OPTIMAL PROCEDURES IN HIGH-DIMENSIONAL VARIABLE SELECTION

by

Qi Zhang

Bachelor of Science in Math and Applied Math, China Agricultural University, 2008

Submitted to the Graduate Faculty of
the Department of Statistics in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2013

UNIVERSITY OF PITTSBURGH

DEPARTMENT OF STATISTICS

This dissertation was presented

by

Qi Zhang

It was defended on

April 12, 2013

and approved by

Prof. Leon Gleser, Department of Statistics, University of Pittsburgh

Prof. Jiashun Jin, Department of Statistics, Carnegie Mellon University

Prof. Yu Cheng, Department of Statistics, University of Pittsburgh

Prof. Robert Krafty, Department of Statistics, University of Pittsburgh

Dissertation Advisors: Prof. Leon Gleser, Department of Statistics, University of Pittsburgh,

Prof. Jiashun Jin, Department of Statistics, Carnegie Mellon University

OPTIMAL PROCEDURES IN HIGH-DIMENSIONAL VARIABLE SELECTION

Qi Zhang, PhD

University of Pittsburgh, 2013

Motivated by the recent trend in “Big data”, we are interested in the case where both p, the number of variables, and n, the number of subjects are large, and probably $p \gg n$. When $p \gg n$, the signals are usually rare and weak, and the observation units are correlated in a complicated way. When the signals are rare and weak, it may be hard to recover them individually. In this thesis, we are interested in the problem of recovering the rare and weak signals with the assistance of correlation structure of the data.

We consider the helps from two types of correlation structures, the correlation structure of the observed units, and the dependency among the unobserved factors. In Chapter 2, in a setting of high dimensional linear regression, we study the variable selection problem when the observed predictors are correlated. In Chapter 3, we consider recovering the sparse mean vector of a Stein’s normal means model, where the elements of the unobserved mean vector are dependent through an Ising model. In each chapter, we study the optimality in variable selection, discover the non-optimality of the conventional methods such as the lasso, subset selection and hard thresholding, and propose *Screen and Clean* type of variable selection procedures which are optimal in terms of the Hamming distance. The theoretical findings is supported by the simulation results and applications.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

Every journey has its ups and downs, and one cannot conquer it without the help and support from his/her mentors, friends and family. The last five years have been the most amazing adventure in my life. Without the contribution and support from my advisors, my friends and family, it could not have been so wonderful, and I am deeply grateful to everybody.

Leon, thank you for providing me a flexible research environment and supportive mentoring. I have been benefited so much from your broad knowledge, deep understanding in statistics, endless trust, life experience, and even your sense of humor.

Jiashun, thank you for leading me into the world of high dimensional inference. Thank you for your frank and straightforward advises in research, and your kindly help in my career development. I have been becoming a better researcher under your guidance. Without you, it may never happen.

Yu and Rob, thank you for being in my committee and your help and advise in my research and job hunting.

Satish, I would like to thank you for your kindness and generousness, and giving me the right support when I need it most.

I would also like to thank all other faculty members in our department and my fellow students who have discussed with me about my research, their research or career development. I have learned a lot from the intellectual conversations of this kind.

Finally, I would like to thank all my friends and my family, who have been keeping me motivated to achieve what I have achieved today and more.

1.0 BACKGROUND AND INTRODUCTION

Nowadays, massive data are collected in many areas of sciences and business on a routine basis (e.g. genomics, cosmology, finance and imaging).

The primary feature of the massive data is that both p, the number of variables, and n, the number of samples are large, and $p \gg n$. For example, in cancer genetics, the sample size may be a few dozen or a few hundred, and for each patient, the expression values of tens of thousands of genes are measured, no mention other clinical measurements. There are two consequences of $p \gg n$, namely, the signal sparsity and the signal weakness. Signal sparsity means that for a specific purpose (e.g. regression, classification), only a small portion of the variables are relevant. For example, in the cancer genetics example, though tens of thousands of gene expressions are measured from each patient, maybe only a few hundreds of them are related to tumor metastasis. While signal sparsity is intuitive and widely accepted, signal weakness is a less understood notion. We say the signals are weak if they are barely separable from the noise, or hard to be detected individually. In the $p \gg n$ regime, the noise level is high, which makes the weak signals dominate in numbers. We remark that $p \gg n$ is not the only source of the signal weakness and there are signals that are intrinsically weak.

A second feature of the massive data set is that the observation units are correlated in a complicated way. For example, in the cancer genetics example, the observed gene expression values are usually correlated, and in imaging, the gray scale of one pixel is usually similar to those of its neighbors. Besides the direct correlations among the observed units, the dependency among the unobserved factors may also contribute to the observed complicated correlation structure. For example, in the cancer genetics example, if one gene is known to be related to metastasis, those genes that work in the same biological function with it are more likely to be related to metastasis as well. Such dependency information is usually in a form of networks and pathways, and known from common sense (e.g. [3]) or previous studies(e.g.[33], [36]).

Motivated by the above observations, we are primarily interested in recovering the rare and weak signals by the assistance of the correlation structure. The key idea is: though the weak signals cannot be detected individually, we may be able to estimate them jointly. We consider both the correlation structure of the observed units and the dependency structure of the unobserved factors.

Consider the following high dimensional linear regression problem as an illustrative example

$$Y = X\beta + z, \quad and \quad z \sim N(0, I_n).$$

Here $X = X_{n\times p}$ where $p \gg n$. β is unknown, but presumably sparse, and the signals may be very weak (its non-zero elements are only moderately large in magnitude). Our goal is to find the locations of the non-zero elements of β. In order to recover the weak signals as accurately as possible, we use the observed correlation structure of the predictors or the dependency among the elements of the unobserved factor β for assistance.

In Chapter 2, we consider the former in the setting of linear regression, and in Chapter 3, we consider the later in a more general setting. In each case, we develop an optimal procedure to perform the task, and also present intensive theoretical discussions on the situations where the conventional methods (e.g. lasso [43], subset selection[1, 42], hard thresholding) are not optimal.

2.0 EXPLOIT THE CORRELATION AMONG THE PREDICTORS

2.1 INTRODUCTION

Consider a linear regression model

$$Y = X\beta + \sigma z, \qquad z \sim N(0, I_n), \tag{2.1}$$

where the design matrix $X = X_{n,p}$ has n rows and p columns. Throughout this chapter, 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 chapter is to identify such nonzero coordinates (i.e., variable selection).

2.1.1 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 chapter, 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 chapter, 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. The goal is to study the rate of the minimax Hamming distance study the optimality of variable selection procedures.

2.1.2 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 [14, 15, 28, 39]. However, in numerous applications, there are relationships between predictors which are important to consider.

In this chapter, 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 [12, 13, 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 [40]. Frequently, the concentration matrix $\Omega = \Sigma^{-1}$ is believed to be sparse, and can be effectively estimated in some cases (e.g. [5, 6]), or can be assumed as known in others, with the so-called “data about data” available [34]. 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 [31] and Factor Analysis [26].

Well-known approaches to variable selection include subset selection, the lasso, SCAD, MC+, greedy search and more [1, 8, 16, 17, 42, 43, 50, 52, 53, 54]. 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 2.3 for detailed discussion. On the other hand, there are two-stage Screen and Clean methods that are optimal, e.g. *Graphlet Screening* (GS). 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. For for more details, see Section 2.2 and the paper [32].

2.1.3 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., [7]). Fix parameters $\epsilon \in (0, 1)$ and $\tau > 0$. Let $b = (b_1, \ldots, b_p)'$ be the $p \times 1$ random vector where

$$b_i \overset{iid}{\sim} \text{Bernoulli}(\epsilon). \tag{2.2}$$

We model the signal vector β as

$$\beta = b \circ \mu, \tag{2.3}$$

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 $(x \circ y)_i = x_i y_i$, $1 \leq i \leq p$), and $\mu \in \Theta_p(\tau)$ with

$$\Theta_p(\tau) = \{\mu \in \mathbb{R}^p : |\mu_i| \geq \tau, 1 \leq i \leq p\}. \tag{2.4}$$

In later sections, we may further restrict μ to a subset of $\Theta_p(\tau)$; see (2.9).

Definition 2.1.1. *We call (2.2)-(2.4) 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 2.1.4 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

$$X = [x_1, x_2, \dots, x_p] = [X_1, X_2, \dots, X_n]' \tag{2.5}$$

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

$$\Omega^* = (\Omega^*(i,j))_{p \times p}, \qquad \Omega^*(i,j) = G(i,j) 1\{|G(i,j)| \geq \delta\}, \tag{2.6}$$

as a *regularized Gram matrix.*

Definition 2.1.2. *The GOSD is the graph $\mathcal{G}^* = (V, E)$, where $V = \{1, 2, \dots, p\}$ 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 2.1.3. *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 chapter.

Definition 2.1.4. *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 in parallel, 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 A.1.1.

Lemma 2.1.1. *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 2.1.1 holds with overwhelming probability. We mention that Lemma 2.1.1 is not tied to Model (2.2)-(2.4) 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 [29]. 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 2.1.1 in Section A.1.1.

2.1.4 Asymptotic Rare and Weak model for regression with random design

We continue our discussion with the Rare and Weak model RW(ϵ, τ, μ) 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

$$\epsilon = \epsilon_p = p^{-\vartheta}. \tag{2.7}$$

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

$$\tau = \tau_p = \sigma\sqrt{2r\log(p)}. \tag{2.8}$$

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

$$\Theta_p^*(\tau_p, a) = \{\mu \in \Theta_p(\tau_p) : |\mu_i| \leq a\tau_p, i = 1, 2, \ldots, p\}, \tag{2.9}$$

and the parameter a is unknown.

Definition 2.1.5. *We call model (2.2)-(2.4) and (2.7)-(2.9) 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:

$$X_i \overset{iid}{\sim} N(0, \frac{1}{n}\Omega). \tag{2.10}$$

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 [11], and other application areas.

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

$$n = n_p = p^{\kappa}. \tag{2.11}$$

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

$$\kappa > (1 - \vartheta), \tag{2.12}$$

so that $n_p \gg p\epsilon_p$. Note $p\epsilon_p$ is approximately the total number of signals. Condition (2.12) is almost necessary for successful variable selection [12, 13].

Definition 2.1.6. *We call Model (2.10)-(2.12) the Random Design model* $\mathrm{RD}(\vartheta, \kappa, \Omega)$.

Come back to (2.9). 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

$$\lambda_k^*(\Omega) = \min\{\lambda(A) : A \text{ is a } k \times k \text{ principle submatrix of } \Omega\}. \tag{2.13}$$

At the same time, fixing a constant $c_0 > 0$, let (ϑ, r) be as in (2.7) and (2.8), respectively, let m be as in the GS-step, and let g be the smallest integer such that

$$g \geq \max\{m, (\vartheta + r)^2/(2r)\}. \tag{2.14}$$

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

$$\big(\theta_*^{(0)}(V_0, V_1), \theta_*^{(1)}(V_0, V_1)\big) = \mathrm{argmax}\{(\theta^{(1)} - \theta^{(0)})'\Omega(\theta^{(1)} - \theta^{(0)})\}, \tag{2.15}$$

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 2.1.2. *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 chapter, unless stated otherwise, we assume

$$\Omega \in \mathcal{M}_p(c_0, g), \qquad a > a_g^*(\Omega). \tag{2.16}$$

In Section 2.2, we further restrict Ω to a subset of $\mathcal{M}_p(c_0, g)$ to foster graph sparsity. Condition (2.16) 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 (2.15), 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 (2.15) has exactly the same solution, which does not depend on a. This explains (2.16).

2.1.5 Measure the performance of variable selection: Hamming distance and phase diagram

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

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

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 (2.2) and (2.10). Finally, the minimax Hamming distance is

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

In the above definitions, $\mathrm{sgn}(x) = 0, 1, -1$ 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. For a lower bound of this minimax Hamming distance, see [32].

For given (Ω, κ, a), the minimax Hamming distance depends on the tuning parameters (ϑ, r), which are associated with the signal sparsity and the signal strength, respectively. 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 $\text{Hamm}_p^*(\vartheta, \kappa, r, a, \Omega) = 1$) 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 $1 \ll \text{Hamm}_p^*(\vartheta, \kappa, r, a, \Omega) \ll p^{1-\vartheta}$, 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 $\text{Hamm}_p^*(\vartheta, \kappa, r, a, \Omega) = o(1)$, and it is possible to exactly recover all signals with overwhelming probability.

We are interested in study whether certain variable selection procedures can achieve exact recovery when the optimal procedure can. If a procedure is optimal, its exact recovery region should be as large as the *Region of Exact Recovery.* We are also interested in study the rate of Hamming distance of some procedures in the *Region of Almost Full Recovery.* More exactly, we further partition this region into two sub-regions, the optimal region where the rate of this procedure is optimal, and the non-optimal region, where it is not optimal. We continue this discussion in Section 2.3.

2.2 GRAPHLET SCREENING AND ITS OPTIMALITY

The decomposability discussed in Section 2.1.3 invites the following two-stage variable selection procedure, which we call the *Graphlet Screening (GS).* We describe the procedure and state its optimality result here, and please see the author and the collaborators' paper [32] for details.

Conceptually, *Graphlet Screening* 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 [47, 18] in that they have a screening and a cleaning stage, but is more sophisticated in nature.

We now describe two steps in details. Recalling (2.5), we have the following definition.

Definition 2.2.1. *For X in Model (2.1) and any $\mathcal{I} \subset \{1, 2, ..., p\}$, 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 (2.6) 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

$$T(Y, \hat{D}, \hat{F}) = \|P^{\mathcal{I}_0} Y\|^2 - \|P^{\hat{F}} Y\|^2 > t(\hat{D}, \hat{F}), \tag{2.17}$$

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 2.1 for a recap of the procedure.

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

$$\mathcal{Q} \equiv \{t(\hat{D}, \hat{F}) : (\hat{D}, \hat{F}) \text{ are as defined in (2.17)}\}.$$

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 [32].

The computational cost of the GS-step hinges on the sparsity of $\mathcal{G}^*$. In Section 2.1.4, 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 [23].

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 in parallel in the GC-step.

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

Definition 2.2.2. *For a $p \times m$ matrix X and $\mathcal{I}_0 \subset \{1, 2, \ldots, p\}$ and $\mathcal{J}_0 \subset \{1, 2, \ldots, m\}$, $X^{\mathcal{I}_0, \mathcal{J}_0}$ denotes the $|\mathcal{I}_0| \times |\mathcal{J}_0|$ 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 $1 \leq j \leq p$, 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:

$$\|P^{\mathcal{I}_0}(Y - X^{*, \mathcal{I}_0}\xi)\|^2 + (u^{gs})^2 \|\xi\|_0. \tag{2.18}$$

Here, ξ is an $|\mathcal{I}_0| \times 1$ vector each nonzero coordinate of which $\geq v^{gs}$ 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, 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, $|\mathcal{I}_0| \leq \ell_0$ for any $\mathcal{I}_0 \lhd \mathcal{U}_p^*$. As a result, the computational cost of the *GC*-step is no greater than $|\mathcal{U}_p^*| \times 2^{\ell_0}$, which is moderate.

For moderately large p, we also propose an iterative Graphlet Screening in the spirit of the refined UPS[31], except with a different initial estimate. See Section 2.4 for details.

Regarding to the performance, it turns out that under mild conditions, Graphlet Screening is optimal in terms of the Hamming distance given the tuning parameters are well-set [32] . The most important regularity condition is the sparsity of Ω, which is presented as the following,

$$\mathcal{M}_p^*(\gamma, c_0, g, A) = \Big\{\Omega \in \mathcal{M}_p(c_0, g) : \sum_{j=1}^{p} |\Omega(i, j)|^\gamma \leq A, \, 1 \leq i \leq p\Big\}. \tag{2.19}$$

where $\gamma \in (0, 1)$ and $A > 0$.

GS-step:	List $\mathcal{G}^*$-connected submodels $\mathcal{I}_{0,k}$ with $\lvert\mathcal{I}_{0,1}\rvert \leq \lvert\mathcal{I}_{0,2}\rvert \leq \cdots \leq m$
	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 (2.17)
	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 (2.18) 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

Table 2.1: Graphet Screening Algorithm

2.3 A BLOCKWISE EXAMPLE: NON-OPTIMAILITY OF SUBSET SELECTION AND THE LASSO AND THE OPTIMALITY OF GRAPHLET SCREENING

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:

$$\frac{1}{2}\|Y - X\beta\|^2 + \frac{1}{2}(\lambda_{ss})^2\|\beta\|_0, \tag{2.20}$$

where $\|\beta\|_q$ denotes the L^q-norm, $q \geq 0$, and $\lambda_{ss} > 0$ is a tuning parameter. The AIC, BIC, and RIC are methods of this type [1, 42, 19]. 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 [9, 43], SCAD [17], MC+ [50], and Dantzig selector [8]. Take the lasso for example. The method selects variables by minimizing the following functional:

$$\frac{1}{2}\|Y - X\beta\|^2 + \lambda_{lasso}\|\beta\|_1, \tag{2.21}$$

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

$$G(i,j) = 1\{i = j\} + h_0 \cdot 1\{|j-i| = 1,\ max(i,j) \text{ is even}\},\ \ |h_0| < 1,\ 1 \leq i, j \leq p. \tag{2.22}$$

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 ARW(ϑ, r, a, μ), 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 $\geq \tau_p$ (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 2.1.4 since the settings are slightly different.

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

$$\rho_{lasso}^{(3)}(\vartheta, r, h_0) = \left\{(2|h_0|)^{-1}[(1-h_0^2)\sqrt{r} - \sqrt{(1-h_0^2)(1-|h_0|)^2 r - 4|h_0|(1-|h_0|)\vartheta}]\right\}^2,$$

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 A.1.2.

Theorem 2.3.1. *Fix $\vartheta \in (0,1)$ and $r > 0$ such that $r > \vartheta$. If G satisfies (2.22), then*

$$\rho_{gs}(\vartheta, r, h_0) = \min\Big\{\frac{(\vartheta+r)^2}{4r}, \vartheta + \frac{(1-|h_0|)}{2}r, 2\vartheta + \frac{\{[(1-h_0^2)r - \vartheta]_+\}^2}{4(1-h_0^2)r}\Big\}, \tag{2.23}$$

$$\rho_{ss}(\vartheta, r, h_0) = \min\Big\{\frac{(\vartheta+r)^2}{4r}, \vartheta + \frac{(1-|h_0|)}{2}r, \rho^{(1)}_{ss}(\vartheta, r, h_0), \rho^{(2)}_{ss}(\vartheta, r, h_0)\Big\}, \tag{2.24}$$

and

$$\rho_{lasso}(\vartheta, r, h_0) = \min\Big\{\frac{(\vartheta+r)^2}{4r}, \vartheta + \frac{(1-|h_0|)r}{2(1+\sqrt{1-h_0^2})}, \rho^{(1)}_{lasso}(\vartheta, r, h_0), \rho^{(2)}_{lasso}(\vartheta, r, h_0)\Big\}. \tag{2.25}$$

$\vartheta/r/h_0$	.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 2.2: The exponents $\rho_\star(\vartheta, r, h_0)$ in Theorem 2.3.1, where $\star = gs, ss, lasso$.

It can be shown that $\rho_{gs}(\vartheta, r, h_0) \geq \rho_{ss}(\vartheta, r, h_0) \geq \rho_{lasso}(\vartheta, r, h_0)$, 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 2.2 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.

Similar to that in Section 2.1.5, 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 2.1.

Subset selection and the lasso are rate non-optimal for they are so-called *one-step* or *non-adaptive* methods [31], 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.

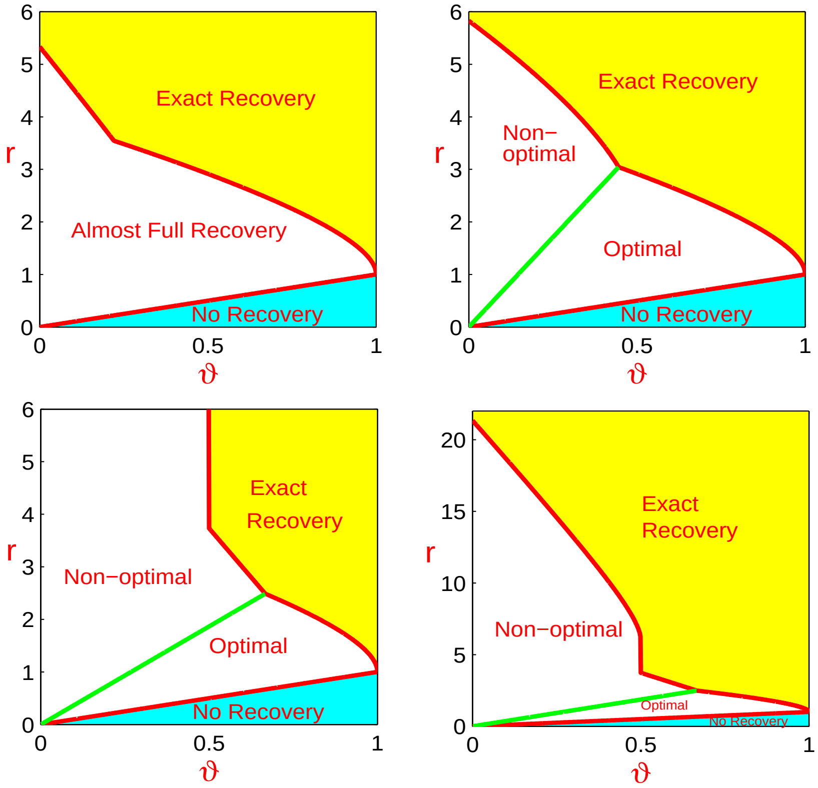


Figure 2.1: Phase diagrams for Graphlet Screening (top left), subset selection (top right), and the lasso (bottom; zoom-in on the left and zoom-out on the right), where $h_0 = 0.5$.

2.4 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 $b_i \overset{iid}{\sim}$ Bernoulli(ϵ_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 [32]). 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 [22] 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 $|\mu_i| \overset{iid}{\sim} 0.8\nu_{\tau_p} + 0.2h$, 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 the optimality conditions of GS. The average Hamming errors for both procedures across 40 repetitions are tabulated in Table 2.3.

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 2.3: 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 2a-2c. For each sub-experiment, the average Hamming errors of 40 repetitions are reported in Figure 2.2.

In Experiment 2a, 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).

In Experiment $2b$, 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 $(.05, -.05)'$.

In Experiment $2c$, 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 results suggest that Graphlet Screening is consistently better than the lasso.

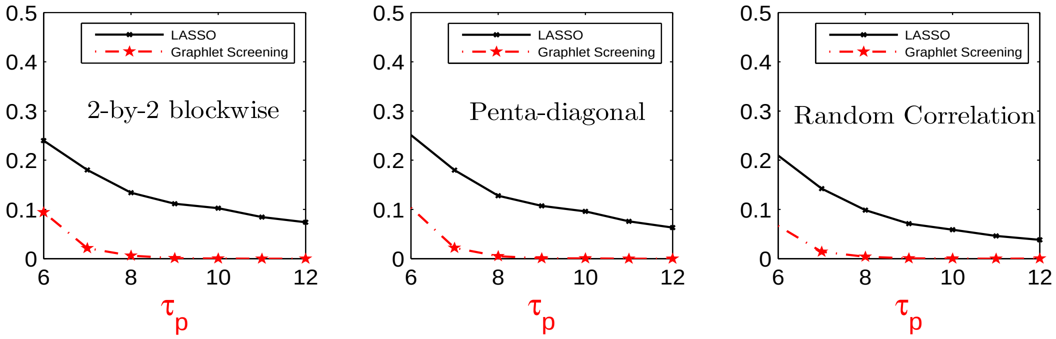


Figure 2.2: x-axis: τ_p. y-axis: ratios between the average Hamming errors and $p\epsilon_p$ (from left to right: Experiment 2a, 2b, and 2c).

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.2.4.

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 2.4: 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 [32]), 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 $4a$, 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 [32]). 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 2.5, 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 2.5: Ratios between the average Hamming errors of Graphlet Screening and $p\epsilon_p$ (Experiment $4a$).

In Experiment $4b$, we use the same settings as in Experiment $4a$, 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 $4c$, we use the same settings as in Experiment $4a$, 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 $4b$-$4c$, 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 2.6, which suggest that the misspecifications have little effect as long as r^*/r and ϑ^*/ϑ are reasonably close to 1.

In Experiment $4d$, we re-examine the misspecification issue with a random design. We use the same settings as in Experiment $4b$ and Experiment $4c$, except for (a) while we use the same Ω as in Experiment $4b$, the design matrix X are generated according to the random design model as in Experiment $2b$, and (b) we only investigate for the case of $r = 2$ and $\vartheta \in \{0.35, 0.6\}$. The results are summarized in Table 2.7, which is consistent with the results in $4b$-$4c$.

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

2.5 DISCUSSION AND FUTURE WORK

In this chapter, we focus on the regime where the Gram matrix is sparse. A further extension of the GS methodology would be the cases where the Gram matrix is not sparse, but sparsifiable. One such case is when there are a few hot hubs in GOSD. Once we find these hubs, the conditional covariance structure of the other variables may be sparse or blockwise. Another interesting direction of future research is the extension of the GS methodology to more general models such as logistic regression.

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

3.0 EXPLOIT THE DEPENDENCY AMONG THE SIGNALS

3.1 INTRODUCTION

Consider a Stein's normal means model

$$Y = \beta + z, \tag{3.1}$$

where Y, β and z are $p \times 1$ vectors, $z \sim N(0, I_p)$. Here $\beta = (\beta_1, \beta_2, ..., \beta_p)'$, if $\beta_i \neq 0$, we call it a signal, otherwise, a noise. The mean vector β is unknown, but presumably sparse, which means most of the coordinates of β are zeroes. We define a length p vector $b = (b_1, \ldots, b_p)'$ as $b_i = 1\{\beta_i \neq 0\}$ for $i = 1, \ldots, p$. Then b is the indicator of the signals. The main goal of this paper is to estimate b, or equivalently, to find all the non-zero coordinates of β.

Though (3.1) is very simple, it has profound implications in many statistical problems.

- Linear regression. Consider $Y = X\beta + z$, where $z \sim N(0, I_n)$, X is the $n \times p$ design matrix, and the Gram matrix $X'X \approx I_p$. Let $\tilde{Y} = X'Y$, then approximately we have $\tilde{Y} = \beta + \tilde{z}$, where $\tilde{z} \sim N(0, I_p)$.
- Classification. Consider a two class classification problem where the training set has n subjects and p variables and $X = [X_1, \ldots, X_p]$ is the $n \times p$ feature matrix. We are interested in building an effective classifier with a small portion of the variables. We can define an length p vector Y where for $i = 1, \ldots, p$, Y_i is the standardized difference of the sample means of X_i between the two classes. When both p and n are large (possibly $p \gg n$), approximately Y satisfies (3.1). Intuitively, if $\beta_i \neq 0$, then X_i should be included to the classifier. If $\beta_i = 0$, X_i may not be necessary for classification.

3.1.1 Structured sparsity in the paradigm of rare and weak signals

We focus on the regime where the signals are both rare and weak. The rarity (or sparsity) is a well-accepted concept. But signal weakness is a much neglected notion until recently. Actually, similar to the sparsity, signal weakness is also a natural phenomenon that usually arise from the issue of $p \gg n$. Take the above high dimensional classification problem for example, when $p \gg n$, some of the standardized two-sample means could be very noisy. In our model, signal weakness implies that the magnitudes of the non-zeros of β are individually small or moderately large, and barely separable with the noise.

Most of the previous work on the rare and weak signals focus on the completely unstructured case and do not take in interaction among the features into account. Recently, in the setting linear regression and classification, some works [21, 31, 32, 35] start to consider the correlation structure of the features. While these works are successful in their own regimes, what is still neglected is the dependency structure among the signals themselves. In this paper, we consider the variable selection problem when the signals depend on each other significantly.

3.1.2 Ising model: modeling the signal dependency

Let $\mathcal{G}$ be an undirected graph of dependency with p nodes, and each node is conditional independent to others given its neighbors. In our model (3.1), the signals depend on each other in a sense that if we know that $\beta_i \neq 0$, then $\beta_{j_1}, \ldots, \beta_{n_i}$ are more likely to be non-zeros, where $\{j_1, \ldots, j_{n_i}\}$ are the neighbors of node i in graph $\mathcal{G}$. We also assume the dependency structure is sparse in a sense that the graph $\mathcal{G}$ is sparse or equivalently, for each node i, the size of its neighborhood n_i is very small. In many application areas, the dependency among the sparse signals naturally exists, and the dependency structure is sparse and known, though the dependency strength may be unknown.

- *Image analysis:* In image analysis, it is of interest to recover the objects in a “dirty picture” [3]. Consider a toy example where there are only two colors, black and white. Assume that most area of the picture is white, then the signals (black pixels) are sparse. In this problem, if we know one pixel is black, the pixels near it are more likely to be black. The simplest graphical presentation of the above observation would be a 2-d lattice, where the pixels are treated as nodes in the graph, and for each pixel, only the four pixels in its direct neighborhood are connected with it in the lattice graph. This graph is sparse.

- *Network-based analysis in biology:* In biology, genes and proteins interact in a complicated way. In the past two decades, various network databases have been built to model these interaction patterns, e.g genetic regulatory network, protein-protein interaction network and biological pathways. In most of the cases, each gene or protein only interact with a small portion of other genes or proteins, hence the network in biology is usually sparse. In data analysis problems such as finding differentially expressed genes(DE) between the tumor patients and the general population, it is believed that only a small portion of the human genes are DE, hence DE is sparse. In this problem, it is intuitive to believe that if we know one gene is related to cancer, then those genes that interact with it are more likely to be also related to cancer. Thus, the existing biological network information may be very useful in such problems, and can be integrated into the analysis as the known "data about data" [34, 48].

An immediate question would be how to incorporate such prior information on signal dependency into data analysis.

We tackle this problem from a generative model perspective. Recall that b is the indicator of the locations of the signals, and as before, our primary goal is to estimate b. We model the dependency among the signals by modeling b as a sample from an Ising model. So that Y satisfies a hidden Ising model.

The Ising model is originally a mathematical model of ferromagnetism in statistical mechanics [29]. Nowadays, we use Ising model to denote a wide class of Binary Markov Random Field with a distribution

$$P(b = \ell) = \frac{1}{Z_p(\eta, \Omega)} \exp\Big(-\eta \ell' \Omega \ell\Big), \tag{3.2}$$

where Ω is the weighted adjacency matrix of the encoded undirected graph $\mathcal{G}$, and its diagonal elements are $1's$. $Z_p(\eta, \Omega)$ is the normalizing constant. We denote it as $b \sim Ising(\eta, \Omega)$. We remark that when $\Omega = I_p$,

$$b_i \overset{iid}{\sim} Bernoulli(\frac{1}{1+e^{\eta}}); \tag{3.3}$$

Hence the Ising model can be viewed as a multivariate bernoulli distribution that takes the pairwise interactions into account.

Roughly speaking, in (3.2), Ω controls the dependency strength among the signals and for given Ω, the parameter η controls the number of signals. In our setting, Ω is sparse and the locations of the non-zeros are known, but the values of the elements of Ω are unknown.

The signals generated by an Ising model live in disjoint signal clusters which we will formally define In Section 3.2.1. In Section 3.2.1, we show that under mild constraints, (3.2) is an appropriate generative model for the sparse signals with sparse dependency structure. It generates sparse signals living in short signal clusters in graph $\mathcal{G}$ that are isolated to each other, and the maximal cluster length has a constant upper bound except a negligible probability. We refer to such models as sparse Ising model.

3.1.3 Objective of the chapter

We are interested in the optimal variable selection for (3.1) where the signal is weak and b is an interdependent signal generated by a sparse Ising model. The objective of the paper is three-fold.

- To study the Ising model, and specify a class of sparse Ising model that appropriately models the sparse interdependent signals.
- To develop a variable selection procedure that is accurate and computationally affordable, and investigate its application in various statistical problems.
- To study the variable selection problem with interdependent signals and the rate of minimax Hamming distance, to characterize the difficulty level of the variable selection problem and the performance of the procedures by Hamming distance, and to show the optimality of the above procedure in the rate of convergence in Hamming distance and the phase diagram.

In the regime of the rare and weak signal, exact recovery of b is usually impossible. Instead of probability of exact support recovery or "oracle property", Hamming distance is a more appropriate measure for the performances of variable selection procedures. Hamming distance is the expected number of errors in estimating the support b. Two related notions are Hamming ratio and the phase diagram. The phase diagram is a graphical criterion for measuring the performance of variable selection procedures, which is appropriate in our rare and weak signal paradigm. Hamming ratio is the ratio of the Hamming distance and the expected number of signals, which we use as the measurement of the variable selection procedures in our numeric study.

3.1.4 Non-optimality of hard thresholding

When b is a sample from a bernoulli model, the signal is unstructured. In this case, as a corollary of the results in [31], the element-wise hard thresholding can separate the signal from the noise optimally in the Hamming distance, given that the uniform threshold is optimally set. The the optimal rate of the hard thresholding is the same as the rate of separating a signal singleton from pure noise. Thus, the success of hard thresholding in the unstructured case can be attributed to the fact that the signals are independent to each other. Generally speaking, hard thresholding could be optimal if most signals live as singletons.

When the signals are generated from a sparse Ising model, there are two different cases. In the first case, the signals are moderately sparse and the dependency strength is strong so that most signals live in short signal clusters with more than one nodes, e.g. pairs, triplets. Thus hard thresholding is not optimal since it ignores the dependency structure. In the second case, the signals are so sparse or the dependency strength is so weak such that most signals are still singletons. Thus hard thresholding may be still optimal in this case.

However, when the signal dependency is very weak, it may not be necessary to incorporate such information into analysis at all, no matter what signal model we intend to use. Only when the signal dependency is strong, a model for interdependent signals such as our sparse Ising model becomes important or even crucial in data analysis. Therefore, as long as the signal dependency is a major concern in data analysis, hard thresholding is not the optimal choice for variable selection.

3.1.5 Intractability and decomposability of the likelihood of the sparse Ising model

Viewing b as hidden states, maximizing the conditional probability $P(b|Y)$ is the "fundamentally correct" way to estimate b. However, even when the parameters of the Ising model are known, maximizing this probability may involve comparing all $b \in \{0,1\}^p$, which is intractable in computation.

Though this problem is seemingly intractable, we can still maximize $P(b|Y)$ approximately. Recall that the sparse Ising model generates sparse signal living in short clusters that are isolated to each other. By the definition of the signals clusters, there is no other signal in the neighborhood of each cluster. Thus if we know the locations of these clusters, these clusters are independent to each other. Then the signal graph of the sparse Ising model is decomposable, and so is the likelihood of the sparse Ising model.

To see this, let $S_i, i = 1, \dots, M$ be a collection of the isolated connected subgraphs of $\mathcal{G}$ that contains all the signal clusters, and for any subgraph B, let $N(B)$ be the neighborhood of B in the graph $\mathcal{G}$. There is no signal outside $\cup_{i=1}^M S_i$, and we have

$$P(b|Y) \propto P(b_j = 0 \text{ for } j \notin \cup_{i=1}^M S_i|Y) \prod_{i=1}^M P(b_{S_i}|Y_{S_i}, b_j = 0 \text{ for } j \in N(S_i)) \tag{3.4}$$

Then we can maximize $P(b|Y)$ by maximizing $P(b_{S_i}|Y_{S_i}, b_j = 0 \text{ for } j \in N(S_i))$ separately for $i = 1, \dots, M$.

Example 1. We illustrate the decomposability with a toy example. In model (3.1)-(3.2), let $p = 4$, and Ω is a diagonal matrix whose non-zero off-diagonal elements are $-1/2$. Then,

$$P(b = \ell) \propto \exp\left[-\eta(\sum_{i=1}^4 \ell_i - \sum_{i=1}^3 \ell_i \ell_{i+1})\right].$$

If we know that $b_2 = 0$, we have

$$P(b = \ell|b_2 = 0) \propto \exp\left(-\eta\ell_1\right) \cdot \exp\left[-\eta(\ell_3 + \ell_4 - \ell_3\ell_4)\right].$$

On the other hand, we know

$$P(b = \ell|Y) \propto P(b = \ell) \prod_{i=1}^4 p(Y_i|b_i)$$

Thus

$$\begin{aligned} P(b = \ell | b_2 = 0, Y) \;&\propto\; p(Y_2|b_2 = 0) \cdot [p(Y_1|b_1 = \ell_1) \exp\left(-\eta\ell_1\right)] \\ &\quad \cdot \left\{p(Y_3|b_3 = \ell_3)p(Y_4|b_4 = \ell_4) \cdot \exp\left[-\eta(\ell_3 + \ell_4 - \ell_3\ell_4)\right]\right\}. \end{aligned}$$

Simplifying the above yields

$$P(b = \ell|b_2 = 0, Y) \propto P(b_1 = \ell_1|b_2 = 0, Y_1) \cdot P(b_3 = \ell_3, b_4 = \ell_4|b_2 = 0, Y_3, Y_4).$$

The decomposability as illustrated in the above example naturally introduces a Screen and Clean procedure for variable selection which we refer to as Graphical Model Assisted Selection(GMAS). GMAS approximately maximizes $P(b|Y)$ with moderate computational cost, and the result is still optimal in a broad context. In the following, we provide a non-technical description of the procedure and a discussion on our key ideas and major innovations.

3.1.6 Our proposal: Graphical Model Assisted Selection (GMAS), a screen and clean approach

Graphical Model Assisted Selection (GMAS) is a Screen and Clean procedure that contains a screening stage and a cleaning stage. In the screening stage, the signal clusters are roughly located; and in the cleaning stage, the isolated pieces of the survivors are investigated separately in more details. The key idea of GMAS is to take advantage of the decomposability of the likelihood of the sparse Ising model. The problem of maximizing the intractable high dimensional data likelihood can be solved by only considering a collection low dimensional tractable likelihoods separately. Because of the this, GMAS enjoys a two-fold advantage: theoretical optimality and modest computational complexity. In the following, we provide a non-technical description of the procedure and a discussion on our key ideas and major innovations.

We first consider the screening stage. Recall that the sparse Ising model generates sparse signals living in isolated short signal clusters. The goal of the screening stage of GMAS is to roughly locate these clusters. It needs to be done in a way that most of the signals are retained while the majority of the noise is removed. Our strategy is m-phase screening. Let $\mathcal{U}$ be the set of retained nodes after screening, and initially we have $\mathcal{U} = \emptyset$. The m-phase screening works as follows.

- In Phase 1, we test whether each node can be a signal individually. For $j = 1, \ldots, p$, we perform the test
$$H_0 : \beta_j = 0 \text{ against } H_1 : \beta_j \neq 0.$$
We update $\mathcal{U}$ by adding node j if H_{0j} is rejected.
- In Phase 2 we test whether each pair of nodes can be a signal pair. For each edge (i, j) in graph $\mathcal{G}$, if $\{i, j\} \subset \mathcal{U}$ even before the test, no action will be taken. If none of $\{i, j\}$ is in $\mathcal{U}$, we perform the test
$$H_0 : \beta_i = \beta_j = 0 \text{ against } H_1 : \beta_i \neq 0, \beta_j \neq 0.$$
If only one of (i, j) is in $\mathcal{U}$, without losing generality, let $i \in \mathcal{U}$, we perform the test
$$H_0 : \beta_j = 0 \text{ against } H_1 : \beta_j \neq 0.$$
In either case, if H_0 is rejected, we update $\mathcal{U}$ by adding $\{i, j\}$.
- In general, in Phase k, $1 \leq k \leq m$, let B be a connected subgraph of $\mathcal{G}$ with k nodes, and we want to test whether all nodes in B are signals or there is no new signals in B other than those that are already in $\mathcal{U}$ before the test. In detail, let $E = B \cap \mathcal{U}$ and $D = B \setminus E$. If $D \neq \emptyset$, we perform the test
$$H_0 : \beta_k \neq 0 \text{ for } k \in E, \text{ against } H_1 : \beta_k \neq 0 \text{ for } k \in B.$$

For each test, we only use the conditional distribution of (b_B, Y_B) given that there is no other signal in the neighborhood of B in graph $\mathcal{G}$. Then we use the likelihood ratio test of this conditional likelihood as the test statistic. For each likelihood ratio test, we set the rejection rule in a way that the Type II error is very low such that the expected number of false negative in all tests is negligible (e.g only $o(1)$).

We now consider the cleaning step. It can be shown that the survivors after the screening stage split into many separated small-size connected subgraphs("islands") in graph $\mathcal{G}$, and each connected subgraph may contain one or more signal clusters. Since the excluded nodes contain almost no signal, by the decomposability of the likelihood of the sparse Ising model, we can investigate each connected subgraph separately to remove the false positives.

The success of GMAS can be attributed to the decomposability of the signal graph. In more detail, this decomposability answers the following four questions.

How does GMAS avoid the problems caused by the intractability of the likelihood of the sparse Ising model? In the screening stage, each test only depends on the conditional likelihood, $L(b_B|Y_B, b_j = 0 \text{ for } j \in N(B))$. If B itself is a signal cluster, the nodes in B can be selected except a negligible probability. In the cleaning stage there is essentially no signal in the excluded nodes. Therefore, by (3.4), the isolated islands of the survivors can be analyzed separately, and for an island S, only the conditional likelihood $L(b_S|Y_S, b_j = 0 \text{ for } j \in N(S))$ is involved. In both stages, we only use the conditional likelihoods of a few coordinates of b, instead of referring to the intractable likelihood.

How many phases does the screening stage of GMAS need? In the m-phase screening, we can include all the signal clusters with no more than m signals. Hence the larger m is, the less false negatives are made. On the other hand, a large m also bring larger computational cost. We wish m to be as small as possible, yet still large enough to keep most signals. Except a negligible probability, the missed signals must be in a signal cluster that is longer than m. Let $\mathcal{G}_i^*$, for $i = 1, \ldots, M$ be all the signal clusters, m need to be set such that

$$P(\max_i(|\mathcal{G}_i^*|) > m) = o(p^{-1}).$$

m can be set as the maximal cluster size as in Theorem 3.2.1.

Why is the computational cost of GMAS moderate? The computational complexity of GMAS contains two parts: screening stage and the cleaning stage. The computational complexity of the screening stage depends on the number of connected subgraphs of $\mathcal{G}$ screened. Now suppose that the maximal degree of $\mathcal{G}$ is K. A neat result in graph theory [23] says that for any integer ℓ, the sparse graph $\mathcal{G}$ at most has $p(eK)^\ell$ connected subgraphs with exactly ℓ nodes. Since K is finite or a multi-log(p) term, and m is finite, the computational complexity of the screening stage is moderate. It could be shown that the survivors after the screening stage splits int isolated "islands", and the size of each is finite. In the cleaning stage, we clean these “islands” separately, and at most there are p islands. Thus the computational complexity of the cleaning stage is also moderate.

How does GMAS achieve the optimal rate of the convergence in the Hamming distance? The Hamming distance is the expected total number of incorrect selection. We construct its lower bound by finding the configuration of the sparse Ising model that is the worst possible one for the greedy search. The signal graph $\mathcal{G}^*$ of the worst possible configuration is still decomposable. Hence this high-dimensional lower bound of the Hamming distance can be matched nearly sharply by only doing the local low-dimensional analysis, as GMAS does.

3.1.7 Contents

In Section 3.2, we state our main results. Section 3 details the properties of GMAS and concludes with the proof of Theory 3.2.4. In Section 3.4, we provides a small scale simulation study. In Section 3.5, we illustrate the application of GMAS by two real data examples in microarray: one is about the identification of the differentially expressed genes in the yeast environmental stress data; and the other is about the variable selection in the human tumor classification.

3.2 MAIN RESULTS

To recap, we consider a Stein's normal means model

$$Y = \beta + z, \quad z \sim N(0, I_p). \tag{3.5}$$

We model β as

$$\beta = u \circ b$$

where b is the indicator vector of being a signal or noise. We model $b \sim Ising(\eta, \Omega)$ and

$$u \in \Theta_p(\tau_p) \text{ where } \Theta_p(\tau_p) = \{u : |u_i| \geq \tau_p, i = 1, \ldots, p\} \text{ and } \tau_p = \sqrt{2r\log(p)}.$$

We are interested finding the non-zero elements of β when the signals are rare, weak and interdependent. In Section 3.2.1, we first specify a class of $Ising(\eta, \Omega)$ as appropriate model for such signals. In the rest of this paper, we study the variable selection problem for (3.5).

3.2.1 Sparse Ising model

As discussed in Section 3.1.1 and 3.1.2, we are interested in modeling the length p vectors β that have the following four properties:

- *Signal sparsity*: the expected number of signals generated by this model is $o(p)$.
- *Graph sparsity*: each coordinate of β only depends on a few other coordinates of β.
- *Clustering effect*: if a signal is found the signals attract each other and tend to live in clusters.
- *Finite cluster size*: with probability $1 - o(p^{-1})$, the lengths of all signal clusters are finite.

The goal of this section is thus to specify a class of Ising model that generates signals with these properties.

Towards this end, we model b as a realization of an Ising model with the distribution

$$P(b = \ell) = \frac{1}{Z_p(\eta, \Omega)} \exp\left(-\eta \ell' \Omega \ell\right).$$

Here Ω is the weighted adjacency matrix of the encoded undirected graph $\mathcal{G}$, and $Z_p(\eta, \Omega)$ is the normalizing constant. Not all Ising models generate signals with the desired properties, and in this section, we specify this class of Ising models by adding some regularity conditions on (η, Ω).

In order to model the graph sparsity of the signals, we only need the sparsity of Ω which is formally defined as follows.

Definition 3.2.1. *A graph $\mathcal{G}$ is K-sparse if for every node in $\mathcal{G}$, there are at most K edges connecting it to other nodes. A symmetric matrix Ω is sparse if and only if its induced graph is sparse.*

Before we proceed to the discussion on the signal sparsity, we need to introduce some notations. For an induced subgraph B of the graph $\mathcal{G}$ with a weighted adjacency matrix Ω, we use Ω_S the denote the submatrix of Ω restricted to the set B, and $V_B = \sum_{i,j\in B} \Omega(i,j)$. In order to characterize the signal sparsity, we set

$$\eta = \vartheta \log(p) \text{ where } \vartheta > 0.$$

Then in the Bernoulli model (3.3), as $p \to \infty$, $1/(1+e^{\eta}) \approx p^{-\vartheta}$, so that the Bernoulli model considered in Ji and Jin 2012, Jin and et al 2012 is a special case of the Ising model in this paper. For ℓ, a realization of the Ising model (3.2), let $S(\ell)$ be the set of all the signals. The probability of ℓ is $P(b = \ell) = \exp(-\eta V_{S(\ell)})/Z_p(\eta, \Omega)$, which depends on $V_{S(\ell)}$. Similar to Bernoulli model, the most likely single realization of a sparse Ising model should be $\ell = 0$, which is a vector with all zeros. Hence we assume $V_S \geq 0$, for all subgraph S. The sparsity level of the Bernoulli model is $p^{1-\vartheta}$, which is not necessarily true in our Ising model setting. In fact, Theorem 3.2.1 shows that $p^{1-\vartheta}$ is the number of the singletons and hence a lower bound of the overall sparsity level.

In general, the sparsity level is $p^{1-c(\Omega)\vartheta}$, where $c(\Omega)$ is a constant between $(0, 1)$. $c(\Omega)$ depends on the clustering effect as described in the following, and illustrated via a simple example in Section 3.2.7. The spirit of our discussion in the following is: when the clustering effect is very weak, $c(\Omega)$ is close to 1; when the clustering is strong, $c(\Omega)$ is small; and we need some regularity condition on Ω so that $c(\Omega)$ is bounded away above zero and the signal sparsity still holds.

Now we formally define the signal cluster, and start the discussion on the clustering effect and the cluster size.

Definition 3.2.2. *Let B be a connected subgraph of the graph $\mathcal{G}$, and ℓ be a realization of the Ising model* (3.2)*. If $\ell_i = 1$ for all $i \in B$, and $\ell_j = 0$ for $j \in N(B)$, we call B a signal cluster of ℓ.*

Let ℓ be a random realization of the Ising model (3.2), and $\ell^\star$ a specific realization of the Ising model where all the signals are singletons. We assume ℓ and $\ell^\star$ have the same number of signals so that $|S(\ell)| = |S(\ell)|$. Clustering effect implies that the signals generated by the Ising model are more likely to live in clusters with more than one signal. So for most choices of ℓ, there is $P(b = \ell) > P(b = \ell^\star)$. Hence $V_{S(\ell)} < V_{S(\ell^\star)} = |S(\ell)|$. This explains why $c(\Omega) < 1$ for Ω that is not an identity matrix. In a special case where $|S(\ell)| = 2$, $V_{S(\ell)} < 2$ implies that all the non-zero off-diagonal elements of Ω are negative.

On the other hand, extremely strong clustering effect may result in long signal clusters. When there are too many such clusters, ℓ may not be sparse. This is why we need a regularity condition to guarantee the finite cluster length. Towards this end, for any connected subgraph B of $\mathcal{G}$, we consider constructing an upper bound of the probability that B is a signal cluster. It is easy to see that

$$P(B \text{ is a signal cluster. }) \leq \frac{\exp(-\eta V_B)}{\sum_{S \subset B} \exp(-\eta V_S)} \leq \exp(-\eta V_B).$$

Thus we only need a regularity condition that characterizes a reasonable lower bound of V_B, especially when $|B|$ is large. In the rest of this paper, we focus on the Ising models where Ω) satisfies the following condition.

Condition 1. *Let $\mathcal{G}$ be a K-sparse graph, and Ω is its weighted adjacency matrix whose diagonal elements are 1, and the non-zero off-diagonal elements are negative. For any connected subgraph S of the graph $\mathcal{G}$,*

$$V_S \geq \begin{cases} \delta & \text{if } |S| = t, 1 \leq t \leq d \\ (t-d)\rho + \delta & \text{if } |S| = t, d \leq t \leq p \end{cases}$$

where $0 < \rho, \delta \leq 1$, $\frac{\rho \log(p)}{\log(k)} = O(1)$, and d is an integer.

Here d represents the length of the most common seen signal cluster length, δ controls the number of clusters with no more than d nodes, and ρ controls the probability of longer clusters. We remark that Bernoulli model obviously satisfies Condition 1 for $d = \delta = \rho = 1$.

The following notation will be used throughout this paper.

Definition 3.2.3. *$L_p > 0$ is a $multi - \log(p)$ term which may change from occurrence from occurrence, such that for for any fixed $\delta > 0$, $\lim_{p\to\infty} L_p \cdot p^\delta = \infty$ and $\lim_{p\to\infty} L_p \cdot p^{-\delta} = 0$*

To summarize, we formally define the *Sparse Ising model* as follows.

Definition 3.2.4. Sparse Ising model *is an Ising model* (3.2) *whose parameters* (η, Ω) *satisfy the following three conditions:*

1. $\eta = \vartheta \log(p)$ *where* $\vartheta > 0$*;*
2. Ω *is* K*-sparse where* K *is at most a* L_p *term;*
3. Ω *satisfies Condition 1.*

As discussed above, Sparse Ising model is an appropriate generative model for the sparse interdependent signals. Its properties are summarized in the following lemma.

Theorem 3.2.1. *Let random variable* b *satisfies a sparse Ising model* $Ising(\eta, \Omega)$*, and* $s_p = \sum_{k=1}^{p} P(b_k = 1)$ *be the sparsity level of the Ising model. Then the following three statements are true.*

- ***Lower bound of the sparsity level:*** *except a probability of* $o(p^{-1})$*,*

$$s_p \geq L_p \cdot p^{1-\vartheta}.$$

- ***Upper bound of the sparsity level:*** *except a probability of* $o(p^{-1})$*,*

$$s_p \leq (eK)^d \cdot p^{1-\delta\vartheta}$$

Especially, when d *is finite,*

$$s_p \leq L_p \cdot p^{1-\delta\vartheta}.$$

- ***Maximal cluster size:*** *let* m_{max} *be the size of the longest signal cluster in a realization of* b*, then except a probability of* $o(p^{-1})$*,*

$$m_{max} \leq \frac{2 - \vartheta\delta}{\rho\vartheta} + d.$$

We remark that the lower bound of the sparsity level holds for a more general class of Ising model, where Condition 1 is substituted by $V_S \geq 0$ for all connected subgraph S.

3.2.2 GMAS: a screen and clean procedure for variable selection

Theorem 3.2.1 reveals the signals generated by a sparse Ising model live in disconnected short clusters, which introduce GMAS, our proposed *Screen and Clean* procedure for variable selection naturally. GMAS has two steps, a Screening step and a Cleaning step. In the screening step, for each connected subgraph of $\mathcal{G}$ no longer than a pre-specified size $m > 1$, we test whether it could be a signal cluster. We perform the tests in a way that all the signals are retained except a negligible probability. The retained nodes live in many small components. In Cleaning step, in order to find all the signals and remove the previously falsely kept signals, we fit with a constrained MLE on each component survived after the screening.

The Screening step works as follows,

- In the first step, we perform a univariate screening. A node β_i is chosen as a signal if

  $$T_i = Y_i^2 > q(\emptyset, \{i\})\tau^2$$

  Let $\hat{S}$ be the set of chosen nodes.

- In the second step, we perform a bivariate screening. Line up all possible connected pairs in some order. For each edge (i, j), if both of them are already in $\hat{S}$, we keep both. If exactly one of them is already in $\hat{S}$, say $i \in \hat{S}$ we will include j if

  $$T_{\{i\},\{i,j\}} = Y_j^2 > q(\{i\}, \{i, j\})\tau^2$$

  If none of them is in $\hat{S}$, we include both if

  $$T_{\emptyset,\{i,j\}} = (Y_j^2 + Y_i^2) > q(\emptyset, \{i, j\}) \cdot 2\tau^2$$

- Next, we test all triplets, quadruples, so on and so forth. In general, consider a connected subgraph B, let $E = B \bigcap \hat{S}$, $D = B \setminus E$. We test

  $$H_0 : \beta_i = 0, i \in D \text{ and } \beta_i \neq 0, i \in E \text{ against } H_1 : \beta_i \neq 0, i \in B \tag{3.6}$$

  We reject H_0 and update $\hat{S}$ by adding the nodes in D if

  $$T_{E,B} = \sum_{i \in D} Y_i^2 \geq q(E, B)|D|\tau^2 \tag{3.7}$$

  Otherwise, we keep $\hat{S}$ unchanged. Here $q(E, B) \in (0, 1)$ are tuning parameters.

- The process stops when we finish screening all size-ℓ_c connected subgraphs and the probability of having a larger connected subgraph is negligible.

The retained subgraph after screening contains most of the signals(SS property) and can be separate into disjoint components(SAS property), and in the Cleaning step, we can conduct more careful analysis in each of the components of the retained subgraph individually.

In the Cleaning step, assume S_c is a component in the surviving indices and $|S_c| = m$. We estimate β_c, the coordinates of β in set S_c, by minimizing the negative log-likelihood

$$L(\beta_c) = \frac{1}{2}\sum_{i \in S_c}(Y_i - \beta_i)^2 + \vartheta V_{S_{c,1}} \log(p), \tag{3.8}$$

subject to $|\beta_i| = 0$ or $|\beta_i| \geq \tau$, where $S_{c,1}$ is the support of β_c.

Later on, we will show GMAS procedure is optimal in variable selection.

3.2.3 Evaluating variable selection procedures: Hamming distance, Hamming ratio and the phase diagram

In order to compare the variable selection procedures and assess their optimality, we choose the Hamming distance as the loss function. In model (3.5), for any true model β and an estimate $\hat{\beta}$, the Hamming distance between the two is the number of differences in their sign vectors:

$$h_p(\beta, \hat{\beta}) = \sum_{j=1}^{p} 1(sgn(\beta_j) \neq sgn(\hat{\beta}_j)),$$

and the overall Hamming distance of a variable selection procedure $\hat{\beta}$ is

$$H_p(\hat{\beta}; \vartheta, r, \Omega) = E[h_p(u \circ b, \hat{\beta}) \mid b \sim Ising(\vartheta, \Omega)].$$

In the worst case scenario, the overall Hamming distance of this procedure is

$$Hamm_p(\hat{\beta}; \vartheta, r, \Omega) = \sup_{u \in \Theta_p(\tau_p)} E[h_p(u \circ b, \hat{\beta}) \mid b \sim Ising(\vartheta, \Omega)],$$

and its minimum across all variable procedures is

$$Hamm_p^*(\vartheta, r, \Omega) = \inf_{\hat{\beta}} \sup_{u \in \Theta_p(\tau_p)} E[h_p(u \circ b, \hat{\beta}) \mid b \sim Ising(\vartheta, \Omega)],$$

which we refer to as the minimax Hamming distance.

If $Hamm_p(\hat{\beta}; \vartheta, r, \Omega) = L_p \cdot Hamm_p^*(\vartheta, r, \Omega)$ for a given variable selection procedure $\hat{\beta}$, then we say this procedure achieves the optimal rate of convergence in Hamming distance. Minimax Hamming distance is a natural measure for the optimality in variable selection for two reasons: (1) it directly measures the errors made in variable selection; (2) it lays its foundation on the classic minimax theory.

Besides comparing variable selection procedures for a given set of (ϑ, r, Ω), we are also interested in comparing the performance of a given variable selection procedure for difference (ϑ, r, Ω). Hamming distance is not appropriate for this task because Hamming error usually increases as the sparsity level increases. Instead, we can use the ratio of the Hamming distance and the sparsity level, which we refer to as Hamming ratio. We remark that Hamming ratio is also appropriate for comparing different procedures for given (ϑ, r, Ω).

In the following, we illustrate how to measure the performance of the variable selection procedures with the Hamming distance and Hamming ratio to . Let $Hamm_p(\vartheta, \Omega, r, \hat{\beta}) = L_p p^{1-\rho(\vartheta, \Omega, r, \hat{\beta})}$. Also, as in Section 3.2.1, let the sparsity level of a sparse Ising model be $s_p = L_p \cdot p^{1-\vartheta_{sp}}$ where $\vartheta_{sp} = c(\Omega)\vartheta$. If $\rho(\vartheta, \Omega, r, \hat{\beta}) \leq \vartheta_{sp}$, then $Hamm_p(\vartheta, \Omega, r, \hat{\beta}) \geq L_p \cdot s_p$ and the Hamming ratio is L_p. In this case, the variable selection procedure makes more errors than the number of signals, and it fails its task. If $\rho(\vartheta, \Omega, r, \hat{\beta}) \geq 1$, then $Hamm_p(\vartheta, \Omega, r, \hat{\beta}) < L_p$. In this case, the variable selection procedure only makes a few errors, and all signals are recovered except negligible many. If $\vartheta_{sp} < \rho(\vartheta, \Omega, r, \hat{\beta}) < 1$, then there is $L_p < Hamm_p(\vartheta, \Omega, r, \hat{\beta}) < L_p \cdot s_p$. In this case, as $p \to \infty$, the Hamming distance goes to infinity but the Hamming ratio goes to zero. Hence the variable selection procedure recovers most signals, though not all of them.

For given Ω and the variable selection procedure, $\rho(\vartheta, \Omega, r, \hat{\beta})$ and s_p only depends on (ϑ, r) and hence only depends on (ϑ_{sp}, r). The above three cases are corresponding to the three regions in the so-called phase diagram. We call the two-dimensional *parameter space* $\{(\vartheta_{sp}, r) : 0 < \vartheta_{sp} < 1, r > 0\}$ the phase space. There are two important curves $r = \gamma_{no}(\vartheta_{sp}, \Omega)$ and $r = \gamma_{et}(\vartheta_{sp}, \Omega)$ in the phase space. The former is the solution of $\rho(\vartheta_{sp}, \Omega, r, \hat{\beta}) = \vartheta_{sp}$ and the later is the solution of $\rho(\vartheta_{sp}, \Omega, r, \hat{\beta}) = 1$. We remark that there is $\gamma_{no}(\vartheta_{sp}, \Omega) \leq \gamma_{et}(\vartheta_{sp}, \Omega)$. Hence these two curves partition the whole phase space into three different regions:

- *Region of No Recovery.* $\{(\vartheta_{sp}, r) : 0 < r < \gamma_{no}(\vartheta_{sp}, \Omega), 0 < \vartheta_{sp} < 1\}$. In this region, as $p \to \infty$, the minimax Hamming error that the variable selection procedure makes as many errors as the the total expected number of signals. In this region, this procedure fails.
- *Region of Almost Full Recovery.* $\{(\vartheta_{sp}, r) : \gamma_{no}(\vartheta_{sp}, \Omega) < r < \gamma_{et}(\vartheta_{sp}, \Omega), 0 < \vartheta_{sp} < 1\}$.. In this region, as $p \to \infty$, the Hamming distance goes to infinity but the Hamming ratio goes to zero. Hence this procedure can recover most of the signals, but not all of them.
- *Region of Exact Recovery.* In this region, as $p \to \infty$, the Hamming distance is $L_p \cdot o(1)$, and this procedure can exactly recover all signals with overwhelming probability.

We can also compare the variable selection procedures by their phase diagrams. If one procedure yields larger exact recovery region and/or smaller no recovery region than the other one, this procedure is considered a better procedure, at least in some regions in the (ϑ_{sp}, r) plane.

Using Hamming distance as the loss function, we are ready to study the optimality of variable selection. We first characterize an lower bound of $Hamm_p^*(\vartheta, r, \Omega)$. Then, for a variable selection procedure $\hat{\beta}$, we calculate an upper bound of $Hamm_p(\hat{\beta}; \vartheta, r, \Omega)$. If the lower bound and the upper bound match up to a L_p level, there must be $Hamm_p(\hat{\beta}; \vartheta, r, \Omega) = L_p \cdot Hamm_p^*(\vartheta, r, \Omega)$ and this procedure is optimal. If they do not match, take a closer look and find the situations where it is not optimal. In the rest of this section, we will show that hard thresholding is not optimal in some cases, while our proposed procedure GMAS is optimal.

3.2.4 Lower bound of the minimax Hamming distance

The goal of this section is to formulate a lower bound of the minimax Hamming distance. This lower bound serves as a benchmark in assessing the variable selection procedures. If the errors a procedure makes is of the same level of this lower bound, we can say this procedure is optimal. The strategy of constructing such a lower bound is: we define the "local risk" around each node and use the sum of the "local risk" across the p nodes as the lower bound of the minimax Hamming distance. In the following, we sketch the way of defining the "local risk" and introduce necessary notations, yet leave the technical details in Section A.2.2.

We start with considering a specific node j and a connected subgraph S that j is in, presumably no signals in $N(S)$. Let $S_0 \subset S$ be the signals in the current realization of the Ising model, and S_1 the set of signals in a tampered model. In the tampered model, we only change S by replacing the signals in S_0 by S_1 while keeping other coordinates unchanged. Without losing generality, let $S = S_0 \cup S_1$, and $j \notin S_0 \cap S_1$. Then, how well we can decide whether node j is signal depends on how well can separate the current model from the tampered model. Since in both models, the signals and noises in subgraph S do not depend on the signals outside, we only need to concern about the coordinates in subgraph S. In the rest of this section, we will use Y, β u and b to denote the coordinates of Y, β u and b in subgraph S, and use Ω to denote the submatrix of Ω in S without confusion. We use β_{S_i} u_{S_i} and b_{S_i}, $i = 0, 1$ to denote β, u and b in the current model and the tampered model, respectively.

Mathematically speaking, how well the two models can be separated could be measured by the minimax risk of the hypothesis testing problem

$$H_0 : Y \sim N(\beta_{S_0}, I_{|S|}) \text{ against } H_1 : Y \sim N(\beta_{S_1}, I_{|S|}) \tag{3.9}$$

Neyman-Pearson's lemma says the most powerful test for these hypotheses is to reject H_0 when

$$T = (Y - \beta_{S_0})^T (Y - \beta_{S_0}) > t$$

for some threshold $t = 2q\log(p)$. Under H_0, $T \sim \chi^2_{|S|}$, and under H_1, $T \sim \chi^2_{|S|}(\tau^2\omega)$, where $\tau^2\omega = (\beta_{S_1} - \beta_{S_0})'(\beta_{S_1} - \beta_{S_0})$. For given S_0, S_1, β_{S_0}, β_{S_1} and q, the risk of this test is

$$P(H_1 \text{ is true}) P(\chi^2_{|S|} \leq t) + P(H_0 \text{ is true}) P(\chi^2_{|S|}(\tau^2\omega) > t) \tag{3.10}$$

Easy to see that this risk is of the level of $L_p \cdot p^{-\rho(S_0, S_1, \beta_{S_0}, \beta_{S_1}, q)}$, where $\rho(S_0, S_1, \beta_{S_0}, \beta_{S_1}, q)$ is a positive constant that depends on $(S_0, S_1, \beta_{S_0}, \beta_{S_1}, q)$. When q is chosen optimally, this risk is minimized. We define

$$\rho(S_0, S_1, \beta_{S_0}, \beta_{S_1}) = \max_{q>0} [\rho(S_0, S_1, \beta_{S_0}, \beta_{S_1}, q)],$$

and the lower bound of the risk of the specific test is $L_p \cdot p^{-\rho(S_0,S_1,\beta_{S_0},\beta_{S_1})}$. The minimax "local risk" around a node j would be the risk of the test in the "worst case scenario" among all (S_0, S_1) such that $j \in S_0 \cup S_1$, and all $(\beta_{S_0}, \beta_{S_1})$ for each given (S_0, S_1) pair. Let A_j be the set of (S_0, S_1) need to be considered at node j, and we define

$$\rho_j^* = \rho_j^*(r, \vartheta, \Omega) = \min_{(S_0,S_1)\in A_j, u_{S_0}, u_{S_1} \in \Theta_p(\tau)} \{\rho(S_0, S_1, \beta_{S_0}, \beta_{S_1})\}.$$

Then the lower bound of the minimax risk at the node j is $L_p \cdot p^{-\rho_j^*}$. The result is formally stated in the following theory. For more details, please refer to Section A.2.2.

Theorem 3.2.2. *Consider Model* (3.5) *and* (3.2)*, and assume the assumptions of Theorem 3.2.1 holds. Define*

$$\rho_j^* = \rho_j^*(r, \vartheta, \Omega) = \min_{(S_0,S_1)\in A_j, u_{S_0}, u_{S_1} \in \Theta_p(\tau)} \{\rho(S_0, S_1, \beta_{S_0}, \beta_{S_1})\}.$$

where

$$A_j = \{(S_0, S_1) \mid S_0 \textit{ and } S_1 \textit{ are distinct subgraphs of } G, j \in S_0 \cup S_1 \equiv S, S \textit{ is connected, } , 0 \le |S| \le 2\ell_c\}.$$

Then, a universal lower bound of the Hamming distance of any variable selection procedure is

$$\inf_{\hat\beta}\{Hamm_p(\vartheta, \Omega, r, \hat\beta)\} \ge L_p \cdot \sum_{j=1}^{p} p^{-\rho_j^*} = Hamm_p^*(\vartheta, u, \Omega)$$

3.2.5 When does hard thresholding fail?

It has been shown that hard thresholding cannot achieve the optimal rate when the noise is highly correlated [31]. On the other hand, when the signals are not independent, hard thresholding may also fail to achieve the optimal rate. In this section, we study whenand why it happens.

Consider (.12), the minimax risk of the hypothesis test problem (3.9). Let $S_0 = \emptyset$ and $S_1 = \{j\}$, then

$$\rho(\emptyset, \{j\}) = \vartheta + \frac{1}{4}(\sqrt{r} - \frac{\vartheta}{\sqrt{r}})_+^2.$$

There must be

$$\rho_j^* \leq \rho(\emptyset, \{j\}),$$

and the equality holds only when $(\emptyset, \{j\})$ is the worst case.

On the other hand, the optimal rate of hard thresholding at node j is

$$\rho_H = \vartheta_{sp} + \frac{1}{4}(\sqrt{r} - \frac{\vartheta_{sp}}{\sqrt{r}})_+^2,$$

where $P(b_j = 1) = L_p \cdot p^{-\vartheta_{sp}}$. Note for any connected subgraph S with $V_S > 0$ and $j \in S$, we have

$$P(b_j = 1) \geq p^{-V_S \vartheta}.$$

Hence $\vartheta_{sp} \leq V_S \vartheta$, especially, $\vartheta_{sp} \leq \vartheta$.

By saying hard thresholding fails in we mean it cannot achieve the optimal rate in variable selection, or equivalently,

$$\rho_H < \rho_{LB}, \tag{3.11}$$

which implies $\rho_H < \rho(\emptyset, \{j\})$. Note that $\rho(\emptyset, \{j\})$ and ρ_H have the same form, and this is an increasing function of ϑ(or ϑ_{sp}). So a necessary condition of hard thresholding's failure is

$$\vartheta_{sp} < \vartheta$$

This is the key insight in understanding hard thresholding, and its interpretation is the following. Hard thresholding intrinsically assumes the independence among the signals and only utilizes the marginal distribution of each node. It will work perfectly when most of the signals live as singletons. However, when $\vartheta_{sp} < \vartheta$, the probability of being a singleton is $O(p^{-\vartheta}) = o(p^{-\vartheta_{sp}})$, and most of the signals live in short clusters(pairs, triplets and etc.). In such cases, hard thresholding may fail due to its ignorance on the dependency among the signals.

In detail, we have the following theorem characterizing the situations when hard thresholding fails.

Theorem 3.2.3. *In the hypothesis problem (3.9), let S_0 and S_1 be the pair of clusters at node j in the worst case scenario, and we write $V_{S_0} = V_0$ and $V_{S_1} = V_1$ for short. Also assume $P(b_j = 1) = L_p \cdot p^{-\vartheta_{sp}}$ and $r > \vartheta_{sp}$. Hard thresholding cannot achieve the optimal rate if and only if one of the following holds.*

1. $\frac{\vartheta_{sp}}{\vartheta} < \frac{r}{\vartheta} \le \frac{|V_0 - V_1|}{|D|}$.
2. $|D| > 1$ *and* $\frac{r}{\vartheta} > \max(\frac{\vartheta_{sp}}{\vartheta}, \frac{|V_0-V_1|}{|D|})$.
3. $|D| = 1$, $\frac{\vartheta_{sp}}{\vartheta} \le V_0 + V_1$ *and* $\frac{r}{\vartheta} > \max(\frac{(\vartheta_{sp}/\vartheta)^2 - (V_0 - V_1)^2}{2[(V_0+V_1) - \vartheta_{sp}/\vartheta]}, \frac{\vartheta_{sp}}{\vartheta}, \frac{|V_0-V_1|}{|D|})$.

3.2.6 Optimal rate of convergence of GMAS

Theorem 3.2.4. *Consider Model (3.5) and (3.2), and assume the assumptions of Theorem.3.2.1 holds.*

- *In the screening step of GMAS procedure, for each test (3.6), tune*

$$0 < q(E, B) \le \left(1 - \sqrt{\frac{(\rho_D^* - V_B\vartheta)_+}{|D|r}}\right)^2 = q^*(E, B),$$

  where $\rho_D^* = \max_{i \in D}(\rho_i)$.

- *In the cleaning step, in each component of the surviving indices, we minimize (3.8).*

Denote the set of the screening parameters as q_s, and the result of GMAS as $\hat{\beta}_{mcs}(\Omega, \vartheta, r, q_s)$. The upper bound of the hamming error of GMAS procedure is

$$Hamm_p(\hat{\beta}_{mcs}(\Omega, \vartheta, r, q_s), \vartheta, \Omega, r) \le L_p \cdot \sum_{j=1}^{p} p^{-\rho_j^*}$$

3.2.7 A tri-diagonal example

The exact sparsity level of a sparse Ising model and the minimax Hamming error are rarely of explicit forms except in some special cases. In this section, we carefully analyze such a simple example. Our result shows that the theoretical properties of the sparse Ising model rely on Ω, especially its large off-diagonal elements.

We consider an Ising model where Ω is a tri-diagonal matrix as follows.

$$\Omega(i,j) = \begin{cases} 1 & if \quad i = j \\ -\theta_1 & if \quad |i-j| = 1 \text{ and } \max(i,j) \text{ is an even number} \\ -\theta_2 & if \quad |i-j| = 1 \text{ and } \max(i,j) \text{ is an odd number} \\ 0 & otherwise \end{cases} \tag{3.12}$$

where $0 < \theta_2 < 1/2 < \theta_1 < 1 - \theta_2$. Here the edges with parameter θ_1 are those of the strong dependency, and θ_2 are those of the weak dependency. Hence the induced graph of Ω contains both strong and weak dependency, which is more realistic than a dependency graph with homogeneous strength.

When Ω is of the form of (3.12), the sparsity level of the Ising model can be explicitly calculated, which is presented in the following corollary, along with result on the maximal cluster size.

Corollary 3.2.1. *For a sparse Ising model where Ω satisfies (3.12), the following two statements hold.*

- ***Sparsity Level:*** *the sparse level of this model is $s_p = L_p \cdot p^{1-\vartheta_{sp}} = L_p \cdot p^{1-\vartheta(2-2\theta_1)}$.*
- ***Maximal cluster size:*** *except probability of $o(p^{-1})$, all the clusters are no longer than $\frac{2(1-\vartheta\theta_2)}{\vartheta(1-\theta_1-\theta_2)}$*

The surprise is that the sparsity level is determined by the strong dependency θ_1, up to a L_p term.

For the Ising model considered in this section, most signals it generates live in pairs due to the strong dependency. Hence hard thresholding is not optimal in the minimax Hamming error, though it separates the singletons and the noises optimally. The follow corollary characterizes the universal lower bound of the minimax Hamming error in variable selection and the non-optimality of hard thresholding.

Corollary 3.2.2. *In model* (3.5) *and* (3.12)*, the following statements hold .*

- *The optimal rate of Hamming distance is* $p^{1-\rho^*_{1d}}$ *where*

$$\rho^*_{1d} = \begin{cases} \vartheta + \frac{1}{4}(\sqrt{r} - \frac{\vartheta}{\sqrt{r}})^2_+, & If \quad \frac{r}{\vartheta} \geq c(\theta_1) \\ 2\vartheta(1-\theta_1) + \frac{1}{2}[\sqrt{r} - \frac{\vartheta(1-\theta_1)}{\sqrt{r}}]^2_+, & If \quad \frac{r}{\vartheta} < c(\theta_1) \end{cases}$$

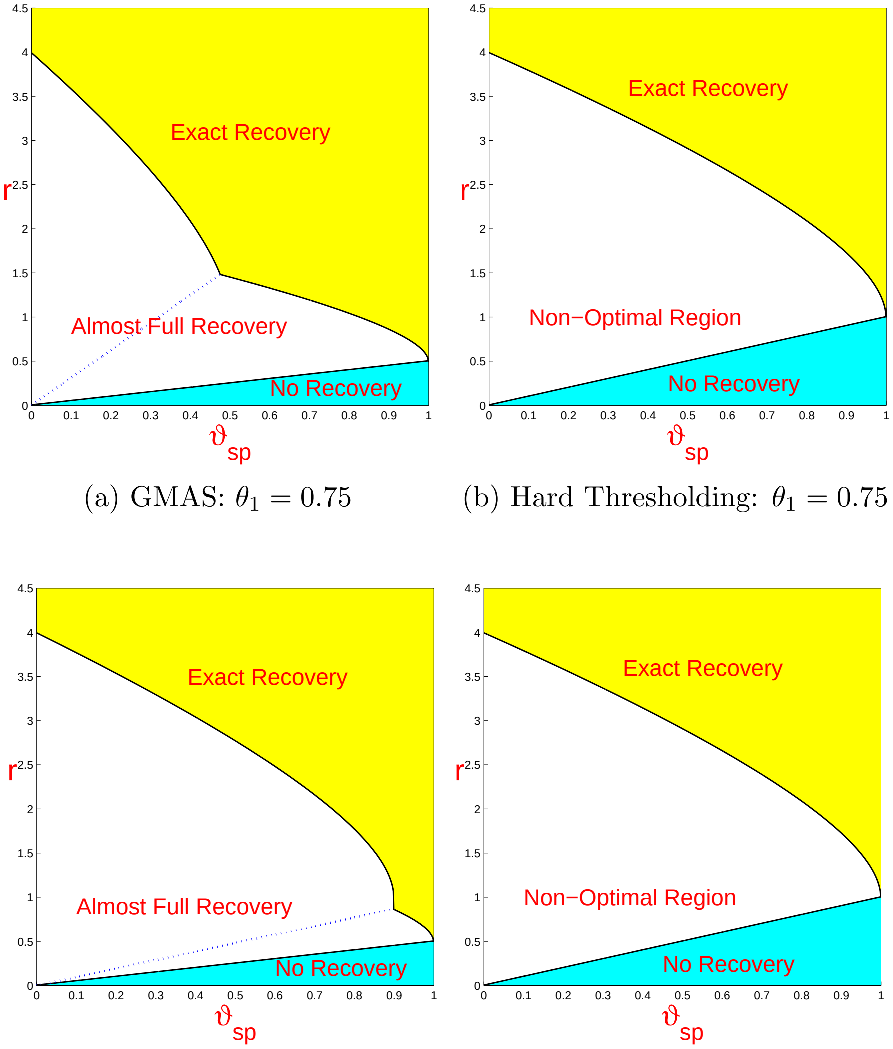
  where

$$c(\theta_1) = (\theta_1 + \sqrt{2\theta_1 - 1})I(\theta_1 \leq 2 - \sqrt{2}) + [\theta_1(2+\sqrt{2}) - 1]I(\theta_1 > 2 - \sqrt{2}).$$

- *Hard thresholding cannot achieve the optimal rate when* $\frac{r}{\vartheta} > 1 - \theta_1$.

In general, the optimal rate of Hamming error depends on (ϑ, r, Ω) in a complicated way, and it is hard to draw the phase diagram for GMAS. However, Corollary 3.2.2 characterizes an example where the optimal rate $p^{1-\rho^*_{1d}}$ only depends on (ϑ, r, θ_1) via a very simple form. In Figure 3.1, we take advantage of this simple form, and compare the optimal phase diagrams (also the phase diagrams of GMAS) and the phase diagrams of the hard thresholding in the tri-diagonal case when $\theta_1 = 0.75$ and 0.55. We remark that in both cases, most signals live in pairs, triplets and longer clusters, but not singletons.

We make two comparisons in Figure 3.1. First, we compare the two phase diagrams in each row. The rate of hard thresholding is non-optimal in both cases. This is because hard thresholding ignores the signal dependency, and only utilizes the sparsity level. It may separate the singletons from the pure noise optimally, but is non-optimal in separating the pairs from the singletons or the pure noise. Second, we compare the phase diagrams in each column. We can see that the rate of hard thresholding does not depend on θ_1. On the other hand, when $\theta_1 = 0.75$, the exact recovery region is larger than that in the case of $\theta_1 = 0.55$, and the no recovery region is also smaller. Therefore, when θ_1 is large and the signal dependency is strong, the variable selection problem is actually easier for the optimal procedures such as GMAS. This is different from the findings in [32]. In their case, when the correlation among the predictors becomes stronger, the optimal rate of Hamming error becomes worse. In our case, when the signal dependency becomes stronger, the optimal rate of Hamming error becomes better. Thus, the signal dependency is rather a “bless” than a “curse”.



(a) GMAS: $\theta_1 = 0.75$

(b) Hard Thresholding: $\theta_1 = 0.75$

(c) GMAS: $\theta_1 = 0.55$

(d) Hard Thresholding: $\theta_1 = 0.55$

Figure 3.1: Phase diagrams of GMAS and hard thresholding in the tri-diagonal case with different θ_1

3.3 SCREEN AND CLEAN ON ISING MODEL

As we have stated, our two-step Screen and Clean procedure work as follows: in the screening step, the majority of the noise are excluded while almost all the signals are retained; and in the cleaning step, the signals are detected. The screening step has two important properties that guarantees the success of this two-step strategy, which are Sure Screening(SS) property and Separable After Screening (SAS) property. The former makes sure that most of the signals are retained, and the latter guarantees the retained nodes live in small components so that the cleaning in each piece is computationally affordable. In this section, we present this idea in detail, and conclude with the proof of Theory 3.2.4.

We first study the properties of the screening step. The key is to set a collection of thresholds Q for the tests. Let $\mathcal{U}$ be the set of the retained nodes. If the thresholds are too large, we will miss a lot of signals; and if the thresholds are too low, $\mathcal{U}$ may be too large and the cleaning step becomes too expensive in computation. The following lemma characterizes the Sure Screening property, and give the maximal thresholds that guarantee this property.

Lemma 3.3.1. *In Model* (3.5)*, assume the assumptions of Theorem 3.2.1 holds. In each test in the the screening step, we test the hypothesis* (3.6) *with the test statistics* (3.7)*. Tune*

$$0 < q(E, B) \leq \left(1 - \sqrt{\frac{(\rho_D^* - V_B \vartheta)_+}{|D| r}}\right)^2 \equiv q^*(E, B),$$

where $\rho_D^* = \max_{i \in D}(\rho_i)$*, then as* $p \to \infty$*,*

$$\sum_{i=1}^{p} P(\beta_i \neq 0, i \neq \mathcal{U}) \leq L_p \cdot \sum_{i=1}^{p} p^{-\rho_i^*}.$$

The next is SAS property.

Lemma 3.3.2. *Consider Model* (3.5)*, and assume the assumptions of Theorem.3.2.1 holds. In each test in the screening step, we test the hypothesis* (3.6) *with the test statistics* (3.7)*, if*

$$q^* = \min\{q(E, B) | B \subseteq G \text{ connected}, \ E \subset B.\} > 0, \tag{3.13}$$

then the the screening step has SAS property except a probability of $o(p^{-1})$*. The set of the retained nodes contains no cluster that is longer than*

$$\max\left(a\ell_c, \frac{2a}{[\sqrt{r}(\sqrt{q^*(a-1)}-\sqrt{q_u})-\sqrt{2}]^2}\right),$$

Here a is a constant and

$$a > 1 + \frac{1}{q^*}(\sqrt{q_u}+\sqrt{2/r})^2,$$

where q_u is the parameter of univariate screening.

We remark here the length of the clusters in the set of retained nodes is usually at most $a\ell_c$, where ℓ_c is the maximum size of the clusters in the underlying Ising Model defined in Theorem.3.2.1, and a is a constant.

3.3.1 Proof of Theorem 3.2.4

Due to the Sure Screening Property of the the screening step, all the signals are kept in the surviving indices, and the SAS property assure the surviving indices split into disjoint components. Since neither the signals nor the noises in one component depend on the those in other components, we can solve the normal means problem in each component separately. Consider the model (3.5) in one component U, which is.

$$Y_U = \beta_U + z_U, \text{ where } z \sim N(0, I_{|U|}) \text{ , } \beta_U = \tau \circ b_U \text{ and } b_U \sim Ising(\Omega_U)$$

In the rest of this proof, we will discard the subscript U without confusion.

(3.8) can be rewritten as

$$L(\beta) = \frac{1}{2} \parallel Y - \beta \parallel_2^2 + \eta V_S, \tag{3.14}$$

where S is the support of β. Recall that $\beta = u \circ b$, and our cleaning step is minimizing (3.14) subject to $|u_i| \geq \tau$ or $= 0$. When S is given, the constrained minimizer of (3.14) is $\hat{\beta} = \hat{u} \circ b$, where

$$\hat{u}_i = sgn(Y_i)\max(|Y_i|, \tau).$$

Let β be the true underlying parameter with support S, $\hat{\beta} = u \circ b$ to be an estimate with the true support, and $\beta^* = u^* \circ b^*$ another estimate of β with a wrong support S^*. We are interested in $P(L(\beta^*) < L(\hat{\beta}))$, the probability that the cleaning step fails to recover the true support. In the cleaning step, this component contributes to Hamming distance for at most $|U| \cdot \max_{\beta^*} P(L(\beta^*) < L(\hat{\beta}))$

Now let us start considering $P(L(\beta^*) < L(\beta))$ for a given true support b. As before, we can view S as the true support and S^* the support of a tampered model. We denote $S \cap S^*$ as E, and $D = (S \cup S^*) \setminus E$. We can split D into four disjoint sets,

$$D_1 = \{i \in S \cup S^*, b_i = 0, b_i^* = 1 \text{ and } Y_i > \tau\}, \qquad D_2 = \{i \in S \cup S^*, b_i = 0, b_i^* = 1 \text{ and } Y_i \leq \tau\}$$
$$D_3 = \{i \in S \cup S^*, b_i = 1, b_i^* = 0 \text{ and } Y_i > \tau\} \qquad \text{and} D_4 = \{i \in S \cup S^*, b_i = 1, b_i^* = 0 \text{ and } Y_i \leq \tau\}$$

It is easy to see, except a probability of $L_p \cdot |U| \cdot p^{-r}$, $D_1 = \emptyset$.

Recall that $Y_i = \beta_i + z_i$ where z_i are iid standard normal. Direct calculation shows

$$L(\beta^*) - L(\hat{\beta}) = \tau(\sum_{i \in D_4} z_i - \sum_{i \in D_2} z_i) - \sum_{i \in D_1} \frac{z_i^2}{2} + \frac{\tau^2 |D_2|}{2} + \sum_{i \in D_3} \frac{Y_i^2}{2} + \tau \sum_{i \in D_4} (\beta_i - \frac{\tau}{2}) + \eta(V_{S^*} - V_S).$$

Except a probability of $L_p \cdot |U| \cdot p^{-r}$, $D_1 = \emptyset$. Hence, Except a probability of $L_p \cdot |U| \cdot p^{-r}$,

$$L(\beta^*) - L(\hat{\beta}) = \tau(\sum_{i \in D_4} z_i - \sum_{i \in D_2} z_i) + \frac{\tau^2 |D_2|}{2} + \sum_{i \in D_3} \frac{Y_i^2}{2} + \tau \sum_{i \in D_4} (\beta_i - \frac{\tau}{2}) + \eta(V_{S^*} - V_S).$$

Use the definitions, we can obtain a lower bound of the right hand side, which is

$$\tau(\sum_{i \in D_4} z_i - \sum_{i \in D_2} z_i) + \frac{\tau^2 |D|}{2} + \eta(V_{S^*} - V_S). \tag{3.15}$$

Note that (3.15) < 0 is a necessary condition of $L(\beta^*) < L(\beta)$, hence $P(L(\beta^*) < L(\beta)) \leq P((3.15) < 0)$.

In (3.15) $\tau(\sum_{i \in D_4} z_i - \sum_{i \in D_2} z_i)$ has a normal distribution

$$\eta \equiv tau(\sum_{i \in D_4} z_i - \sum_{i \in D_2} z_i) \sim N(0, (|D_2| + |D_4|)\tau^2).$$

Hence,

$$P((3.15) < 0) \leq P(\eta < -[\frac{\tau^2 |D|}{2} + \eta(V_{S^*} - V_S)]).$$

By mill's ratio,

$$P(\eta < -[\frac{\tau^2 |D|}{2} + \eta(V_{S^*} - V_S)]) = L_p \cdot \exp\left\{ -\frac{[\tau^2 |D|/2 + \eta(V_{S^*} - V_S)]^2}{2\tau^2 (|D_2| + |D_4|)} \right\}.$$

Obviously,

$$\exp\left\{-\frac{[\tau^2|D|/2+\eta(V_{S^*}-V_S)]^2}{2\tau^2(|D_2|+|D_4|)}\right\} \leq \exp\left\{-\frac{[\tau^2|D|/2+\eta(V_{S^*}-V_S)]^2}{2|D|\tau^2}\right\}.$$

The right hand side the above is

$$\exp\left\{-\frac{1}{4}\left(\sqrt{|D|r}-\frac{\vartheta|V_S-V_{S^*}|}{\sqrt{|D|r}}\right)_+^2 + \vartheta\max(V_S,V_{S^*})\right\} = p^{-\rho(S,S^*)}.$$

Therefore, we have

$$P(L(\beta^*)<L(\beta)) \leq L_p \cdot p^{-\rho(S,S^*)}.$$

If $S \cup S^*$ is connected, by Lemma A.2.3, $|S \cup S^*| \leq 2\ell_c$. If it is not connected, we can always only consider its smallest component. Hence we can always find $j \in S \cup S^*$, s.t. $(S, S^*) \in A_j$. Therefore $P(L(\beta^*) < L(\beta)) \leq L_p \cdot \rho_j^*$, which concludes the proof.

3.4 SIMULATION

In this section, we conduct a series of small scale simulation experiments to investigate the numerical properties of GMAS and compare it with other procedures such as hard thresholding and the lasso. We consider the experiments where the $b \sim Ising(\eta_p, \Omega)$, and $u \in \Theta_p(\tau_p)$. As before (η_p, τ_p) are tied to (ϑ, r) by $\eta = \vartheta\log(p)$ and $\tau_p = \sqrt{2r\log(p)}$. Each experiment except Experiment 3 contains the following steps.

1. Fix $(p, \vartheta, r, \mu, \Omega)$ such that $\mu \in \Theta_p(\tau_p)$. Generate a vector $b = (b_1, b_2, \ldots, b_p)' \sim Ising(\eta, \Omega)$ from a Gibbs sampler and set $\beta = b \circ \mu$.
2. Generate $Y \sim N(\beta, I_p)$, and apply GMAS and the hard thresholding.
3. Repeat 1-2 independently, and record the average Hamming distances.

In the two sub-experiments in Experiment 3, we investigate the application of GMAS in linear regression and classification, hence the simulation follows slightly different steps, which we will describe therein.

Experiment 1: In this experiment, we investigate how the choices of the signal strength vector u and Ω affect the comparison of GMAS and the hard thresholding. This experiment include 4 sub-experiments, each of which use a different combination of (u, Ω). In experiments 1(a) and 1(b), we consider signals with equal strength by setting $u_i = \pm\tau_p$ with equal probability. In experiments 1(c) and 1(d), we consider the unequal strength signals and set the signs of u_i to be ± 1 with equal probability, and $|u_i| = \tau_p(0.8 + h_i)$ where $h_i \overset{iid}{\sim} Gamma(1, 0.2)$. In experiments 1(a) and 1(c), we use Ω as in Section 3.2.7 with $\theta_1 = 0.75$ and $\theta_2 = 0.1$. In experiments 1(b) and 1(d), we use Ω whose encoded graph is a random graph with a power law distribution. The procedure of generating Ω we use is as the follow.

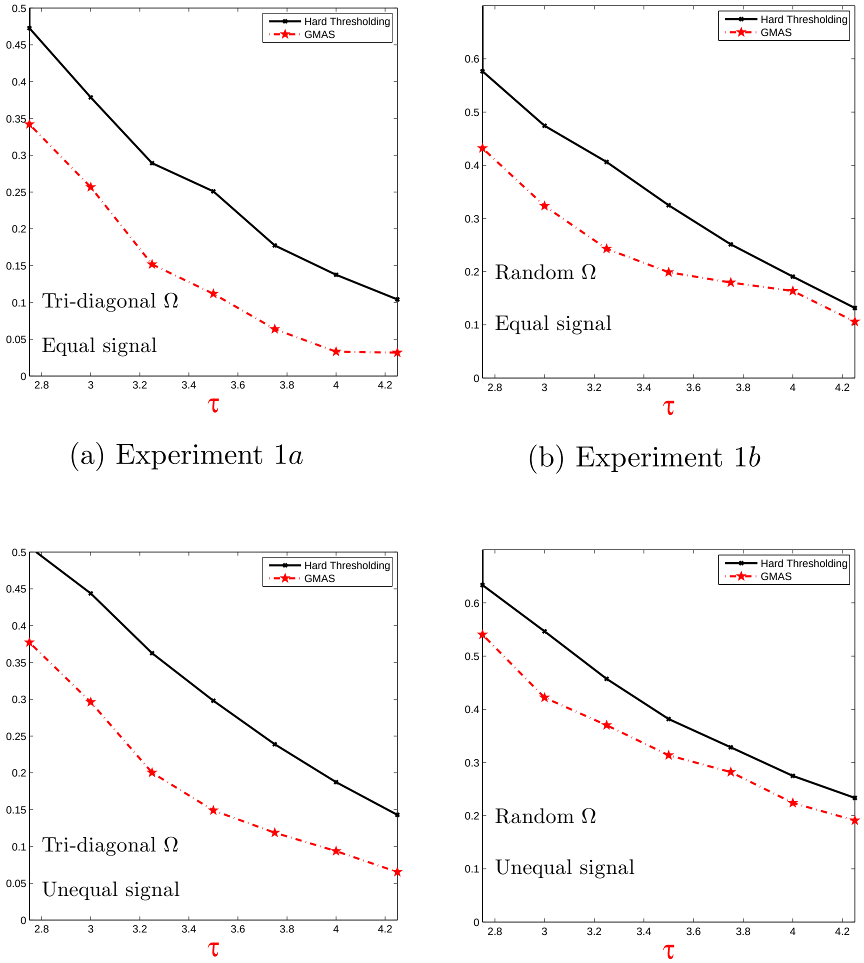
1. Generate the degree parameter sequence:
   a. Generate $\eta^{(0)} = (\eta_1^{(0)}, \ldots, \eta^{(p)})'$ $\eta_i^{(0)} \overset{iid}{\sim} \eta^{-a} 1\{1 \leq \eta \leq c\}$.
   b. Normalize $\eta^{(0)}$ to $\eta^{(1)}$ such that $\eta_i^{(1)} = [\eta_i^{(0)} - 0.9 \cdot \min(\eta^{(0)})]/[\max(\eta^{(0)}) - 0.9 \cdot \min(\eta^{(0)})]$. Note that now the range of $\eta^{(1)}$ is in $(0, 1)$.
   c. Set $\eta = p_0 \cdot \eta^{(1)}$, where p_0 is a positive constant such that $\frac{1}{p}\left(\sum_{i=1}^{p} \eta_i\right)^2 = f$.

   η is the degree parameter sequence.
2. Generate a random graph $\mathcal{G}_{rand} = (V_{rand}, E_{rand})$ for given degree parameters η: For $1 \leq i < j \leq p$, assigning (i, j) as an edge in E_{rand} randomly with probability $P((i, j) \in E_{rand}) = \eta_i \eta_j$.
3. Generate Ω by randomly assigning weights to the edges of $\mathcal{G}_{rand}$: for $1 \leq i < j \leq p$, we let $\Omega(i, i) = 1$, $\Omega(j, i) = \Omega(i, j)$ and $\Omega(i, j) = -g_{ij} 1\{(i, j) \in E_{rand}\}$ where $g_{ij} \overset{iid}{\sim} \sum_{\ell=1}^{L} p_\ell \nu_{\theta_\ell}$. Here $\nu_\star$ is a point mass at the value "$\star$", $p_\ell \in (0, 1)$ is the weight of ν_{θ_ℓ} and $\sum_{\ell=1}^{L} p_\ell = 1$.

We remark that the graph $\mathcal{G}_{rand}$ is drawn from a Degree-corrected Block Model with only one block. Its degree parameter sequence follows a power law x^{-a} and its average degree is around f. In our experiment, we use $c = 10^{1/4}$, $a = 5$, $f = 3$, $L = 3$, $(\theta_1, \theta_2, \theta_3) = (0.6, 0.2, 0.1)$ and $(p_1, p_2, p_3) = (0.1, 0.1, 0.8)$. For all experiments, we fix $(p, \vartheta) = (2000, 0.65)$. For each sub-experiment, we let $\tau_p \in 0.25 + 0.25 \times \{1, 2, 3, 4, 5, 6, 7\}$. We run the simulation for 40 repetitions. The average Hamming ratios are reported in Figure 3.2. The results suggest that GMAS outperforms hard thresholding consistently.

Experiment 2: The goal of this experiment is two fold. First, we examine how the optimal rate of Hamming error depends on the sparsity level, the signal strength and the strength of the signal dependency. Second, we want to thoroughly compare GMAS and the hard thresholding in various combinations of signal sparsity, signal strength and dependency strength. In this experiment, we fix $p = 2000$ and Ω as the tri-diagonal example in Section 3.2.7. We also fix $\theta_2 = 0.1$ since by Lemma 3.2.1 and Corollary 3.2.2, θ_2 at most has a minor effect on the Hamming ratio. Let $\theta_1 = 0.55, 0.75$, $\vartheta = 0.4, 0.65$, $\tau = 2.5, 3.5, 4.5$, and for each τ, $r = \tau^2/\log(p)$. We consider all the 8 combinations of $(\vartheta, \tau, \theta_1)$. These cases include all combinations of moderately sparse/very sparse, strong/weak signals, strong/weak signal dependency, e.g. $(\vartheta, \tau, \theta_1, \theta_2) = (0.4, 3.75, 0.75, 0.05)$ is a case where the signal is moderately sparse, but very strong, the dependency among the signals is a mixture of very strong dependency ($\theta_1 = 0.75$), and very weak ones ($\theta_2 = 0.05$). The average Hamming ratios are recorded in Table 3.1. Our results suggest that GMAS is better than hard thresholding when the signal is weak or the signal dependency is strong. Regarding to the behavior of the Hamming ratio of GMAS (also the optimal Hamming ratio), we have the following three observations: (1) For fixed (ϑ, θ_1), it decreases with τ_p, which is consistent with the intuition that the variable selection is easier when the signal becomes stronger; (2) For fixed (θ_1, τ_p), our recorded Hamming ratio of GMAS actually increases with ϑ in most cases; (3) For fixed (ϑ, τ_p), the Hamming ratio decreases with θ_1. That is to say, variable selection is easier when the signal dependency is strong, and hence the signal dependency is a "bless" that we should take advantage of instead of a "curse" that we need to break.

Experiment 3: In this experiment, we investigate the sensitivity of GMAS with respect to the parameters (ϑ, r, Ω). It includes three sub-experiments. In each sub-experiment, we perturb each of (ϑ, r, Ω) by a reasonable small amount, respectively, and investigate how the perturbation affect the performance of GMAS. Throughout the whole experiment, we adopt the random Ω and equal strength u as in Experiment 1b. Let $\vartheta = 0.65, 0.85$ and $r = 0.6, 1.4$, and we consider all the four combinations of (ϑ, r) in all sub-experiments.



(a) Experiment 1a

(b) Experiment 1b

(c) Experiment 1c

(d) Experiment 1d

Figure 3.2: Hamming ratio results in Experiment 1

τ_p		2.5		3.5		4.5	
θ_1		0.75	0.55	0.75	0.55	0.75	0.55
$\vartheta = 0.4$	GMAS	0.3610	0.5215	0.1010	0.2019	0.0304	0.0781
	hard thresholding	0.4191	0.5437	0.1637	0.2166	0.0433	0.0694
$\vartheta = 0.65$	GMAS	0.4886	0.7970	0.1242	0.2909	0.0195	0.1463
	hard thresholding	0.5906	0.7366	0.2473	0.3289	0.0661	0.1350

Table 3.1: Ratios between the average Hamming errors and the average number of signals.

In Experiment $3a$, we investigate the sensitivity of GMAS against the mis-specification of ϑ. We assume ϑ is unknown (r and Ω are assumed as known, however), and let ϑ^* be the misspecified value of ϑ. We perform GMAS with the parameters (ϑ^*, r, Ω). In this sub-experiment, we take $\vartheta^* \in \vartheta \times \{0.9, 0.95, 1, 1.05, 1.1\}$.

In Experiment $3b$, we investigate the sensitivity of GMAS against the mis-specification of r. We assume r is unknown (ϑ and Ω are assumed as known, however), and let r^* be the misspecified value of r. We perform GMAS with the parameters (ϑ, r^*, Ω). In this sub-experiment, we take $r^* \in r \times \{0.8, 0.9, 1, 1.1, 1.2\}$. The results of Experiment $3a$-$3b$ are tabulated in Table 3.2.

In Experiment $3c$, we investigate the sensitivity of GMAS against the mis-specification of Ω. We assume Ω is noisy but (ϑ, r) are known. We construct the mis-specified Ω^* by randomly deleting a small proportion of edges encoded in Ω, and then randomly adding some edges with randomly assigned weight. In detail, the diagonal elements of Ω^* are ones, and for $1 \leq i < j \leq p$, we set

$$\Omega^*(i,j) = \Omega(i,j) 1\{\Omega(i,j) \neq 0\} h_0(i,j) - g_1(i,j) \cdot h_1(i,j), \text{ and } \Omega^*(j,i) = \Omega^*(i,j)$$

where $h_0(i,j) \overset{iid}{\sim} p_{del}\nu_0 + (1 - p_{del})\nu_1$, $h_1(i,j) \overset{iid}{\sim} (1 - p_{add}f/p)\nu_0 + (p_{add}f/p)\nu_1$. Here p_{del} is the proportion of the edges deleted, and p_{add} is the ratio of expected number of randomly added edges and the expected number of edges encoded in Ω. It is reasonable to believe

ϑ^*/ϑ	0.9	0.95	1	1.05	1.1
$(\vartheta, r) = (0.65, 0.6)$	0.3038	0.2968	0.2937	0.2945	0.3019
$(\vartheta, r) = (0.85, 0.6)$	0.4039	0.4270	0.4076	0.4057	0.4057
$(\vartheta, r) = (0.65, 1.4)$	0.1008	0.1012	0.0984	0.0992	0.1016
$(\vartheta, r) = (0.85, 1.4)$	0.1174	0.1174	0.1183	0.1203	0.1193

r^*/r	0.8000	0.9000	1.0000	1.1000	1.2000
$(\vartheta, r) = (0.65, 0.6)$	0.3131	0.3009	0.3046	0.3087	0.3516
$(\vartheta, r) = (0.85, 0.6)$	0.3884	0.3805	0.3510	0.3510	0.3805
$(\vartheta, r) = (0.65, 1.4)$	0.0729	0.0785	0.0898	0.0959	0.1007
$(\vartheta, r) = (0.85, 1.4)$	0.0900	0.1052	0.1081	0.1194	0.1327

Table 3.2: Hamming ratio results in Experiment 3a-3b.

that both proportions are small. In our analysis, we choose $(p_{del}, p_{add}) \in \{0, 0.05, 0.1\}^2$. $g_1(i, j)$ is the randomly assigned weight of of (i, j) if it is a randomly added edge to the perturbed graph. Its distribution is $\sum_{t=1}^{T} \pi_t \nu_{w_t}$ where $T = 3$, $(w_1, w_2, w_3) = (0.1, 0.2, 0.3)$ and $(\pi_1, \pi_2, \pi_3) = (0.6, 0.2, 0.2)$. The result of Experiment 3c is tabulated in Table 3.3.

Experiment 4: In this experiment, we explore the prediction performance of GMAS. It contains two sub-experiments. In Experiment 4a, we explore the application of GAMS in linear regression, and in Experiment 4b, in classification. In both sub-experiments, we use equal strength signals, and use both of the tri-diagonal Ω and the random Ω as in the Experiment 1. In each setting, we consider two combinations of (ϑ, r), and compare GMAS with hard thresholding and lasso. For GMAS, we use tuning parameters (ϑ, r, Ω); for lasso, we use $\lambda \in 0.002\{1, \dots, 100\}$ and report the minimal prediction error; and for hard thresholding, we use s_h, the number of selected variables as the tuning parameter, and report the minimal prediction error. For each setting, we run 40 times.

Experiment 4a: The steps of this experiment is as the follows.

1. For fixed (p, n, n_t), generate an $n \times p$ matrix X and a $n_t \times p$ matrix $X^{(test)}$, by generate their columns as iid samples of $N(0, \frac{1}{n} I_p)$
2. Generate $z \sim N(0, I_n)$, and $z^{(test)} \sim N(0, I_{n_t})$
3. For fixed $(p, \vartheta, r, \Omega)$, generate b and construct β as in Experiment $1a$-$1b$.
4. $Y = X\beta + z$, and $Y^{(test)} = X^{(test)}\beta + z^{(test)}$.
5. Perform lasso on the training set (X, Y), and calculate the prediction error on the test set $(X^{(test)}, Y^{(test)})$.
6. Perform GMAS and hard thresholding on $\tilde{Y} = X'Y$.
7. Use the GMAS/hard thresholding estimate of b to estimate β by OLS on the training set.
8. Calculate the prediction errors on the test set.

In our experiment, we use $(p, n, n_t) = (5000, 4000, 500)$. For hard thresholding, we consider $s_h \in \{1, \ldots, 500\}$. In the case with tri-diagonal Ω, we set $\vartheta = 0.8$, and for the random Ω, we set $\vartheta = 0.65$. For both cases, we consider $r \in \{0.6, 1.5\}$. For each method, we report a triplet that includes the means and the standard deviations of the prediction errors and the average number of variables selected. For reference, we also report the means and the standard deviations of $var(Y)$, and the actual number of signals. We refer this row as "Ref". The results are recorded in Table 3.4. In almost all cases, we can see that GMAS is the best in terms of predictor error and hard thresholding performs the worst. GMAS also recover the true sparsity better than the hard thresholding (in most cases) and the lasso (in all cases).

Experiment 4b: The steps of this experiment is as follows.

1. For fixed (p, n, n_t), generate an $(n + n_t) \times p$ matrix Z whose elements are iid samples of $N(0, 1)$.
2. Fix the first $(n+n_t)/2$ elements of the length Y to be 1, and the last $(n+n_t)/2$ elements be -1.
3. For fixed $(p, \vartheta, r, \Omega)$, generate b and construct β as in Experiment $1a$-$1b$.
4. $X = Y\beta' + z$.
5. random sample n_t rows of (X, Y) as the test set $(X^{(test)}, Y^{(test)})$, and the rest as the training set $(X^{(train)}, Y^{(train)})$.

6. Perform lasso on the training set $(X^{(train)}, Y^{(train)})$, and calculate the prediction error rate on the test set $(X^{(test)}, Y^{(test)})$.
7. Perform GMAS and hard thresholding on the standardized difference of the means of the two groups.
8. Use the variables selected by GMAS/hard thresholding to train linear classifiers on the training set.
9. Calculate the classification errors on the test set.

In our experiment, we use $(p, n, n_t) = (5000, 400, 100)$. For hard thresholding, we consider $s_h \in \{1, \ldots, 250\}$. In the case with tri-diagonal Ω, we set $\vartheta = 0.8$, and for the random Ω, we set $\vartheta = 0.65$. For both cases, we consider $r \in \{0.5, 1\}$. For each method, we report a triplet that includes the means and the standard deviations of the balanced classification errors and the average number of variables selected. For reference, we also report the actual number of signals. We refer this row as "Ref". The results are recorded in Table 3.5. Due to our model setup, lasso is expected to perform poorly, as in the table. Comparing with the hard thresholding, GMAS is better both in terms of predictor errors and in the sparsity recovery.

3.5 REAL DATA APPLICATIONS

3.5.1 Identify differentially expressed genes: an yeast data example

We analyze the yeast heat shock data in [24]. In the authors' heat shock experiment, they measure the gene expression values from two cultures. In one culture(hs-1), the researchers add heated material (heat shock). In the other culture(hs-2), the cells are grown in a constant temperature, and serve as a reference group. For hs-1, samples were removed at 5, 10, 15, 20, 30, 40, 50, and 60 minutes, and total RNA was harvested for array analysis. For hs-2, the same thing was done at at 5, 15, 30, 45, and 60 minutes.

Fixing time t, we can obtain Y, the gene expression difference vector of hs-1 and hs-2, and use it to identify the differentially expressed (DE) genes between the two conditions. We expect them also to be associated to the yeast heat shock response, and finding such genes is our ultimate goal. In literature, researchers have identified some genes responsive to heat shock. In the online data base at http://www.yeastgenome.org, we find 80 genes whose Gene Ontology Biological Process (GOBP) terms contain key word ”heat”, 68 of which are also in our data set. We use this set as a benchmark, and a good DE genes identification procedure should pick more genes in this set(GOBP-heat) without using the information about this set. In literature, this is sometimes referred to as GO term enrichment(reference).

For simplicity, we only use the data at 60 minutes and 30 minutes, respectively. For each data, we apply hard thresholding and GMAS and compare the number of genes selected. For the hard thresholding, we use the number of genes selected as the tuning parameter. For GMAS, we use two parameters (s, θ) to estimate (ϑ, r, Ω). s is an integer that represents the number of genes picked by hard thresholding. Let $(Y_1, \ldots, Y_s)$ be the largest elements in Y in magnitude, and we estimate $\tau_p = median(Y_1, \ldots, Y_s)$, $r = \tau_p^2/(2\log(p))$ and $\vartheta = 1 - \log(s)/\log(p)$. In order to obtain Ω, we use the network of curated protein interaction extracted from [33], as part of their evidence in building their probabilistic model based unified network. For a positive tuning parameter θ, we set Ω as $\Omega(i,j) = -\frac{\theta}{d_i d_j} \cdot 1\{\frac{\theta}{d_i d_j} < 0.9\} - 0.9 \cdot 1\{\frac{\theta}{d_i d_j} \geq 0.9\}$ where d_i is the degree of the node i in the network. θ is the parameter that characterizing the dependency strength, and we use $\theta = 10, 25, 50$.

The results are graphed in Figure 3.3, which are the plots of the number of genes selected against the number of genes in GOBP-heat. In general, more GOBP-heat genes are selected in the 60 minutes data than in the 30 minutes data, which is consistent with the intuition in biology. In each analysis, GMAS performs better than hard thresholding, and GMAS with larger θ performs better than GMAS with smaller θ. This is a sign that the improvement made by GMAS is attributed to the signal dependency.

3.5.2 Variable selection for classification: a human tumor classification example

In this section, we explore the application of GMAS in classification. We use the breast cancer data compiled in [44] and used in [10] and [30]. This data includes 78 metastasis tumor patients and 217 non-metastasis patients, and records the expression of 8141 genes. The network I used is also from [10], which it contains 7910 nodes, and 42587 edges. For computational simplicity, we focus on the $p = 576$ genes whose variance are of the top 1000 in all genes, and it is connected to some other genes in this network.

We first calculate the standardized difference in the gene expression mean between the two groups Y. Then we apply hard thresholding and GMAS on it, separately, and then we build a Fisher's linear classifier for each selected model. We compare hard thresholding and GMAS by comparing the classification errors of their resultant classifiers. We use 5-fold cross-validation to calculate the error rate, and for each training set, we use 10-fold cross-validation to select the tuning parameters.

We use the same method to set the tuning parameter of the hard thresholding, and Ω for GMAS. In order to estimate (ϑ, r) for GMAS, we impute the tuning parameter $t > 0$, and set $r = t^2/(2\log(p))$ and $\vartheta = 1 - \log(s)/\log(p)$ where $s = 2\sum_{i=1}^{p} 1\{|Y_i| > t\}$. For tuning parameter selection of GMAS, we consider all combinations of (t, θ) where $t \in 3.5 + 0.25 \cdot \{0, \ldots, 12\}$ and $\theta \in 0.25 \cdot \{0, \ldots, 20\}$. The median error rate and the model size are reported in Table 3.6. Though the improvement is slim, but the model size of GMAS is much smaller than that of the hard thresholding. This is an evidence that GMAS build a more efficient classifier by using the network structure.

3.6 CONNECTION TO EXISTING LITERATURE, DISCUSSION AND FUTURE WORK

Our idea of using a sparse graph for variable selection is similar to [34], in which the authors propose a structured Elastic Net type of penalty for regularization in linear regression. However, our goal and approach are very different from that in [34]. First, they use dependency graph heuristically, and our approach is a generative model approach. Second, we developed the optimality theory of variable selection for signals generated by an Ising model. Third, their proposed procedure is a one-step non-adaptive approach, and ours is a two-step Screen and Clean type of method.

Our use of Ising model is close related to [48], where the authors propose an Ising model based approach for detecting differentially expressed genes. Our research concerns variable selection for regression and classification, so the scope of our research is much wider than that of [48]. Also, though the authors claim the use of an Ising model, in their algorithm, they actually use a pseudo likelihood that is not compatible with the likelihood of the Ising model, so their approach is not a generative model approach in nature. Furthermore, our paper presents a real data example where the superiority of GMAS is justified by accuracy of the detected DE's, while in the real data application in [48], the authors justify their approach by the biological interpretability of the selected subnetworks.

Our research is close related to recent works by Jin and his collaborators ([31, 32]), and these papers propose similar Screen and Clean type of variable selection methods. However, the motivation, the scope and the innovation of our research are significantly different from theirs in many ways. The motivation of our research is how to integrate the known signal dependency information into data analysis, while [31, 32] are motivated by the affect of the observed correlation structure of the predictors. Our research concerns the variable selection problem in general, including both regression and classification, while [31, 32] focus on the linear regression. The success of GMAS is attributed to the decomposability of the likelihood of the sparse Ising model, while the major innovation of [32] is to guide variable selection by the graph of strong dependency among the predictors.

The theoretical implications of our research are also very different from those in [31, 32]. For example, in Section 3.2.7, for fixed (ϑ_{sp}, r), we can see that the lower bound of the minimax Hamming distance decreases when the signal dependency becomes stronger. So signal dependency is a "bless", and GMAS is optimal because it take advantage of it. On the other hand, in [32], their lower bound of the Hamming distance increases when the correlation among the predictors increases. Hence the correlation among the predictors is a "curse", and their major contribution is to overcome the challenge of *signal cancelation* [47] brought by the predictor correlation structure.

There are many possible extensions of our research. While we focus on Ising model, we also believe that the procedure and the theories we develop here can be extend to other graphical models, e.g. Potts model [41], a multi-class extension of Ising model. Another interesting direction of future research lies in the implication of our Screen and Clean research in the problem of maximizing intractable likelihood. Previous literature in this field focus on approaches such as composite likelihood [45] or simulation based approaches [4]. In our research, our Stein's normal means model can be viewed as a hidden Ising sparse model, which is intractable. Viewing the mean vector β as the parameters, GMAS approximately maximize its intractable likelihood by taking advantage of the decomposability of this sparse likelihood. It would be interesting to investigate the application of Screen and Clean methodology on a wider class of sparse models with intractable likelihood.

(p_{del}, p_{add})	(0,0)	(0, 0.05)	(0,0.1)
$(\vartheta, r) = (0.65, 0.6)$	0.2985	0.2985	0.2985
$(\vartheta, r) = (0.85, 0.6)$	0.4536	0.4536	0.4536
$(\vartheta, r) = (0.65, 1.4)$	0.0909	0.0909	0.0909
$(\vartheta, r) = (0.85, 1.4)$	0.1136	0.1136	0.1136

(p_{del}, p_{add})	(0.05,0)	(0.05, 0.05)	(0.05,0.1)
$(\vartheta, r) = (0.65, 0.6)$	0.3097	0.2985	0.3246
$(\vartheta, r) = (0.85, 0.6)$	0.4536	0.4536	0.5052
$(\vartheta, r) = (0.65, 1.4)$	0.0909	0.0988	0.1304
$(\vartheta, r) = (0.85, 1.4)$	0.1136	0.1136	0.1364

(p_{del}, p_{add})	(0.1,0)	(0.1, 0.05)	(0.1,0.1)
$(\vartheta, r) = (0.65, 0.6)$	0.3134	0.3284	0.3321
$(\vartheta, r) = (0.85, 0.6)$	0.5052	0.4639	0.5052
$(\vartheta, r) = (0.65, 1.4)$	0.0988	0.0949	0.0949
$(\vartheta, r) = (0.85, 1.4)$	0.1364	0.1136	0.1136

Table 3.3: Hamming ratio results in Experiment 3*c*.

Tri-diagonal Ω		
(ϑ, r)	$(0.8, 0.6)$	$(0.8, 1.5)$
Ref	$(1.4450, 0.0827, 197.95)$	$(2.099, 0.179, 188.725)$
GMAS	$(1.1963, 0.0729, 294.2)$	$(1.1742, 0.0656, 227.7)$
Hard thresholding	$(1.3014, 0.0863, 254)$	$(1.3347, 0.0716, 347)$
lasso	$(1.2295, 0.0717, 658.825)$	$(1.2784, 0.0746, 771.25)$

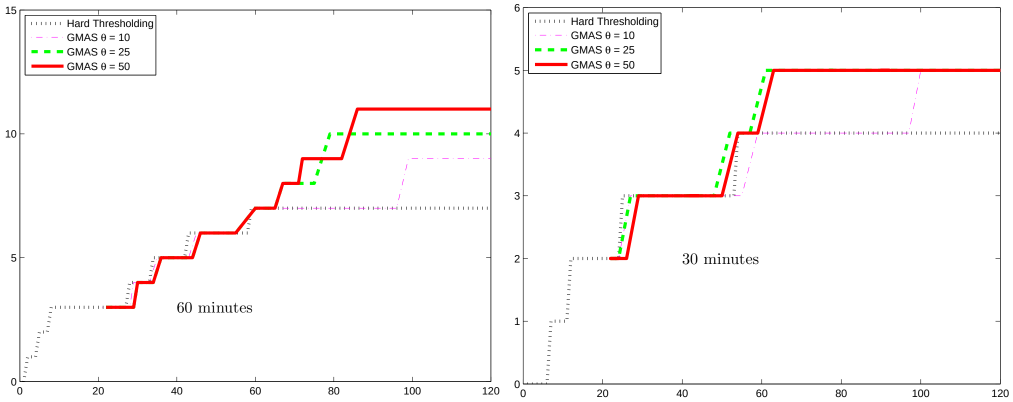
Random Ω		
(ϑ, r)	$(0.65, 0.6)$	$(0.65, 1.5)$
Ref	$(1.2029, 0.0638, 81.775)$	$(1.4822, 0.1056, 83.325)$
GMAS	$(1.1166, 0.0515, 84.125)$	$(1.1154, 0.0577, 90.45)$
Hard thresholding	$(1.1641, 0.0544, 56)$	$(1.1356, 0.0572, 102)$
lasso	$(1.1288, 0.0559, 307.575)$	$(1.1441, 0.0541, 387.425)$

Table 3.4: The means and the standard deviations of the prediction errors and the average number of variables selected by GMAS, the hard thresholding and the lasso. The means and the standard deviations of $var(Y)$, and the real sparsity level are also recorded in the first row titled "Ref".

Tri-diagonal Ω		
(ϑ, r)	$(0.8, 0.5)$	$(0.8, 1)$
Ref	186.525	197.3
GMAS	$(0.1733, 0.0448, 191.325)$	$(0.032, 0.0178, 193.5)$
Hard thresholding	$(0.1865, 0.0507, 57)$	$(0.046, 0.0173, 145)$
lasso	$(0.5919, 0.0318, 118.95)$	$(0.52, 0.0198, 172.15)$

Random Ω		
(ϑ, r)	$(0.65, 0.5)$	$(0.65, 1)$
Ref	87.1	91.925
GMAS	$(0.2106, 0.0442, 69.87)$	$(0.1063, 0.0446, 74.4)$
Hard thresholding	$(0.2771, 0.0394, 37)$	$(0.1237, 0.0324, 62)$
lasso	$(0.6, 0.2025, 0)$	$(0.5669, 0.0222, 120.85)$

Table 3.5: The means and the standard deviations of the classification errors and the average number of variables selected by GMAS, the hard thresholding and the lasso. The real sparsity level are also recorded in the first row titled “Ref”.



(a) number of GOBP-heat genes selected at 60 minutes

(b) number of GOBP-heat genes selected at 30 minutes

Figure 3.3: Differentially expressed genes in heat shock

	GMAS	hard thresholding
error rate	0.2276	0.2309
model size	93.5	136.5

Table 3.6: The result of the hard thresholding and GMAS

4.0 CURRENT AND FUTURE WORK: CLUSTERING AND NETWORK

In our previous research, we have focus on combining the correlation structure and the data in supervised learning. In this process, we have been becoming more and more interested in analyzing relational data such as network. We tackle this problem in the perspective of clustering, and our primary result seems promising.

4.1 CURRENT RESEARCH: CLUSTER EXTRACTION AND ITS APPLICATION IN COMMUNITY EXTRACTION FOR NETWORK DATA

Motivated by network data analysis in political science, we consider the partial clustering problem when at least one of the following three conditions holds:

- The sizes of the clusters are heterogenous.
- The tightness of the clusters are heterogenous.
- Not all points are in meaningful clusters.

In these situations, our goal is to extract these moderately small/reasonably tight/meaningful clusters instead of fully cluster all data.

If we view all data that are not in a meaningful cluster as one of several *"clusters of not clustered"*, conventional clustering procedures are also applicable, but usually fail. Take k-means for example: our empirical studies show that k-means may split the data into clusters of similar sizes. Hence it may split large clusters into several smaller ones, or merge small clusters into a larger one. On the other hand, part of the clusters identified by k-means may be real meaningful clusters. Hence it is crucial to develop inner measures for cluster validation and pick the meaningful clusters from the false clusters. It is also important to give the identified meaningful clusters a second chance to refine themselves. We have solved the first problem empirically, and have been working on its theory and the second problem.

4.2 CURRENT RESEARCH: THRESHOLDING IN CLUSTERING AND NETWORK CONSTRUCTION

Most of current networks are not original data, but constructed from observed relational data. In most cases, such construction involves thresholding a correlation-like matrix. The two key questions here are: **Why threshold?** and **How to threshold?**. Similar discussions have been seen in the literature of covariance matrix estimation and the gene coexpression network analysis.

We attack the first question from the perspective of clustering. More specifically, in certain regimes, thresholding may preserve the underlying group structure better. We have carefully analyzed a few examples. The results indicate that hard thresholding helps when the signals are weak and the noise level is heterogenous. This implication is supported by simulation studies and data analysis. This problem is close related to high dimensional clustering. Our theoretical analysis shows that when the signals are very weak, thresholding on the attributes can improve the signal noise ratio and the clustering result. Our simulation result even shows that aggregating the information from the thresholded variables may improve the clustering result more than a careful variable selection.

We have been working on data-driven optimal thresholding schemes for both clustering and network construction, and the theories behind our examples and empirical evidences.

4.3 IMPORTANCE OF THE CURRENT WORK AND THE FUTURE WORK

The two current research projects consider clustering and network analysis together, and connect network analysis, this modern area with the traditional clustering. This is part of the effort of understanding the structure of social and biological networks we have observed. Especially in Section 4.2, we start to consider the origin of the networks. Future work will include more theoretical studies of the relational data and network construction, and exploring new network models in this perspective.

APPENDIX

PROOFS

A.1 PROOFS OF THE RESULTS IN CHAPTER 2

A.1.1 Proof of Lemma.2.1.1

When $\mathcal{G}_S^*$ contains a connected subgraph of size $\geq m+1$, it must contain a connected subgraph with size $m+1$. By [23], 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. □

A.1.2 Proof of Theorem 2.3.1

Since σ is known, for simplicity, we assume $\sigma = 1$. First, consider (2.23). By [32], $\rho_{gs} = \min_{\{(D,F): D\cap F=\emptyset, D\neq\emptyset, D\cup F\subset\{1,2\}\}} \rho(D,F;\Omega)$, 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 (2.24). 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 [31, Lemma 4.3].

Lemma A.1.1. *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 $f_{ss}(q) = f_{ss}(q, \vartheta, r, h_0) = \min\big\{\vartheta + (1-|h_0|)r/2, q, f_{ss}^{(1)}(q), f_{ss}^{(2)}(q), f_{ss}^{(3)}(q)\big\}$.*

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

Last, consider (2.25). The proof is very similar to that of the subset selection, except for that we need to use [31, Lemma 4.1], instead of [31, Lemma 4.3]. For this reason, we omit the proof. □

A.1.2.1 Proof of Lemma A.1.1 By the symmetry in (2.20)-(2.21) when G is given by (2.22), 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

$$= L_p(I + II + III), \tag{.1}$$

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 $(\tilde{Y}_1, \tilde{Y}_2)$, $D_0 = \{\max(|\tilde{Y}_1|, |\tilde{Y}_2|) > \lambda \text{ } or \text{ } \tilde{Y}_1^2 + \tilde{Y}_2^2 - 2h_0\tilde{Y}_1\tilde{Y}_2 > 2\lambda^2(1-h_0^2)\}$, $D_1 = \{|\tilde{Y}_1| < \lambda\,,\ \tilde{Y}_1 < \tilde{Y}_2 \text{ } or \text{ } |\tilde{Y}_2 - h_0\tilde{Y}_1| > \lambda\sqrt{1-h_0^2}\}$, $D_{21} = \{\tilde{Y}_2 - h_0\tilde{Y}_1 < \lambda\sqrt{1-h_0^2} \text{ } or \text{ } \tilde{Y}_1 - h_0\tilde{Y}_2 < \lambda\sqrt{1-h_0^2}\}$ and $D_{22} = \{\tilde{Y}_2 - h_0\tilde{Y}_1 > -\lambda\sqrt{1-h_0^2} \text{ } or \text{ } \tilde{Y}_1 - h_0\tilde{Y}_2 > \lambda\sqrt{1-h_0^2} \text{ } or \text{ } \tilde{Y}_1^2 + \tilde{Y}_2^2 - 2h_0\tilde{Y}_1\tilde{Y}_2 < 2\lambda^2(1-h_0^2)\}$. Using [31, 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 $\mu \in \Theta_p(\tau_p)$,

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

$$II \leq L_p \cdot P(N((\tau_p, h_0\tau_p)', H) \in D_1) = L_p \cdot p^{-\vartheta - \min[(\sqrt{r} - \sqrt{q})^2, (1-h_0)r/2, q]}, \tag{.3}$$

and when $\beta_1 = \tau_p$ and $\beta_2 = 0$, the equality holds in (.3). 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 $\mu \in \Theta_p(\tau_p)$,

$$P(A_{21} \cap B_{21}^c) = P((\tilde{Y}_1, \tilde{Y}_2)' \in D_{21}) \leq L_p \cdot p^{-2\vartheta - (\sqrt{r(1-h_0^2)} - \sqrt{q})_+^2}, \tag{.4}$$

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 $\mu \in \Theta_p(\tau_p)$,

$$P(A_{22} \cap B_{22}^c) = P((\tilde{Y}_1, \tilde{Y}_2)' \in D_{22}) \leq L_p \cdot p^{-2\vartheta - \min([(\sqrt{r(1-h_0^2)} - \sqrt{q})+]^2, 2\{[\sqrt{r(1-h_0)} - \sqrt{q}]+\}^2)}, \tag{.5}$$

and the equality holds when $\beta_1 = -\beta_2 = \tau_p$. Inserting (.2)-(.5) into (.1) gives the claim. □

A.2 PROOFS OF THE RESULTS IN CHAPTER 3

A.2.1 Proof of Theorem 3.2.1

The goal of this section is to find $\ell_0 = p^{1-\delta_0}$ and a constant g of a L_p level, s.t.

$$P(\text{ there are no more than } \frac{\ell_0}{g} \text{ signals }) = o(p^{-1}).$$

We want to find the smallest δ_0. We will only assume $V_S \geq 0$ for any subgraph S.

More specifically, we want to find ℓ_0 and g such that

$$P(\text{ there are no more than } \frac{\ell_0}{g} \text{ signals }) = \frac{1}{Z_p(\eta)} \sum_{|S| \leq \ell_0/g} \exp\left(-\eta V_S\right) = o(p^{-1})$$

It is obvious that

$$\frac{1}{Z_p(\eta)} \sum_{|S| \leq \ell_0/g} \exp\left(-\eta V_S\right) \leq \frac{1}{Z_p(\eta)} \sum_{|S| \leq \ell_0/g} 1.$$

The right hand side can be written as $\frac{1}{Z_p(\eta)} \sum_{j=0}^{\ell_0/g} \binom{p}{j}$ whose upper bound is $\frac{\ell_0}{g}\binom{p}{\ell_0/g}$. Now we have

$$P(\text{ there are no more than } \frac{\ell_0}{g} \text{ signals }) \leq \frac{\ell_0}{g Z_p(\eta)} \binom{p}{\ell_0/g}$$

The following lemma is Lemma 1.1 in [21],

Lemma A.2.1. *Fix a sufficiently large p and $1 \leq K < p$ and suppose $G = (V, E)$ is a K-sparse graph. There is a constant $C > 0$ such that the graph decomposes into at most $CK \log(p)$ different disjoint subsets, where in each subset, there is no edge between any pair of nodes.*

By the above lemma, we can find a constant $C > 0$ such that

$$\tfrac{1}{Z_p(\eta)} \tbinom{p/(CK\log(p))}{\ell_0} \exp\left(-\ell_0 \vartheta \log(p)\right) \leq P(\text{ exactly } \ell_0 \text{ independent signals exist }) \leq 1$$

So it is enough to show

$$\binom{p}{\ell_0/g} \leq \frac{g}{\ell_0 p} \binom{p/(CK\log(p)}{\ell_0} \exp\Big(-\ell_0 \vartheta \log(p)\Big)$$

for a certain range of δ_0 and g. Apply $\ell_0 = p^{1-\delta_0}$ to the inequality and apply Stirling's approximation, we have

$$p^{1-\delta_0}\Big\{\log(p)[\delta_0(1-1/g) - \vartheta] + (1+\log(g))/g + \log(CK\log(p)) - 1/(CK\log(p))\Big\} + o(p^{1-2\delta_0}) + L_p \geq 0$$

Hence it is equivalent to show

$$\delta_0(1-1/g) - \vartheta + \frac{1}{\log(p)}[1/g + \log(g)/g + \log(CK\log(p)) - 1/(CK\log(p))] \geq 0,$$

which can be rewrote as

$$\delta_0 \geq \vartheta + \left(\frac{\vartheta}{g-1} - \frac{g \log(CK\log(p) + \log(g) + 1}{(g-1)\log(p)} + \frac{g}{C(g-1)K\log^2(p)} \right).$$

If g is bounded, then $\delta_0 \geq \frac{\vartheta}{1-1/g} + o(1)$. If $g \to \infty$ as $p \to \infty$, then $\delta_0 \geq \vartheta + o(1)$. Therefore, the minimal δ_0 is $\delta_0 = \vartheta$. We can choose any $g \geq CK\log(p)$, or just let $g = CK\log(p)$ for simplicity. Finally, we conclude

$$P(\text{ there are no more than } \frac{p^{1-\vartheta}}{CK\log(p)} \text{ signals }) = o(p^{-1}).$$

If a node I_k is a signal, it has to be in a cluster. Hence we have

$$P(I_k = 1) = \sum_{j=1}^{p} P(I_k \text{ is a signal in a cluster with length j}) \tag{.6}$$

Each term on the right hand side can be further decomposed as

$$P(I_k \text{ is in a cluster of j nodes}) = \sum_{k \in S, |S|=j, S \text{ is connected}} P(S \text{ is a cluster }). \tag{.7}$$

The number of items on the right hand side of (.7) is well controlled by the following lemma proved in [23]

Lemma A.2.2. *In a K sparse graph, fix a node I_k, the number of the connected graphs that contain I_k and have ℓ nodes is at most $(eK)^{\ell-1}$.*

What is left is to control each item on the right hand side of (.7), and the sum is thus controlled.

Consider a node I_k in the Ising model, and assume this node is in a cluster S of j signals, where $j \geq d$. we have

$$P(S \text{ is a cluster }) \leq \frac{\exp\Bigg(-1_S^T \Omega 1_S \vartheta \log(p)\Bigg)}{1 + \exp\Bigg(-1_S^T \Omega 1_S \vartheta \log(p)\Bigg)},$$

and an upper bound of the right hand side is $\exp\Bigg(-[(j-d)\rho + \delta]\vartheta\log(p)\Bigg)$. Apply this and Lemma A.2.2 to (.7), and we have, for $j \geq d$

$$P(I_k \text{ is in a cluster of j nodes}) \leq p^{-\delta\vartheta}(eK)^{d-1}[eKp^{-\rho\vartheta}]^{j-d} \tag{.8}$$

Similarly, for $1 \leq j < d$, we have

$$P(I_k \text{ is in a cluster of j nodes}) \leq p^{-\delta\vartheta}(eK)^{j-1}$$

Combine the above two with (.6), and we have

$$P(I_k = 1) \leq p^{-\delta\vartheta}(eK)^d + p^{-\delta\vartheta}(eK)^{d-1}\sum_{j=d}^{p} K^{j-d}[eKp^{-\rho\vartheta}]^{j-d}$$

When $\rho > \frac{\log(K)+1}{\vartheta\log(p)}$, $eKp^{-\rho\vartheta} < 1$. Then we can have

$$P(I_k = 1) \leq (eK)^d \cdot p^{-\delta\vartheta}.$$

It holds for any nodes in an Ising model. Hence

$$s_p = \sum_{k=1}^{p} P(I_k = 1) \leq (eK)^d \cdot p^{1-\delta\vartheta}$$

When $d \preceq \log(L_p^*)$,

$$s_p \leq L_p \cdot p^{1-\delta\vartheta}$$

where L_p and L_p^* are multi-log(p) terms. Especially, when $d = 1$,

$$s_p \leq O(L_p \cdot p^{1-\vartheta})$$

Our strategy here is similar to that of the proof of Lemma **??**, we have

$$P(\text{there exist a cluster longer than } \ell) \leq \sum_{k=1}^{p} P(I_k \text{ is in a cluster longer than } \ell) \qquad (.9)$$

and the right hand side can be further decomposed as

$$P(I_k \text{ is in a cluster longer than } \ell) = \sum_{j=\ell+1}^{p} P(I_k \text{ is a signal in a cluster with length j}) \quad (.10)$$

We only need to control each item on the right hand side of (.10).

Apply (.8) to (.10), and we have

$$P(I_k \text{ is a signal in a cluster with length longer than } \ell) \leq \sum_{j=\ell+1}^{p} p^{-\delta\vartheta}(eK)^{d-1}[eKp^{-\rho\vartheta}]^{j-d}$$

When $\rho > \frac{\log(K)+1}{\vartheta \log(p)}$, we have $eKp^{-\rho\vartheta} < 1$. Then

$$P(I_k \text{ is a signal in a cluster with length longer than } \ell) \leq L_p \cdot p^{-\delta\vartheta}(eK)^{d-1}[eKp^{-\rho\vartheta}]^{\ell-d}$$

Set the right hand side to be $o(L_p \cdot p^{-2})$, and we found it is satisfied when

$$\ell \geq \frac{2 - \delta\vartheta + \rho\vartheta d}{\rho\vartheta + \log(ek)/\log(p)}$$

$\log(ek)/\log(p) = o(1)$, hence a simplified version of the lower bound is

$$\ell \geq \frac{2 - \delta\vartheta}{\rho\vartheta} + d \tag{.11}$$

Note that (.11) does not depend on the specific node index k. Applying (.11) to (.9) concludes the proof. □

A.2.2 Proof of Theorem 3.2.2

We start at (3.10), and direct calculation gives

$$\rho(S_0, S_1, \beta_{S_0}, \beta_{S_1}, q) = \min\Big[V_{S_0}\vartheta + q, V_{S_1} + (\sqrt{\omega r} - \sqrt{q})_+^2\Big].$$

From this, it is easy to find the optimal choice of q is

$$q = \left(\frac{\sqrt{\omega r}}{2} + \frac{\vartheta(V_{S_1} - V_{S_0})}{2\sqrt{\omega r}}\right)_+^2.$$

When q is optimal, the risk is minimal, and $\rho(S_0, S_1, \beta_{S_0}, \beta_{S_1}, q) = \rho(S_0, S_1, \beta_{S_0}, \beta_{S_1})$. Hence the minimax risk for given S_0, S_1, β_{S_0} and β_{S_1} is $L_p \cdot p^{-\rho(S_0,S_1,\beta_{S_0},\beta_{S_1})}$ where

$$\rho(S_0, S_1, \beta_{S_0}, \beta_{S_1}) = V_{S_0}\vartheta + \left(\frac{\sqrt{\omega r}}{2} + \frac{\vartheta(V_{S_1} - V_{S_0})}{2\sqrt{\omega r}}\right)_+^2 = \vartheta \max(V_{S_0}, V_{S_1}) + \left(\frac{\sqrt{\omega r}}{2} - \frac{\vartheta|V_{S_1} - V_{S_0}|}{2\sqrt{\omega r}}\right)_+^2$$

Worst case scenario requires to minimize $\rho(S_0, S_1, \beta_{S_0}, \beta_{S_1})$ over $(S_0, S_1, \beta_{S_0}, \beta_{S_1})$. We minimize it over $(\beta_{S_0}, \beta_{S_1})$ for given S_0 and S_1 first, and then minimize over (S_0, S_1). Obviously, $\rho(S_0, S_1, \beta_{S_0}, \beta_{S_1})$ is minimized when

$$\omega =\mid V_{S_0} - V_{S_1} \mid \vartheta/r,$$

and as a function of ω, $\rho(S_0, S_1, \beta_{S_0}, \beta_{S_1})$ is nondecreasing. Let $E = S_0 \bigcap S_1$ and $D = S \setminus E$. By the definition of ω, we have

$$\omega = \frac{1}{\tau^2}\Bigg[\sum_{i\in D}(\beta_{S_1}(i) - \beta_{S_0}(i))^2 + \sum_{i\in E}(\beta_{S_1}(i) - \beta_{S_0}(i))^2\Bigg] \geq \frac{1}{\tau^2}[\sum_{i\in D}(\beta_{S_1}(i) - \beta_{S_0}(i))^2 \geq |D|.$$

Hence the worst-case-scenario choice of ω is

$$\omega = \max[\mid V_{S_0} - V_{S_1} \mid \vartheta/r, |D|],$$

and the minimax risk of the hypothesis testing problem (3.9) can be written as $L_p \cdot p^{-\rho(S_0,S_1)}$ where

$$\rho(S_0, S_1) \equiv \min\nolimits_{u_{S_0}, u_{S_1} \in \Theta_p(\tau_p)}[\rho(S_0, S_1, \beta_{S_0}, \beta_{S_1})] = \vartheta \max(V_{S_0}, V_{S_1}) + \tfrac{1}{4}\Bigg(\sqrt{|D|r} - \tfrac{\vartheta|V_{S_1} - V_{S_0}|}{\sqrt{|D|r}}\Bigg)_+^2. \tag{.12}$$

Hence it only contributes to the Hamming distance by $L_p \cdot p^{-\rho(S_0,S_1)}$.

The only issue left is how to define the range of (S_0, S_1) that are necessary to be considered. First of all, since we are considering the local risk at the node j, we only need to consider those such that (S_0, S_1). Second, if S is not connected, we can only consider its component that contains node j. Hence we only need to consider those (S_0, S_1) such that $S_0 \cup S_1$ is connected. At the last, we assume $S = S_0 \cup S_1$ and (S_0, S_1) are signals in the current model and the tempered model, respectively. Obviously, at least in one of the current and the tempered models, there are at least $|S|/2$ signals.

The following lemma shows that because of this, $|S|$ is small except for a negligible probability.

Lemma A.2.3. *Assume b is a realization of Ising Model (3.2), and b_S stands for its coordinates in subgraph S. Also assume the assumptions of Lemma.*?? *hold. For a finite number $a > 1$, and $\ell_c = \frac{2-\delta\vartheta}{\rho\vartheta} + d$, there is no connected subgraph S satisfies both $|S| \geq a\ell_c$ and $|b_S|_0 > |S|/a$, except a probability of $o(p^{-1})$.*

Since $S = S_0 \bigcup S_1$, immediately we know $|S| \leq |S_0| + |S_1|$. In Lemma A.2.3, set $a > 2$. Thus for $|S| > a\ell_c$, we have $|S_0| < |S|/2$ in the current model and $|S_1| < |S|/2$ in the tampered model, except a probability of $o(p^{-1})$. This contradicts with $|S| \leq |S_0| + |S_1|$, hence $|S| \leq 2\ell_c$, except a probability of $o(p^{-1})$. This concludes the definition of A_j.

Easy to check $|A_j| \leq 2(2ek)^{2\ell_c}$ which is at most a L_p term. Then we can write the minimax "local risk" at the node j as $L_p \cdot p^{-\rho_j^*}$ where

$$\rho_j^* = \rho_j^*(r, \vartheta, \Omega) = \min_{(S_0,S_1)\in A_j} \{\rho(S_0, S_1)\}.$$

Finally, summing up these local risks of the p nodes yields the universal lower bound of the Hamming distance

$$\inf_{\hat{\beta}}\{Hamm_p(\vartheta, \Omega, r, \hat{\beta})\} \geq L_p \cdot \sum_{j=1}^{p} p^{-\rho_j^*} = Hamm_p^*(\vartheta, u, \Omega).$$

A.2.3 Proof of Theorem 3.2.3

To recap.

$$\rho_j^* = \vartheta \max(V_{S_0}, V_{S_1}) + \frac{1}{4}\left(\sqrt{|D|r} - \frac{\vartheta \mid V_{S_1} - V_{S_0} \mid}{\sqrt{|D|r}}\right)_+^2$$

and

$$\rho_H = \vartheta_{sp} + \frac{1}{4}(\sqrt{r} - \frac{\vartheta_{sp}}{\sqrt{r}})_+^2.$$

We only need to find the conditions that

$$\vartheta \max(V_{S_0}, V_{S_1}) + \frac{1}{4}\left(\sqrt{|D|r} - \frac{\vartheta \mid V_{S_1} - V_{S_0} \mid}{\sqrt{|D|r}}\right)_+^2 \geq \vartheta_{sp} + \frac{1}{4}(\sqrt{r} - \frac{\vartheta_{sp}}{\sqrt{r}})_+^2. \qquad (.13)$$

When $r \leq \vartheta_{sp}$, the right hand side of (.13) is ϑ_{sp}, which is in the non-recovery region of hard thresholding. This is the uninteresting case. In the following, we focus on the case where $r > \vartheta_{sp}$. We consider the case of $r \leq \frac{\vartheta|V_{S_1} - V_{S_0}|}{|D|}$, and the case of $r > \frac{\vartheta|V_{S_1} - V_{S_0}|}{|D|}$ separately.

(.13) holds iff

$$\frac{r}{4} + \frac{\vartheta_{sp}}{2} + \frac{\vartheta_{sp}^2}{4r} < \vartheta \max(V_0, V_1).$$

For given $(\vartheta, \vartheta_{sp})$, the above holds when

$$r < (\sqrt{\vartheta \max(V_0, V_1)} + \sqrt{\vartheta \max(V_0, V_1) - \vartheta_{sp}})^2.$$

It is easy to check that the right hand side of the above is larger than $\frac{\vartheta|V_{S_1}-V_{S_0}|}{|D|}$. Hence in this case, what is need is only $\vartheta_{sp} < r \leq \frac{\vartheta|V_{S_1}-V_{S_0}|}{|D|}$.

(.13) holds iff

$$\vartheta \max(V_{S_0}, V_{S_1}) + \frac{1}{4}\left(\sqrt{|D|r} - \frac{\vartheta \mid V_{S_1} - V_{S_0} \mid}{\sqrt{|D|r}}\right)^2 \geq \vartheta_{sp} + \frac{1}{4}(\sqrt{r} - \frac{\vartheta_{sp}}{\sqrt{r}})^2.$$

The above holds iff

$$|D|(|D|-1)r^2 + 2r|D|[\vartheta(V_1+V_0) - \vartheta_{sp}] + \vartheta^2(V_1-V_0)^2 - |D|\vartheta_{sp}^2 > 0 \tag{.14}$$

When $|D| = 1$, (.14) holds iff $\vartheta(V_1+V_0) > \vartheta_{sp}$ and $r > \frac{\vartheta_{sp}^2 - \vartheta^2(V_0-V_1)^2}{2[\vartheta(V_0+V_1)-\vartheta_{sp}]}$. When $|D| > 1$, the left hand side of (.14) increases monotonically with r for $r > \frac{\vartheta_{sp}-\vartheta(V_0+V_1)}{|D|-1}$, hence it does for $r > \vartheta_{sp}$. When $r = \vartheta_{sp}$, (.14) holds, hence it holds for all $r > \vartheta_{sp}$. □

A.2.4 Proof of Corollary 3.2.1

For each node j,

$$P(b_j = 1) = P(b_1 = \ldots = b_p = 1) + \sum_{i_L=2}^{j} \sum_{i_R=j}^{p-1} P(b_{i_L} = \ldots = b_{i_R} = 1, b_{i_L-1} = b_{i_R+1} = 0)$$

An upper bound of the right hand side is

$$\sum_{i_L=1}^{j} \sum_{i_R=j}^{p} \exp\left[-\vartheta \log(p) \left(\sum_{i=i_L}^{i_R} b_i - 2\theta_1 \sum_{i=i_L, i\, odd}^{i_R-1} b_i b_{i+1} - 2\theta_2 \sum_{i=i_L, i\, even}^{i_R-1} b_i b_{i+1}\right)\right]$$

We can further simply it as

$$\frac{p^{-\vartheta} + p^{-\vartheta(2-2\theta_1)} + p^{-\vartheta(2-2\theta_2)}}{(1-p^{-\vartheta(1-\theta_1-\theta_2)})^2}$$

and it is $L_p \cdot p^{-\vartheta(2-2\theta_1)}$.

On the other hand, without losing generality, let j be an odd number. We then have

$$P(b_j = b_{j+1} = 1 | b_{j-1} = b_{j+2} = 0) P(b_{j-1} = b_{j+2} = 0) \geq \tfrac{p^{-\vartheta(2-2\theta_1)}}{p^{-\vartheta(2-2\theta_1)} + 2p^{-\vartheta} + 1}(1 - 2L_p \cdot p^{-\vartheta(2-2\theta_1)})$$

and it is also of the level of $L_p \cdot p^{-\vartheta(2-2\theta_1)}$. Hence $P(b_j = 1) = L_p \cdot p^{-\vartheta(2-2\theta_1)}$, and the sparse level is $L_p \cdot p^{1-\vartheta(2-2\theta_1)}$.

For a specific integer ℓ, an upper bound of

$$P(b_{j+1} = \ldots = b_{j+\ell} = 1, b_j = b_{j+\ell+1} = 0)$$

is

$$\exp\left[-\vartheta \log(p) \left(\sum_{i=j+1}^{j+\ell} b_i - 2\theta_1 \sum_{i=j+1, i\, odd}^{j+\ell-1} b_i b_{i+1} - 2\theta_2 \sum_{i=j+1, i\, even}^{j+\ell-1} b_i b_{i+1} \right)\right].$$

A simplified upper bound of it is $L_p \cdot p^{-2\vartheta[1-\theta_1+(\ell-2)(1-\theta_1-\theta_2)]}$. Hence

$$P(\text{node j is in a cluster longer than } d) = \textstyle\sum_{\ell=d}^{p} L_p \cdot p^{-2\vartheta[1-\theta_1+(\ell-2)(1-\theta_1-\theta_2)]} = \frac{L_p \cdot p^{-2\vartheta[1-\theta_1+(d-2)(1-\theta_1-\theta_2)]}}{1-L_p \cdot p^{-\vartheta(1-\theta_1-\theta_2)}}$$

We set the above probability to be $O(L_p \cdot p^{-2})$, and we get $d = \frac{2(1-\vartheta\theta_2)}{\vartheta(1-\theta_1-\theta_2)}$. For this d,

$$P(\text{ there exists a cluster longer than d}) = O(L_p \cdot p^{-1})$$

□

A.2.5 Proof of Corollary 3.2.2

In this proof, we compare $\rho(S_0, S_1)$ with $\rho(\emptyset, \{j\})$ and $\rho(\emptyset, \{j, j+1\})$, and eliminate (S_0, S_1) that obviously produce a larger rate. Then we compare the (S_0, S_1) pairs left and define the optimal rate in the worst case scenario.

We first compare $\rho(\emptyset, \{j\})$ and $\rho(S_0, S_1)$. In (.12), let $(S_0, S_1) = (\emptyset, \{j\})$, and we have

$$\rho(\emptyset, \{j\}) = \vartheta + \frac{1}{4}(\sqrt{r} - \frac{\vartheta}{\sqrt{r}})_+^2.$$

When $r \leq \vartheta$, $\rho(\emptyset, \{j\}) \leq \rho(S_0, S_1)$ implies $\vartheta \leq \vartheta \max(V_0, V_1)$. Hence we only need $\max(V_0, V_1) \geq 1$. When $r > \vartheta$, apply the fact that

$$|V_0 - V_1| \leq |1'_{S_0 \bigcup S_1} \Omega_{S_0 \bigcup S_1} 1_{S_0 \bigcup S_1} - 1'_{S_0 \bigcap S_1} \Omega_{S_0 \bigcap S_1} 1_{S_0 \bigcap S_1}| \leq |D|$$

to

$$\sqrt{r} - \frac{\vartheta}{\sqrt{r}} \leq \sqrt{|D|r} - \frac{\vartheta |V_0 - V_1|}{\sqrt{|D|r}},$$

and we can find that it holds for all $r > \vartheta$. Hence $\rho(\emptyset, \{j\}) \leq \rho(S_0, S_1)$ if $\max(V_0, V_1) \geq 1$.

We then compare $\rho(\emptyset, \{j, j+1\})$ and $\rho(S_0, S_1)$. Here

$$\rho(\emptyset, \{j, j+1\}) = 2\vartheta(1-\theta_1) + \frac{1}{2}(\sqrt{r} - \frac{\vartheta}{\sqrt{r}})_+^2.$$

We only need to consider (S_0, S_1) pairs where $\max(V_0, V_1) < 1$. Note that in this tri-diagonal example, if $|S_0|$ is an odd number, then $V_0 \geq 1$. Thus we only need to consider those (S_0, S_1) pairs where $|S_0|$ and $|S_1|$ are both even numbers. In such cases, for $i = 0, 1$, V_i has a form of $2t_i(1-\theta_1) + 2(|S_i|/2 - t_i)(1-\theta_1-\theta_2)$ where s_i is the number of components in S_i. Thus,

$$\frac{|V_0 - V_1|}{|D|} = \frac{|2(t_0 - t_1)(1-\theta_1) + 2(|S_0|/2 - |S_1|/2 - t_0 + t_1)(1-\theta_1-\theta_2)|}{|D|} \leq 1 - \theta_1,$$

and we can get

$$(\sqrt{|D|r} - \frac{\vartheta|V_0 - V_1|}{\sqrt{|D|r}})^2 \geq |D|(\sqrt{r} - (1-\theta_1)\vartheta/\sqrt{r})^2 \geq 2(\sqrt{r} - (1-\theta_1)\vartheta/\sqrt{r})^2.$$

Combine this with the fact that $\vartheta \max(V_0, V_1) > 2\vartheta(1-\theta_1)$, we can find that for all (S_0, S_1) such that $\max(V_0, V_1) < 1$,

$$\rho(S_1, S_0) \geq \rho(\emptyset, \{j, j+1\}).$$

To summarize, for any pair of (r, ϑ), there is

$$\rho_{1d}^* = \min(\rho(\emptyset, \{j\}), \rho(\emptyset, \{j, j+1\})).$$

We now compare $\rho(\emptyset, \{j\})$ and $\rho(\emptyset, \{j, j+1\})$. When $r/\vartheta \leq 1 - \theta_1$, $\rho(\emptyset, \{j, j+1\}) = 2(1-\theta_1)\vartheta < \vartheta = \rho(\emptyset, \{j\})$. When $1 - \theta_1 < r/\vartheta \leq 1$, $\rho(\emptyset, \{j\}) \leq \rho(\emptyset, \{j, j+1\})$ is equivalent to

$$\vartheta \leq (1-\theta_1)\vartheta + \frac{r}{2} + \frac{(1-\theta_1)^2\vartheta^2}{2r}.$$

It holds when

$$\frac{r}{\vartheta} \geq \theta_1 + \sqrt{2\theta_1 - 1}.$$

Thus it holds when $\theta_1 + \sqrt{2\theta_1 - 1} \leq \frac{r}{\vartheta} \leq 1$, which is not empty iff $\theta_1 \leq 2 - \sqrt{2}$. When $r/\vartheta > 1$, $\rho(\emptyset, \{j\}) \leq \rho(\emptyset, \{j, j+1\})$ is equivalent to

$$\frac{\vartheta}{2} + \frac{r}{4} + \frac{\vartheta^2}{4r} \leq (1-\theta_1)\vartheta + \frac{r}{2} + \frac{(1-\theta_1)^2\vartheta^2}{2r}.$$

It holds when

$$\frac{r}{\vartheta} \geq (2+\sqrt{2})\theta_1 - 1$$

When $\theta_1 > 2-\sqrt{2}$, $(2+\sqrt{2})\theta_1 - 1 > 1$. Summarize the above 2 cases and the optimal rate in the tri-diagonal case follows.

The optimal rate of hard thresholding is

$$\rho_H = 2(1-\theta_1)\vartheta + \frac{1}{4}(\sqrt{r} - 2(1-\theta_1)\vartheta/\sqrt{r})_+^2.$$

Compare ρ_H with $\rho(\emptyset, \{j\})$ and $\rho(\emptyset, \{j, j+1\})$. It is trivial to see that $\rho_H < \rho(\emptyset, \{j\})$ for all (r, ϑ) and that $\rho_H < \rho(\emptyset, \{j, j+1\})$ for $r/\vartheta > 1-\theta_1$. Since $\rho_{1d}^* = \min(\rho(\emptyset, \{j\}), \rho(\emptyset, \{j, j+1\}))$, we immediately obtain $\rho_H < \rho_{1d}^*$ for $r/\vartheta > 1-\theta_1$. □

A.2.6 Proof of Lemma 3.3.1

For each test in the screening step, use p^{-II} to denote its Type II error. We want

$$p^{-II} = P(T_{E,B} \leq q(E,B) | D_j|\tau^2|H_1 \text{ is true.}) P(H_1 \text{ is true.}) \leq L_p \cdot \min_{i\in D}(p^{-\rho_i^*}),$$

Define

$$\rho_D^* = \max_{i\in D}(\rho_i),$$

we then only need

$$II = V_B\vartheta + |D|r(1-\sqrt{q(E,B)})^2 \geq \rho_D^*.$$

Solving the above inequality gives the upper bound of the tuning parameters

$$q(E,B) \leq \left(1 - \sqrt{\frac{(\rho_D^* - V_B\vartheta)_+}{|D_j|r}}\right)^2 = q^*(E,B),$$

and the upper bound of the Type II error of this single test is

$$L_p \cdot \min_{i\in D}(p^{-\rho_i^*}) = L_p \cdot p^{-\rho_D^*}.$$

Assume there are M tests in the screening step, we know $M = L_p \cdot p$, then there must be

$$\sum_{j=1}^{M} p^{-II_j} \leq L_p \cdot \sum_{j=1}^{M} p^{-\rho_D^*} \leq L_p \cdot \sum_{i=1}^{p} p^{-\rho_i^*}.$$

It holds for any

$$0 < q(E,B) \leq q^*(E,B)$$

□

A.2.7 Proof of Lemma 3.3.2

$\hat{S}$ is the set of the retained nodes after multivariate screening. Follow the first part(before formula(A.22)) of the proof of Lemma.2.4 in [31] it is sufficient to show that we can find a sufficient large integer m, s.t. for a fixed connected subgraph S_0 with $|S_0| = m$,

$$P(S_0 \subset \hat{S}) = o(p^{-2}(eK)^{-m}) = o(p^{-\delta_0}),$$

where $\delta_0 = 2 + \frac{m \log(eK)}{\log(p)}$.

Without losing generality, assume $S_0 \subset \hat{S}$ and there is no other connected subgraph S_0^*, s.t. $S_0 \subset S_0^* \subset \hat{S}$(or say, S_0 is a component of $\hat{S}$). $S_0 \subset \hat{S}$ only if there is a sequence of hypothesis tests

$$H_{0j} : \beta_i = 0, i \in D_j \text{ and } \beta_i \neq 0, i \in E_j \text{ against } H_{1j} : \beta_i \neq 0, i \in B_j, \text{ for } j = 1, ..., J$$

where B_j is a connected subgraph of $\mathcal{G}$, $\hat{S}_j$ is the set of chosen nodes after the first $j-1$ tests, $E_j = B_j \bigcap \hat{S}_j$, $D_j = B_j \setminus E_j$, $E_1 = \emptyset$ and $S_0 = \bigcup_{j=1}^{J} B_j = \bigcup_{j=1}^{J} D_j$. For each test, we have

$$T_{E_j, B_j} > q(E_j, B_j)|D_j|\tau^2.$$

Then, after these tests, we can find S_0 as a component of $\hat{S}$.

Since we only care about the order of m, and intuitively, there should be $m > \ell_c$, it does not matter if we assume $m > a\ell_c$ for a constant a. Use Lemma.A.2.3, for any connected subgraph S with $|S| = m > a\ell_c$, there are at most m/a signals. If we can find a sufficiently large constant a, such that in each component of $\hat{S}$, there are at least $[\frac{m}{a}]$ signals, then components longer than $a\ell_c$ do not exist.

We can split S_0 into two sets, S_U and S_M, which are the sets of the nodes that are added in univariate screening and multivariate screening, respectively. Set $m_1 = |S_U|$ and $m_2 = |S_M|$, then $m_1 + m_2 = m$, and the m_1 tests are univariate screenings. Denote the tuning parameter of the univariate screening as q_u, obviously $q_u \geq q^*$. The number of noise that univariate screening retains in S_0 is at most $m \cdot p^{-q_u}$. Obviously, univariate screening can find m/a signals at most, hence

$$m_1 \leq m/a + m \cdot p^{-q_u} \text{ and } m_2 \geq m(1 - a^{-1}) + m \cdot p^{-q_u}$$

We remark here that for a constant $a > 1$ and sufficiently large m, S_M is not empty.

In set S_M, since each node fails in the univariate screening, we have $Y_i^2 \leq q_u \tau^2$ for $i \in S_M$. Direct calculation gives

$$|\beta_i| \leq \tau\sqrt{q_u} + |z_i|, z_i \sim N(0,1) \text{ for } i \in S_M$$

By mill's ratio, except a probability of $o(p^{-2})$, $|z_i| \leq 2\sqrt{\log(p)}$. Let $\tau_u = \tau\sqrt{q_u} + 2\sqrt{\log(p)}$, then except a probability of $o(p^{-1})$, for all $i \in S_M$, $|\beta_i| \leq \tau_u$. When $\tau_u < \tau$, this says there is essentially no signal left after univariate screening, But we can always find $q_u \in (0,1)$, s.t.$\tau_u > \tau$.

By the definition of q^*, we have

$$T_{E_j,B_j} > q(E_j,B_j)|D_j|\tau^2 > q^*|D_j|\tau^2,$$

therefore

$$T_M = \sum_{i \in S_M} Y_i^2 = \sum_{j=m_1+1}^{T} T_{E_j,B_j} > \sum_{j=m_1+1}^{T} q^*|D_j|\tau^2 = q^* m_2 \tau^2.$$

The distribution of T_M is

$$T_M \sim \chi^2_{m_2}(\| \beta_M \|_2^2),$$

where β_M is composed with the coordinates of β in S_M. The support of β_M is at most m/a, hence

$$\| \beta_M \|_2^2 \leq m\tau_u^2/a.$$

By the properties of non-central χ^2 distribution, the survival probability increase with the degree of freedom and the non-central parameter, we have

$$P(T_M \geq q^* m_2 \tau^2 \mid T_M \sim \chi^2_{|S_M|}(\| \beta_M \|_2^2)) \leq P(T_M \geq q^* m_2 \tau^2 \mid T_M \sim \chi^2_m(m\tau_u^2/a)),$$

For large m, $m_2 \geq m(1-a^{-1})$, hence

$$P(T_M \geq q^* m_2 \tau^2 \mid T_M \sim \chi^2_m(m\tau_u^2/a)) \leq P(T_M \geq q^* m \tau^2 (1-a^{-1}) \mid T_M \sim \chi^2_m(m\tau_u^2/a))$$

By the properties of non-central χ^2 and the mill's ratio,

$$P(T_M \geq q^* m \tau^2(1-a^{-1}) \mid T_M \sim \chi^2_m(m\tau_u^2/a)) = L_p \cdot \exp\left\{-\frac{m}{2a}(\tau\sqrt{q^*(a-1)} - \tau_u)_+^2\right\}.$$

Therefore, we only need to find (a, m) such that

$$\frac{m}{2a}(\tau\sqrt{q^*(a-1)} - \tau_u)_+^2 \geq \delta_0 = 2 + \frac{m\log(eK)}{\log(p)} \tag{.15}$$

Let $\tau\sqrt{q^*(a-1)} - \tau_u > 0$, and we get

$$a > 1 + \frac{1}{q^*}(\sqrt{q_u} + \sqrt{2/r})^2.$$

For each a that satisfies the above inequality, (.15) implies

$$m \geq \frac{2a}{[\sqrt{r}(\sqrt{q^*(a-1)} - \sqrt{q_u}) - \sqrt{2}]^2}.$$

□

A.2.8 Proof of Lemma A.2.3

Let $m_a > 0$, by Lemma A.2.2,

$$P(\text{there is a subgraph } S, s.t. |S| \geq m_a, |b_S| > |S|/a) \leq \textstyle\sum_{m\geq m_a}^{p} p(eK)^m P(|S| = m \text{ and } |b_S| > m/a).$$

For a subgraph S with m nodes, there is

$$P(|b_S| = \ell) \leq \binom{m}{\ell} \exp -\vartheta\log(p)[\rho(\ell - d) + \delta].$$

Combine the above two, only need to prove

$$\sum_{m\geq m_a}^{p} p(eK)^m \sum_{\ell=[\frac{m}{a}]+1}^{m} \binom{m}{\ell} \exp -\vartheta\log(p)[\rho(\ell - d) + \delta] = o(p^{-1})$$

for $m_a = a\ell_c$.

When $\frac{\log(K)}{\log(p)} \ll \rho$, for constant a, $2eKp^{-\rho\vartheta/a} < 1$. Then the left hand side of the above equation is simplified as

$$p^{1+d\rho\vartheta-\delta\vartheta} \sum_{m=m_a+1}^{p} (2eKp^{-\rho\vartheta/a})^m = L_p \cdot p^{1+d\rho\vartheta-\delta\vartheta}(2eKp^{-\rho\vartheta/a})^{m_a}.$$

Thus we only need to prove that for $m_a = a\ell_c$,

$$m_a[\frac{\rho\vartheta}{a}\log(p) - \log(2eK)] - (1 + d\rho\vartheta - \delta\vartheta)\log(p) \geq \log(p).$$

Finally we get

$$m_a \geq \frac{2 + d\rho\vartheta - \delta\vartheta}{\rho\vartheta/a - \log(2eK)/\log(p)} = a\ell_c + o(1).$$

□

BIBLIOGRAPHY

[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] Besag, J. (1986). On the Statistical Analysis of Dirty Pictures. *J. Roy. Soc. Statist. B*, **48**(3), 259-302.

[4] Beskos, A., Papaspiliopoulos, O., Roberts, G. O., Fearnhead, P. (2006). Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). *J. Roy. Soc. Statist. B*, **68**(3), 333-382

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

[6] 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.

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

[8] 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.

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

[10] Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Molecular systems biology*, **3**(1).

[11] 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.

[12] 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.

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

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

[15] 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.

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

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

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

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

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

[21] Fan, Y., Jin, J. and Yao, Z. (2012). Optimal classification in sparse Gaussian graphic model *arXiv*:1212.5332.

[22] 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`

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

[24] Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., ... and Brown, P. O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Science Signaling*, **11**(12), 4241.

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

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

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

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

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

[30] Jacob, L., Obozinski, G., and Vert, J. P. (2009, June). Group lasso with overlap and graph lasso. *In Proceedings of the 26th Annual International Conference on Machine Learning*, 433-440. ACM.

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

[32] Jin, J., Zhang, C.-H. and Zhang, Q. (2012). Optimality of Graphlet Screening in high dimensional variable selection. *arXiv*:1204.6452.

[33] Lee, I., Li, Z. and Marcotte, E. M. (2007). An improved, bias-reduced probabilistic functional gene network of baker's yeast, Saccharomyces cerevisiae. *PLoS ONE*, **2**(10), e988

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

[35] Ke, T., Jin, J., and Fan. J (2012). Covariance assisted screening and estimation. *arXiv*:1205.4645.

[36] Kanehisa, Minoru and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, **28**(1), 27-30.

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

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

[39] 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.

[40] 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.

[41] Potts, R. B. (1952, January). Some generalized order-disorder transformations. *Proceedings of the Cambridge Philosophical Society* **48**(2), 106-109.

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

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

[44] Van De Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., ... and Bernards, R. (2002). A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, **347**(25), 1999-2009

[45] Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, **21**(1), 5-42.

[46] 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.

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

[48] Wei, Z. and Li, H. (2007). A Markov random field model for network-based analysis of genomic data *Bioinformatics*, **23**(12), 1537C1544.

[49] 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.

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

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

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

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

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