^{1/2}\beta, I_p)$, where $\beta = \sqrt{n}\alpha$ and $G \approx \Omega$, and both are sparse.

Other examples can be found in Computer Security [25] and Factor Analysis [21].

Well-known approaches to variable selection include subset selection, the lasso, SCAD, MC+, greedy search and more [1, 6, 13, 14, 31, 32, 36, 38, 39, 40]. While these approaches may exploit the *signal sparsity* effectively, they are not designed to take advantage of the sparsity of the graphical structure of the design variables. It is therefore of great interest to study how to exploit such *graph sparsity* to substantially improve variable selection. This is particularly important in the "rare and weak" paradigm, where it is so easy to be fooled by noise.

In fact, in such a paradigm, even the 'optimal' penalized least squares methods (including exhaustive subset selection) are non-optimal. The exhaustive subset selection is non-optimal because it is a *one-stage* and *non-adaptive* method that does not fully utilize the graphical structure among the design variables. See Section 1.9 for detailed discussion.

In this paper, we propose a new approach to variable selection which we call the *Graphlet Screening* (GS). This is a two-stage Screen and Clean method, where the main methodological innovation is the use of a *graph of strong dependence* (GOSD), constructed from the Gram matrix, to guide both the screening and the cleaning processes. The procedure limits the attention to strong correlated substructures only, and has a two-fold advantage: modest computational cost and theoretic optimality. Below, we begin our discussion with the so-called phenomenon of *interaction of signal sparsity and graph sparsity*, which plays a central role in the proposed method.

1.3 Sparse signal model, interplay of signal sparsity and graph sparsity

Motivated by the above examples, we adopt a sparse signal model as follows (e.g., [5]). Fix parameters $\epsilon \in (0,1)$ and $\tau > 0$. Let $b = (b_1, \ldots, b_p)'$ be the $p \times 1$ random vector where

$$b_i \stackrel{iid}{\sim} \text{Bernoulli}(\epsilon). \tag{1.2}$$

We model the signal vector β as

$$\beta = b \circ \mu, \tag{1.3}$$

where $\circ$ denotes the Hadamard product (i.e., for any $p \times 1$ vectors x and y, $x \circ y$ is the $p \times 1$ vector such that $(x \circ y)_i = x_i y_i$, $1 \le i \le p$), and $\mu \in \Theta_p(\tau)$ with

$$\Theta_p(\tau) = \{\mu \in \mathbb{R}^p : \, |\mu_i| \ge \tau, 1 \le i \le p\}. \tag{1.4}$$

In later sections, we may further restrict μ to a subset of $\Theta_p(\tau)$; see (1.11).

Definition 1.1 *We call (1.2)-(1.4) the Rare and Weak signal model (*$\text{RW}(\epsilon, \tau, \mu)$*).*

In this model, β_i is either 0 or a signal with at least strength τ. The parameter ϵ is unknown to us, but is presumably small so the signals are sparse. At the same time, we take τ to be moderately large (see Section 1.5 for details) so that the signals are barely separable from the noise. This models a situation where the signals are both rare and weak.

Naturally, a sparse Gram matrix induces a sparse graph among design vectors, which we call the *graph of strong dependence* (GOSD). Towards this end, write

$$X = [x_1, x_2, \ldots, x_p] = [X_1, X_2, \ldots, X_n]' \tag{1.5}$$

so that x_j is the j-th column of X and X_i' is the i-th row of X. For a tuning parameter $\delta > 0$ ($\delta = 1/\log p$ or other small values of logarithmic order), we introduce

$$\Omega^* = (\Omega^*(i,j))_{p \times p}, \qquad \Omega^*(i,j) = G(i,j) 1\{|G(i,j)| \ge \delta\}, \tag{1.6}$$

as a *regularized Gram matrix*.

Definition 1.2 *The GOSD is the graph $\mathcal{G}^* = (V, E)$, where $V = \{1, 2, \ldots, p\}$ and nodes i and j are connected if and only if $\Omega^*(i,j) \neq 0$.*

If each row of Ω^* has no more than K nonzeros, then the graph $\mathcal{G}^*$ is *K-sparse*.

Definition 1.3 *A graph $\mathcal{G} = (V, E)$ is K-sparse if the degree of each node is no greater than K.*

At first glance, it is unclear how the sparsity of $\mathcal{G}^*$ may help in variable selection. In fact, for any fixed node i, even when K is as small as 2, it is possible to have a very long path that connects node i to another node j. Therefore, it is unclear how to remove the influence of other nodes when we attempt to make inference about node i.

However, on a second thought, we note that what is crucial to variable selection is not the graph $\mathcal{G}^*$, but the subgraph of $\mathcal{G}^*$ formed by all the signal nodes. Compared to the whole graph, this subgraph not only has a much smaller size, but also has a much simpler structure: It decomposes into many components, each of which is small in size, and different components are disconnected (a component is a maximal connected subgraph). The following notation is frequently used in this paper.

Definition 1.4 *Fixing a graph $\mathcal{G}$, we say $\mathcal{I}_0 \lhd \mathcal{G}$ if $\mathcal{I}_0$ is a component of $\mathcal{G}$.*

In other words, due to the interplay between the signal sparsity and the graph sparsity, the original regression problem is *decomposable*: the signals live in isolated units, each is small in size (if only we know where they are!), and different units are disconnected to each other. So to solve the original regression problem, it is sufficient to solve many small-size regression problems parallely, where one problem has little influence over the others.

Formally, denote the support of the signal vector by

$$S = S(\beta) = \{1 \leq i \leq p : \beta_j \neq 0\}.$$

Let $\mathcal{G}_S^*$ be the subgraph of $\mathcal{G}^*$ formed by all the nodes in S. The following lemma is proved in Section 5.

Lemma 1.1 *Fixing $K \geq 1$, $m \geq 1$, $\epsilon > 0$, $\tau > 0$, suppose $\mathcal{G}^*$ is K-sparse and β is from the Rare and Weak model $RW(\epsilon, \tau, \mu)$. Then, with at least probability $1 - p(e\epsilon K)^{m+1}$, $\mathcal{G}_S^*$ decomposes into many components, each has a size $\leq m$, and different ones are disconnected.*

For moderately sparse signals (e.g. in an asymptotic framework where as $p \to \infty$, $\epsilon = \epsilon_p \leq p^{-\vartheta}$ for some fixed parameter $\vartheta > 0$), $p(e\epsilon K)^{m+1}$ is small so that the decomposability in Lemma 1.1 holds with overwhelming probability. We mention that Lemma 1.1 is not tied to Model (1.2)-(1.4) and holds in much broader settings. For example, a similar claim can be drawn if the vector b in $\beta = b \circ \mu$ satisfies a certain Ising model [24]. The decomposability of $\mathcal{G}_S^*$ is mainly due to the interplay of the signal sparsity and the graph sparsity, not the specific model of the signals. For further elaboration on this point, see the proof of Lemma 1.1 in Section 5.

1.4 Graphlet screening

The aforementioned decomposability invites the following two-stage variable selection procedure, which we call the *Graphlet Screening (GS)*. Conceptually, the procedure contains a graphical screening step (GS-step) and a graphical cleaning step (GC-step).

- *GS-step.* This is an m-stage χ^2-screening process, where $m \geq 1$ is a preselected integer. In this process, we investigate all connected subgraphs of $\mathcal{G}^*$ of no more than m nodes. For each of them, we test whether some of the nodes in the connected subgraph are signals, or none of them is a signal. We then retain all those which we believe to contain one or more signals.

- *GC-step.* The surviving nodes decompose into many components, each of which has no more than ℓ_0 nodes, where ℓ_0 is a fixed small number. We then fit each component with penalized MLE, in hopes of removing all falsely kept signals.

In philosophy, the GS is similar to [34, 15] in that they have a screening and a cleaning stage, but is more sophisticated in nature.

We now describe two steps in details. Recalling (1.5), we have the following definition.

Definition 1.5 *For X in Model (1.1) and any subset $\mathcal{I} \subset \{1, 2, ..., p\}$, let $P^{\mathcal{I}} = P^{\mathcal{I}}(X)$ be the projection from $\mathbb{R}^n$ to the span of $\{x_j, j \in \mathcal{I}\}$.*

Consider the *GS*-step first. Let $\mathcal{G}^*$ be as in (1.6) and fix $m \geq 1$. The *m-stage χ^2-screening* is as follows.

- *Initial sub-step.* Let $\mathcal{U}_p^* = \emptyset$. List all connected subgraphs of $\mathcal{G}^*$, say $\mathcal{I}_0$, in ascending order of the number of nodes $|\mathcal{I}_0|$, with ties broken lexicographically, subject to $|\mathcal{I}_0| \leq m$. Since a node is thought of as connected to itself, the first p connected subgraphs on the list are simply the nodes $1, 2, \ldots, p$. We screen all connected subgraphs in the order they are listed.

- *Updating sub-step.* Let $\mathcal{I}_0$ be the connected subgraph under consideration, and let $\mathcal{U}_p^*$ be the current set of retained indices. We update $\mathcal{U}_p^*$ with a χ^2 tests as follows. Let $\hat{F} = \mathcal{I}_0 \cap \mathcal{U}_p^*$ and $\hat{D} = \mathcal{I}_0 \setminus \mathcal{U}_p^*$, so that $\hat{F}$ is the set of nodes in $\mathcal{I}_0$ that have already been accepted, and $\hat{D}$ is the set of nodes in $\mathcal{I}_0$ that is currently under investigation. Note that no action is needed if $\hat{D} = \emptyset$. For a threshold $t(\hat{D}, \hat{F}) > 0$ to be determined, we update $\mathcal{U}_p^*$ by adding all nodes in $\hat{D}$ to it if

$$T(Y, \hat{D}, \hat{F}) = \|P^{\mathcal{I}_0} Y\|^2 - \|P^{\hat{F}} Y\|^2 > t(\hat{D}, \hat{F}), \tag{1.7}$$

  and we keep $\mathcal{U}_p^*$ the same otherwise (by default, $\|P^{\hat{F}} Y\| = 0$ if $\hat{F} = \emptyset$). We continue this process until we finish screening all connected subgraphs on the list.

In the *GS*-step, once a node is kept in any sub-stage of the screening process, it remains there until the end of the *GS*-step (however, it may be killed in the *GC*-step). This has a similar flavor to that of the Forward regression. See Table 1 for a recap of the procedure.

The GS-step uses the following set of tuning parameters:

$$\mathcal{Q} \equiv \{t(\hat{D}, \hat{F}) : (\hat{D}, \hat{F}) \text{ are as defined in (1.7)}\}.$$

A convenient way to set these parameters is to let $t(\hat{D}, \hat{F}) = 2\sigma^2 q \log p$ for a fixed $q > 0$ and all $(\hat{D}, \hat{F})$. More sophisticated choices are given in Section 1.7.

The computational cost of the *GS*-step hinges on the sparsity of $\mathcal{G}^*$. In Section 1.5, we show that with a properly chosen δ, for a wide class of design matrices, $\mathcal{G}^*$ is K-sparse for some $K = K_p \leq (\log(p))^\alpha$ as $p \to \infty$, where $\alpha > 0$ is a constant. As a result, the computational cost of the *GS*-step is moderate, because for any K-sparse graph, there are at most $p(eK)^m$ subgraphs with size m [19].

The *GS*-step has two important properties: *Sure Screening* and *Separable After Screening (SAS)*. With tuning parameters $\mathcal{Q}$ properly set, the Sure Screening property says that $\mathcal{U}_p^*$ retains all but a negligible fraction of the signals. Viewing $\mathcal{U}_p^*$ as a subgraph of $\mathcal{G}^*$, the SAS property says that this subgraph decomposes into many disconnected components, each has a size $\leq \ell_0$ for a fixed small integer ℓ_0. Together, these two properties enable us to reduce the original large-scale regression problem to many small-size regression problems that can be solved parallelly in the *GC*-step. See Section 2 for elaboration on these ideas.

We now discuss the *GC*-step. The following notation is frequently used in this paper.

Table 1: Graphet Screening Algorithm

GS-step:	List $\mathcal{G}^*$-connected submodels $\mathcal{I}_{0,k}$ with $\lvert\mathcal{I}_{0,1}\rvert \le \lvert\mathcal{I}_{0,2}\rvert \le \cdots \le m$
	Initialization: $\mathcal{U}_p^* = \emptyset$ and $k = 1$
	Test $H_0 : \mathcal{I}_{0,k} \cap \mathcal{U}_p^*$ against $H_1 : \mathcal{I}_{0,k}$ with χ^2 test (1.7)
	Update: $\mathcal{U}_p^* \leftarrow \mathcal{U}_p^* \cup \mathcal{I}_{0,k}$ if H_0 rejected, $k \leftarrow k+1$
GC-step:	As a subgraph of $\mathcal{G}^*$, $\mathcal{U}_p^*$ decomposes into many components $\mathcal{I}_0$
	Use the L^0-penalized test (1.8) to select a subset $\hat{\mathcal{I}}_0$ of each $\mathcal{I}_0$
	Return the union of $\hat{\mathcal{I}}_0$ as the selected model

Definition 1.6 *For a $p \times m$ matrix X and subsets $\mathcal{I}_0 \subset \{1, 2, \ldots, p\}$ and $\mathcal{J}_0 \subset \{1, 2, \ldots, m\}$, $X^{\mathcal{I}_0, \mathcal{J}_0}$ denotes the $\lvert\mathcal{I}_0\rvert \times \lvert\mathcal{J}_0\rvert$ sub-matrix of X formed by restricting the rows of X to $\mathcal{I}_0$ and columns to $\mathcal{J}_0$. For short, in the case where $\mathcal{J}_0 = \{1, 2, \ldots, m\}$, we write it as $X^{\mathcal{I}_0}$, and in the case where $\mathcal{I}_0 = \{1, 2, \ldots, p\}$, we write it as $X^{*,\mathcal{J}_0}$. When $m = 1$, X is a vector, and $X^{\mathcal{I}_0}$ is the sub-vector of X formed by restricting the rows of X to $\mathcal{I}_0$.*

For any $1 \le j \le p$, we have either $j \notin \mathcal{U}_p^*$, or that there is a unique connected subgraph $\mathcal{I}_0$ such that $j \in \mathcal{I}_0 \lhd \mathcal{U}_p^*$ (see Definition 1.3 for the notation). In the first case, we estimate β_j as 0. In the second case, for two tuning parameters $u^{gs} > 0$ and $v^{gs} > 0$, we estimate the whole set of variables $\beta^{\mathcal{I}_0}$ by minimizing the following functional:

$$\|P^{\mathcal{I}_0}(Y - X^{*,\mathcal{I}_0}\xi)\|^2 + (u^{gs})^2\|\xi\|_0. \tag{1.8}$$

Here, ξ is an $\lvert\mathcal{I}_0\rvert \times 1$ vector each nonzero coordinate of which $\ge v^{gs}$ in magnitude, and $\|\xi\|_0$ is the L^0-norm of ξ. The resultant estimator is the final estimate of Graphlet Screening which we denote by $\hat{\beta}^{gs} = \hat{\beta}^{gs}(Y; \delta, \mathcal{Q}, u^{gs}, v^{gs}, X, p, n)$.

The computational cost of the GC-step hinges on maximal size of the components of $\mathcal{U}_p^*$. By the SAS property of the GS-step (see Lemma 2.3 for details), for a broad class of design matrices, with the tuning parameters chosen properly, there is a *fixed* integer ℓ_0 such that with overwhelming probability, $\lvert\mathcal{I}_0\rvert \le \ell_0$ for any $\mathcal{I}_0 \lhd \mathcal{U}_p^*$. As a result, the computational cost of the GC-step is no greater than $\lvert\mathcal{U}_p^*\rvert \times 2^{\ell_0}$, which is moderate.

How does Graphlet Screening behave? Surprisingly well. In sections below, we show that Graphlet Screening achieves the minimax Hamming errors over a wide class of Rare and Weak models. Towards this end, we invoke the random design model. The use of random design model is mainly for simplicity in presentation. In particular, it is much easier to elaborate the sparsity of the Gram matrix in a random design model than a fixed design model. On the other hand, the main results in this section can be translated to fixed design models with careful modification of the notation. See for example Corollary 1.1 and Section 4.

1.5 Asymptotic Rare and Weak model for regression with random design

We continue our discussion with the Rare and Weak model $\mathrm{RW}(\epsilon, \tau, \mu)$ by introducing an asymptotic framework. In this framework, we let p be the driving asymptotic parameter, and parameters (ϵ, τ) are tied to p through some fixed parameters. In detail, fixing $0 < \vartheta < 1$, we model

$$\epsilon = \epsilon_p = p^{-\vartheta}. \tag{1.9}$$

For any fixed ϑ, the signals become increasingly sparser as $p \to \infty$. Also, as ϑ ranges, the sparsity level ranges from very dense to very sparse, and covers most interesting cases.

It turns out that the most interesting range for τ is $\tau = \tau_p = O(\sqrt{\log(p)})$. In fact, when $\tau_p \ll \sigma\sqrt{\log(p)}$, the signals are simply too rare and weak so that successful variable selection is impossible. On the other hand, when τ_p is sufficiently large, it is possible to exactly recover the support of β under proper conditions on the design. In light of this, we fix $r > 0$ and calibrate τ by

$$\tau = \tau_p = \sigma\sqrt{2r\log(p)}. \tag{1.10}$$

At the same time, fixing a constant $a > 1$, in the $RW(\epsilon, \tau, \mu)$, we further restrict the vector μ to a subset of $\Theta_p(\tau_p)$, denoted by $\Theta_p^*(\tau_p, a)$, where

$$\Theta_p^*(\tau_p, a) = \{\mu \in \Theta_p(\tau_p) : |\mu_i| \leq a\tau_p, i = 1, 2, \ldots, p\}, \tag{1.11}$$

and the parameter a is unknown. The constraint of $|\mu_i| \leq a\tau_p$ is mainly for technical reasons (only needed in the proof of Lemma 2.3). Hopefully, in the near future, such a constraint can be removed. See later part of this subsection for more discussion on the role of a.

Definition 1.7 *We call model (1.2)-(1.4) and (1.9)-(1.11) the Asymptotic Rare Weak signal model* $\mathrm{ARW}(\vartheta, r, a, \mu)$.

We now introduce the random design model. Fix a correlation matrix Ω that is presumably unknown to us (however, for simplicity, we assume that Ω has unit diagonals). In the random design model, we assume that the rows of X as iid samples from a p-variate zero means Gaussian vector with correlation matrix Ω:

$$X_i \overset{iid}{\sim} N(0, \frac{1}{n}\Omega). \tag{1.12}$$

The factor $1/n$ is chosen so that the diagonal elements of the Gram matrix G are approximately one. In the literature, this is called the Gaussian design, which can be found in Compressive Sensing [2], Computer Security [8], and other application areas.

At the same time, fixing $\kappa \in (0, 1)$, we model the sample size n by

$$n = n_p = p^{\kappa}. \tag{1.13}$$

As $p \to \infty$, n_p becomes increasingly large but is still much smaller than p. We assume

$$\kappa > (1 - \vartheta), \tag{1.14}$$

so that $n_p \gg p\epsilon_p$. Note $p\epsilon_p$ is approximately the total number of signals. Condition (1.14) is almost necessary for successful variable selection [9, 10].

Definition 1.8 *We call Model (1.12)-(1.14) the Random Design model* $\mathrm{RD}(\vartheta, \kappa, \Omega)$.

Come back to (1.11). From a practical point of view, it is preferable to assume a moderately large (but fixed) a, since we usually don't have sufficient knowledge on μ. For this reason, we are primarily interested in the case where a is "appropriately" large. This will make $\Theta_p^*(\tau_p, a)$ sufficiently broad so that neither the minimax rate nor any variable selection procedure needs to adapt to a.

Towards this end, we impose some mild "local" regularity conditions on Ω. In detail, for any positive definite matrix A, let $\lambda(A)$ be the smallest eigenvalue, and let

$$\lambda_k^*(\Omega) = \min\{\lambda(A) : A \text{ is a } k \times k \text{ principle submatrix of } \Omega\}. \tag{1.15}$$

At the same time, fixing a constant $c_0 > 0$, let (ϑ, r) be as in (1.9) and (1.10), respectively, let m be as in the GS-step, and let g be the smallest integer such that

$$g \geq \max\{m, (\vartheta + r)^2/(2r)\}. \tag{1.16}$$

Introduce

$$\mathcal{M}_p(c_0, g) = \{\Omega : p \times p \text{ correlation matrix, } \lambda_k^*(\Omega) \geq c_0, 1 \leq k \leq g\}.$$

For any two subsets V_0 and V_1 of $\{1, 2, \ldots, p\}$, consider the optimization problem

$$\big(\theta_*^{(0)}(V_0, V_1), \theta_*^{(1)}(V_0, V_1)\big) = \text{argmax}\{(\theta^{(1)} - \theta^{(0)})'\Omega(\theta^{(1)} - \theta^{(0)})\}, \tag{1.17}$$

subject to the constraints that for $k = 0, 1$, $\theta^{(k)}$ are $p \times 1$ vectors satisfying $|\theta_i^{(k)}| \geq 1$ for $i \in V_k$ and $\theta_i^{(k)} = 0$ otherwise, and that the sign vectors of $\theta^{(0)}$ and $\theta^{(1)}$ are unequal. Introduce

$$a_g^*(\Omega) = \max_{\{(V_0, V_1): |V_0 \cup V_1| \leq g\}} \max\{\|\theta_*^{(0)}(V_0, V_1)\|_\infty, \|\theta_*^{(1)}(V_0, V_1)\|_\infty\}.$$

We have the following lemma, the proof of which is elementary and thus omitted.

Lemma 1.2 *For any $\Omega \in \mathcal{M}_p(c_0, g)$, there is a constant $C = C(c_0, g)$ such that $a_g^*(\Omega) \leq C$.*

In this paper, unless stated otherwise, we assume

$$\Omega \in \mathcal{M}_p(c_0, g), \qquad a > a_g^*(\Omega). \tag{1.18}$$

One exception is Theorem 1.1, where the result holds without such a constraint. In Section 1.7, we further restrict Ω to a subset of $\mathcal{M}_p(c_0, g)$ to foster graph sparsity. Condition (1.18) is mild for it involves only small-size principle sub-matrices of Ω, and we assume $a > a_g^*(\Omega)$ mostly for simplicity. For insight, imagine that in (1.17), we further require that $|\theta_i^{(k)}| \leq a\tau_p$, $i \in V_k$, $k = 0, 1$. Then as long as $a > a_g^*(\Omega)$, the optimization problem in (1.17) has exactly the same solution, which does not depend on a. This explains (1.18).

For any fixed β and any variable selection procedure $\hat{\beta}$, we measure the performance by the Hamming distance between the sign vectors $\text{sgn}(\hat{\beta})$ and $\text{sgn}(\beta)$:

$$h_p(\hat{\beta}, \beta\big|X) = E\Big[\sum_{j=1}^{p} 1\big(\text{sgn}(\hat{\beta}_j) \neq \text{sgn}(\beta_j)\big)\big|X\Big].$$

In the Asymptotic Rare Weak model, $\beta = b \circ \mu$, and (ϵ_p, τ_p) depend on p through (ϑ, r), so the overall Hamming distance for $\hat{\beta}$ is

$$H_p(\hat{\beta}; \epsilon_p, n_p, \mu, \Omega) = E_{\epsilon_p} E_\Omega\big[h_p(\hat{\beta}, \beta\big|X)\big] \equiv E_{\epsilon_p} E_\Omega\big[h_p(\hat{\beta}, b \circ \mu\big|X)\big],$$

where E_{ϵ_p} is the expectation with respect to the law of b, and E_Ω is the expectation with respect to the law of X; see (1.2) and (1.12). Finally, the minimax Hamming distance is

$$\text{Hamm}_p^*(\vartheta, \kappa, r, a, \Omega) = \inf_{\hat{\beta}} \sup_{\mu \in \Theta_p^*(\tau_p, a)} \big\{H_p(\hat{\beta}; \epsilon_p, n_p, \mu, \Omega)\big\}.$$

In the above definitions, $\text{sgn}(x) = 0, 1, -1$ if $x = 0$, $x > 0$, and $x < 0$ correspondingly. Note that the Hamming distance is no smaller than the sum of the expected number of signal components that are misclassified as noise and the expected number of noise components that are misclassified as signal.

1.6 Lower bound for the minimax Hamming distance, and GOLF

In this section, we construct a lower bound for the minimax Hamming distance. We focus our discussion on the Random Design model, but will also address deterministic design. The main insight is that, the lower bound depends on the structure of small-size connected subgraphs of $\mathcal{G}^*$, rather than the large-scale structures of $\mathcal{G}^*$. This is a direct result of the *decomposability* of $\mathcal{G}_S^*$ mentioned earlier.

It is noteworthy that the constructed lower bound is useful provided that the so-called *graph of least favorables* (GOLF) (to be introduced below) is sparse. The GOLF can be sparse even if $\mathcal{G}^*$ is heavily non-sparse, so the lower bound can be useful in rather general context, with only minimal conditions imposed on Ω.

Below, we first derive such a lower bound without assuming (1.18). We then discuss the case where (1.18) holds, and derive a useful alternative expression for the lower bound. The key in the construction is to derive lower bounds for the so-called "local risk". The "local risk" at an index j, $1 \leq j \leq p$, is the risk of estimating the set of variables $\{\beta_k : d(k,j) \leq g\}$, where g is defined in (1.16) and $d(j,k)$ denotes the geodesic distance between j and k in the graph $\mathcal{G}^*$. Aggregating the lower bounds for the "local risk" at different j gives a lower bound for the global risk, where the effect of repetitive counting is controlled by the sparsity of the GOLF.

We now construct a lower bound for the "local risk" at site j. The "local risk" is characterized by the exponent $\rho_j^*(\vartheta, r, a, \Omega)$, which depends on Ω in a complicated way, and it takes relatively long preparations to describe it.

In detail, for a lower bound for the "local risk" at site j, the goal is to construct two subsets V_0 and V_1 (maybe equal) of $\{1, 2, \ldots, p\}$ and two realization of β, $\beta^{(0)}$ and $\beta^{(1)}$, such that their sign vectors are unequal, and

- (a) $j \in V_1 \cup V_0$,
- (b) $\beta^{(0)}$ and $\beta^{(1)}$ take same values at any $k \notin V_0 \cup V_1$,
- (c) for any $k \in V_0 \cup V_1$, $\beta_k^{(i)} \neq 0$ if and only if $k \in V_i$, $i = 0, 1$.

In the literature, it is known that how well we can estimate $\{\beta_k : d(k,j) \leq g\}$ depends on how well we can test the model $Y = X\beta^{(0)} + \sigma z$ against the model $Y = X\beta^{(1)} + \sigma z$, where $z \sim N(0, I_n)$. These two models can be thought of as an original model and a tampered one, respectively. In the context of lower bound construction, we usually assume that $\beta^{(0)}$ and $\beta^{(1)}$ are known, but we don't know which of the two models is true. The least favorable scenario corresponds to the "worst-case" configuration of $(V_0, V_1, \beta^{(0)}, \beta^{(1)})$, for which two regression models are the most difficult to separate.

Now, for any subset $V \subset \{1, 2, \ldots, p\}$, let I_V be the $p \times 1$ vector such that $(I_V)_k = 1$ if $k \in V$ and 0 otherwise, and let B_V be the set of vectors

$$B_V = \{\theta = I_V \circ \mu : \mu \in \Theta_p^*(\tau_p, a)\}.$$

Rewriting $\beta^{(1)} - \beta^{(0)} = \theta^{(1)} - \theta^{(0)}$ such that $\theta^{(i)} \in B_{V_i}$, $i = 0, 1$, we introduce

$$\alpha(\theta^{(0)}, \theta^{(1)}) = \tau_p^{-2}(\theta^{(0)} - \theta^{(1)})'\Omega(\theta^{(0)} - \theta^{(1)}).$$

In the aforementioned testing problem, it is seen that the optimal test is to reject the original model if and only if $(\theta^{(1)} - \theta^{(0)})'X'(Y - X\beta^{(0)}) \geq t\sigma\tau_p\sqrt{\alpha(\theta^{(0)}, \theta^{(1)})}$ for some

threshold $t > 0$ to be determined, and the sum of Type I and Type II error of any test is no smaller than (up to some negligible differences)

$$\inf_t \big[\epsilon_p^{|V_0|}\bar{\Phi}(t) + \epsilon_p^{|V_1|}\Phi\big(t - (\tau_p/\sigma)[\alpha(\theta^{(0)}, \theta^{(1)})]^{1/2}\big)\big]. \tag{1.19}$$

This gives a lower bound for the "local risk" at index j. Here $\bar{\Phi} = 1 - \Phi$ is the survival function of $N(0,1)$. In the above calculations, we have used that in the Random Design model,

$$(\theta^{(0)} - \theta^{(1)})'G(\theta^{(0)} - \theta^{(1)}) \approx (\theta^{(0)} - \theta^{(1)})'\Omega(\theta^{(0)} - \theta^{(1)}),$$

since the support of $\theta^{(0)} - \theta^{(1)}$ is contained in $V_0 \cup V_1$, which is a small-size set.

To tighten the lower bound in (1.19), we look for the "worst-case" configuration of $(V_0, V_1, \theta^{(0)}, \theta^{(1)})$. Towards this end, first, we fix (V_0, V_1) and look for the "worst-case" $(\theta^{(0)}, \theta^{(1)})$. To do so, define $\alpha^*(V_0, V_1; \Omega) = \alpha^*(V_0, V_1; \vartheta, r, a, \Omega, p)$ by

$$\alpha^*(V_0, V_1; \Omega) = \min\big\{\alpha(\theta^{(0)}, \theta^{(1)}) : \ \theta^{(i)} = I_{V_i} \circ \mu^{(i)}, \mu^{(i)} \in \Theta_p^*(\tau_p, a), i = 0, 1, \mathrm{sgn}(\theta^{(0)}) \neq \mathrm{sgn}(\theta^{(1)})\big\},$$

where $\mathrm{sgn}(\theta)$ denotes the sign vector of θ. By (1.19) and monotonicity, a lower bound for the "local risk" at index j is then

$$\sup_{\{(V_0,V_1):\, j \in V_0 \cup V_1\}} \big\{\inf_t \big[\epsilon_p^{|V_0|}\bar{\Phi}(t) + \epsilon_p^{|V_1|}\Phi\big(t - (\tau_p/\sigma)[\alpha^*(V_0, V_1; \Omega)]^{1/2}\big)\big\}. \tag{1.20}$$

Next, we find the "locally worst-case" (V_0, V_1) for (1.20). The following shorthand notation is frequently used in this paper, which stands for a generic multi-$\log(p)$ term that may vary from one occurrence to another.

Definition 1.9 *$L_p > 0$ denotes a multi-$\log(p)$ term such that when $p \to \infty$, for any $\delta > 0$, $L_p p^{\delta} \to \infty$ and $L_p p^{-\delta} \to 0$.*

Now, introduce $\eta(V_0, V_1; \Omega) = \eta(V_0, V_1; \vartheta, r, a, \Omega, p)$ by

$$\eta(V_0, V_1; \Omega) = \max\{|V_0|, |V_1|\}\vartheta + \frac{1}{4}\left[\left(\sqrt{\alpha^*(V_0, V_1; \Omega)r} - \frac{\big|(|V_1| - |V_0|)\big|\vartheta}{\sqrt{\alpha^*(V_0, V_1; \Omega)r}}\right)_+\right]^2.$$

Recalling $\tau_p/\sigma = \sqrt{2r\log p}$ and $\epsilon_p = p^{-\vartheta} = \exp(-\vartheta \log p)$, by Mills's ratio the lower bound (1.20) can be equivalently written as

$$L_p \exp\Big(-\min_{\{(V_0,V_1):\, j \in V_1 \cup V_0\}} \eta(V_0, V_1; \Omega)\log p\Big).$$

In light of this, for any $1 \le j \le p$, the *least favorable configuration* at index j is the triplet $(\rho_j^*, V_{0j}^*, V_{1j}^*)$, where

$$\rho_j^*(\vartheta, r, a, \Omega) = \min_{\{(V_0,V_1):j \in V_1 \cup V_0\}} \eta(V_0, V_1; \Omega), \tag{1.21}$$

and

$$(V_{0j}^*, V_{1j}^*) = \mathrm{argmin}_{\{(V_0,V_1):j \in V_1 \cup V_0\}} \eta(V_0, V_1; \Omega).$$

When there is a tie, pick the pair that appears first lexicographically. Therefore, for any $1 \le j \le p$, $V_{0j}^* \cup V_{1j}^*$ is uniquely defined. The following lemma is proved in Section 5.

Lemma 1.3 *Fix p, ϑ, r, Ω, and $1 \le j \le p$, $\max\{|V_{0j}^* \cup V_{1j}^*|\} \le (\vartheta + r)^2/(2\vartheta r)$.*

So far, we have been focused on the "worst-case" of the "local risk", which is characterized by $L_p\exp(-\rho_j^*(\vartheta,r,a,\Omega)\log(p))$. In order for the sum of "local risk" to reflect the global risk, we need the sparsity of the following graph.

Definition 1.10 *The graph of least favorables (GOLF) is the graph $\mathcal{G}^\diamond=(V,E)$, where $V=\{1,2,\ldots,p\}$ and nodes j and k are connected if and only if $(V_{0j}^*\cup V_{1j}^*)$ and $(V_{0k}^*\cup V_{1k}^*)$ have non-empty intersections.*

Since $k\in(V_{0k}^*\cup V_{1k}^*)$, j and k are connected if $k\in(V_{0j}^*\cup V_{1j}^*)$. Let $d_p(\mathcal{G}^\diamond)$ be the maximum degree of $\mathcal{G}^\diamond$. When $d_p(\mathcal{G}^\diamond)$ is relatively small, the least favorable configurations do not contain a hub: that is, any node k can appear in $(V_{0j}\cup V_{1j}^*)$ for relatively few j. Such a property holds for a broad class of Ω, including even those ill-posed ones. In this very general context, we have the following theorem, which establishes the lower bound.

Theorem 1.1 *As $p\to\infty$, $\mathrm{Hamm}_p^*(\vartheta,\kappa,r,a,\Omega)\geq L_p\big[d_p(\mathcal{G}^\diamond)\big]^{-1}\sum_{j=1}^p p^{-\rho_j^*(\vartheta,r,a,\Omega)}$ in the Random Design model $RD(\vartheta,\kappa,\Omega)$.*

A similar conclusion can be drawn for deterministic design models, the proof of which is similar so we omit it.

Corollary 1.1 *For deterministic design models, the parallel lower bound holds for the minimax Hamming distance with Ω replaced by G in the calculation of ρ_j^* and $d_p(\mathcal{G}^\diamond)$.*

In Theorem 1.1, we have not imposed much conditions on Ω. In the remaining part of this subsection, we assume (1.18) holds. In this case, $\rho_j^*(\vartheta,r,a,\Omega)$ does not depend on a, and have an alternative expression. In detail, for any subsets D and F of $\{1,2,\ldots,p\}$, define $\omega(D,F;\Omega)=\omega(D,F;\vartheta,r,a,\Omega,p)$ by

$$\omega(D,F;\Omega)=\min_{\xi\in\mathbb{R}^{|D|},\min_{i\in D}|\xi_i|\geq 1}\Big\{\xi'\big(\Omega^{D,D}-\Omega^{D,F}(\Omega^{F,F})^{-1}\Omega^{F,D}\big)\xi\Big\}, \tag{1.22}$$

and $\rho(D,F;\Omega)=\rho(D,F;\vartheta,r,a,\Omega,p)$ by

$$\rho(D,F;\Omega)=\frac{(|D|+2|F|)\vartheta}{2}+\begin{cases}\frac{1}{4}\omega(D,F;\Omega)r, & |D| \text{ is even},\\ \frac{\vartheta}{2}+\frac{1}{4}\big[(\sqrt{\omega(D,F;\Omega)r}-\frac{\vartheta}{\sqrt{\omega(D,F;\Omega)r}})_+\big]^2, & |D| \text{ is odd}.\end{cases} \tag{1.23}$$

The following lemma is proved in Section 5.

Lemma 1.4 *Fix $\vartheta,r,g,a,\Omega,c_0$. If (1.18) holds, then $\rho_j^*(\vartheta,r,a,\Omega)$ does not depend on a, and satisfies*

$$\rho_j^*(\vartheta,r,a,\Omega)=\min_{\{(D,F):j\in D\cup F,D\cap F=\emptyset,D\neq\emptyset,|D\cup F|\leq g\}}\rho(D,F;\Omega).$$

Lemma 1.4 provides an alternative way to calculate $\rho_j^*(\vartheta,r,a,\Omega)$, and is particularly useful in proving Corollaries 1.2-1.4.

1.7 Upper bound and optimality of Graphlet Screening

We are now ready for the main result of this paper. When Ω is sparse—so that the GOSD $\mathcal{G}^*$ is sparse—the Hamming distance of Graphlet Screening achieves the rate of convergence prescribed by the lower bound we derived in the preceding section, provided the tuning

parameters are properly set. Therefore, the lower bound is tight and Graphlet Screening is rate optimal.

In detail, fix constants $\gamma \in (0,1)$ and $A > 0$. To foster sparsity of the Gram matrix, we shift our attention from $\mathcal{M}_p(c_0, g)$ (where only a lower bound c_0 for the lower eigenvalue of dimension g is imposed) to the following subset:

$$\mathcal{M}_p^*(\gamma, c_0, g, A) = \Big\{\Omega \in \mathcal{M}_p(c_0, g) : \sum_{j=1}^p |\Omega(i,j)|^\gamma \le A,\ 1 \le i \le p\Big\}. \tag{1.24}$$

Note that any $\Omega \in \mathcal{M}_p^*(\gamma, c_0, g, A)$ is sparse in the sense that each row of Ω has relatively few large coordinates. The sparsity of Ω implies the sparsity of the Gram matrix G, since small-size sub-matrices of G approximately equal to their counterparts of Ω.

In Graphlet Screening, when we regularize GOSD, we set the threshold δ as $1/\log(p)$; see (1.6). Such a choice for threshold is mainly for convenience, and can be replaced by any term that tends to 0 logarithmically fast as $p \to \infty$.

We choose the tuning parameters in the GS-step in a way such that

$$t(\hat{D}, \hat{F}) = 2\sigma^2 q(\hat{D}, \hat{F}) \log p, \tag{1.25}$$

where $q = q(\hat{D}, \hat{F}) > 0$ are chosen as follows: For fixed $q_0 > 0$,

$$\begin{cases} \sqrt{q_0} \le \sqrt{q} \le \sqrt{\omega r} - \sqrt{\frac{(\vartheta+\omega r)^2}{4\omega r} - \frac{|\hat{D}|+1}{2}}\vartheta, & |\hat{D}| \text{ is odd } \&\ \omega r/\vartheta > |\hat{D}| + (|\hat{D}|^2-1)^{1/2}, \\ \sqrt{q_0} \le \sqrt{q} \le \sqrt{\omega r} - \sqrt{\frac{1}{4}\omega r - \frac{1}{2}|\hat{D}|\vartheta}, & |\hat{D}| \text{ is even } \&\ \omega r/\vartheta \ge 2|\hat{D}|, \end{cases} \tag{1.26}$$

and q can be any other number no smaller than q_0 in other cases. Here $\omega = \omega(\hat{D}, \hat{F}; \Omega)$ are defined in the same way as (1.22).

We set the GC-step tuning parameters by

$$u^{gs} = \sigma\sqrt{2\vartheta \log p}, \qquad v^{gs} = \tau_p = \sigma\sqrt{2r \log p}. \tag{1.27}$$

The main result in this paper is the following theorem.

Theorem 1.2 *Consider Model (1.1) where β is modeled by* $\mathrm{ARW}(\vartheta, r, a, \mu)$ *and X is modeled by* $\mathrm{RD}(\vartheta, \kappa, \Omega)$. *Suppose that for sufficiently large p, $a > a_g^*(\Omega)$ and $\Omega \in \mathcal{M}_p^*(\gamma, c_0, g, A)$. Let $\hat{\beta}^{gs} = \hat{\beta}^{gs}(Y; \delta, \mathcal{Q}, u^{gs}, v^{gs}, X, p, n)$ be the Graphlet Screening procedure defined in Section 1.3. If the tuning parameters are set as in (1.25)-(1.27), then as $p \to \infty$,*

$$\sup_{\mu \in \Theta_p^*(\tau_p, a)} H_p(\hat{\beta}^{gs}; \epsilon_p, n_p, \mu, \Omega) \le L_p\Big[p^{1-(m+1)\vartheta} + \sum_{j=1}^p p^{-\rho_j^*(\vartheta, r, a, \Omega)}\Big] + o(1).$$

Note that $\rho_j^* = \rho_j^*(\vartheta, r, a, \Omega)$ does not depend on a. Also, note that in the most interesting range, $\sum_{j=1}^p p^{-\rho_j^*} \gg 1$. So if we choose m properly large (e.g. $(m+1)\vartheta > 1$), then

$$\sup_{\mu \in \Theta_p^*(\tau_p, a)} H_p(\hat{\beta}^{gs}; \epsilon_p, n_p, \mu, \Omega) \le L_p \sum_{j=1}^p p^{-\rho_j^*(\vartheta, r, a, \Omega)}.$$

Together with Theorems 1.1, this says that Graphlet Screening achieves the optimal rate of convergence, adaptively to all Ω in $\mathcal{M}_p^*(\gamma, c_0, g, A)$ and $\beta \in \Theta_p^*(\tau_p, a)$. We call this property

optimal adaptivity. Note that since the diagonals of Ω are scaled to 1 approximately, $\kappa \equiv \log(n_p)/\log(p)$ does not have a major influence over the convergence rate, as long as (1.14) holds.

Theorem 1.2 addresses the case where $a > a_g^*(\Omega)$. We now briefly discuss the case where $a < a_g^*(\Omega)$. In this case, the set $\Theta_p^*(\tau_p, a)$ becomes sufficiently narrow and a starts to have some influence over the optimal rate of convergence, at least for some choices of (ϑ, r). To reflect the role of a, we modify Graphlet Screening as follows: (a) in the GC-step (1.8), limit ξ to the class where either $\xi_i = 0$ or $\tau_p \le |\xi_i| \le a\tau_p$, and (b) in the GS-step, replacing the χ^2-screening by the likelihood based screening procedure; that is, when we screen $\mathcal{I}_0 = \hat{D} \cup \hat{F}$, we accept nodes in $\hat{D}$ only when $h(\hat{F}) > h(\mathcal{I}_0)$, where for any subset $D \subset \{1, 2, \ldots, p\}$,

$$h(D) = \min_{\{\xi = I_D \circ \mu, \mu \in \Theta_p^*(\tau_p, a)\}} \frac{1}{2}\|P^D(Y - X^{*,D}\xi)\|^2 + \vartheta\sigma^2\log(p)|D|.$$

From a practical point of view, this modified procedure depends more on the underlying parameters and is harder to implement. However, this is the price we need to pay when a is small. Since we are primarily interested in the case of relatively larger a (so that $a > a_g^*(\Omega)$ holds), we skip further discussion along this line.

1.8 Phase diagram and examples where $\rho_j^*(\vartheta, r, a, \Omega)$ have simple forms

In general, the exponents $\rho_j^*(\vartheta, r, a, \Omega)$ may depend on Ω in a complicated way. Still, from time to time, one may want to find a simple expression for $\rho_j^*(\vartheta, r, a, \Omega)$. It turns out that in a wide class of situations, simple forms for $\rho_j^*(\vartheta, r, a, \Omega)$ are possible. The surprise is that, in many examples, $\rho_j^*(\vartheta, r, a, \Omega)$ depends more on the trade-off between the parameters ϑ and r (calibrating the signal sparsity and signal strength, respectively), rather than on the large coordinates of Ω.

We begin with the following theorem, which is proved in [25, Theorem 1.1].

Theorem 1.3 *Fix $\vartheta \in (0,1)$, $r > 0$, and $a > 1$. Then for any correlation matrix Ω,*

$$\frac{\mathrm{Hamm}_p^*(\vartheta, \kappa, r, a, \Omega)}{p^{1-\vartheta}} \gtrsim \begin{cases} 1, & 0 < r < \vartheta, \\ L_p p^{-(r-\vartheta)^2/(4r)}, & r > \vartheta. \end{cases}$$

Note that $p^{1-\vartheta}$ is approximately the number of signals. Therefore, when $r < \vartheta$, the number of selection errors can not get substantially smaller than the number of signals. This is the most difficult case where no variable selection method can be successful.

In this section, we focus on the case $r > \vartheta$, so that successful variable selection is possible. In this case, Theorem 1.3 says that a *universal* lower bound for the Hamming distance is

$$L_p p^{1-(\vartheta+r)^2/(4r)}.$$

An interesting question is, to what extend, this lower bound is tight.

Recall that $\lambda_k^*(\Omega)$ denotes the minimum of smallest eigenvalues across all $k \times k$ principle submatrices of Ω, as defined in (1.15). The following corollaries are proved in Section 5.

Corollary 1.2 *Fix $\vartheta \in (0,1)$ and $r > 0$ such that $1 < r/\vartheta < 3 + 2\sqrt{2} \approx 5.828$. Suppose that in addition to the conditions of Theorem 1.2, $|\Omega(i,j)| \le 4\sqrt{2} - 5 \approx 0.6569$, for all $1 \le i, j \le p$, $i \ne j$, then $\mathrm{Hamm}_p^*(\vartheta, \kappa, r, a, \Omega) = L_p p^{1-(\vartheta+r)^2/(4r)}$.*

Corollary 1.3 *Fix $\vartheta \in (0,1)$ and $r > 0$ such that $1 < r/\vartheta < 5 + 2\sqrt{6} \approx 9.898$. Suppose that in addition to the conditions of Theorem 1.2, $\lambda_3^* \geq 2(5 - 2\sqrt{6}) \approx 0.2021$, $\lambda_4^* \geq 5 - 2\sqrt{6} \approx 0.1011$, and for all $1 \leq i, j \leq p$, $i \neq j$, $|\Omega(i,j)| \leq 8\sqrt{6} - 19 \approx 0.5959$, then $\mathrm{Hamm}_p^*(\vartheta, \kappa, r, a, \Omega) = L_p p^{1-(\vartheta+r)^2/(4r)}$.*

The conditions in these corollaries are rather relaxed. Somewhat surprisingly, the off-diagonals of Ω do not necessarily have a major influence on the optimal rate convergence, as one might have expected.

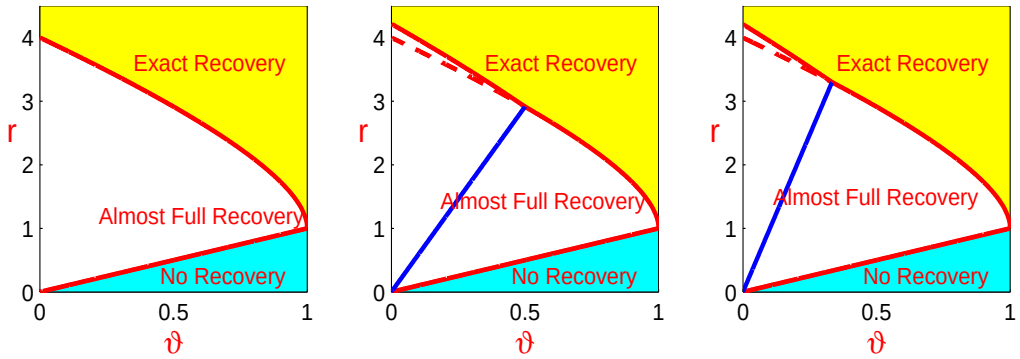

Figure 1: Phase diagram for $\Omega = I_p$ (left), for Ω satisfying conditions of Corollary 1.2 (middle), and for Ω satisfying conditions of Corollary 1.3 (right). Red line: $r = \vartheta$. Solid red curve: $r = \rho(\vartheta, \Omega)$. In each of the last two panels, the blue line intersects with the red curve at $(\vartheta, r) = (1/2, [3 + 2\sqrt{2}]/2)$ (middle) and $(\vartheta, r) = (1/3, [5 + 2\sqrt{6}]/3)$ (right), which splits the red solid curve into two parts; the part to the left is illustrative for it depends on Ω in a complicated way; the part to the right, together with the dashed red curve, represent $r = (1 + \sqrt{1 - \vartheta})^2$ (in the left panel, this is illustrated by the red curve).

Together, Theorem 1.3 and Corollaries 1.2-1.3 have an interesting implication on the so-called *phase diagram*. Call the two-dimensional *parameter space* $\{(\vartheta, r) : 0 < \vartheta < 1, r > 0\}$ the phase space. There are two curves $r = \vartheta$ and $r = \rho(\vartheta, \Omega)$ (the latter can be thought of as the solution of $\sum_{j=1}^p p^{-\rho_j^*(\vartheta, r, a, \Omega)} = 1$; recall that $\rho_j^*(\vartheta, r, a, \Omega)$ does not depend on a) that partition the whole phase space into three different regions:

- *Region of No Recovery.* $\{(\vartheta, r) : 0 < r < \vartheta, 0 < \vartheta < 1\}$. In this region, as $p \to \infty$, for any Ω and any procedures, the minimax Hamming error equals approximately to the total expected number of signals. This is the most difficult region, in which no procedure can be successful in the minimax sense.

- *Region of Almost Full Recovery.* $\{(\vartheta, r) : \vartheta < r < \rho(\vartheta, \Omega)\}$. In this region, as $p \to \infty$, the minimax Hamming distance satisfies $1 \ll \mathrm{Hamm}_p^*(\vartheta, \kappa, r, a, \Omega) \ll p^{1-\vartheta}$, and it is possible to recover most of the signals, but it is impossible to recover all of them.

- *Region of Exact Recovery.* In this region, as $p \to \infty$, the minimax Hamming distance $\mathrm{Hamm}_p^*(\vartheta, \kappa, r, a, \Omega) = o(1)$, and it is possible to exactly recover all signals with overwhelming probability.

In general, the function $\rho(\vartheta, \Omega)$ depends on Ω in a complicated way. However, by Theorem 1.3 and Corollaries 1.2-1.3, we have the following conclusions. First, for all Ω and $a > 1$,

$$\rho(\vartheta, \Omega) \geq (1 + \sqrt{1 - \vartheta})^2, \qquad \text{for all } 0 < \vartheta < 1.$$

Second, in the simplest case where $\Omega = I_p$, $\mathrm{Hamm}_p^*(\vartheta, \kappa, r, a, \Omega) = L_p p^{1-(\vartheta+r)^2/(4r)}$, and

$$\rho(\vartheta, \Omega) = (1 + \sqrt{1-\vartheta})^2, \qquad \text{for all } 0 < \vartheta < 1.$$

Third, under the conditions of Corollary 1.2,

$$\rho(\vartheta, \Omega) = (1 + \sqrt{1-\vartheta})^2, \qquad \text{if } 1/2 < \vartheta < 1,$$

Last, under the conditions of Corollary 1.3,

$$\rho(\vartheta, \Omega) = (1 + \sqrt{1-\vartheta})^2, \qquad \text{if } 1/3 < \vartheta < 1.$$

The phase diagram for the last three cases are illustrated in Figure 1. The blue lines are $r/\vartheta = 3 + 2\sqrt{2}$ (middle) and $r/\vartheta = 5 + 2\sqrt{6}$ (right).

Corollaries 1.2-1.3 can be extended to more general situations, where r/ϑ may get arbitrary large, but consequently, we need stronger conditions on Ω. Towards this end, we note that for any (ϑ, r) such that $r > \vartheta$, we can find a unique integer $N = N(\vartheta, r)$ such that

$$2N - 1 \le (\vartheta/r + r/\vartheta)/2 < 2N + 1.$$

Suppose that for any $2 \le k \le 2N-1$,

$$\lambda_k^*(\Omega) \ge \max_{\{(k+1)/2 \le j \le \min\{k, N\}\}} \Big\{ \frac{(r/\vartheta + \vartheta/r)/2 - 2j + 2 + \sqrt{[(r/\vartheta + \vartheta/r)/2 - 2j + 2]^2 - 1}}{(2k - 2j + 1)(r/\vartheta)} \Big\}, \tag{1.28}$$

and that for any $2 \le k \le 2N$,

$$\lambda_k^*(\Omega) \ge \max_{\{k/2 \le j \le \min\{k-1, N\}\}} \Big\{ \frac{(r/\vartheta + \vartheta/r)/2 + 1 - 2j}{(k-j)(r/\vartheta)} \Big\}. \tag{1.29}$$

Then we have the following corollary.

Corollary 1.4 *Fix $\vartheta \in (0,1)$ and $r > 0$ such that $r > \vartheta$. Suppose (1.28)-(1.29) hold, additional to the conditions of Theorem 1.2, then* $\mathrm{Hamm}_p^*(\vartheta, \kappa, r, a, \Omega) = L_p p^{1-(\vartheta+r)^2/(4r)}$.

Corollary 1.4 implies a similar partition of the phase diagram as do Corollaries 1.2-1.3, say, that (1.28)-(1.29) hold for all (ϑ, r) satisfying $r/\vartheta \le s_0$, for some fixed constant $s_0 > 0$.

1.9 Non-optimaility of subset selection and the lasso

Subset selection (also called the L^0-penalization method) is a well-known method for variable selection, which selects variables by minimizing the following functional:

$$\frac{1}{2}\|Y - X\beta\|^2 + \frac{1}{2}(\lambda_{ss})^2\|\beta\|_0, \tag{1.30}$$

where $\|\beta\|_q$ denotes the L^q-norm, $q \ge 0$, and $\lambda_{ss} > 0$ is a tuning parameter. The AIC, BIC, and RIC are methods of this type [1, 31, 16]. Subset selection is believed to have good "theoretic property", but the main drawback of this method is that it is computationally NP hard. To overcome the computational challenge, many *relaxation* methods are proposed, including but are not limited to the lasso [7, 32], SCAD [14], MC+ [36], and Dantzig selector [6]. Take the lasso for example. The method selects variables by minimizing the following functional:

$$\frac{1}{2}\|Y - X\beta\|^2 + \lambda_{lasso}\|\beta\|_1, \tag{1.31}$$

where the L^0-penalization is replaced by the L^1-penalization, so the functional is convex and the optimization problem is solvable in polynomial time under proper conditions.

Somewhat surprisingly, subset selection is generally *rate non-optimal* in terms of selection errors. This sub-optimality of subset selection is due to its lack of flexibility in adapting to the "local" graphic structure of the design variables. Similarly, other global relaxation methods are sub-optimal as well, as the subset selection is the "idol" these methods try to mimic. To save space, we only discuss subset selection and the lasso, but a similar conclusion can be drawn for SCAD, MC+, and Dantzig selector.

For mathematical simplicity, we illustrate the point with an idealized regression model where the Gram matrix $G = X'X$ is diagonal block-wise and has the following form

$$G(i,j) = 1\{i=j\} + h_0 \cdot 1\{|j-i|=1,\, max(i,j) \text{ is even}\}, \quad |h_0|<1,\ 1\le i,j\le p. \quad (1.32)$$

Using an idealized model is mostly for technical convenience, but the non-optimality of subset selection or the lasso holds much more broadly than what is considered here. Since our goal is to show such methods are non-optimal, using a simple model is sufficient: if a procedure is non-optimal in an idealized case, we can not expect it to be optimal in a more general context.

At the same time, we continue to model β with the Asymptotic Rare and Weak model $\mathrm{ARW}(\vartheta, r, a, \mu)$, but where we relax the assumption of $\mu \in \Theta_p^*(\tau_p, a)$ to that of $\mu \in \Theta_p(\tau_p)$ so that the strength of each signal $\ge \tau_p$ (but there is no upper bound on the strength). Consider a variable selection procedure $\hat{\beta}^\star$, where $\star = gs, ss, lasso$, representing Graphlet Screening, subset selection, and the lasso (where the tuning parameters for each method are ideally set; for the worst-case risk considered below, the ideal tuning parameters depend on (ϑ, r, p, h_0) but do not depend on μ). For some exponents $\rho_\star = \rho_\star(\vartheta, r, h_0)$ that does not depend on p, it is seen that for large p, the worst-case Hamming selection error of $\hat{\beta}^\star$ has the form of

$$\sup_{\{\mu\in\Theta_p(\tau_p)\}} H_p(\hat{\beta}^\star; \epsilon_p, \mu, G) = L_p p^{1-\rho_\star(\vartheta,r,h_0)}.$$

Here, H_p is slightly different from that in Section 1.5 since the settings are slightly different.

We now study $\rho_\star(\vartheta, r, h_0)$. Towards this end, we first introduce

$$\rho^{(3)}_{lasso}(\vartheta,r,h_0) = \big\{(2|h_0|)^{-1}[(1-h_0^2)\sqrt{r} - \sqrt{(1-h_0^2)(1-|h_0|)^2 r - 4|h_0|(1-|h_0|)\vartheta}]\big\}^2,$$

and

$$\rho^{(4)}_{lasso}(\vartheta,r,h_0) = \vartheta + \frac{(1-|h_0|)^3(1+|h_0|)}{16h_0^2}\big[(1+|h_0|)\sqrt{r} - \sqrt{(1-|h_0|)^2 r - 4|h_0|\vartheta/(1-h_0^2)}\big]^2.$$

We then let

$$\rho^{(1)}_{ss}(\vartheta,r,h_0) = \begin{cases} 2\vartheta, & r/\vartheta \le 2/(1-h_0^2) \\ [2\vartheta + (1-h_0^2)r]^2/[4(1-h_0^2)r], & r/\vartheta > 2/(1-h_0^2) \end{cases},$$

$$\rho^{(2)}_{ss}(\vartheta,r,h_0) = \begin{cases} 2\vartheta, & r/\vartheta \le 2/(1-|h_0|) \\ 2[\sqrt{2(1-|h_0|)r} - \sqrt{(1-|h_0|)r - \vartheta}]^2, & r/\vartheta > 2/(1-|h_0|) \end{cases},$$

$$\rho^{(1)}_{lasso}(\vartheta,r,h_0) = \begin{cases} 2\vartheta, & r/\vartheta \le 2/(1-|h_0|)^2 \\ \rho^{(3)}_{lasso}(\vartheta,r,h_0), & r/\vartheta > 2/(1-|h_0|)^2 \end{cases},$$

and

$$\rho^{(2)}_{lasso}(\vartheta, r, h_0) = \begin{cases} 2\vartheta, & r/\vartheta \le (1+|h_0|)/(1-|h_0|)^3 \\ \rho^{(4)}_{lasso}(\vartheta, r, h_0), & r/\vartheta > (1+|h_0|)/(1-|h_0|)^3 \end{cases}.$$

The following theorem is proved in Section 5.

Theorem 1.4 *Fix $\vartheta \in (0,1)$ and $r > 0$ such that $r > \vartheta$. If G satisfies (1.32), then*

$$\rho_{gs}(\vartheta, r, h_0) = \min\Big\{\frac{(\vartheta+r)^2}{4r},\ \vartheta + \frac{(1-|h_0|)}{2}r,\ 2\vartheta + \frac{\{[(1-h_0^2)r - \vartheta]_+\}^2}{4(1-h_0^2)r}\Big\}, \tag{1.33}$$

$$\rho_{ss}(\vartheta, r, h_0) = \min\Big\{\frac{(\vartheta+r)^2}{4r},\ \vartheta + \frac{(1-|h_0|)}{2}r,\ \rho^{(1)}_{ss}(\vartheta, r, h_0), \rho^{(2)}_{ss}(\vartheta, r, h_0)\Big\}, \tag{1.34}$$

and

$$\rho_{lasso}(\vartheta, r, h_0) = \min\Big\{\frac{(\vartheta+r)^2}{4r},\ \vartheta + \frac{(1-|h_0|)r}{2(1+\sqrt{1-h_0^2})},\ \rho^{(1)}_{lasso}(\vartheta, r, h_0),\ \rho^{(2)}_{lasso}(\vartheta, r, h_0)\Big\}. \tag{1.35}$$

It can be shown that $\rho_{gs}(\vartheta, r, h_0) \ge \rho_{ss}(\vartheta, r, h_0) \ge \rho_{lasso}(\vartheta, r, h_0)$, where depending on the choices of (ϑ, r, h_0), we may have equality or strict inequality (note that a larger exponent means a better error rate). This fits well with our expectation, where as far as the convergence rate is concerned, Graphlet Screening is optimal for all (ϑ, r, h_0), so it beats the subset selection, which in turn beats the lasso. Table 2 summarizes the exponents for some representative (ϑ, r, h_0). It is seen that differences between these exponents become increasingly prominent when h_0 increase and ϑ decrease.

$\vartheta/r/h_0$	.1/11/.8	.3/9/.8	.5/4/.8	.1/4/.4	.3/4/.4	.5/4/.4	.1/3/.2	.3/3/.2
$\star = gs$	1.1406	1.2000	0.9000	0.9907	1.1556	1.2656	0.8008	0.9075
$\star = ss$	0.8409	0.9047	0.9000	0.9093	1.1003	1.2655	0.8007	0.9075
$\star = lasso$	0.2000	0.6000	0.7500	0.4342	0.7121	1.0218	0.6021	0.8919

Table 2: The exponents $\rho_\star(\vartheta, r, h_0)$ in Theorem 1.4, where $\star = gs, ss, lasso$.

Similar to that in Section 1.8, each of these methods has a phase diagram, where the phase space partitions into three regions: *Region of Exact Recovery*, *Region of Almost Full Recovery*, and *Region of No Recovery*. Interestingly, the separating boundary for the last two regions are the same for three methods, which is the line $r = \vartheta$. The boundary that separates the first two regions, however, vary significantly for different methods. For any $h_0 \in (-1, 1)$ and $\star = gs, ss, lasso$, the equation for this boundary can be obtained by setting $\rho_\star(\vartheta, r, h_0) = 1$ (the calculations are elementary so we omit them). Note that the lower the boundary is, the better the method is, and that the boundary corresponding to the lasso is discontinuous at $\vartheta = 1/2$. Compare the phase diagrams in Figure 2.

Subset selection and the lasso are rate non-optimal for they are so-called *one-step* or *non-adaptive* methods [25], which use only one tuning parameter, and which do not adapt to the local graphic structure. The non-optimality can be best illustrated with the diagonal block-wise model presented here, where each block is a 2×2 matrix. Correspondingly, we can partition the vector β into many size 2 blocks, each of which is of the following three types (i) those have no signal, (ii) those have exactly one signal, and (iii) those have two signals. Take the subset selection for example. To best separate (i) from (ii), we need to

set the tuning parameter ideally. But such a tuning parameter may not be the "best" for separating (i) from (iii). This explains the non-optimality of subset selection.

Seemingly, more complicated penalization methods that use multiple tuning parameters may have better performance than the subset selection and the lasso. However, it remains open how to design such extensions to achieve the optimal rate for general cases. To save space, we leave the study along this line to the future.

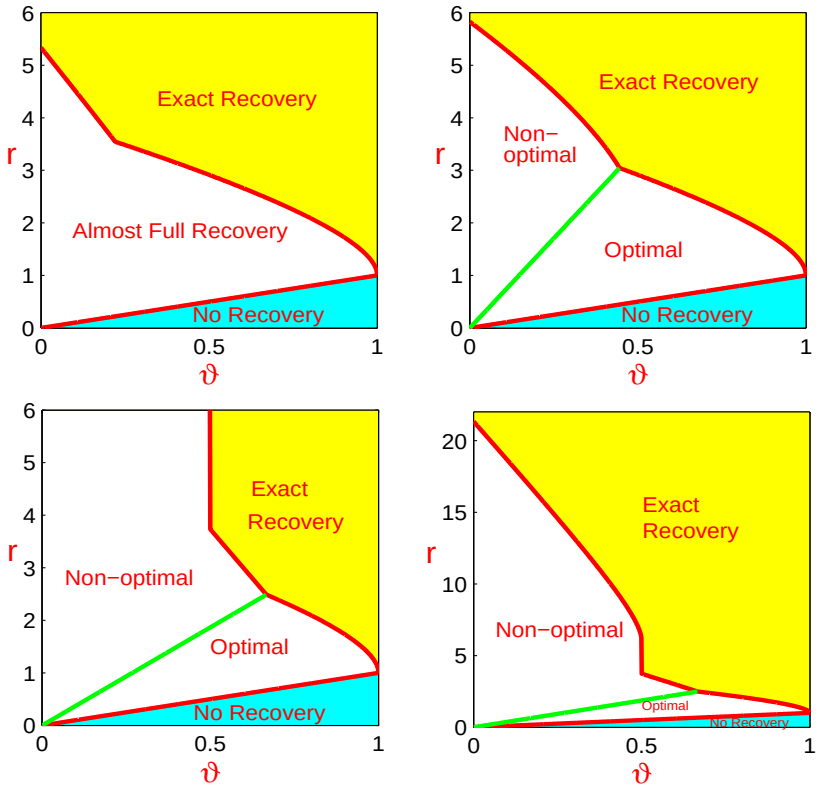

Figure 2: Phase diagrams for Graphlet Screening (top left), subset selection (top right), and the lasso (bottom; zoom-in on the left and zoom-out on the right), where $h_0 = 0.5$.

1.10 Summary

We propose Graphlet Screening as a new approach to variable selection. The key methodological innovation is to use the GOSD to guide the multivariate screening. While a brute-force m-variate screening has a computational cost of $O(p^m)$, Graphlet Screening only has a computation cost of $L_p p$, by utilizing graph sparsity.

We use *asymptotic minimaxity of the Hamming distance* as the criterion for assessing optimality. Compared with existing literature on variable selection where the *oracle property* or *probability of exact support recovery* is used to assess optimality, our approach is both mathematically more demanding and scientifically more relevant.

We have proved that Graphlet Screening achieves the optimal rate of convergence of Hamming errors, especially when signals are rare and weak, provided that the Gram matrix

is sparse. Somewhat surprisingly, the well-known methods of subset selection and the lasso are not rate optimal, even with very simple Gram matrix G and even when the tuning parameters are ideally set. The sub-optimality of these methods is due to that they do not take advantage of the 'local' graphical structure as Graphlet Screening does.

The GS methodology has three key tuning parameters: the parameter q for the threshold level $t(\hat{D}, \hat{F}) = 2\sigma^2 q \log p$ in the GS-step, and $(u^{gs}, v^{gs}) = (\sigma\sqrt{2\vartheta \log p}, \sigma\sqrt{2r \log p})$ in the GC-step. While the choice of q is reasonably flexible and a sufficiently small fixed $q > 0$ is usually adequate, the choice of u^{gs} and v^{gs} are more directly tied to the signal sparsity and signal strength. Adaptive choice of these tuning parameters is a challenging direction of further research. One of our ideas to be developed in this direction is a subsampling scheme similar to the Stability Selection [28]. On the other hand, as shown in our numeric results in Section 3, the performance of the GS is relatively insensitive to mis-specification of (ϵ_p, τ_p); see details therein.

1.11 Content

The remaining part of the paper is organized as follows. In Section 2, we introduce the so-called *Sure Screening* property and the *Separable After Screening* property of Graphlet Screening, and use these two properties to prove Theorem 1.2. Section 3 contains numeric results, Section 4 discusses more connections to existing literature and possible extensions of Graphlet Screening, and Section 5 contains technical proofs.

For a vector ξ, $\|\xi\|_q$ denotes the L^q-norm, and when $q = 2$, we drop q for simplicity. For a $p \times p$ matrix A, $\|A\|_\infty$ denotes the matrix L^∞-norm, and $\|A\|$ denotes the spectral norm [22]. Also, $\mathcal{G}^*$ is the graph of strong dependence (GOSD), $\mathcal{G}^*_S$ is its subgraph formed by all nodes in $S = S(\beta)$, and $\mathcal{G}^\diamond$ is the graph of least favorables (GOLF).

2 Properties of Graphlet Screening, proof of Theorem 1.2

Graphlet Screening attributes the success to two important properties: the Sure Screening property and the *Separable After Screening* (SAS) property.

The Sure Screening property means that in the m-stage χ^2 screening, by picking an appropriate threshold, the set $\mathcal{U}_p^*$ (which is the set of retained indices after the GS-step) contains all but a small fraction of true signals. Asymptotically, this fraction is comparably smaller than the minimax Hamming errors, and so it is negligible. The SAS property means that except for a negligible probability, as a subgraph of the GOSD, $\mathcal{U}_p^*$ decomposes into many disconnected components of the GOSD, where the size of each component does not exceed a fixed integer. Together, these two properties ensure that the original regression problem reduces to many small-size regression problems, and thus pave the way for the GC-step.

Below, we explain these ideas in detail, and conclude the section by the proof of Theorem 1.2. Since the only place we need the knowledge of σ is in setting the tuning parameters, so without loss of generality, we assume $\sigma = 1$ throughout this section.

First, we discuss the GS-step. For short, we write $\hat{\beta} = \hat{\beta}^{gs}(Y; \delta, \mathcal{Q}, u^{gs}, v^{gs}, X, p, n)$ throughout this section. We first discuss the computational cost of the GS-step. As in Theorem 1.2, we take the threshold δ in the definition of $\mathcal{G}^*$ (see (1.6)) to be $\delta = \delta_p = 1/\log(p)$. The proof of the following lemma is similar to that [25, Lemma 2.2], so we omit it.

Lemma 2.1 *As $p \to \infty$, for any $\Omega \in \mathcal{M}_p^*(\gamma, c_0, g, A)$, with probability $1 - o(1/p^2)$, $\|\Omega - \Omega^*\|_\infty \leq C(\log(p))^{-(1-\gamma)}$, and $\mathcal{G}^*$ is K-sparse, where $K \leq C(\log(p))^{1/\gamma}$.*

Combining Lemma 2.1 and [19], it follows that with probability $1-o(1/p^2)$, $\mathcal{G}^*$ has at most

$$p(Ce(\log(p))^{1/\gamma})^m$$

connected subgraphs of size $\leq m$, where we note that the second factor is at most logarithmically large. Therefore, the computational cost in the GS-step is at most $L_p p$ flops.

We now consider the performance of the GS-step. The goal of this step is two-fold: on one hand, it tries to retain as many signals as possible during the screening; on the other hand, it tries to minimize the computational cost of the GC-step by controlling the maximum size of all components of $\mathcal{U}_p^*$. The key in the GS-step is to set the collection of thresholds $\mathcal{Q}$. The tradeoff is that, setting the thresholds too high may miss too many signals during the screening, and setting the threshold too low may increase the maximum size of the components in $\mathcal{U}_p^*$, and so increase the computational burden of the GC-step. The following lemma characterizes the Sure Screening property of GS, and is proved in Section 5.

Lemma 2.2 *(Sure Screening). Fix $m \geq 1$, $A > 0$, $r > 0$, $(\vartheta, \gamma) \in (0,1)^2$ and $\kappa > 1 - \vartheta$. In the m-stage χ^2 screening of the GS-step, if we set the thresholds $t(\hat{D}, \hat{F})$ as in (1.25) and the conditions of Theorem 1.2 hold, then as $p \to \infty$, for any $\Omega \in \mathcal{M}_p^*(\gamma, c_0, g, A)$*

$$\sum_{j=1}^p P(\beta_j \neq 0, j \notin \mathcal{U}_p^*) \leq L_p[p^{1-(m+1)\vartheta} + \sum_{j=1}^p p^{-\rho_j^*}] + o(1).$$

Next, we formally state the SAS property. Viewing it as a subgraph of $\mathcal{G}^*$, $\mathcal{U}_p^*$ decomposes into many disconnected components $\mathcal{I}^{(k)}$, $1 \leq k \leq N$, where N is an integer that may depend on the data.

Lemma 2.3 *(SAS). Fix $m \geq 1$, $A > 0$, $r > 0$, $(\vartheta, \gamma) \in (0,1)^2$ and $\kappa > 1 - \vartheta$. In the m-stage χ^2 screening in the GS-step, suppose we set the thresholds $t(\hat{D}, \hat{F})$ as in (1.25) such that $q(\hat{D}, \hat{F}) \geq q_0$ for some constant $q_0 = q_0(\vartheta, r) > 0$. As $p \to \infty$, under the conditions of Theorem 1.2, for any $\Omega \in \mathcal{M}_p^*(\gamma, c_0, g, A)$, there is a constant $\ell_0 = \ell_0(\vartheta, r, \kappa, \gamma, A, c_0, g) > 0$ such that with probability at least $1 - o(1/p)$,*

$$|\mathcal{I}^{(k)}| \leq \ell_0, \qquad 1 \leq k \leq N.$$

We remark that a more convenient way of picking q is to let

$$\begin{cases} q_0 \leq q \leq (\frac{\omega r + \vartheta}{2\omega r})^2 \omega r, & |\hat{D}| \text{ is odd } \& \ \omega r/\vartheta > |\hat{D}| + (|\hat{D}|^2 - 1)^{1/2}, \\ q_0 \leq q \leq \frac{1}{4}\omega r, & |\hat{D}| \text{ is even } \& \ \omega r/\vartheta \geq 2|\hat{D}|, \end{cases} \tag{2.36}$$

and let q be any other number otherwise, with which both lemmas continue to hold with this choice of q. Here, for short, $\omega = \omega(\hat{D}, \hat{F}; \Omega)$. Note that numerically this choice is comparably more conservative.

Together, the above two lemmas say that the GS-step makes only negligible false non-discoveries, and decomposes $\mathcal{U}_p^*$ into many disconnected components, each has a size not exceeding a fixed integer. As a result, the computational cost of the following GC-step is moderate, at least in theory.

We now discuss the GC-step. The key to understanding the GC-step is that the original regression problem reduces to many disconnected small-size regression problem. To see the point, define $\tilde{Y} = X'Y$ and recall that $G = X'X$. Let $\mathcal{I}_0 \lhd \mathcal{U}_p^*$ be a component, we limit our attention to $\mathcal{I}_0$ by considering the following regression problem:

$$\tilde{Y}^{\mathcal{I}_0} = G^{\mathcal{I}_0}\beta + (X'z)^{\mathcal{I}_0}, \tag{2.37}$$

where $(X'z)^{\mathcal{I}_0} \sim N(0, G^{\mathcal{I}_0,\mathcal{I}_0}) \approx N(0, \Omega^{\mathcal{I}_0,\mathcal{I}_0})$, and $G^{\mathcal{I}_0}$ is a $|\mathcal{I}_0| \times p$ matrix according to our notation. What is non-obvious here is that, the regression problem still involves the whole vector β, and is still high-dimensional. To see the point, letting $V = \{1, 2, \ldots, p\} \setminus \mathcal{U}_p^*$, we write

$$G^{\mathcal{I}_0}\beta = G^{\mathcal{I}_0,\mathcal{I}_0}\beta^{\mathcal{I}_0} + I + II, \qquad I = \sum_{\mathcal{J}_0:\mathcal{J}_0 \lhd \mathcal{U}_p^*, \mathcal{J}_0 \neq \mathcal{I}_0} G^{\mathcal{I}_0,\mathcal{J}_0}\beta^{\mathcal{J}_0}, \qquad II = G^{\mathcal{I}_0,V}\beta^V.$$

First, by Sure Screening property, β^V contains only a negligible number of signals, so we can think II as negligible. Second, for any $\mathcal{J}_0 \neq \mathcal{I}_0$ and $\mathcal{J}_0 \lhd \mathcal{U}_p^*$, by the SAS property, $\mathcal{I}_0$ and $\mathcal{J}_0$ are disconnected and so the matrix $G^{\mathcal{I}_0,\mathcal{J}_0}$ is a small size matrix whose coordinate are uniformly small. This heuristic is made precise in the proof of Theorem 1.2. It is now seen that the regression problem in (2.37) is indeed low-dimensional:

$$\tilde{Y}^{\mathcal{I}_0} \approx G^{\mathcal{I}_0,\mathcal{I}_0}\beta^{\mathcal{I}_0} + (X'z)^{\mathcal{I}_0} \approx N(\Omega^{\mathcal{I}_0,\mathcal{I}_0}\beta^{\mathcal{I}_0}, \Omega^{\mathcal{I}_0,\mathcal{I}_0}), \tag{2.38}$$

The above argument is made precise in Lemma 2.4, see details therein. Finally, approximately, the GC-step is to minimize

$$\frac{1}{2}(\tilde{Y}^{\mathcal{I}_0} - \Omega^{\mathcal{I}_0,\mathcal{I}_0}\xi)'(\Omega^{\mathcal{I}_0,\mathcal{I}_0})^{-1}(\tilde{Y}^{\mathcal{I}_0} - \Omega^{\mathcal{I}_0,\mathcal{I}_0}\xi) + \frac{1}{2}(u^{gs})^2\|\xi\|_0,$$

where each coordinate of ξ is either 0 or $\geq v^{gs}$ in magnitude. Comparing this with (2.38), the procedure is nothing but the penalized MLE of a low dimensional normal model, and the main result follows by exercising basic statistical inferences.

We remark that in the GC-step, removing the constraints on the coordinates of ξ will not give the optimal rate of convergence. This is one of the reasons why the classical subset selection procedure is rate non-optimal. Another reason why the subset selection is non-optimal is that, the procedure has only one tuning parameter, but Graphlet Screening has the flexibility of using different tuning parameters in the GS-step and the GC-step. See Section 1.9 for more discussion.

We are now ready for the proof of Theorem 1.2.

2.1 Proof of Theorem 1.2

By Lemma 2.2,

$$\sum_{j=1}^{p} P(\beta_j \neq 0, j \notin \mathcal{U}_p^*) \leq L_p[p^{1-(m+1)\vartheta} + \sum_{j=1}^{p} p^{-\rho_j^*}] + o(1). \tag{2.39}$$

So to show the claim, it is sufficient to show

$$\sum_{j=1}^{p} P(j \in \mathcal{U}_p^*, \mathrm{sgn}(\beta_j) \neq \mathrm{sgn}(\hat{\beta}_j)) \leq L_p[\sum_{j=1}^{p} p^{-\rho_j^*} + p^{1-(m+1)\vartheta}] + o(1). \tag{2.40}$$

Towards this end, let $S(\beta)$ be the support of β, Ω^* be as in (1.6), and $\mathcal{G}^*$ be the GOSD. Let $\mathcal{U}_p^*$ be the set of retained indices after the GS-step. Note that when $\mathrm{sgn}(\hat{\beta}_j) \neq 0$, there is a unique component $\mathcal{I}_0$ such that $j \in \mathcal{I}_0 \lhd \mathcal{U}_p^*$. For any connected subgraph $\mathcal{I}_0$ of $\mathcal{G}^*$, let

$$B(\mathcal{I}_0) = \{k\colon\ k \notin \mathcal{I}_0,\ \Omega^*(k,\ell) \neq 0 \text{ for some } \ell \in \mathcal{I}_0,\ 1 \leq k \leq p\}.$$

Note that when $\mathcal{I}_0$ is a component of $\mathcal{U}_p^*$, we must have $B(\mathcal{I}_0) \cap \mathcal{U}_p^* = \emptyset$ as any node in $B(\mathcal{I}_0)$ is connected to some nodes in the component $\mathcal{I}_0$. As a result,

$$P(j \in \mathcal{I}_0 \lhd \mathcal{U}_p^*, B(\mathcal{I}_0) \cap S(\beta) \neq \emptyset) \leq \sum_{\mathcal{I}_0: j \in \mathcal{I}_0} \sum_{k \in B(\mathcal{I}_0)} P(k \notin \mathcal{U}_p^*, \beta_k \neq 0), \tag{2.41}$$

where the first summation is over all connected subgraphs that contains node j. By Lemma 2.3, with probability at least $1 - o(1/p)$, $\mathcal{G}^*$ is K-sparse with $K = C(\log(p))^{1/\gamma}$, and there is a finite integer ℓ_0 such that $|\mathcal{I}_0| \leq \ell_0$. As a result, there are at most finite $\mathcal{I}_0$ such that the event $\{j \in \mathcal{I}_0 \lhd \mathcal{U}_p^*\}$ is non-empty, and for each of such $\mathcal{I}_0$, $B(\mathcal{I}_0)$ contains at most L_p nodes. Using (2.41) and Lemma 2.2, a direct result is

$$\sum_{j=1}^{p} P(j \in \mathcal{I}_0 \lhd \mathcal{U}_p^*, B(\mathcal{I}_0) \cap S(\beta) \neq \emptyset) \leq L_p[\sum_{j=1}^{p} p^{-\rho_j^*} + p^{1-(m+1)\vartheta}] + o(1). \tag{2.42}$$

Comparing (2.42) with (2.40), to show the claim, it is sufficient to show that

$$\sum_{j=1}^{p} P(\mathrm{sgn}(\beta_j) \neq \mathrm{sgn}(\hat{\beta}_j), j \in \mathcal{I}_0 \lhd \mathcal{U}_p^*, B(\mathcal{I}_0) \cap S(\beta) = \emptyset) \leq L_p[\sum_{j=1}^{p} p^{-\rho_j^*} + p^{1-(m+1)\vartheta}] + o(1). \tag{2.43}$$

Fix $1 \leq j \leq p$ and a connected subgraph $\mathcal{I}_0$ such that $j \in \mathcal{I}_0$. For short, let S be the support of $\beta^{\mathcal{I}_0}$ and $\hat{S}$ be the support of $\hat{\beta}^{\mathcal{I}_0}$. The event $\{\mathrm{sgn}(\beta_j) \neq \mathrm{sgn}(\hat{\beta}_j), j \in \mathcal{I}_0 \lhd \mathcal{U}_p^*\}$ is identical to the event of $\{\mathrm{sgn}(\beta_j) \neq \mathrm{sgn}(\hat{\beta}_j), j \in S \cup \hat{S}\}$. Moreover, Since $\mathcal{I}_0$ has a finite size, both S and $\hat{S}$ have finite possibilities. So to show (2.43), it is sufficient to show that for any fixed $1 \leq j \leq p$, connected subgraph $\mathcal{I}_0$, and subsets $S_0, S_1 \subset \mathcal{I}_0$ such that $j \in S_0 \cup S_1$,

$$P(\mathrm{sgn}(\beta_j) \neq \mathrm{sgn}(\hat{\beta}_j), S = S_0, \hat{S} = S_1, j \in \mathcal{I}_0 \lhd \mathcal{U}_p^*, B(\mathcal{I}_0) \cap S(\beta) = \emptyset) \leq L_p[p^{-\rho_j^*} + p^{-(m+1)\vartheta}]. \tag{2.44}$$

We now show (2.44). The following lemma is proved in [25, A.4].

Lemma 2.4 *Over the event* $\{j \in \mathcal{I}_0 \lhd \mathcal{U}_p^*\} \cap \{B(\mathcal{I}_0) \cap S(\beta) = \emptyset\}$, $\|(\Omega\beta)^{\mathcal{I}_0} - \Omega^{\mathcal{I}_0,\mathcal{I}_0}\beta^{\mathcal{I}_0}\|_\infty \leq C\tau_p(\log(p))^{-(1-\gamma)}$.

Write for short $\hat{M} = G^{\mathcal{I}_0,\mathcal{I}_0}$ and $M = \Omega^{\mathcal{I}_0,\mathcal{I}_0}$. By definitions, $\hat{\beta}^{\mathcal{I}_0}$ is the minimizer of the following functional

$$Q(\xi) \equiv \frac{1}{2}(\tilde{Y}^{\mathcal{I}_0} - \hat{M}\xi)'\hat{M}^{-1}(\tilde{Y}^{\mathcal{I}_0} - \hat{M}\xi) + \frac{1}{2}(u^{gs})^2\|\xi\|_0,$$

where ξ is an $|\mathcal{I}_0| \times 1$ vector whose coordinates are either 0 or $\geq v^{gs}$ in magnitude, $u^{gs} = \sqrt{2\vartheta\log(p)}$, and $v^{gs} = \sqrt{2r\log(p)}$. In particular,

$$Q(\beta^{\mathcal{I}_0}) \geq Q(\hat{\beta}^{\mathcal{I}_0}),$$

or equivalently

$$(\hat{\beta}^{\mathcal{I}_0} - \beta^{\mathcal{I}_0})'(\tilde{Y}^{\mathcal{I}_0} - \hat{M}\beta^{\mathcal{I}_0}) \geq \frac{1}{2}(\hat{\beta}^{\mathcal{I}_0} - \beta^{\mathcal{I}_0})'\hat{M}(\hat{\beta}^{\mathcal{I}_0} - \beta^{\mathcal{I}_0}) + (|S_1| - |S_0|)\vartheta\log(p). \tag{2.45}$$

Now, write for short $\delta = \tau_p^{-2}(\hat{\beta}^{\mathcal{I}_0} - \beta^{\mathcal{I}_0})'M(\hat{\beta}^{\mathcal{I}_0} - \beta^{\mathcal{I}_0})$. First, by Schwartz inequality, $[(\hat{\beta}^{\mathcal{I}_0} - \beta^{\mathcal{I}_0})'(\tilde{Y}^{\mathcal{I}_0} - \hat{M}\beta^{\mathcal{I}_0})]^2 \leq \delta\tau_p^2(\tilde{Y}^{\mathcal{I}_0} - \hat{M}\beta^{\mathcal{I}_0})'M^{-1}(\tilde{Y}^{\mathcal{I}_0} - \hat{M}\beta^{\mathcal{I}_0})$. Second, by Lemma 2.4, $\tilde{Y}^{\mathcal{I}_0} = w + M\beta^{\mathcal{I}_0} + rem$, where $w \sim N(0, M)$ and with probability $1 - o(1/p)$, $|rem| \leq C(\log(p))^{-(1-\gamma)}\tau_p$. Last, with probability at least $(1 - o(1/p))$, $|\hat{M} - M|_\infty \leq C\sqrt{\log(p)}p^{-[\kappa-(1-\vartheta)]/2}$. Inserting these into (2.45) gives that with probability at least $(1 - o(1/p))$

$$w'M^{-1}w \geq \frac{1}{4}\Big[\big(\sqrt{\delta r} + \frac{(|S_1| - |S_0|)\vartheta}{\sqrt{\delta r}}\big)_+\Big]^2(2\log(p)) + O((\log(p))^\gamma).$$

Since $\gamma < 1$, $O((\log(p))^\gamma)$ is negligible. We note that $w'M^{-1}w \sim \chi^2_{|\mathcal{I}_0|}(0)$. Inserting this back to (2.44), the left hand side

$$\leq \epsilon_p^{|S_0|}P(\chi^2_{|\mathcal{I}_0|}(0) \geq [(\sqrt{\delta r} + (|S_1| - |S_0|)\vartheta/\sqrt{\delta r})_+]^2(\log(p)/2)) + o(1/p).$$

Assume $\mathrm{sgn}(\beta_j) \neq \mathrm{sgn}(\hat{\beta}_j)$, and fix all parameters except δ, S_0 and S_1. By arguments similar to the proof of Lemma 1.4, the above quantity cannot achieve its maximum in the cases where $S_0 = S_1$. Hence we only need to consider the cases where $S_0 \neq S_1$. We also only need to consider the cases where $\max(|S_0|, |S_1|) \leq m$, since the sum of the probabilities of other cases is well controlled by $p^{1-(m+1)\vartheta}$. The claim follows by the definitions of ρ_j^*. □

3 Simulations

We conducted a small-scale simulation study to investigate the numerical performance of Graphlet Screening and compare it with the lasso. The subset selection is not included for comparison since it is computationally NP hard. We consider experiments for both random design and fixed design, where as before, the parameters (ϵ_p, τ_p) are tied to (ϑ, r) by $\epsilon_p = p^{-\vartheta}$ and $\tau_p = \sqrt{2r\log(p)}$ (we assume $\sigma = 1$ for simplicity in this section). The experiments with random design contain the following steps.

1. Fix $(p, \vartheta, r, \mu, \Omega)$ such that $\mu \in \Theta_p(\tau_p)$. Generate a vector $b = (b_1, b_2, \ldots, b_p)'$ such that $b_i \overset{iid}{\sim} \mathrm{Bernoulli}(\epsilon_p)$, and set $\beta = b \circ \mu$.
2. Fix κ and let $n = n_p = p^\kappa$. Generate an $n \times p$ matrix with iid rows from $N(0, (1/n)\Omega)$.
3. Generate $Y \sim N(X\beta, I_p)$, and apply Graphlet Screening and the lasso.
4. Repeat 1-3 independently, and record the average Hamming distances.

The steps for fixed design experiments are similar, except for that $n_p = p$ and $X = \Omega^{1/2}$.

Graphlet Screening uses tuning parameters $(m, \mathcal{Q}, u^{gs}, v^{gs})$. We set $m = 3$ for our experiments, which is usually large enough due to signal sparsity. The choice of $\mathcal{Q}$ is not critical, as long as the corresponding parameter q satisfies (1.26). Numerical studies below (e.g. Experiment 4a) support this point. In principle, the optimal choices of $(\mathcal{Q}, u^{gs}, v^{gs})$ depend on the unknown parameters (ϵ_p, τ_p), and how to estimate them in general settings is a lasting open problem (even for linear models with orthogonal designs). Fortunately, our studies (e.g. Experiment 4b-4d) show that mis-specifying parameters (ϵ_p, τ_p) by a reasonable amount does not significantly affect the performance of the procedure. For this reason, in most experiments below, we set the tuning parameters in a way by assuming

(ϵ_p, τ_p) as known. To be fair in comparison, we also set the tuning parameters of the lasso ideally assuming (ϵ_p, τ_p) as known. We use *glmnet* package [18] to perform lasso.

The simulations contain 4 different experiments which we now describe separately.

Experiment 1. In this experiment, we investigate how different choices of signal vector β affect the comparisons of two methods. We use a random design model, and Ω is a symmetric tri-diagonal correlation matrix where the vector on each sub-diagonal consists of blocks of $(.4, .4, -.4)'$. Fix $(p, \kappa) = (0.5 \times 10^4, 0.975)$ (note $n = p^\kappa \approx 4,000$). We let $\epsilon_p = p^{-\vartheta}$ with $\vartheta \in \{0.35, 0.5\}$ and let $\tau_p \in \{6, 8, 10\}$. For each combination of (ϵ_p, τ_p), we consider two choices of μ. For the first choice, we let μ be the vector where all coordinates equal to τ_p (note β is still sparse). For the second one, we let μ be the vector where the signs of $\mu_i = \pm 1$ with equal probabilities, and $|\mu_i| \stackrel{iid}{\sim} 0.8\nu_{\tau_p} + 0.2h$, where ν_{τ_p} is the point mass at τ_p and $h(x)$ is the density of $\tau_p(1+V/6)$ with $V \sim \chi^2_1$. For Graphlet Screening, the tuning parameters (m, u^{gs}, v^{gs}) are set as $(3, \sqrt{2\log(1/\epsilon_p)}, \tau_p)$, and the tuning parameter q in $\mathcal{Q}$ are set as maximal possible value satisfying (1.26). The average Hamming errors for both procedures across 40 repetitions are tabulated in Table 3.

τ_p		6		8		10	
Signal Strength		Equal	Unequal	Equal	Unequal	Equal	Unequal
$\vartheta = 0.35$	Graphic Screening	0.0810	0.0825	0.0018	0.0034	0	0.0003
	lasso	0.2424	0.2535	0.1445	0.1556	0.0941	0.1109
$\vartheta = 0.5$	Graphic Screening	0.0315	0.0297	0.0007	0.0007	0	0
	lasso	0.1107	0.1130	0.0320	0.0254	0.0064	0.0115

Table 3: Ratios between the average Hamming errors and $p\epsilon_p$ (Experiment 1), where "Equal" and "Unequal" stand for the first and the second choices of μ, respectively.

Experiment 2. In this experiment, we generate β the same way as in the second choice of Experiment 1, and investigate how different choices of design matrices affect the performance of the two methods. Setting $(p, \vartheta, \kappa) = (0.5 \times 10^4, 0.35, 0.975)$ and $\tau_p \in \{6, 7, 8, 9, 10, 11, 12\}$, we use Gaussian random design model for the study. For each method, the tuning parameters are set in the same way as in Experiment 1. The experiment contains 3 sub-experiments 2*a*-2*c*.

In Experiment 2*a*, we set Ω as the symmetric diagonal block-wise matrix, where each block is a 2×2 matrix, with 1 on the diagonals, and ± 0.5 on the off-diagonals (the signs alternate across different blocks). The average Hamming errors of 40 repetitions are reported in Figure 3.

In Experiment 2*b*, we set Ω as a symmetric penta-diagonal correlation matrix, where the main diagonal are ones, the first sub-diagonal consists of blocks of $(.4, .4, -.4)'$, and the second sub-diagonal consists of blocks of $(.05, -.05)'$. The average Hamming errors across 40 repetitions are reported in Figure 3.

In Experiment 2*c*, we generate Ω as follows. First, we generate Ω using the function *sprandsym(p,K/p)* in *matlab*. We then set the diagonals of Ω to be zero, and remove some of entries so that Ω is K-sparse for a pre-specified K. We then normalize each non-zero entry by the sum of the absolute values in that row or that column, whichever is larger, and multiply each entry by a pre-specified positive constant A. Last, we set the diagonal elements to be 1. We choose $K = 3$ and $A = 0.7$, draw 5 different Ω with this method, and for each of them we repeat the simulation 10 times independently. The average Hamming errors are reported in Figure 3.

The results suggest that Graphlet Screening is consistently better than the lasso.

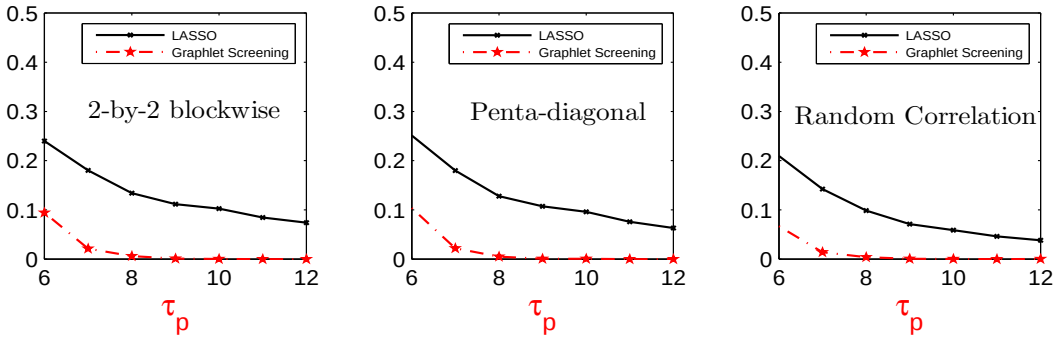

Figure 3: x-axis: τ_p. y-axis: ratios between the average Hamming errors and $p\epsilon_p$ (from left to right: Experiment 2a, 2b, and 2c).

Experiment 3. In this experiment, we investigate what are the minimum signal strength levels τ_p required by Graphlet Screening and the lasso to yield exact recovery, respectively. Fixing $p = 10^4$, we let $\epsilon_p = p^{-\vartheta}$ for $\vartheta = 0.25, 0.45, 0.65$, and let $\tau_p \in \{5, 6, 7, 8, 9, 10, 11, 12\}$. We use a fixed design model where Ω is the block-wise matrix as in Experiment 2a. For each pair of (ϵ_p, τ_p), we generate β as in the second choice of Experiment 1. The tuning parameters for Graphlet Screening and the lasso are set in the same way as in Experiment 1. The average Hamming errors across 20 repetitions are tabulated in Table.4.

Suppose we say a method yields 'exact recovery' if the average Hamming error ≤ 3. Then the minimum τ_p for Graphlet Screening to yield exact recovery is $\tau_p \approx 9$, but that for the lasso is much larger (≥ 12). For larger ϑ, the differences are less prominent, but the minimum τ_p for Graphlet Screening to yield exact recovery is consistently smaller than that of the lasso.

	τ_p	5	6	7	8	9	10	11	12
$\vartheta = 0.25$	Graphic Screening	58	21.4	9.2	3.5	3	1.8	1.0	0.6
	lasso	75.2	34.4	21.6	15	14.3	12.6	10.1	8.9
$\vartheta = 0.45$	Graphic Screening	11	3.7	0.7	0.2	0.1	0	0	0
	lasso	13.9	5.2	1.3	0.6	0.1	0.4	0.2	0.2
$\vartheta = 0.65$	Graphic Screening	3.4	0.8	0.1	0.1	0	0	0	0
	lasso	3.7	1	0.3	0.1	0	0	0	0

Table 4: Comparison of average Hamming errors (Experiment 3).

Experiment 4. In this experiment, we investigate how sensitive Graphlet Screening is with respect to the tuning parameters. The experiment contains 4 sub-experiments, 4a-4d. In Experiment 4a, we investigate how sensitive the procedure is with respect to the tuning parameter q in $\mathcal{Q}$ (recall that the main results hold as long as q fall into the range given in (1.26)), where we assume (ϵ_p, τ_p) are known. In Experiment 4b-4d, we mis-specify (ϵ_p, τ_p) by a reasonably small amount, and investigate how the mis-specification affect the performance of the procedure. For the whole experiment, we choose β the same as in the second choice of Experiment 1, and Ω the same as in Experiment 2b. We use a fixed design model in Experiment 4a-4c, and a random design model in Experiment 4d. For each sub-experiment,

the results are based on 40 independent repetitions. We now describe the sub-experiments with details.

In Experiment 4*a*, we choose $\vartheta \in \{0.35, 0.6\}$ and $r \in \{1.5, 3\}$. In Graphlet Screening, let $q_{max} = q_{max}(\hat{D}, \hat{F})$ be the maximum value of q satisfying (1.26). For each combination of (ϑ, r) and $(\hat{D}, \hat{F})$, we choose $q(\hat{D}, \hat{F}) = q_{max}(\hat{D}, \hat{F}) \times \{0.7, 0.8, 0.9, 1, 1.1, 1.2\}$ for our experiment. The results are tabulated in Table 5, which suggest that different choices of q have little influence over the variable selection errors. We must note that the larger we set $q(\hat{D}, \hat{F})$, the faster the algorithm.

$q(\hat{F}, \hat{D})/q_{max}(\hat{F}, \hat{D})$	0.7	0.8	0.9	1	1.1	1.2
$(\vartheta, r) = (0.35, 1.5)$	0.0782	0.0707	0.0661	0.0675	0.0684	0.0702
$(\vartheta, r) = (0.35, 3)$	0.0066	0.0049	0.0036	0.0034	0.0033	0.0032
$(\vartheta, r) = (0.6, 1.5)$	0.1417	0.1417	0.1417	0.1417	0.1417	0.1417
$(\vartheta, r) = (0.6, 3)$	0.0089	0.0089	0.0089	0.0089	0.0089	0.0089

Table 5: Ratios between the average Hamming errors of Graphlet Screening and $p\epsilon_p$ (Experiment 4*a*).

In Experiment 4*b*, we use the same settings as in Experiment 4*a*, but we assume ϑ (and so ϵ_p) is unknown (the parameter r is assumed as known, however), and let ϑ^* is the misspecified value of ϑ. We take $\vartheta^* \in \vartheta \times \{0.85, 0.925, 1, 1.075, 1.15, 1.225\}$ for the experiment.

In Experiment 4*c*, we use the same settings as in Experiment 4*a*, but we assume r (and so τ_p) is unknown (the parameter ϑ is assumed as known, however), and let r^* is the misspecified value of r. We take $r^* = r \times \{0.8, 0.9, 1, 1.1, 1.2, 1.3\}$ for the experiment.

In Experiment 4*b*-4*c*, we run Graphlet Screening with tuning parameters set as in Experiment 1, except ϑ or r are replaced by the misspecified counterparts ϑ^* and r^*, respectively. The results are reported in Table 6, which suggest that the misspecifications have little effect as long as r^*/r and ϑ^*/ϑ are reasonably close to 1.

ϑ^*/ϑ	0.85	0.925	1	1.075	1.15	1.225
$(\vartheta, r) = (0.35, 1.5)$	0.0799	0.0753	0.0711	0.0710	0.0715	0.0746
$(\vartheta, r) = (0.35, 3)$	0.0026	0.0023	0.0029	0.0030	0.0031	0.0028
$(\vartheta, r) = (0.6, 1.5)$	0.1468	0.1313	0.1272	0.1280	0.1247	0.1296
$(\vartheta, r) = (0.6, 3)$	0.0122	0.0122	0.0139	0.0139	0.0130	0.0147
r^*/r	0.8	0.9	1	1.1	1.2	1.3
$(\vartheta, r) = (0.35, 1.5)$	0.0843	0.0731	0.0683	0.0645	0.0656	0.0687
$(\vartheta, r) = (0.35, 3)$	0.0062	0.0039	0.0029	0.0030	0.0041	0.0054
$(\vartheta, r) = (0.6, 1.5)$	0.1542	0.1365	0.1277	0.1237	0.1229	0.1261
$(\vartheta, r) = (0.6, 3)$	0.0102	0.0076	0.0085	0.0059	0.0051	0.0076

Table 6: Ratios between of the average Hamming error of the Graphlet Screening and $p\epsilon_p$ (Experiment 4*b* (top) and Experiment 4c (bottom)).

In Experiment 4*d*, we re-examine the misspecification issue with a random design. We use the same settings as in Experiment 4*b* and Experiment 4*c*, except for (a) while we use the same Ω as in Experiment 4*b*, the design matrix X are generated according to the

random design model as in Experiment 2*b*, and (b) we only investigate for the case of $r = 2$ and $\vartheta \in \{0.35, 0.6\}$. The results are summarized in Table 7, which is consistent with the results in 4*b*-4*c*.

ϑ^*/ϑ	0.85	0.925	1	1.075	1.15	1.225
$(\vartheta, r) = (0.35, 2)$	0.1730	0.1367	0.1145	0.1118	0.0880	0.0983
$(\vartheta, r) = (0.6, 2)$	0.0583	0.0591	0.0477	0.0487	0.0446	0.0431
r^*/r	0.8	0.9	1	1.1	1.2	1.3
$(\vartheta, r) = (0.35, 2)$	0.1881	0.1192	0.1275	0.1211	0.1474	0.1920
$(\vartheta, r) = (0.6, 2)$	0.0813	0.0515	0.0536	0.0397	0.0442	0.0510

Table 7: Ratios between the average Hamming errors of Graphlet Screening and $p\epsilon_p$ (Experiment 4*d*).

4 Connection to existing literature and possible extensions

Our idea of utilizing graph sparsity is related to the graphical lasso [27, 17], which also attempts to exploit graph structure. However, the setting we consider here is different from that in [27, 17], and our emphasis on precise optimality and calibration is also very different. Our method allows nearly optimal detection of very rare and weak effects, because they are based on careful analysis that has revealed a number of subtle high-dimensional effects (e.g. phase transitions) that we properly exploit. Existing methodologies are not able to exploit or capture these phenomena, and can be shown to fail at the levels of rare weak effects where we are successful.

The paper is closely related to the recent work by Ji and Jin [25] (see also [15, 20]), and two papers use a similar rare and weak signal framework and a similar random design model. However, they are different in important ways, since the technical devise developed in [25] can not be extended to the current study. For example, the lower bound derived in this paper is different and sharper than that in [25]. Also, the procedure in [25] relies on marginal regression for screening. The limitation of marginal regression is that it neglects the graph structure of GOSD for the regularized Gram matrix (1.5), so that it is incapable of picking variables that have *weak marginal* correlation but *significant joint* correlation to Y. Correct selection of such hidden significant variables, termed as the challenge of *signal cancellation* [34], is the difficulty at the heart of the variable selection problem. One of the main innovation of Graphlet Screening is that it uses the graph structure to guide the screening, so that it is able to successfully overcome the challenge of signal cancelation.

Additionally, two papers have very different objectives, and consequently the underlying analyses are very different. The main results of each of these two papers can not be deduced from the other. For example, to assess optimality [25] uses the criterion of the partition of the phase diagram, while the current paper uses the minimax Hamming distance. Given the complexity of the high dimensional variable selection, one type of optimality does not imply the other, and vice versa. Also, the main result in [25] focuses on conditions under which the optimal rate of convergence is $L_p p^{1-(\vartheta+r)^2/(4r)}$ for the *whole* phase space. While this overlaps with our Corollaries 1.2 and 1.3, we must note that [25] deals with the much more difficult cases where r/ϑ can get arbitrary large; and to ensure the success in that case, they assume very strong conditions on the design matrix and the range of the signal

strength. On the other hand, the main focus of the current paper is on optimal variable selection under conditions (of the Gram matrix G as well as the signal vector β) that are as general as possible.

While the study in this paper has been focused on the Random Design model $\mathrm{RD}(\vartheta, \kappa, \Omega)$, extensions to deterministic design models are straightforward (in fact, in Corollary 1.1, we have already stated some results on deterministic design models), and the omission of discussion on the latter is largely for technical simplicity and the sake of space. In fact, for models with deterministic designs, since the likelihood ratio test in the derivation of the lower bound matches the penalized MLE in the cleaning step of the GS, the optimality of the GS follows from the *Sure Screening and Separable After Screening* properties of the GS. The proof of these properties, and therefore the optimality of the GS, follows the same line as those for random design as long as $\max_j |\sum_i \beta_i G(i,j) I\{\Omega^*(i,j) = 0\}|/\tau_p$ is small. This last condition on G holds when $p^{1-\vartheta}\|G-\Omega\|_\infty = o(1)$ with a certain $\Omega \in \mathcal{M}_p^*(\gamma, c_0, g, A)$. Alternatively, this condition holds when $p^{1-\vartheta}\|G-\Omega\|_\infty^2 \log p = o(1)$ with $\Omega \in \mathcal{M}_p^*(\gamma, c_0, g, A)$, provided that $\mathrm{sgn}(\beta_j)$ are iid symmetric random variables as in [5].

Another interesting direction of future research is the extension of the GS methodology to more general models such as logistic regression. The extension of the lower bound in Theorem 1.1 is relatively simple since the degree of GOLF can be bounded using the true β. This indicates the optimality of the GS method in logistic and other generalized linear models as long as proper generalized likelihood ratio or Bayes tests are used in both the *GS*- and *GC*-steps.

5 Proofs

In this section, we provide all technical proofs. For simplicity, we assume $\sigma = 1$ in this section.

5.1 Proof of Lemma.1.1

When $\mathcal{G}_S^*$ contains a connected subgraph of size $\geq m+1$, it must contain a connected subgraph with size $m+1$. By [19], there are $\leq p(eK)^{m+1}$ connected subgraph of size $m+1$. Therefore, the probability that $\mathcal{G}_S^*$ has a connected subgraph of size $(m+1)$ $\leq p(eK)^{m+1}\epsilon_p^{m+1}$. Combining these gives the claim. □

5.2 Proof of Theorem 1.1

Write for short $\rho_j^* = \rho_j^*(\vartheta, r, a, \Omega)$. Without loss of generality, assume $\rho_1^* \leq \rho_2^* \leq \ldots \leq \rho_p^*$. We construct indices $i_1 < i_2 < \ldots < i_m$ as follows. (a) start with $B = \{1, 2, \ldots, p\}$ and let $i_1 = 1$, (b) updating B by removing i_1 and all nodes j that are connected to i_1 in GOLF, let i_2 be the smallest indices, (c) defining $i_3, i_4, \ldots, i_m$ by repeating (b), and terminates the process when no indices is left in B. Since each time we remove at most $d_p(\mathcal{G}^\diamond)$ nodes, it follows that

$$\sum_{j=1}^{p} p^{-\rho_j^*} \leq d_p(\mathcal{G}^\diamond) \sum_{k=1}^{m} p^{-\rho_{i_k}^*}. \tag{5.46}$$

For each $1 \leq j \leq p$, as before, let (V_{0j}^*, V_{1j}^*) be the least favorable configuration, and let $(\theta_{*j}^{(0)}, \theta_{*j}^{(1)}) = \mathrm{argmin}_{\{\theta^{(0)} \in B_{V_{0j}^*}, \theta^{(1)} \in B_{V_{1j}^*}, \mathrm{sgn}(\theta^{(0)}) \neq \mathrm{sgn}(\theta^{(1)})\}} \alpha(\theta^{(0)}, \theta^{(1)}; \Omega)$. By our notations,

it is seen that

$$\rho_j^* = \eta(V_{0j}^*, V_{1j}^*; \Omega), \qquad \alpha^*(V_{0j}^*, V_{1j}^*; \Omega) = \alpha(\theta_{*j}^{(0)}, \theta_{*j}^{(1)}; \Omega). \tag{5.47}$$

We construct a $p \times 1$ vector μ^* as follows. Fix $j \in \{i_1, \cdots, i_m\}$. For all indices in V_{0j}^*, set the constraint of μ^* on these indices to be $\theta_{*j}^{(0)}$. For any index $i \notin \cup_{k=1}^m V_{0i_k}^*$, set $\mu_i^* = \tau_p$. Since

$$\mathrm{Hamm}_p^*(\vartheta, \kappa, r, a, \Omega) \geq \inf_{\hat{\beta}} H_p(\hat{\beta}; \epsilon_p, n_p, \mu^*, \Omega) = \inf_{\hat{\beta}} \sum_{i=1}^p P(\mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j)), \tag{5.48}$$

it follows that

$$\mathrm{Hamm}_p^*(\vartheta, \kappa, r, a, \Omega) \geq \sum_{k=1}^m \sum_{j \in V_{0i_k} \cup V_{1i_k}} P(\mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j)), \tag{5.49}$$

where $\beta = b \circ \mu^*$ as in (5.48)-(5.49). Combining (5.46) and (5.49), to show the claim, we only need to show that for any $1 \leq k \leq m$ and any procedure $\hat{\beta}$,

$$\sum_{j \in V_{0i_k} \cup V_{1i_k}} P(\mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j)) \geq L_p p^{-\rho_{i_k}^*}. \tag{5.50}$$

Towards this end, we write for short $V_0 = V_{0i_k}$, $V_1 = V_{1i_k}$, $V = V_0 \cup V_1$, $\theta^{(0)} = \theta_{*i_k}^{(0)}$, and $\theta^{(1)} = \theta_{*i_k}^{(1)}$. Note that by Lemma 1.3,

$$|V| \leq (\vartheta + r)^2/(2\vartheta r). \tag{5.51}$$

Consider a test setting where under the null H_0, $\beta = \beta^{(0)} = b \circ \mu^*$ and $I_V \circ \beta^{(0)} = I_V \circ \theta^{(0)}$, and under the alternative H_1, $\beta = \beta^{(1)}$ which is constructed by keeping all coordinates of $\beta^{(0)}$ unchanged, except those coordinates in V are perturbed in a way so that $I_V \circ \beta^{(1)} = I_V \circ \theta^{(1)}$. In this construction, both $\beta^{(0)}$ and $\beta^{(1)}$ are assumed as known, but we don't know which of H_0 and H_1 is true. In the literature, it is known that $\inf_{\hat{\beta}} \sum_{j \in V} P(\mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j))$ is not smaller than the minimum sum of Type I and Type II errors associated with this testing problem.

Note that by our construction and (5.47), the right hand side is $\alpha^*(V_0, V_1; \Omega)$. At the same time, it is seen the optimal test statistic is $Z \equiv (\theta^{(1)} - \theta^{(0)})' X' (Y - X\beta^{(0)})$. It is seen that up to some negligible terms, $Z \sim N(0, \alpha^*(V_0, V_1; \Omega)\tau_p^2)$ under H_0, and $Z \sim N(\alpha^*(V_0, V_1; \Omega)\tau_p^2, \alpha^*(V_0, V_1; \Omega)\tau_p^2)$ under H_1. The optimal test is to reject H_0 when $Z \geq t[\alpha^*(V_0, V_1; \Omega)]^{1/2}\tau_p$ for some threshold t, and the minimum sum of Type I and Type II error is

$$\inf_t \big\{ \epsilon_p^{|V_0|} \bar{\Phi}(t) + \epsilon_p^{|V_1|} \Phi(t - [\alpha^*(V_0, V_1; \Omega)]^{1/2}\tau_p) \big\}, \tag{5.52}$$

Here, we have used $P(H_0) \sim \epsilon_p^{|V_0|}$ and $P(H_1) \sim \epsilon_p^{|V_1|}$, as a result of the Binomial structure in β. It follows that

$$\sum_{j \in V} P(\mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j)) \gtrsim \inf_t \big\{ \epsilon_p^{|V_0|} \bar{\Phi}(t) + \epsilon_p^{|V_1|} \Phi(t - [\alpha^*(V_0, V_1; \Omega)]^{1/2}\tau_p) \big\}.$$

Using Mills' ratio and definitions, the right hand side $\geq L_p p^{-\eta(V_0, V_1; \Omega)}$, and (5.50) follows by recalling (5.47). □

5.3 Proof of Lemma 1.3

Let $V_0 = \emptyset$ and $V_1 = \{j\}$. It is seen that $\alpha^*(V_0, V_1; \Omega) = 1$, and $\eta(V_0, V_1; \Omega) \leq (\vartheta + r)^2/(4r)$. Using this and the definitions of V_{0j}^* and V_{1j}^*, $\max\{|V_{0j}^*|, |V_{1j}^*|\}\vartheta \leq (\vartheta + r)^2/(4r)$ and the claim follows. □

5.4 Proof of Lemma 1.4

Let sets V_0 and V_1 and vectors $\theta^{(0)}$ and $\theta^{(1)}$ be as in Section 1.6, and let $V = V_0 \cup V_1$. By the definition of $\rho_j^*(\vartheta, r, a, \Omega)$,

$$\rho_j^*(\vartheta, r, a, \Omega) = \min(I, II),$$

where

$$I = \min_{\{(V_0,V_1): j \in V_1 \cup V_0, V_0 \neq V_1\}} \eta(V_0, V_1; \Omega), \qquad II = \min_{\{V_0: j \in V_0 \cup V_1, V_0 = V_1\}} \eta(V_0, V_1; \Omega).$$

So to show the claim, it is sufficient to show

$$I = \min_{\{(D,F): j \in D \cup F, D \cap F = \emptyset, D \neq \emptyset, |D \cup F| \leq g\}} \rho(D, F; \Omega), \qquad II \geq I. \tag{5.53}$$

Consider the first claim in (5.53). Write for short $F = F(V_0, V_1) = V_0 \cap V_1$ and $D = D(V_0, V_1) = V \setminus F$. By the definitions, $D \neq \emptyset$. The key is to show that when $|V_0 \cup V_1| \leq g$,

$$\alpha^*(V_0, V_1; \Omega) = \omega(D, F; \Omega). \tag{5.54}$$

Towards this end, note that by definitions, $\alpha^*(V_0, V_1; \Omega) = \alpha(\theta_*^{(0)}, \theta_*^{(1)})$, where

$$(\theta_*^{(0)}, \theta_*^{(1)}) = \mathrm{argmin}_{\{\theta^{(0)} \in B_{V_0}, \theta^{(1)} \in B_{V_1}\}} \alpha(\theta^{(0)}, \theta^{(1)}).$$

By $a > a_g^*(\Omega)$ and the way $a_g^*(\Omega)$ is defined, $(\theta_*^{(0)}, \theta_*^{(1)})$ remains as the solution of the optimization problem if we relax the conditions $\theta^{(i)} \in B_{V_i}$ to that of $\theta^{(i)} = I_{V_i} \circ \mu^{(i)}$, where $\mu^{(i)} \in \Theta_p(\tau_p)$ (so that upper bounds on the signal strengths are removed), $i = 0, 1$. As a result,

$$\alpha^*(V_0, V_1; \Omega) = \min_{\{\theta^{(i)} \in I_{V_i} \circ \mu^{(i)}, \mu^{(i)} \in \Theta_p(\tau_p), i = 0,1,\}} \alpha(\theta^{(0)}, \theta^{(1)}). \tag{5.55}$$

We now study (5.55). For short, write $\xi = \tau_p^{-1}(\theta^{(1)} - \theta^{(0)})^V$, $\Omega_{VV} = \Omega^{V,V}$, $\xi_D = \tau_p^{-1}(\theta^{(1)} - \theta^{(0)})^D$, and similarly for Ω_{DD}, Ω_{DF}, Ω_{FD}, Ω_{FF}, and ξ_F. Without loss of generality, assume the indices in D come first in V. It follows

$$\Omega_{VV} = \begin{pmatrix} \Omega_{DD} & \Omega_{DF} \\ \Omega_{FD} & \Omega_{FF} \end{pmatrix},$$

and

$$\alpha(\theta^{(0)}, \theta^{(1)}) = \xi' \Omega_{VV} \xi = \xi_D' \Omega_{DD} \xi_D + 2\xi_D' \Omega_{DF} \xi_F + \xi_F' \Omega_{FF} \xi_F. \tag{5.56}$$

By definitions, it is seen that there is no constraint on the coordinates of ξ_F, so to optimize the quadratic form in (5.54), we need to choose ξ is a way such that $\xi_F = -\Omega_{FF}^{-1} \Omega_{FD} \xi_D$, and that ξ_D minimizes

$$\xi_D'(\Omega_{DD} - \Omega_{DF}\Omega_{FF}^{-1}\Omega_{FD})\xi_D,$$

where every coordinate of $\xi_D \geq 1$ in magnitude. Combining these with (5.55) gives (5.54).

At the same time, we rewrite

$$I = \min_{\{(D,F):j\in D\cup F, D\neq\emptyset, D\cap F=\emptyset\}} \Big\{ \min_{\{(V_0,V_1):V_0\cup V_1=D\cup F, V_0\cap V_1=F\}} \eta(V_0,V_1;\Omega)\Big\}. \tag{5.57}$$

By similar arguments as in the proof of Lemma 1.3, the subsets (V_0, V_1) that achieve the minimum of $\eta(V_0, V_1; \Omega)$ must satisfy $|V_0 \cup V_1| \leq g$. Using (5.54), for any fixed D and F such that $|D \cup F| \leq g$, $D \neq \emptyset$ and $D \cap F = \emptyset$, the term in the big bracket on the right hand side is

$$\min_{\{(V_0,V_1):V_0\cup V_1=D\cup F, V_0\cap V_1=F\}} \{\frac{(2|F|+|D|)\vartheta}{2} + \frac{\big||V_1|-|V_0|\big|\vartheta}{2} + \frac{1}{4}[(\sqrt{\rho(D,F;\Omega)r} - \frac{\big||V_1|-|V_0|\big|\vartheta}{\sqrt{\rho(D,F;\Omega)r}})_+]^2\},$$

It is worth noting that for fixed D and F, the above quantity is monotone increasing with $\big||V_1| - |V_0|\big|$. When $|D|$ is even, the minimum is achieved at (V_0, V_1) with $|V_0| = |V_1|$, and when $|D|$ is odd, the minimum is achieved at (V_0, V_1) with $\big||V_1| - |V_0|\big| = 1$, and in both cases, the minimum is $\rho(D, F; \Omega)$. Inserting this to (5.57), it is seen that

$$I = \min_{\{(D,F):j\in D\cup F, D\cap F=\emptyset, D\neq\emptyset, |D\cup F|\leq g\}} \rho(D,F;\Omega), \tag{5.58}$$

which is the first claim in (5.53).

Consider the second claim of (5.53). In this case, by definitions, $V_0 = V_1$ but $\mathrm{sgn}(\theta^{(0)}) \neq \mathrm{sgn}(\theta^{(1)})$. Redefine D as the subset of V_0 where the signs of the coordinates of $\theta^{(0)}$ do not equal to those of $\theta^{(1)}$, and let $F = V \setminus D$. By definitions, it is seen that $\alpha^*(V_0, V_0; \Omega) = 4\alpha^*(F, V_0; \Omega)$, where we note $D \neq \emptyset$ and $F \neq V_0$. By the definition of $\eta(V_0, V_1; \Omega)$, it follows that $\eta(V_0, V_0; \Omega) \geq \eta(F, V_0; \Omega)$, and the claim follows. □

5.5 Proof of Corollaries 1.2, 1.3, and 1.4

Write for short $\omega = \rho(D, F; \Omega)$ and $T = r/\vartheta$. The following inequality is frequently used below, the proof of which is elementary so we omit it:

$$\omega \geq \lambda_k^*|D|, \qquad \text{where } k = |D| + |F|. \tag{5.59}$$

To show these corollaries, it is sufficient to show for all subsets D and F of $\{1, 2, \ldots, p\}$,

$$\rho(D,F;\Omega) \geq (\vartheta + r)^2/(4r), \qquad |D| \geq 1. \tag{5.60}$$

where $\rho(D, F; \Omega)$ is as in (1.23). By basic algebra, (5.60) is equivalent to

$$\begin{cases} (\omega T + 1/(\omega T) - 2)1\{\omega T \geq 1\} \geq (T + 1/T - 2(|D| + 2|F|)), & |D| \text{ is odd}, \\ \omega \geq \frac{2}{T}[(T + 1/T)/2 + 1 - (|D| + 2|F|)], & |D| \text{ is even}. \end{cases} \tag{5.61}$$

Note that when $(|D|, |F|) = (1, 0)$, this claim holds trivially, so it is sufficient to consider the case where

$$|D| + |F| \geq 2. \tag{5.62}$$

We now show that (5.61) holds under the conditions of each of corollaries.

5.5.1 Proof of Corollary 1.2

In this corollary, $1 < (T + 1/T)/2 \leq 3$, and if either (a) $|D| + 2|F| \geq 3$ and $|D|$ is odd or (b) $|D| + 2|F| \geq 4$ and $|D|$ is even, the right hand side of (5.61) ≤ 0, so the claim holds trivially. Therefore, all we need to show is the case where $(|D|, |F|) = (2, 0)$. In this case, since each off-diagonal coordinate $\leq 4\sqrt{2} - 5 \equiv \rho_0$, it follows from definitions and basic algebra that $\omega \geq 2(1 - \rho_0) = 4(3 - 2\sqrt{2})$, and (5.61) follows by noting that $\frac{2}{T}[(T + 1/T)/2 + 1 - (|D| + 2|F|)] = (1 - 1/T)^2 \leq 4(3 - 2\sqrt{2})$. □

5.5.2 Proof of Corollary 1.3

In this corollary, $1 < (T + 1/T)/2 \leq 5$. First, we consider the case where $|D|$ is odd. By similar argument, (5.61) holds trivially when $|D| + 2|F| \geq 5$, so all we need to consider is the case $(|D|, |F|) = (1, 1)$ and the case $(|D|, |F|) = (3, 0)$. In both cases, $|D| + 2|F| = 3$. By (5.59), when $\omega T < 1$, there must be $T < 1/\min(\lambda_2^*, 3\lambda_3^*)$. By the conditions of this corollary, it follows $T < (5 + 2\sqrt{6})/4 < 3 + 2\sqrt{2}$. When $1 < T < 3 + 2\sqrt{2}$, there is $T + 1/T - 6 < 0$, and thus (5.61) holds for $\omega T < 1$. When $\omega T \geq 1$, (5.61) holds if and only if $\omega T + \frac{1}{\omega T} - 2 \geq T + 1/T - 6$. By basic algebra, this holds if

$$\omega \geq \frac{1}{4}\big[(1 - 1/T) + \sqrt{(1 - 1/T)^2 - 4/T}\big]^2. \tag{5.63}$$

Note that the right hand of (5.63) is a monotone in T and has a maximum of $(3+2\sqrt{2})(5-2\sqrt{6})$ at $T = (5 + 2\sqrt{6})$. Now, on one other hand, when $(|D|, |F|) = (1, 0)$, by (5.59) and conditions of the corollary, $\omega \geq 3\lambda_3^* > (3 + 2\sqrt{2})(5 - 2\sqrt{6})$. On the other hand, when $(|D|, |F|) = (1, 1)$, by basic algebra and that each off-diagonal coordinate of $\Omega \leq \sqrt{1 + (\sqrt{6} - \sqrt{2})}/(1 + \sqrt{3/2}) \equiv \rho_1$ in magnitude, $\omega \geq 1 - \rho_1^2 = (3 + 2\sqrt{2})(5 - 2\sqrt{6})$. Combining these gives (5.61).

We now consider the case where $|D|$ is even. By similar argument, (5.61) holds when $|D| + 2|F| \geq 6$, so all we need is to show is that (5.61) holds for the following three cases: $(|D|, |F|) = (4, 0), (2, 1), (2, 0)$. Equivalently, this is to show that $\omega \geq \frac{2}{T}[(T + 1/T)/2 - 3]$ in the first two cases and that $\omega \geq \frac{2}{T}[(T + 1/T)/2 - 1]$ in the last case. Similarly, by the monotonicity of the right hand side of these inequalities, all we need to show is $\omega \geq 4(5 - 2\sqrt{6})$ in the first two cases, and $\omega \geq 8(5 - 2\sqrt{6})$ in the last case. Now, on one hand, using (5.59), $\omega \geq 4\lambda_4^*$ in the first case, and $\omega \geq 2\lambda_3^*$ in the second case, so by the conditions of the corollary, $\omega \geq 4(5 - 2\sqrt{6})$ in the first two cases. On the other hand, in the last case, since all off-diagonal coordinates of $\Omega \leq 8\sqrt{6} - 19 \equiv \rho_0$ in magnitude, and $\omega \geq 2(1 - \rho_0) = 8(5 - 2\sqrt{6})$. Combining these gives (5.61). □

5.5.3 Proof of Corollary 1.4

Let N be the unique integer such that $2N - 1 \leq (T + 1/T)/2 < 2N + 1$. First, we consider the case where $|D|$ is odd. Note that when $|D| + 2|F| \geq 2N + 1$, the right hand side of (5.61) ≤ 0, so all we need to consider is the case $|D| + 2|F| \leq 2N - 1$. Write for short $k = k(D, F) = |D| + |F|$ and $j = j(D, F) = (|D| + 2|F| + 1)/2$. By (5.62), definitions, and that $|D| + 2|F| \leq 2N - 1$, it is seen that $2 \leq k \leq 2N - 1$ and $(k + 1)/2 \leq j \leq \min\{k, N\}$. By the condition of the corollary,

$$\lambda_k^* \geq \frac{(T + 1/T)/2 - 2j + 2 + \sqrt{[(T + 1/T)/2 - 2j + 2]^2 - 1}}{T(2k - 2j + 1)}$$

Note that $|D| = 2k - 2j + 1$. Combining these with (5.59) gives

$$\omega T \geq (2k-2j+1)\lambda_k^* T \geq (T+1/T)/2 - 2j + 2 + \sqrt{[(T+1/T)/2 - 2j + 2]^2 - 1} \geq 1.$$

and (5.61) follows by basic algebra.

We now consider the case where $|D|$ is even. Similarly, the right hand side of (5.61) is negative when $|D| + 2|F| \geq 2(N+1)$, so we only need to consider the case where $|D| + 2|F| \leq 2N$. Similarly, write for short $k = k(D,F) = |D| + |F|$ and $j = (|D| + 2|F|)/2$. It is seen that $2 \leq k \leq 2N$ and $k/2 \leq j \leq \min\{k-1, N\}$. By the conditions of the corollary,

$$\lambda_k^* \geq \frac{(T+1/T)/2 + 1 - 2j}{T(k-j)}.$$

Note that $|D| = k - j$. It follows from (5.59) that $\omega \geq 2(k-j)\lambda_k^* \geq \frac{2}{T}[(T+1/T)/2 + 1 - 2j]$, and (5.61) follows. □

5.6 Proof of Lemma 2.2

To show the claim, it is sufficient to show that for any fixed $1 \leq j \leq p$,

$$P(j \notin \mathcal{U}_p^*, \beta_j \neq 0) \leq L_p[p^{-\rho_j^*} + p^{-(m+1)\vartheta} + o(1/p)]. \tag{5.64}$$

Using Lemma 2.1 and [25, Lemma 3.1], there is an event A_p that depends on (X, β) such that $P(A_p^c) \leq o(1/p)$ and that over the event, Ω^* is K-sparse with $K = C(\log(p))^{1/\gamma}$,

$$\|\Omega^* - \Omega\|_\infty \leq (\log(p))^{-(1-\gamma)}, \qquad \|(X'X - \Omega)\beta\|_\infty \leq C\|\Omega\|\sqrt{2\log(p)}p^{-[(\kappa-(1-\vartheta)]/2},$$

and for all subset B with size $\leq m$,

$$\|G^{B,B} - \Omega^{B,B}\|_\infty \leq L_p p^{-\kappa/2}.$$

Recall that $\mathcal{G}^*$ is the GOSD and $\mathcal{G}_S^*$ is the subgraph of the GOSD formed by the nodes in the support of β, $S(\beta) = \{1 \leq j \leq p : \beta_j \neq 0\}$. When $\beta_j \neq 0$, there is a unique component $\mathcal{I}_0$ such that $j \in \mathcal{I}_0 \lhd \mathcal{G}_S^*$ ($A \lhd B$ means that A is component or maximal connected subgraph of B). Let B_p be the event $|\mathcal{I}_0| \leq m$. By Frieze [19], it is seen that

$$P(B_p^c \cap A_p) \leq L_p p^{-(m+1)\vartheta}.$$

So to show (5.64), it is sufficient to show that

$$P(j \notin \mathcal{U}_p^*, j \in \mathcal{I}_0 \lhd \mathcal{G}_S^*, A_p \cap B_p) \leq L_p p^{-\rho_j^*}. \tag{5.65}$$

Now, in the screening procedure, when we screen $\mathcal{I}_0$, we have $\mathcal{I}_0 = \hat{D} \cup \hat{F}$ as in (1.7). Since the event $\{j \notin \mathcal{U}_p^*, j \in \mathcal{I}_0 \lhd \mathcal{G}_S^*\}$ is contained in the event $\{T(Y, \hat{D}, \hat{F}) < t(\hat{D}, \hat{F})\}$,

$$P(j \notin \mathcal{U}_p^*, j \in \mathcal{I}_0 \lhd \mathcal{G}_S^*, A_p \cap B_p) \leq P(T(Y, \hat{D}, \hat{F}) \leq t(\hat{D}, \hat{F}), j \in \mathcal{I}_0 \lhd \mathcal{G}_S^*, A_p \cap B_p),$$

where the right hand side does not exceed

$$\sum_{(\mathcal{I}_0, D, F): j \in \mathcal{I}_0 \ \&\ \mathcal{I}_0 = D \cup F \text{ is a partition}} P(T(Y, D, F) \leq t(D, F), j \in \mathcal{I}_0 \lhd \mathcal{G}_S^*, A_p \cap B_p); \tag{5.66}$$

note that $(\mathcal{I}_0, D, F)$ do not depend on z (but may still depend on (X, β)). First, note that over the event A_p, there are at most $(eK)^{m+1}$ $\mathcal{I}_0$ such that $j \in \mathcal{I}_0$ and $|\mathcal{I}_0| \le m$. Second, note that for each $\mathcal{I}_0$, there are only finite ways to partition it to D and F. Last, note that for any fixed j and $\mathcal{I}_0$, $P(j \in \mathcal{I}_0 \lhd \mathcal{G}_S^*) \le \epsilon_p^{|\mathcal{I}_0|}$. Combining these observations, to show (5.65), it is sufficient to show that for any such triplet $(\mathcal{I}_0, D, F)$,

$$\epsilon_p^{|\mathcal{I}_0|} P\big(T(Y, D, F) \le t(D, F) \big| \{j \in \mathcal{I}_0 \lhd \mathcal{G}_S^*\} \cap A_p \cap B_p\big) \le L_p p^{-\rho_j^*}. \tag{5.67}$$

We now show (5.67). Since $\lambda_m \ge C > 0$, it follows from the definition of A_p and basic algebra that for any realization of (X, β) in $A_p \cap B_p$,

$$\|(G^{\mathcal{I}_0, \mathcal{I}_0})^{-1}\|_\infty \le C. \tag{5.68}$$

Recall that $\tilde{Y} = X'Y$ and denote for short $y = (G^{\mathcal{I}_0, \mathcal{I}_0})^{-1} \tilde{Y}^{\mathcal{I}_0}$. It is seen that

$$y = \beta^{\mathcal{I}_0} + w + rem, \qquad w \sim N(0, (G^{\mathcal{I}_0, \mathcal{I}_0})^{-1}), \qquad rem \equiv (G^{\mathcal{I}_0, \mathcal{I}_0})^{-1} G^{\mathcal{I}_0, \mathcal{I}_0^c} \beta^{\mathcal{I}_0^c}. \tag{5.69}$$

Since $\mathcal{I}_0$ is a component of $\mathcal{G}_S^*$, $(\Omega^*)^{\mathcal{I}_0, \mathcal{I}_0^c} \beta^{\mathcal{I}_0^c} = 0$. Therefore, we can write $rem = (G^{\mathcal{I}_0, \mathcal{I}_0})^{-1}(I + II)$, where $I = (G^{\mathcal{I}_0, \mathcal{I}_0^c} - \Omega^{\mathcal{I}_0, \mathcal{I}_0^c})\beta^{\mathcal{I}_0^c}$ and $II = [\Omega^{\mathcal{I}_0, \mathcal{I}_0^c} - (\Omega^*)^{\mathcal{I}_0, \mathcal{I}_0^c}]\beta^{\mathcal{I}_0^c}$. By the definition of A_p, $\|I\|_\infty \le C\sqrt{2\log(p)} p^{-[\kappa - (1-\vartheta)]/2}$, and $\|II\|_\infty \le \|\Omega - \Omega^*\|_\infty \|\beta^{\mathcal{I}_0^c}\|_\infty \le C\tau_p (\log(p))^{-(1-\gamma)}$. Combining these with (5.68) gives $\|rem\|_\infty \le C\tau_p(\log(p))^{-(1-\gamma)}$.

At the same time, let y_1, w_1, and rem^1 be the restriction of y, w, and rem to indices in D, correspondingly, and let $H = [G^{D,D} - G^{D,F}(G^{F,F})^{-1} G^{F,D}]$. By (5.69) and direct calculations,

$$T(Y, D, F) = y_1' H y_1, \qquad y_1 \sim N(\beta^D + rem^1, H^{-1}),$$

and so $T(Y, D, F)$ is distributed as $\chi^2_{|D|}(\delta)$, where the non-central parameter is

$$(\beta^D + rem^1)' H (\beta^D + rem^1) = \delta + O((\log(p))^\gamma), \qquad \delta \equiv (\beta^D)' H \beta^D;$$

where since $\lambda_m^* \ge C$, $\delta \ge C\tau_p^2$ and is the dominating terms. It follows that

$$P(T(Y, D, F) \le t(D, F) \big| \{j \in \mathcal{I}_0 \lhd \mathcal{G}_S^*\} \cap A_p \cap B_p\big) \lesssim P\big(\chi^2_{|D|}(\delta) \le t(D, F)\big). \tag{5.70}$$

Now, first, by definitions, $\delta \ge 2\omega(D, F; \Omega) r \log(p)$, so by basic knowledge on non-central χ^2,

$$P(\chi^2_{|D|}(\delta) \le t(D, F)) \le P(\chi^2_{|D|}(2\omega(D, F; \Omega) r \log(p)) \le t(D, F)). \tag{5.71}$$

Second, recalling $t(D, F) = 2q\log(p)$, we have

$$P(\chi^2_{|D|}(2\omega(D, F; \Omega) r \log(p)) \le t(D, F)) \le L_p p^{-[(\sqrt{\omega(D,F;\Omega)r} - \sqrt{q})_+]^2}. \tag{5.72}$$

Inserting (5.71)-(5.72) into (5.70) and recalling $\epsilon_p = p^{-\vartheta}$,

$$\epsilon_p^{|\mathcal{I}_0|} P(T(Y, D, F) \le t(D, F) \big| \{j \in \mathcal{I}_0 \lhd \mathcal{G}_S^*\} \cap A_p \cap B_p\big) \le L_p p^{-(|\mathcal{I}_0|\vartheta + [(\sqrt{\omega(D,F;\Omega)r} - \sqrt{q})_+]^2)}. \tag{5.73}$$

By the choice of q and direct calculations,

$$|\mathcal{I}_0|\vartheta + [(\sqrt{\omega(D, F; \Omega) r} - \sqrt{q})_+]^2 \ge \rho(D, F; \Omega) \ge \rho_j^*, \tag{5.74}$$

where $\rho(D, F; \Omega)$ and ρ_j^* are as in and (1.23) and (1.21). Combining (5.73)-(5.74) gives (5.67). □

5.7 Proof of Lemma 2.3

In the screening stage, suppose we pick the threshold $t(\hat{D}, \hat{F}) = 2q\log(p)$ in a way such that there is a constant $q_0(\vartheta, r) > 0$ such that

$$q = q(\hat{D}, \hat{F}) \geq q_0(\vartheta, r, \kappa) > 0,$$

Recall that $\mathcal{G}^*$ denotes the GOSD. Let $\mathcal{U}_p^*$ be the set of retained indices. Viewing it as a subgraph of $\mathcal{G}^*$, $\mathcal{U}_p^*$ decomposes into many components

$$\mathcal{U}_p^* = \mathcal{I}^{(1)} \cup \mathcal{I}^{(2)} \ldots \cup \mathcal{I}^{(N)}.$$

Recall that $\tilde{Y} = X'Y$. The following lemma is proved below.

Lemma 5.1 *There is a constant $c_1 = c_1(\vartheta, r, \kappa, \gamma, A) > 0$ such that with probability at least $1 - o(1/p)$, for any component $\mathcal{I}_0 \lhd \mathcal{U}_p^*$, $\|\tilde{Y}^{\mathcal{I}_0}\|^2 \geq 2c_1|\mathcal{I}_0|\log(p)$.*

The remaining part of the proof is similar to that of [25, Lemma 2.3] so we omit it. We note that however Lemma 5.1 is new and needs a much harder proof. □

5.7.1 Proof of Lemma 5.1

First, we need some notations. Let $\mathcal{I}_0$ be a component of $\mathcal{U}_p^*$, and let $\mathcal{I}_0^{(i)}$, $1 \leq i \leq N_0$, be all connected subgraphs with size $\leq m$, listed lexicographically, where N_0 is an integer that may depend on (X, Y). For each $1 \leq i \leq N_0$, let $\mathcal{I}_0^{(i)} = \hat{D}^{(i)} \cup \hat{F}^{(i)}$ be the *exactly the same* partition when we screen $\mathcal{I}_0^{(i)}$ in the m-stage χ^2-screening of the GS-step. In out list, we only keep $\mathcal{I}_0^{(i)}$ such that $\hat{D}^{(i)} \cap \mathcal{I}_0 \neq \emptyset$. Since $\mathcal{I}_0$ is a component of $\mathcal{U}_p^*$ and $\mathcal{I}_0^{(i)}$ is a connected subgraph, it follows from the way that the χ^2-screening is designed and the definition of $\hat{D}^{(i)}$ that

$$\mathcal{I}_0^{(i)} \subset \mathcal{I}_0, \qquad \text{and} \qquad \hat{D}^{(i)} = \mathcal{I}_0^{(i)} \setminus (\cup_{j=1}^{i-1}\mathcal{I}_0^{(j)}), \qquad 1 \leq i \leq N_0. \tag{5.75}$$

and

$$\mathcal{I}_0 = \hat{D}^{(1)} \cup \hat{D}^{(2)} \ldots \cup \hat{D}^{(N_0)} \text{ is a partition}, \tag{5.76}$$

where $\hat{F}^{(1)}$ is empty.

Now, for each $1 \leq i \leq N_0$, recall that as long as $G^{\mathcal{I}_0^{(i)}, \mathcal{I}_0^{(i)}}$ is non-singular, the χ^2-test score in Graphlet Screening is $T(Y, \hat{D}^{(i)}, \hat{F}^{(i)}) = T(Y, \hat{D}^{(i)}, \hat{F}^{(i)}; \mathcal{I}_0^{(i)}, X, p, n)$:

$$T(Y, \hat{D}^{(i)}, \hat{F}^{(i)}) = (\tilde{Y}^{\mathcal{I}_0^{(i)}})'(G^{\mathcal{I}_0^{(i)}, \mathcal{I}_0^{(i)}})^{-1}\tilde{Y}^{\mathcal{I}_0^{(i)}} - (\tilde{Y}^{\hat{F}^{(i)}})'(G^{\hat{F}^{(i)}, \hat{F}^{(i)}})^{-1}\tilde{Y}^{\hat{F}^{(i)}}.$$

By basic algebra and direct calculations, it can be verified that

$$T(Y, \hat{D}^{(i)}, \hat{F}^{(i)}) = \|W_i\|^2,$$

where $W_i = W(\tilde{Y}, \hat{D}^{(i)}, \hat{F}^{(i)}; \mathcal{I}_0^{(i)}, X, p, n)$ is defined as $W_i = V_i^{-1/2}y_i$, and for short,

$$V_i = G^{\hat{D}^{(i)}, \hat{D}^{(i)}} - G^{\hat{D}^{(i)}, \hat{F}^{(i)}}(G^{\hat{F}^{(i)}, \hat{F}^{(i)}})^{-1}G^{\hat{F}^{(i)}, \hat{D}^{(i)}}, \qquad y_i = \tilde{Y}^{\hat{D}^{(i)}} - G^{\hat{D}^{(i)}, \hat{F}^{(i)}}(G^{\hat{F}^{(i)}, \hat{F}^{(i)}})^{-1}\tilde{Y}^{\hat{F}^{(i)}}.$$

At the same time, for a constant $\delta > 0$ to be determined, define $\tilde{\Omega}$ by

$$\tilde{\Omega}(i, j) = G(i, j) \cdot 1\{|G(i, j)| \geq \delta\}.$$

The definition of $\tilde{\Omega}$ is the same as that of Ω^*, except for that the threshold δ would be selected differently. We introduce a counterpart of W_i which we call W_i^*,

$$W_i^* = V_i^{-1/2} y_i^*. \tag{5.77}$$

where

$$y_i^* = \tilde{Y}^{\hat{D}^{(i)}} - \tilde{\Omega}^{\hat{D}^{(i)},\hat{F}^{(i)}} (\tilde{\Omega}^{\hat{F}^{(i)},\hat{F}^{(i)}})^{-1} \tilde{Y}^{\hat{F}^{(i)}}.$$

Let $W^* = ((W_1^*)', (W_2^*)', \ldots, (W_{N_0}^*)')'$, and define $|\mathcal{I}_0| \times |\mathcal{I}_0|$ matrices H_1 and H_2 as follows: H_1 is a diagonal block-wise matrix where the i-th block is $V_i^{-1/2}$, and $H_2 = \tilde{H}_2^{\mathcal{I}_0,\mathcal{I}_0}$, where $\tilde{H}_2$ is a $p \times p$ matrix such that for every component $\mathcal{I}_0$ of $\mathcal{U}_p^*$, and $\hat{D}^{(i)}$ and $\hat{F}^{(i)}$ defined on each component,

$$\tilde{H}_2^{\hat{D}^{(i)},\hat{F}^{(i)}} = -(\tilde{\Omega})^{\hat{D}^{(i)},\hat{F}^{(i)}} [(\tilde{\Omega})^{\hat{F}^{(i)},\hat{F}^{(i)}}]^{-1}, \qquad \tilde{H}_2^{\hat{D}^{(i)},\hat{D}^{(i)}} = I_{|\hat{D}^{(i)}|},$$

and that the coordinates of $\tilde{H}_2$ are zero elsewhere. Here I_k stands for $k \times k$ identity matrix. From the definitions, it is seen that

$$W^* = H_1 H_2 \tilde{Y}^{\mathcal{I}_0}. \tag{5.78}$$

Compared to W_i, W_i^* is relatively easier to study, for it induces column-sparsity of H_2. In fact, using [25, Lemma 2.2, 3.1], there is an event A_p that depends on (X, β) such that $P(A_p^c) \leq o(1/p^2)$ and that over the event, for all subset B with size $\leq m$,

$$\|G^{B,B} - \Omega^{B,B}\|_\infty \leq L_p p^{-\kappa/2}. \tag{5.79}$$

The following lemma is proved below.

Lemma 5.2 *Fix $\delta > 0$. Over the event A_p, there is a constant $C > 0$ such that each row and column of $\tilde{H}_2$ has no more than C nonzero coordinates.*

We are now ready to show Lemma 5.1. To begin with, note that since we accept $\hat{D}^{(i)}$ when we graphlet-screen $\mathcal{I}_0^{(i)}$,

$$\|W_i\|^2 \geq 2q_0 |\hat{D}^{(i)}| \log(p). \tag{5.80}$$

At the same time, by basic algebra, $\|W_i - W_i^*\| \leq \|V_i^{-1/2}\| \|y_i - y_i^*\|$, and

$$\|y_i - y_i^*\| \leq \|G^{\hat{D}^{(i)},\hat{F}^{(i)}} (G^{\hat{F}^{(i)},\hat{F}^{(i)}})^{-1} - (\tilde{\Omega})^{\hat{D}^{(i)},\hat{F}^{(i)}} ((\tilde{\Omega})^{\hat{F}^{(i)},\hat{F}^{(i)}})^{-1}\|_\infty \cdot \|\tilde{Y}^{\hat{F}^{(i)}}\|.$$

First, since $\lambda_m^* \geq C$, it is seen that over the event A_p, $\|V_i^{-1/2}\| \leq C$. Second, by similar reasons, it is not hard to see that except for probability $o(p^{-2})$, $\|G^{\hat{D}^{(i)},\hat{F}^{(i)}} (G^{\hat{F}^{(i)},\hat{F}^{(i)}})^{-1} - (\tilde{\Omega})^{\hat{D}^{(i)},\hat{F}^{(i)}} ((\tilde{\Omega})^{\hat{F}^{(i)},\hat{F}^{(i)}})^{-1}\|_\infty \leq C\delta^{1-\gamma}$, and $\|\tilde{Y}^{\hat{F}^{(i)}}\| \leq C\sqrt{\log(p)} \leq C\tau_p$. Combining these gives

$$\|W_i - W_i^*\| \leq C\delta^{1-\gamma}\tau_p, \tag{5.81}$$

Inserting this to (5.80), if we choose δ to be a sufficiently small constant,

$$\|W_i^*\|^2 \geq \frac{1}{2}\|W_i\|^2 \geq q_0 |\hat{D}^{(i)}| \log(p).$$

At the same time, by definitions, it follows from $\|V_i^{-1/2}\| \leq C$ that $\|H_1\| \leq C$. Also, since over the event A_p, each coordinate of H_2 is bounded from above by a constant in magnitude, it follows from Lemma 5.2 that $\|H_2\| \leq C$. Combining this with (5.76)-(5.78), it follows from basic algebra that except for probability $o(p^{-2})$,

$$q_0 |\mathcal{I}_0| \log(p) \leq \|W^*\|^2 \leq \|H_1 H_2 \tilde{Y}^{\mathcal{I}_0}\|^2 \leq C\|\tilde{Y}^{\mathcal{I}_0}\|^2,$$

and the claim follows. □

5.7.2 Proof of Lemma 5.2

By definitions, it is equivalent to show that over the event A_p, each row and column of $\tilde{H}_2$ has finite nonzero coordinates. It is seen that each row of $\tilde{H}_2$ has $\leq m$ nonzeros, so all we need to show that each column of $\tilde{H}_2$ has finite nonzeros.

Towards this end, we introduce a new graph $\tilde{\mathcal{G}} = (V, E)$, where $V = \{1, 2, \ldots, p\}$ and nodes i and j are connected if and only if $|\tilde{\Omega}(i,j)| \neq 0$. This definition is the same as GOSD, except that Ω^* is substituted by $\tilde{\Omega}$. It is seen that over the event A_p, for any $\Omega \in \mathcal{M}_p^*(\gamma, c_0, g, A)$, $\tilde{\mathcal{G}}$ is K-sparse with $K \leq C\delta^{-1/\gamma}$. The key for the proof is to show that for any $k \neq \ell$ such that $\tilde{H}_2(k,\ell) \neq 0$, there is a path with length $\leq (m-1)$ in $\tilde{\mathcal{G}}$ that connects k and ℓ.

To see the point, we note that when $\tilde{H}_2(k,\ell) \neq 0$, there must be an i such that $k \in \hat{D}^{(i)}$ and $\ell \in \hat{F}^{(i)}$. We claim that there is a path in $\mathcal{I}_0^{(i)}$ (which is regarded as a subgraph of $\tilde{\mathcal{G}}$) that connects k and ℓ. In fact, if k and ℓ are not connected in $\mathcal{I}_0^{(i)}$, we can partition $\mathcal{I}_0^{(i)}$ into two separate sets of nodes such that one contains k and the other contains ℓ, and two sets are disconnected. In effect, both the matrix $\tilde{\Omega}^{\hat{D}^{(i)},\hat{D}^{(i)}}$ and $\tilde{\Omega}^{\hat{D}^{(i)},\hat{F}^{(i)}}$ can be visualized as two by two blockwise matrix, with off-diagonal blocks being 0. As a result, it is seen that $\tilde{H}_2(k,\ell) = 0$. This contradiction shows that whenever $\tilde{H}_2(k,\ell) \neq 0$, k and ℓ are connected by a path in $\mathcal{I}_0^{(i)}$. Since $|\mathcal{I}_0^{(i)}| \leq m$, there is a path $\leq m-1$ in $\tilde{\mathcal{G}}$ that connects k and ℓ where $k \neq \ell$.

Finally, since $\tilde{\mathcal{G}}$ is K-sparse with $K = C\delta^{-1/\gamma}$, for any fixed ℓ, there are at most finite k connecting to ℓ by a path with length $\leq (m-1)$. The claim follows. □

5.8 Proof of Theorem 1.4

Since σ is known, for simplicity, we assume $\sigma = 1$. First, consider (1.33). By Theorem 1.2 and (1.23), $\rho_{gs} = \min_{\{(D,F):D\cap F=\emptyset, D\neq\emptyset, D\cup F\subset\{1,2\}\}} \rho(D,F;\Omega)$, where we have used that G is a diagonal block-wise matrix, each block is the same 2×2 matrix. To calculate $\rho(D,F;\Omega)$, we consider three cases (a) $(|D|,|F|) = (2,0)$, (b) $(|D|,|F|) = (1,0)$, (c) $(|D|,|F|) = (1,1)$. By definitions and direct calculations, it is seen that $\rho(D,F;\Omega) = \vartheta + [(1-|h_0|)r]/2$ in case (a), $\rho(D,F;\Omega) = (\vartheta+r)^2/(4r)$ in case (b), and $\rho(D,F;\Omega) = 2\vartheta + [(\sqrt{(1-h_0^2)r} - \vartheta/\sqrt{(1-h_0^2)r})_+]^2/4$ in case (c). Combining these gives the claim.

Next, consider (1.34). Similarly, by the block-wise structure of G, we can restrict our attention to the first two coordinates of β, and apply the subset selection to the size 2 subproblem where the Gram matrix is the 2×2 matrix with 1 on the diagonals and h_0 on the off-diagonals. Fix $q > 0$, and let the tuning parameter $\lambda_{ss} = \sqrt{2q_{ss}\log(p)}$. Define

$$f_{ss}^{(1)}(q) = f_{lasso,1}(q) = \vartheta + [(\sqrt{r} - \sqrt{q})_+]^2, \qquad f_{ss}^{(2)}(q) = 2\vartheta + [(\sqrt{r(1-h_0^2)} - \sqrt{q})_+]^2,$$

and

$$f_{ss}^{(3)}(q) = 2\vartheta + 2[(\sqrt{r(1-|h_0|)} - \sqrt{q})_+]^2,$$

where $x_+ = \max\{x, 0\}$. The following lemma is proved below, where the key is to use [25, Lemma 4.3].

Lemma 5.3 *Fix $q > 0$, and apply the subset selection to the aforementioned size 2 subproblem with $\lambda_{ss} = \sqrt{2q\log(p)}$. As $p \to \infty$, the worst-case Hamming error rate is $L_p p^{-f_{ss}(q)}$, where $f_{ss}(q) = f_{ss}(q,\vartheta,r,h_0) = \min\{\vartheta + (1-|h_0|)r/2, q, f_{ss}^{(1)}(q), f_{ss}^{(2)}(q), f_{ss}^{(3)}(q)\}$.*

By direct calculations, $\rho_{ss}(\vartheta, r, h_0) = \max_{\{q>0\}} f_{ss}(\vartheta, r, h_0)$ and the claim follows.

Last, consider (1.35). The proof is very similar to that of the subset selection, except for that we need to use [25, Lemma 4.1], instead of [25, Lemma 4.3]. For this reason, we omit the proof. □

5.8.1 Proof of Lemma 5.3

By the symmetry in (1.30)-(1.31) when G is given by (1.32), we only need to consider that case where $h_0 \in [0, 1)$ and $\beta_1 \geq 0$. Introduce events, $A_0 = \{\beta_1 = \beta_2 = 0\}$, $A_1 = \{\beta_1 \geq \tau_p, \beta_2 = 0\}$, $A_{21} = \{\beta_1 \geq \tau_p, \beta_2 \geq \tau_p\}$,$A_{22} = \{\beta_1 \geq \tau_p, \beta_2 \leq -\tau_p\}$, $B_0 = \{\hat{\beta}_1 = \hat{\beta}_2 = 0\}$, $B_1 = \{\hat{\beta}_1 > 0, \hat{\beta}_2 = 0\}$, $B_{21} = \{\hat{\beta}_1 > 0, \hat{\beta}_2 > 0\}$ and $B_{22} = \{\hat{\beta}_1 > 0, \hat{\beta}_2 < 0\}$. It is seen that the Hamming error

$$= L_p(I + II + III), \tag{5.82}$$

where $I = P(A_0 \cap B_0^c)$, $II = P(A_1 \cap B_1^c)$ and $III = P(A_{21} \cap B_{21}^c) + P(A_{22} \cap B_{22}^c)$.

Let H be the 2×2 matrix with ones on the diagonals and h_0 on the off-diagonals, $\alpha = (\beta_1, \beta_2)'$, and $w = (\tilde{Y}_1, \tilde{Y}_2)$, where we recall $\tilde{Y} = X'Y$. It is seen that $w \sim N(H\alpha, H)$. Write for short $\lambda = \sqrt{2q\log(p)}$. Define regions on the plane of $(\tilde{Y}_1, \tilde{Y}_2)$, $D_0 = \{\max(|\tilde{Y}_1|, |\tilde{Y}_2|) > \lambda \text{ or } \tilde{Y}_1^2 + \tilde{Y}_2^2 - 2h_0\tilde{Y}_1\tilde{Y}_2 > 2\lambda^2(1-h_0^2)\}$, $D_1 = \{|\tilde{Y}_1| < \lambda \ , \ \tilde{Y}_1 < \tilde{Y}_2 \text{ or } |\tilde{Y}_2 - h_0\tilde{Y}_1| > \lambda\sqrt{1-h_0^2}\}$, $D_{21} = \{\tilde{Y}_2 - h_0\tilde{Y}_1 < \lambda\sqrt{1-h_0^2} \text{ or } \tilde{Y}_1 - h_0\tilde{Y}_2 < \lambda\sqrt{1-h_0^2}\}$ and $D_{22} = \{\tilde{Y}_2 - h_0\tilde{Y}_1 > -\lambda\sqrt{1-h_0^2} \text{ or } \tilde{Y}_1 - h_0\tilde{Y}_2 > \lambda\sqrt{1-h_0^2} \text{ or } \tilde{Y}_1^2 + \tilde{Y}_2^2 - 2h_0\tilde{Y}_1\tilde{Y}_2 < 2\lambda^2(1-h_0^2)\}$. Using [25, Lemma 4.3], we have $B_0^c = \{(\tilde{Y}_1, \tilde{Y}_2)' \in D_0\}$, $B_1^c = \{(\tilde{Y}_1, \tilde{Y}_2)' \in D_1\}$, $B_{21}^c) = \{(\tilde{Y}_1, \tilde{Y}_2)' \in D_{21}\}$, and $B_{22}^c = \{(\tilde{Y}_1, \tilde{Y}_2)' \in D_{22}\}$. By direct calculation and Mills' ratio, it follows that for all $\mu \in \Theta_p(\tau_p)$,

$$I = L_p \cdot (P(N(0,1) > \lambda) + P(\chi_2^2 > 2\lambda^2)) = L_p \cdot p^{-q}, \tag{5.83}$$

$$II \leq L_p \cdot P(N((\tau_p, h_0\tau_p)', H) \in D_1) = L_p \cdot p^{-\vartheta - \min[(\sqrt{r}-\sqrt{q})^2, (1-h_0)r/2, q]}, \tag{5.84}$$

and when $\beta_1 = \tau_p$ and $\beta_2 = 0$, the equality holds in (5.84). At the same time, note that over the event A_{21}, the worst case scenario, is where $\beta_1 = \beta_2 = \tau_p$. In such a case, $(\tilde{Y}_1, \tilde{Y}_2)' \sim N(((1+h_0)\tau_p, (1+h_0)\tau_p)', H)$. Combining this with Mills' ratio, it follows that for all $\mu \in \Theta_p(\tau_p)$,

$$P(A_{21} \cap B_{21}^c) = P((\tilde{Y}_1, \tilde{Y}_2)' \in D_{21}) \leq L_p \cdot p^{-2\vartheta - (\sqrt{r(1-h_0^2)} - \sqrt{q})_+^2}, \tag{5.85}$$

and the equality holds when $\beta_1 = \beta_2 = \tau_p$. Similarly, note that over the event A_{22}, in the worst case scenario, $\beta_1 = -\beta_2 = \tau_p$. In such a case, $(\tilde{Y}_1, \tilde{Y}_2)' \sim N(((1-h_0)\tau_p, -(1-h_0)\tau_p)', H)$. Combining this with Mills' ratio, it follows that for all $\mu \in \Theta_p(\tau_p)$,

$$P(A_{22} \cap B_{22}^c) = P((\tilde{Y}_1, \tilde{Y}_2)' \in D_{22}) \leq L_p \cdot p^{-2\vartheta - \min([(\sqrt{r(1-h_0^2)} - \sqrt{q})_+]^2, 2\{[\sqrt{r(1-h_0)} - \sqrt{q}]_+\}^2)}, \tag{5.86}$$

and the equality holds when $\beta_1 = -\beta_2 = \tau_p$. Inserting (5.83)-(5.86) into (5.82) gives the claim. □

References

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

[2] Bajwa, W. U., Haupt, J. D., Raz, G. M., Wright, S. J. and Nowak, R. D. (2007). Toeplitz-structured compressed sensing matrices. *Proc. SSP' 07, Madison, WI, Aug. 2007*, 294–298.

[3] Bickel, P. and Levina, E. (2008). Covariance regularization by thresholding. *Ann. Statist.*, **36**(6), 2577–2604.

[4] Cai, T., Liu, W. and Luo, X. (2010). A constrained ℓ^1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.*, **106**(494), 594-607.

[5] Candes, E. and Plan, Y. (2009). Near-ideal model selection by ℓ^1-minimization. *Ann. Statist.*, **37**(5), 2145–2177.

[6] Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n (with discussion). *Ann. Statist.*, **35**(6), 2313–2404.

[7] Chen, S., Donoho, D. and Saunders, M. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Computing*, **20**(1), 33–61.

[8] Dinur, I. and Nissim, K. (2003). Revealing information while preserving privacy. *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 202–210, ACM Press.

[9] Donoho, D. (2006a). For most large underdetermined systems of linear equations the minimal ℓ^1-norm solution is also the sparsest solution. *Comm. Pure Appl. Math.*, **59**(7), 907–934.

[10] Donoho, D. (2006b). Compressed sensing. *IEEE Trans. Inform. Theory*, **52**(4), 1289–1306.

[11] Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, **32**(3), 962–994.

[12] Donoho, D. and Jin, J. (2008). Higher criticism thresholding: optimal feature selection when useful features are rare and weak. *Proc. Natl. Acad. Sci.*, **105**(39), 14790–14795.

[13] Efron, B., Hastie, H., Johnstone, I. and Tibshirani, R. (2004). Least angle regression (with discussion). *Ann. Statist.*, **32**(2), 407-499.

[14] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, **96**(456), 1349–1360.

[15] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. Roy. Soc. Statist. B*, **70**(5), 849–911.

[16] Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.*, **22**(4), 1947–1975.

[17] Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432–441.

[18] Friedman, J. H., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**(1). `http://cran.r-project.org/web/packages/glmnet/index.html`

[19] Frieze, A.M. and Molloy, M. (1999). Splitting an expander graph. *J. Algorithms*, **33**(1), 166–172.

[20] Genovese, C., Jin, J. and Wasserman, L. (2012). Revisiting marginal regression. *Manuscript*.

[21] Fan, J., Han, X. and Gu, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *To appear in J. Amer. Statist. Assoc.*

[22] Horn, R. and Johnson (1990). Matrix analysis. *Cambridge University Press*.

[23] Ingster, Y., Pouet, C. and Tsybakov, A. (2009). Classification of sparse high-dimensional vectors. *Phil. Trans. R. Soc. A*, **367**, 4427-4448.

[24] Ising, E. (1925). A contribution to the theory of ferromagnetism. *Z.Phys.*, **31**(1), 253-258.

[25] Ji, P. and Jin, J. (2011). UPS delivers optimal phase diagram in high dimensional variable selection. *To appear in Ann. Statist.*

[26] Li, C. and Li, H. (2011). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, **24**(9), 1175-1182.

[27] Meinsausen, N. and Buhlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Ann. Statist.*, **34**(3), 1436–1462.

[28] Meinsausen, N. and Buhlmann, P. (2010). Stability Selection (with discussion). *J. Roy. Soc. Statist. B*, **72**, 417-473.

[29] Meinsausen, N. and Rice, J. (2006). Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Statist.*, **34**(1), 323-393.

[30] Peng, J., Wang, P., Zhou, N. and Zhu, J. (2010). Partial correlation estimation by joint sparse regression model. *J. Amer. Statist. Assoc.*, **104**(486),735-746.

[31] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**(2), 461–464.

[32] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. B*, **58**(1), 267–288.

[33] Wainwright, M. (2009). Sharp threshold for high-dimensional and noisy recovery of sparsity using ℓ_1 constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory*, **52**(5), 2183–2202.

[34] Wasserman, L. and Roeder, K. (2009). High-dimensional variable selection. *Ann. Statist.*, **37**(5), 2178–2201.

[35] Ye, F. and Zhang, C-H. (2010). Rate minimaxity of the Lasso and Dantzig selection for the ℓ_q loss in ℓ_r balls. *J. Mach. Learn. Res.*, **11**, 3519-3540.

[36] Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, **38**, 894-942.

[37] Zhang, C.-H. and Zhang, T. (2011). A general theory of concave regularization for high dimensional sparse estimation problems. *arXiv*:1108.4988

[38] Zhang, T. (2011). Adaptive forward-backward greedy algorithm for learning sparse representations. *To appear in IEEE Trans. Inform. Theory.*

[39] Zhao, P. and Yu, B. (2006). On model selection consistency of LASSO. *J. Mach. Learn. Res.*, **7**, 2541–2567.

[40] Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, **101**(476), 1418–1429.