UPS Delivers Optimal Phase Diagram in High Dimensional Variable Selection

Pengsheng Ji[1] and Jiashun Jin[2]
[1]Department of Statistical Science, Cornell University
[2]Department of Statistics, Carnegie Mellon University

Abstract

We consider a linear regression model

$$Y = X\beta + z, \qquad z \sim N(0, I_n), \qquad X = X_{n,p},$$

where both p and n are large but $p > n$. The vector β is unknown but is sparse in the sense that only a small proportion of its coordinates is nonzero, and we are interested in identifying these nonzero ones. We model the coordinates of β as samples from a two-component mixture $(1-\epsilon)\nu_0 + \epsilon\pi$, and the rows of X as samples from $N(0, \frac{1}{n}\Omega)$, where ν_0 is the point mass at 0, π is a distribution, and Ω is a p by p correlation matrix which is unknown but is presumably sparse.

We propose a two-stage variable selection procedure which we call the *UPS*. This is a Screen and Clean procedure [32], in which we screen with the Univariate thresholding, and clean with the Penalized MLE. In many situations, the UPS possesses two important properties: Sure Screening and Separable After Screening (SAS). These properties enable us to reduce the original regression problem to many small-size regression problems that can be fitted separately. As a result, the UPS is effective both in theory and in computation.

We measure the performance of variable selection procedure by the Hamming distance, and use an asymptotic framework where $p \to \infty$ and $(\epsilon, \pi, n, \Omega)$ depend on p. We find that in many situations, the UPS achieves the optimal rate of convergence. We also find that in the (ϵ_p, π_p) space, there is a three-phase diagram shared by many choices of Ω. In the first phase, it is possible to recover all signals. In the second phase, exact recovery is impossible, but it is possible to recover most of the signals. In the third phase, successful variable selection is impossible. The UPS partitions the phase space in the same way that the optimal procedures do, and recovers most of the signals as long as successful variable selection is possible.

The lasso and the subset selection (also known as the L^1- and L^0-penalization methods, respectively) are well-known approaches to variable selection. However, somewhat surprisingly, there are regions in the phase space where neither the lasso nor the subset selection is rate optimal, even for very simple Ω. The lasso is non-optimal because it is too loose in filtering out fake signals (i.e. noise that is highly correlated with a signal), and the subset selection is non-optimal because it tends to kill one or more signals when signals appear in pairs, triplets, etc..

Keywords: Graph, Hamming distance, lasso, Stein's normal means model, penalization methods, phase diagram, regression, Screen and Clean, subset selection, variable selection.

AMS 2000 subject classifications: Primary-62J05, 62J07; secondary 62G20, 62C05.

Acknowledgments: JJ thanks David Donoho and Alan Friez for valuable pointers and discussion. PJ and JJ were partially supported by NSF CAREER award DMS-0908613. PJ was also partially supported by NSF grant DMS-08-05632.

1 Introduction

Consider a sequence of regression problems indexed by the number of variables p:

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

where $X^{(p)}$ is an n_p by p matrix, and $\beta^{(p)}$ is a p by 1 vector of interest. We assume both p and n_p are large but $p > n_p$. The vector $\beta^{(p)}$ is unknown to us, but is sparse in the sense that it has s_p nonzeros where $s_p \ll p$. We are interested in variable selection: determining which components of $\beta^{(p)}$ are nonzero. In this paper, we use p as the driving asymptotic parameter. We suppress the superscript $^{(p)}$ and subscript p whenever there is no confusion.

A well-known approach to variable selection is *subset selection*, also known as the L^0-penalization method (see for example AIC [2], BIC [27], and RIC [19]). This approach selects variables by minimizing the following functional:

$$\frac{1}{2}\|Y - X\beta\|_2^2 + \frac{(\lambda^{ss})^2}{2}\|\beta\|_0, \tag{1.2}$$

where $\lambda^{ss} > 0$ is a tuning parameter and $\|\cdot\|_q$ denotes the L^q-norm. The approach has some good properties, but the optimization problem (1.2) is known to be NP hard, which prohibits the use of this approach when p is large.

In the middle 90's, Chen *et al.* [7] (see also Tibshirani [29]) proposed a trail-breaking approach which is now known as the Basis Pursuit or the lasso. This approach selects variables by minimizing a similar functional, but $\|\beta\|_0$ is replaced by $\|\beta\|_1$:

$$\frac{1}{2}\|Y - X\beta\|_2^2 + \lambda^{lasso}\|\beta\|_1. \tag{1.3}$$

A major advantage of the lasso is that (1.3) can be efficiently solved by the Interior Point method [7] even when p is relatively large. Additionally, in a series of papers [10, 12, 13, 16], it was shown that in the noiseless case (i.e. $z = 0$), the lasso solution is also the subset selection solution, provided that β is sufficiently sparse. For these reasons, the lasso procedure is passionately embraced by statisticians, engineers, biologists, and many others.

With that being said, we note that an obvious shortcoming of these methods is that, the penalization term does not reflect the correlation structure in X, which prohibits the method from fully capturing the essence of the regression model (e.g. Zou [36]). However, on a second thought, this shortcoming is largely due to the fact that these methods are *one-stage* procedures. The limitation can be overcome if we use a *two-stage* or a *multi-stage* procedure.

1.1 Screen and Clean: a two-stage method

An idea introduced in the 1960's, Screen and Clean has seen a revival in recent years (see for example Wasserman and Roeder [32] and Fan and Lv [18]). Screen and Clean is a two-stage method. At the first stage, we screen in hope of removing as many irrelevant variables as possible while keeping all relevant ones. At the second stage, we re-investigate the surviving variables in hope of removing all false positives. The screening stage has advantages in the following aspects, some of which are elaborated in the literature:

- *Dimension reduction.* We remove many irrelevant variables from the model, reducing the dimension from p to a much smaller number [18, 32].

- *Correlation complexity reduction.* A variable may be significantly correlated to many other variables. But few of the correlated variables would survive the screening. A surviving variable is only correlated with a few other surviving ones.

- *Computation complexity reduction.* Under certain conditions (to be elaborated in Section 2), the set of surviving variables splits into many units. Each unit is small in size (e.g. $\leq K$), and correlation between units is very weak. Therefore, these units can be fitted separately, with the total computational cost $\leq$ # of units $\times 2^K$.

Despite the perceptive vision and philosophical importance in these works [18, 32], substantial vagueness remains: How to screen? How to clean? Does Screen and Clean really have an advantage over the lasso and the subset selection? This is where the UPS comes in.

1.2 UPS: a Screen and Clean method

UPS stands for *a two-stage method which screens by* **U***nivariate thresholding and cleans by* **P***enalized MLE for* **S***electing variables.* UPS contains a U-step and a P-step.

In the U-step, we screen with Univariate thresholding [11] (also known as marginal regression [22] and Sure Screening [18]). This method is originated in the 1970's, but has seen a revival recently [18, 22, 32]. Fix a threshold $t > 0$ and let x_j be the j-th column of X. We remove the j-th variable from the regression model if and only if $|(x_j, Y)| < t$. The set of surviving indices is then

$$\mathcal{U}_p(t) = \mathcal{U}_p(t; Y, X) = \{j : |(x_j, Y)| \geq t, \ 1 \leq j \leq p\}.$$

Despite its simplicity, the U-step can be effective in many situations. In Section 2, we show that $\mathcal{U}_p(t)$ has the following important properties that ensure the success of the P-step.

- *Sure Screening.* With overwhelming probability, $\mathcal{U}_p(t)$ includes all but a negligible proportion of the signals (i.e. indices at which β has a nonzero coordinate). While it might be more accurate to call this *Screening with acceptable loss*, we call it *Sure Screening* for consistency with the literature [18].

- *Separable After Screening (SAS).* View each index $(1, 2, \ldots, p)$ as a node of a graph. We say two nodes j and k are connected if the corresponding columns x_j and x_k are "significantly" correlated. With overwhelming probability, $\mathcal{U}_p(t)$ splits into many disconnected subsets, each of which is a small-size maximal connected subgraph of $\mathcal{U}_p(t)$.

We now explain how these properties pave the way for the P-step. Let $\mathcal{I}_0 = \{i_1, \ldots, i_K\}$ and $\mathcal{J}_0 = \{j_1, \ldots, j_L\}$ be two subsets of $\{1, 2, \ldots, p\}$, $1 \leq K, L \leq p$. We have the following definition.

Definition 1.1 *For any p by 1 vector Y, $Y^{\mathcal{I}_0}$ denotes the K by 1 vector such that $Y^{\mathcal{I}_0}(k) = Y_{i_k}$, $1 \leq k \leq K$. For any p by p matrix Ω, $\Omega^{\mathcal{I}_0, \mathcal{J}_0}$ denotes the K by L matrix such that $\Omega^{\mathcal{I}_0, \mathcal{J}_0}(k, \ell) = \Omega(i_k, j_\ell)$, where $1 \leq k \leq K, 1 \leq \ell \leq L$.*

Note that the regression model is closely related to the model

$$X'Y = X'X\beta + X'z.$$

Restricting the attention to $\mathcal{U} = \mathcal{U}_p(t)$, we have

$$(X'Y)^{\mathcal{U}} = (X'X\beta)^{\mathcal{U}} + (X'z)^{\mathcal{U}} = (X'X)^{\mathcal{U},\mathcal{V}}\beta + (X'z)^{\mathcal{U}},$$

where $\mathcal{V} = \{1, 2, \ldots, p\}$. Three key observations are:

- By $z \sim N(0, I_n)$, $(X'z)^{\mathcal{U}} \sim N(0, (X'X)^{\mathcal{U},\mathcal{U}})$.
- By the Sure Screening property, $(X'X)^{\mathcal{U},\mathcal{V}}\beta \approx (X'X)^{\mathcal{U},\mathcal{U}}\beta^{\mathcal{U}}$.
- By the SAS property, $(X'X)^{\mathcal{U},\mathcal{U}}$ approximately equals a block diagonal matrix, where each block corresponds to a maximal connected subgraph contained in $\mathcal{U}_p(t)$.

As a result, the regression problem reduces to many small-size regression problems, each corresponding to a maximal connected subgraph. These small-size regression problems can be solved separately, each at a modest computation cost.

In detail, fix two parameters $\lambda^{ups} > 0$ and $u^{ups} > 0$. Let $\mathcal{I}_0 = \{i_1, i_2, \ldots, i_K\} \subset \mathcal{U}_p(t)$ be a maximal connected subgraph, and let μ be a K by 1 vector the coordinates of which are either 0 or u^{ups}. Let $\hat{\mu}(\mathcal{I}_0) = \hat{\mu}(\mathcal{I}_0; Y, X, t, \lambda^{ups}, u^{ups}, p)$ be the minimizer of the following functional:

$$\frac{1}{2}\big((X'Y)^{\mathcal{I}_0} - (X'X)^{\mathcal{I}_0,\mathcal{I}_0}\mu\big)'\big((X'X)^{\mathcal{I}_0,\mathcal{I}_0}\big)^{-1}\big((X'Y)^{\mathcal{I}_0} - (X'X)^{\mathcal{I}_0,\mathcal{I}_0}\mu\big) + \frac{1}{2}(\lambda^{ups})^2\|\mu\|_0. \quad (1.4)$$

Combining all such estimates across different maximal connected subgraphs in $\mathcal{U}_p(t)$ gives the final estimate of the UPS, denoted by $\hat{\beta}^{ups} = \hat{\beta}^{ups}(Y, X; t, \lambda^{ups}, u^{ups}, p)$:

$$\hat{\beta}_j^{ups} = \begin{cases} (\hat{\mu}(\mathcal{I}_0))_k, & \text{if } j = i_k \in \mathcal{I}_0 \text{ for some } \mathcal{I}_0 = \{i_1, i_2, \ldots, i_K\} \subset \mathcal{U}_p(t), \\ 0, & \text{if } j \notin \mathcal{U}_p(t_p). \end{cases}$$

The UPS uses three tuning parameters, one in the U-step, and two in the P-step. The tuning parameter in the U-step t has a similar role to that in Sure Screening [18]. We find that in many situations, the performance of the UPS is relatively insensitive to the choice of t, as long as it falls in a certain range. The tuning parameter in the P-step λ^{ups} has a similar role to that of the lasso λ^{lasso} or that of the subset selection λ^{ss}. However, there is a major difference: λ^{ups} can be conveniently estimated using the data, whereas how to set (λ^{ss}, λ^{lasso}) remains a challenging problem in the literature. See Section 2 for more discussion.

Come back to the questions raised in Section 1.1. The UPS indeed has advantages over the lasso and the subset selection. In Sections 1.3–1.7, we establish a theoretic framework and investigate these procedures closely. The main finding is: for a wide range of design matrix X, the Hamming distance of the UPS achieves the optimal rate of convergence (i.e. rate optimal). In contrast, the lasso and the subset selection may be rate non-optimal, even for very simple design matrices.

1.3 Sparse signal model, Hamming distance, and universal lower bound

Fix $\epsilon \in (0,1)$ and a distribution π. We write $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ and suppose

$$\beta_j \overset{iid}{\sim} (1-\epsilon)\nu_0 + \epsilon\pi, \qquad 1 \leq j \leq p, \tag{1.5}$$

where ν_0 is the point mass at 0, and π has no mass at 0. We allow (ϵ, π) to depend on p, and let

$$\epsilon = \epsilon_p = p^{-\vartheta}, \qquad 0 < \vartheta < 1, \tag{1.6}$$

so that the expected number of signals is

$$s_p = p\epsilon_p = p^{1-\vartheta}. \tag{1.7}$$

Note that as p grows, the vector β gets increasingly sparse. For any variable selection procedure $\hat{\beta} = \hat{\beta}(Y|X)$, we measure the loss by the Hamming distance

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

If the design matrix X is random and has a multivariate distribution F, then the overall Hamming distance is

$$\mathrm{Hamm}_p(\hat{\beta}, \beta) = E_F[h_p(\hat{\beta}|X)].$$

In the context of variable selection, the Hamming distance is a natural choice for loss function. The main results in this paper can be extended if we replace the Hamming distance by the L^q-loss. The following notation is frequently used in this paper.

Definition 1.2 *$L_p > 0$ is a multi-$\log(p)$ term which may change from occurrence to occurrence, such that for any fixed $\delta > 0$, $\lim_{p\to\infty} L_p \cdot p^{\delta} = \infty$ and $\lim_{p\to\infty} L_p p^{-\delta} = 0$.*

Somewhat surprisingly, it turns out that for all sample size n and design matrix X, there is a *universal* lower bound for the Hamming distance. In detail, fixing $r > 0$, we introduce

$$\tau_p = \tau_p(r) = \sqrt{2r\log p}, \tag{1.8}$$

and

$$\lambda_p = \lambda_p(\epsilon_p, \tau_p) = \frac{1}{\tau_p}\big[\log(\frac{1-\epsilon_p}{\epsilon_p}) + \frac{\tau_p^2}{2}\big]. \tag{1.9}$$

Let $\bar{\Phi} = 1 - \Phi$ be the survival function of $N(0,1)$. The following theorem is proved in Section 6.

Theorem 1.1 *(Universal Lower bound). Fix $\vartheta \in (0,1)$, $r > 0$, and a sufficiently large p. Let ϵ_p, s_p, and τ_p be as in (1.6)–(1.8), and suppose the support of π_p is contained in $[-\tau_p, 0) \cup (0, \tau_p]$. For any fixed n and matrix $X^{(p)}$ such that $(X^{(p)})'X^{(p)}$ has unit diagonals,*

$$\frac{h_p(\hat{\beta}, \beta|X)}{s_p} \geq \Big[\frac{1-\epsilon_p}{\epsilon_p}\bar{\Phi}(\lambda_p) + \Phi(\tau_p - \lambda_p)\Big]. \tag{1.10}$$

Note that as $p \to \infty$,

$$\left[\frac{1-\epsilon_p}{\epsilon_p}\bar{\Phi}(\lambda_p) + \Phi(\tau_p - \lambda_p)\right] \geq \begin{cases} L_p \cdot p^{-\frac{(r-\vartheta)^2}{4r}}, & r > \vartheta, \\ (1+o(1)), & r < \vartheta. \end{cases} \tag{1.11}$$

The case $r = \vartheta$ is more delicate, and we skip the discussion. Theorem 1.1 continues to hold if $X \sim F$ is random, provided that we replace $h_p(\hat{\beta}, \beta|X)$ by $\text{Hamm}_p(\hat{\beta}, \beta)$ in (1.10).

It may seem counter-intuitive that the lower bound does not depend on n, but this is due to the way we normalize X. Also, in the case of orthogonal design (i.e., $X'X = I_p$), the lower bound can be achieved by either the lasso or marginal regression [22]. Therefore, among all design matrices, the orthogonal design is among the best in terms of the convergence rate.

Theorem 1.1 implies that if we have $p^{1-\vartheta}$ signals and the maximal signal strength is slightly smaller than $\sqrt{2\vartheta \log(p)}$, then asymptotically the Hamming distance of any procedure can not be substantially smaller than s_p, and so successful variable selection is impossible. In the sections below, we focus on the case where the signal strength is larger than $\sqrt{2\vartheta \log(p)}$, so that successful variable selection is possible.

The universality of the lower bound hints that it might not be tight once X is not orthogonal. Fortunately, it turns out that in many interesting cases, the lower bound is tight. To facilitate the analysis, we invoke the random design model.

1.4 Random design and connections to Stein's normal means model

Write

$$X = (x_1, x_2, \ldots, x_p) = \begin{pmatrix} X_1' \\ \ldots \\ X_n' \end{pmatrix}.$$

We model X_i as iid samples from a p-variate zero-mean Gaussian distribution,

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

The p by p matrix $\Omega = \Omega^{(p)}$ is unknown but for simplicity we assume that it has unit diagonals. The normalizing constant $1/n$ is chosen so that the diagonals of the Gram matrix $X'X$ are approximately 1. Fixing $\theta \in (1-\vartheta, 1)$, we let

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

Note that $s_p \ll n_p \ll p$ as $p \to \infty$. For successful variable selection, it is almost necessary to have $s_p \ll n_p$ [11].

In the literature, Model (1.1) and (1.12) is referred to as the *random design model*, which may arise in the following application areas.

- *Compressive Sensing.* The vector of interest, β, is very high dimensional and presumably sparse. We plan to measure n general linear combinations of β and then reconstruct it. For $1 \leq i \leq n$, we choose a p by 1 coefficient vector X_i and observe $Y_i = X_i'\beta + z_i$, where $z_i \sim N(0, \sigma^2)$ is measurement error. For computation and storage concerns, one usually chooses X_i's as simple as possible. Popular choices of X_i include Gaussian design, Bernoulli design, Circulant design, etc. [11, 3]. Model (1.12) corresponds to the Gaussian design.

- *Privacy-preserving data mining.* The vector β may contain some confidential information (e.g. HIV-diagnosis results of a community) that we must protect. While we can not release the whole vector, we must allow data mining to some extent, because, for example, the study is of public interest and is supported by federal funding. To compromise, we allow queries as follows. For each query, the database randomly generates a p by 1 vector X_i, and releases both X_i and $Y_i = X_i'\beta + z_i$ to the querier, where $z_i \sim N(0, \sigma^2)$ is a noise term. For privacy concern, the number of allowed queries is much smaller than p. Popular choices of X_i include Gaussian design and Bernoulli design [8].

Random design model is closely related to the well-known Stein's normal means model

$$W \sim N(\beta, \Sigma), \tag{1.14}$$

where $\Sigma = \Omega^{-1}$ is a covariance matrix. Stein's normal means model is of interest in many areas including nonparametric estimation, wavelets, and large-scale multiple testing [33]. To see how two models are related, recall that Model (1.1) is closely related to the model $X'Y = X'X\beta + X'z$. Since the rows of X are iid samples from $N(0, \frac{1}{n}\Omega)$ and $s_p \ll n_p \ll p$, we expect to see that

$$X'X\beta \approx \Omega\beta, \qquad X'z \approx (\frac{\sqrt{n}}{\|z\|}) \cdot X'z \sim N(0, \Omega).$$

It follows that $X'Y \approx N(\Omega\beta, \Omega)$, which is Model (1.14). See Lemma 3.1 for further discussion. Therefore, Stein's normal means model can be viewed as an idealized version of the random design model. The close relationship between the two models suggests that solving the variable selection problem opens doors for solving other problems, and vice versa. In the forthcoming manuscript [5], we forge a link between variable selection and large-scale multiple testing.

1.5 Optimality of the UPS and upper bound for the Hamming distance

We now come to the main results of this paper. To state such results (Theorems 2.1–2.2) in precise terms, we need relatively long preparations. Below, we sketch the main results, and leave the formal statements to Section 2.

In Model (1.1), (1.5), and (1.12), let (s_p, τ_p, n_p) be as in (1.7), (1.8), and (1.13). Suppose

- *Sparsity of* Ω. Each row of Ω satisfies a certain summability condition, so it has relatively few large coordinates.
- The support of π_p is contained in $[\tau_p, (1+\eta)\tau_p]$, where η is a constant to be defined later. Suppose $r > \vartheta$, so that the strengths of all signals are greater than $\sqrt{2\vartheta \log(p)}$ (this is the critical value below which successful variable selection is impossible; see Theorem 1.1).
- Either all coordinates of Ω are positive, or that $r/\vartheta \le 3 + 2\sqrt{2}$ (this ensures that the so-called "signal cancellation" phenomenon does not have a major effect [32]).

If we fix $0 < q \le (\vartheta + r)^2/(4r)$, and set the UPS tuning parameters by

$$t_p^* = t_p^*(q) = \sqrt{2q\log p}, \qquad \lambda^{ups} = \lambda_p^{ups} = \sqrt{2\vartheta \log(p)}, \qquad u^{ups} = u_p^{ups} = \tau_p,$$

then the ratio between the Hamming error of the UPS and s_p does not exceed $L_p p^{-(\vartheta-r)^2/(4r)}$, where the exponent matches that of the lower bound in Theorem 1.1. Therefore, the lower bound in Theorem 1.1 is tight and the UPS is rate optimal.

Furthermore, if $\vartheta < q \leq (\vartheta + r)^2/(4r)$, then we can estimate the parameters λ_p^{ups} and u_p^{ups}, and plug them into the UPS. The plug-in UPS continues to be rate optimal. This says that the UPS needs essentially only one tuning parameter t_p^*. The behavior of the UPS is relatively insensitive to the choice of t_p^*. While this is largely asymptotic, numerical study for finite p confirms such insensitivity. See Theorems 2.1–2.2 for details.

1.6 Phase diagram for high dimensional variable selection

Theorems 1.1, 2.1, and 2.2 reveal a watershed phenomenon. Suppose we have roughly $s_p = p^{1-\vartheta}$ signals. Then if the maximal signal strength is slightly smaller than $\sqrt{2\vartheta \log p}$, then the Hamming distance of any procedure can not be substantially smaller than s_p, hence successful variable selection is impossible. If the minimal signal strength is slightly larger than $\sqrt{2\vartheta \log p}$, then there exist procedures (UPS is one of them) whose Hamming distances are substantially smaller than s_p, and they manage to recover most of the signals.

The phenomenon is best caricatured in the special case where π_p is a point mass at τ_p,

$$\pi_p = \nu_{\tau_p},$$

where $\tau_p = \sqrt{2r \log p}$ is as in (1.8). If we call the two-dimensional domain $\{(\vartheta, r) : \ 0 < \vartheta < 1, \ r > 0\}$ the *phase space*, then the theorems say that the phase space is partitioned into three sub-regions as follows:

- *Region of No Recovery* $(0 < \vartheta < 1, 0 < r < \vartheta)$. In this region, the Hamming distance of any procedure $\geq (1 + o(1))s_p$, and successful variable selection is impossible.
- *Region of Almost Full Recovery* $(0 < \vartheta < 1, \ \vartheta < r < (1 + \sqrt{1-\vartheta})^2)$. In this region, there are procedures (and UPS is one of them) whose Hamming distances are much larger than 1 but are smaller than $L_p \cdot s_p \cdot p^{-(\vartheta-r)^2/(4r)}$. In this region, it is possible to recover most of the signals, but not all of them.
- *Region of Exact Recovery* $(0 < \vartheta < 1, \ r > (1 + \sqrt{1-\vartheta})^2)$. In this region, there are procedures (and UPS is one of them) that recover all signals with probability ≈ 1.

Note that the partitions are the same for many choices of Ω. Figure 1 shows these sub-regions. Because of the partition of the phases, we call this the phase diagram. The UPS is optimal in the sense that it partitions the phase space in exactly the same way as do the optimal procedures.

The phase diagram provides a benchmark for all variable selection procedures. The lasso would be optimal if it partitions the phase space in the same way as in Figure 1. Unfortunately, this is not the case, even when Ω has relatively simple structures. Below we investigate a case where $X'X$ is a tridiagonal matrix, and identity precisely the regions where the lasso is rate optimal and where it is rate non-optimal. What might be more surprising is that there is a region in the phase space where the subset selection is also rate non-optimal.

1.7 Non-optimal region for the lasso

In this and the next section, we leave the random design model (1.12) and consider a Stein's normal means model, which can be viewed as an idealized version of the former. While

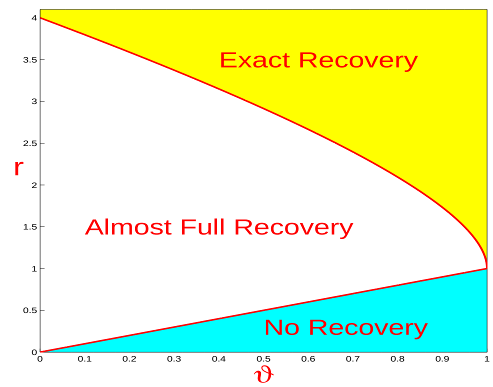

Figure 1: Phase diagram. In Region of Exact Recovery, the UPS recovers all signals with high probability. In Region of Almost Full Recovery, it is possible (e.g. by the UPS) to recover almost all signals, but impossible to recover all of them. In Region of No Recovery, the Hamming distance of any procedure $\gtrsim s_p$ and successful variable selection is impossible.

considering the idealized version is mainly for mathematical convenience, the insight gained is valid in much broader settings. Also, since the goal here is to illustrate that the lasso and the subset selection are rate non-optimal, considering an idealized version is sufficient (we use a broader setting in previous sections as the goal is to show the optimality of the UPS). The point is that, if a procedure is non-optimal in simple cases, we should not expect them to be optimal in more complicated cases.

In this spirit, we consider a Stein's normal means model

$$\tilde{Y} \equiv X'Y \sim N(\Omega\beta, \Omega), \tag{1.15}$$

where β is as in (1.5) with $\tau_p = \nu_{\pi_p}$ and $\pi_p = \sqrt{2r\log(p)}$. To further simplify the study, we take Ω as the tridiagonal matrix:

$$T(a) = \begin{pmatrix} 1 & a & \dots & 0 \\ a & 1 & \dots & 0 \\ 0 & 0 & \dots & 1 \end{pmatrix}, \qquad 0 < a < 1/2. \tag{1.16}$$

The case $a < 0$ can be discussed similarly; see Section 4 for details. In this case, the UPS partitions the phase space optimally.

We now discuss the phase diagram of the lasso. As the discussion is relatively long, we briefly describe the results below, and leave the details to Section 4. We focus on the region $\{(\vartheta, r) : 0 < \vartheta < 1, r > \vartheta\}$ in the phase space. This lies above the Region of No Recovery, where successful variable selection is possible. The region can be further partitioned into three sub-regions as follows (see Figure 2).

- *Non-optimal region of the lasso*: $0 < \vartheta < 2a(1+a)^{-1}$ and $\frac{1}{a}(1+\sqrt{1-a^2})\vartheta < r < \left(1+\sqrt{\frac{1+a}{1-a}}\right)^2(1-\vartheta)$. In this region, the lasso is rate non-optimal (i.e., the Hamming distance is $L_p \cdot p^c$ with constant $c > 1-(\vartheta+r)^2/(4r))$, even when the tuning parameter is set ideally.

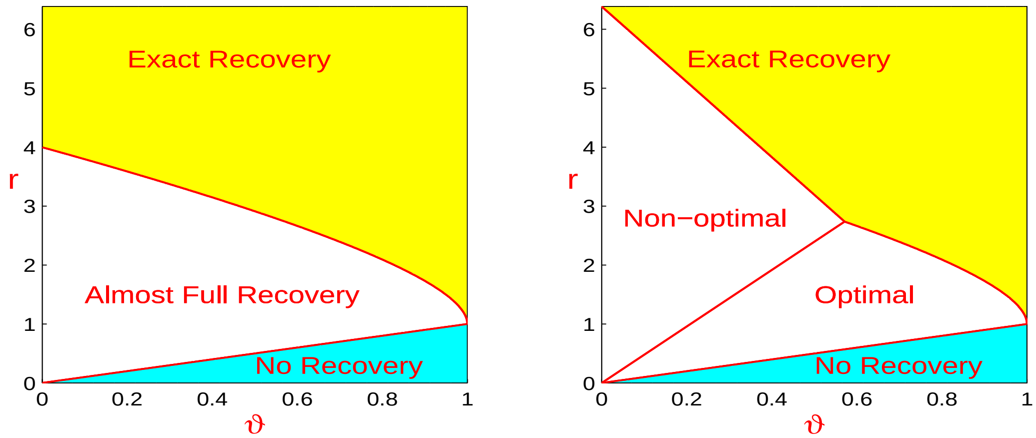

Figure 2: Left: a re-display of Figure 1 for comparison. Right: partition of the phase space by the lasso for the tridiagonal model (1.15)-(1.16) ($a = 0.4$). The lasso is rate non-optimal in the Non-optimal region. The Region of Exact Recovery by the lasso is substantially smaller than that of the optimal procedure, displayed on the left.

- *Optimal region of the lasso*: $0 < \vartheta < 1$ and $\vartheta < r < \frac{1}{a}(1 + \sqrt{1 - a^2})\vartheta$ and $r < (1 + \sqrt{1 - \vartheta})^2$. In this region, if additionally $a \geq 1/3$, then the lasso may be rate optimal in this region provided that the tuning parameter is set ideally. The case $0 < a < 1/3$ is relatively messy. As the focus is on when the lasso is non-optimal, we skip further discussion in this case.
- *Region of Exact Recovery by the lasso*: $0 < \vartheta < 1$ and $r > (1 + \sqrt{1 - \vartheta})^2$ and $r > \left(1 + \sqrt{\frac{1+a}{1-a}}\right)^2 (1 - \vartheta)$. In this region, provided that the tuning parameter is set ideally, the lasso may have a Hamming error of $o(1)$ and yield exact recovery with high probability. Region of the Exactly Recovery by the lasso is substantially smaller than that of the optimal procedure. There is a sub-region in the phase space where the optimal procedure yields exact recovery, but the lasso does not even when the tuning parameter is set ideally.

For discussions in the case where Ω is the identity matrix, compare [22, 31]. The results above are based on Theorem 4.1 in Section 4, where we derive a lower bound for the Hamming errors by the lasso. In [26], we show that the lower bound is tight for properly large ϑ, but is not when ϑ is small. It is, however, tight for all $\vartheta \in (0, 1)$ if we replace Model (1.5) by a closely related model, namely (2.2)–(2.3) in [23]. For these reasons, the non-optimal region of the lasso may be larger than that illustrated in Figure 2 (the Region of Exact Recovery and the optimal region are slightly smaller than those illustrated in Figure 2, respectively). The calculation of the exact optimal rate of convergence for the lasso is tedious. Since the main purpose of this section is to show that the lasso is non-optimal in certain regions, we skip further discussion along this line.

We now shed some light on why the lasso is non-optimal.

Definition 1.3 *We say that $\tilde{Y}_j$ is a signal if $\beta_j \neq 0$, is a fake signal if $(\Omega\beta)_j \neq 0$ and $\beta_j = 0$, and is a (pure) noise if $\beta_j = (\Omega\beta)_j = 0$.*

Intellectually, a fake signal is a noise that may look like a signal due to correlation. With the tuning parameter set ideally, the lasso is able to distinguish signals from pure noise, but it may fail to filter out fake signals efficiently, which then causes its non-optimality. In the optimal region of the lasso, the number of falsely kept fake signals is much smaller than the optimal Hamming distance, so it only has a negligible effect. In the non-optimal region of the lasso, the number of falsely kept fake signals is much larger than the optimal Hamming distance, and so is non-negligible.

The above insight suggests that when $X'X$ moves away from the tridiagonal case, the partitions of the sub-regions by the lasso may change, but the non-optimal region of the lasso continues to exist in rather general situations.

1.8 Non-optimal region for the subset selection

We now move to subset selection. The discussion is similar to that for the lasso so we keep it brief. Introduce

$$v_1(a) = \frac{2 - \sqrt{1-a^2}}{\sqrt{1-a^2}(1-\sqrt{1-a^2})}, \qquad v_2(a) = 2\sqrt{1-a^2} - 1.$$

- *Non-optimal region of the subset selection:* $0 < \vartheta < \frac{4v_1(a)}{(v_1(a)+1)^2}$ and $v_1(a)\vartheta < r < \left[\frac{1}{v_2(a)}\left(\sqrt{1-2\vartheta} + \sqrt{1-2\vartheta+\vartheta v_2(a)}\right)\right]^2$. In this region, the subset selection is rate non-optimal even when the tuning parameter is set ideally.
- *Optimal region of the subset selection:* $0 < \vartheta < 1$ and $\vartheta < r < v_1(a)\vartheta$ and $r < (1+\sqrt{1-\vartheta})^2$. In this region, the subset selection may be rate optimal provided that the tuning parameter is set ideally.
- *Exact Recovery region of the subset selection:* $0 < \vartheta < 1$, $r > (1+\sqrt{1-\vartheta})^2$ and $r > \left[\frac{1}{v_2(a)}\left(\sqrt{1-2\vartheta} + \sqrt{1-2\vartheta+\vartheta v_2(a)}\right)\right]^2$. In this region, provided that the tuning parameter is set ideally, the subset selection may have Hamming errors of the order $o(1)$ and yield exact recovery with high probability. The region is substantially smaller than its counterpart of the optimal procedure.

See Figure 3. The results are based on Theorem 4.2 where we derive a lower bound for the subset selection. Similar to the remarks in Section 1.7, the Region of Exact Recovery and the optimal region of the subset selection may be smaller than those illustrated in Figure 3.

Why is the subset selection non-optimal? The reason is very different from that for the lasso: the lasso is non-optimal for it is too loose on fake signals, but the subset selection is non-optimal for it is too harsh on signal clusters (pairs/triplets, etc.). With the tuning parameter set ideally, the subset selection is effective in filtering out fake signals, but it also tends to kill one or more signals when the true signals appear in clusters. These falsely killed signals account for the non-optimality of the subset selection. See Section 4.2 for details.

1.9 Connection to recent literature

This work is related to recent literature on oracle property [36, 25], but is different in important ways. A procedure is said to have the oracle property if it recovers all signals. However, exact recovery is rarely seen in practice, especially in applications where $p \gg n$

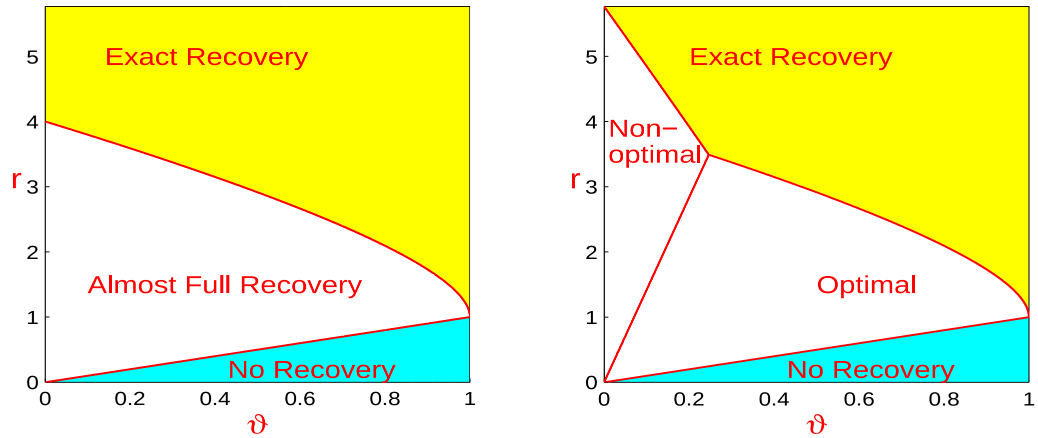

Figure 3: Left: a re-display of Figure 1 for comparison. Right: partition of the phase space by the subset selection in the tridiagonal model (1.15)-(1.16) ($a = 0.4$). The subset selection is not rate optimal in the Non-optimal region. The Exact Recovery region by the subset selection is substantially smaller than that of the optimal procedure, displayed on the left.

(e.g. genomics). In such applications, a large p usually means that signals are sparse, and a small n usually means signals are weak. When the signals are sparse and weak, exact recovery is usually impossible; see Figure 1. As a result, it is more relevant to compare error rates of different procedures than to investigate when they satisfy the oracle property. After all, if the signals are strong enough, many procedures may satisfy the oracle property, even if they are not rate optimal.

The work is also related to [6, 34]. These papers adopted an asymptotic minimax framework and showed that the lasso is rate optimal. While their results seem to contradict with ours, the difference can be easily reconciled. In the minimax approach, the asymptotic least favorable distribution of β is given by $\beta_j \stackrel{iid}{\sim} (1-\epsilon_p)\nu_0 + \epsilon_p \nu_{\tau_p}$, where $\epsilon_p = p^{-\vartheta}$, $\tau_p = \sqrt{2r \log p}$, and

$$\vartheta = r,$$

which is the boundary line of the Region of No Recovery in the phase space. See for example [34, Page 18-19], and [1, Section 3]. This suggests that the minimax approach has a limitation: it reduces the analysis to the worst scenario, but the worst scenario may be outside the range of interest. How to overcome this limitation has been the theme of recent meetings (e.g. [24]). In our approach, we let (ϑ, r) range freely, and evaluate a procedure based on how it partitions the phase space. The approach has a similar spirit to that in [16].

This approach is also related to the adaptive lasso [36]. The adaptive lasso is similar to the lasso, but the L^1-penalty $\lambda^{lasso}\|\beta\|_1$ is replaced by the weighted L^1-penalty $\sum_{j=1}^p w_j|\beta_j|$, where $w = (w_1, \ldots, w_p)'$ is the weight vector. The adaptive lasso can be viewed as a Screen and Clean method, where we first set weights, and then clean. Our approach is different from the adaptive lasso in important ways. First, Zou [36] suggested weight choices by the least squares estimate, which may work when p is small, but may not when p is large. In the case of $p \gg n$, our results suggest the weights should be very sparse, but the least squares estimates are usually dense. Second, for the surviving indices, we first partition them into many disjoint units of small sizes, and then fit them individually. The adaptive lasso fits all surviving variables together, which is computationally more expensive. Last, we use Penalized MLE in the clean step while the adaptive lasso uses L^1-penalization. As pointed out in Section 1.7, the L^1-penalty in the clean step is too loose on fake signals, which prohibits the procedure from being rate optimal.

1.10 Contents

In summary, we propose the UPS as a two-stage method for variable selection. The UPS is optimal in the sense that it partitions the phase space in the same way as the optimal procedure. We use Univariate thresholding in the screening step for its exceptional convenience in computation, and we use Penalized MLE in the cleaning step because it is the only procedure we know so far that yields the optimal rate of convergence. The lasso and even the subset selection do not partition the phase space optimally. We identify regions in the phase space where such methods are rate non-optimal.

The remaining sections are organized as follows. Section 2 discusses the UPS procedure and the upper bound for the optimal rate of convergence. The section also addresses how to estimate the tuning parameters of the UPS, and the convergence rate of the resultant plug-in procedure. Section 3 discusses a refinement of the UPS for moderately large p. Section 4 discusses the behavior of the lasso and the subset selection. Section 5 discusses numerical results where we compare the UPS with the lasso (the subset selection is computationally infeasible for large p so is not included for comparison). Section 6 contains proofs of the results.

Below are some notations we use in this paper. Fix $0 \leq q \leq \infty$. For a p by 1 vector x, $\|x\|_q$ denotes the L^q-norm of x and we omit the subscript when $q = 2$. For a p by p matrix M, $\|M\|_q$ denotes the matrix L^q-norm, and $\|M\|$ denotes the spectral norm.

2 UPS and upper bound for the Hamming distance

In this section, we establish the upper bound for the Hamming distance and show that the UPS is rate optimal. We begin by discussing necessary notations and the U-step, followed by discussion on the Sure Screening property and the SAS property of the UPS. We then show how the regression problem reduces to many separate small-size regression problems, and explain the rationale of using the Penalized MLE in the P-step. We conclude the section by the rate optimality of the UPS, where the tuning parameters are either set ideally or estimated.

Since different parts of our model are introduced gradually in different subsections, we summarize them as follows. The model we consider is

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

where

$$X_i \overset{iid}{\sim} N(0, \frac{1}{n}\Omega), \qquad \beta_j \overset{iid}{\sim} (1 - \epsilon_p)\nu_0 + \epsilon_p \pi_p, \qquad 1 \leq i \leq n, \qquad 1 \leq j \leq p. \tag{2.2}$$

Fixing $\theta > 0$, $\vartheta > 0$, and $r > 0$, we calibrate

$$\epsilon_p = p^{-\vartheta}, \qquad \tau_p = \sqrt{2r \log p}, \qquad n_p = p^{\theta}, \tag{2.3}$$

assuming that

$$\theta < (1 - \vartheta). \tag{2.4}$$

Recall that the optimal rate of convergence is $L_p p^{1-(\vartheta+r)^2/(4r)}$. In this section, we focus on the case where the exponent $1-(\vartheta+r)^2/(4r)$ falls between 0 and $(1-\vartheta)$, or equivalently,

$$\vartheta < r < (1+\sqrt{1-\vartheta})^2. \tag{2.5}$$

In the phase space, this corresponds to the Region of Almost Full Recovery. The case $r < \vartheta$ corresponds to Region of the No Recovery and is studied in Theorem 1.1. The case $r > (1+\sqrt{1-\vartheta})^2$ corresponds to the Exact Recovery Region. The discussion of this case is similar but is much simpler, so we omit it.

Next, fixing $A > 0$ and $\gamma \in (0,1)$, introduce

$$\mathcal{M}_p(\gamma, A) = \{\Omega: \quad \Omega \text{ is a } p \text{ by } p \text{ correlation matrix}, \sum_{j=1}^{p} |\Omega(i,j)|^\gamma \leq A, \ \forall\, 1 \leq i \leq p\}.$$

For any $\Omega \in \mathcal{M}_p(\gamma, A)$, let $U = U(\Omega)$ be the upper part of Ω excluding the diagonals, that is, $U(i,j) = \Omega(i,j)1_{\{i<j\}}$, and let

$$d(\Omega) = \max\{\|U(\Omega)\|_1, \ \|U(\Omega)\|_\infty\}.$$

Fixing $\omega_0 \in (0, 1/2)$, we introduce

$$\mathcal{M}_p^*(\omega_0, \gamma, A) = \{\Omega \in \mathcal{M}_p(\gamma, A)\colon \ d(\Omega) \leq \omega_0\},$$

and a subset of $\mathcal{M}_p^*(\omega_0, \gamma, A)$,

$$\mathcal{M}_p^+(\omega_0, \gamma, A) = \{\Omega \in \mathcal{M}_p^*(\omega_0, \gamma, A): \ \Omega(i,j) \geq 0 \text{ for all } 1 \leq i, j \leq p\}.$$

For any $\Omega \in \mathcal{M}_p^*(\omega_0, \gamma, A)$, the eigenvalues are contained in $(1-2\omega_0, 1+2\omega_0)$, so Ω is positive definite. When $\omega_0 > 1/2$, Ω may not be positive definite. For example, let $\Omega = T(a)$ as in (1.16). In this case, $\omega_0 = a$, and $T(a)$ is not necessarily positive definite when $a > 1/2$.

Last, denote

$$\eta = \eta(\vartheta, r, \omega_0) = \frac{\sqrt{\vartheta r}}{(\vartheta + r)\sqrt{1+2\omega_0}} \min\left\{\frac{2\vartheta}{r}, 1 - \frac{\vartheta}{r}, \sqrt{2(1-\omega_0)} - 1 + \frac{\vartheta}{r}\right\}. \tag{2.6}$$

We suppose the support of signal distribution π_p is contained in

$$[\tau_p, (1+\eta)\tau_p], \tag{2.7}$$

where $\tau_p = \sqrt{2r\log(p)}$ as in (1.8). This assumption is only needed for proving the main lemma of the P-step (Lemma 6.5), and can be relaxed for proving other lemmas. Also, we assume the signals are one-sided mainly for simplicity. The results can be extended to the case where the signals are two-sided.

We now discuss the U-step. As mentioned before, the benefits of the U-step are three-fold: dimension reduction, correlation complexity reduction, and computation cost reduction. The U-step is able to achieve these goals simultaneously because it satisfies the Sure Screening property and the SAS property, which we now discuss separately.

2.1 The Sure screening property of the U-step

Recall that in the U-step, we remove the j-th variable if and only if $|(x_j, Y)| < t$ for some threshold $t > 0$. For simplicity, we make a slight change and remove the j-th variable if and only if

$$(x_j, Y) < t.$$

When the signals are one-sided, the change makes negligible difference. Set the threshold t by

$$t_p^* = t_p^*(q) = \sqrt{2q\log(p)}, \qquad \text{where} \qquad 0 < q \leq \frac{(\vartheta + r)^2}{4r}. \tag{2.8}$$

The following lemma is proved in Section 6.

Lemma 2.1 *(Sure Screening). Consider Model (2.1)-(2.2) where (2.3)-(2.7) hold. Let t_p^* be as in* (2.8)*. If $\Omega^{(p)} \in \mathcal{M}_p^+(\omega_0, \gamma, A)$ for sufficiently large p, then as $p \to \infty$,*

$$\sum_{j=1}^{p} P(x_j'Y < t_p^*, \beta_j \neq 0) \leq L_p p^{1 - \frac{(\vartheta+r)^2}{4r}}.$$

The claim remains valid if $r/\vartheta \leq 3 + 2\sqrt{2}$ and $\Omega^{(p)} \in \mathcal{M}_p^*(\omega_0, \gamma, A)$ for sufficiently large p.

The implication is that the Hamming errors we make in the U-step are at most comparable to the optimal rate of convergence, and are therefore acceptable.

2.2 The SAS property of the U-step

We need some terminology in Graph Theory. A network or graph $N = (V, E)$ [9] consists of two finite sets V and E, where V is the set of *vertices* or *nodes*, and E is the set of *edges* or *links*. We need the following definitions.

Definition 2.1 *A walk in $N = (V, E)$ is a sequence of nodes where each adjacent pair are linked, and a path in $N = (V, E)$ is a walk where all nodes are distinct.*

Definition 2.2 *A component of a network (V, E) is a subnetwork $N' = (V', E')$ such that (a) for any pair of nodes $i \in V'$ and $j \in V'$, there is a path in N' connecting them, and (b) for any nodes $i \in V'$ and $k \notin V'$, there is no link between them in N.*

In other words, a component of a network is a maximal connected subnetwork. We can always divide a network into components each of which is disconnected to others.

For a subset $V' \subset V$, the graph (V, E) naturally induces a graph (V', E') where two nodes in V' are connected if and only if they are connected in (V, E). For simplicity, we write such (V', E') as (V', E). When we say $\mathcal{I}_0$ is a connected graph, we mean that $\mathcal{I}_0$ is a connected subgraph of (V, E) unless otherwise stated.

Definition 2.3 *Fixing a subset $\mathcal{I}_0 \subset V'$, we write $\mathcal{I}_0 \lhd V'$ if $\mathcal{I}_0$ is a component of (V', E).*

For any node $v \in V'$, there is a unique connected subgraph $\mathcal{I}_0$ such that $v \in \mathcal{I}_0 \lhd V'$.

Fix a p by p symmetric matrix Ω_0 which is presumably sparse. If we let $V_0 = \{1, 2, \ldots, p\}$ and say nodes i and j are linked if and only if $\Omega_0(i,j) \neq 0$, then we have a network $N = (V_0, \Omega_0)$. Fix $t > 0$. Recall that $\mathcal{U}_p(t)$ is the set of surviving indices in the U-step:

$$\mathcal{U}_p(t) = \mathcal{U}_p(t, Y, X) = \{j : \ (x_j, Y) \geq t, \ 1 \leq j \leq p\}. \tag{2.9}$$

Note that the induced graph $(\mathcal{U}_p(t), \Omega_0)$ splits into many different components.

Definition 2.4 *Fix an integer $K \geq 1$. We say that $\mathcal{U}_p(t)$ has the Separable After Screening (SAS) property with respect to (V_0, Ω_0, K) if each component of the graph $(\mathcal{U}_p(t), \Omega_0)$ has no more than K nodes.*

The SAS property is monotone in t. The following lemma is elementary, so we omit the proof.

Lemma 2.2 *Suppose $\mathcal{U}_p(t)$ has the SAS property with respect to (V_0, Ω_0, K). Then for all $s > t$, $\mathcal{U}_p(s)$ also has the SAS property with respect to (V_0, Ω_0, K).*

Back to Model (2.1)-(2.2), we hope to relate the regression setting to a graph (V_0, Ω_0), and use it to spell out the SAS property. Towards this end, we set

$$V_0 = \{1, 2, \ldots, p\}.$$

As for Ω_0, a natural choice is the matrix Ω in (2.2). However, the SAS property makes more sense if Ω_0 is sparse and known, while Ω is neither. To overcome this difficulty, we first estimate Ω by the empirical covariance matrix, and then regularize the latter with hard thresholding [33].

In detail, let $\hat{\Omega} = X'X$ be the empirical covariance matrix. Recall that

$$X = (X_1, X_2, \ldots, X_n)', \qquad X_i \sim N(0, \frac{1}{n}\Omega).$$

It is known [4] that there is a constant $C > 0$ such that with probability $1 - o(1/p^2)$, for all $1 \leq i, j \leq p$,

$$|\hat{\Omega}(i,j) - \Omega(i,j)| \leq C\sqrt{\log(p)}/\sqrt{n}. \tag{2.10}$$

When p is large, $\hat{\Omega}$ is a very noisy estimate, so we regularize it by hard thresholding:

$$\Omega^*(i,j) = \hat{\Omega}(i,j) 1_{\{|\hat{\Omega}(i,j)| \geq \log^{-1}(p)\}}. \tag{2.11}$$

The threshold $\log^{-1}(p)$ is chosen mainly for simplicity and can be replaced by $\log^{-a}(p)$, where $a > 0$ is a constant. The following lemma is a direct result of (2.10); we omit the proof.

Lemma 2.3 *Fix $A > 0$, $\gamma \in (0,1)$, and $\omega_0 \in (0, 1/2)$. As $p \to \infty$, for any $\Omega \in \mathcal{M}_p^*(\omega_0, \gamma, A)$, with probability of $1 - o(1/p^2)$, each row of Ω^* has no more than $2\log(p)$ nonzero coordinates, and $\|\Omega^* - \Omega\|_\infty \leq C(\log(p))^{-(1-\gamma)}$.*

Taking $\Omega_0 = \Omega^*$, we form a graph (V_0, Ω^*). The following lemma is proved in Section 6, which says that except for a negligible probability, $\mathcal{U}_p(t_p^*)$ has the SAS property.

Lemma 2.4 *(SAS). Consider Model (2.1)-(2.2) where (2.3)-(2.7) hold. Let t_p^* be as in (2.8). As $p \to \infty$, there is a constant $K > 0$ such that with probability $1 - L_p p^{-(\vartheta + r)^2/(4r)}$, $\mathcal{U}_p(t_p^*)$ has the SAS property with respect to (V_0, Ω^*, K).*

The constant K depends on the parameters $(\vartheta, \theta, r, q)$ and the parameters that define $\mathcal{M}_p^*(\omega_0, \gamma, A)$, but does not depend on p.

2.3 Reduction to many small-size regression problems

Together, the Sure Screening property and the SAS property make sure that the original regression problem reduces to many separate small-size regression problems. In detail, the SAS property implies that $\mathcal{U}_p(t_p^*)$ splits into many connected subgraphs, each is small in size, and different ones are disconnected. Given two disjoint connected subgraphs $\mathcal{I}_0$ and $\mathcal{J}_0$ where $\mathcal{I}_0 \lhd \mathcal{U}_p(t)$ and $\mathcal{J}_0 \lhd \mathcal{U}_p(t)$,

$$\Omega^*(i,j) = 0, \qquad \forall\, i \in \mathcal{I}_0,\ j \in \mathcal{J}_0. \tag{2.12}$$

Recall that the regression model (1.1) is closely related to the model $X'Y = X'X\beta + X'z$. Fixing a connected subgraph $\mathcal{I}_0 \lhd \mathcal{U}_p(t_p^*)$, we restrict our attention to $\mathcal{I}_0$ by considering

$$(X'Y)^{\mathcal{I}_0} = (X'X\beta)^{\mathcal{I}_0} + (X'z)^{\mathcal{I}_0}.$$

See Definition 1.1 for notations. Since $X_i \overset{iid}{\sim} N(0, \frac{1}{n}\Omega)$ and $\mathcal{I}_0$ has a small size, we expect to see $(X'X\beta)^{\mathcal{I}_0} \approx (\Omega\beta)^{\mathcal{I}_0}$ and $(X'z)^{\mathcal{I}_0} \approx N(0, \Omega^{\mathcal{I}_0,\mathcal{I}_0})$. Therefore,

$$(X'Y)^{\mathcal{I}_0} \approx N((\Omega\beta)^{\mathcal{I}_0}, \Omega^{\mathcal{I}_0,\mathcal{I}_0}).$$

A key observation is

$$(\Omega\beta)^{\mathcal{I}_0} \approx \Omega^{\mathcal{I}_0,\mathcal{I}_0}\beta^{\mathcal{I}_0}. \tag{2.13}$$

In fact, letting $\mathcal{I}_0^c = \{j : 1 \le j \le p, j \notin \mathcal{I}_0\}$, it is seen that

$$(\Omega\beta)^{\mathcal{I}_0} - \Omega^{\mathcal{I}_0,\mathcal{I}_0}\beta^{\mathcal{I}_0} = (\Omega^*)^{\mathcal{I}_0,\mathcal{I}_0^c}\beta^{\mathcal{I}_0^c} + (\Omega - \Omega^*)^{\mathcal{I}_0,\mathcal{I}_0^c}\beta^{\mathcal{I}_0^c} = I + II. \tag{2.14}$$

First, by Lemma 2.3, $|II| \le C\|\Omega - \Omega^*\|_\infty \|\beta\|_\infty = o\big(\sqrt{\log(p)}\big)$ coordinate-wise, hence II is negligible. Second, by the Sure Screening property, signals that are falsely screened out in the U-step are smaller than $L_p p^{1-(\vartheta+r)^2/(4r)}$ in number, and therefore have a negligible effect. To bring out the intuition, we assume $\mathcal{U}_p(t_p^*)$ contains all signals for a moment (see Lemma 6.4 for a formal treatment). This, together with (2.12), implies that $I = 0$, and (2.13) follows.

As a result, the original regression problem reduces to many regression problems of the form

$$(X'Y)^{\mathcal{I}_0} \approx N(\Omega^{\mathcal{I}_0,\mathcal{I}_0}\beta^{\mathcal{I}_0},\ \Omega^{\mathcal{I}_0,\mathcal{I}_0}), \tag{2.15}$$

which are small in size and can be fitted separately. Note that $\Omega^{\mathcal{I}_0,\mathcal{I}_0}$ can be accurately estimated by $(X'X)^{\mathcal{I}_0,\mathcal{I}_0}$, due to the small size of $\mathcal{I}_0$. We are now ready for the P-step.

2.4 P-step

The goal of the P-step is that, for each fixed connected subgraph $\mathcal{I}_0 \lhd \mathcal{U}_p(t_p^*)$, we fit Model (2.15) with an error rate $\le L_p p^{-(\vartheta+r)^2/(4r)}$. This turns out to be rather delicate, and many methods (including the lasso and the subset selection) do not achieve the desired rate of convergence.

For this reason, we proposed a Penalized-MLE approach. The idea can be explained as follows. Given that $\mathcal{I}_0 \lhd \mathcal{U}_p(t_p^*)$, as a priori, the chance that $\mathcal{I}_0$ contains k signals is $\sim \epsilon_p^k$, $k \ge 0$. This motivates us to fit Model (2.15) by maximizing the likelihood function

$$\epsilon_p^k \cdot \exp\Big[-\frac{1}{2}[(X'Y)^{\mathcal{I}_0} - (X'X)^{\mathcal{I}_0}\mu]' \left((X'X)^{\mathcal{I}_0,\mathcal{I}_0}\right)^{-1}[(X'Y)^{\mathcal{I}_0} - (X'X)^{\mathcal{I}_0}\mu]\Big], \quad \text{subject to } \|\mu\|_0 = k.$$

Recalling $(X'X)^{\mathcal{I}_0} \approx \Omega^{\mathcal{I}_0,\mathcal{I}_0}$, this is proportional to the density of $(X'Y)^{\mathcal{I}_0}$ in (2.15), hence the name of Penalized MLE. Recalling that $\epsilon_p = p^{-\vartheta}$, it is equivalent to minimizing

$$\frac{1}{2}\big[(X'Y)^{\mathcal{I}_0} - (X'X)^{\mathcal{I}_0}\mu\big]' \left((X'X)^{\mathcal{I}_0,\mathcal{I}_0}\right)^{-1} \big[(X'Y)^{\mathcal{I}_0} - (X'X)^{\mathcal{I}_0}\mu\big] + \frac{(\lambda_p^{ups})^2}{2} \cdot \|\mu\|_0, \quad (2.16)$$

where $\lambda_p^{ups} = \sqrt{2\vartheta \log p}$.

Unfortunately, (2.16) does not achieve the desired rate of convergence as expected. The reason is that we have not taken full advantage of the information provided: given that all coordinates in $\mathcal{I}_0$ survive the screening, each signal in $\mathcal{I}_0$ should be relatively strong. Motivated by this, for some tuning parameter $u^{ups} > 0$, we force all nonzero coordinates of μ to equal u^{ups}. This is the UPS procedure we introduced in Section 1. The UPS yields the desired rate of convergence provided that u^{ups} is properly set. See Theorem 2.1 for details.

One may think that forcing all nonzero coordinates of μ to be equal is too restrictive, since the nonzero coordinates of $\beta^{\mathcal{I}_0}$ are unequal. Nevertheless, the UPS achieves the desired error rate. The reason is that, knowing the exact values of the nonzero coordinates is not crucial, noticing that the main goal is to identify all nonzero coordinates of $\beta^{\mathcal{I}_0}$.

Similarly, since knowing the signal distribution π_p may be very helpful, one may choose to estimate π_p using the data first, then combine the estimated distribution with the P-step. However, this has two drawbacks. First, Model (2.15) is very small in size, and can be easily over fit if we introduce too many degrees of freedom. Second, estimating π_p usually involves deconvolution, which generally has relatively slow rate of convergence; see [33] for example. Note that a noisy estimate of π_p may hurt rather than help in fitting Model (2.15).

2.5 Upper bound

We are now ready for the upper bound. To recap, the proposed procedure is as follows.

- With fixed tuning parameters $(t, \lambda^{ups}, u^{ups})$, obtain $\mathcal{U}_p(t) = \{j : 1 \leq j \leq p,\ (x_j, Y) \geq t\}$.
- Obtain Ω^* as in (2.11), and form a graph (V_0, Ω_0) with $V_0 = \{1, 2, \ldots, p\}$, and $\Omega_0 = \Omega^*$.
- Split $\mathcal{U}_p(t)$ into connected subgraphs where different ones are disconnected. For each connected subgraph $\mathcal{I}_0 = \{i_1, i_2, \ldots, i_K\}$, obtain the minimizer of (2.16), where each coordinate of μ is either 0 or u^{ups}. Denote the estimate by $\hat{\mu}(\mathcal{I}_0) = \hat{\mu}(\mathcal{I}_0; Y, X, t, \lambda^{ups}, u^{ups}, p)$.
- For any $1 \leq j \leq p$, if $j \notin \mathcal{U}_p(t)$, set $\hat{\beta}_j = 0$. Otherwise, there is a unique $\mathcal{I}_0 = \{i_1, i_2, \ldots, i_K\} \lhd \mathcal{U}_p(t)$, where $i_1 < i_2 < \ldots < i_K$, such that j is the k-th coordinate of $\mathcal{I}_0$. Set $\hat{\beta}_j = (\hat{\mu}(\mathcal{I}_0))_k$.

Denote the resulting estimator by $\hat{\beta}(Y, X; t, \lambda^{ups}, u^{ups})$. The following theorem establishes the upper bound for the Hamming distance and the rate optimality of the UPS.

Theorem 2.1 *Consider Model (2.1)-(2.2) where (2.3)-(2.7) hold, and fix* $0 < q \leq (\vartheta + r)^2/(4r)$. *For sufficiently large* p, *if* $\Omega^{(p)} \in \mathcal{M}_p^+(\omega_0, \gamma, A)$, *and we set the tuning parameters of the UPS at*

$$t = t_p^* = \sqrt{2q\log(p)}, \qquad \lambda^{ups} = \lambda_p^{ups} = \sqrt{2\vartheta \log p}, \qquad u^{ups} = u_p^{ups} = \tau_p,$$

then as $p \to \infty$,

$$\frac{\text{Hamm}_p(\hat{\beta}^{ups}(Y, X; t_p^*, \lambda_p^{ups}, u_p^{ups}), \vartheta, r, \Omega^{(p)})}{s_p} \leq L_p \cdot p^{-\frac{(r-\vartheta)^2}{4r}}.$$

The claim remains valid if $r/\vartheta \leq 3 + 2\sqrt{2}$ *and* $\Omega^{(p)} \in \mathcal{M}_p^*(\omega_0, \gamma, A)$ *for sufficiently large* p.

Except for the L_p term, the upper bound matches the lower bound in Theorem 1.1. Therefore, both bounds are tight and the UPS is rate optimal.

2.6 Tuning parameters of the UPS

The UPS uses three tuning parameters, one in the U-step and two in the P-step. In this section, we show that under certain conditions, the P-step tuning parameters can be conveniently estimated from the data. As a result, the UPS only needs one tuning parameter.

In detail, denote $\tilde{Y} = X'Y$. For any $t > 0$, introduce

$$\bar{F}_p(t) = \frac{1}{p}\sum_{j=1}^p 1_{\{\tilde{Y}_j > t\}}, \qquad \mu_p(t) = \frac{1}{p}\sum_{j=1}^p \tilde{Y}_j \cdot 1_{\{\tilde{Y}_j > t\}}.$$

Denote the largest off-diagonal coordinate of Ω by

$$\delta_0 = \delta_0(\Omega) = \max_{\{1 \leq i,j \leq p, i \neq j\}} |\Omega(i,j)|.$$

Recalling that the support of π_p is contained in $[\tau_p, (1+\eta)\tau_p]$, we suppose

$$2\delta_0(1+\eta) - 1 \leq \vartheta/r, \qquad \text{so that} \qquad \delta_0^2(1+\eta)^2 r < \frac{(\vartheta + r)^2}{4r}. \tag{2.17}$$

Denote the mean of π_p by $\mu_p^* = \mu_p^*(\pi_p)$. The following lemma is proved in Section 6.

Lemma 2.5 *Fix* q *such that* $\max\{\delta_0^2(1+\eta)^2 r, \vartheta\} < q \leq \frac{(\vartheta+r)^2}{4r}$, *and let* $t_p^* = \sqrt{2q\log p}$. *Suppose the conditions in Theorem 2.1 hold. As* $p \to \infty$, *with probability of* $1 - o(1/p)$,

$$\left|\frac{\bar{F}_p(t_p^*)}{\epsilon_p} - 1\right| \leq o(1), \qquad \textit{and} \qquad \left|\frac{\mu_p(t_p^*)}{\epsilon_p \mu_p^*(\pi_p)} - 1\right| \leq o(1). \tag{2.18}$$

Motivated by Lemma 2.18, we propose to estimate (λ^{ups}, u^{ups}) by

$$\hat{\lambda}_p^{ups} = \hat{\lambda}_p^{ups}(q) = \sqrt{-2\log(\bar{F}_p(t_p^*))}, \qquad \hat{u}_p^{ups} = \hat{u}_p^{ups}(q) = \frac{\mu_p(t_p^*)}{\bar{F}_p(t_p^*)}. \tag{2.19}$$

The following theorem is proved in Section 6.

Theorem 2.2 *Fix* q *such that* $\max\{\delta_0^2(1+\eta)^2 r, \vartheta\} < q \leq \frac{(\vartheta+r)^2}{4r}$, *and let* $t_p^* = \sqrt{2q\log p}$. *Suppose the conditions of Theorem 2.1 hold. As* $p \to \infty$, *if additionally* $\mu_p^*(\pi_p) \leq (1 + o(1))\tau_p$, *then*

$$\frac{\text{Hamm}_p(\hat{\beta}^{ups}(Y, X; t_p^*, \hat{\lambda}_p^{ups}, \hat{u}_p^{ups}), \vartheta, r, \Omega^{(p)})}{s_p} \leq L_p \cdot p^{-\frac{(r-\vartheta)^2}{4r}}.$$

Therefore, t_p^* is the only tuning parameter needed by the UPS. Theorem 2.2 states that, the performance of the UPS is relatively insensitive to the choice of t_p^*, as long as it falls in a certain range. Numerical studies in Section 5 confirm such insensitivity for finite p. The numerical study also reveals that the lasso is comparably more sensitive to its tuning parameter λ^{lasso}. Note that how to set λ^{lasso} remains an unsolved problem in the literature.

The success of the UPS relies on its Sure Screening property and its SAS property. These properties hold in much broader settings than those presented here. In a forthcoming manuscript [17], we explore a strongly dependent case where each row of Ω decays slowly. In that setting, we find that (up to some modifications) these properties continue to hold, and the UPS continues to work well.

3 A refinement for moderately large p

In this section, we introduce a refinement for the UPS when p is moderately large. We begin by investigating the relationship between the regression model and Stein's normal means model.

Recall that the regression model (1.1) is closely related to the following model:

$$X'Y = X'X\beta + X'z, \qquad z \sim N(0, I_n), \tag{3.1}$$

which in turn is approximately equivalent to Stein's normal means model as follow:

$$X'Y \approx \Omega\beta + N(0,\Omega) \qquad \Longleftrightarrow \qquad \Omega^{-1}X'Y \approx N(\beta, \Omega^{-1}). \tag{3.2}$$

In the literature, Stein's normal means model has been extensively studied, but the focus has been on the case where Ω is diagonal (e.g. [33]). When Ω is not diagonal, Stein's normal means model is intrinsically a regression problem. To see how close Models (3.1) and (3.2) are, write

$$X'Y = \left(\Omega\beta + \frac{\sqrt{n}}{\|z\|}X'z\right) + \left((X'X-\Omega)\beta + (\frac{\|z\|}{\sqrt{n}} - 1)\frac{\sqrt{n}}{\|z\|}X'z\right) = I + II. \tag{3.3}$$

Note that $I \sim N(\Omega\beta, \Omega)$. For II, we have the following lemma, which is proved in Section 6.

Lemma 3.1 *Consider Model (2.1)-(2.2) where (2.2)-(2.4) hold. As $p \to \infty$, there is a constant $C > 0$ such that except for a probability of $o(1/p)$,*

$$\left|\frac{\|z\|}{\sqrt{n}} - 1\right| \leq C(\sqrt{\log p})p^{-\theta/2}, \qquad \|(X'X-\Omega)\beta\|_\infty \leq C\|\Omega\|(\sqrt{2\log p})p^{-(\theta-(1-\vartheta))/2}.$$

It follows that $|II| \leq C\sqrt{2\log(p)}\cdot p^{-[\theta-(1-\vartheta)]/2}$ coordinate-wise. Therefore, asymptotically, Models (3.1) and (3.2) have negligible difference. When p is moderately large, the difference between Models (3.1) and (3.2) may be non-negligible. In Table 1, we tabulate the values of $\sqrt{2\log(p)}\cdot p^{-[\theta-(1-\vartheta)]/2}$, which are relatively large for moderately large p (e.g. $p = 2000$).

Moreover, for moderately large p, we observe that the random design model is much noisier than Stein's normal means model. In the U-step, we tend to falsely keep more noise terms in the former than in the latter; some of these noise terms are large in magnitude,

p	400	5×400	$5^2 \times 400$	$5^3 \times 400$	$5^4 \times 400$	$5^5 \times 400$
$(\theta, \vartheta) = (0.91, 0.65)$	0.65	0.46	0.33	0.22	0.15	0.10
$(\theta, \vartheta) = (0.91, 0.5)$	1.01	0.82	0.65	0.51	0.39	0.30

Table 1: The values of $\sqrt{2\log(p)}p^{-[\theta-(1-\vartheta)]/2}$ for different p and (θ, ϑ).

and it is hard to clean all of them in the P-step. To see how the problem can be fixed, we write

$$X'X\beta = (X'X - \Omega^*)\beta + \Omega^*\beta. \tag{3.4}$$

On one hand, the term $(X'X - \Omega^*)\beta$ causes the random design model to be much noisier than Stein's normal means model. On the other hand, this term can be easily removed from the model if we have a reasonably good estimate of β. This motivates a refinement as follows.

For any p by 1 vector y, let $S^2(y) = \frac{1}{p-1}\sum_{j=1}^p (y_j - \bar{y})^2$ where $\bar{y} = \frac{1}{p}\sum_{j=1}^p y_j$. We propose the following procedure:

- Run the UPS and obtain an estimate of β, say, $\hat{\beta}$. Let $W^{(0)} = X'Y$ and $\hat{\beta}^{(0)} = \hat{\beta}$.
- For $j = 1, 2, 3$, respectively, let $W^{(j)} = X'Y - (X'X - \Omega^*)\hat{\beta}^{(j-1)}$. If $S(W^{(j)})/S(W^{(j-1)}) \leq 1.05$, run the UPS with $X'Y$ replaced by $W^{(j)}$ and other parts unchanged, and let $\hat{\beta}^{(j)}$ be the new estimate. Stop otherwise.

Numerical studies in Section 5 suggest that the refinement is beneficial for moderately large p. When p is sufficiently large (e.g. $\sqrt{2\log(p)} \cdot p^{-[\theta-(1-\vartheta)]/2} \leq 0.4$), the original UPS is usually good enough. In this case, refinements are not necessary, but may still offer improvements.

4 Understanding the lasso and subset selection

In this section, we continue our discussion in Sections 1.7-1.8 on the performance of the lasso and the subset selection. We show that there is a region in the phase space where the lasso is rate non-optimal (and similarly for the subset selection). For technical convenience, we consider Stein's normal means model instead of the random design model. As mentioned in Section 1.7, since the goal is to understand the non-optimality of the lasso and the subset selection, focusing on a simpler model enjoys mathematical convenience, yet is also sufficient.

To recap, the model we consider in this section is

$$\tilde{Y} \sim N(\Omega\beta, \Omega), \tag{4.1}$$

where $\tilde{Y}$ is the counterpart of $X'Y$ in the random design model. Fix $a \in (-1/2, 1/2)$. As in Section 1.7, we let Ω be the tridiagonal matrix as in (1.16), and π_p be the point mass at $\tau_p = \sqrt{2r\log p}$. In other words,

$$\beta_j \overset{iid}{\sim} (1-\epsilon_p)\nu_0 + \epsilon_p \nu_{\tau_p}, \qquad \epsilon_p = p^{-\vartheta}, \qquad \tau_p = \sqrt{2r\log p}. \tag{4.2}$$

Throughout this section, we assume $r > \vartheta$ so that successful variable selection is possible. Somewhat surprisingly, even in this simple case and even when (ϵ_p, τ_p) are known, there is a region in the phase space where the lasso (or the subset selection) is non-optimal. We now take a heuristic approach in hope of shedding light on this. Formal statements are given in later sections.

4.1 Understanding the lasso

The vector $\tilde{Y}$ consists of three main components: true signals, fake signals, and pure noise (see Definition 1.3). According to (4.2), true signals may appear as singletons, pairs, triplets, ..., but singletons are the most common and therefore have the major effect. For each signal singleton, with no other signal nearby, since Ω is tridiagonal, we have two fake signals, one to the left and one to the right. Given a site j, $1 \leq j \leq p$, the lasso may make three types of errors:

- *Type I.* $\tilde{Y}_j$ is a pure noise, but the lasso mistakes it as a signal.
- *Type II.* $\tilde{Y}_j$ is a signal singleton, but the lasso mistakes it as a noise.
- *Type III.* $\tilde{Y}_j$ is a fake signal next to a signal singleton, but the lasso mistakes it as a signal.

There are other types of errors, but these are the major ones.

To minimize the sum of these errors, the lasso needs to choose the tuning parameter λ^{lasso} carefully. To shed light, we first consider the uncorrelated case where Ω is the identity matrix. In this case, we do not have fake signals and it is understood that the lasso is equivalent to the soft-thresholding procedure [33], where the sum of Type I and Type II errors is

$$p\big[(1-\epsilon_p)\bar{\Phi}(\lambda^{lasso}) + \epsilon_p \Phi(\lambda^{lasso} - \tau_p)\big]. \tag{4.3}$$

Recall that $\bar{\Phi} = 1 - \Phi$ is the survival function of $N(0,1)$. Fix $0 < q < 1$, and take

$$\lambda^{lasso} = \lambda_p^{lasso} = \sqrt{2q\log(p)}.$$

It follows from Mills' ratio [33] that

$$p\big[(1-\epsilon_p)\bar{\Phi}(\lambda_p^{lasso}) + \epsilon_p\Phi(\lambda_p^{lasso} - \tau_p)\big] \sim \begin{cases} L_p\big[p^{1-q} + p^{1-(\vartheta+(\sqrt{q}-\sqrt{r})^2)}\big], & \text{if } 0 < q < r, \\ p^{1-q} + p^{1-\vartheta}, & \text{if } q > r. \end{cases}$$

The right-hand side is minimized at $q = (\vartheta + r)^2/(4r)$ at which $\lambda_p^{lasso} = \frac{\vartheta+r}{2r}\tau_p$, and the sum of errors is $L_p p^{-(\vartheta+r)^2/(4r)}$, which is the optimal rate of convergence. For a smaller q, the lasso keeps too many noise terms. For a larger q, the lasso kills too many signals.

Come back to the correlated case. The vector $\tilde{Y}$ is at least as noisy as that in the uncorrelated case. As a result, to control the Type I errors, we should choose λ_p^{lasso} to be at least as large as $\frac{\vartheta+r}{2r}\tau_p$. This is confirmed in Lemma 4.2 below.

In light of this, we fix $q \geq (\vartheta + r)^2/(4r)$ and let $\lambda_p^{lasso} = \sqrt{2q\log(p)}$ from now on. We observe that except for a negligible probability, the support of $\hat{\beta}^{lasso}$, denoted by $\hat{S}_p^{lasso}$, splits into many small clusters (i.e. block of adjacent indices). There is an integer K not dependent on p that has the following effects.

- If $\tilde{Y}_j$ is a pure noise, and there is no signal within a distance of K from it (i.e. $\beta_k = 0$ for all k satisfying $1 \leq |k-j| \leq K$), then either $\hat{\beta}_j^{lasso} = 0$, or $\hat{\beta}_j^{lasso} \neq 0$ but $\hat{\beta}_{j\pm1}^{lasso} = 0$.
- If $\tilde{Y}_j$ is a signal singleton, and there is no other signal within a distance of K from it (i.e. $\beta_k = 0$ for all k satisfying $1 \leq |k - j| \leq K$), then either $\hat{\beta}_j^{lasso} = 0$, or $\hat{\beta}_j^{lasso} \neq 0$ but $\hat{\beta}_{j\pm2} = 0$ and at least one of $\{\hat{\beta}_{j+1}^{lasso}, \hat{\beta}_{j-1}^{lasso}\}$ is 0.

This is justified in a forthcoming thesis [26]. The proofs are relatively long so we omit them. We use these results to provide insight, but do not use them to prove the results below (though technically they are connected).

At the same time, let $\mathcal{I}_0 = \{j-k+1,\ldots,j\} \subset \hat{S}_p^{lasso}$ be a cluster, so that $\hat{\beta}_{j-k}^{lasso} = \hat{\beta}_{j+1}^{lasso} = 0$. Since Ω is tridiagonal, the restriction of $\hat{\beta}^{lasso}$ to $\mathcal{I}_0$, denoted by $(\hat{\beta}^{lasso})^{\mathcal{I}_0}$, is equivalent to solving a small-scale minimization problem:

$$\frac{1}{2}\mu'(\Omega^{\mathcal{I}_0,\mathcal{I}_0})\mu - \mu'\tilde{Y}^{\mathcal{I}_0} + \lambda^{lasso}\|\mu\|_1, \qquad \text{where } \mu \text{ is a } k \text{ by 1 vector.} \tag{4.4}$$

See Definition of 1.1 for notations. Two special cases are noteworthy. In the first case, $\mathcal{I}_0 = \{j\}$, and the solution of (4.4) is given by

$$\hat{\beta}_j^{lasso} = \mathrm{sgn}(\tilde{Y}_j)(|\tilde{Y}_j| - \lambda^{lasso})^+,$$

which is the soft-thresholding [33]. In the second case, $\mathcal{I}_0 = \{j-1, j\}$. We call the solution of (4.4) in this case the *bivariate lasso*. We have the following lemma, where Regions I–$IIId$ are illustrated in Figure 4.

Lemma 4.1 *Denote $\lambda = \lambda^{lasso}$. The solution of the bivariate lasso is given by*

$$(\hat{\beta}_{j-1}^{lasso}, \hat{\beta}_j^{lasso}) = \begin{cases} (\mathrm{sgn}(\tilde{Y}_{j-1})(|\tilde{Y}_{j-1}| - \lambda)^+,\ \mathrm{sgn}(\tilde{Y}_j)(|\tilde{Y}_j| - \lambda)^+), & \textit{Regions I, IIa-IId}, \\ \frac{1}{1-a^2}((\tilde{Y}_{j-1} - \lambda) - a(\tilde{Y}_j - \lambda),\ (\tilde{Y}_j - \lambda) - a(\tilde{Y}_{j-1} - \lambda)), & \textit{Region IIIa}, \\ \frac{1}{1-a^2}((\tilde{Y}_{j-1} + \lambda) - a(\tilde{Y}_j - \lambda),\ (\tilde{Y}_j - \lambda) - a(\tilde{Y}_{j-1} + \lambda)), & \textit{Region IIIb}, \\ \frac{1}{1-a^2}((\tilde{Y}_{j-1} + \lambda) - a(\tilde{Y}_j + \lambda),\ (\tilde{Y}_j + \lambda) - a(\tilde{Y}_{j-1} + \lambda)), & \textit{Region IIIc}, \\ \frac{1}{1-a^2}((\tilde{Y}_{j-1} - \lambda) - a(\tilde{Y}_j + \lambda),\ (\tilde{Y}_j + \lambda) - aP(\tilde{Y}_{j-1} - \lambda)), & \textit{Region IIId}. \end{cases}$$

In the white region of Figure 4, both $\hat{\beta}_{j-1}^{lasso}$ and $\hat{\beta}_j^{lasso}$ are 0. In the blue regions, exactly one of them is 0. In the yellow regions, both are nonzero. Lemma 4.1 is proved in Section 6.

Combining these observations, the following hold except for a negligible probability.

- *Type I.* There are $O(p)$ indices j where $\tilde{Y}_j$ is a pure noise, and no signal appears within a distance of K from it. For each of such j, the lasso acts on $\tilde{Y}_j$ as soft-thresholding, and $\hat{\beta}_j^{lasso} \neq 0$ if and only if $|\tilde{Y}_j| \geq \lambda_p^{lasso}$.

- *Types II-III.* There are $O(p\epsilon_p)$ indices where $\tilde{Y}_j$ is a signal singleton, and no other signal appears within a distance of K from it. The lasso either acts on $\tilde{Y}_j$ as soft-thresholding, or acts on both $\tilde{Y}_j$ and one of its neighbors as the bivariate lasso. As a result, $\hat{\beta}_j^{lasso} = 0$ if and only if $|\tilde{Y}_j| \leq \lambda_p^{lasso}$ (Type II), and both $\hat{\beta}_j^{lasso}$ and $\hat{\beta}_{j-1}^{lasso}$ are nonzero if and only if $(\tilde{Y}_{j-1}, \tilde{Y}_j)'$ falls in Regions $IIIa$–$IIId$, with $IIIa$ and $IIIb$ being the most likely (Type III).

Note that $\tilde{Y}_j \sim N(0,1)$ when it is a pure noise and $\tilde{Y}_j \sim N(\tau_p, 1)$ if it is a signal singleton. It follows that the sum of Type I and Type II errors is

$$L_p p\big[P(N(0,1) \geq \lambda_p^{lasso}) + \epsilon_p P(N(\tau_p,1) < \lambda_p^{lasso})\big] = L_p p\big[\bar{\Phi}(\lambda_p^{lasso}) + \epsilon_p \Phi(\lambda_p^{lasso} - \tau_p)\big]. \tag{4.5}$$

Note that when $\tilde{Y}_j$ is a signal singleton, $(\tilde{Y}_{j-1}, \tilde{Y}_j)'$ is distributed as a bivariate normal with means $a\tau_p$ and τ_p, variances 1, and correlation a. Denote such a bivariate normal distribution by W for short. The Type III error is

$$L_p p \cdot P\big(\beta_{j-1} = 0, \beta_j = \tau_p, (\tilde{Y}_{j-1}, \tilde{Y}_j)' \in \text{Regions } IIIa \text{ or } IIIb\big) \sim L_p p \epsilon_p \cdot P\big(W \in \text{Regions } IIIa \text{ or } IIIb\big).$$

It follows that the sum of three types of errors is

$$L_p p \cdot \big[\bar{\Phi}(\lambda_p^{lasso}) + \epsilon_p \Phi(\lambda_p^{lasso} - \tau_p) + \epsilon_p P\big(W \in \text{Regions } IIIa \text{ or } IIIb\big)\big], \tag{4.6}$$

which can be evaluated directly by elementary calculus and Mills's ratio [33]. Compare (4.6) with (4.3). It is interesting to note that the sum of Type I and Type II errors is the same in the uncorrelated case and in the correlated case. Recall that the sum of Type I and Type II errors is minimized at $\lambda_p^{lasso} = (\vartheta + r)/(2r)\tau_p$. Therefore, whether the lasso is optimal or not depends on whether the Type III error is smaller than the optimal rate of convergence or not. It turns out that in certain regions in the phase space, the Type III error can be significantly larger than the optimal rate, and so the lasso is rate non-optimal. In other words, provided that the tuning parameters are properly set, the lasso is able to separate the signal singletons from the pure noise. However, it may not be efficient in filtering out the fake signals. This is why the lasso may be rate non-optimal.

The following lemma is proved in Section 6, which confirms that the heuristics above are valid.

Lemma 4.2 *Fix $\vartheta \in (0,1)$, $r > \vartheta$, $q > 0$ and $a \in (-1/2, 1/2)$. Set the lasso tuning parameter as $\lambda_p^{lasso} = \sqrt{2q \log p}$. As $p \to \infty$,*

$$\frac{\text{Hamm}(\hat{\beta}^{lasso}(\lambda_p^{lasso}); \epsilon_p, \tau_p, a)}{s_p} \geq \begin{cases} L_p p^{-\min\{\frac{1-|a|}{1+|a|}q,\ q-\vartheta\}}, & \text{if } 0 < q < \frac{(\vartheta+r)^2}{4r}, \\ L_p p^{-\min\{\frac{1-|a|}{1+|a|}q,\ (\sqrt{r}-\sqrt{q})^2\}}, & \text{if } \frac{(\vartheta+r)^2}{4r} < q < r, \\ (1+o(1)), & \text{if } q > r. \end{cases} \tag{4.7}$$

The exponent on the right-hand side of (4.7) is minimized at

$$q = \begin{cases} \frac{(\vartheta+r)^2}{4r}, & \text{if } r < \frac{1+\sqrt{1-a^2}}{|a|}\vartheta, \\ \frac{(1+|a|)(1-\sqrt{1-a^2})}{2a^2} r, & \text{if } r > \frac{1+\sqrt{1-a^2}}{|a|}\vartheta, \end{cases} \tag{4.8}$$

where we note that $r < \frac{1+\sqrt{1-a^2}}{|a|}\vartheta$ and $r > \frac{1+\sqrt{1-a^2}}{|a|}\vartheta$ correspond to the optimal and non-optimal regions of the lasso, respectively. This shows that in the optimal region of the lasso, $\lambda_p^{lasso} = (\vartheta + r)/(2r)\tau_p$ remains the optimal tuning parameter, at which the sum of Type I and Type II errors is minimized, and the Type III error has a negligible effect. In the non-optimal region of the lasso, at $\lambda_p^{lasso} = (\vartheta + r)/(2r)\tau_p$, the Type III error is larger than the sum of Type I and Type II errors, so the lasso needs to raise the tuning parameter slightly to minimize the sum of all three types of errors (but the resultant Hamming error is still larger than that of the optimal procedure). The following theorem is a direct result of Lemma 4.2 and (4.8); we omit the proof.

Theorem 4.1 *Set the tuning parameter $\lambda_p^{lasso} = \sqrt{2q \log p}$. For all choices of $q > 0$, the Hamming distance of the lasso satisfies*

$$\frac{\text{Hamm}_p(\hat{\beta}^{lasso}(\lambda_p^{lasso}); \epsilon_p, \tau_p, a)}{s_p} \geq \begin{cases} L_p p^{-\frac{(\vartheta-r)^2}{4r}}, & \text{if } r/\vartheta < (1+\sqrt{1-a^2})/|a|, \\ L_p p^{\vartheta - \frac{(1-|a|)(1-\sqrt{1-a^2})}{2a^2} r}, & \text{if } r/\vartheta > (1+\sqrt{1-a^2})/|a|. \end{cases}$$

In [26], we show that when $r/\vartheta \leq 3 + 2\sqrt{2}$, the lower bound in Theorem 4.1 is tight. Together, these give the phase diagram in Figure 2. The proof of the upper bound involves considerable effort in characterizing the lasso solution and is relatively long. Since the main interest in this section is the non-optimal region of the lasso, we leave further discussion to [26].

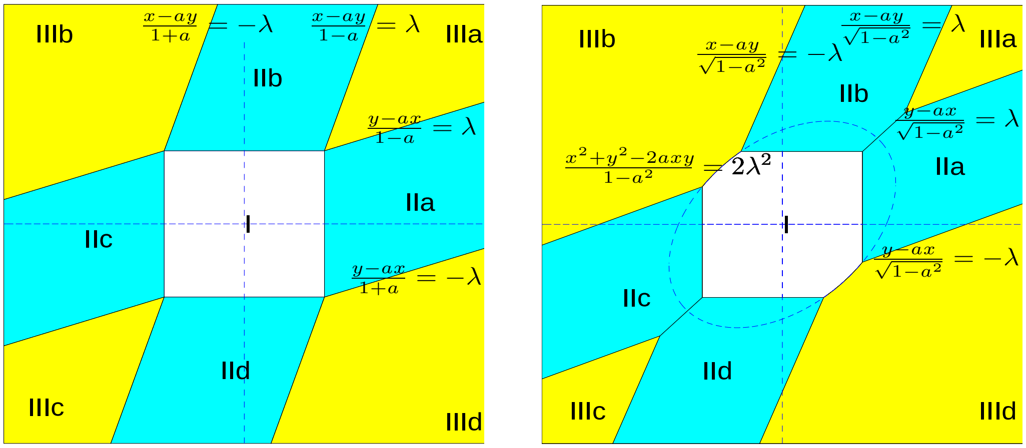

Figure 4: Partition of regions as in Lemma 4.1 (left) and in Lemma 4.3 (right).

4.2 Understanding the subset selection

The discussion for the subset selection is mostly similar, so we keep it brief. Fix a site j, $1 \leq j \leq p$. The subset selection has three major types of errors.

- *Type I.* $\tilde{Y}_j$ is a pure noise, but the subset selection mistakes it as a signal.
- *Type II.* $\tilde{Y}_j$ is a signal singleton, but the subset selection mistakes it as a noise.
- *Type III.* $(\tilde{Y}_{j-1}, \tilde{Y}_j)$ is a signal pair, but the subset selection mistakes one of them as a noise.

Compared to the lasso, the subset selection is more efficient in filtering out fake signals. However, it tends to kill one or more signals when signals appear in pairs, triplets, etc.. Type III error here is different from that in the preceding section.

Suppose that $\tilde{Y}_j$ is either a pure noise or a signal singleton, and for an appropriately large K, no other signal appears within a distance of K from it. In this case, except for a negligible probability, $\hat{\beta}_{j\pm1}^{lasso} = 0$, and the subset selection acts on site j as hard thresholding [33]:

$$\hat{\beta}_j^{ss} = \tilde{Y}_j \cdot 1_{\{|\tilde{Y}_j| \geq \lambda^{ss}\}}.$$

Note that $\tilde{Y}_j \sim N(0,1)$ if $\tilde{Y}_j$ is a pure noise, and $\tilde{Y}_j \sim N(\tau_p, 1)$ if it is a signal singleton. Take $\lambda^{ss} = \lambda_p^{ss} = \sqrt{2q \log p}$ as before. Similarly, the sum of Type I and Type II errors is

$$L_p p \cdot [\bar{\Phi}(\lambda_p^{ss}) + p^{-\vartheta} \Phi(\lambda_p^{ss} - \tau_p)] = \begin{cases} L_p \cdot (p^{1-q} + p^{1-\vartheta-(\sqrt{q}-\sqrt{r})^2}), & \text{if } 0 < q < r, \\ L_p \cdot (p^{1-q} + p^{1-\vartheta}), & \text{if } q > r. \end{cases} \tag{4.9}$$

On the right-hand side, the exponent is minimized at $q = (\vartheta + r)^2/4r$, at which the rate is $L_p p^{1-(\vartheta+r)^2/(4r)}$, which is the optimal rate of convergence.

Next, consider the Type III error. Suppose $(\tilde{Y}_{j-1}, \tilde{Y}_j)$ is a signal pair and no other signal appears within a distance of K for a properly large K. Similarly, since Ω is tridiagonal, $(\hat{\beta}_{j-1}^{ss}, \hat{\beta}_j^{ss})'$ is the minimizer of the following functional:

$$\frac{1}{2}\beta_{j-1}^2 + \frac{1}{2}\beta_j^2 + a\beta_{j-1}\beta_j - (\tilde{Y}_{j-1}\beta_{j-1} + \tilde{Y}_j\beta_j) + \frac{(\lambda_p^{ss})^2}{2}\left(I\{\beta_{j-1} \neq 0\} + I\{\beta_j \neq 0\}\right).$$

We call the resultant procedure *bivariate subset selection*. The following lemma is proved in Section 6, where the regions are illustrated in Figure 4.

Lemma 4.3 *The solution of the bivariate subset selection is given by*

$$(\hat{\beta}^{ss}_{j-1}, \hat{\beta}^{ss}_j) = \begin{cases} (0,0), & \textit{if } (\tilde{Y}_{j-1}, \tilde{Y}_j) \textit{ in Region I}, \\ (\tilde{Y}_{j-1}, 0), & \textit{if } (\tilde{Y}_{j-1}, \tilde{Y}_j) \textit{ in Regions IIa, IIc}, \\ (0, \tilde{Y}_j), & \textit{if } (\tilde{Y}_{j-1}, \tilde{Y}_j) \textit{ in Regions IIb, IId}, \\ (\frac{\tilde{Y}_{j-1} - a\tilde{Y}_j}{1-a^2}, \frac{\tilde{Y}_j - a\tilde{Y}_{j-1}}{1-a^2}), & \textit{if } (\tilde{Y}_{j-1}, \tilde{y}_j) \textit{ in Regions IIIa to IIId}. \end{cases}$$

When $(\tilde{Y}_{j-1}, \tilde{Y}_j)$ falls in Regions I, IIa or IIb, either $\hat{\beta}^{ss}_{j-1}$ or $\hat{\beta}^{ss}_j$ is 0, and the subset selection makes a Type III error. Note that there are $O(p\epsilon_p^2)$ signal pairs, and that $(\tilde{Y}_{j-1}, \tilde{Y}_j)'$ is jointly distributed as a bivariate normal with means $(1+a)\tau_p$, variances 1, and correlation a. The Type III error is then

$$L_p p \epsilon_p^2 P((\tilde{Y}_{j-1}, \tilde{Y}_j)' \in \text{Regions } I, \, IIa \text{ or } IIb) = L_p p^{1-(2\vartheta + \min\left\{[(\sqrt{r(1-a^2)} - \sqrt{q})^+]^2, 2[(\sqrt{r(1+a)} - \sqrt{q})^+]^2\right\}}. \tag{4.10}$$

Combining (4.9)-(4.10) and using Mills' ratio give the sum of all three types of errors. Formally speaking, we have the following lemma, which is proved in Section 6.

Lemma 4.4 *Set the tuning parameter $\lambda_p^{ss} = \sqrt{2q \log p}$. The Hamming error for the subset selection satisfies*

$$\frac{\text{Hamm}_p(\hat{\beta}^{ss}(\lambda_p^{ss}); \epsilon_p, \tau_p, a)}{s_p} \geq \begin{cases} L_p p^{-\min\{q-\vartheta, \, \vartheta + [(\sqrt{r(1-a^2)} - \sqrt{q})^+]^2\}}, & \textit{if } 0 < q < \frac{(\vartheta+r)^2}{4r}, \\ L_p p^{-\min\{(\sqrt{r} - \sqrt{q})^2, \, \vartheta + [(\sqrt{r(1-a^2)} - \sqrt{q})^+]^2\}}, & \textit{if } \frac{(\vartheta+r)^2}{4r} < q < r, \\ (1+o(1)), & \textit{if } q > r. \end{cases}$$

The exponents on the right-hand side are minimized at

$$q = \begin{cases} \frac{(\vartheta+r)^2}{4r}, & \text{if } \frac{r}{\vartheta} < \frac{2-\sqrt{1-a^2}}{\sqrt{1-a^2}(1-\sqrt{1-a^2})}, \\ \frac{[2\vartheta + r(1-a^2)]^2}{4r(1-a^2)}, & \text{if } \frac{r}{\vartheta} > \frac{2-\sqrt{1-a^2}}{\sqrt{1-a^2}(1-\sqrt{1-a^2})}. \end{cases} \tag{4.11}$$

As a result, we have the following theorem, the proof of which is omitted.

Theorem 4.2 *Set the tuning parameter $\lambda_p^{ss} = \sqrt{2q \log p}$. Then for all $q > 0$, the Hamming error of the subset selection satisfies*

$$\frac{\text{Hamm}_p(\hat{\beta}^{ss}(\lambda_p^{ss}); \epsilon_p, \tau_p, a)}{s_p} \geq \begin{cases} L_p p^{-(\vartheta - r)^2/(4r)}, & \textit{if } \frac{r}{\vartheta} < \frac{2-\sqrt{1-a^2}}{\sqrt{1-a^2}(1-\sqrt{1-a^2})}, \\ L_p p^{-\frac{[2\vartheta + r(1-a^2)]^2}{4r(1-a^2)} + \vartheta}, & \textit{if } \frac{r}{\vartheta} > \frac{2-\sqrt{1-a^2}}{\sqrt{1-a^2}(1-\sqrt{1-a^2})}. \end{cases}$$

This gives the phase diagram by the subset selection in Figure 2, where $r/\vartheta < \frac{2-\sqrt{1-a^2}}{\sqrt{1-a^2}(1-\sqrt{1-a^2})}$ and $r/\vartheta > \frac{2-\sqrt{1-a^2}}{\sqrt{1-a^2}(1-\sqrt{1-a^2})}$ correspond to the optimal and non-optimal regions of the subset selection, respectively. Similar to the lasso, the subset selection is able to separate signal singletons from the pure noise provided that the tuning parameter is properly set. But the subset selection is too harsh on signal pairs, triplets, etc., which costs its rate optimality. In [26], we further show that in certain regions of the phase space, the lower bound in Theorem 4.1 is tight.

5 Simulations

We have conducted a small-scale empirical study of the performance of the UPS. The idea is to select a few interesting combinations of $(\vartheta, \theta, \pi_p, \Omega)$ and study the behavior of the UPS for finite p. Fixing $(p, \pi_p, \Omega, \vartheta, \theta)$, let $n_p = p^\theta$ and $\epsilon_p = p^{-\vartheta}$. We investigate both the random design model and Stein's normal means model. In the former, the experiment contains the following steps.

1. Generate a p by 1 vector β by $\beta_j \stackrel{iid}{\sim} (1-\epsilon_p)\nu_0 + \epsilon_p \pi_p$, and an n_p by 1 vector $z \sim N(0, I_{n_p})$.
2. Generate an n_p by p matrix X the rows of which are samples from $N(0, \frac{1}{n_p}\Omega)$; let $Y = X\beta + z$.
3. Apply the UPS and the lasso. For the lasso, we use the *glmnet* package by Friedman *et al.* [20] (Ω is assumed unknown in both procedures).
4. Repeat 1–3 for 100 independent cycles, and calculate the average Hamming distances.

In the latter, the settings are similar, except for (i) $n_p = p$, (ii) $Y \sim N(\Omega^{1/2}\beta, I_p)$ in Step 2, and (iii) Ω is assumed as known in Step 3 (otherwise valid inference is impossible). We include Stein's normal means model for study because the *glmnet* package is inefficient for the random design model when p is large. In fact, for large p, the efficiency of the *glmnet* package critically depends on the sparsity level of the correlation structure. For Stein's normal means model, by choosing a sparse Ω, the *glmnet* can be efficient even when $p \geq 10^5$. For the random design model, the matrix $X'X$ is dense, and the *glmnet* package becomes inefficient as soon as $p, n \geq 2000$.

We have conducted 4 different experiments, which we now describe. Both the UPS and the lasso need some tuning parameters, the choices of which are discussed separately below.

Experiment 1. In this experiment, we use a Stein's normal means model to investigate the boundaries of Region of Exact Recovery by the UPS and that by the lasso. Fixing $p = 10^4$ and Ω as the tridiagonal matrix in (1.16) with $a = 0.45$, we let ϑ range in $\{0.25, 0.5, 0.65\}$, and let $\pi_p = \nu_{\tau_p}$ with $\tau_p = \sqrt{2r \log p}$, where r is chosen such that $\tau_p \in \{5, 6, \ldots, 12\}$. For both procedures, we use the ideal threshold introduced in Section 2 and Section 4, respectively. That is, the tuning parameters of the UPS are set as $(t_p^*, \lambda_p^{ups}, u_p^{ups}) = (\frac{\vartheta+r}{2r}\tau_p, \sqrt{2\vartheta\log(p)}, \tau_p)$, and the tuning parameter of the lasso is set as $\lambda_p^{lasso} = \max\{\frac{\vartheta+r}{2r}, (1+\sqrt{(1-a)/(1+a)})^{-1}\}\tau_p$.

The results are reported in Table 2, where the UPS outperforms consistently over the lasso, most prominently in the case of $\vartheta = 0.25$. Also, for $\vartheta = 0.25, 0.5$, or 0.65, the Hamming errors of the UPS start to fall below 1 when $\tau_p \geq 8, 7, 7$, but that of the lasso won't fall below 1 until $\tau_p \geq 12, 8, 7$. In Section 1, we show that the UPS yields exact recovery when $\tau_p > (1+\sqrt{1-\vartheta})\sqrt{2\log p}$, where the right-hand side equals $(8.01, 7.32, 7.01)$ with the current choices of (p, ϑ). The numerical results fit well with the theoretic results.

Experiment 2. We use a random design model where $(p, \vartheta, \theta) = (10^4, 0.65, 0.91)$. The experiment contains three sub-experiments 2a–2c. In 2a, we use the same tuning parameters Ω and π_p as in Experiment 1, but $\tau_p \in \{1, 2, \ldots, 7\}$. In 2b–2c, we use the same tuning parameters and setups as in 2a, except that in 2b, for each τ_p, we change π_p from the point mass ν_{τ_p} to the uniform distribution $U(\tau_p - 0.5, \tau_p + 0.5)$, and that in 2c, we replace Ω by the penta-diagonal matrix $\Omega(i,j) = 1_{\{i=j\}} + 0.4 \cdot 1_{\{|i-j|=1\}} + 0.1 \cdot 1_{\{|i-j|=2\}}$. The results

	τ_p	5	6	7	8	9	10	11	12
$\vartheta = 0.25$	UPS	49	11.1	1.79	0.26	0.02	0	0	0
	lasso	186.7	99.35	58.26	38.53	25.97	18.18	12.94	10.57
$\vartheta = 0.50$	UPS	10.06	2.11	0.37	0.09	0	0	0	0
	lasso	16.36	5.11	1.47	0.51	0.28	0.33	0.26	0.09
$\vartheta = 0.65$	UPS	5.49	1.29	0.33	0.06	0	0	0	0
	lasso	7.97	2.43	0.69	0.18	0.07	0.03	.02	.01

Table 2: Hamming errors as in Experiment 1. To yield the exact recovery, the lasso needs stronger signals than the UPS does.

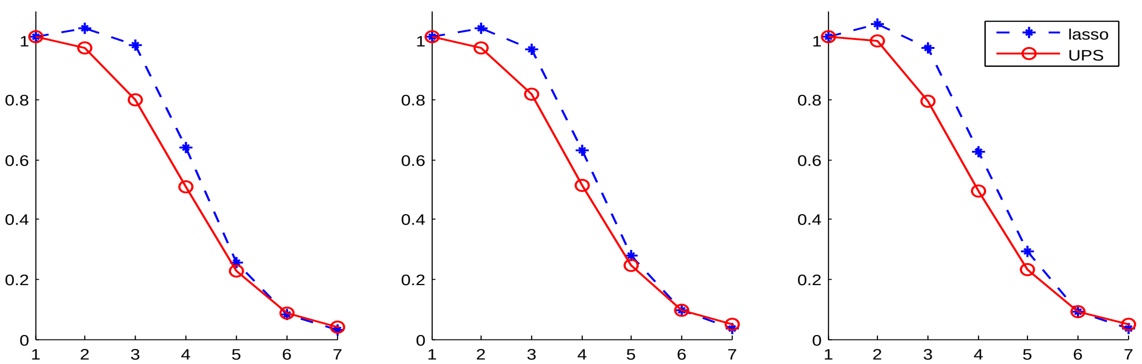

Figure 5: Hamming errors as in Experiment 2a–2c (left, middle, right). The x-axis is τ_p, and the y-axis is the ratio between the Hamming error and the total number of signals $p\epsilon_p$.

are reported in Figure 5, which suggest that the UPS outperforms the lasso over the whole range of τ_p, especially when τ_p is moderately large.

Experiment 3. The goal of this experiment is two-fold. First, we investigate the sensitivity of the UPS and the lasso with respect to their tuning parameters. Second, we investigate the refined UPS introduced in Section 3. For the lasso, we take $\lambda_p^{lasso} = \sqrt{2q\log(p)}$. For the UPS, set the U-step tuning parameter as $t_p^* = \sqrt{2q\log(p)}$ and let the P-step tuning parameters be estimated as in (2.19). Theorem 2.2 predicts that the UPS performs well provided that $q \in (\max\{\vartheta, \delta_0^2(1+\eta)^2 r\}, (\vartheta+r)^2/(4r))$, so both the lasso and the UPS are driven by one tuning parameter q. We now investigate how the choice of q affects the performances of the UPS and the lasso. The experiment contains three sub-experiments 3a–3c.

In 3a, we use a Stein's normal means model where $(p, r) = (10^4, 3)$, $\pi_p = \nu_{\tau_p}$ with $\tau_p = \sqrt{2r\log p}$, Ω is the penta-diagonal matrix satisfying $\Omega(i,j) = 1_{\{i=j\}} + 0.45\cdot 1_{\{|i-j|=1\}} + 0.05\cdot 1_{\{|i-j|=2\}}$, and $\vartheta \in \{0.2, 0.5, 0.65\}$. Note that when $\vartheta = 0.65$, $(\max\{\vartheta, \delta_0^2(1+\eta)^2 r\}, (\vartheta+r)^2/(4r)) = (0.65, 1)$ (and similar for other choices of ϑ), so we let $q \in \{0.7, 0.8, \ldots, 1.1\}$. The results are reported in Figure 6, which suggest that, first, the UPS consistently outperforms the lasso, and second, the UPS is relatively less sensitive to different choices of q.

In 3b, we use a random design model where (p, r, π_p, Ω, q) and the tuning parameters are the same as in 3a, but $\theta = 0.8$ and $\vartheta \in \{0.5, 0.65\}$ (the case $\vartheta = 0.2$ is relatively challenging in computation so is omitted). We compare the lasso with the refined UPS where in each iteration, we use the same tuning parameters as in 3a. The results are reported in Figure

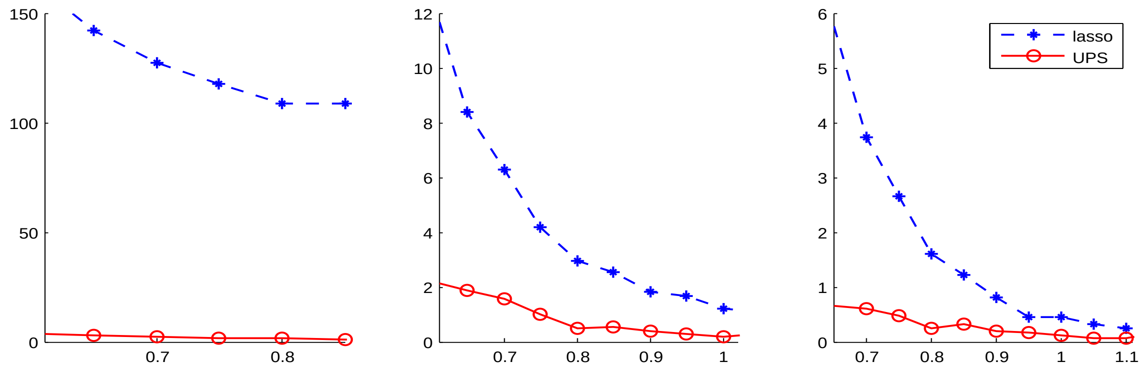

Figure 6: Experiment 3a. The x-axis is q, and the y-axis is the Hamming error. Left to right: $\vartheta = 0.2, 0.5, 0.65$.

7, which suggest that, first, the UPS consistently outperforms the lasso, and, second, the UPS is relatively less sensitive to different choices of q.

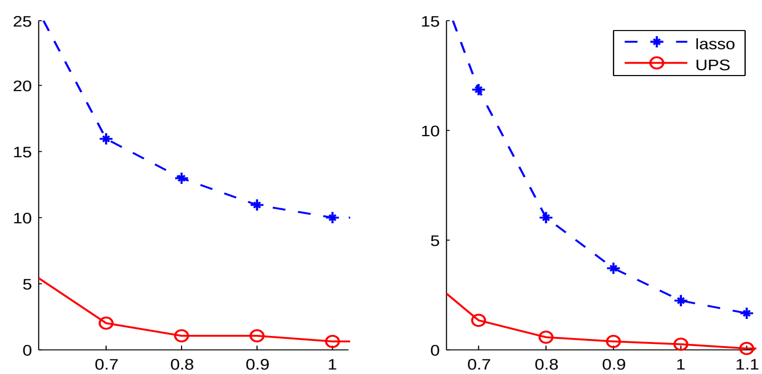

Figure 7: Experiment 3b. The x-axis is q, and the y-axis is the Hamming error. Left: $\vartheta = 0.5$. Right: $\vartheta = 0.65$.

In 3c, we use the same setup as in 3b, except that we fix $q = 1$ and let τ_p range in $\{6, 6.5, \ldots, 9\}$. The results are reported in Figure 8, which suggest that the UPS consistently outperforms the lasso.

Experiment 4. In this experiment, we investigate the effect of larger p and n, respectively. The experiment includes two sub-experiments 4a and 4b.

In 4a, we use a Stein's normal means model where $(\vartheta, r) = (0.5, 3)$, Ω as in Experiment 2c, $\pi_p = \nu_{\tau_p}$ with $\tau_p = \sqrt{2r \log p}$, and we let $p = 100 \times \{1, 10, 10^2, 10^3, 10^4\}$. The lasso and the UPS are implemented as in Experiment 3a, where $q = 1$. The results are reported in the upper part of Table 3, where the second line displays the ratios between the Hamming errors by the lasso and that by the UPS. Theoretic results (Sections 1.7 and 4) predict that for (ϑ, r) in the non-optimal region of the lasso, such ratios diverge as p tends to ∞. The numerical results fit well with the theory.

In 4b, we illustrate that in a random design model, if we fix p and let the sample size n increase, then the random design models get increasingly close to a Stein's normal means model. In detail, we take a random design model where $(p, \vartheta, r) = (10^4, 0.5, 3)$, Ω and π_p as in Experiment 2c, and $n_p = 300 \times \{1, 3, 3^2, 3^3, 3^4\}$. We take also a Stein's normal means

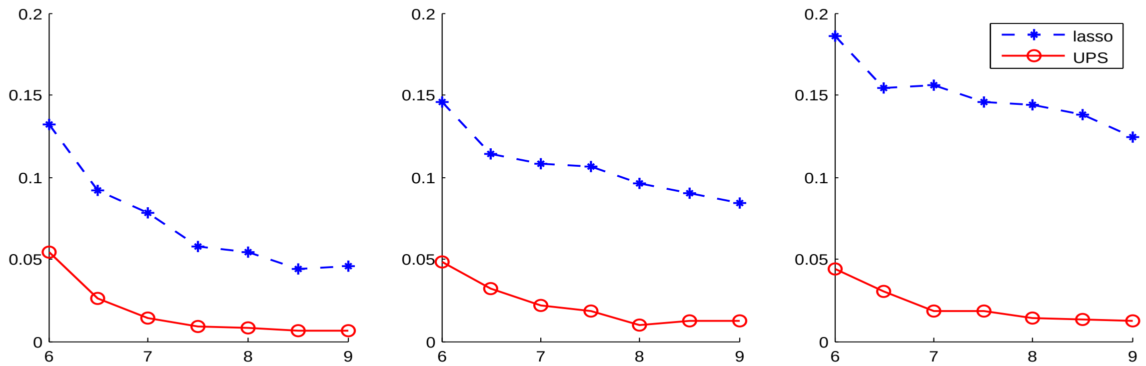

Figure 8: Experiment 3c. The x-axis is τ_p, and the y-axis is the ratio between the Hamming error and $p\epsilon_p$. Left to right: $\vartheta = 0.2, 0.5, 0.65$.

model with the same $(p, \vartheta, r, \Omega, \pi_p)$. We implement the UPS to both models. The results are reported in the lower part of Table 3, where the last line displays the ratios between the Hamming errors by the UPS for the random design model and that for the Stein's normal means model. The ratios effectively converge to 1 as n increases.

p	10^2	10^3	10^4	10^5	10^6
Hamm Ratio	2.43	5.81	6.25	8.80	10.37

n	300	900	2700	8100	24000
Hamm Ratio	479.25	54.04	12.66	1.08	1.01

Table 3: Upper: Ratios between the Hamming errors by the UPS and that by the lasso (Experiment 4a). Lower: Ratios between the Hamming errors by the UPS for the random design model and that for Stein's normal means model (Experiment 4b).

6 Proofs

6.1 Proof of Theorem 1.1.

Fixing $1 \leq j \leq p$, by basic algebra,

$$P\big(\mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j)\big) \geq P(\beta_j = 0, \hat{\beta}_j \neq 0) + P(\beta_j \neq 0, \hat{\beta}_j = 0). \tag{6.1}$$

Consider the hypothesis testing

$$H_{0,j}: \qquad \beta_j = 0, \qquad \text{vs.} \qquad H_{1,j}: \qquad \beta_j \neq 0.$$

Note that any variable selection procedure $\hat{\beta}$ can be viewed as a test which rejects $H_{0,j}$ if and only if $\hat{\beta}_j \neq 0$. Let $f_0^{(j)}(y)$ and $f_1^{(j)}(y)$ be the joint densities of Y under $H_{0,j}$ and $H_{1,j}$, respectively. The superscript $^{(j)}$ is tedious, so we suppress it. Recall that $P(\beta_j \neq 0) = \epsilon_p$. By Neyman-Pearon's fundamental lemma,

$$P(\beta_j = 0, \hat{\beta}_j \neq 0) + P(\beta_j \neq 0, \hat{\beta}_j = 0) \geq \frac{1}{2}\big[1 - \|(1-\epsilon_p)f_0 - \epsilon_p f_1\|_1\big], \tag{6.2}$$

where $\|\cdot\|_1$ denotes the L^1 distance. Combining (6.1) and (6.2) gives

$$P\big(\text{sgn}(\hat{\beta}_j) \neq \text{sgn}(\beta_j)\big) \geq \frac{1}{2}\big[1 - \|(1-\epsilon_p)f_0 - \epsilon_p f_1\|_1\big]. \tag{6.3}$$

We now study $\|(1-\epsilon_p)f_0 - \epsilon_p f_1\|_1$. For any realization of the mean vector β, let $\tilde{\beta} = \beta - \beta_j e_j$, where e_j is j-th basis of $\mathcal{R}^p$. Let $h(y;\tilde{\beta},\alpha)$ be the joint density of $Y \sim N(X(\tilde{\beta}+\alpha e_j), I_n)$. It follows that

$$h(y;\tilde{\beta},\alpha) = h(y,\tilde{\beta},0)\cdot e^{\alpha x_j'(y-X\tilde{\beta}) - \alpha^2 x_j' x_j/2}, \tag{6.4}$$

and that

$$f_0(y) = \int h(y;\tilde{\beta},0)dF(\tilde{\beta}), \qquad f_1(y) = \int h(y;\tilde{\beta},\alpha)d\pi_p(\alpha)dF(\tilde{\beta}), \tag{6.5}$$

where $F(\tilde{\beta})$ denotes the cdf of $\tilde{\beta}$. Using elementary calculus and Fubini's Theorem,

$$\begin{aligned}
\|(1-\epsilon_p)f_0 - \epsilon_p f_1\|_1 &= \int\left|\int\Big((1-\epsilon_p)h(y,\tilde{\beta},0) - \epsilon_p h(y,\tilde{\beta},\alpha)\Big)d\pi_p(\alpha)dF(\tilde{\beta})\right|dy \\
&\leq \int\int\big|(1-\epsilon_p)h(y,\tilde{\beta},0) - \epsilon_p h(y,\tilde{\beta},\alpha)\big|d\pi_p(\alpha)dF(\tilde{\beta})dy \\
&= \int\Big[\int\big|(1-\epsilon_p)h(y,\tilde{\beta},0) - \epsilon_p h(y,\tilde{\beta},\alpha)\big|dy\Big]d\pi_p(\alpha)dF(\tilde{\beta}) \\
&= \int H(\tilde{\beta},\alpha)d\pi_p(\alpha)dF(\tilde{\beta}), \qquad (6.6)
\end{aligned}$$

where $H(\tilde{\beta},\alpha) = H(\tilde{\beta},\alpha;\epsilon_p) = \int\big|(1-\epsilon_p)h(y,\tilde{\beta},0) - \epsilon_p h(y;\tilde{\beta},\alpha)\big|dy$. For any fixed $\tilde{\beta}$, it is seen that $H(\tilde{\beta},\alpha) = H(\tilde{\beta},-\alpha)$ and that $H(\tilde{\beta},\alpha)$ is monotonely increasing in $\alpha \in (0,\infty)$. Therefore, for all $\alpha \in [-\tau_p,0)\cup(0,\tau_p]$,

$$H(\tilde{\beta},\alpha) \leq H(\tilde{\beta},\tau_p). \tag{6.7}$$

Recall that the support of π_p is contained in $[-\tau_p,0)\cup(0,\tau_p]$. Inserting (6.7) into (6.6) gives

$$\|(1-\epsilon_p)f_0 - \epsilon_p f_1\|_1 \leq \int H(\tilde{\beta},\tau_p)dF(\tilde{\beta}). \tag{6.8}$$

The following lemma is proved in Section 6.1.1.

Lemma 6.1 *Suppose the same conditions as in Theorem 1.1 hold. For any realization of* $\tilde{\beta}$*,*

$$\frac{1}{2}\Big[1 - \int\big|(1-\epsilon_p)h(y,\tilde{\beta},0) - \epsilon_p h(y,\tilde{\beta},\tau_p)\big|dy\Big] = (1-\epsilon_p)\bar{\Phi}(\lambda_p) + \epsilon_p\Phi(\lambda_p - \tau_p),$$

where λ_p *is defined as in* (1.9).

Using Lemma 6.1, it follows from (6.8) and definitions that

$$\frac{1}{2}\big[1 - \|(1-\epsilon_p)f_0 - \epsilon_p f_1\|_1\big] \geq (1-\epsilon_p)\bar{\Phi}(\lambda_p) + \epsilon_p\Phi(\lambda_p - \tau_p). \tag{6.9}$$

Inserting (6.9) into (6.3) and noting $s_p = p\epsilon_p$ give the first claim.

Additionally, plugging in $\epsilon_p = p^{-\vartheta}$ and $\tau_p = \sqrt{2r\log p}$ and using Mills' ratio [33] give that as $p \to \infty$,

$$\frac{1-\epsilon_p}{\epsilon_p}\bar{\Phi}(\lambda_p) = L_p p^{-\frac{(r-\vartheta)^2}{4r}}, \qquad \Phi(\lambda_p - \tau_p) = \begin{cases} L_p p^{-\frac{(r-\vartheta)^2}{4r}}, & r > \vartheta, \\ (1+o(1)), & r < \vartheta, \end{cases} \tag{6.10}$$

and the second claim follows. □

6.1.1 Proof of Lemma 6.1

For any realization of $\tilde{\beta}$, let $D_p(\tilde{\beta}) = D_p(\tilde{\beta}; \epsilon_p, \tau_p, X) = \{y: \ \epsilon_p e^{\tau_p x_j'(y - X\tilde{\beta}) - \tau_p^2/2} > (1-\epsilon_p)\}$. By (6.4), $y \in D_p(\tilde{\beta})$ if and only if $\epsilon_p h(y, \tilde{\beta}, \tau_p) > (1-\epsilon_p) h(x, \tilde{\beta}, 0)$. It follows that

$$\int \big|(1-\epsilon_p)h(y,\tilde{\beta},0) - \epsilon_p h(y,\tilde{\beta},\tau_p)\big| dy$$
$$= -\int_{D_p(\tilde{\beta})} [(1-\epsilon_p)h(y,\tilde{\beta},0) - \epsilon_p h(y,\tilde{\beta},\tau_p)]dy + \int_{D_p^c(\tilde{\beta})} [(1-\epsilon_p)h(y,\tilde{\beta},0) - \epsilon_p h(y,\tilde{\beta},\tau_p)]dy.$$

At the same time,

$$1 = \int [(1-\epsilon_p)h(y,\tilde{\beta},0) + \epsilon_p h(y,\tilde{\beta},\tau_p)]dy$$
$$= \int_{D_p(\tilde{\beta})} [(1-\epsilon_p)h(y,\tilde{\beta},0) + \epsilon_p h(y,\tilde{\beta},\tau_p)]dy + \int_{D_p^c(\tilde{\beta})} [(1-\epsilon_p)h(y,\tilde{\beta},0) + \epsilon_p h(y,\tilde{\beta},\tau_p)]dy.$$

Combining these gives

$$\frac{1}{2}\big[1 - \int \big|(1-\epsilon_p)h(y,\tilde{\beta},0) - \epsilon_p h(y,\tilde{\beta},\tau_p)\big|dy\big] = (1-\epsilon_p)\int_{D_p(\tilde{\beta})} h(y,\tilde{\beta},0)dy + \epsilon_p \int_{D_p^c(\tilde{\beta})} h(y,\tilde{\beta},\tau_p)dy. \tag{6.11}$$

Let $W_j(\tilde{\beta}) = x_j'(Y - X\tilde{\beta})$. Note that $Y \in D_p(\tilde{\beta})$ if and only if $W_j(\tilde{\beta}) > \lambda_p$. It follows that

$$\int_{D_p(\tilde{\beta})} h(y,\tilde{\beta},0)dy = P_0(W_j > \lambda_p), \qquad \int_{D_p^c(\tilde{\beta})} h(y,\tilde{\beta},\tau_p)dy = P_1(W_j \leq \lambda_p), \tag{6.12}$$

where P_0 and P_1 denote the law $Y \sim N(X\tilde{\beta}, I_n)$ and $Y \sim N(X(\tilde{\beta} + \tau_p e_j), I_n)$, respectively. Recall that $X'X$ has unit diagonals. It follows that $W_j \sim N(0,1)$ under P_0 and $W_j \sim N(\tau_p, 1)$ under P_1. Combining these with (6.12) gives

$$\int_{D_p(\tilde{\beta})} h(y,\tilde{\beta},0)dy = \bar{\Phi}(\lambda_p), \qquad \int_{D_p^c(\tilde{\beta})} h(y,\tilde{\beta},\tau_p)dy = \Phi(\lambda_p - \tau_p). \tag{6.13}$$

The claim follows by inserting (6.13) into (6.11). □

6.2 Proof of Lemma 2.1

Let D_p be the event

$$\{\|(X'X - \Omega)\beta\|_\infty \leq C\|\Omega\|\sqrt{\log(p)}\, p^{-(\theta-(1-\vartheta))/2}, \ |\frac{\|z\|}{\sqrt{n}} - 1| \leq C\sqrt{\log(p)} p^{-\theta/2}\}. \tag{6.14}$$

By Lemma 3.1, $P(D_p^c) \leq o(1/p)$ for a properly large constant $C > 0$.

Consider the first claim. In this case, $\Omega(i,j) \geq 0$ for all $1 \leq i, j \leq p$. It is sufficient to show for each $1 \leq j \leq p$,

$$P(x_j'Y < t_p^*, \beta_j \neq 0, D_p) \leq L_p p^{-(\vartheta+r)^2/(4r)}.$$

Let e_j be the j-th basis of the $\mathcal{R}^p$. It is seen that over the event D_p, $x_j'Y \approx e_j'\Omega\beta + \sqrt{n}x_j'z/\|z\|$, where the error is algebraically small. Note that $\sqrt{n}x_j'z/\|z\| \sim N(0,1)$, and that when $\beta_j \neq 0$, $e_j'\Omega\beta \geq \beta_j \geq \tau_p$. It follows that

$$P(x_j'Y < t_p^*, \beta_j \neq 0, D_p) \lesssim p^{-\vartheta} P(e_j'\Omega\beta + \sqrt{n}x_j'z/\|z\| < t_p^* | \beta_j \geq \tau_p) \leq p^{-\vartheta}\Phi(t_p^* - \tau_p).$$

Recall that $t_p^* \leq ((\vartheta + r)/(2r))\tau_p$ and $\tau_p = \sqrt{2r \log p}$. The claim follows from Mills' ratio [33].

Consider the second claim. In this case, $r/\vartheta \leq 3 + 2\sqrt{2}$. Fix $1 \leq j \leq p$, let

$$S_j = S_j(\Omega) = \{k : 1 \leq k \leq p, |\Omega(k,j)| \geq \log^{-1}(p)\},$$

and let B_j be the event $\{\beta_k = 0 \text{ for all } k \neq j \text{ and } k \in S_j\}$. By the definition of $\mathcal{M}_p^*(\omega_0, \gamma, A)$, $|S_j| \leq 2\log(p)$, so

$$P(\beta_j \neq 0, B_j^c) \leq \sum_{k \in S_j, k \neq j} P(\beta_j \neq 0, \beta_k \neq 0) \leq 2\log(p)\epsilon_p^2 = 2\log(p)p^{-2\vartheta}. \tag{6.15}$$

Since $r/\vartheta \leq 3 + 2\sqrt{2}$, $2\vartheta \geq (\vartheta + r)^2/(4r)$. Compare (6.15) with the desired claim, it is sufficient to show

$$P(x_j'Y < t_p^*, \beta_j \neq 0, B_j) \leq L_p p^{-(\vartheta+r)^2/(4r)}. \tag{6.16}$$

Towards this end, write $e_j'\Omega\beta = \sum_{k=1}^p \Omega(j,k)\beta_k = \sum_{k \in S_j} \Omega(j,k)\beta_k + \sum_{k \notin S_j} \Omega(j,k)\beta_k$. Over the event $\{\beta_j \neq 0\} \cap B_j$, note that first, $\sum_{k \in S_j} \Omega(j,k)\beta_k = \beta_j \geq \tau_p$, and second,

$$|\sum_{k \notin S_j} \Omega(j,k)\beta_k| \leq C\sqrt{\log(p)} \sum_{k \notin S_j} |\Omega(j,k)| \leq C\sqrt{\log(p)}(\log^{-1}(p))^{1-\gamma} \sum_{k \notin S_j} |\Omega(j,k)|^\gamma,$$

where by the summability condition of Ω, the right-hand side $= o(\sqrt{2\log p})$. It follows that $e_j'\Omega\beta \gtrsim \tau_p$ over the event $\{\beta_j \neq 0\} \cap B_j$. By similar argument as in the proof of the first case, (6.16) follows. □

6.3 Proof of Lemma 2.3.

Write for short $\delta_p = \log^{-1}(p)$. Let D_p be the event $\{|\hat{\Omega}(i,j) - \Omega(i,j)| \leq C\sqrt{\log p} \cdot p^{-\theta/2}, \text{ for all } 1 \leq i,j \leq p\}$. By (2.10), for an appropriately large constant $C > 0$, $P(D_p^c) \leq o(1/p^2)$. It is sufficient to show that for sufficiently large p, both claims hold over D_p.

Consider the first claim. By the definition of $\mathcal{M}_p^*(\omega_0, \gamma, A)$, each row of Ω has at most $2\log(p)$ coordinates exceeding $(1/2 + \omega_0)\delta_p$ in magnitude, where $(1/2 + \omega_0) < 1$. It follows that for sufficiently large p, each row of $\hat{\Omega}$ has at most $2\log(p)$ coordinates exceeding δ_p in magnitude over the event D_p. The claim follows from the definition of Ω^*.

Consider the second claim. The goal is to show that over the event D_p, $\sum_{j=1}^p |\Omega(i,j) - \Omega^*(i,j)| \leq C\delta_p^{(1-\gamma)}$, for all $1 \leq i \leq p$. Write

$$\sum_{j=1}^p |\Omega(i,j) - \Omega^*(i,j)| = I + II, \tag{6.17}$$

where $I = \sum_{\{j: |\Omega^*(i,j)| > \delta_p\}} |\Omega(i,j) - \Omega^*(i,j)|$, and $II = \sum_{\{j: |\Omega^*(i,j)| \leq \delta_p\}} |\Omega(i,j)|$. First, by the definition of D_p and the first claim,

$$I \leq 2\log(p) \max_{1 \leq i,j \leq p} \{|\hat{\Omega}(i,j) - \Omega(i,j)|\} \leq L_p\, p^{-\theta/2}. \tag{6.18}$$

Second, note that over the event D_p, $|\Omega(i,j)| \geq 2\delta_p$ whenever $|\Omega^*(i,j)| \geq \delta_p$. It follows that

$$II \leq \sum_{\{j: |\Omega(i,j)| \leq 2\delta_p\}} |\Omega(i,j)| \leq \sum_{\{j: |\Omega(i,j)| \leq 2\delta_p\}} (|\Omega(i,j)|^\gamma)(|\Omega(i,j)|^{1-\gamma}), \tag{6.19}$$

where by the definition of $\mathcal{M}_p^*(\omega_0, \gamma, A)$, the last term $\leq (2\delta_p)^{1-\gamma} \sum_{j=1}^p |\Omega(i,j)|^\gamma \leq C\delta_p^{1-\gamma}$. Inserting (6.18)-(6.19) into (6.17) gives the claim. □

6.4 Proof of Lemma 2.4

Denote all size ℓ Connected sub-Graph (CG) with respect to (V_0, Ω^*) that contain j by

$$\mathcal{N}_j(\ell) = \{\mathcal{I}_0 = \big\{i_1, i_2, \ldots, i_\ell\big\} \text{ is a CG}: \; i_1 < i_2 < \ldots < i_\ell, \; j \in \mathcal{I}_0\big\}.$$

The following lemma is proved in Frieze and Molloy [21].

Lemma 6.2 *Fix* $1 \leq j \leq p$ *and* $1 \leq k \leq p-1$. *If each row of* Ω^* *has at most* $(k+1)$ *nonzeros, then* $|\mathcal{N}_j(\ell)| \leq (ek)^{\ell-1}$.

For any $\ell \geq 1$, since a CG with size $(\ell+1)$ always contains a CG with size ℓ,

$$P(\mathcal{U}_p(t_p^*) \text{ contains a CG with size} \geq \ell) \leq P(\mathcal{U}_p(t_p^*) \text{ contains a CG with size } \ell).$$

To show the claim, it is sufficient to show that for a constant ℓ_0 to be determined,

$$P(\mathcal{U}_p(t_p^*) \text{ contains a CG with size } \ell_0) \leq o(1/p). \tag{6.20}$$

Recall $\hat{\Omega} = X'X$. Introduce events $D_p^{(1)} = \{|\hat{\Omega}(i,j) - \Omega(i,j)| \leq C\sqrt{\log(p)}p^{-\theta/2}, \, 1 \leq i,j \leq p\}$, $D_p^{(2)} = \{\big|\frac{\sqrt{n}}{\|z\|} - 1\big| \leq C(\sqrt{\log p})p^{-\theta/2}, \|(X'X - \Omega)\beta\|_\infty \leq C(\sqrt{\log p})p^{-(\theta-(1-\vartheta))/2}\}$, and $D_p = D_p^{(1)} \cap D_p^{(2)}$. By (2.10) and Lemma 3.1, $P(D_p^c) \leq o(1/p)$ for a properly large constant $C > 0$. So to show (6.20), it is sufficient to show

$$P(\mathcal{U}_p(t_p^*) \text{ contains a CG with size } \ell_0, \, D_p) \leq o(1/p). \tag{6.21}$$

Recall that by Lemma 2.3, each row of Ω^* has at most $2\log(p)$ nonzero coordinates. Using Lemma 6.2, there are at most $p(2e\log(p))^{\ell_0}$ CG with size ℓ_0. So to show (6.21), it is sufficient to show for any fixed CG of size ℓ_0, say $\mathcal{I}_0 = \{i_1, i_2, \ldots, i_{\ell_0}\}$,

$$P(\mathcal{I}_0 \subset \mathcal{U}_p(t_p^*), D_p) \leq o(1/p^2). \tag{6.22}$$

We now show (6.22). Let $\mathcal{J}_0 = \{1, 2, \ldots, p\}$, and write for short $M = \Omega^{\mathcal{I}_0, \mathcal{J}_0}$, $W = (X'Y)^{\mathcal{I}_0}$, $\eta = (\sqrt{n}X'z/\|z\|)^{\mathcal{I}_0}$, and $\Omega_0 = \Omega^{\mathcal{I}_0, \mathcal{I}_0}$. Note that η is independent of β and $\eta \sim N(0, \Omega_0)$, so

$$\eta'\Omega_0^{-1}\eta \sim \chi^2(\ell_0). \tag{6.23}$$

Note that $W \approx M\beta + \eta$, or more precisely, by definitions and Schwartz inequality,

$$\|\eta\|^2 \geq \frac{1}{2}\|W\|^2 - \|M\beta\|^2 - rem, \qquad \text{over the event } D_p, \tag{6.24}$$

where the reminder term rem is non-stochastic and algebraically small, and so has a negligible effect. Since the largest eigenvalue of Ω_0 does not exceed that of Ω, where the latter ≤ 2,

$$\eta'\Omega_0^{-1}\eta \geq \frac{1}{2}\|\eta\|^2. \tag{6.25}$$

Recall $t_p^* = \sqrt{2q\log(p)}$. By definitions, if $\mathcal{I}_0 \subset \mathcal{U}_p(t_p^*)$, then

$$\|W\|^2 \geq \ell_0 t_p^{*2} \geq 2q\ell_0\log(p). \tag{6.26}$$

Combining (6.24)-(6.26) gives that over the event $\{\mathcal{I}_0 \subset \mathcal{U}_p(t_p^*)\} \cap D_p$,

$$\eta'\Omega_0^{-1}\eta \geq \frac{1}{2}\|\eta\|^2 \geq \frac{1}{2}[q\ell_0\log(p) - \|M\beta\|^2 - rem]. \tag{6.27}$$

The following lemma is proved in Section 6.4.1.

Lemma 6.3 *Fix $k \geq 1$. As $p \to \infty$, there is a constant $C > 0$ such that*

$$P(\|M\beta\|^2 \geq (1+\eta)^2(4k + C\ell_0(\log(p))^{-2(1-\gamma)})\tau_p^2,\, D_p) \leq 2(2\ell_0 \log^\gamma(p))^k p^{-\vartheta k}.$$

Let $k_0 = k_0(\ell_0; q, \gamma, \eta, r, p)$ be the largest k satisfying $(1+\eta)^2(4k + C\ell_0(\log(p))^{-2(1-\gamma)})\tau_p^2 \leq \frac{1}{2}q\ell_0\log(p)$. Denote the event $\{\|M\beta\|^2 \geq (1+\eta)^2(4k_0 + C\ell_0(\log(p))^{-2(1-\gamma)})\tau_p^2\}$ by $\tilde{D}_p$. By Lemma 6.3 and (6.27),

$$P(D_p \cap \tilde{D}_p) \leq L_p p^{-\vartheta k_0}, \qquad \text{and} \qquad \eta'\Omega_0^{-1}\eta \gtrsim \tfrac{1}{4}q\ell_0\log(p) \text{ over } D_p \cap \tilde{D}_p^c. \tag{6.28}$$

As a result,

$$P(\mathcal{I}_0 \subset \mathcal{U}_p(t_p^*), D_p) \leq P(\eta'\Omega_0^{-1}\eta \gtrsim \frac{1}{4}q\ell_0\log(p)) + P(\tilde{D}_p \cap D_p).$$

Using (6.23) and (6.28), it follows from basic statistics that

$$P(\mathcal{I}_0 \subset \mathcal{U}_p(t_p^*), D_p) \leq L_p(p^{-\frac{1}{8}q\ell_0} + p^{-\vartheta k_0}). \tag{6.29}$$

By definitions, $(k_0+1)/\ell_0 \gtrsim q/(16(1+\eta)^2 r)$. Choosing ℓ_0 sufficiently large ensures the existence of k_0, the right-hand side of (6.29) $\leq o(1/p^2)$ and then gives (6.22). □

6.4.1 Proof of Lemma 6.3

Let $S = \{j : 1 \leq j \leq p,\, \Omega^*(i,j) \neq 0 \text{ for some } i \in \mathcal{I}_0\}$. Recall that over the event D_p, each row of Ω^* has at most $2\log(p)$ nonzero coordinates. Since $|\mathcal{I}_0| = \ell_0$,

$$|S| \leq 2\ell_0 \log(p). \tag{6.30}$$

Denote for short $M_1 = \Omega^{\mathcal{I}_0,S}$ and $\xi = \beta^S$. Note that $M\beta - M_1\xi = \Omega^{\mathcal{I}_0,S^c}\beta^{S^c} = (\Omega - \Omega^*)^{\mathcal{I}_0,S^c}\beta^{S^c}$. By Lemma 2.3 and assumptions, $\|\Omega^* - \Omega\|_\infty \leq C(\log(p))^{-(1-\gamma)}$ and $\|\beta\|_\infty \leq (1+\eta)\tau_p$. Therefore, $\|M\beta - M_1\xi\|_\infty \leq C(1+\eta)(\log(p))^{-(1-\gamma)}\tau_p$, and

$$\|M\beta - M_1\xi\|^2 \leq C(1+\eta)^2\ell_0(\log(p))^{-2(1-\gamma)}\tau_p^2. \tag{6.31}$$

At the same time, by basic algebra, the largest eigenvalue of $M_1'M_1$ does not exceed that of Ω^2, where the latter ≤ 4. By $\|\xi\|_\infty \leq \|\beta\|_\infty \leq (1+\eta)\tau_p$,

$$\|M_1\xi\|^2 \leq 4\|\xi\|^2 \leq 4\|\xi\|_0(1+\eta)^2\tau_p^2. \tag{6.32}$$

Combining (6.31)–(6.32) gives

$$\|M\beta\|^2 \leq (1+\eta)^2(4\|\xi\|_0 + C\ell_0(\log(p))^{-2(1-\gamma)})\tau_p^2.$$

Recall that $\epsilon_p = p^{-\vartheta}$ and $\|\xi\|_0$ is distributed as Binomial$(|S|, \epsilon_p)$ (see (2.2)). Using (6.30),

$$P(\|\xi\|_0 \geq k) = \sum_{j=k}^{|S|}\binom{|S|}{j}\epsilon_p^j(1-\epsilon_p)^{|S|-j} \leq \sum_{j=k}^{|S|}(2\ell_0\log(p))^j p^{-\vartheta j} \leq 2(2\ell_0\log(p))^k p^{-\vartheta k}. \tag{6.33}$$

Combining (6.33)-(6.32), the claim follows by recalling $\tau_p = \sqrt{2r\log p}$. □

6.5 Proof of Theorem 2.1

Using (2.10), except for a probability of $o(1/p)$,

$$|(X'X)(i,j) - \Omega(i,j)| \leq L_p p^{-\theta/2}, \qquad \forall\, 1 \leq i,j \leq n. \tag{6.34}$$

Fix $K \geq 1$. It is seen that for all connected subgraph $\mathcal{I}_0 = \{i_1, i_2, \ldots, i_\ell\}$ with size $\ell \leq K$,

$$\|(X'X)^{\mathcal{I}_0,\mathcal{I}_0} - \Omega^{\mathcal{I}_0,\mathcal{I}_0}\|_\infty \leq L_p p^{-\theta/2}. \tag{6.35}$$

Write for short $\hat{\beta} = \hat{\beta}^{ups}(Y, X; t_p^*, \lambda_p^{ups}, u_p^{ups})$. By definitions, $\text{Hamm}_p(\hat{\beta}, \beta) = E[h_p(\hat{\beta}|X)]$, where $h_p(\hat{\beta}|X) \leq p$ for all X. So the event where X does not satisfy either (6.34) or (6.35) only has a negligible effect on the claim. All we need to show is that, for any X satisfying (6.34)-(6.35),

$$h_p(\hat{\beta}|X) \leq L_p p^{1-(\vartheta+r)^2/(4r)}, \tag{6.36}$$

where the right-hand side does not depend on X.

We now show (6.36). Given X satisfying (6.34) and (6.35), write

$$h_p(\hat{\beta}|X) = \sum_{j=1}^p P(\text{sgn}(\hat{\beta}_j) \neq \text{sgn}(\beta_j)|X) = I + II, \tag{6.37}$$

where $I = \sum_{j=1}^p P(\text{sgn}(\hat{\beta}_j) \neq \text{sgn}(\beta_j), j \notin \mathcal{U}_p(t_p^*)|X)$ and $II = \sum_{j=1}^p P(\text{sgn}(\hat{\beta}_j) \neq \text{sgn}(\beta_j), j \in \mathcal{U}_p(t_p^*)|X)$. The dependence on X is tedious and we drop the "$|X$" part below. Consider I. When $j \notin \mathcal{U}_p(t_p^*)$, $x_j'Y < t_p^*$, and $\hat{\beta}_j = 0$. Combining this with Lemma 2.1 gives

$$I \leq \sum_{j=1}^p P(x_j'Y < t_p^*, \beta_j \neq 0) \leq L_p p^{1-(\vartheta+r)^2/(4r)}.$$

It remains to show

$$II \leq L_p p^{1-(\vartheta+r)^2/(4r)}. \tag{6.38}$$

By Lemma 2.4, there are constant $K > 0$ and event A_p such that $P(A_p^c) \leq L_p p^{-(\vartheta+r)^2/(4r)}$ and that $\mathcal{U}_p(t_p^*)$ has the SAS property with respect to (V_0, Ω^*, K) over the event A_p. So to show (6.38), it is sufficient to show that for all $1 \leq j \leq p$,

$$P(\text{sgn}(\hat{\beta}_j) \neq \text{sgn}(\beta_j), j \in \mathcal{U}_p(t_p^*), A_p) \leq L_p p^{-(\vartheta+r)^2/(4r)}. \tag{6.39}$$

By the definition of the SAS property, over the event $\{j \in \mathcal{U}_p(t_p^*)\} \cap A_p$, there exists a unique component $\mathcal{I}_0 = \{i_1, i_2, \ldots, i_\ell\}$ with size $\ell \leq K$ satisfying $j \in \mathcal{I}_0 \lhd \mathcal{U}_p(t_p^*)$. In other words,

$$P(\text{sgn}(\hat{\beta}_j) \neq \text{sgn}(\beta_j), j \in \mathcal{U}_p(t_p^*), A_p) \leq \sum_{\mathcal{I}_0} P(\text{sgn}(\hat{\beta}_j) \neq \text{sgn}(\beta_j), j \in \mathcal{I}_0 \lhd \mathcal{U}_p(t_p^*), A_p), \tag{6.40}$$

where the summation is over all connected subgraphs $\mathcal{I}_0$ of (V_0, Ω^*) that contains j and that has a size $\leq K$. By Lemma 2.3, each row of Ω^* has no more than $2\log(p)$ nonzero coordinates. It follows from Lemma 6.2 that there are at most $C(2e\log(p))^K$ of such $\mathcal{I}_0$. To show (6.38), it is sufficient to show for any fixed connected subgraph $\mathcal{I}_0$ of (V_0, Ω^*) that contains j,

$$P(\text{sgn}(\hat{\beta}_j) \neq \text{sgn}(\beta_j), j \in \mathcal{I}_0 \lhd \mathcal{U}_p(t_p^*)) \leq L_p p^{-\frac{(\vartheta+r)^2}{4r}}. \tag{6.41}$$

Introduce the event $B_p(\mathcal{I}_0) = B_p(\mathcal{I}_0, \beta; X, j)$ through its complement

$$B_p^c(\mathcal{I}_0) = \{\text{There are indices } i \notin \mathcal{I}_0 \text{ and } k \in \mathcal{I}_0 \text{ such that } \beta_i \neq 0,\ \Omega^*(i,k) \neq 0\}.$$

In the event $B_p^c(\mathcal{I}_0) \cap \{j \in \mathcal{I}_0 \lhd \mathcal{U}_p(t_p^*)\} \cap A_p$, we must have $i \notin \mathcal{U}_p(t_p^*)$ and so that $X_i'Y < t_p^*$. In other words, the event $B_p^c(\mathcal{I}_0) \cap \{j \in \mathcal{I}_0 \lhd \mathcal{U}_p(t_p^*)\} \cap A_p$ is contained in the following event:

$$\{\text{There are indices } i \notin \mathcal{I}_0 \text{ and } k \in \mathcal{I}_0 \text{ such that } \beta_i \neq 0,\ \Omega^*(i,k) \neq 0, \text{ and } x_i'Y < t_p^*\}.$$

It follows that

$$P(j \in \mathcal{I}_0 \lhd \mathcal{U}_p(t_p^*), B_p^c \cap A_p) \leq \sum_i P(\beta_i \neq 0, x_i'Y < t_p^*), \tag{6.42}$$

where the summation is over all indices i satisfying that $\Omega^*(i,k) \neq 0$ for some index $k \in \mathcal{I}_0$. Since each row of Ω^* has at most $2\log(p)$ nonzero coordinates, there are at most $2K\log(p)$ such indices i. Additionally, for any fixed i, by the Sure Screening property,

$$P(\beta_i \neq 0, x_i'Y < t_p^*) \leq L_p p^{-(\vartheta+r)^2/(4r)}.$$

Combining these with (6.42) gives that

$$P(j \in \mathcal{I}_0 \lhd \mathcal{U}_p(t_p^*), B_p^c \cap A_p) \leq L_p p^{-(\vartheta+r)^2/(4r)}. \tag{6.43}$$

Comparing (6.43) and (6.41), all that remains to show is

$$P(\mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j), j \in \mathcal{I}_0 \lhd \mathcal{U}_p(t_p^*), B_p \cap A_p) \leq L_p p^{-(\vartheta+r)^2/(4r)}. \tag{6.44}$$

A key fact is that, over the event $\{j \in \mathcal{I}_0 \lhd \mathcal{U}_p(t_p^*)\} \cap B_p \cap A_p$, $(\Omega\beta)^{\mathcal{I}_0} \approx \Omega^{\mathcal{I}_0,\mathcal{I}_0}\beta^{\mathcal{I}_0}$. This is the following lemma, which is proved in Section 6.5.1.

Lemma 6.4 *Over the event* $\{j \in \mathcal{I}_0 \lhd \mathcal{U}_p(t_p^*)\} \cap A_p \cap B_p$,

$$\|(\Omega\beta)^{\mathcal{I}_0} - \Omega^{\mathcal{I}_0,\mathcal{I}_0}\beta^{\mathcal{I}_0}\|_\infty \leq C\tau_p(\log(p))^{-(1-\gamma)}.$$

We now relate the event

$$\{\mathrm{sgn}(\beta_j) \neq \mathrm{sgn}(\hat{\beta}_j), j \in \mathcal{I}_0 \lhd \mathcal{U}_p(t_p^*), B_p \cap A_p\} \tag{6.45}$$

to the P-step. Consider the Penalized MLE

$$\frac{1}{2}\big[\tilde{Y}^{\mathcal{I}_0} - (X'X)^{\mathcal{I}_0,\mathcal{I}_0}\mu\big]'\big((X'X)^{\mathcal{I}_0}\big)^{-1}\big[\tilde{Y}^{\mathcal{I}_0} - (X'X)^{\mathcal{I}_0}\mu\big] + \frac{(\lambda_p^{ups})^2}{2}\|\mu\|_0, \tag{6.46}$$

where the coordinates of μ take values from $\{0, u_p^{ups}\}$, $\lambda_p^{ups} = \sqrt{2\vartheta\log p}$, and $u_p^{ups} = \tau_p = \sqrt{2r\log p}$. Let $\hat{\mu}(\mathcal{I}_0) = \hat{\mu}(\mathcal{I}_0; Y, X, t_p^*, \lambda_p^{ups}, u_p^{ups}, p)$ be the minimizer of (6.45). By the definition of the UPS, the event in (6.45) is contained in the event

$$\{\mathrm{sgn}(\hat{\mu}(\mathcal{I}_0)) \neq \mathrm{sgn}(\beta^{\mathcal{I}_0}), j \in \mathcal{I}_0 \lhd \mathcal{U}_p(t_p^*), B_p \cap A_p\}, \tag{6.47}$$

where $\mathrm{sgn}(\beta)$ is the vector of signs of β. The claim follows from the following lemma, which is proved in Section 6.5.2.

Lemma 6.5 *Suppose the conditions of Theorem 2.1 hold. Fix $1 \leq j \leq p$. As $p \to \infty$, for any fixed $\mathcal{I}_0$ with size $\leq K$ that contains j,*

$$P\big(\mathrm{sgn}(\hat{\mu}(\mathcal{I}_0; Y, X, t_p^*, \lambda_p^{ups}, u_p^{ups}, p)) \neq \mathrm{sgn}(\beta^{\mathcal{I}_0}), j \in \mathcal{I}_0 \lhd \mathcal{U}_p(t_p^*), B_p \cap A_p\big) \leq L_p p^{-(\vartheta+r)^2/(4r)} + p^{-2\vartheta}.$$

If furthermore all coordinates of $\Omega^{\mathcal{I}_0,\mathcal{I}_0}$ are non-negative, then

$$P\big(\mathrm{sgn}(\hat{\mu}(\mathcal{I}_0; Y, X, t_p^*, \lambda_p^{ups}, u_p^{ups}, p)) \neq \mathrm{sgn}(\beta^{\mathcal{I}_0}), j \in \mathcal{I}_0 \lhd \mathcal{U}_p(t_p^*), B_p \cap A_p\big) \leq L_p p^{-(\vartheta+r)^2/(4r)}.$$

□

6.5.1 Proof of Lemma 6.4

Let $\mathcal{I}_0^c = \{j : 1 \leq j \leq p, j \notin \mathcal{I}_0\}$. It is seen that

$$(\Omega\beta)^{\mathcal{I}_0} - \Omega^{\mathcal{I}_0,\mathcal{I}_0}\beta^{\mathcal{I}_0} = \Omega^{\mathcal{I}_0,\mathcal{I}_0}\beta^{\mathcal{I}_0} + \Omega^{\mathcal{I}_0,\mathcal{I}_0^c}\beta^{\mathcal{I}_0^c} - \Omega^{\mathcal{I}_0,\mathcal{I}_0}\beta^{\mathcal{I}_0} = \Omega^{\mathcal{I}_0,\mathcal{I}_0^c}\beta^{\mathcal{I}_0^c}. \tag{6.48}$$

Since $\mathcal{I}_0 \lhd \mathcal{U}_p(t_p^*)$, and over the event B_p, $k \in \mathcal{I}_0$ and $i \in \mathcal{I}_0^c$ imply that either $\beta_i = 0$ or $\Omega^*(k,i) = 0$, we have

$$(\Omega^*)^{\mathcal{I}_0,\mathcal{I}_0^c}\beta^{\mathcal{I}_0^c} = 0. \tag{6.49}$$

Combining (6.48)-(6.49) gives

$$(\Omega\beta)^{\mathcal{I}_0} - \Omega^{\mathcal{I}_0,\mathcal{I}_0}\beta^{\mathcal{I}_0} = (\Omega - \Omega^*)^{\mathcal{I}_0,\mathcal{I}_0^c}\beta^{\mathcal{I}_0^c}.$$

By assumptions and Lemma 2.3, $\|\beta^{\mathcal{I}_0^c}\|_\infty \leq (1+\eta)\tau_p$ and $\|\Omega - \Omega^*\|_\infty \leq C(\log(p))^{-(1-\gamma)}$, so

$$\|(\Omega - \Omega^*)^{\mathcal{I}_0,\mathcal{I}_0^c}\beta^{\mathcal{I}_0^c}\|_\infty \leq \|(\Omega - \Omega^*)^{\mathcal{I}_0,\mathcal{I}_0^c}\|_\infty \cdot \|\beta^{\mathcal{I}_0^c}\|_\infty \leq C\tau_p(\log(p))^{-(1-\gamma)}. \tag{6.50}$$

The claim follows. □

6.5.2 Proof of Lemma 6.5

Write for short $\hat{\mu}(\mathcal{I}_0) = \hat{\mu}(\mathcal{I}_0; Y, X, t_p^*, \lambda_p^{ups}, u_p^{ups}, p)$, $\beta^* = \tau_p \,\mathrm{sgn}(\beta)$ and $\lambda = \lambda_p^{ups} = \sqrt{2\vartheta \log p}$. Introduce the event

$$\tilde{D}_p = \tilde{D}_p(z, X) = \{\|X'z\|_\infty \leq C\sqrt{\log p}\}.$$

Choosing the constant C appropriately large, $P(\tilde{D}_p^c) \leq o(1/p)$. So all we need to show is

$$P(\mathrm{sgn}(\hat{\mu}(\mathcal{I}_0)) \neq \mathrm{sgn}(\beta^{\mathcal{I}_0}), j \in \mathcal{I}_0 \lhd \mathcal{U}_p(t_p^*), B_p \cap A_p \cap \tilde{D}_p) \leq L_p p^{-(\vartheta+r)^2/(4r)}. \tag{6.51}$$

Now, if the sign vector of $\hat{\mu}(\mathcal{I}_0)$ does not match that of $\beta^{\mathcal{I}_0}$, it does not match that of $(\beta^*)^{\mathcal{I}_0}$. By the definitions of $\hat{\mu}(\mathcal{I}_0)$,

$$\begin{aligned}
&\frac{1}{2}(\tilde{Y}^{\mathcal{I}_0} - (X'X)^{\mathcal{I}_0,\mathcal{I}_0}\hat{\mu}(\mathcal{I}_0))'((X'X)^{\mathcal{I}_0,\mathcal{I}_0})^{-1}(\tilde{Y}^{\mathcal{I}_0} - (X'X)^{\mathcal{I}_0,\mathcal{I}_0}\hat{\mu}(\mathcal{I}_0)) + \frac{\lambda^2}{2}\|\hat{\mu}(\mathcal{I}_0)\|_0 \\
\leq &\frac{1}{2}(\tilde{Y}^{\mathcal{I}_0} - (X'X)^{\mathcal{I}_0,\mathcal{I}_0}(\beta^*)^{\mathcal{I}_0})'((X'X)^{\mathcal{I}_0,\mathcal{I}_0})^{-1}(\tilde{Y}^{\mathcal{I}_0} - (X'X)^{\mathcal{I}_0,\mathcal{I}_0}(\beta^*)^{\mathcal{I}_0}) + \frac{\lambda^2}{2}\|(\beta^*)^{\mathcal{I}_0}\|_0.
\end{aligned}$$

By (6.35), $\|(X'X)^{\mathcal{I}_0,\mathcal{I}_0} - \Omega^{\mathcal{I}_0,\mathcal{I}_0}\|_\infty$ is algebraically small. So up to a negligible effect,

$$\begin{aligned}
&\frac{1}{2}(\tilde{Y}^{\mathcal{I}_0} - \Omega^{\mathcal{I}_0,\mathcal{I}_0}\hat{\mu}(\mathcal{I}_0))'(\Omega^{\mathcal{I}_0,\mathcal{I}_0})^{-1}(\tilde{Y}^{\mathcal{I}_0} - \Omega^{\mathcal{I}_0,\mathcal{I}_0}\hat{\mu}(\mathcal{I}_0)) + \frac{\lambda^2}{2}\|\hat{\mu}(\mathcal{I}_0)\|_0 \\
\leq &\frac{1}{2}(\tilde{Y}^{\mathcal{I}_0} - \Omega^{\mathcal{I}_0,\mathcal{I}_0}(\beta^*)^{\mathcal{I}_0})'(\Omega^{\mathcal{I}_0,\mathcal{I}_0})^{-1}(\tilde{Y}^{\mathcal{I}_0} - \Omega^{\mathcal{I}_0,\mathcal{I}_0}(\beta^*)^{\mathcal{I}_0}) + \frac{\lambda^2}{2}\|(\beta^*)^{\mathcal{I}_0}\|_0.
\end{aligned} \tag{6.52}$$

Denote $d = d(\mathcal{I}_0) = \|(\beta^*)^{\mathcal{I}_0}\|_0 - \|\hat{\mu}(\mathcal{I}_0)\|_0$. Reorganizing, it follows from (6.52) that

$$((\beta^*)^{\mathcal{I}_0} - \hat{\mu}(\mathcal{I}_0))'\tilde{Y}^{\mathcal{I}_0} \leq \frac{1}{2}\big[d\lambda^2 + ((\beta^*)^{\mathcal{I}_0})'\Omega^{\mathcal{I}_0,\mathcal{I}_0}(\beta^*)^{\mathcal{I}_0} - \hat{\mu}'(\mathcal{I}_0)\Omega^{\mathcal{I}_0,\mathcal{I}_0}\hat{\mu}(\mathcal{I}_0)\big], \tag{6.53}$$

where by Lemma 6.4, there is an $|\mathcal{I}_0|$ by 1 vector $\tilde{z} \sim N(0, \Omega^{\mathcal{I}_0,\mathcal{I}_0})$ independent of $\beta^{\mathcal{I}_0}$ such that

$$\tilde{Y}^{\mathcal{I}_0} = \Omega^{\mathcal{I}_0,\mathcal{I}_0}\beta^{\mathcal{I}_0} + \tilde{z} + rem, \qquad \|rem\|_\infty \leq o(\sqrt{\log p}). \tag{6.54}$$

Now, for notational simplicity, we drop $\mathcal{I}_0$ everywhere in (6.52)–(6.54). This is a slight misuse of the notations. Note that β and Ω below are low-dimensional. Write

$$\beta - \hat{\mu} = \tau_p(\Delta_1 + \Delta_2), \qquad \text{where} \qquad \Delta_1 = \frac{1}{\tau_p}(\beta^* - \hat{\mu}), \qquad \Delta_2 = \frac{1}{\tau_p}(\beta - \beta^*). \tag{6.55}$$

Plug (6.54)-(6.55) into (6.53) and re-organize. We conclude that over the event (6.51),

$$-\frac{\Delta_1' \tilde{z}}{\sqrt{\Delta_1'\Omega\Delta_1}} \geq \frac{1}{2\sqrt{\Delta_1'\Omega\Delta_1}}\big(-d(\vartheta/r) + 2\Delta_1'\Omega\Delta_2 + \Delta_1'\Omega\Delta_1\big)\sqrt{2r\log p} + o(\sqrt{\log p}), \tag{6.56}$$

where the $o(\sqrt{2\log(p)})$ term is non-stochastic and has a negligible effect.

Let B_{nn} be the number of zero coordinates of β estimated as 0, B_{ns} be the number of those estimated as τ_p. Let B_{sn} be the number of nonzero coordinates of β that are estimated as 0, and B_{ss} be the number of those estimated as τ_p. Note that, first, over the event in (6.51), $B_{ns} + B_{sn} \geq 1$. Otherwise, the sign vector of $\hat{\mu}$ matches that of β. Second, the probability that $\mathcal{I}_0$ contains $B_{sn} + B_{ss}$ signals $\sim \epsilon_p^{B_{sn}+B_{ss}} = p^{-\vartheta(B_{sn}+B_{ss})}$. Third, since $\tilde{z} \sim N(0, \Omega)$, $(\Delta_1'\tilde{z}/\sqrt{\Delta_1'\Omega\Delta_1}) \sim N(0,1)$. Combining these with (6.56), to show (6.51), it is sufficient to show

$$p^{-\vartheta(B_{sn}+B_{ss})}\bar{\Phi}\Big(\frac{\big(-d(\vartheta/r) + 2\Delta_1'\Omega\Delta_2 + \Delta_1'\Omega\Delta_1\big)}{2\sqrt{\Delta_1'\Omega\Delta_1}}\sqrt{2r\log p}\Big)$$
$$\leq \begin{cases} L_p p^{-\frac{(\vartheta+r)^2}{4r}}, & \text{if } \Omega \text{ only has non-negative coordinates,} \\ L_p p^{-\frac{(\vartheta+r)^2}{4r}} + p^{-2\vartheta}, & \text{if } \Omega \text{ may have negative coordinates,} \end{cases} \tag{6.57}$$

where $\bar{\Phi} = 1 - \Phi$ is the survival function of $N(0,1)$.

First, we consider (6.57) for the case where Ω only has non-negative coordinates. Before we proceed further, we note that, first, when a zero coordinate of β is estimated as 0, it has no effect on the desired inequality. So without loss of generality, we assume $B_{nn} = 0$. Second, the proof for the case $B_{sn} + B_{ss} \geq (\vartheta + r)^2/(4\vartheta r)$ is trivial, so we assume $B_{sn} + B_{ss} < (\vartheta + r)^2/(4\vartheta r)$. Third, the case $B_{sn} + B_{ss} = 0$ is easy. In fact, note that $d = B_{sn} - B_{ns} \leq -1$, $\Delta_1'\Omega\Delta_1 \geq 1$, and $\Delta_2 = 0$. So

$$B_{sn} + B_{ss} = 0, \qquad \frac{-d(\vartheta/r) + 2\Delta_1'\Omega\Delta_2 + \Delta_1'\Omega\Delta_1}{2\sqrt{\Delta_1'\Omega\Delta_1}} \geq \frac{1 + (\vartheta/r)}{2}.$$

The left-hand side of (6.57) is

$$p^{-\vartheta(B_{sn}+B_{ss})} \cdot \bar{\Phi}\left(\frac{-d(\vartheta/r) + 2\Delta_1'\Omega\Delta_2 + \Delta_1'\Omega\Delta_1}{2\sqrt{\Delta_1'\Omega\Delta_1}}\sqrt{2r\log p}\right) \leq \bar{\Phi}(\frac{1 + (\vartheta/r)}{2}\sqrt{2r\log p}),$$

and the claim follows from Mills' ratio [33]. Last, the case $B_{ns} = 0$ but $B_{sn} + B_{ss} \leq 1$ is also relatively easy. In this case, as $\mathrm{sgn}(\hat{\mu}) \neq \mathrm{sgn}(\beta)$, B_{ns} and B_{sn} can not be 0 at the same time, and we must have $B_{sn} = 1$ and $B_{ss} = 0$. It follows that $d = 1$, $\Delta_1 = 1$, $\Delta_2 \geq 0$, and $\Omega = 1$. So

$$B_{sn} + B_{ss} = 1, \qquad \frac{-d(\vartheta/r) + 2\Delta_1'\Omega\Delta_2 + \Delta_1'\Omega\Delta_1}{2\sqrt{\Delta_1'\Omega\Delta_1}} \geq \frac{1 - (\vartheta/r)}{2}.$$

Using Mills' ratio [33], the claim follows from

$$p^{-\vartheta(B_{sn}+B_{ss})}\cdot\bar{\Phi}\left(\frac{-d(\vartheta/r)+2\Delta_1'\Omega\Delta_2+\Delta_1'\Omega\Delta_1}{2\sqrt{\Delta_1'\Omega\Delta_1}}\sqrt{2r\log p}\right)\leq\epsilon_p\bar{\Phi}(\frac{1-(\vartheta/r)}{2}\sqrt{2r\log p}).$$

In light of these observations, below, we assume $B_{nn}=0$ and

$$1\leq B_{sn}+B_{ss}\leq(\vartheta+r)^2/(4\vartheta r),\qquad\text{and}\qquad\text{when } B_{ns}=0,\ B_{ss}+B_{sn}\geq 2. \tag{6.58}$$

The following lemma is proved in Section 6.5.3.

Lemma 6.6 *Fix $\omega_0\in[0,1/2)$. Suppose that Ω has unit diagonals and only non-negative coordinates, and that*

$$\max\{\|U(\Omega)\|_\infty,\|U(\Omega)\|_1\}\leq\omega_0. \tag{6.59}$$

Then

$$\Delta_1'\Omega\Delta_1\geq\begin{cases}2B_{sn}-2\omega_0(2B_{sn}-1), & B_{sn}=B_{ns}\geq 1,\\ (B_{sn}+B_{ns})-4\omega_0\min\{B_{sn},B_{ns}\}, & B_{sn}\neq B_{ns}.\end{cases}$$

By Cauchy-Schwartz inequality,

$$|\Delta_1'\Omega\Delta_2|\leq\sqrt{\Delta_1'\Omega\Delta_1}\sqrt{\Delta_2'\Omega\Delta_2}. \tag{6.60}$$

First, by assumptions, the largest eigenvalue of Ω is bounded by $1+2\omega_0$, so

$$\Delta_2'\Omega\Delta_2\leq(1+2\omega_0)\|\Delta_2\|_2^2. \tag{6.61}$$

Second, recall that the support of π_p is contained in $[\tau_p,(1+\eta)\tau_p]$. By definitions, Δ_2 has $(B_{ss}+B_{sn})$ nonzero coordinates, each of which $\leq\eta$ in magnitude. It follows that

$$\Delta_2'\Omega\Delta_2\leq(1+2\omega_0)\|\Delta_2\|_2^2\leq(1+2\omega_0)(B_{ss}+B_{sn})\eta^2. \tag{6.62}$$

Recall that $B_{sn}+B_{ss}\leq(\vartheta+r)^2/(4r\vartheta)$. Combining this with (6.60)-(6.62) gives

$$|\Delta_1'\Omega\Delta_2|\leq\sqrt{(1+2\omega_0)\frac{(\vartheta+r)^2}{4\vartheta r}\eta^2}\cdot\sqrt{\Delta_1'\Omega\Delta_1}. \tag{6.63}$$

Write for short $c=c(\eta;\vartheta,r,\omega_0)=(1+2\omega_0)\frac{(\vartheta+r)^2}{4\vartheta r}\eta^2$. By the definition of η (i.e. (2.6)),

$$2\sqrt{c}\leq\min\{\frac{2\vartheta}{r},1-\frac{\vartheta}{r},\sqrt{2-2\omega_0}-1+\frac{\vartheta}{r}\}. \tag{6.64}$$

Combining these with (6.63) gives

$$\frac{-d(\vartheta/r)+2\Delta_1'\Omega\Delta_2+\Delta_1'\Omega\Delta_1}{2\sqrt{\Delta_1'\Omega\Delta_1}}\geq\frac{-d(\vartheta/r)+\Delta_1'\Omega\Delta_1}{2\sqrt{\Delta_1'\Omega\Delta_1}}-\sqrt{c}. \tag{6.65}$$

We now discuss three different cases (a) $B_{ns}=B_{sn}\geq 1$, (b) $B_{ns}>B_{sn}$, and (c) $B_{ns}<B_{sn}$ separately.

Consider (a). In this case, $d=0$, and by Lemma 6.6, $\Delta_1'\Omega\Delta_1\geq 2B_{sn}(1-2\omega_0)+2\omega_0\geq 2-2\omega_0$. It follows that

$$\frac{-d(\vartheta/r)+\Delta_1'\Omega\Delta_1}{2\sqrt{\Delta_1'\Omega\Delta_1}}-\sqrt{c}=\frac{1}{2}\sqrt{\Delta_1'\Omega\Delta_1}-\sqrt{c}\geq\frac{1}{2}(\sqrt{2-2\omega_0}-2\sqrt{c}). \tag{6.66}$$

By (6.64),

$$2\sqrt{c} \leq \sqrt{2-2\omega_0} - 1 + (\vartheta/r). \tag{6.67}$$

Combining (6.65)-(6.67) gives

$$\frac{-d(\vartheta/r) + 2\Delta_1'\Omega\Delta_2 + \Delta_1'\Omega\Delta_1}{2\sqrt{\Delta_1'\Omega\Delta_1}} \geq \frac{1}{2}(1 - \frac{\vartheta}{r}).$$

Inserting this into (6.57) and noting $B_{ss} + B_{sn} \geq 1$, the claim follows by Mills' ratio [33].

Consider (b). In this case, $B_{ns} > B_{sn}$ and so $d \leq -1$. First, by (6.64), $\sqrt{c} \leq \vartheta/r$. Second, note that the function $[\frac{(\vartheta/r)+x}{2\sqrt{x}} - \sqrt{c}]$ is positive and monotonely increasing in the range of $x \geq 1$, and that by Lemma 6.6, $\Delta_1'\Omega\Delta_1 \geq 1$. It follows that

$$\frac{-d(\vartheta/r) + \Delta_1'\Omega\Delta_1}{2\sqrt{\Delta_1'\Omega\Delta_1}} - \sqrt{c} \geq \frac{1}{2}(1 + \frac{\vartheta}{r}) - \frac{\vartheta}{r} = \frac{1}{2}(1 - \frac{\vartheta}{r}).$$

By (6.58), $B_{sn} + B_{ss} \geq 1$. Inserting these into (6.57), the claim follows by Mills' ratio [33].

Consider (c). In this case, $B_{ns} < B_{sn}$. We have either $B_{ns} = 0$ or $B_{ns} \geq 1$. By (6.58), we have that in either case, $B_{sn} + B_{ss} \geq 2$. First, suppose $\vartheta/r \geq 1/3$. In this case, $2\vartheta \geq (\vartheta + r)^2/(4r)$, and the claim follows by $p^{-\vartheta(B_{sn}+B_{ss})} \leq p^{-2\vartheta}$. Next, suppose $0 < \vartheta/r < 1/3$. Note that $d = B_{sn} - B_{ns} \geq 1$. By Lemma 6.6, $\Delta_1'\Omega\Delta_1 \geq B_{sn} - B_{ns}$. Recall that, for given $d \geq 1$ and $r > \vartheta$, the function $\frac{-d(\vartheta/r)+x}{2\sqrt{x}}$ is positive and monotonely increasing in the range of $x \geq d$. Combining these gives

$$\frac{-d(\vartheta/r) + \Delta_1'\Omega\Delta_1}{2\sqrt{\Delta_1'\Omega\Delta_1}} \geq \frac{-d(\vartheta/r) + d}{2\sqrt{d}} \geq \frac{1}{2}(1 - \frac{\vartheta}{r}).$$

At the same time, by (6.64), $\sqrt{c} \leq \theta/r$. It follows that

$$\frac{-d(\vartheta/r) + \Delta_1'\Omega\Delta_1}{2\sqrt{\Delta_1'\Omega\Delta_1}} - \sqrt{c} \geq \frac{1}{2}(1 - \vartheta/r) - \vartheta/r = \frac{1}{2}(1 - 3\vartheta/r).$$

Inserting this into (6.57) and recalling $B_{sn} + B_{ss} \geq 2$, the claim follows from

$$p^{-2\vartheta}\bar{\Phi}(\frac{1}{2}(1 - 3\vartheta/r)\sqrt{2r\log p}) = L_p p^{-2\vartheta-(r-3\vartheta)^2/(4r)} \leq L_p p^{-(\vartheta+r)^2/(4r)},$$

where we have used Mills' ratio [33]. This proves (6.57) for the case where Ω has only non-negative coordinates.

Next, consider (6.57) for the case where Ω may have negative coordinates. The proof for the case $B_{sn} + B_{ss} \geq 2$ is trivial, so we only consider the case $B_{sn} + B_{ss} \leq 1$. By similar arguments as in Lemma 6.6,

$$\Delta_1'\Omega\Delta_1 \geq 1. \tag{6.68}$$

We now consider three cases (a) $B_{sn} + B_{ss} = 0$, (b) $B_{sn} = 1$ and $B_{ss} = 0$, and (c) $B_{sn} = 0$ and $B_{ss} = 1$, separately.

Consider (a). In this case, $B_{ns} \geq 1$ and so $d \leq -1$. Also, we must have $\Delta_2 = 0$. By (6.68) and the monotonicity of the function $((\vartheta/r) + x)/\sqrt{x}$ in $x \in [1, \infty)$,

$$\frac{-d(\vartheta/r) + \Delta_1'\Omega\Delta_1 + 2\Delta_1'\Omega\Delta_2}{2\sqrt{\Delta_1'\Omega\Delta_1}} \geq \frac{\vartheta/r + \Delta_1'\Omega\Delta_1}{2\sqrt{\Delta_1'\Omega\Delta_1}} \geq \frac{1}{2}(1 + \frac{\vartheta}{r}),$$

and the claim follows by similar arguments.

Consider (b). In this case, $d \leq 1$ and $\Delta_1'\Omega\Delta_2 \geq 0$. By (6.68) and the monotonicity of the function $(-(\vartheta/r)+x)/\sqrt{x}$ in $x \in [1,\infty)$,

$$\frac{-d(\vartheta/r) + \Delta_1'\Omega\Delta_1 + 2\Delta_1'\Omega\Delta_2}{2\sqrt{\Delta_1'\Omega\Delta_1}} \geq \frac{-(\vartheta/r) + \Delta_1'\Omega\Delta_1}{2\sqrt{\Delta_1'\Omega\Delta_1}} \geq \frac{1}{2}(1 - \vartheta/r).$$

Noting that $B_{sn} + B_{ss} = 1$, the claim follows by similar arguments.

Consider (c). In this case, $\Delta_1'\Omega\Delta_2 \geq -\omega_0\eta \geq -\vartheta/r$, where we have used the condition $\eta \leq 2\vartheta/r$. Note that in this case, we must have $B_{ns} \geq 1$, so $d \leq -1$. By (6.68) and the monotonicity of the function $(-(\vartheta/r)+x)/\sqrt{x}$ in $x \in [1,\infty)$,

$$\frac{-d(\vartheta/r) + \Delta_1'\Omega\Delta_1 + 2\Delta_1'\Omega\Delta_2}{2\sqrt{\Delta_1'\Omega\Delta_1}} \geq \frac{-(\vartheta/r) + \Delta_1'\Omega\Delta_1}{2\sqrt{\Delta_1'\Omega\Delta_1}} \geq \frac{1}{2}(1 - \vartheta/r),$$

and the claim follows similarly. □

6.5.3 Proof of Lemma 6.6

Without loss of generality, assume all coordinates of Δ_1 are nonzero. Write for short $A_1 = B_{sn}$, $A_2 = B_{ns}$ and $k = A_1 + A_2$. Introduce a k by k diagonal matrix Λ such that $\Lambda(i,i)$ is the sign of the i-th coordinate of Δ_1. For notational simplicity, we write $\Delta = \Delta_1$, and let Δ_i be the i-th coordinate of Δ, $1 \leq i \leq k$. Let $\tilde{\Omega} = \Lambda'\Omega\Lambda$. Note that $|\tilde{\Omega}(i,j)| = |\Omega(i,j)|$ for all $1 \leq i,j \leq k$, and so $\max\{\|U(\tilde{\Omega})\|_\infty, \|U(\tilde{\Omega})\|_1\} \leq \omega_0$. It is seen that

$$\Delta'\Omega\Delta = 1'\Lambda'\Omega\Lambda 1 = 1'\tilde{\Omega}1, \tag{6.69}$$

where 1 is the k by 1 vector of ones. We discuss the case $A_1 = A_2 \geq 1$ and the case $A_1 \neq A_2$ separately.

In the first case, $A_1 = A_2 \geq 1$. By the assumptions of the lemma and direct calculations,

$$1\tilde{\Omega}1 = \sum_{i=1}^{k} \tilde{\Omega}(i,i) + 2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} \tilde{\Omega}(i,j) \geq k - 2\sum_{i=1}^{k-1}\omega_0 \geq k - 2(k-1)\omega_0. \tag{6.70}$$

In the second case, $A_1 \neq A_2$. By symmetry, we only show the case $A_1 > A_2$. Let $S_1 = \{1 \leq i \leq k: \ \Delta_i = 1\}$ and $S_2 = \{1 \leq i \leq p, \ \Delta_i = -1\}$. Note that $|S_1| = A_1$ and $|S_2| = A_2$, and that $\tilde{\Omega}(i,j) \leq 0$ if and only if $i \in S_1$ and $j \in S_2$, or $i \in S_2$ and $j \in S_1$. It follows that

$$1'\tilde{\Omega}1 = \sum_{i=1}^{k} \tilde{\Omega}(i,i) + \sum_{i \neq j} \tilde{\Omega}(i,j) \geq k + (I + II), \tag{6.71}$$

where $I = \sum_{i \in S_1, j \in S_2} \tilde{\Omega}(i,j)$ and $II = \sum_{i \in S_2, j \in S_1} \tilde{\Omega}(i,j)$. By the assumptions of the lemma and the symmetry of $\tilde{\Omega}$, for each fixed $j \in S_2$, $\sum_{i \in S_1} |\tilde{\Omega}(i,j)| \leq 2\omega_0$. Similarly, for each fixed $i \in S_2$, $\sum_{j \in S_1} |\tilde{\Omega}(i,j)| \leq 2\omega_0$. Inserting these into (6.71) gives $1'\tilde{\Omega}1 \geq (A_1 + A_2) - 4\omega_0 A_2$, and the claim follows. □

6.6 Proof of Lemma 2.5

In this section and Sections 6.6.1 and 6.6.2, we denote $t = t_p^*$ for simplicity. Since the proofs are similar, we only show the first claim. Note that except for a probability of $o(1/p)$, $|\tilde{Y}_j| \leq C\sqrt{2\log p}$ for some constant $C > 0$. Write for short $\delta_p = 1/\log(p)$, and let $\tilde{\Omega}$ be the matrix where $\tilde{\Omega}(i,j) = \Omega(i,j)1_{\{|\Omega(i,j)| \geq \delta_p\}}$, $1 \leq i,j \leq p$. By the summability assumption of Ω and elementary algebra, we have (i) each row of $\tilde{\Omega}$ has no more than $2\log(p)$ nonzero coordinates, (ii) $\|\Omega - \tilde{\Omega}\|_\infty \leq C(\log(p))^{-(1-\gamma)}$, and (iii) there is a non-stochastic term $a_p = (1 + o(1))$ such that $a_p\tilde{\Omega} - \Omega$ is positive semi-definite (note $\|\tilde{\Omega} - \Omega\|_\infty = o(1)$). Recall that $\tilde{Y} = X'X\beta + X'z$, where $\sqrt{n}X'z/\|z\| \sim N(0,\Omega)$. Let $\eta \sim N(0, a_p\tilde{\Omega} - \Omega)$ be a Gaussian random vector that is independent of $\sqrt{n}X'z/\|z\|$. Introduce

$$W = \tilde{\Omega}\beta + \frac{1}{\sqrt{a_p}}(\sqrt{n}X'z/\|z\| + \eta).$$

It is seen that $W \sim N(\tilde{\Omega}\beta, \tilde{\Omega})$. Additionally, there is a non-stochastic term $b_p = o(1)$ such that except for a probability of $o(1/p)$,

$$\|W - \tilde{Y}\|_\infty \leq b_p \cdot \sqrt{2\log(p)}. \tag{6.72}$$

In fact, letting $\tilde{W} = \Omega\beta + \sqrt{n}X'z/\|z\|$, we write

$$\|W - \tilde{Y}\|_\infty \leq \|W - \tilde{W}\|_\infty + \|\tilde{W} - \tilde{Y}\|_\infty. \tag{6.73}$$

First, by Lemma 3.1, except for a probability of $o(1/p)$,

$$\|\tilde{Y} - \tilde{W}\|_\infty \leq C\sqrt{\log(p)}(p^{-(\theta-(1-\vartheta))/2} + p^{-\theta/2}). \tag{6.74}$$

Second, by definitions, $\|W - \tilde{W}\|_\infty \leq \|(\Omega - \tilde{\Omega})\beta\|_\infty + (|\frac{1}{\sqrt{a_p}} - 1|)\|\frac{\sqrt{n}X'z}{\|z\|}\|_\infty + \frac{1}{\sqrt{a_p}}\|\eta\|_\infty$. It follows from (i)–(iii) and elementary statistics that except for a probability of $o(1/p)$,

$$\|W - \tilde{W}\|_\infty \leq o(\sqrt{2\log(p)}). \tag{6.75}$$

Inserting (6.74)-(6.75) into (6.73) gives (6.72).

Now, introduce event $A_p = \{\|\tilde{Y} - W\|_\infty \leq b_p\sqrt{2\log(p)}\}$, and $\bar{F}_p^{\pm}(t) = \frac{1}{p}\sum_{j=1}^p 1_{\{W_j \pm b_p\sqrt{2\log p} \geq t\}}$. Comparing $\bar{F}_p^{\pm}(t)$ with $\bar{F}_p(t)$, it is seen that over the event A_p,

$$\bar{F}_p^-(t) \leq \bar{F}_p(t) \leq \bar{F}_p^+(t).$$

The claim follows from the following lemma, which is proved in Section 6.6.1.

Lemma 6.7 *Under the conditions of Lemma 2.5, there is a constant $c = c(\vartheta, r) > 0$ such that, with probability $1 - o(1/p)$,*

$$\left|\frac{1}{p\epsilon_p}\sum_{j=1}^p 1_{\{W_j \geq t\}} - 1\right| \leq L_p p^{-c(\vartheta,r)}.$$

6.6.1 Proof of Lemma 6.7

Let $\tilde{\Omega}$ be defined as above. The following lemma is proved below in Section 6.6.2.

Lemma 6.8 *Suppose $Y \sim N(0, \tilde{\Omega})$, and $S_p(t) = \sum_{j=1}^p 1_{\{Y_j \geq t\}}$. Fixing an integer $m > 0$,*

$$E[(S_p(t))^m] \leq C(m)(1 + 2ep\log(p)\bar{\Phi}(t))^m.$$

As a result, for any fixed constant $c_0 > 0$, $P(S_p(t) \geq p^{c_0} E[S_p(t)]) \leq o(1/p)$.

We now proceed to prove Lemma 6.7. Write $W = \tilde{\beta} + \tilde{z}$, where we bear in mind that (i) $\tilde{\beta} = \tilde{\Omega}\beta$ and $\tilde{z} \sim N(0, \tilde{\Omega})$, (ii) $\tilde{\beta}$ and $\tilde{z}$ are independent, (iii) each row of $\tilde{\Omega}$ has no more than $2\log(p)$ nonzero coordinates, and (iv) if $\beta_j \neq 0$, then $\tau_p \leq \beta_j \leq (1+\eta)\tau_p$. For each $1 \leq j \leq p$, let $D_j = \{1 \leq k \leq p: \ \tilde{\Omega}(j,k) \neq 0\}$, and let A_{0j}, A_{1j}, and A_{2j} be correspondingly the events where there are none, one, and two or more indices $k \in D_j$ such that $\beta_k \neq 0$. Write

$$\frac{1}{p}\sum_{j=1}^p 1_{\{W_j \geq t\}} = \frac{1}{p}(I + II + III),$$

where $I = \sum_{j=1}^p 1_{\{W_j \geq t\}} 1_{\{A_{0j}\}}$, $II = \sum_{j=1}^p 1_{\{W_j \geq t\}} 1_{\{A_{2j}\}}$, and $III = \sum_{j=1}^p 1_{\{W_j \geq t\}} 1_{\{A_{1j}\}}$.

Consider I first. Note that over the event A_{0j}, $\tilde{\beta}_j = 0$. It follows from (i) that $I \leq \sum_{j=1}^p 1_{\{W_j \geq t, \tilde{\beta}_j = 0\}} \leq \sum_{j=1}^p 1_{\{\tilde{z}_j \geq t\}}$. By Lemma 6.8, for any fixed $c_0 > 0$, as $p \to \infty$, except for a probability of $o(1/p)$,

$$I \leq p^{c_0} \sum_{j=1}^p P(\tilde{z}_j \geq t) = p^{1+c_0}\bar{\Phi}(t). \tag{6.76}$$

Consider II. Introduce the set $H = \{(k,\ell): \ k < \ell, \text{and } \tilde{\Omega}(j,k) \neq 0, \tilde{\Omega}(j,\ell) \neq 0 \text{ for some } 1 \leq j \leq p\}$. It is seen that $|H| \leq 4\log^2(p)p$, and that

$$\sum_{j=1}^p 1_{\{A_{2j}\}} \leq \sum_{j=1}^p \sum_{\{k \in D_j, \ell \in D_j, k < \ell\}} 1_{\{\beta_k \neq 0, \beta_\ell \neq 0\}} = \sum_{\{(k,\ell) \in H\}} 1_{\{\beta_k \neq 0, \beta_\ell \neq 0\}}.$$

Define a graph where each element of H is a node, and two nodes (k,ℓ) and (k',ℓ') are connected if and only if $\{k,\ell\} \cap \{k',\ell'\} \neq \emptyset$. Fixing a node (k,ℓ), we calculate the number of nodes (k',ℓ') that are connected to (k,ℓ). Note that two nodes are connected if and only if $k = k'$, $k = \ell'$, $\ell = k'$, or $\ell = \ell'$. Take the first case for example. By definition, there is a j such that $\tilde{\Omega}(j,k) \neq 0$ and $\tilde{\Omega}(j,\ell') \neq 0$. By (iii), for a given k, there are $2\log(p)$ different choices of j, and for a given j, there are $2\log(p)$ different choices of ℓ'. It follows that there are no more than $4\log^2(p)$ nodes (k',ℓ') that may be connected to (k,ℓ). By similar argument as in the proof of Lemma 6.8, except for a probability of $o(1/p)$,

$$\sum_{\{(k,\ell) \in H\}} 1_{\{\beta_k \neq 0, \beta_\ell \neq 0\}} \leq p^{c_0} E[\sum_{\{(k,\ell) \in H\}} 1_{\{\beta_k \neq 0, \beta_\ell \neq 0\}}] \leq 4\log^2(p)p^{1+c_0}\epsilon_p^2, \tag{6.77}$$

where we have used $|H| \leq 4p\log^2(p)$. It follows that

$$II \leq \sum_{j=1}^p 1_{\{A_{2j}\}} \leq 4\log^2(p)p^{1+c_0}\epsilon_p^2. \tag{6.78}$$

Consider III. Write $III = IIIa + IIIb - IIIc$, where $IIIa = \sum_{j=1}^p 1_{\{W_j \geq t\}} 1_{\{A_{1j}\}} 1_{\{\beta_j = 0\}}$, $IIIb = \sum_{j=1}^p 1_{\{A_{1j}\}} 1_{\{\beta_j \neq 0\}}$, and $IIIc = \sum_{j=1}^p 1_{\{W_j < t\}} 1_{\{A_{1j}\}} 1_{\{\beta_j \neq 0\}}$. Consider $IIIa$. Write for short $\delta_0 = \delta_0(\Omega)$. Note that over the event $A_{1j} \cap \{\beta_j = 0\}$, $\tilde{\beta}_j \leq \delta_0(1+\eta)\tau_p$. Fix a realization of β, let $j_1 < j_2 < \ldots j_\ell$ be all the indices at which $1_{\{A_{1j}\}} 1_{\{\beta_j = 0\}} = 1$. Using (i)-(ii),

$$IIIa \leq \sum_{j=1}^p 1_{\{\tilde{z}_j \geq t - \delta_0(1+\eta)\tau_p\}} 1_{\{A_{1j}\}} 1_{\{\beta_j = 0\}} \leq \sum_{k=1}^\ell 1_{\{\tilde{z}_{j_k} \geq t - \delta_0(1+\eta)\tau_p\}}.$$

Using Lemma 6.8, for any $c_0 > 0$, as $p \to \infty$, except for a probability of $o(1/p)$,

$$\sum_{k=1}^\ell 1_{\{\tilde{z}_{j_k} \geq t - \delta_0(1+\eta)\tau_p\}} \leq p^{c_0} \sum_{k=1}^\ell P(\tilde{z}_{j_k} \geq t - \delta_0(1+\eta)\tau_p) \leq p^{c_0} \ell \bar{\Phi}(t - \delta_0(1+\eta)\tau_p). \quad (6.79)$$

Since (6.79) holds for all the realizations of β, and that except for a probability of $o(1/p)$, $\ell \leq \|\beta\|_0 \leq 2p\epsilon_p$, it follows that

$$IIIa \leq 2p^{1+c_0}\epsilon_p \bar{\Phi}(t - \delta_0(1+\eta)\tau_p). \quad (6.80)$$

Consider $IIIb$. Write

$$IIIb = \sum_{j=1}^p 1_{\{\beta_j \neq 0\}} - \sum_{j=1}^p 1_{\{\beta_j \neq 0\}} 1_{\{A_2 j\}}, \quad (6.81)$$

where we have used the fact $1_{\{A_{0j}\}} 1_{\{\beta_j \neq 0\}} = 0$. Note that except for a probability of $o(1/p)$, $|\sum_{j=1}^p 1_{\{\beta_j \neq 0\}} - p\epsilon_p| \leq C\sqrt{\log(p)/(p\epsilon_p)}$, and that by (6.78), $\sum_{j=1}^p 1_{\{\beta_j \neq 0\}} 1_{\{A_2 j\}} \leq \sum_{j=1}^p 1_{\{A_{2j}\}} \leq 4\log^2(p) p^{1+c_0}\epsilon_p^2$. It follows that except for a probability of $o(1/p)$,

$$|IIIb - p\epsilon_p| \leq C/\sqrt{p\epsilon_p} + 4\log^2(p)p^{1+c_0}\epsilon_p^2. \quad (6.82)$$

Consider $IIIc$. Note that over the event $A_{1j} \cap \{\beta_j \neq 0\}$, $\tilde{\beta}_j = \beta_j \geq \tau_p$. By (i)-(ii), $IIIc \leq \sum_{j=1}^p 1_{\{W_j < t\}} 1_{\{\tilde{\beta}_j \geq \tau_p\}} \leq \sum_{j=1}^p 1_{\{\tilde{z}_j < t - \tau_p\}} 1_{\{\tilde{\beta}_j \neq 0\}}$. Note that except for a probability of $o(1/p)$, $\sum_{j=1}^p 1_{\{\tilde{\beta}_j \neq 0\}} \leq 2\log(p) \sum_{j=1}^p 1_{\{\beta_j \neq 0\}} \leq 4\log(p)p\epsilon_p$. By similar arguments as in the proof of $IIIa$, for any fixed $c_0 > 0$, as $p \to \infty$, except for a probability of $o(1/p)$,

$$IIIc \leq \sum_{j=1}^p 1_{\{\tilde{z} < t - \tau_p\}} 1_{\{\tilde{\beta}_j \neq 0\}} \leq 4\log(p)p^{1+c_0}\epsilon_p \bar{\Phi}(\tau_p - t). \quad (6.83)$$

Combining (6.80), (6.82), and (6.83) gives that except for a probability of $o(1/p)$,

$$|III - p\epsilon_p| \leq C\log^2(p)\left[p^{1+c_0}\epsilon_p^2 \bar{\Phi}(t - \delta_0(1+\eta)\tau_p) + p^{1+c_0}\epsilon_p \bar{\Phi}(\tau_p - t) + \sqrt{\frac{1}{p\epsilon_p}} + p^{1+c_0}\epsilon_p^2\right]. \quad (6.84)$$

Recall $t = t_p^* = \sqrt{2q\log p}$ where $\max\{\delta_0^2(1+\eta)^2 r, \vartheta\} < q \leq \frac{(\vartheta+r)^2}{4r}$. Combining (6.76), (6.78), and (6.84), the claim follows by Mill's ratio [33].

6.6.2 Proof of Lemma 6.8

The second claim follows directly by Chebyshev's inequality, so we only show the first claim. Write

$$E[S_p^m(t)] = \sum_{k=1}^{m} \sum_{a_1+\ldots a_k=m} \sum_{i_1<\ldots<i_k} E\Big[(1_{\{Y_{i_1}\geq t\}})^{a_1}\ldots(1_{\{Y_{i_k}\geq t\}})^{a_k}\Big],$$

where $a_i \geq 1$ are integers, $1 \leq i \leq k$. By basic combinatorics,

$$\begin{aligned} E[S_p^m(t)] &= \sum_{k=1}^{m} \sum_{a_1+\ldots a_k=m} \sum_{i_1<\ldots<i_k} E\Big[(1_{\{Y_{i_1}\geq t\}})\ldots(1_{\{Y_{i_k}\geq t\}})\Big] \\ &\leq \sum_{k=1}^{m} \binom{m-1}{k-1} \sum_{i_1<\ldots<i_k} E\Big[(1_{\{Y_{i_1}\geq t\}})\ldots(1_{\{Y_{i_k}\geq t\}})\Big]. \end{aligned} \tag{6.85}$$

Form a graph where $\{1, 2, \ldots, p\}$ are the nodes and nodes $\{i, j\}$ are connected if and only if $\tilde{\Omega}(i, j) \neq 0$. For $1 \leq \ell \leq k$, let $\mathcal{M}(\ell; k) = \{\{i_1 < \ldots < i_k\} : \{i_1, \ldots, i_k\}$ splits into ℓ different CG$\}$, where CG stands for connected subgraph as before. First, by Lemma 6.2 and basic combinatorics, $|\mathcal{M}(\ell; k)| \leq \binom{p}{\ell}\binom{k-1}{\ell-1}(2e\log(p))^k \leq C(m)p^\ell(2e\log(p))^k$. Second, note that for any $\{i_1, \ldots, i_k\} \in \mathcal{M}(\ell; k)$, $E[(1_{\{Y_{i_1}\geq t\}})\ldots(1_{\{Y_{i_k}\geq t\}})] \leq (\bar{\Phi}(t))^\ell$. Combining these gives that for each $1 \leq k \leq m$,

$$\begin{aligned} \sum_{i_1<\ldots<i_k} E[(1_{\{Y_{i_1}\geq t\}})\ldots(1_{\{Y_{i_k}\geq t\}})] &= \sum_{\ell=1}^{k} \sum_{\{\{i_1,\ldots,i_k\}\in\mathcal{M}(\ell;k)\}} E[(1_{\{Y_{i_1}\geq t\}})\ldots(1_{\{Y_{i_k}\geq t\}})] \\ &\leq \sum_{\ell=1}^{k}(2e\log(p))^k \sum_{\ell=1}^{k}(p\bar{\Phi}(t))^\ell \leq k(2e\log(p)\bar{\Phi}(t))^k. \end{aligned}$$

Inserting this into (6.85) gives the claim. □

6.7 Proof of Theorem 2.2.

Let $(\lambda_p^{ups}, u_p^{ups})$ be the tuning parameters as in Theorem 2.1. Write for short $(\lambda_p, u_p) = (\lambda_p^{ups}, u_p^{ups})$ and $(\hat{\lambda}_p, \hat{u}_p) = (\hat{\lambda}_p^{ups}, \hat{u}_p^{ups})$. The proof is similar to that of Theorem 2.1 except one difference: the non-stochastic tuning parameters (λ_p, u_p) are replaced by stochastic tuning parameters $(\hat{\lambda}_p, \hat{u}_p)$. By a close investigation of the proof of Theorem 2.1, it is sufficient to show that Lemma 6.5 continues to hold if we replace (λ_p, u_p) by $(\hat{\lambda}_p, \hat{u}_p)$, except for that the generic logarithmic term L_p may be different. Towards this end, note that by Lemma 2.5, there is a positive number $\delta_p = o(1)$ such that except for a probability of $o(1/p)$,

$$(1-\delta_p)\lambda_p \leq \hat{\lambda}_p \leq (1+\delta_p)\lambda_p, \qquad (1-\delta_p)u_p \leq \hat{u}_p \leq (1+\delta_p)u_p. \tag{6.86}$$

Note that Lemma 6.5 continues to hold if we replace λ_p by $(1 \pm \delta_p)\lambda_p$ and u_p by $(1 \pm \delta_p)u_p$. The claim follows by (6.86) and a close investigation of the proof of Lemma 6.5. □

6.8 Proof of Lemma 3.1

The first claim follows directly from [4], so we only show the second claim. Let e_j be the j-th basis of the $\mathcal{R}^p$. All we need to show is that for each $1 \leq j \leq p$, except for a probability of $o(1/p^2)$, $|e_j'(X'X - \Omega)\beta| \leq C\|\Omega\|\sqrt{\log p}\; p^{-[\theta-(1-\vartheta)]/2}$. By symmetry, it is sufficient to show this for $j = 1$ only. Denote $a = (X'X - \Omega)e_1$ and write $e_1'(X'X - \Omega)\beta = \sum_{i=1}^p a_i\beta_i$. It is sufficient to show that except for a probability of $o(1/p^2)$,

$$|\sum_{i=1}^p a_i\beta_i| \leq C\|\Omega\|\sqrt{\log p}\; p^{-[\theta-(1-\vartheta)]/2}. \tag{6.87}$$

Towards this end, let $\mu_p = \mu_p(a, \pi_p) = \frac{1}{p}\sum_{i=1}^p E[a_i\beta_i]$ and $\sigma_p^2 = \sigma_p^2(a, \pi_p) = \frac{1}{p}\sum_{i=1}^p a_i^2 \mathrm{Var}(\beta_i)$. Direct calculation shows that

$$p\mu_p \asymp \epsilon_p\sqrt{\log p}\sum_{i=1}^p a_i, \qquad p\sigma_p^2 \asymp \epsilon_p \log(p)\sum_{i=1}^p a_i^2. \tag{6.88}$$

First, let $Z = X\Omega^{-1/2}$, $\xi = \Omega^{1/2}e_1$, and $\eta = \frac{1}{\sqrt{p}}\Omega^{1/2}1_p/\sqrt{\|\Omega\|}$. Note that $\|\xi\|^2 = e_1'\Omega e_1 = 1$ and $\|\eta\|^2 = \frac{1}{\|\Omega\|}(\frac{1}{p}1_p'\Omega 1_p) \leq 1$. It follows that

$$\sum_{i=1}^p a_i = e_1'(X'X - \Omega)1_p = \sqrt{p\|\Omega\|}(\xi'Z'Z\eta - \xi'\eta). \tag{6.89}$$

Write $Z = (Z_1, Z_2, \ldots, Z_n)'$ and $\xi'Z'Z\eta - \xi'\eta = \frac{1}{n}\sum_{i=1}^n(\sqrt{n}\xi'Z_i)(\sqrt{n}\eta'Z_i) - \xi'\eta$. Note that for $1 \leq i \leq n$, $(\sqrt{n}\xi'Z_i, \sqrt{n}\eta'Z_i)'$ are iid samples from a bivariate normal with variances $\|\xi\|^2$ and $\|\eta\|^2$, and covariance $\xi'\eta$. By similar arguments as in [4] and that $n = p^\theta$, except for a probability of $o(1/p^2)$,

$$|\xi'Z'Z\eta - \xi'\eta| \leq C\sqrt{\log(n)}/\sqrt{n} \leq Cp^{-\theta/2}\sqrt{\log(p)}. \tag{6.90}$$

Combining (6.88)-(6.90), we have that except for a probability of $o(1/p^2)$,

$$p\mu_p \leq C\epsilon_p\log(p)\sqrt{p\|\Omega\|}p^{-\theta/2} \leq C\log(p)\sqrt{\|\Omega\|}p^{-\frac{\vartheta}{2}-[\theta-(1-\vartheta)]/2}. \tag{6.91}$$

Second, write

$$\sum_{i=1}^p a_i^2 = e_1'(X'X - \Omega)(X'X - \Omega)e_1 = \xi'(Z'Z - I_p)\Omega(Z'Z - I_p)\xi. \tag{6.92}$$

It is known [30] that except for a probability of $o(1/p^2)$, the largest eigenvalue of $(Z'Z - I_p)$ is no greater than $C\sqrt{p/n}$ in absolute value. Recalling $\|\xi\| = 1$,

$$\xi'(Z'Z - I_p)\Omega(Z'Z - I_p)\xi \leq \|\Omega\|\xi'(Z'Z - I_p)(Z'Z - I_p)\xi \leq C\|\Omega\|(p/n). \tag{6.93}$$

Combining (6.88), (6.92) and (6.93) gives

$$p\sigma_p^2 \leq C\|\Omega\|\log(p)\epsilon_p p/n \leq C\|\Omega\|\log(p)p^{-[\theta-(1-\vartheta)]}. \tag{6.94}$$

Last, since $\beta_i \leq C\sqrt{\log p}$, using Bennett's lemma [28], for any $\lambda > 0$,

$$P(\sum_{i=1}^p a_i\beta_i \geq p\mu_p + \sqrt{p}\lambda) \leq \exp\big(-\frac{\lambda^2}{2\sigma_p^2}\psi(\frac{\lambda C\sqrt{\log p}}{\sigma_p^2\sqrt{p}})\big), \tag{6.95}$$

where $\psi(x) > 0$ and $x\psi(x)$ is monotonely increasing in $x \in (0, \infty)$. Choose λ such that

$$\sqrt{p}\lambda = C\|\Omega\|\sqrt{\log(p)\epsilon_p(p/n)} = C\|\Omega\|\sqrt{\log(p)}\ p^{-[\theta-(1-\vartheta)]/2}.$$

Using (6.94), it follows from (6.95) that

$$P(\sum_{i=1}^{p} a_i\beta_i \geq p\mu_p + \sqrt{p}\lambda) = o(1/p^2). \tag{6.96}$$

Combining (6.96) with (6.91) and (6.94) gives (6.87). □

6.9 Proof of Lemma 4.1

For national simplicity, write for short $\beta_1 = \beta_{j-1}$, $\beta_2 = \beta_j$, $\hat{\beta}_1 = \hat{\beta}_{j-1}$, $\hat{\beta}_2 = \hat{\beta}_j$, $\tilde{y}_1 = \tilde{Y}_{j-1}$, and $\tilde{y}_2 = \tilde{Y}_j$. By the KKT condition [31], $(\hat{\beta}_1, \hat{\beta}_2)'$ minimizes the functional if and only if there is a sub-gradient $\alpha = (\alpha_1, \alpha_2)'$ such that

$$\begin{pmatrix} 1 & a \\ a & 1 \end{pmatrix}\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} - \begin{pmatrix} \tilde{y}_1 \\ \tilde{y}_2 \end{pmatrix} + \lambda\alpha = 0, \qquad \text{and} \qquad \begin{cases} \alpha_i = \mathrm{sgn}(\hat{\beta}_i), & \text{if } \hat{\beta}_i \neq 0, \\ |\alpha_i| \leq 1, & \text{otherwise.} \end{cases} \tag{6.97}$$

Since the proofs are similar, we only show that for Regions I, IIa, and $IIIa$.

Consider Region I. For $i = 1, 2$, construct $\hat{\beta}_i = 0$ and $\alpha_i = \tilde{y}_i/\lambda$. It is seen that the first requirement in (6.97) is satisfied. Moreover, note that $|\tilde{y}_i| \leq \lambda$ in the current region. It follows that $|\alpha_i| \leq 1$, and the constructions satisfy the second requirement in (6.97) as well. So in this case, the minimizer is $(\hat{\beta}_1, \hat{\beta}_2) = (0, 0)$.

Consider Region IIa. Construct $\hat{\beta}_1 = \tilde{y}_1 - \lambda$, $\hat{\beta}_2 = 0$, $\alpha_1 = 1$, and $\alpha_2 = [(\tilde{y}_2 - a\tilde{y}_1) + a\lambda)]/\lambda$. Direct calculations show that these satisfy the first requirement of (6.97). Moreover, since $-(1 + a)\lambda < (\tilde{y}_2 - a\tilde{y}_1) < (1 - a)\lambda$, $|\alpha_2| \leq 1$, so this construction also satisfies the second requirement of (6.97). So in this case, $(\hat{\beta}_1, \hat{\beta}_2) = (\tilde{y}_1 - \lambda, 0)$.

Consider Region $IIIa$. Set $\alpha_1 = \alpha_2 = 1$ and

$$\hat{\beta}_1 = \frac{1}{1 - a^2}[(\tilde{y}_1 - \lambda) - a(\tilde{y}_2 - \lambda)], \qquad \hat{\beta}_2 = \frac{1}{1 - a^2}[(\tilde{y}_2 - \lambda) - a(\tilde{y}_1 - \lambda)].$$

Direct calculations show that these constructions satisfy the first requirement of (6.97). Moreover, by the definition of Region $IIIa$, $\hat{\beta}_1 > 0$ and $\hat{\beta}_2 > 0$, so $\alpha_i = \mathrm{sgn}(\hat{\beta}_i)$ and the second requirement of (6.97) is also satisfied. Combining these gives the claim. □

6.10 Proof of Lemma 4.2

Write for short $\hat{\beta} = \hat{\beta}^{lasso}$ and $\lambda_p = \lambda_p^{lasso} = \sqrt{2q\log(p)}$. Introduce events $A_{0j} = \{\beta_k = 0,\ j - 2 \leq k \leq j + 1\}$, $A_{1j} = \{\beta_{j-2} = \beta_{j-1} = \beta_{j+1} = 0, \beta_j = \tau_p\}$, $B_{0j} = \{\hat{\beta}_{j-2} = \hat{\beta}_{j-1} = \hat{\beta}_j = \hat{\beta}_{j+1} = 0\}$, and $B_{1j} = \{\hat{\beta}_{j-2} = \hat{\beta}_{j-1} = \hat{\beta}_{j+1} = 0, \hat{\beta}_j \neq 0\}$. The Hamming distance satisfies

$$\sum_{j=1}^{p} P(\mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j)) \geq \sum_{j=3}^{p-1}\big[P(\hat{\beta}_j \neq 0, \beta_j = 0) + P(\hat{\beta}_j = 0, \beta_j \neq 0)\big] \geq \frac{1}{7}\sum_{j=3}^{p-1}(I_j + II_j),$$

where

$$I_j = \sum_{k=j-2}^{j+1} P(\hat{\beta}_k \neq 0, \beta_k = 0), \qquad II_j = P(\hat{\beta}_j = 0, \beta_j = \tau_p) + \sum_{k\in\{j-2,j-1,j+1\}} P(\hat{\beta}_k \neq 0, \beta_k = 0).$$

By basic algebra and definitions, $I_j \geq \sum_{k=j-2}^{j+1} P(\hat{\beta}_k \neq 0, A_{0j}) \geq P(A_{0j} \cap B_{0j}^c)$, and $II_j \geq P(\hat{\beta}_j = 0, A_{1j}) + \sum_{k\in\{j-2,j-1,j+1\}} P(\hat{\beta}_k \neq 0, A_{1j}) \geq P(A_{1j} \cap B_{1j}^c)$. It follows that

$$\sum_{j=1}^{p} P(\text{sgn}(\hat{\beta}_j) \neq \text{sgn}(\beta_j)) \geq \frac{1}{7}\sum_{j=3}^{p-1}[P(A_{0j} \cap B_{0j}^c) + P(A_{1j} \cap B_{1j}^c)]. \tag{6.98}$$

Let R be a two-dimensional region as follows

$$\left\{(x, y) : \frac{x - ay}{1 - a} > \lambda_p \text{ and } \frac{y - ax}{1 - a} > \lambda_p, \quad \text{or} \quad \frac{y - ax}{1 + a} > \lambda_p \text{ and } \frac{x - ay}{1 + a} < -\lambda_p\right\}.$$

We introduce the events

$$D_{0j} = \{|\tilde{Y}_j| > \lambda_p\}, \qquad D_{1j} = \{|\tilde{Y}_j| \leq \lambda_p\}, \qquad \tilde{D}_{1j} = \{(\tilde{Y}_{j-1}, \tilde{Y}_j)' \in R\}.$$

Note that $D_{1j} \cap \tilde{D}_{1j} = \emptyset$. We now show that

$$B_{0j}^c \supseteq \{|\tilde{Y}_j| > \lambda_p\}, \qquad B_{1j}^c \supseteq (D_{1j} \cup \tilde{D}_{1j}). \tag{6.99}$$

This is equivalent to show that

$$B_{0j} \cap D_{0j} = \emptyset, \qquad B_{1j} \cap (D_{1j} \cup \tilde{D}_{1j}) = \emptyset. \tag{6.100}$$

Towards this end, we note that by the KKT condition [31],

$$\Omega\hat{\beta} = \tilde{Y} - \lambda_p \alpha, \tag{6.101}$$

where α is the vector of sub-gradients (i.e. $\alpha_j = \text{sgn}(\hat{\beta}_j)$ if $\hat{\beta}_j \neq 0$ and $|\alpha_j| \leq 1$ otherwise). Consider the first claim in (6.100). Recall that Ω is a tridiagonal matrix. When B_{0j} happens, it follows from (6.101) that $0 = \hat{\beta}_j = \tilde{Y}_j - \lambda_p\alpha_j$. Therefore, $|\tilde{Y}_j| \leq \lambda_p$, and the claim follows. Consider the second claim of (6.100). When B_{1j} happens, it follows from Lemma 4.1 that

$$\begin{cases} \tilde{Y}_j < -\lambda_p, \\ -(1-a)\lambda_p \leq \tilde{Y}_{j-1} - a\tilde{Y}_j \leq (1+a)\lambda_p, \end{cases} \quad \text{or} \quad \begin{cases} \tilde{Y}_j > \lambda_p, \\ -(1+a)\lambda_p \leq \tilde{Y}_{j-1} - a\tilde{Y}_j \leq (1-a)\lambda_p. \end{cases}$$

Then (6.100) follows by noting that

$$\begin{aligned} &\{|\tilde{Y}_j| > \lambda_p\} \cap D_{1j} = \emptyset, \\ &\{\tilde{Y}_j < -\lambda_p, -(1-a)\lambda_p \leq \tilde{Y}_{j-1} - a\tilde{Y}_j \leq (1+a)\lambda_p\} \cap \tilde{D}_{1j} = \emptyset, \\ &\{\tilde{Y}_j > \lambda_p, -(1+a)\lambda_p \leq \tilde{Y}_{j-1} - a\tilde{Y}_j \leq (1-a)\lambda_p\} \cap \tilde{D}_{1j} = \emptyset. \end{aligned}$$

Next, note that $D_{1j} \cap \tilde{D}_{1j} = \emptyset$. Combining (6.98) and (6.99) gives

$$\sum_{j=1}^{p} P(\text{sgn}(\hat{\beta}_j) \neq \text{sgn}(\beta_j)) \geq \frac{1}{7}\sum_{j=3}^{p-1}[P(A_{0j} \cap D_{0j}) + P(A_{1j} \cap D_{1j}) + P(A_{1j} \cap \tilde{D}_{1j})]. \tag{6.102}$$

By definitions, $P(A_{0j}) = (1 - \epsilon_p)^4$, $P(A_{1j}) = (1 - \epsilon_p)^3\epsilon_p$, that conditional on A_{0j}, $\tilde{Y}_j \sim N(0, 1)$, and that conditional on A_{1j}, $\tilde{Y}_j \sim N(\tau_p, 1)$. It follows from elementary statistics and definitions that

$$P(A_{0j} \cap D_{0j}) = (1 - \epsilon_p)^4 P(N(0, 1) \geq \lambda_p) = L_p p^{-q}, \tag{6.103}$$

and that

$$P(A_{1j} \cap D_{1j}) = (1-\epsilon_p)^3 \epsilon_p P(N(\tau_p, 1) \le \lambda_p) = \begin{cases} L_p p^{-[\vartheta + (\sqrt{q}-\sqrt{r})^2]}, & q < r, \\ p^{-\vartheta}(1+o(1)), & q > r. \end{cases} \quad (6.104)$$

At the same time, $P(A_{1j}) = (1-\epsilon_p)^3\epsilon_p$, so

$$P(A_{1j} \cap \tilde{D}_{1j}) = (1-\epsilon_p)^3 \epsilon_p P\big((\tilde{Y}_{j-1}, \tilde{Y}_j)' \in R | A_{1j}\big).$$

Note that conditional on A_{1j}, $\tilde{Y}_{j-1} \sim N(a\tau_p, 1)$, $\tilde{Y}_j \sim N(\tau_p, 1)$, and $\mathrm{Cov}(\tilde{Y}_{j-1}, \tilde{Y}_j) = a$. Directly evaluating $P\big((\tilde{Y}_{j-1}, \tilde{Y}_j)' \in R|A_{1j}\big)$ gives

$$P(A_{1j} \cap \tilde{D}_{1j}) = \begin{cases} L_p p^{-\vartheta - \frac{1-|a|}{1+|a|} q}, & 0 < q < r, \\ L_p p^{-\vartheta \frac{1}{1+|a|}(2q + (1+|a|)r - 2(1+|a|)\sqrt{qr})}, & r < q. \end{cases} \quad (6.105)$$

Inserting (6.103)-(6.105) into (6.102) gives the claim. □

6.11 Proof of Lemma 4.3

For simplicity, write for short $\lambda_p = \lambda_p^{ss}$, $\beta_1 = \beta_{j-1}$, $\beta_2 = \beta_j$, $\hat{\beta}_1 = \hat{\beta}_{j-1}$, $\hat{\beta}_2 = \hat{\beta}_j$, $\tilde{y}_1 = \tilde{Y}_{j-1}$, and $\tilde{y}_2 = \tilde{Y}_j$. Direct calculations show that the minimum of the functional is

$$\begin{cases} 0, & \text{if } \beta_1 = 0 \ \&\ \beta_2 = 0, \\ (\lambda_p^2 - \tilde{y}_1^2)/2, & \text{if } \beta_1 \ne 0 \ \&\ \beta_2 = 0, \\ (\lambda_p^2 - \tilde{y}_2^2)/2, & \text{if } \beta_1 = 0 \ \&\ \beta_2 \ne 0, \\ \lambda_p^2 - (\tilde{y}_1^2 + \tilde{y}_2^2 - 2a\tilde{y}_1\tilde{y}_2)/(2(1-a^2)), & \text{if } \beta_1 \ne 0 \ \&\ \beta_2 \ne 0, \end{cases} \quad (6.106)$$

obtained at $(\beta_1, \beta_2)' = (0,0)$, $(\tilde{y}_1, 0)'$, $(0, \tilde{y}_2)'$, and $((\tilde{y}_1 - a\tilde{y}_2)/(1-a^2), (\tilde{y}_2 - a\tilde{y}_1)/(1-a^2))'$, correspondingly. Write for short $A_{1a} = (\lambda_p^2 - \tilde{y}_1^2)/2$, $A_{1b} = (\lambda_p^2 - \tilde{y}_1^2)/2$, and $A_2 = \lambda_p^2 - (\tilde{y}_1^2 + \tilde{y}_2^2 - 2a\tilde{y}_1\tilde{y}_2)/(2(1-a^2))$. We now discuss the regions one by one. By symmetry, we only show that for Regions I, IIa and $IIIa$.

In Region I, it is seen that $A_{1a} > 0$, $A_{1b} > 0$, and $A_2 > 0$. By (6.106), the minimum of the functional is achieved at $(\beta_1, \beta_2)' = (0,0)$, and the claim follows. In Region IIa, we have $|\tilde{y}_1| > \lambda_p$, $|\tilde{y}_2| < |\tilde{y}_1|$, and $|a\tilde{y}_1 - \tilde{y}_2| < \lambda_p\sqrt{1-a^2}$. Correspondingly, it follows that $A_{1a} < 0$, $A_{1a} < A_{1b}$, and $A_{1a} < A_2$, and the claim follows. In Region $IIIa$, we have $\tilde{y}_1^2 + \tilde{y}_2^2 - 2a\tilde{y}_1\tilde{y}_2 - 2\lambda_p^2(1-a^2) > 0$, $|a\tilde{y}_1 - \tilde{y}_2| > \lambda_p\sqrt{1-a^2}$, and $|a\tilde{y}_2 - \tilde{y}_1| > \lambda_p\sqrt{1-a^2}$. Correspondingly, it follows that $A_2 < 0$, $A_2 < A_{1a}$, and $A_2 < A_{1b}$, and the claim follows. □

6.12 Proof of Lemma 4.4

Write for short $\hat{\beta} = \hat{\beta}^{ss}$ and $\lambda_p = \lambda_p^{ss} = \sqrt{2q\log(p)}$. Introduce events $A_{0j} = \{\beta_{j-2} = \beta_{j-1} = \beta_j = \beta_{j+1} = 0\}$, $A_{1j} = \{\beta_{j-2} = \beta_{j-1} = \beta_{j+1} = 0, \beta_j = \tau_p\}$, $A_{2j} = \{\beta_{j-2} = \beta_{j+1} = 0, \beta_{j-1} = \beta_j = \tau_p\}$, $B_{0j} = \{\hat{\beta}_{j-2} = \hat{\beta}_{j-1} = \hat{\beta}_j = \hat{\beta}_{j+1} = 0\}$, $B_{1j} = \{\hat{\beta}_{j-2} = \hat{\beta}_{j-1} = \hat{\beta}_{j+1} = 0, \hat{\beta}_j \ne 0\}$, and $B_{2j} = \{\hat{\beta}_{j=1} = \hat{\beta}_{j+1} = 0, \hat{\beta}_{j-1} \ne 0, \hat{\beta}_j \ne 0\}$. The Hamming distance is

$$\sum_{j=1}^{p} P(\mathrm{sgn}(\hat{\beta}_j) \ne \mathrm{sgn}(\beta_j)) \ge \sum_{j=3}^{p-1} P(\hat{\beta}_j \ne 0, \beta_j = 0) + P(\hat{\beta}_j = 0, \beta_j \ne 0) \ge \frac{1}{9}\sum_{j=3}^{p-2}(I_j + II_j + III_j),$$

where

$$I_j = \sum_{k=j-2}^{j+1} P(\hat{\beta}_k \neq 0, \beta_k = 0), \qquad II_j = P(\hat{\beta}_j = 0, \beta_j = \tau_p) + \sum_{k \in \{j-2, j-1, j+1\}} P(\hat{\beta}_k \neq 0, \beta_k = 0),$$

and

$$III_j = \sum_{k \in \{j-2, j+1\}} P(\hat{\beta}_k \neq 0, \beta_k = 0) + \sum_{k \in \{j-1, j\}} P(\hat{\beta}_k = 0, \beta_k = \tau_p).$$

By basic algebra and definitions,

$$I_j \geq \sum_{k=j-2}^{j+1} P(\hat{\beta}_k \neq 0, A_{0j}) \geq P(A_{0j} \cap B_{0j}^c),$$

$$II_j \geq P(\hat{\beta}_j = 0, A_{1j}) + \sum_{k \in \{j-2, j-1, j+1\}} P(\hat{\beta}_k \neq 0, A_{1j}) \geq P(A_{1j} \cap B_{1j}^c),$$

and

$$III_j \geq \sum_{k \in \{j-2, j+1\}} P(\hat{\beta}_k \neq 0, A_{2j}) + \sum_{k \in \{j-1, j\}} P(\hat{\beta}_k = 0, A_{2j}) \geq P(A_{2j} \cap B_{2j}^c).$$

It follows that

$$\sum_{j=1}^{p} P(\mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j)) \geq \frac{1}{9} \sum_{j=3}^{p-1} [P(A_{0j} \cap B_{0j}^c) + P(A_{1j} \cap B_{1j}^c) + P(A_{2j} \cap B_{2j}^c)]. \quad (6.107)$$

Introduce the events

$$D_{0j} = \{|\tilde{Y}_j| > \lambda_p\}, \quad D_{1j} = \{|\tilde{Y}_j| < \lambda_p\}, \quad H_j = \{\tilde{Y}_{j-1}^2 + \tilde{Y}_j^2 - 2a\tilde{Y}_{j-1}\tilde{Y}_j < 2\lambda_p^2(1-a^2)\},$$

and

$$D_{2j} = H_j \cup \{|a\tilde{Y}_{j-1} - \tilde{Y}_j| < \lambda_p\sqrt{1-a^2}, |\tilde{Y}_{j-1}| > \lambda_p\} \cup \{|a\tilde{Y}_j - \tilde{Y}_{j-1}| < \lambda_p\sqrt{1-a^2}, |\tilde{Y}_j| > \lambda_p\}.$$

We now show that

$$B_{0j}^c \supseteq D_{0j}, \qquad B_{1j}^c \supseteq D_{1j}, \qquad B_{2j}^c \supseteq D_{2j}, \quad (6.108)$$

or equivalently, that

$$B_{0j} \cap D_{0j} = \emptyset, \qquad B_{1j} \cap D_{1j} = \emptyset, \qquad B_{2j} \cap D_{2j} = \emptyset.$$

Consider the first claim. Recall that Ω is a tridiagonal matrix. When B_{0j} or B_{1j} happens, $\hat{\beta}_{j-2} = \hat{\beta}_{j-1} = \hat{\beta}_{j+1} = 0$, and $\hat{\beta}_j$ minimizes the functional

$$\frac{1}{2}u^2 - u\tilde{Y}_j + \frac{\lambda_p^2}{2} 1_{\{u \neq 0\}}.$$

Elementary calculus shows that the minimum is achieved at $u = 0$ if and only if $|\tilde{Y}_j| < \lambda_p$. Therefore, when B_{0j} happens, $\hat{\beta}_j = 0$, the minimum is achieved at $u = 0$. Therefore, $|\tilde{Y}_j| \leq \lambda_p$, and the claim follows. Consider the second claim. Similarly, when B_{1j} happens, $\hat{\beta}_j \neq 0$ and $|\tilde{Y}_j| \geq \lambda_p$, and the claim follows. Consider the third claim. Let $W_j = (\hat{\beta}_{j-1}, \hat{\beta}_j)'$ and

u be a two-dimensional vector. Similarly, when B_{2j} happens, W_j minimizes the following functional

$$\frac{1}{2}u' \begin{pmatrix} 1 & a \\ a & 1 \end{pmatrix} u - u' \begin{pmatrix} \tilde{Y}_{j-1} \\ \tilde{Y}_j \end{pmatrix} + \frac{\lambda_p^2}{2}\|u\|_0.$$

By Lemma 4.3, both coordinates of the minimizing vector u are nonzero if and only if

$$(\tilde{Y}_{j-1}, \tilde{Y}_j) \in \{|a\tilde{Y}_{j-1} - \tilde{Y}_j| > \lambda_p\sqrt{1-a^2}, \quad |a\tilde{Y}_j - \tilde{Y}_{j-1}| > \lambda_p\sqrt{1-a^2}, \quad \tilde{Y}_{j-1}^2 + \tilde{Y}_j^2 - 2a\tilde{Y}_{j-1}\tilde{Y}_j > 2\lambda_p^2(1-a^2)\}.$$

When B_{2j} happens, both coordinates of W_j are nonzero. This implies that $(\tilde{Y}_{j-1}, \tilde{Y}_j) \in D_{2j}^c$, and the claim follows.

Now, combining (6.107) into (6.108) gives

$$\sum_{j=1}^{p} P(\mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j)) \geq \frac{1}{3}\sum_{j=3}^{p-1}[P(A_{0j} \cap D_{0j}) + P(A_{1j} \cap D_{1j}) + P(A_{2j} \cap D_{2j})]. \quad (6.109)$$

By definitions, $P(A_{0j}) = (1-\epsilon_p)^4$, $P(A_{1j}) = (1-\epsilon_p)^3\epsilon_p$, that conditional on A_{0j}, $\tilde{Y}_j \sim N(0,1)$, and that conditional on A_{1j}, $\tilde{Y}_j \sim N(\tau_p, 1)$. It follows from elementary statistics and definition that

$$P(A_{0j} \cap D_{0j}) = (1-\epsilon_p)^4 P(N(0,1) \geq \lambda_p) = L_p p^{-q}, \quad (6.110)$$

and that

$$P(A_{1j} \cap D_{1j}) = (1-\epsilon_p)^3\epsilon_p P(N(\tau_p, 1) \leq \lambda_p) = \begin{cases} L_p p^{-[\vartheta + (\sqrt{q} - \sqrt{r})^2]}, & q < r, \\ p^{-\vartheta}(1 + o(1)), & q > r. \end{cases} \quad (6.111)$$

Furthermore, we have that $P(A_{2j}) = (1-\epsilon_p)^2\epsilon_p^2$ and that conditional on A_{2j}, $(\tilde{Y}_{j-1}, \tilde{Y}_j)$ is distributed as a bivariate normal with equal means $(1+a)\tau_p$, unit variances and correlation a. Let R denote the region in the two-dimensional Euclidean space

$$\begin{aligned} R \;=\; & \{(x,y) : |ay - x| < \lambda_p\sqrt{1-a^2} \text{ and } |y| > \lambda_p\} \\ & \cup\{(x,y) : |ax - y| < \lambda_p\sqrt{1-a^2} \text{ and } |x| > \lambda_p\} \\ & \cup\{(x,y) : x^2 + y^2 - 2axy < 2\lambda_p^2(1-a^2)\}. \end{aligned}$$

By direct calculations,

$$P(A_{2j} \cap D_{2j}) = (1-\epsilon_p)^2\epsilon_p^2 P((\tilde{Y}_{j-1}, \tilde{Y}_j) \in R) = L_p p^{-2\vartheta - \min\{[(\sqrt{r(1-a^2)} - \sqrt{q})^+]^2, 2[(\sqrt{r(1+a)} - \sqrt{q})^+]^2\}}. \quad (6.112)$$

Inserting (6.110)-(6.112) into (6.109) gives the claim. □

References

[1] Abramovich, F., Benjamini, Y., Donoho, D. and Johnstone, I. (2008). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.*, **34**(2), 584–653.

[2] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**(6), 716–723.

[3] 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.

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

[5] Cai, T. and Jin, J. (2010). Optimal procedure for large-scale multiple testing in the presence of dependence. *Unpublished Manuscript.*

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

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

[8] Dinur, I. and Nissim, K. (2003). Revealing information while preserving privacy. *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 202–210, ACM Press.

[9] Diestel, R. (2005). *Graph Theory*, 2nd edition. Springer, NY.

[10] Donoho, D. (2006). 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.

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

[12] Donoho, D. and Elad, M. (2003). Optimally sparse representation in general (non-orthogonal) dictionaries via ℓ^1 minimization. *Proc. Natl. Acad. Sci.*, **100**(5), 2197–2202.

[13] Donoho, D. L. and Huo, X. (2001). Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inform. Theory*, **47**, 2845–2862.

[14] Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, **32**, 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**, 14790–14795.

[16] Donoho, D. and Tanner, J. (2005). Sparse nonnegative solution of underdetermined linear equations by linear programming. *Proc. Natl. Acad. Sci.*, **102**(27), 9446–9451.

[17] Fan, J., Jin, J. and Zheng, K. (2010). Optimal procedure for variable selection in the presence of strong dependence. *Unpublished Manuscript.*

[18] Fan, J. and Lv, J. (2008). Sure independence screening in generalized linear models with NP-dimensionality. *J. Roy. Soc. Statist. B*, **70**, 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. 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`

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

[22] Genovese, C., Jin, J. and Wasserman, L. (2010). Revisiting marginal regression. *Unpublished Manuscript.*

[23] Hall, P. and Jin, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *Ann. Statist.*, **38**(3), 1686–1732.

[24] Mathematisches Forschungsinstitut Oberwolfach Workshop on "Modern non-parametric statistics: going beyond asymptotic minimaxity". `http://www.mfo.de/programme/schedule/2010/13/OWR_2010_16.pdf`

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

[26] Ji, P. (2010). *Selected Topics in Nonparametric Testing for Spectral Densities and Variable Selection for High Dimensional Data.* Ph.D Thesis, Department of Statistical Science, Cornell University.

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

[28] Shorack, G. R. and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics.* John Wiley and Sons, Inc.

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

[30] Vershynin, R. (2010). *Introduction to the non-asymptotic analysis of random matrices.* Lecture notes, Department of Mathematics, University of Michigan.

[31] Wainwright, M. (2006). Sharp threshold for high-dimensional and noisy recovery of sparsity. Technical report, Department of Statistics, University of Berkeley.

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

[33] Wasserman, L. (2005). *All of Nonparametric Statistics.* Springer, NY.

[34] Ye, F. and Zhang, C.-H. (2009). Rate minimaxity of the lasso and Dantzig estimators. Technical Report, Department of Statistics and Biostatistics, Rutgers University.

[35] Zhao, P. and Yu, B. (2006). On model selection consistency of LASSO. *J. Machine Learning Research*, **7**, 2541–2567.

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