SELECTED TOPICS IN NONPARAMETRIC TESTING AND VARIABLE SELECTION FOR HIGH DIMENSIONAL DATA

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Pengsheng Ji

August 2012

SELECTED TOPICS IN NONPARAMETRIC TESTING AND VARIABLE SELECTION FOR HIGH DIMENSIONAL DATA

Pengsheng Ji, Ph.D.

Cornell University 2012

Part I:

The Gaussian white noise model has been used as a general framework for nonparametric problems. The asymptotic equivalence of this model to density estimation and nonparametric regression has been established by Nussbaum (1996), Brown and Low (1996).

In Chapter 1, we consider testing for presence of a signal in Gaussian white noise with intensity $n^{-1/2}$, when the alternatives are given by smoothness ellipsoids with an L_2-ball of radius ρ removed. It is known that, for a fixed Sobolev type ellipsoid $\Sigma(\beta, M)$ of smoothness β and size M, the radius rate $\rho \asymp n^{-4\beta/(4\beta+1)}$ is the critical separation rate, in the sense that the minimax error of second kind over α-tests stays asymptotically between 0 and 1 strictly (Ingster, 1982). In addition, Ermakov (1990) found the sharp asymptotics of the minimax error of second kind at the separation rate. For adaptation over both β and M in that context, it is known that a log log-penalty over the separation rate for ρ is necessary for a nonzero asymptotic power. Here, following an example in nonparametric estimation related to the Pinsker constant, we investigate the adaptation problem over the ellipsoid size M only, for fixed smoothness degree β. It is established that the Ermakov type sharp asymptotics can be preserved in that adaptive setting, if $\rho \to 0$ slower than the separation rate. The penalty for adapation in that setting turns out to be a sequence tending to infinity arbitrarily

slowly.

In Chapter 2, motivated by the sharp asymptotics of nonparametric estimation for non-Gaussian regression (Golubev and Nussbaum, 1990), we extend Ermakov's sharp asymptotics for the minimax testing errors to the nonparametric regression model with nonnormal errors. The paper entitled "Sharp Asymptotics for Risk Bounds in Nonparametric Testing with Uncertainty in Error Distributions" is in preparation.

This part is joint work with Michael Nussbaum.

Part II:

Consider a linear model $Y = X\beta + z$, $z \sim N(0, I_n)$. Here, $X = X_{n,p}$, where both p and n are large but $p > n$. We model the rows of X as *iid* samples from $N(0, \frac{1}{n}\Omega)$, where Ω is a $p \times p$ correlation matrix, which is unknown to us but is presumably sparse. The vector β is also unknown but has relatively few nonzero coordinates, and we are interested in identifying these nonzeros.

We propose the **U**nivariate **P**enalization **S**creeing (UPS) for variable selection. This is a Screen and Clean method where we screen with Univariate thresholding, and clean with Penalized MLE. It has two important properties: Sure Screening and Separable After Screening. These properties enable us to reduce the original regression problem to many small-size regression problems that can be fitted separately. The UPS is effective both in theory and in computation.

We measure the performance of a procedure by the Hamming distance, and use an asymptotic framework where $p \to \infty$ and other quantities (e.g., n, sparsity level and strength of signals) are linked to p by fixed parameters. We find that in many cases, the UPS achieves the optimal rate of convergence. Al-

so, for many different Ω, there is a common three-phase diagram in the two-dimensional phase space quantifying the signal sparsity and signal strength. In the first phase, it is possible to recover all signals. In the second phase, it is possible to recover most of the signals, but not all of them. In the third phase, successful variable selection is impossible. 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 are well-known approaches to variable selection. However, somewhat surprisingly, there are regions in the phase space where neither of them is rate optimal, even in very simple settings such as Ω is tridiagonal, and when the tuning parameter is ideally set.

This part is joint work with Jiashun Jin, and has appeared in Annals of Statistics.

BIOGRAPHICAL SKETCH

Pengsheng Ji was born in 1981 in Shandong, China. He received B.S. in the Special Class of Mathematics and M.S. in statistics at Nankai University.

In 2006, Pengsheng came to Cornell to pursue his PhD degree in statistics.

Dedicated to my parents, Guangxia Ji and Xiuying Li.

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to my advisors, Prof. Michael Nussbaum and Prof. Jiashun Jin. Their guidance, encouragement and friendship have given me the greatest experience.

Thanks to Prof. Jim Booth and Prof. Martin Wells for being on my PhD committee. Their suggestions and advice have been a valuable contribution to my work.

Thanks to Professors Robert Strawderman, J.T. Gene Hwang, Harry Zhou, David Ruppert, John Bunge and Giles Hooker for their support and tremendous help during my years of PhD study.

Thanks to my friends and fellow students Zhigen Zhao, Yingxing Li, Michael Grabchak, Bret Hanlon, Vadim Zipunnikov, Haizhi Jeff Lin, Gongfu Zhou and others.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

SHARP ADAPTIVE NONPARAMETRIC HYPOTHESIS TESTING FOR SOBOLEV ELLIPSOIDS

1.1 Introduction and main result

Consider the Gaussian white noise model in sequence space, where observations are

$$Y_j = f_j + n^{-1/2}\xi_j, \qquad j = 1, 2, ..., \tag{1.1.1}$$

with unknown, nonrandom signal $f = (f_j)_{j=1}^{\infty}$, and noise variables ξ_j which are i.i.d. $N(0, 1)$. It can also be written in the form of the stochastic differential equation

$$dY(t) = f(t)dt + n^{-1/2}dW(t), \qquad t \in [0, 1],$$

where W is a standard Wiener process on $[0, 1]$, given an orthonormal basis. The asymptotic equivalence to nonparametric regression and density estimation has been established by Brown and Low (1996), and Nussbaum (1996).

We intend to test the null hypothesis of "no signal" against nonparametric alternatives described as follows. For some $\beta > 0$ and $M > 0$, let $\Sigma(\beta, M)$ be the set of sequences

$$\Sigma(\beta, M) = \{f = (f_j)_{j=1}^{\infty} : \sum_{j=1}^{\infty} j^{2\beta} f_j^2 \leq M\};$$

this might be called a Sobolev type ellipsoid with smoothness parameter β and size parameter M. Consider further the complement of an open ball in the sequence space l_2: if $\|f\|_2^2 = \sum_{j=1}^{\infty} f_j^2$ is the squared norm then

$$B_\rho = \{f \in l_2 : \|f\|_2^2 \geq \rho\}.$$

Here $\rho^{1/2}$ is the radius of the open ball; by an abuse of language we call ρ itself the "radius". We study the hypothesis testing problem

$$H_0 : f = 0 \quad \text{against} \quad H_a : f \in \Sigma(\beta, M) \cap B_\rho.$$

Assuming that $n \to \infty$, implying that the noise size $n^{-1/2}$ tends to zero, we expect that for a fixed radius ρ, consistent α-testing in that setting is possible. More precisely, there exist α-tests with type II error tending to zero uniformly over the nonparametric alternative $f \in \Sigma(\beta, M) \cap B_\rho$. If now the radius $\rho = \rho_n$ tends to zero as $n \to \infty$, the problem becomes more difficult and if $\rho_n \to 0$ too quickly, all α-tests will have the trivial asymptotic (worst case) power α. According to fundamental results of Ingster (1982, 1984), there is a critical rate for ρ_n, the so-called *separation rate*

$$\rho_n \asymp n^{-4\beta/(4\beta+1)} \tag{1.1.2}$$

at which the transition in the power behaviour occurs. More precisely, consider a (possibly randomized) α-test ϕ_n in the model (1.1.1) with respect to $H_0 : f = 0$, that is, a test fulfilling $E_{n,0}\phi_n \le \alpha$ where $E_{n,f}(\cdot)$ denotes expectation in the model (1.1.1). For given ϕ_n, we define the worst case type II error over the alternative $f \in \Sigma(\beta, M) \cap B_\rho$ as

$$\Psi(\phi_n, \rho, \beta, M) := \sup_{f \in \Sigma(\beta,M) \cap B_\rho} \left(1 - E_{n,f}\phi_n\right).$$

The search for a best α-test in this sense leads to the minimax type II error

$$\pi_n(\alpha, \rho, \beta, M) := \inf_{\phi_n : E_{n,0}\phi_n \le \alpha} \Psi(\phi_n, \rho, \beta, M).$$

An α-test which attains the inf above for a given n is minimax with respect to type II error. Ingster's separation rate result can now be formulated as follows: if $\rho_n \asymp n^{-4\beta/(4\beta+1)}$ and $0 < \alpha < 1$ then

$$0 < \underline{\lim_n}\, \pi_n(\alpha, \rho_n, \beta, M) \text{ and } \overline{\lim_n}\, \pi_n(\alpha, \rho_n, \beta, M) < 1 - \alpha.$$

Moreover, if $\rho_n \gg n^{-4\beta/(4\beta+1)}$ then $\pi_n(\alpha, \rho_n, \beta, M) \to 0$, and if $\rho_n \ll n^{-4\beta/(4\beta+1)}$ then $\pi_n(\alpha, \rho_n, \beta, M) \to 1 - \alpha$.

These minimax rates in nonparametric testing, presented here in the simplest case of an l_2-setting, have been extended in two ways. Firstly, Ermakov (1990) found the exact asymptotics of the minimax type II error $\pi_n(\alpha, \rho, \beta, M)$ (equivalently, of the maximin power) at the separation rate. The shape of that result and its derivation from an underlying Bayes-minimax theorem on ellipsoids exhibit an analogy to the Pinsker constant in nonparametric estimation. Secondly, Spokoiny (1996) considered the adaptive version of the minimax nonparametric testing problem, where both β and M are unknown, and showed that the rate at which $\rho_n \to 0$ has to be slowed by a $\log\log n$-factor if nontrivial asymptotic power is to be achieved. Thus an "adaptive minimax rate" was specified, analogous to Ingster's nonadaptive separation rate (1.1.2), where the additional $\log\log n$-factor is interpreted as a penalty for adaptation. However a corresponding sharp adaptive type II error asymptotics in the sense of Ermakov (1990) has not been obtained.

It is noteworthy that in nonparametric estimation over $f \in \Sigma(\beta, M)$ with l_2-loss (as opposed to testing), where the risk asymptotics is given by the Pinsker constant, there is a multitude of results showing that adaptation is possible with neither a penalty in the rate nor in the constant, cf. Efromovich and Pinsker (1984), Golubev (1987, 1992), Tsybakov (2009). The present paper deals with the question of whether the sharp risk asymptotics for testing in the sense of Ermakov (1990) can be reproduced in an adaptive setting, in the context of a possible rate penalty for adaptation.

Let us first present the well known results on sharp risk asymptotics for

testing in the nonadaptive setting. Let Φ be the distribution function of the standard normal, and for $\alpha \in (0,1)$ let z_α be the upper α-quantile, such that $\Phi(z_\alpha) = 1-\alpha$. Write $a_n \gg b_n$ (or $b_n \ll a_n$) iff $b_n = o(a_n)$, and $a_n \sim b_n$ iff $\lim_n a_n/b_n = 1$.

Proposition 1. *(Ermakov, 1990) Suppose $\alpha \in (0,1)$,and that the radius ρ_n tends to zero at the separation rate, more precisely*

$$\rho_n \sim c \cdot n^{-4\beta/(4\beta+1)},$$

for some constant $c > 0$.

(i) For any sequence of tests ϕ_n satisfying $E_{n,0}\phi_n \le \alpha + o(1)$, we have the following lower bound

$$\Psi(\phi_n, \rho_n, \beta, M) \ge \Phi(z_\alpha - \sqrt{A(c,\beta,M)/2}) + o(1) \text{ as } n \to \infty,$$

where

$$A(c,\beta,M) = A_0(\beta) M^{-1/(2\beta)} c^{2+1/(2\beta)}$$

and $A_0(\beta)$ is Ermakov's constant

$$A_0(\beta) = \frac{2(2\beta+1)}{(4\beta+1)^{1+1/(2\beta)}}. \tag{1.1.3}$$

(ii) For given β and $M > 0$ there exists a sequence of tests ϕ_n satisfying $E_{n,0}\phi_n \le \alpha + o(1)$ such that

$$\Psi(\phi_n, \rho_n, \beta, M) \le \Phi(z_\alpha - \sqrt{A(c,\beta,M)/2}) + o(1).$$

This gives the sharp asymptotics for the minimax type II error at the separation rate, analogous to the Pinsker constant for nonparametric estimation. The optimal test attaining the bound of (ii) above, as given by Ermakov (1990), depends on β and M. As regards adaptivity in both of these unknown parameters, a test can not depend on them and the following result is known.

Proposition 2. *(Spokoiny, 1996). Let $\mathcal{T}$ be a subset of $(0,\infty)\times(0,\infty)$ such that there exist $M>0, \beta_2>\beta_1>0$ and*

$$\mathcal{T} \supseteq \{(\beta, M) : \beta_1 \le \beta \le \beta_2\}.$$

(i) If $t_n \ll (\log\log n)^{1/2}$ and $\rho_n \sim c\cdot(n/t_n)^{-4\beta/(4\beta+1)}$, then for any constant $c>0$ and any adaptive test ϕ_n satisfying $E_{n,0}\phi_n \le \alpha + o(1)$, we have

$$\sup_{(\beta,M)\in\mathcal{T}} \Psi(\phi_n, \rho_n, \beta, M) \ge 1-\alpha+o(1).$$

(ii) For any $\beta^*>1/2$ and $0<M_1\le M_2$, let

$$\mathcal{T} = \{(\beta, M) : 1/2 < \beta \le \beta^*, M_1 \le M \le M_2\}.$$

Then there exist a constant $c_1 = c_1(\beta^*, M_1, M_2)$ and an adaptive test ϕ_n satisfying $E_{n,0}\phi_n = o(1)$, such that, if

$$\rho_n \sim c_1\left(\frac{n}{(\log\log n)^{1/2}}\right)^{-4\beta/(4\beta+1)} \tag{1.1.4}$$

then

$$\sup_{(\beta,M)\in\mathcal{T}} \Psi(\phi_n, \rho_n, \beta, M) = o(1). \tag{1.1.5}$$

Here the criterion to evaluate a test sequence has changed, to include the worst case type II error over a whole range of β, M. Hence the critical radius rate (1.1.4) has to be interpreted as an *adaptive separation rate.* It differs by a factor $(\log\log n)^{2\beta/(4\beta+1)}$ from the nonadaptive separation rate (1.1.2); this factor is an example of the well-known phenomenon of a penalty for adaptation. Furthermore, as noted in Spokoiny (1996), a degenerate behaviour occurs here, in that both error probabilities at the critical rate tend to zero. Thus any sequence ϕ_n of tests fulfilling (1.1.5) should be seen as *adaptive rate optimal*,

comparable to rate optimal tests in the nonadaptive case (that is, tests fulfilling $\overline{\lim}_n \Psi(\phi_n, \rho_n, \beta, M) < 1 - \alpha$ at ρ_n given by (1.1.2)). In Ingster and Suslina (2003), chap. 7, the worst case adaptive error (1.1.5) is further analyzed, with a view to a sharp asymptotics, but the results are not conclusive with regard discriminating between different adaptive rate optimal sequences of tests.

In this paper we address the question of whether an exact type II error asymptotics in the sense of Ermakov(1990) is possible in an adaptive setting. In our approach β is kept fixed and known, while we aim for adaptation over the ellipsoid size M. First, we present a negative result for adaptation at Ingster's separation rate.

Theorem 1. *Suppose $c > 0$, $0 < M_1 < M_2 < \infty$ and $\rho_n \sim c \cdot n^{-4\beta/(4\beta+1)}$. Then there is no adaptive test ϕ_n satisfying $E_{n,0}\phi_n \le \alpha + o(1)$, such that*

$$\Psi_n(\phi_n, \rho_n, \beta, M_i) \le \Phi(z_\alpha - \sqrt{A(c, \beta, M_i)/2}) + o(1),$$

for $i = 1, 2$.

This result states sharp adaptation even just for M at the separation rate is impossible, and the adaptation for even just M is not trivial as some might think. Instead, we enlarge the radius slightly and examine how the minimax error approaches zero. To be specific, we replace the constant c in $\rho_n \sim c \cdot n^{-4\beta/(4\beta+1)}$ by a sequence c_n tending to infinity slowly. In that case the minimax type II error bound of Proposition 1, namely $\Phi(z_\alpha - \sqrt{A(c, \beta, M)/2})$ will tend to zero. To this error probability we apply a log-asymptotics as in moderate and large deviation theory and show that in this sense, adaptation to Ermakov´s constant is possible.

Theorem 2. *Assume $c_n \to \infty$ and $c_n = o(n^K)$ for every constant $K > 0$. If $\rho_n =$*

$c_n \cdot n^{-4\beta/(4\beta+1)}$, *there exists a test* ϕ_n *not depending on* M *such that*

$$E_{n,0}\phi_n \le \alpha + o(1),$$

and for all $M > 0$,

$$\overline{\lim_n} \frac{1}{c_n^{2+1/(2\beta)}} \log \Psi(\phi_n, \rho_n, \beta, M) \le -\frac{A_0(\beta) M^{-1/(2\beta)}}{4}$$

However now, since the optimality criterion has been changed, a formal argument is needed that no α-test can be better in the sense of the log-asymptotics for the error of second kind. Such a result is implied by Theorem 3 in Ermakov (2008), where the nondaptive sharp asymptotics was studied in a setting where $\rho_n = c_n \cdot n^{-4\beta/(4\beta+1)}$ with $c_n \to \infty$, hence error probabilities tending to zero. Since the nonadaptive minimax lower risk bound for fixed c is based on a Gaussian limit argument, the case of $c_n \to \infty$ (sufficiently slowly) should be treated with the methodology of moderate deviations.

Theorem 3. *Under the same assumptions as in the last theorem, if* $\rho_n = c_n \cdot n^{-4\beta/(4\beta+1)}$, *then for any test* ϕ_n *(possibly depending on* M*) satisfying* $E_{n,0}\phi_n \le \alpha + o(1)$, *we have*

$$\underline{\lim_n} \frac{1}{c_n^{2+1/(2\beta)}} \log \Psi(\phi_n, \rho_n, \beta, M) \ge -\frac{A_0(\beta) M^{-1/(2\beta)}}{4}.$$

This result is implied by Theorem 3 in Ermakov (2008), and hence the proof is omitted.

We have a few remarks to address the relation to the literature.

1.) Ermakov (2008) also shows that, for nonadaptive testing, these asymptotics of moderate deviation probabilities are valid in a sharper sense, i.e., there

are tests ϕ_n depending on M such that

$$\Psi(\phi_n, \rho_n, \beta, M) = (1 + o(1)) \cdot \Phi\left(z_\alpha - \sqrt{A_0 M^{-1/(2\beta)} c_n^{2+1/(2\beta)}/2}\right).$$

This is in the same spirit of the Cramer type of moderate deviation, as discussed for the central limit theorem by Chen et al. (2011), chap. 11. The question whether this sharper asymptotics can be replicated in the adaptive setting is open.

2.) Ingster(1993a, 1993b, 1993c, 1998) discussed the "detection" problem for ℓ^p ball using the sum of the type I error and the maximum type II error. The conditions for which the sum of is bounded away from 0 and 1 are given. But no sharp or adaptive asymptotics is obtained.

3.) Golubev(1987) studied the adaptive estimation with β fixed, but incorporates local aspects (twofold, both local with respect to $x \in [0, 1]$ and smoothness class) to construct the optimal test. In the current paper, we do not use the local aspects.

4.) In the literature, the testing problem considered here has sometimes been connected with the estimation problem of the quadratic functional $Q(f) = \sum_{i=1}^{\infty} f_j^2$, Ibragimov and Hasminskii(1980), Bickel and Ritov(1988) found that when the unknown function is sufficiently smooth, quadratic functionals can be estimated with parametric $\sqrt{n}$ rate, otherwise the rate is slower. More precisely, the minimax rate with the parameter space $\Sigma(\beta, M)$ is n^{-r} with an exponent $r = \frac{4\beta}{4\beta+1} < \frac{1}{2}$ when $0 < \beta < 1/4$, but when $\beta \geq 1/4$, the minimax rate becomes $n^{-1/2}$. Efromovich(1994) showed that at the point $\beta = 1/4$ the optimal adaptive rate is $n^{-1/2} c_n$ where $c_n \to \infty$ slower than any power function of n, and for $\beta > 1/4$, the optimal adaptive rate is $n^{-1/2}$. Efromovich and Low(1996) showed that, in the case $\beta < 1/4$, the optimal adaptive rate is $(n \log \sqrt{n})^{-r}$, which is larger

than the nonadaptive rate by a logarithmic factor. See Tsybakov (1998) for more discussion of the adaptive rates and of the boundary effects. Klemela (2006) found sharp adaptive estimators for the irregular case $\beta < 1/4$. That is, first, the constant $K_{\beta,M}$ is found such that

$$\lim_{n\to\infty} \inf_{\hat{Q}} \sup_{(\beta,M)\in B} \left(K_{\beta,M}(n \log \sqrt{n})^{-r}\right)^{-p} \sup_{f\in\Sigma(\beta,M)} |\hat{Q} - Q(f)|^p = 1,$$

where $B = [\beta_1, \beta_2] \times [M_1, M_2]$, $\beta_2 < 1/4$ and $p \geq 1$. Second, the estimators which do not depend on (β, M) and achieve the infimum are obtained.

The rest of the paper is organized as follows. In Section 2, we show that the minimax quadratic test is asymptotically minimax over all tests, and provide the idea of the proofs in Sections 3, 4 and 5. Finally, in the appendix, we recap the ideas for adaptive estimation of Golubev (1992) to help the readers to understand the idea of adaptive testing.

1.2 The Bayes-minimax problem for nonparametric testing

The purpose of this expository section is to elucidate the analogy between the Pinsker constant for L_2-estimation over ellipsoids and the constant found by Ermakov (1990) for nonparametric testing over ellipsoids with an L_2-ball removed. We draw on the backgound explanation given in Ingster and Suslina (2003), sec. 4.1, but we focus specifically on the fact that very similar Bayes-minimax problems are at the root of the estimation and testing variants. For the theory underlying the Pinsker constant cf. Belitser and Levit (1995), Nussbaum (1999), Tsybakov (2009).

For this exposition, we shall assume that observations (1.1.1) are for $j =$

$1, \ldots, n$; we will thus assume $f \in \mathbb{R}^n$ and understand the sets $\Sigma(\beta, M)$ and B_ρ accordingly, i.e. they refer only to the first n coefficents of f. By $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$ we denote euclidean norm and inner product in $\mathbb{R}^n$. Since most expressions will depend on n, for this discussion we shall often suppress dependence on n in the notation. Assume that the radius ρ tends to zero at the critical rate, that is $\rho \asymp n^{-4\beta/(4\beta+1)}$. Let $\mathbb{R}^n_+ = [0, \infty)^n$; for a certain $d \in \mathbb{R}^n_+$, consider a quadratic statistic of the form $\tilde{T} = n \sum_{j=1}^n d_j Y_j^2$. Under H_0, we have $E_{n,0}\tilde{T} = \sum_{j=1}^n d_j$ and $\mathrm{Var}_{0,n}\tilde{T} = 2\|d\|^2$. Since we will work with the normalized test statistic, obtained by centering and dividing by the standard deviation, it is obvious that we need only consider coefficients d fulfilling $\|d\|^2 = 1$. Accordingly define, for such coefficients d, the statistic

$$T = \frac{1}{\sqrt{2}} \left(\tilde{T} - \sum_{j=1}^n d_j \right). \tag{1.2.6}$$

Under H_0, we now have $E_0 T = 0$ and $\mathrm{Var}_0 T = 1$. We will consider quadratic tests

$$\psi_d = \mathbf{1}\{T > z_\alpha\}. \tag{1.2.7}$$

A further condition on d is imposed by requiring $d \in \mathcal{D}$, a set which is defined for a given sequence $\delta = (\log n)^{-1}$ as

$$\mathcal{D} = \{d \in \mathbb{R}^n_+ : \|d\|^2 = 1 \text{ and } \sup_j d_j^2 \leq \frac{\delta}{n\rho}\}. \tag{1.2.8}$$

For any test, we are interested in the worst case type II error under the constraint $f \in \Sigma(\beta, M) \cap B_\rho$. A monotonicity argument shows that for every ψ_d, this is attained when $\|f\|^2$ is minimal, i.e. at $\|f\|^2 = \rho$. It follows that for quadratic tests ψ_d, we may replace the restriction $f \in B_\rho$ by $f \in B'_\rho$ where

$$B'_\rho = \{f \in \mathbb{R}^n : \rho \leq \|f\|^2 \leq 2\rho\}.$$

For $f \in \mathbb{R}^n$ we set $f^2 := \left(f_j^2\right)_{j=1}^n$. For $d \in \mathcal{D}$ and $g \in \mathbb{R}^n_+$ define the functional

$$L(d, g) = \frac{n}{\sqrt{2}} \langle d, g \rangle.$$

Lemma 1.2.1. *(a) Under H_0, we have $T \rightsquigarrow N(0,1)$ uniformly over $d \in \mathcal{D}$.*
(b) The statistic T given by (1.2.6) fulfills

$$T - L(d, f^2) \rightsquigarrow N(0,1)$$

uniformly over $d \in \mathcal{D}$ and $f \in B'_\rho$.
(c) Suppose f is random such that $f_j \sim N\left(0, \sigma_j^2\right)$ for a certain $\sigma \in \mathbb{R}^n$. Then the statistic T given by (1.2.6) fulfills

$$T - L(d, \sigma^2) \rightsquigarrow N(0,1)$$

uniformly over $d \in \mathcal{D}$ and $\sigma \in B'_\rho$.

Denote the expectation under the model of (c) by E^*_σ. The lemma implies that for uniformly over $d \in \mathcal{D}$ and $f \in \{0\} \cup \left(\Sigma(\beta, M) \cap B'_\rho\right)$

$$E_f(1 - \psi_d) = \Phi(z_\alpha - L(d, f^2)) + o(1) \tag{1.2.9}$$

$$= E^*_f(1 - \psi_d) + o(1). \tag{1.2.10}$$

In particular, all quadratic tests ψ_d with $d \in \mathcal{D}$ are aymptotic α-tests under $H_0 : f = 0$. To characterize the worst case error under the alternative $H_a : f \in \Sigma(\beta, M) \cap B_\rho$, we use (1.2.9) and the strict monotonicity of Φ and look for a saddlepoint of the functional $L(d, f^2)$.

Lemma 1.2.2. *For n large enough, there exists a saddlepoint $d_0 \in \mathcal{D}$, $f_0 \in \Sigma(\beta, M) \cap B'_\rho$ of the functional $L(d, f^2)$ such that*

$$L(d, f_0^2) \le L(d_0, f_0^2) \le L(d_0, f^2)$$

for all $d \in \mathcal{D}$ and all $f \in \Sigma(\beta, M) \cap B'_\rho$.

The normal distribution on the signal f postulated in (c) will be interpreted as a prior distribution. The next result shows that the Bayesian tests in this context are quadratic tests ψ_d, and in particular, if the σ^2 is taken at the saddlepoint ($\sigma_0^2 = f_0^2$) then $d \in \mathcal{D}$, i.e. it fulfills the infinitesimality condition $d_j^2 \leq \delta/n\rho$.

Lemma 1.2.3. *(a) For any $\sigma^2 \in \mathbb{R}_+^n$, the Neyman-Pearson α-test for simple hypotheses*

$$H_0 : \ Y_j \sim N(0, n^{-1}), \ j = 1, \ldots, n$$
$$H_a^* : \ Y_j \sim N(0, \sigma_j^2 + n^{-1}), \ j = 1, \ldots, n$$

is equivalent to a quadratic test pof form $\psi_d = \mathbf{1}\{T > t\}$ where $T = \sum_{j=1}^n d_j Y_j^2$, $d \in \mathbb{R}_+^n$, $\|d\| = 1$.

(b) If $\sigma^2 = f_0^2$ then the pertaining d is in $\mathcal{D}$ for n large enough, and $t \to z_\alpha$.

Part (b) implies that

$$\inf_{\phi : E_0\phi \leq \alpha} E_{f_0}^*(1 - \phi) = \inf_{d \in \mathcal{D}} E_{f_0}^*(1 - \psi_d) + o(1). \tag{1.2.11}$$

We are now ready to present the essence of the argument underlying the result of Ermakov (1990). Recall that $\pi_n(\alpha, \rho, \beta, M)$ denotes the minimax type II error over all α-tests. Denote the value of $L(d, f^2)$ at the saddlepoint

$$L_0 := L(d_0, f_0^2) = \sup_{d \in \mathcal{D}} \inf_{f \in \Sigma(\beta, M) \cap B'_\rho} L_n(d, f^2) = \inf_{f \in \Sigma(\beta, M) \cap B'_\rho} \sup_{d \in \mathcal{D}} L_n(d, f^2). \tag{1.2.12}$$

We begin with an $\alpha' > \alpha$ such that asymptotic α-tests are α'-tests for n large

enough. Then

$$\pi_n(\alpha', \rho, \beta, M) = \inf_{\phi: E_0\phi \le \alpha'} \sup_{f \in \Sigma(\beta,M)\cap B_\rho} E_f\,(1-\phi) \le \inf_{d\in\mathcal{D}} \sup_{f\in\Sigma(\beta,M)\cap B_\rho} E_f(1-\psi_d) \qquad (1.2.13)$$

$$= \inf_{d\in\mathcal{D}} \sup_{f\in\Sigma(\beta,M)\cap B'_\rho} E_f(1-\psi_d)$$

$$= \inf_{d\in\mathcal{D}} \sup_{f\in\Sigma(\beta,M)\cap B'_\rho} \Phi(z_\alpha - L_n(d, f^2)) + o(1) \text{ [relation (1.2.9)]}$$

$$= \Phi(z_\alpha - L_n(d_0, f_0^2)) + o(1) \text{ [monotonicity of } \Phi \text{ and (1.2.12)]}$$

$$= \inf_{d\in\mathcal{D}} E^*_{f_0}(1-\psi_d) + o(1) \text{ [relation (1.2.10)]}$$

$$= \inf_{\phi: E_0\phi\le\alpha} E^*_{f_0}(1-\phi) + o(1) \text{ [relation (1.2.11)]}.$$

The main term of the last expression is the Bayes risk for a prior distribution $f_j \sim N(0, f_{0j}^2)$ in the original model $Y_j \sim N\left(f_j, n^{-1}\right)$. Since $f_0 \in \Sigma(\beta, M) \cap B'_\rho$ and is extremal there, it fulfills

$$\sum_{j=1}^n f_{0j}^2 j^{2\beta} = M, \ \sum_{j=1}^n f_{0j}^2 = \rho$$

(see the precise description of the saddlepoint (d_0, f_0) in Lemma 1.5.1 below). It can therefore be shown that (as in the original Pinsker [1980] result) that this prior distribution asymptotically concentrates on every set of the form $\Sigma(\beta, M(1+\varepsilon)) \cap B'_{\rho(1-\varepsilon)}$ for $\varepsilon > 0$. A standard reasoning by truncation shows that in this case, for a certain probability measure G strictly concentrated on $\Sigma(\beta, M(1+\varepsilon)) \cap B'_{\rho(1-\varepsilon)}$

$$\inf_{\phi: E_0\phi\le\alpha} E^*_{f_0}(1-\phi) \le \inf_{\phi: E_0\phi\le\alpha} \int E_f(1-\phi) dG(f) + o(1).$$

However, by the relation between Bayes and minimax risk

$$\inf_{\phi: E_0\phi\le\alpha} \int E_f(1-\phi) dG(f) \le \pi_n(\alpha, \rho(1-\varepsilon), \beta, M(1+\varepsilon)). \qquad (1.2.14)$$

Summarizing (1.2.13)-(1.2.14) we have obtained for every $\varepsilon > 0$

$$\pi_n(\alpha(1+\varepsilon), \rho, \beta, M) \le \Phi(z_\alpha - L_n(d_0, f_0^2)) + o(1) \le \pi_n(\alpha, \rho(1-\varepsilon), \beta, M(1+\varepsilon)) + o(1)$$

Below in Lemma 8, it is shown that if $\rho = c \cdot n^{-4\beta/(4\beta+1)}$, where c is constant, then

$$L(d_0, f_0^2) \sim \sqrt{A_0 M^{-1/(2\beta)} c^{2+1/(2\beta)}/2}.$$

Since the right side is continuous in M and c , the result of Proposition 1 follows.

1.3 Proof of Theorem 1

For brevity we write $A_i = A(c, \beta, M_i), i = 1, 2$ in this section. Assume there exists a test ϕ_n not depending on i such that

$$E_{n,0}\phi_n \le \alpha + o(1), \tag{1.3.15}$$

$$\sup_{f \in \Sigma(\beta, M_i) \cap B_\rho} E_{n,f}(1 - \phi_n) \le \Phi(z_\alpha - \sqrt{A_i/2}) + o(1), \tag{1.3.16}$$

for $i = 1$ or 2. Let G_{n,M_i} be the Gaussian prior for f with $f_j \sim N(0, \sigma_j^{*2})$ independently, where

$$\sigma_j^{*2}(M_i) = (\lambda - \mu j^{2\beta})_+, \quad j = 1, 2, \ldots$$

and where λ and μ are determined by

$$\sum j^{2\beta} \sigma_j^{*2} = M_i \quad \text{and} \quad \sum \sigma_j^{*2} = \rho.$$

It can be shown that G_{n,M_i} asymptotically concentrates on $\Sigma(\beta, M_i)$. Then

$$\sup_{f \in \Sigma(\beta, M_i) \cap B_\rho} E_{n,f}(1 - \phi_n) \ge (1 + o(1)) \cdot \int E_{n,f}(1 - \phi_n)\, G_{n,M_i}(df). \tag{1.3.17}$$

p Recall $Y_j = f_j + n^{-1/2}\xi_j$. Let the joint distributions of $(Y_j)_0^\infty$ under the priors $G_{n,0}, G_{n,M_1}$ and G_{n,M_2} be $Q_{0,n}, Q_{1,n}$ and $Q_{2,n}$, respectively, i.e.,

$$Q_{0,n} : Y_j \sim N(0, n^{-1}), \ j = 1, 2, \ldots$$

$$Q_{1,n} : Y_j \sim N(0, n^{-1} + \sigma_j^{*2}(M_1)), \ j = 1, 2, \ldots$$

$$Q_{2,n} : Y_j \sim N(0, n^{-1} + \sigma_j^{*2}(M_2)), \ j = 1, 2, \ldots$$

Therefore,

$$E_{Q_{0,n}}\phi_n = E_{n,0}\phi_n,$$
$$E_{Q_{i,n}}(1-\phi_n) = \int E_{n,f}(1-\phi_n)\, G_{n,M_i}(df), \quad i = 1, 2.$$

Combining these with (1.3.16) and (1.3.17) gives

$$E_{Q_{0,n}}\phi_n \le \alpha + o(1), \tag{1.3.18}$$
$$E_{Q_{i,n}}(1-\phi_n) \le \Phi(z_\alpha - \sqrt{A_i/2}) + o(1), \quad i = 1, 2. \tag{1.3.19}$$

The likelihood ratio of $Q_{i,n}$ against $Q_{0,n}$ is

$$\begin{aligned}\frac{dQ_{i,n}}{dQ_{0,n}} &= \exp\left(-\frac{1}{2}\sum_j\left(\frac{Y_j^2}{n^{-1}+\sigma_j^{*2}(M_i)} - \frac{Y_j^2}{n^{-1}}\right)\right)\cdot\prod_j\left(\frac{n^{-1}}{n^{-1}+\sigma_j^{*2}(M_i)}\right)^{1/2}\\ &= \exp\left(\frac{1}{2}\sum_j\frac{n^2\sigma_j^{*2}(M_i)}{1+n\sigma_j^{*2}(M_i)}Y_j^2\right)\cdot\prod_j\left(\frac{n^{-1}}{n^{-1}+\sigma_j^{*2}(M_i)}\right)^{1/2}.\end{aligned}$$

Therefore, by the factorization theorem, it is seen that the bivariate vector

$$T_n = \left(\sum_j \frac{n^2\sigma_j^{*2}(M_1)(Y_j^2-n^{-1})}{(1+n\sigma_j^{*2}(M_1))\sqrt{2n^2\sum_k\sigma_k^{*4}(M_1)}}, \sum\frac{n^2\sigma_j^{*2}(M_2)(Y_j^2-n^{-1})}{(1+n\sigma_j^{*2}(M_2))\sqrt{2n^2\sum_k\sigma_k^{*4}(M_2)}}\right)$$

is a sufficient statistic for the family of distributions $\{Q_{0,n}, Q_{1,n}, Q_{2,n}\}$. Write the induced family for T_n as $\{Q_{0,n}^T, Q_{1,n}^T, Q_{2,n}^T\}$ and take the conditional expectation $\phi_n^*(T_n) = E_{Q_{i,n}}(\phi_n|T_n)$. By sufficiency (Bahadur´s theorem, cf. Lehmann and Romano, 2005, chap. 11), the (possibly randomized) test $\phi_n^*(T_n)$ for $\{Q_{0,n}^T, Q_{1,n}^T, Q_{2,n}^T\}$ is as good as ϕ_n, i.e.,

$$E_{Q_{0,n}^T}\phi_n^* = E_{n,0}\phi_n \le \alpha + o(1), \tag{1.3.20}$$
$$E_{Q_{i,n}^T}(1-\phi_n^*) = E_{Q_{1,n}}\phi_n \le \Phi(z_\alpha - \sqrt{A_i/2}) + o(1), \quad i = 1, 2. \tag{1.3.21}$$

Then we have the following lemma, which is proved later.

Lemma 1.3.1. *Under* $\{Q_{0,n}, Q_{1,n}, Q_{2,n}\}$*, the law of the statistic* T_n *converges in total variation to* $N(0,\Sigma)$*,* $N(\mu_1,\Sigma)$ *and* $N(\mu_2,\Sigma)$ *respectively, where*

$$\begin{aligned}
\mu_1 &= (\sqrt{A_1/2}, r\sqrt{A_1/2})', \\
\mu_2 &= (r\sqrt{A_2/2}, \sqrt{A_2/2})', \\
\Sigma &= \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}, \\
r &= \left(\frac{M_1}{M_2}\right)^{1/(4\beta)} \cdot \frac{4\beta + 1 - M_1/M_2}{4\beta}.
\end{aligned} \tag{1.3.22}$$

Then by the weak compactness theorem (c.f. Lehmann and Romano, 2005, Appendix), there exists a test ϕ^* and a subsequence $\phi^*_{n_k}$ such that $\phi^*_{n_k}$ converges weakly to ϕ^*. Thus

$$\begin{aligned}
&E_{Q^T_{0,n}}\phi^* \le \alpha, \\
&E_{Q^T_{i,n}}(1-\phi^*) \le \Phi(z_\alpha - \sqrt{A_i/2}), \quad i = 1, 2.
\end{aligned}$$

By the Neyman-Pearson lemma and some direct calculations, the right hand side of the previous inequality is the type II error of the uniformly most powerful test for $N(0,\Sigma)$ against $N(\mu_i,\Sigma)$, for $i = 1, 2$, respectively. Therefore, ϕ^* is a uniformly most powerful test for $N(0,\Sigma)$ against $\{N(\mu_1,\Sigma), N(\mu_2,\Sigma)\}$.

On the other hand , note that r in Lemma 1.3.1 is monotone increasing with respect to M_1/M_2, and then $0 < r < 1$ for $M_2 > M_1 > 0$. Thus, μ_1, μ_2 and the origin are not on the same line. For $i = 1, 2$ respectively, the log-likelihood ratio for $N(\mu_i,\Sigma)$ against $N(0,\Sigma)$ is $T'^{-1}\mu_i = T_i \cdot A_i$. Then by the necessity part of the Neyman-Pearson lemma, (cf. Lehmann and Romano, 2005, chap. 3), the uniformly most powerful test for $N(0,\Sigma)$ against $N(\mu_i,\Sigma)$ has the form of $\mathbf{1}\{T_i > k_i\}$. But since these two types of tests can never coincide, there is no uniformly

most powerful test for $N(0,\Sigma)$ against $\{N(\mu_1,\Sigma), N(\mu_2,\Sigma)\}$. By this contradiction, Theorem 1 is proved.

*Proof of Lemma **1.3.1**.* For simplicity, we only show the result for the first coordinate of T_n. The proof can be extended to T_n naturally. Under $Q_{0,n}$, the characteristic function of $\frac{n(Y_j^2-1/n)}{\sqrt{2}} \sim N(0,1)$ is $g(t) = \exp(-t^2/2)$. Note $g(t) = 1 - \frac{1}{2}t^2 + o(t^2)$, as $t \to 0$ and $\int |g(t)| < \infty$. The density of $T_{n,1}$ can be written as

$$p_n(x) = \frac{1}{2\pi} \int e^{-itx} \prod g\left(\frac{\sigma_j^{*2}(M_1) \cdot t}{(1 + n\sigma_j^{*2}(M_1)) \sqrt{\sum_k \sigma_k^{*4}(M_1)}} \right),$$

where, by the central limit theorem and Levy's continuity theorem, the integrand converges to $e^{-itx} \exp\{-t^2/2\}$. By splitting the integral into two parts and using dominated convergence, it can be shown that the integral converges to

$$\frac{1}{2\pi} \int e^{-itx} e^{-t^2/2}\, dt = \frac{e^{-x^2/2}}{\sqrt{2\pi}}.$$

Then an application of Scheffé's theorem (c.f. van der Vaart, 1998) establishes convergence in total variation. The correlation r can be calculated directly. □

1.4 Proof of Theorem 2

Choose $\tilde{N}$ and $\gamma_n = o(1)$ such that

$$\gamma_n^{1/2\beta} \cdot n^{2/(4\beta+1)} \gg \tilde{N} \gg c_n^{-1/(2\beta)} \cdot n^{2/(4\beta+1)}, \tag{1.4.23}$$

e.g. $\gamma_n = c_n^{-1/2}$, $\tilde{N} = c_n^{-1/(3\beta)} \cdot n^{2/(4\beta+1)}$. Define

$$\begin{aligned}
M_0 = M_0(f) &= \sum_{j=1}^{\tilde{N}} j^{2\beta} f_j^2 + \gamma_n, \\
N = N(M_0) &= \left(\frac{(4\beta+1)M_0}{\rho} \right)^{1/(2\beta)}, \\
\tilde{\lambda} = \tilde{\lambda}(M_0) &= \frac{2\beta+1}{2\beta} \left(\frac{1}{M_0(4\beta+1)} \right)^{1/(2\beta)} \rho^{(2\beta+1)/(2\beta)}, \\
\tilde{d}_j = \tilde{d}_j(M_0) &= \tilde{\lambda}[1 - (j/N)^{2\beta}]_+,
\end{aligned}$$

which all depend on the unknown f. Define the oracle statistic

$$T_n^* = \frac{n^2 \sum_j \tilde{d}_j(M_0) Y_j^2 - n \sum_j \tilde{d}_j(M_0)}{\sqrt{2n^2 \sum_j \tilde{d}_j^2(M_0)}},$$

and the oracle test $\phi_n^* = \mathbf{1}\{T_n^* > z_\alpha\}$. The following lemma holds; it is proved later.

Lemma 1.4.1. *Under the assumptions of Theorem 2, the oracle test ϕ_n^* is an asymptotic α-test and*

$$\overline{\lim_n} \frac{1}{c_n^{2+1/(2\beta)}} \log \Psi(\phi_n^*, \rho_n, \beta, M) \le -\frac{A_0(\beta) M^{-1/(2\beta)}}{4}$$

Define

$$\hat{M} = \sum_{j=1}^{\tilde{N}} (Y_j^2 - 1/n) j^{2\beta} + \gamma_n$$

and introduce the statistic

$$T_n = \frac{n^2 \sum \tilde{d}_j(\hat{M}) Y_j^2 - n \sum \tilde{d}_j(\hat{M})}{\sqrt{2n^2 \sum \tilde{d}_j^2(\hat{M})}}$$

and also the test

$$\phi_n = \mathbf{1}\{T_n > z_\alpha\}.$$

For $\hat{M}$, we have the following lemma, which is proved later.

Lemma 1.4.2. *Under the assumptions of Theorem 2, we have*

$$\frac{\hat{M}}{M_0(f)} - 1 = o_p(1),$$

uniformly for $f \in \Sigma(\beta, M) \cap B_\rho$.

Now rewrite

$$T_n = \sum_j \frac{\tilde{d}_j(\hat{M})}{\sqrt{\sum \tilde{d}_j^2(\hat{M})}} \cdot \frac{Y_j^2 - 1/n}{\sqrt{2n^{-2}}},$$

where $\tilde{d}_j(\hat{M}) = \tilde{\lambda}(1 - (j/N(\hat{M}))^{2\beta})_+$. Since $\tilde{\lambda}$ in the last display can be canceled, for simplicity we write $\tilde{d}_j(\hat{M}) = (1 - (j/N(\hat{M}))^{2\beta})_+$ from now on in this section. First, since $N(\hat{M}) \geq N(\gamma_n)$, we have

$$\begin{aligned}
\sum \tilde{d}_j^2(\hat{M}) &= \sum \left(1 - \left(\frac{j}{N(\hat{M})}\right)^{2\beta}\right)_+^2 \\
&\sim N(\hat{M}) \int_0^1 (1 - t^{2\beta})_+^2 \, dt \\
&= N(\hat{M}) K(\beta).
\end{aligned}$$

Therefore,

$$T_n = (1 + o(1)) \sum \frac{\tilde{d}_j(\hat{M})}{\sqrt{N(\hat{M})K(\beta)}} \cdot \frac{Y_j^2 - 1/n}{\sqrt{2n^{-2}}}.$$

By Lemma 1.4.2,

$$T_n = (1 + o(1)) \sum_j \frac{\tilde{d}_j(\hat{M})}{\sqrt{N(M_0(f))K(\beta)}} \cdot \frac{Y_j^2 - 1/n}{\sqrt{2n^{-2}}}.$$

At this point, make $\hat{M}$ independent of Y_j^2 by sample splitting. Set $n = \tau n + (1-\tau)n$, where τ is close to 1 but fixed, and $n_1 = \tau n, n_2 = (1 - \tau)n$. Assume two sets of observations

$$Y_{1j} = f_j + n_1^{-1/2} \xi_{1j}, j = 1, 2, \ldots \tag{1.4.24}$$

$$Y_{2j} = f_j + n_2^{-1/2} \xi_{2j}, j = 1, 2, \ldots \tag{1.4.25}$$

Use $\{Y_{2j}\}$ to obtain $\hat{M}$, and now replace T_n by

$$T_n^s = (1 + o(1)) \sum_j \frac{\tilde{d}(\hat{M})}{\sqrt{N(M_0(f))K(\beta)}} \cdot \frac{Y_{1j}^2 - n^{-1}}{\sqrt{2}n^{-1}}.$$

Denote the difference of coefficients by $\Delta_j = \tilde{d}_j(\hat{M}) - \tilde{d}_j(M_0(f))$. Note the largest difference is obtained at $j \approx \min\{N(\hat{M}), N(M_0(f))\}$. Then

$$|\Delta_j| \leq \frac{|\hat{M} - M_0(f)|}{\gamma_n}$$

uniformly for all j. Note in T_1 there are at most $C_2 c_n^{-1/(2\beta)} n^{2/(4\beta+1)}$ nonzero coefficients. Then

$$T_n^s = (1 + o(1)) \sum_{j=1}^{C_2 c_n^{-1/(2\beta)} n^{2/(4\beta+1)}} \frac{\tilde{d}_j(M_0(f))}{\sqrt{N(M_0(f))K(\beta)}} \eta_j + r_n$$

where $\eta_j = \frac{Y_{1j}^2 - n_1^{-1}}{\sqrt{2}n_1^{-1}}$, and

$$r_n = \sum_{j=1}^{C_2 c_n^{-1/(2\beta)} n^{2/(4\beta+1)}} \frac{\Delta_j \eta_j}{\sqrt{N(M_0(f))K(\beta)}}.$$

Under H_0, the r.v.´s η_j are independent of $\hat{M}$ and $E\eta_j = 0$, $\mathrm{Var}(\eta_j) = 1$. Thus $\mathrm{Var}(r_n) = Er_n^2 = EE(r_n^2|\{Y_{2j}\})$ and

$$E(r_n^2|\{Y_{2j}\}) = E \sum_{j=1}^{C_2 c_n^{-1/(2\beta)} n^{2/(4\beta+1)}} \frac{\Delta_j^2}{N(M_0(f))K(\beta)} \leq \frac{|\hat{M} - M_0(f)|^2}{\gamma_n^{2+1/(2\beta)}}.$$

Therefore, by the result for $\mathrm{Var}(\hat{M})$ in the proof of Lemma 1.4.2,

$$\mathrm{Var}(r_n) \leq \frac{E|\hat{M} - M_0(f)|^2}{\gamma_n^{2+1/(2\beta)}} = \frac{\mathrm{Var}(\hat{M})}{\gamma_n^{2+1/(2\beta)}} \leq \frac{2K(\beta)\tilde{N}^{4\beta+1}}{n^2\gamma_n^{2+1/(2\beta)}} + \frac{4\tilde{N}^{2\beta}M}{n\gamma_n^{2+1/(2\beta)}},$$

where the last two terms converge to 0 by the first inequality in (1.4.23). Hence, under H_0, the r.v.´s T_n and T_n^s converge to $N(0,1)$ in law.

Next, we consider T_n or T_n^s under the alternative. The worst case type II error is determined by the following quantity

$$L_n = \frac{n}{\sqrt{2}} \inf_{f \in \Sigma(\beta,M) \cap B_\rho} \frac{\sum_{j=1}^{\tilde{N}} f_j^2 \tilde{d}_j(\hat{M})}{\left(\sum \tilde{d}_j(\hat{M})\right)^{1/2}}.$$

First, since $N(\hat{M}) \geq \left(\frac{\gamma_n}{c_n}\right)^{1/(2\beta)} \cdot n^{2/(4\beta+1)} \to \infty$,

$$\begin{aligned}\tilde{d}_j^2 &= \sum_{j=1}^{\tilde{N}} \left(1 - (j/N)^{2\beta}\right)_+^2 \\ &= (1 + o(1))N \int_0^1 (1 - t^{2\beta})^2 dt \\ &= (1 + o(1))N \cdot \frac{8\beta^2}{(2\beta+1)(4\beta+1)}. \end{aligned} \tag{1.4.26}$$

Second, consider

$$\sum_{j=1}^{\tilde{N}} f_j^2 \tilde{d}_j(\hat{M}) = \sum_{j=1}^{\tilde{N}} f_j^2 (1 - (j/N)^{2\beta})_+.$$

Note

$$\begin{aligned}\sum_{j=1}^{\tilde{N}} f_j^2 &= \sum_{j=1}^{\infty} f_j^2 - \sum_{j=\tilde{N}+1}^{\infty} f_j^2 \\ &\geq \rho - \tilde{N}^{-2\beta} M \\ &= \rho \left(1 - \frac{M}{\rho \tilde{N}^{2\beta}}\right) \\ &= \rho(1 + o(1)), \end{aligned} \tag{1.4.27}$$

where the last step is refers to the second inequality of (1.4.23). On the other hand, since $\tilde{N} \gg N$ and $N(\hat{M}) = [(4\beta+1)\hat{M}\rho^{-1}]^{1/(2\beta)}$,

$$\begin{aligned}\sum_{j=1}^{N} f_j^2 (j/N)^{2\beta} + \sum_{j=N+1}^{\tilde{N}} f_j^2 &\leq \sum_{j=1}^{\tilde{N}} f_j^2 (j/N)^{2\beta} \\ &\leq N^{-2\beta} M_0(f) \\ &= \rho(1 + 4\beta)^{-1}. \end{aligned} \tag{1.4.28}$$

Combining (1.4.30)-(1.4.32) gives

$$\sum_{j=1}^{\tilde{N}} f_j^2 \tilde{d}_j \geq (1 + o(1))\tilde{\lambda}\rho \cdot \frac{4\beta}{4\beta+1}.$$

Combining this with (1.4.29) gives

$$\begin{aligned}\frac{n\sum_{j=1}^{\tilde{N}} f_j^2\tilde{d}_j}{(2\sum\tilde{d}_j^2)^{1/2}} &\geq (1+o(1))\frac{n}{\sqrt{2}}\sqrt{\frac{2(2\beta+1)}{4\beta+1}\rho^2/N} \\ &\geq (1+o(1))\sqrt{\frac{(2\beta+1)c_n^{2+1/(2\beta)}}{(4\beta+1)^{1+1/(2\beta)}(M+\gamma_n)^{1/(2\beta)}}} \\ &\geq (1+o(1))\sqrt{\frac{1}{2}A_0(\beta)c_n^{2+1/(2\beta)}M^{-1/(2\beta)}}\end{aligned}$$

Theorem 2 is proved.

Proof of Lemma 1.4.1. Rewrite

$$T_n^* = \sum_j \frac{\tilde{d}_j(M_0(f))}{\sqrt{\sum\tilde{d}_j^2(M_0(f))}}\cdot\frac{Y_j^2-1/n}{\sqrt{2n^{-2}}}.$$

Under H_0, we have $f=0$, and $M_0(f)=\gamma_n$. Since

$$\sum[1-(j/N)^{2\beta}]_+^2 \sim N\cdot\int_0^1(1-t^{2\beta})^2\,dt = K(\beta)\cdot(\gamma_n/c_n)^{1/2\beta}n^{2/(4\beta+1)},$$

then

$$\left|\frac{\tilde{d}_j(M_0(f))}{\sqrt{\sum\tilde{d}_j^2(M_0(f))}}\right| \leq \frac{1}{\sqrt{K(\beta)\cdot(\gamma_n/c_n)^{1/2\beta}n^{2/(4\beta+1)}}} = o(1),$$

uniformly for all j. It can be shown that T_n^* converges to $N(0,1)$ in law.

By similar arguments, the worst type II error is $(1+o(1))\Phi(z-L_n)$ where

$$L_n = \inf_{f\in\Sigma(\beta,M)\cap B_\rho}\frac{n\sum f_j^2\tilde{d}_j}{(2\sum\tilde{d}_j^2)^{1/2}}.$$

Note $\tilde{d}_j = \tilde{d}_j(M_0(f))$ depending on f. By the second inequality of (1.4.23), we have $\tilde{N}\gg N(M_0(f))$ and $\tilde{d}_j=0$, for $j\geq\tilde{N}$,

$$L_n = \frac{n}{\sqrt{2}}\inf_{f\in\Sigma(M)\cap B_\rho}\frac{\sum_{j=1}^{\tilde{N}}f_j^2\tilde{d}_j}{(\sum\tilde{d}_j^2)^{1/2}}.$$

First, since $N(M_0(f)) \geq \left(\frac{\gamma_n}{c_n}\right)^{1/(2\beta)} \cdot n^{2/(4\beta+1)} \to \infty$ uniformly for $f \in \Sigma(\beta, M) \cap B_\rho$,

$$\begin{aligned}
\tilde{d}_j^2 &= \tilde{\lambda}^2 \sum_{j=1}^{\tilde{N}} \left(1 - (j/N)^{2\beta}\right)_+^2 \\
&= (1 + o(1))\tilde{\lambda}^2 N \int_0^1 (1 - t^{2\beta})^2 dt \\
&= (1 + o(1))\tilde{\lambda}^2 N \cdot \frac{8\beta^2}{(2\beta + 1)(4\beta + 1)},
\end{aligned} \tag{1.4.29}$$

uniformly for $f \in \Sigma(\beta, M) \cap B_\rho$. Second, consider

$$\sum_{j=1}^{\tilde{N}} f_j^2 \tilde{d}_j = \tilde{\lambda} \sum_{j=1}^{\tilde{N}} f_j^2 (1 - (j/N)^{2\beta})_+ = \tilde{\lambda} \left[\sum_{j}^{\tilde{N}} f_j^2 - \left(\sum_{j}^{N} f_j^2 (j/N)^{2\beta} + \sum_{j=N+1}^{\tilde{N}} f_j^2 \right) \right]. \tag{1.4.30}$$

Note

$$\begin{aligned}
\sum_{j=1}^{\tilde{N}} f_j^2 &= \sum_{j=1}^{\infty} f_j^2 - \sum_{j=\tilde{N}+1}^{\infty} f_j^2 \\
&\geq \rho - \tilde{N}^{-2\beta} M \\
&= \rho \left(1 - \frac{M}{\rho \tilde{N}^{2\beta}}\right) \\
&= \rho(1 + o(1)),
\end{aligned} \tag{1.4.31}$$

where the last step is due to the second inequality of (1.4.23). On the other hand, since $\tilde{N} \gg N$ and $N = [\rho^{-1}(4\beta + 1)M_0(f)]^{1/(2\beta)}$,

$$\begin{aligned}
\sum_{j=1}^{N} f_j^2 (j/N)^{2\beta} + \sum_{j=N+1}^{\tilde{N}} f_j^2 &\leq \sum_{j=1}^{\tilde{N}} f_j^2 (j/N)^{2\beta} \\
&\leq N^{-2\beta} M_0(f) \\
&= \rho(1 + 4\beta)^{-1}
\end{aligned} \tag{1.4.32}$$

Combining (1.4.30)-(1.4.32) gives

$$\sum_{j=1}^{\tilde{N}} f_j^2 \tilde{d}_j \geq (1 + o(1))\tilde{\lambda}\rho \cdot \frac{4\beta}{4\beta + 1}$$

uniformly for $f \in \Sigma(\beta, M) \cap B_\rho$. Combining this with (1.4.29) gives

$$\begin{aligned}\frac{n\sum_{j=1}^{\tilde{N}} f_j^2 \tilde{d}_j}{(2\sum \tilde{d}_j^2)^{1/2}} &\geq (1+o(1))\frac{n}{\sqrt{2}}\sqrt{\frac{2(2\beta+1)}{4\beta+1}\rho^2/N} \\ &\geq (1+o(1))\sqrt{\frac{(2\beta+1)c_n^{2+1/(2\beta)}}{(4\beta+1)^{1+1/(2\beta)}(M+\gamma_n)^{1/(2\beta)}}} \\ &\geq (1+o(1))\sqrt{\frac{(2\beta+1)c_n^{2+1/(2\beta)}}{(4\beta+1)^{1+1/(2\beta)}M^{1/(2\beta)}}},\end{aligned}$$

uniformly for $f \in \Sigma(\beta, M) \cap B_\rho$. Therefore,

$$L_n \geq (1+o(1))\sqrt{\frac{1}{2}A_0(\beta)c_n^{2+1/(2\beta)}M^{-1/(2\beta)}},$$

and the result follows. □

Proof of Lemma 1.4.2. Since

$$\begin{aligned}\mathrm{Var}(\hat{M}) &= \sum_{j=1}^{\tilde{N}}\left(\frac{2}{n^2}+\frac{4f_j^2}{n}\right)j^{4\beta} \\ &\leq (1+o(1))\frac{2K(\beta)\tilde{N}^{4\beta+1}}{n^2}+\frac{4\tilde{N}^{2\beta}M}{n},\end{aligned}$$

by the first inequality of (1.4.23),

$$\frac{\mathrm{Var}(\hat{M})}{\gamma_n^2} = o(1)$$

uniformly for $f \in \Sigma \cap V_\rho$. Combining with $E\hat{M} = M_0(f)$ and using Chebyshev's inequality give

$$\frac{\left|\hat{M}-M_0(f)\right|}{\gamma_n} = o_p(1),$$

and then

$$\left|\frac{\hat{M}}{M_0(f)}-1\right| \leq \frac{\left|\hat{M}-M_0(f)\right|}{\gamma_n} = o_p(1),$$

uniformly for $f \in \Sigma \cap V_\rho$. □

1.5 Appendix

1.5.1 Ideas on adaptive estimation

Consider the estimation problem for the Gaussian sequence model

$$Y_j = f_j + n^{-1/2}\xi_j$$

with $\sum j^{2\beta} f_j^2 \le M$. It is known, for given M, the optimal filter is $(1-\mu j^\beta)_+$, where μ is determined by

$$\frac{1}{n}\sum j^\beta (1-\mu j^\beta)_+ = \mu M.$$

Since

$$\mu \sim \left(\frac{\beta \cdot n^{-1}}{M(\beta+1)(2\beta+1)}\right)^{\beta/(2\beta+1)},$$

the optimal truncation is of the order $n^{1/(2\beta+1)}$.

Choose $n^{1/(2\beta+1/2)} \gg \tilde{N} \gg n^{1/(2\beta+1)}$ and $1 \gg \gamma_n \gg \tilde{N}^{2\beta+1/2}/n$, and define

$$M_{0,f} = \sum_{j=1}^{\tilde{N}} j^{2\beta} f_j^2 + \gamma_n.$$

Define $N = N(M_{0,f}) = \alpha \cdot n^{1/(2\beta+1)} M_{0,f}^{1/(2\beta+1)}$, where α is a constant to be chosen. Define

$$d_j = d(j/N) = \left(1-(j/N)^\beta\right)_+.$$

Consider the oracle estimator $(d_j Y_j)_1^\infty$. Its risk is

$$\begin{aligned}
&\sum (1-d_j)^2 f_j^2 + \frac{1}{n}\sum d_j^2 \\
=&\sum_{j=1}^{\tilde{N}} (1-d_j)^2 f_j^2 + \sum_{j>\tilde{N}} (1-d_j)^2 f_j^2 + \frac{1}{n}\sum d_j^2 \\
:=&A_1 + A_2 + A_3.
\end{aligned}$$

First,

$$A_1 \le \sup_{j \le \tilde{N}} (1-d_j)^2 j^{-2\beta} M_{0,f} \le N^{-2\beta} M_{0,f} = \alpha^{-2\beta} n^{-2\beta/(2\beta+1)} (M+\gamma_n)^{1/(2\beta+1)}$$

Second, $A_2 = \sum_{j>\tilde{N}} f_j^2 \le \tilde{N}^{-2\beta} M = o(n^{-2\beta/(2\beta+1)})$. Third,

$$\begin{aligned} A_3 &= \frac{N}{n}\frac{1}{N}\sum (1-(j/N)^\beta)_+^2 \\ &\le \alpha n^{-2\beta/(2\beta+1)} (M+\gamma_n)^{1/(2\beta+1)} \int_0^\infty (1-t^\beta)_+^2 dt \\ &= \alpha n^{-2\beta/(2\beta+1)} (M+\gamma_n)^{1/(2\beta+1)} \cdot \frac{2\beta^2}{(\beta+1)(2\beta+1)} \end{aligned}$$

These results hold uniformly over f. Combine these and choose $\alpha = \left(\frac{(\beta+1)(2\beta+1)}{\beta}\right)^{1/(2\beta+1)}$, and we have the supremum risk of the oracle estimator over f is at most

$$c(m) \cdot n^{-2\beta/(2\beta+1)} M^{1/(2\beta+1)},$$

where

$$c(m) = \left(\frac{\beta}{\beta+1}\right)^{2\beta/(2\beta+1)} \cdot (1+2\beta)^{1/(2\beta+1)}$$

is the Pinsker constant.

Let $\hat{M}_n = \sum_{j=1}^{\tilde{N}_n} j^{2\beta} \hat{f}_j^2 + \gamma_n$, where $\hat{f}_j^2 = y_j^2 - n^{-1}$. Then

$$E(\hat{M}) = \sum_{j=1}^{\tilde{N}_n} j^{2\beta} f_j^2 = M_{0,f} \le M + \gamma_n$$

and

$$\begin{aligned} \mathrm{Var}(\hat{M}) &= \sum_{j=1}^{\tilde{N}} j^{4\beta} \mathrm{Var}(Y_j^2) \\ &= \sum_{j=1}^{\tilde{N}} j^{4\beta} n^{-2} (2+4nf_j^2) \\ &= 2n^{-2} \sum_{j=1}^{\tilde{N}} j^{4\beta} + 4n^{-1} \sum_{j=1}^{\tilde{N}} j^{4\beta} f_j^2 \\ &= J_1 + J_2, \end{aligned}$$

where the first term

$$J_1 = 2n^{-2}\tilde{N}^{4\beta+1} \cdot \frac{1}{\tilde{N}} \sum_{j=1}^{\tilde{N}} \left(j/\tilde{N}\right)^{4\beta} \sim 2n^{-2}\tilde{N}^{4\beta+1} \cdot K = o(1)$$

since $\tilde{N} = o(n^{1/(2\beta+1/2)})$, and the second term

$$J_2 \leq 4n^{-1}\tilde{N}^{2\beta} \sum_{j=1}^{\tilde{N}} j^{2\beta} f_j^2 = 4n^{-1}\tilde{N}^{2\beta} M_{0,f} \lesssim 4KMn^{-2}\tilde{N}^{4\beta+1} \cdot \frac{n}{\tilde{N}^{2\beta+1}} = o(J_1)$$

uniformly for $f \in \Sigma(\beta, M)$ since $\tilde{N} \gg n^{1/(2\beta+1)}$. Combining these gives $\mathrm{Var}(\hat{M}) = o(1)$ uniformly for $f \in \Sigma(\beta, M)$. Recalling $\gamma_n \gg \tilde{N}^{2\beta+1/2}/n$ gives

$$\mathrm{Var}\left(\frac{\hat{M} - M_{0,f}}{\gamma_n}\right) \sim \frac{2Kn^{-2}\tilde{N}^{4\beta+1}}{\gamma_n^2} = o(1),$$

and then

$$\left|\frac{\hat{M}}{M_{0,f}} - 1\right| \leq \left|\frac{\hat{M} - M_{0,f}}{\gamma_n}\right| = o_p(1) \tag{1.5.33}$$

uniformly.

The last result is crucial for the next step, i.e. showing that the difference between oracle estimator $\left(d_j Y_j\right)_1^\infty$ and the estimator $\left(d(\frac{j}{N(\hat{M})})Y_j\right)_1^\infty$ is negligible. Re call that $N = N(M_{0,f})$; now (1.5.33) is used to replace $M_{0,f}$ by estimate $\hat{M}$. The remainder of the proof consists of showing

$$E \sum \left(d(j/N(M_{0,f})) - d(j/N(\hat{M}))\right)^2 Y_j^2 = o(n^{-2\beta/(2\beta+1)}).$$

1.5.2 Proofs for Section 1.2

Proof of Lemma 1.2.1. (a) Under the null hypothesis we have $Y_j^2 = n^{-1}\xi_j^2$, hence $T = \sum d_j\left(\xi_j^2 - 1\right)/\sqrt{2}$. Then it follows from (1.2.8) and $n\rho \to \infty$ that the CLT infinitesimality condition

$$\sup_j d_j^2 = o(1)$$

holds uniformly over $d \in \mathcal{D}$, proving the assertion.

(b) Since $Y_j^2 = f_j^2 + 2n^{-1/2} f_j \xi_j + n^{-1}\xi_j^2$, we have

$$T = \frac{1}{\sqrt{2}} \sum d_j \left(n f_j^2 + 2n^{1/2} f_j \xi_j + \left(\xi_j^2 - 1 \right) \right), \tag{1.5.34}$$

$$T - L(d, f) = \frac{1}{\sqrt{2}} \sum d_j \left(2n^{1/2} f_j \xi_j + \left(\xi_j^2 - 1 \right) \right). \tag{1.5.35}$$

An easy calculation gives

$$\mathrm{Var}_f T = \frac{1}{2} \sum d_j^2 \left(4n f_j^2 + 2 \right) = 1 + 2n \sum d_j^2 f_j^2$$

where in view of (1.2.8) we have for $f \in B'_\rho$

$$n \sum d_j^2 f_j^2 \le \delta \rho^{-1} \sum f_j^2 \le 2\delta = o(1).$$

Consequently, $\mathrm{Var}_f T \to 1$ uniformly. Now the CLT infinitesimality condition on the sum (1.5.35) amounts to

$$\sup_j d_j^2 \left(n f_j^2 + 1 \right) = o(1). \tag{1.5.36}$$

For $f \in B'_\rho$ we have $f_j^2 \le 2\rho$, hence in view of (1.2.8)

$$d_j^2 \left(n f_j^2 + 1 \right) \le d_j^2 \left(2n\rho + 1 \right) \le 2\delta$$

for n sufficiently large. Hence (1.5.36) is fulfilled uniformly over $d \in \mathcal{D}$ and $f \in B'_\rho$, and the claim follows.

(c) Set $f_j \sim N(0, \sigma_j^2)$; then in view of (1.5.34)

$$T - L(d, \sigma) = \frac{1}{\sqrt{2}} \sum d_j \left(2n^{1/2} f_j \xi_j + \left(\xi_j^2 - 1 \right) \right) + \frac{n}{\sqrt{2}} \sum d_j \left(f_j^2 - \sigma_j^2 \right). \tag{1.5.37}$$

An easy calculation gives

$$\begin{aligned} \mathrm{Var}_f T &= \frac{1}{2} \sum d_j^2 \left(4n\sigma_j^2 + 2 \right) + n \sum d_j^2 \\ &= 1 + n \sum d_j^2 \left(2\sigma_j^2 + \sigma_j^4 \right) \end{aligned}$$

where in view of (1.2.8) we have for $\sigma \in B'_\rho$

$$n \sum d_j^2 \sigma_j^2 \le \delta \rho^{-1} \sum \sigma_j^2 \le 2\delta = o(1),$$
$$n \sum d_j^2 \sigma_j^4 \le 2\rho n \sum d_j^2 \sigma_j^2 \le 4\rho\delta = o(1).$$

Consequently, $\mathrm{Var}_f T \to 1$ uniformly. Now the infinitesimality condition on the sum (1.5.37) amounts to

$$\sup_j d_j^2 \left(1 + n\sigma_j^2 + n\sigma_j^4\right) = o(1). \tag{1.5.38}$$

For $\sigma \in B'_\rho$ we have $\sigma_j^2 \le 2\rho$, hence in view of (1.2.8)

$$d_j^2 \left(1 + n\sigma_j^2 + n\sigma_j^4\right) \le d_j^2 \left(1 + n\rho + n\rho^2\right) \le 3\delta$$

for n sufficiently large. Hence (1.5.38) is fulfilled uniformly over $d \in \mathcal{D}$ and $\sigma \in B'_\rho$, and the claim follows. □

Proof of Lemma 1.2.2 . Let $\tilde{\mathcal{D}}$ be defined as $\mathcal{D}$ in (1.2.8) but with condition $\|d\|^2 = 1$ replaced by $\|d\|^2 \le 1$. Then, since $L(d, f)$ is linear in d, for every $\tilde{d} \in \tilde{\mathcal{D}}$ there is a $d \in \mathcal{D}$ such that $L(\tilde{d}, f^2) \le L(d, f^2)$ for every f. Hence it suffices to prove the claim for $\mathcal{D}$ replaced by the compact convex set $\tilde{\mathcal{D}}$. The restriction $f \in \Sigma(\beta, M) \cap B'_\rho$ is equivalent to f^2 being in the set

$$\left\{g \in \mathbb{R}_+^n : \sum g_j j^{2\beta} \le M, \rho \le \sum g_j \le 2\rho\right\} \tag{1.5.39}$$

which is convex and compact (and nonempty for large enough n since $\rho \to 0$). The functional L is bilinear in d and f^2; the standard minimax theorem now furnishes the result. □

Lemma 1.5.1. *For n large enough, the saddlepoint d_0, f_0 of Lemma 1.2.2 is given by*

$$d_0 = \frac{f_0^2}{\left\|f_0^2\right\|}, \quad f_{0,j}^2 = \left(\lambda - \mu j^{2\beta}\right)_+, \; j = 1, \ldots, n$$

where λ, μ are the unique positive solutions of the equations

$$\sum_{j=1}^{n} j^{2\beta}\left(\lambda - \mu j^{2\beta}\right)_+ = M, \; \sum_{j=1}^{n}\left(\lambda - \mu j^{2\beta}\right)_+ = \rho. \tag{1.5.40}$$

The value of L at the saddlepoint is

$$L_0 = L(d_0, f_0) = \frac{n}{\sqrt{2}}\left\|f_0^2\right\|. \tag{1.5.41}$$

Proof. Ignore initially the restriction $\sup_j d_j^2 \le \delta/n\rho$ and consider maximizing $L(d, f^2)$ in d for given f. Under the sole restriction $\|d\| = 1$, by Cauchy-Schwartz the solution is found as

$$d(f) = \frac{f^2}{\left\|f^2\right\|}.$$

It remains to minimize $L(d(f), f) = n\left\|f^2\right\| / \sqrt{2}$ under the restrictions on f^2. Setting $g_j = f_j^2$, one has to minimize $\|g\|$ on the convex set (1.5.39). This is solved using Lagrange multipliers λ, μ.

To show that the solution d_0 fulfills the restriction $\sup_j d_j^2 \le \delta/n\rho$, we note that

$$f_{0,j}^2 = \left(\lambda - \mu j^{2\beta}\right)_+ = \lambda\left(1 - \mu\lambda^{-1} j^{2\beta}\right)_+ \le \lambda; \tag{1.5.42}$$

below (cf. (1.5.49), Lemma 1.5.2) it is shown that $\lambda \asymp n^{-1-1/(4\beta+1)}$ and $n\left\|f_0^2\right\| \asymp L_{n,0} \asymp 1$. This implies

$$\begin{aligned} n\rho d_{0,n,j}^2 &= n\rho \cdot O\left(n^2\lambda^2\right), \\ n^3\rho\lambda^2 &\asymp n \cdot n^{-4\beta/(4\beta+1)} \cdot n^{-2/(4\beta+1)} = n^{-1/(4\beta+1)}; \end{aligned} \tag{1.5.43}$$

thus for $\delta = (\log n)^{-1}$ we have that $d_0 \in \mathcal{D}$ for n large enough. □

Proof of Lemma 1.2.3. The log-likelihood ratio is

$$\log \frac{\left(n^{-1}\right)^{n/2}}{\left(\sigma_j^2 + n^{-1}\right)^{n/2}} \exp\left(-\frac{1}{2}\sum_{j=1}^{n}\left(\frac{Y_j^2}{\sigma_j^2 + n^{-1}} - \frac{Y_j^2}{n^{-1}}\right)\right)$$
$$= \frac{1}{2}\sum_{j=1}^{n} nY_j^2\left(\frac{n\sigma_j^2}{n\sigma_j^2 + 1}\right) - \frac{n}{2}\sum_{j=1}^{n}\log\left(n\sigma_j^2 + 1\right).$$

This shows (a) by setting $d = \tilde{d}/\left\|\tilde{d}\right\|$ for $\tilde{d}_j = \frac{n\sigma_j^2}{n\sigma_j^2+1}$. Now for $\sigma_j^2 = f_{0j}^2$ we have, as $\lambda \asymp n^{-1-1/(4\beta+1)}$,

$$nf_{0j}^2 = n\lambda\left(1 - \lambda^{-1}\mu j^{2\beta}\right)_+ \le n\lambda \asymp n \cdot n^{-1-1/(4\beta+1)} = n^{-1/(4\beta+1)} = o(1),$$

hence $\tilde{d}_j \sim nf_{0j}^2$ uniformly over $j = 1, \ldots, n$. This implies $\left\|\tilde{d}\right\| \sim n\left\|f_0^2\right\| \asymp n$ and

$$d_j = \frac{\tilde{d}_j}{\left\|\tilde{d}\right\|} \asymp f_{0j}^2$$

uniformly in $j \le n$. The proof of $n\rho d_{0,n,j}^2 \le \delta$ now exactly follows (1.5.42), (1.5.43) (CHECK). The convergence $t \to z_\alpha$ now is a consequence of Lemma 1.2.1 (a). □

Lemma 1.5.2. *Suppose $\rho = c \cdot n^{-4\beta/(4\beta+1)}$, c constant. Then the saddlepoint value L_0 of (1.2.12) fulfills*

$$L_0 = L(d_0, f_0^2) \sim \sqrt{A_0 M^{-1/(2\beta)} c^{2+1/(2\beta)}/2}.$$

Proof. The proof of Lemma 1.5.1 shows that $L(d_0, f_0^2)$ is also the saddlepoint value under the weaker restrictions $\|d\|^2 \le 1$, $f \in \Sigma(\beta, M) \cap B_\rho$. Let us sketch a derivation of the asymptotics by a renormalization technique. Suppose that $d_j = h^{1/2}d(hj)$, $j \le n$ where h is a bandwidth parameter tending to 0, and the continuous function $d : [0, \infty) \to [0, \infty)$ satisfies

$$\int_0^\infty d^2(x)\,dx \le 1. \tag{1.5.44}$$

Consider another continuous function $\sigma : [0, \infty) \to [0, \infty)$ satisfying

$$\int_0^\infty x^{2\beta}\sigma^2(x)\,dx \le 1 \text{ and } \int_0^\infty \sigma^2(x)\,dx \ge 1 \tag{1.5.45}$$

and set $\sigma_j^2 = Mh^{2\beta+1}\sigma^2(hj)$, $j \le n$. Choose $h = (\rho/M)^{1/(2\beta)}$. The coefficient vector $d = \left(d_j\right)_{j=1}^n$ satisfies

$$\|d\|^2 = h\sum_{j=1}^{n} d(hj) \to \int_0^\infty d(x)dx \le 1.$$

Identifying $f^2 \in \mathbb{R}_+^n$ with $(\sigma_j^2)_{j=1}^n$, the restriction $f \in \Sigma(\beta, M)$ is asymptotically satisfied since

$$\sum_{j=1}^{\infty} j^{2\beta}\sigma_j^2 = Mh\sum_{j=1}^{\infty}(jh)^{2\beta}\sigma^2(jh) \to M\int_0^\infty x^{2\beta}\sigma^2(x)\,dx \le M, \quad h \to 0.$$

The restriction $f \in B_\rho$ is also asymptotically satisfied since

$$\sum_{j=1}^{\infty}\sigma_j^2 = Mh^{2\beta+1}\sum_{j=1}^{\infty}\sigma^2(jh) = \rho h\sum_{j=1}^{\infty}\sigma^2(jh) \sim \rho\int_0^\infty \sigma^2(x)\,dx \ge \rho.$$

Therefore,

$$\begin{aligned}\frac{n}{\sqrt{2}}\sum_{j=1}^{n} d_j\sigma_j^2 &= \frac{n}{\sqrt{2}}Mh^{2\beta+1/2}h\sum_{j=1}^{\infty} d(jh)\sigma^2(jh)\\ &\sim \frac{c^{1+1/(4\beta)}M^{-1/(4\beta)}}{\sqrt{2}}\int_0^\infty d(x)\sigma^2(x)\,dx.\end{aligned}$$

The saddle point problem (1.2.12) for each n is thus asymptotically expressed in terms of a fixed continuous problem with constraints (1.5.44) and (1.5.45). There is unique positive solution (λ^*, μ^*) for the equations (cp. Golubev, 1982),

$$\int_0^\infty x^{2\beta}(\lambda - \mu x^{2\beta})\,dx = 1, \tag{1.5.46}$$

$$\int_0^\infty (\lambda - \mu x^{2\beta})\,dx = 1. \tag{1.5.47}$$

Let $\|\cdot\|_2$ and $\langle\cdot,\cdot\rangle_2$ denote norm and scalar product in $L_2(\mathbb{R}_+)$. Then the saddle point (d^*, σ^{*2}) is given by

$$d^* = \frac{\sigma^{*2}}{\left\|\sigma^{*2}\right\|_2},\ \sigma^{*2}(x) = (\lambda^* - \mu^* x^{2\beta})_+. \tag{1.5.48}$$

Then the value of the game is

$$\sup_{d \text{ in (1.5.44)}} \inf_{\sigma \text{ in (1.5.45)}} \left\langle d, \sigma^2 \right\rangle_2 = \inf_{\sigma \text{ in (1.5.45)}} \sup_{d \text{ in (1.5.44)}} \left\langle d, \sigma^2 \right\rangle_2$$
$$= \left\langle d^*, \sigma^{*2} \right\rangle_2 = \left\| \sigma^{*2} \right\|_2 = \sqrt{A_0(\beta)},$$

where the sup is taken for d satisfying (1.5.44), the inf is taken for σ satisfying (1.5.45), and $A_0(\beta)$ is Ermakov's constant in (1.1.3). The continuous saddlepoint problem arises naturally in a continuous Gaussian white noise setting and a parameter space described by the continuous Fourier transformation, e.g. a Sobolev class of functions on the whole real line (cf. Golubev 1982, 1987).

The above argument provides the guideline for a more rigourous proof, based on calculating the sharp asymptotics of λ and μ directly from (1.5.40). The rough order of λ can be found as follows. By equating $f_0^2 = \sigma_j^{*2}$, we find

$$\left(\lambda - \mu j^{2\beta}\right)_+ = Mh^{2\beta+1}\sigma^{*2}(hj),$$
$$= \lambda\left(1 - \left((\mu/\lambda)^{1/2\beta}\, j\right)^{2\beta}\right)_+$$

we find $\lambda \asymp h^{2\beta+1}$, $h \asymp (\mu/\lambda)^{1/2\beta}$ and thus

$$\lambda \asymp h^{2\beta+1} \asymp (\rho)^{(2\beta+1)/(2\beta)} \asymp n^{-1-1/(4\beta+1)}. \tag{1.5.49}$$

□

Remark 1.5.1. *The paper of Ermakov (1990), when calculating the asymptotics of λ, μ in (1.5.40) and of $A = 2L_0^2$ (in a more general framework where $\sum a_j f_j^2 \le P_0$, $\sum b_j f_j^2 \ge \rho$), contains an error for λ. Here is the correction using the notations therein. Let $a_j = Lj^{2\gamma}$, $b_j = Mj^{2\nu}$, where $\gamma > \nu \ge 0$, L and M are positive constants, and set $\epsilon = n^{-1/2}$. Then as $\epsilon \to 0$ we have that*

$$\lambda \sim \frac{(2\gamma + 2\nu + 1)}{2(\gamma - \nu)} \left(\frac{L}{P_0(4\gamma+1)}\right)^{\frac{4\nu+1}{2(\gamma-\nu)}} \left(\frac{1}{M}\right)^{\frac{4\gamma+1}{2(\gamma-\nu)}} [\rho(4\nu+1)]^{\frac{2(\gamma+\nu)+1}{2(\gamma-\nu)}},$$

$$\mu \sim \frac{(4\nu+1)\rho\lambda}{P_0(4\gamma+1)}, \qquad A \sim \epsilon^{-4}\rho\lambda\frac{4\gamma-4\nu}{4\gamma+1}.$$

Bibliography

[1] BELITSER, E., LEVIT, B. (1995). On minimax filtering on ellipsoids. *Math. Meth. Statist.* **4** 259-273

[2] BROWN, L. D., and LOW, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24** 2384–2398.

[3] CHEN, L. H. Y., GOLDSTEIN, L. and SHAO, Q.-M. (2011). *Normal Approximation by Stein's method.* Springer, Heidelberg.

[4] EFROMOVICH, S and PINSKER, M.S. (1984). A learning algorithm for nonparametric filtering. *Automat. Remote Control* **11** 1434–1440.

[5] ERMAKOV, M. S. (1990). Minimax detection of a signal in a white gaussian noise, *Theory Probab. Appl* **35** 667–679.

[6] ERMAKOV, M. S. (2008). Nonparametric hypothesis testing for small type I and type II error probabilities, *Probl. Inform. Transmission* **44** no. 2, 54–74.

[7] GOLUBEV, G. K. (1987). Adaptive asymptotically minimax estimates of smooth signals. *Probl. Inform. Transmission* **23** 57–67.

[8] GOLUBEV, G. K. (1992). Nonparametric estimation of smooth probability densities in L_2. **28** 44–54.

[9] GOLUBEV, G. K. and NUSSBAUM, M. (1990). A risk bound in Sobolev class regression, *Ann. Statist*. **18** 758–778.

[10] INGSTER, Yu. I. (1982). Minimax nonparametric detection of signals in white Gaussian noise. *Probl. Inform. Transmission* **18** , no. 2, p. 61

[11] INGSTER, Y. I. (1993a). Asymptotically minimax hypothesis testing for nonparametric alternatives. I. *Mathematical Methods of Statistics* **2** 85C114.

[12] INGSTER, Y. I. (1993b). Asymptotically minimax hypothesis testing for nonparametric alternatives. II. *Mathematical Methods of Statistics* **2** 171C189.

[13] INGSTER, Y. I. (1993c). Asymptotically minimax hypothesis testing for nonparametric alternatives, III. *Mathematical Methods of Statistics* **2** 249C268.

[14] INGSTER, Y. I. (1998). Minimax detection of a signal for ℓ_n-balls. *Mathematical Methods of Statistics* **7** 401C428.

[15] INGSTER, Yu. I., SUSLINA, I. A. (2003). *Nonparametric Goodness-of-fit Testing under Gaussian models. Lecture Notes in Statistics, 169*. Springer-Verlag, New York.

[16] INGSTER, Yu. I and SUSLINA, I. A. (2004). Nonparametric hypothesis testing for small type I errors. I. *Math. Methods Statist.* **13** no. 4, 409–459.

[17] INGSTER, Yu. I and SUSLINA, I. A. (2005). Nonparametric hypothesis testing for small type I errors. II. *Math. Methods Statist.* **14** no. 1, 28–52.

[18] LEHMANN, E. L. and ROMANO, J. P. (2005). *Testing Statistical Hypotheses.* Springer, NY.

[19] NUSSBAUM, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.* **24** 2399–2430.

[20] NUSSBAUM, M. (1999). Minimax risk: Pinsker bound. In: *Encyclopedia of Statistical Sciences, Update Volume 3*, 451-460 (S. Kotz, Ed.). Wiley, New York.

[21] ROHDE, A. (2008). Adaptive goodness-of-fit tests based on signed ranks. *Ann. Statist*. **36** 1346–1374.

[22] SPOKOINY, V. G. (1996) Adaptive hypothesis testing using wavelets. *Ann. Statist*. **24** 2477–2498.

[23] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, NY

[24] van der VAART, A. W. (1998). *Asymptotic Statistics.* Cambridge Univ. Press

CHAPTER 2

SHARP ASYMPTOTICS FOR RISK BOUNDS IN NONPARAMETRIC TESTING WITH UNCERTAINTY IN ERROR DISTRIBUTIONS

2.1 Introduction

We are interested in the hypothesis testing problems for nonparametric regression. Consider the observations

$$y_i = f(x_i) + \xi_i, \quad x_i \in [0,1], \quad i = 1, 2, ..., n, \tag{2.1.1}$$

where $\{\xi_{i,n}\}$ are independent random variables with zero expectation, and the function f is to be tested. The nonradom points x_i are assumed to be generated by a density g on $[0,1]$ such that

$$\int_0^{x_{i,n}} g(t)dt = i/n.$$

We define some smoothness class of functions. Let $L_2 = L_2(0,1)$ be the Hilbert space of square integrable functions on $[0,1]$ and let $\|\cdot\|$ denote the usual norm therein. Let, for natural m and $f \in L_2$, $D^m f$ denote the derivative of order m in the distributional sense and let

$$W(m) = \{f \in L_2 : \ D^m f \in L_2\}$$

be the corresponding Sobolev space on the unit interval. The Sobolev class of order m and radius M is defined by

$$W(m, M) = \{f \in W(m) : \ \|D^m f\|^2 \leq M\}$$

for given m and $M > 0$. The periodic Sobolev class is

$$\widetilde{W}(m, M) = \{f \in W(m, M) : \ D^j f(0) = D^j f(1), \ j = 0, 1, ..., m-1\}.$$

Let $\{\psi_j, j = 1, ...\}$ be a orthonormal basis such that

$$\widetilde{W}(m, M) = \{f : f = \sum_{j=1}^{\infty} f_j \psi_j, \sum_{j=1}^{\infty} a_j f_j^2 \leq M\},$$

where $a_j \sim (\pi j)^{2m}$.

Define the complement of an open ball in L_2

$$B_\rho = \{f \in L_2 : \ \|f\|^2 \geq \rho\}.$$

We consider the hypothesis testing problem of

$$H_0 : f = 0 \quad \text{against} \quad H_a : f \in W(m, M) \cap B_\rho.$$

If the radius ρ_n tends to zero too quickly, all tests will have tivial asymptotic power; if it tends to zero too slowly, there exists an α-test such that the type II error tends to zero and consistent testing is possible. Brown and Low [3] established the asymptotic equivalence of the regression model to the Gaussian white noise model,

$$dY = f(t)dt + \sigma dW(t),$$

where W(t) is the Browning motion. In the more general framework of the Gaussian white noise model, Ingser [7, 6] established the separation rate of

$$\rho_n \asymp n^{-4m/(4m+1)}$$

for Sobolev ellipsoids, for which the asymptotic type II error is bounded away from 0 and 1. More precisely, define the minimax type II error as

$$\pi_n(\alpha, \rho_n, m, M) := \inf_{\phi_n : E_{n,0}\phi_n \leq \alpha} \sup_{f \in W(m,M) \cap B_\rho} (1 - E_{n,f}\phi_n).$$

If $\rho_n \asymp n^{-4\beta/(4\beta+1)}$ then

$$0 < \underline{\lim_n} \pi_n(\alpha, \rho_n, m, M) \quad \text{and} \quad \overline{\lim_n} \pi_n(\alpha, \rho_n, \beta, M) < 1 - \alpha.$$

Ermakov [4] found the exact asymptotics of the minimax type II error at the separation rate. More precisely, if

$$\rho_n \sim c \cdot n^{-4\beta/(4\beta+1)},$$

for some $c > 0$, then

$$\pi_n(\alpha, \rho_n, m, M) = \Phi(z_\alpha - \sqrt{A(c, \beta, M)/2}) + o(1) \quad \text{as } n \to \infty,$$

where $A(c, \beta, M) = A_0(\beta) M^{-1/(2\beta)} c^{2+1/(2\beta)}$ and $A_0(\beta)$ is the Ermakov's constant

$$A_0(\beta) = \frac{2(2\beta + 1)}{(4\beta + 1)^{1+1/(2\beta)}}. \tag{2.1.2}$$

In the present paper we consider the sharp asymptotics of the minimax type II error for the model (2.1.1) with uncertain error distributions. This notation of uncertainty is related with robustness, e.g. [2]. As shown below, the model giving meaning meaningful results here is one the nonidentical distributed errors. The distributions of ξ_i will still vary in a small neighborhood of some (unknown) central measure Q_0, but will in general be different.

In contrast, it is noteworthy that substantial attention has been devoted to asymptotically minimax estimation for integrated mean square error. For the Sobolev class of problems, it has been possible to improve the results on best obtainable rates of convergence by find the exact asymptotic value of the minimax risk in the class of all estimators. The key original result is due to Pinkser [9] for a filtering problem over ellipsoids in Hilbert space. Nussbaum [8] and Speckman [10] considered the regression model with Gaussian errors. Golubev and Nussbaum [5] extended the sharp asymptotics to non-Guassian regression for which the error distributions are from a neighborhood of some central measure, and may be nonidentical.

2.2 The lower bound

Our main interest is the testing problem for regression with uncertain error distributions. We adopt the same formulation for the uncertain error distributions as in [5].

Let, for distributions Q_0 and Q,

$$H(Q_0, Q) = \left(\int \left((dQ_0)^{1/2} - (dQ)^{1/2} \right)^2 \right)^{1/2}$$

be the Hellinger distance. Consider a sequence τ_n such that

$$\tau_n \to 0, \quad \tau_n n^{1/2} \to \infty \text{ as } n \to \infty.$$

Introduce the set of probability measures on the real line:

$$\mathbb{Q}_n^H = \{Q; H(Q_0, Q) \le \tau_n, E_Q \xi = 0\}. \tag{2.2.3}$$

The central measure Q_0 has zero expectation, second moment σ^2 and fulfills the following regularity condition: If Q_{0t} denotes the shifted measure $Q_{0t} = Q_0(\cdot + t)$, then

$$H(Q_{0t}, Q_0) = O(t) \quad \text{as } t \to 0. \tag{2.2.4}$$

We assume $\alpha \in (0, 1)$.

Theorem 2.2.1. *Assume in the model* (2.1.1), *ξ_i are independent with distribution $Q \in \mathbb{Q}_n^H$, where the central measure Q_0 has zero expectation, second moment σ^2 and fulfills* (2.2.4).

(i) If $\delta^2 = c \cdot n^{-4m/(4m+1)}$, then, for any test ϕ_n satisfying $E_0 \phi_n = \alpha + o(1)$,

$$\underline{\lim}_n \beta(\phi_n) \ge \Phi(z_\alpha - (A/2)^{1/2}),$$

where

$$A = A(M, m, c, \sigma) = \frac{A_0(m)c^{2+1/(2m)}}{\sigma^4 M^{1/(2m)}}, \tag{2.2.5}$$

and $A_0(m) = \frac{2\pi(1+2m)}{(1+4m)^{1+1/(2m)}}$ *is Ermakov's constant.*

(ii) If $\delta^2 = o(n^{-4m/(4m+1)})$. *Then for any test* ϕ_n *satisfying* $E_0\phi_n = \alpha + o(1)$, *we have* $\overline{\lim}_n \beta(\phi_n) = 1 - \alpha$.

2.3 Attainment

A complete argument for attainment is beyond the scope of the paper, but we provide theoretical backings for our claim that the lower bounds are indeed attainable.

Consider first the regression model (2.1.1) with $g \equiv 1$ and normal noise with variance σ^2. It is known that the error bound in Theorem 2.2.1 is attained by a quadratic statistic given in the frequency domain, [4]. In the time domain 2.1.1, this corresponds to a quadratic statistic given by some linear spline smoothing procedure.

In the nonnormal case, when the noise in 2.1.1 is uncorrelated with zero expectation and variance σ^2, the risk behavior of the quadratic statistics mentioned above remains valid. Actually, the proofs show that the error II error depends only on the first two moments of the noise. The noise distribution model in Theorem 2.2.1 ensures that $\mathrm{Var}\xi \sim \sigma^2$. Indeed, for $Q \in \mathbb{Q}_n^H \cap \mathbb{Q}_c^M$, we

have

$$|E_Q x^2 - \sigma^2|^2 = |\int x^2 \, d(Q - Q_0)|^2 \tag{2.3.6}$$

$$\leq \left(\int x^4 ((dQ)^{1/2} + (dQ_0)^{1/2})^2 \right) H^2(Q_0, Q) \tag{2.3.7}$$

$$\leq 4cH^2(Q_0, Q) = o(1). \tag{2.3.8}$$

Thus it obvious that the bound of Theorem 2.2.1 is attainable for $g \equiv 1$ and known M and σ^2.

For adaptation for unknown M, we conjecture the plug-in-type method developed in the last chapter still works, and some sharp asymptotics can be established by an adaptive smoother. But we leave this study to the future.

2.4 Proofs

Consider the Sobolev space $\widetilde{W}_2^m$ with boundary conditions on $[0, 1]$:

$$\widetilde{W}_2^m = \{f \in W_2^m; (D^k f)(0) = (D^k) f(1) = 0, k = 0, ..., m - 1\}.$$

It is a Hilbert subspace of W_2^m with respect to the norm $(\|f\|^2 + \|D^m f\|^2)^{1/2}$. We will make use of the results on the spectral theory of differential operators; see, e.g., Agmon [1].

There exists a basis φ_j, $j = 1, 2, ...$, in $\widetilde{W}_2^m$ such that, if $(\cdot, \cdot)$ denotes the inner product in $L_2(0, 1)$,

$$(\varphi_i, \varphi_j) = \delta_{ij}, \tag{2.4.9}$$

$$(D^m \varphi_i, D^m \varphi_j) = \lambda_j \delta_{ij}, \quad i, j = 1, 2, ..., \tag{2.4.10}$$

where

$$0 < \lambda_1 < \lambda_2 < \cdots$$

and the asymptotics of the eigenvalues λ_j is given by

$$\lambda_j \sim (\pi j)^{2m}, \quad j \to \infty.$$

The boundary conditions ensure that, when the functions φ_j are continued by zero outside [0,1], the functions belong to the Sobolev space of order m on any interval containing [0,1]. Furthermore, this property allows the construction of another orthogonal system in $\widetilde{W}_2^m$ which is obtained by a change of scale. Fix a natural number q. Later we will let q tend to infinity with n. Define functions

$$\varphi_{jkq} = q^{1/2}\varphi_j(qx - k + 1), \quad k = 1, ..., q, j = 1, 2....$$

Each function φ_{jkq} is in $\widetilde{W}_2^m$, has support $[(k-1)q^{-1}, kq^{-1}]$ and

$$(\varphi_{jkq}, \varphi_{ikq}) = \delta_{ij} \tag{2.4.11}$$

$$(D^m\varphi_{ikq}, D^m\varphi_{jkq}) = q^{2m}\lambda_j\delta_{ij}. \tag{2.4.12}$$

Furthermore, fix a natural s and define $W(q, s, M)$ as the intersection of the linear span of φ_{jkq}, $j = 1, ..., s,\ k = 1, ..., q$, with $\widetilde{W}_2^m(M)$. From (2.4.12), we obtain that for $f \in W(q, s, M)$,

$$\|D^m f\|^2 = \sum_{j=1}^{s}\sum_{k=1}^{q} q^{2m}\lambda_j(\varphi_{jkq}, f)^2$$

and obviously $W(q, s, p)$ is nonempty. Restricting f to this set, we reduce the problem to the one of testing the local Fourier coefficients $f_{jkq} = (\varphi_{jkq}, f)$. The indices q and n will frequently be dropped from the notation in the sequel.

By restricting f to the subset $W(q, s, M)$, we achieve that the observations y_i have a structure

$$y_i = \sum_{j=1}^{s}\varphi_{jk}(x_i)f_{jk} + \xi_i, \quad i = 1, 2, ..., n, \tag{2.4.13}$$

where k is uniquely defined by $i \in \mathcal{F}(k) := \{i; x_i \in q^{-1}(k-1, k]\}$. This may be constructed as a collection of q linear regression models, each accounting for observations in the interval $q^{-1}(k-1, k]$ and having s parameters. The parameters f_{jk} satisfy

$$\sum_{j=1}^{s} \sum_{k=1}^{q} q^{2m} \lambda_j f_{jk}^2 \leq M,$$

and

$$\sum_{j=1}^{s} \sum_{k=1}^{q} f_{jk}^2 \geq \delta^2.$$

Let

$$\delta^2 = c \cdot n^{-4m/(4m+1)},$$

$$\lambda = a \cdot n^{-1/(4m+1)},$$

$$q = [b \cdot n^{2/(4m+1)}],$$

where a and b do not depend on n and will be selected later.

Introduce the column vectors, for $k = 1, 2, ..., q$,

$$h_k = n^{1/2}(f_{1k}, ..., f_{sk})', \tag{2.4.14}$$

$$\bar{\varphi}_i = n^{-1/2}(\varphi_{1k}(x_i), .., \varphi_{sk}(x_i))'. \tag{2.4.15}$$

Then the model (2.4.13) transforms to

$$y_i = \bar{\varphi}_i' h_k + \xi_i, \quad i \in \mathcal{F}(k), \tag{2.4.16}$$

for $k = 1, 2, ..., q$. Now we will select the disturbance distributions in $\mathbb{Q}^H \cap \mathbb{Q}_L^M$ in accordance with the method of least favorable parametric subfamilies. Consider a bounded function ψ on $\mathbb{R}$ such that, if u is the identity map in $\mathbb{R}$,

$$\int \psi dQ_0 = 0, \quad \int u\psi dQ_0 = 1.$$

Set $h_k = \lambda^{1/2} g_k$. For $g \in \mathbb{R}^s$, let $Q_i(g)$ be the measure defined by

$$dQ_i(g) = (1 + \lambda^{1/2} \bar{\varphi}_i' g \psi) dQ_0.$$

Let $Q_i^*(g)$ be the shifted measure

$$Q_i^*(g)(\cdot) = Q_i(g)(\cdot + \lambda^{1/2} \bar{\varphi}_i' g).$$

Lemma 2.4.1. *Let τ_n be the sequence in the definition $\mathbb{Q}_n^H$ and let t_n be such that $t_n \to \infty$, $t_n = o(\tau_n n^{(1-r)/2})$ as $n \to \infty$. Then for sufficiently large n, the set of measures $\{Q_i^*(g); \|g\| \le t_n, i \in \{1, 2, ..., n\}\}$ is contained in $\mathbb{Q}_n^H \cap \mathbb{Q}_L^M$.*

Define

$$\kappa_j^2 = \left(1 - (j/s)^{2m}\right)_+, \quad j = 1, 2, ...$$

and introduce the prior distribution on g_{jk} as independent $N(0, \kappa_j^2)$, $j = 1, 2, ...,$. Let ν_k be the product prior on g_k. Obviously $\nu_k, k = 1, 2, ..., q$ are iid. The induced product prior on all f_{jk}, $j = 1, 2, ..., q = 1, 2, ..., q$ will be denoted by Π_n.

Lemma 2.4.2. *For given $c > 0$ and a small number $\tau > 0$, set*

$$a = \left(\frac{c(1+\tau)}{K_1}\right)^{(2m+1)/(2m)} \cdot \left(\frac{K_2}{M(1-\tau)}\right)^{1/(2m)}, \tag{2.4.17}$$

$$b = \frac{1}{s}\left(\frac{MK_1(1-\tau)}{cK_2(1+\tau)}\right)^{1/(2m)}, \tag{2.4.18}$$

where $K_1 = \frac{2m}{2m+1}$ and $K_2 = \frac{2m}{(2m+1)(4m+1)}$. Then, for sufficiently large s, we have $\Pi_n(\widetilde{W}(m, M) \cap B_\rho) \to 1$ as $n \to \infty$.

Consider the corresponding Bayesian testing problem

$$H_0^* : y_i \sim Q_0, \text{ iid.}$$

$$H_1^* : y_i \sim Q_i(g_k), \text{ where } g_k \sim \nu_k, \ i \in \mathcal{F}(k), \ k = 1, 2, ..., q.$$

We have the log likelihood ratio

$$\Lambda = \sum_{k=1}^{q} \log \int \prod_{i \in \mathcal{F}(k)} (1 + \lambda^{1/2} \bar{\varphi}_i' g \psi(y_i)) d\nu(g).$$

Lemma 2.4.3. *Under the previous assumptions,* Λ *converges to* $N(-\Delta/2, \Delta)$ *weakly under* H_0^**, and to* $N(\Delta/2, \Delta)$ *under* H_a^**, where*

$$\Delta = \Delta(s, \tau, \sigma, m, M, c) \to \frac{A}{2}, \text{ as } \tau \to 0 \text{ and } s \to \infty,$$

and A is as defined in (2.2.5)*.*

The limit experiments give the lower bound for the minimax type II error and concludes the proof.

2.4.1 Proof of Lemma 2.4.1

For the expectation,

$$\int u \, dQ_i^*(g) = \int u \, dQ_i(g) - \bar{\phi} i' g = 0.$$

Let $Q^{**}(g)$ be the shifted measure $Q_0(\cdot + \bar{\phi}_i' g)$. Then for the Hellinger distance,

$$H(Q_i^*(g), Q_0) \leq H(Q^*)_i(g), Q_i^{**}(g)) + H(Q_(^{**} g), Q_0) \tag{2.4.19}$$

Here the first term on the right hand side equals $H(Q_i(g), Q_0)$ and can be bounded by

$$O\left((\lambda \bar{\phi}_i g)^{1/2}\right) = O\left((\lambda t_n n^{(r-1)/2})^{1/2}\right) = o(\tau_n^{1/2}).$$

The second term on the right hand side of (2.4.19) can be bounded similarly in view of condition (2.2.4). Hence all $Q_i^*(g)$ are in Q_n^H, for n sufficiently large, $\|g\| \leq t_n$.

For the fourth moment, we have

$$\int u^4 \, dQ_i^*(g) = \int (u - \bar{\phi}_i(g))^4 (1 - + \lambda \bar{\phi}_i' g \, dQ_0 = \int u^4 dQ_0 + O(\bar{\phi}_i' g)),$$

so that all $Q_i^*(g)$ are in $\mathcal{Q}_c^M$ for sufficiently large n. □

2.4.2 Proof of Lemma 2.4.2

First, we show for for sufficiently large s,

$$P(\sum_{j=1}^{s} \sum_{k=1}^{q} f_{jk}^2 < \delta^2) \to 0, \text{ as } n \to \infty.$$

We have

$$E \sum_{j=1}^{s} \sum_{k=1}^{q} f_{jk}^2 = q\lambda n^{-1} \sum_{j=1}^{s} \left(1 - (j/s)^{2m}\right)_+ \tag{2.4.20}$$

$$= abs \cdot n^{-4m/(4m+1)} \cdot \frac{1}{s} \sum_{j=1}^{s} \left(1 - (j/s)^{2m}\right)_+ \tag{2.4.21}$$

$$\tag{2.4.22}$$

Since

$$\lim_{s \to \infty} \frac{1}{s} \sum_{j=1}^{s} \left(1 - (j/s)^{2m}\right)_+ = \int_0^\infty (1 - t^{2m})_+ dt = K_1,$$

we have, for sufficiently large s,

$$\frac{1}{s} \sum_{j=1}^{s} \left(1 - (j/s)^{2m}\right)_+ > (1 - \frac{\tau}{2}) K_1.$$

Then plugging (2.4.17) and (2.4.18) in gives

$$E \sum_{j=1}^{s} \sum_{k=1}^{q} f_{jk}^2 > (1 + \tau)(1 - \tau/2)\delta^2 > (1 + \tau/4)\delta^2. \tag{2.4.23}$$

The variance is

$$\mathrm{Var}\left(\sum_{j=1}^{s}\sum_{k=1}^{q} f_{jk}^2\right) = 2q\lambda^2 n^{-2}\sum_{j=1}^{s}\left(1-(j/s)^{2m}\right)_+^2) = O(n^{-2}),$$

where we used the fact

$$\lim_{s\to\infty}\frac{1}{s}\sum_{j=1}^{s}\left(1-(j/s)^{2m}\right)_+^2) = \int_0^\infty (1-t^{2m})_+^2 dt = K.$$

Therefore, by Chebyshev's inequality, we have

$$P(\sum_{j=1}^{s}\sum_{k=1}^{q} f_{jk} < \delta^2) < \frac{O(n^{-2})}{\delta^4(1+\tau/4)^2} = O(n^{-2/(4m+1)}) \to 0.$$

Second, we show for sufficiently large s

$$P(\sum_{j=1}^{s}\sum_{k=1}^{q} q^{2m}\lambda_j f_{j_k}^2 > M) \to 0.$$

For the mean, we have

$$E\sum_{j=1}^{s}\sum_{k=1}^{q} q^{2m}\lambda_j f_{j_k}^2 = q^{2m+1}\lambda n^{-1}\sum_{j=1}^{s}\lambda_j\left(1-(j/s)^{2m}\right)_+ \tag{2.4.24}$$

$$= ab^{2m+1}s^{2m+1}\cdot\frac{1}{s}\sum_{j=1}^{s}\frac{\lambda_j}{s^{2m}}\left(1-(j/s)^{2m}\right)_+. \tag{2.4.25}$$

Since $\lambda_j \sim (j\pi)^{2m}$ and it is seen that

$$\lim_{s\to\infty}\frac{1}{s}\sum_{j=1}^{s}\frac{\lambda_j}{s^{2m}}\left(1-(j/s)^{2m}\right)_+ = \int_0^\infty t^{2m}\left(1-t^{2m}\right)_+ dt = K_2,$$

we have for sufficiently large s

$$\frac{1}{s}\sum_{j=1}^{s}\frac{\lambda_j}{s^{2m}}\left(1-(j/s)^{2m}\right)_+ < (1+\tau/2)K_2.$$

Combining these with (2.4.17) and (2.4.18) gives

$$E\left(\sum_{j=1}^{s}\sum_{k=1}^{q} q^{2m}\lambda_j f_{j_k}^2\right) < (1-\tau)(1+\tau/2)M < (1-\tau/2)M.$$

The variance is

$$\begin{aligned}\mathrm{Var}\left(\sum_{j=1}^{s}\sum_{k=1}^{q} q^{2m}\lambda_j f_{j_k}^2\right) &= 2\lambda^2 n^{-2} q^{4m+1} s^{4m+1}\frac{1}{s}\sum_{j=1}\frac{\lambda_j^2}{s^{4m}}\left(1-(j/s)^{2m}\right)_+^2 \\ &= O(a^2 b^{4m+1} s^{4m+1} n^{-2/(4m+1)}) \\ &= o(1),\end{aligned}$$

where we used the fact

$$\lim_{s\to\infty}\frac{1}{s}\sum_{j=1}\frac{\lambda_j^2}{s^{4m}}\left(1-(j/s)^{2m}\right)_+^2 = \int_0^\infty (\pi t)^{4m}(1-t^{2m})_+^2 dt = K.$$

Then by Chebyshev's, we have

$$P(\sum_{j=1}^{s}\sum_{k=1}^{q} q^m \lambda_j f_{j_k}^2 > M) \to 0.$$

□

2.4.3 Proof of Lemma 2.4.3

Without loss of generality, assume $d = n/q$ is an integer.

Recall the logarithm likelihood ratio is

$$\Lambda = \sum_{k=1}^{q}\log\int \Pi_{i\mathcal{F}(k)}\left(1+\lambda^{1/2}\phi(y_i)\bar{\varphi}_i' g\right) d\nu(g) \tag{2.4.26}$$

$$= \sum_{k=1}^{q}\log\int\left(1+\sum_{i\in\mathcal{F}(k)}\lambda^{1/2}\phi(y_i)\bar{\varphi}_i' g + \sum_{i>j,\mathcal{F}(k)}\lambda\psi(y_i)\phi(y_j)\bar{\varphi}_i' g g'\bar{\varphi}_i + Rem\right) d\nu(g). \tag{2.4.27}$$

Write $\Phi = (\bar{\varphi}_1,\ldots,\bar{\varphi}_d)'$ as a matrix, and rewrite the quadratic term above as

$$\begin{aligned}&\sum_{i>j,\mathcal{F}(k)}\int \lambda\psi(y_i)\phi(y_j)\bar{\varphi}_i' g g'\bar{\varphi}_i\, d\nu(g) \\ :=&\frac{\lambda}{2}(J_{1,k}-J_{2,k}),\end{aligned}$$

where

$$J_{1,k} = \psi(y_k)'\Phi\left(E_v g g'\right)\Phi'\psi(y_k) \quad \text{and} \quad J_{2,k} = \sum_{i\in\mathcal{F}(k)} \psi(y_i)^2 \bar{\varphi}_i' \left(E_v g g'\right)\bar{\varphi}_i.$$

Recall $E_v g = 0$ and $E_v g g' = R_s = \text{diag}\{\kappa_1, \ldots, \kappa_s\}$. Then the log likelihood can be expanded further as

$$\begin{aligned}\Lambda &= \sum_{k=1}^{q} \log\left(1 + \frac{\lambda}{2}(J_{1,k} - J_{2,k}) + Rem\right) \\ &= \sum_{k=1}^{q}\left(\frac{\lambda}{2}(J_{1,k} - J_{2,k}) - \frac{\lambda^2}{4}(J_{1,k} - J_{2,k})^2 + Rem\right). \end{aligned} \tag{2.4.28}$$

Consider the means and variances of $J_{1,k}$ and $J_{2,k}$ under H_0^*. As $n \to \infty$,

$$\begin{aligned} EJ_{1,k} &= (1 + o(1)) \cdot \text{tr}(\Phi R_s \Phi') \\ &= (1 + o(1)) \sum_{i\in\mathcal{F}(k)} \sum_{j=1}^{s} \frac{1}{n}\phi_{jk}^2(x_i)\kappa_j, \\ &= (1 + o(1)) \sum_{j=1}^{s} \kappa_j, \end{aligned}$$

$$\begin{aligned} \text{Var}(J_{1,k}) &= (1 + o(1))2\text{tr}(\Phi R_s \Phi' \Phi R_s \Phi') \\ &= 2\text{tr}(\Phi R_s^2 \Phi') \\ &= 2(1 + o(1)) \sum_{j=1}^{s} \kappa_j^2, \end{aligned}$$

$$\begin{aligned} EJ_{2,k} &= \sum_{i\in\mathcal{F}(k)} \sum_{j=1}^{s} \frac{1}{n}\phi_{jk}^2(x_i)\kappa_j, \\ &= \sum_{j=1}^{s} \frac{1}{n} \sum_{i\in\mathcal{F}(k)} \phi_{jk}^2(x_i)\kappa_j \\ &= (1 + o(1)) \sum_{j=1}^{s} \kappa_j, \end{aligned}$$

$$\begin{aligned}\mathrm{Var}(J_{2,k}) &= O(d \cdot (\sum_{j=1}^{s} \frac{1}{n}\phi_{jk}(x)\kappa_j)^2) \\ &= O(dq^2/n^2) \\ &= O(1/d) = o(1).\end{aligned}$$

Therefore, by the law of large numbers and the central limit theorem (cf. [11]), under H_0^* the log likelihood ratio converges weakly to $N(-\Delta/2, \Delta)$, and under H_a^* to $N(\Delta/2, \Delta)$, where

$$\begin{aligned}\Delta &= \frac{1}{2}a^2 b \sum_{j=1}^{s} \kappa_j^2 \\ &= \frac{1}{2}\frac{c^{2+1/(2m)} \cdot (1+\tau)^{2-1/(2m)}}{M^{1/(2m)} \cdot (1-\tau)^{-1/(2m)}} \cdot \frac{1}{s}\sum_{j=1}^{s} \kappa_j^2.\end{aligned}$$

Letting $s \to \infty$ and $\tau \to 0$ gives

$$\Delta \to \frac{1}{2}\frac{A_0(m)c^{2+1/(2m)}}{\sigma^4 M^{1/(2m)}}$$

as claimed. □

Bibliography

[1] AGMON, S. (1968). Asymptotic formulas with remainder estimates for eigenvalues of elliptic operators. *Arch. Rational Mech. Anal.*, **28** 165–183.

[2] BERAN, R. (1982). Robust estimation in models for independent nonidentically distributed data. *Ann. Statist.*, **10** 415–428.

[3] BROWN, L. D. and LOW, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.*, **24** 2384–2398. URL http://dx.doi.org/10.1214/aos/1032181159.

[4] ERMAKOV, M. S. (1990). Minimax detection of a signal in Gaussian white noise. *Teor. Veroyatnost. i Primenen.*, **35** 704–715. URL http://dx.doi.org/10.1137/1135098.

[5] GOLUBEV, G. K. and NUSSBAUM, M. (1990). A risk bound in Sobolev class regression. *Ann. Statist.*, **18** 758–778. URL http://dx.doi.org/10.1214/aos/1176347624.

[6] INGSTER, Y. (1984). Asymptotical minimax testing hypotheses on the distribution density of an independent sample. *Zap. Nauchn. Sem. Leningr. Otdel. Mat. Inst. Steklov.*, **136** 74–96.

[7] INGSTER, Y. I. (1982). Minimax nonparametric detection of signals in white Gaussian noise. *Problems Inform. Transmission*, **18** 130–140.

[8] NUSSBAUM, M. (1985). Spline smoothing in regression models and asymptotic efficiency in L_2. *Ann. Statist.*, **13** 984–997. URL http://dx.doi.org/10.1214/aos/1176349651.

[9] PINSKER, M. (1980). Optimal filtration of square-integrable signals in Gaussian noise. *Probl. Inf. Transm.*, **16** 120–133.

[10] SPECKMAN, P. (1985). Spline smoothing and optimal rates of convergence in nonparametric regression models. *Ann. Statist.*, **13** 970–983. URL http://dx.doi.org/10.1214/aos/1176349650.

[11] VAN DER VAART, A. W. (1998). *Asymptotic statistics*, vol. 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

CHAPTER 3

UPS DELIVERS OPTIMAL PHASE DIAGRAM IN HIGH DIMENSIONAL VARIABLE SELECTION

3.1 Introduction

Consider a sequence of regression problems:

$$Y^{(p)} = X^{(p)}\beta^{(p)} + z^{(p)}, \qquad z^{(p)} \sim N(0,\ I_n), \qquad n = n_p. \tag{3.1.1}$$

Here, $X^{(p)}$ is an $n_p \times p$ matrix, where both p and n_p are large but $p > n_p$. The $p \times 1$ 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. For notational simplicity, 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 (e.g., AIC [2], BIC [24], and RIC [13]). This approach selects variables by minimizing the following functional:

$$\frac{1}{2}\|Y - X\beta\|_2^2 + \frac{(\lambda^{ss})^2}{2}\|\beta\|_0, \tag{3.1.2}$$

where $\lambda^{ss} > 0$ is a tuning parameter and $\|\cdot\|_q$ denotes the L^q-norm. The approach has good properties, but the optimization problem (3.1.2) is known to be NP hard, which prohibits the use of the approach when p is large.

In the middle 90's, Tibshirani [26] and Chen *et al*. [6] proposed a trail-breaking approach which is now known as the lasso or the Basis Pursuit. This approach selects variables by minimizing a similar functional, but $\|\beta\|_0$ is re-

placed by $\|\beta\|_1$:

$$\frac{1}{2}\|Y - X\beta\|_2^2 + \lambda^{lasso}\|\beta\|_1. \tag{3.1.3}$$

A major advantage of the lasso is that (3.1.3) can be efficiently solved by the Interior Point method [6] even when p is relatively large. Additionally, in a series of papers (e.g. [9, 10]), 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, 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 data (e.g. Zou [33]). However, this shortcoming is largely due to that these methods are *one-stage* procedures. This calls for a *two-stage* or *multi-stage* procedure.

3.1.1 Screen and Clean

An idea introduced in the 1960's, Screen and Clean has seen a revival recently [30, 12]. This is a two-stage method, where at the first stage, we remove as many irrelevant variables as possible while keeping all relevant ones. At the second stage, we reinvestigate the surviving variables in hope of removing all false positives. The screening stage has the following advantages, some of which are elaborated in the literature:

- *Dimension reduction*. We remove many irrelevant variables, reducing the dimension from p to a much smaller number [12, 30].

- *Correlation complexity reduction*. A variable may be correlated to many other variables, but few of which will survive the screening; it is only correlated with a few other surviving variables.
- *Computation complexity reduction*. Under some conditions (e.g. Section 3.2), surviving variables can be grouped into many small units, each has a size $\leq K$, and correlation between units is weak. These units can be fitted separately, with computational cost $\leq$ # of units $\times 2^K$.

Despite the perceptive vision and philosophical importance in these works [12, 30], substantial vagueness remains: How to screen? How to clean? Is Screen and Clean really better than the lasso and the subset selection? This is where the **U**nivariate **P**enalization **S**creening (UPS) comes in.

3.1.2 UPS

The UPS is a two-stage method which contains an U-step and a P-step. In the U-step, we screen with Univariate thresholding [9] (also known as marginal regression [16] and Sure Screening [12]). 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. The key insight is that, $\mathcal{U}_p(t)$ has the following important properties.

- *Sure Screening (SS)*. With overwhelming probability, $\mathcal{U}_p(t)$ includes all but a negligible proportion of the signals (i.e. nonzero coordinates of β). The

terminology is slightly different from that in [12].

- *Separable After Screening (SAS).* Define a graph where $\{1, 2, \ldots, p\}$ is the set of nodes, and nodes j and k are connected if and only if $|(x_j, x_k)|$ is large (i.e., columns j and k are "significantly" correlated). The SAS property refers to as that, with overwhelming probability, $\mathcal{U}_p(t)$ splits into many disconnected small-size components (a component is a 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 3.1.1. *For any $p \times 1$ vector Y, $Y^{\mathcal{I}_0}$ denotes the $K \times 1$ vector such that $Y^{\mathcal{I}_0}(k) = Y_{i_k}$, $1 \leq k \leq K$. For any $p \times p$ matrix Ω, $\Omega^{\mathcal{I}_0,\mathcal{J}_0}$ denotes the $K \times L$ matrix such that $\Omega^{\mathcal{I}_0,\mathcal{J}_0}(k, \ell) = \Omega(i_k, j_\ell)$, $1 \leq k \leq K, 1 \leq \ell \leq L$.*

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 the following: (a) since $z \sim N(0, I_n)$, $(X'z)^{\mathcal{U}} \sim N(0, (X'X)^{\mathcal{U},\mathcal{U}})$, (b) by the Sure Screening property, $(X'X)^{\mathcal{U},\mathcal{V}}\beta \approx (X'X)^{\mathcal{U},\mathcal{U}}\beta^{\mathcal{U}}$, and (c) 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 original regression problem reduces to many small-size regression problems that can be solved separately, each at a modest computational cost.

In detail, fix two parameters λ^{ups} and u^{ups}. Let $\mathcal{I}_0 = \{i_1, i_2, \ldots, i_K\} \subset \mathcal{U}_p(t)$ be a component, and let μ be a $K \times 1$ vector the coordinates of which are either 0 or u^{ups}. Write $A = (X'X)^{\mathcal{I}_0, \mathcal{I}_0}$ for short. Let $\hat{\mu}(\mathcal{I}_0) = \hat{\mu}(\mathcal{I}_0; Y, X, t, \lambda^{ups}, u^{ups}, p)$ be the minimizer of the functional:

$$\frac{1}{2}((X'Y)^{\mathcal{I}_0} - A\mu)'A^{-1}((X'Y)^{\mathcal{I}_0} - A\mu) + \frac{1}{2}(\lambda^{ups})^2\|\mu\|_0. \tag{3.1.4}$$

Combining all such estimates across different components of $\mathcal{U}_p(t)$ gives the UPS estimator, 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 $(t, \lambda^{ups}, u^{ups})$. In many cases, the performance of the UPS is relatively insensitive to the choice of t, as long as it falls in a certain range. The parameter λ^{ups} has a similar role to those of the lasso and the subset selection, but there is a major difference: the former can be conveniently estimated using the data, whereas how to set the latter remains an open problem. See Section 3.2 for more discussion.

We are now ready to answer the questions raised in the end of Section 3.1.1: UPS indeed has advantages over the lasso and the subset selection. In Sections 3.1.3-3.1.7, we establish a theoretic framework and investigate these procedures closely. The main finding is the following: for a wide range of design matrices X, the Hamming distance of the UPS achieves the optimal rate of convergence. In contrast, the lasso and the subset selection may be rate non-optimal, even for very simple design matrices.

3.1.3 Sparse signal model and universal lower bound

We model β by

$$\beta_j \overset{iid}{\sim} (1-\epsilon)\nu_0 + \epsilon\pi, \qquad 0 < \epsilon < 1, \qquad 1 \le j \le p, \tag{3.1.5}$$

where ν_0 is the point mass at 0 and π is a distribution that has no mass at 0. We use p as the driving asymptotic parameter, and allow (ϵ, π) to depend on p. Fix $0 < \vartheta < 1$ and recall that s_p is the number of signals. We calibrate

$$\epsilon = \epsilon_p = p^{-\vartheta}, \qquad \text{so that } s_p \sim p\epsilon_p = p^{1-\vartheta}. \tag{3.1.6}$$

For any variable selection procedure $\hat{\beta} = \hat{\beta}(Y|X)$, we measure the loss by the Hamming distance

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

where $\mathrm{sgn}(0) = 0$. In the context of variable selection, the Hamming distance is a natural choice for loss function. While the focus of this paper is on selection error where we use L_0-loss, the idea can be extended to the estimation setting where we use L_q-loss ($0 < q < \infty$), but we have to perform an additional step of least square fitting after the selection.

Somewhat surprisingly, there is a lower bound for the Hamming distance that holds for all sample size n and design matrix X (and so "universal lower bound"). The following notation is frequently used in this paper.

Definition 3.1.2. *$L_p > 0$ 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 $\lim_{p\to\infty} L_p p^{-\delta} = 0$.*

Now, fixing $r > 0$, we introduce

$$\tau_p = \tau_p(r) = \sqrt{2r\log p}, \tag{3.1.7}$$

and $\lambda_p = \lambda_p(\epsilon_p, \tau_p) = \frac{1}{\tau_p}[\log(\frac{1-\epsilon_p}{\epsilon_p}) + \frac{\tau_p^2}{2}]$. Let $\bar{\Phi} = 1 - \Phi$ be the survival function of $N(0,1)$. The following theorem is proved later.

Theorem 3.1.1. *(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 (3.1.6)-(3.1.7), and suppose the support of π_p is contained in $[-\tau_p, 0) \cup (0, \tau_p]$. For any fixed n and matrix $X = X^{(p)}$ such that $X'X$ has unit diagonals, $h_p(\hat{\beta}, \beta|X) \geq s_p \cdot [(1-\epsilon_p)\bar{\Phi}(\lambda_p)/\epsilon_p + \Phi(\tau_p - \lambda_p)]$.*

Note that as $p \to \infty$,

$$\frac{1-\epsilon_p}{\epsilon_p}\bar{\Phi}(\lambda_p) + \Phi(\tau_p - \lambda_p) \geq \begin{cases} L_p \cdot p^{-(r-\vartheta)^2/(4r)}, & r > \vartheta, \\ (1+o(1)), & r < \vartheta. \end{cases} \tag{3.1.8}$$

It may seem counter-intuitive that the lower bound does not depend on n, but this is due to the way we normalize X. In the case of orthogonal design (i.e., coordinates of X and iid from $N(0, 1/n)$), the lower bound can be achieved by either the lasso or marginal regression [16]. Therefore, the orthogonal design is among the best in terms of the error rate.

Theorem 3.1.1 says that if we have $p^{1-\vartheta}$ signals and 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, 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 it may not be tight for nonorthogonal X. Fortunately, it turns out that in many interesting cases, the lower bound is tight. To facilitate the analysis, we invoke the random design model.

3.1.4 Random design, connection to Stein's normal means model

Write $X = (x_1, x_2, \ldots, x_p) = (X_1, X_2, ..., X_n)'$. We model X_i as iid samples from a p-variate zero-mean Gaussian distribution,

$$X_i \overset{iid}{\sim} N(0, \frac{1}{n}\Omega). \tag{3.1.9}$$

The $p \times p$ matrix $\Omega = \Omega^{(p)}$ is unknown but for simplicity we assume 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

$$n = n_p = p^\theta. \tag{3.1.10}$$

Note that $s_p \ll n_p \ll p$ as $p \to \infty$. For successful variable selection, it is almost necessary to have $s_p \ll n_p$ [9]. Also, denoting the distribution of X by $F = F_p$, note that for any variable selection procedure, the *overall Hamming distance* is $\mathrm{Hamm}_p(\hat{\beta}, \beta) = E_F[h_p(\hat{\beta}|X)]$.

Model (3.1.9) is called the *random design model* which may be found in the following application areas.

- *Compressive Sensing.* We are interested in a p-dimensional sparse vector β. We measure n general linear combinations of β and then reconstruct it. For $1 \le i \le n$, choose a $p \times 1$ coefficient vector X_i and observe $Y_i = X_i'\beta + z_i$, where $z_i \sim N(0, \sigma^2)$ is noise. For computational 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. [9, 3]. Model (3.1.9) belongs to 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 $p \times 1$ 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 a Stein's normal means model $W \sim N(\beta, \Sigma)$, where $\Sigma = \Omega^{-1}$. To see the point, recall that Model (3.1.1) 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 $X'X\beta \approx \Omega\beta$ and $X'z \approx N(0, \Omega)$, and so that $X'Y \approx N(\Omega\beta, \Omega)$. Therefore, Stein's normal means model can be viewed as an idealized version of the random design model. This suggests that solving variable selection problem opens doors for solving Stein's normal means problem, and vice versa.

3.1.5 Optimality of the UPS

The main results of this paper are Theorems 3.2.1-3.2.2 in Section 3.2. To state such results, we need relatively long preparations. Therefore, we sketch these results below, but leave the formal statements to later. In Model (3.1.1), (3.1.5), and (3.1.9), let (s_p, τ_p, n_p) be as in (3.1.6), (3.1.7), and (3.1.10). Suppose

- 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 $\tau_p = \sqrt{2r\log(p)}$ and η is a constant to be defined later. We suppose $r > \vartheta$, so that successful variable selection is possible; see Theorem 3.1.1.
- Either all coordinates of Ω are positive, or that $r/\vartheta \le 3 + 2\sqrt{2}$ (so that we won't have too many "signal cancellations" [30]).

Fix $0 < q \le (\vartheta + r)^2/(4r)$, and set the tuning parameters $(t, \lambda^{ups}, u^{ups})$ by

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

The main result is that, as $p \to \infty$, the ratio between the Hamming error of the UPS and s_p is no grater than $L_p p^{-(\vartheta - r)^2/(4r)}$. Comparing this with Theorem 3.1.1 gives that, the lower bound is tight and the UPS is rate optimal.

3.1.6 Phase diagram for high dimensional variable selection

The above results reveal a watershed phenomenon as follows. Suppose we have roughly $s_p = p^{1-\vartheta}$ signals. 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 signals.

The phenomenon is best described in the special case where $\pi_p = \nu_{\tau_p}$ is the point mass at τ_p, with $\tau_p = \sqrt{2r\log p}$ as in (3.1.7). 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 regions:

- *Region of No Recovery* $(0 < \vartheta < 1, 0 < r < \vartheta)$. In this region, the Hamming distance of any procedure $\gtrsim s_p$, 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 (e.g. UPS) whose Hamming errors are much larger than 1 but are also much smaller than s_p. 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 (e.g., UPS) that recover all signals with probability ≈ 1.

See Figure 3.1 (left panel) for these regions. Note that the partitions are the same for many choices of Ω. 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 variable selection. The lasso would be optimal if it partitions the phase space in the same way as in the left panel of Figure 3.1. Unfortunately, this is not the case, even for very simple Ω. Below we investigate the case where $X'X$ is a tridiagonal matrix, and identify precisely the regions where the lasso is rate optimal and where it is rate non-optimal. More surprisingly, there is a region in the phase space where the subset selection is also rate non-optimal.

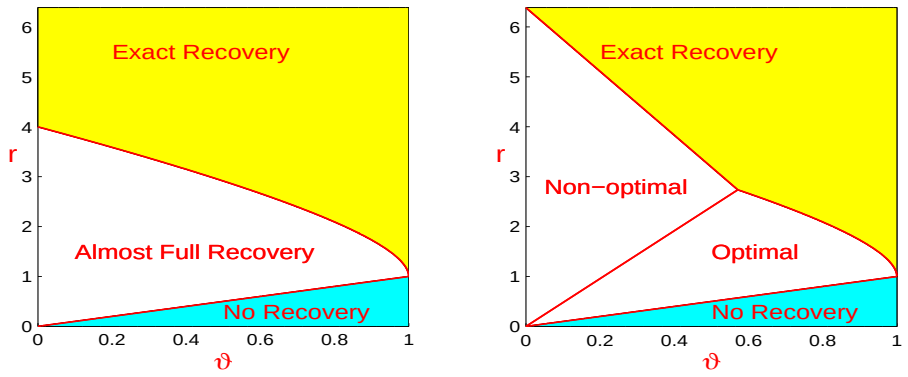

Figure 3.1: Left: Phase diagram. In the yellow region, the UPS recovers all signals with high probability. In the white region, it is possible (i.e., UPS) to recover almost all signals, but impossible to recover all of them. In the cyan region, successful variable selection is impossible. Right: partition of the phase space by the lasso for the tridiagonal model (3.1.11)-(3.1.12) ($a = 0.4$). The lasso is rate non-optimal in the Non-optimal region. The Region of Exact Recovery by the lasso is substantially smaller than that displayed on the left.

3.1.7 Non-optimal region for the lasso

In Section 3.1.7-3.1.8, we temporarily leave the random design model and consider a Stein's normal means model, which is an idealized version of the former. Using an idealized version is mainly for mathematical convenience, but the gained insight is valid in much broader settings: 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

$$\tilde{Y} \equiv X'Y \sim N(\Omega\beta, \Omega), \tag{3.1.11}$$

where β is as in (3.1.5) with $\tau_p = \nu_{\pi_p}$ and $\pi_p = \sqrt{2r\log(p)}$. To further simplify the

study, we fix $a \in (0, 1/2)$ and take Ω as the tridiagonal matrix $T(a)$:

$$T(a)(i, j) = 1\{i = j\} + a \cdot 1\{|i - j| = 1\}, \qquad 1 \le i, j \le p. \tag{3.1.12}$$

Note that in this case, the UPS partitions the phase space optimally.

We now discuss the phase diagram of the lasso. The region $\{(\vartheta, r) : 0 < \vartheta < 1, r > \vartheta\}$ is partitioned into three regions as follows (see Figure 3.1).

- *Non-optimal region*: $0 < \vartheta < 2a(1 + a)^{-1}$ and $\frac{1}{a}(1 + \sqrt{1 - a^2})\vartheta < r < (1 + \sqrt{\frac{1+a}{1-a}})^2(1 - \vartheta)$. 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.
- *Optimal region*: $0 < \vartheta < 1$ and $\vartheta < r < \frac{1}{a}(1 + \sqrt{1 - a^2})\vartheta$ and $r < (1 + \sqrt{1 - \vartheta})^2$. In this region, if additionally $a \ge 1/3$, then the lasso may be rate optimal if the tuning parameter is set ideally. The discussion on the case $0 < a < 1/3$ is tedious so we skip it.
- *Region of Exact Recovery*: $0 < \vartheta < 1$ and $r > (1 + \sqrt{1 - \vartheta})^2$ and $r > (1 + \sqrt{\frac{1+a}{1-a}})^2(1 - \vartheta)$. In this region, if the tuning parameter is set ideally, the lasso may yield exact recovery with high probability. Region of Exactly Recovery by the lasso is substantially smaller than that of the UPS. There is a sub-region in the phase space where the UPS yields exact recovery, but the lasso could not even when the tuning parameter is set ideally.

For discussions in the case where Ω is the identity matrix, compare [16, 28]. The above results are proved in Theorem 3.4.1, where we derive a lower bound for the Hamming errors by the lasso. In a manuscript, 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 (3.1.5) by a closely related model, namely (2.2)-(2.3) in [17]. For these reasons, the non-optimal region of the lasso may be larger than that illustrated in Figure 3.1. The discussion on the exact optimal rate of convergence for the lasso is tedious and we skip it.

Why the lasso is non-optimal? To gain insight, we introduce the term of *fake signal*, a noise coordinate that may look like a signal due to correlation.

Definition 3.1.3. *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 $\beta_j = (\Omega\beta)_j = 0$.*

With the tuning parameter set ideally, the lasso is able to distinguish signals from pure noise, but it does not filter out fake signals efficiently. In the optimal region of the lasso, the number of falsely kept fake signals is much smaller than the optimal rate, so it is negligible; in the non-optimal region, the number becomes much larger than the optimal rate, and so is non-negligible. This suggests that when $X'X$ moves away from the tridiagonal case, the partitions of the regions by the lasso may change, but the non-optimal region of the lasso continues to exist in rather general situations.

The non-optimality of the lasso is largely due to that it is a one-stage method. An interesting question is whether UPS continues to work well if we replace the univariate thresholding by the lasso in the screening stage. The disadvantage of this proposal is that, compared to the univariate thresholding, the lasso is both slower in computation and harder to analyze in theory. Still, one would hope the lasso could perform well in screening.

With that being said, we note that the implementation of the lasso only needs minimal assumption on the model, which makes it very attractive, especially

in complicated situations. In comparison, we need both signal sparsity and graph sparsity to implement the UPS, and how to extend it to more general settings remains unknown. The exploration along this line is continued in our forthcoming manuscripts [20, 21, 11]; see details therein.

3.1.8 Non-optimal region for the subset selection

The discussion on the subset selection is similar to that for the lasso so we keep it brief. Introduce $v_1(a) = \frac{2-\sqrt{1-a^2}}{\sqrt{1-a^2}(1-\sqrt{1-a^2})}$ and $v_2(a) = 2\sqrt{1-a^2} - 1$. Similarly, the phase space partitions into three regions as follows.

- *Non-optimal region:* $0 < \vartheta < \frac{4v_1(a)}{(v_1(a)+1)^2}$ and $v_1(a)\vartheta < r < [\frac{1}{v_2(a)}(\sqrt{1-2\vartheta} + \sqrt{1-2\vartheta+\vartheta v_2(a)})]^2$.
- *Optimal region:* $0 < \vartheta < 1$ and $\vartheta < r < v_1(a)\vartheta$ and $r < (1+\sqrt{1-\vartheta})^2$.
- *Exact Recovery region:* $0 < \vartheta < 1$, $r > (1+\sqrt{1-\vartheta})^2$ and $r > [\frac{1}{v_2(a)}(\sqrt{1-2\vartheta} + \sqrt{1-2\vartheta+\vartheta v_2(a)})]^2$.

See Theorem 3.4.2 for proofs and Figure 3.2 for illustration. Similar to the remarks in Section 3.1.7, the Region of Exact Recovery and the optimal region of the subset selection may be smaller than those illustrated in Figure 3.2.

The reason why the subset selection is non-optimal is almost the *opposite* to that of 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 signal-

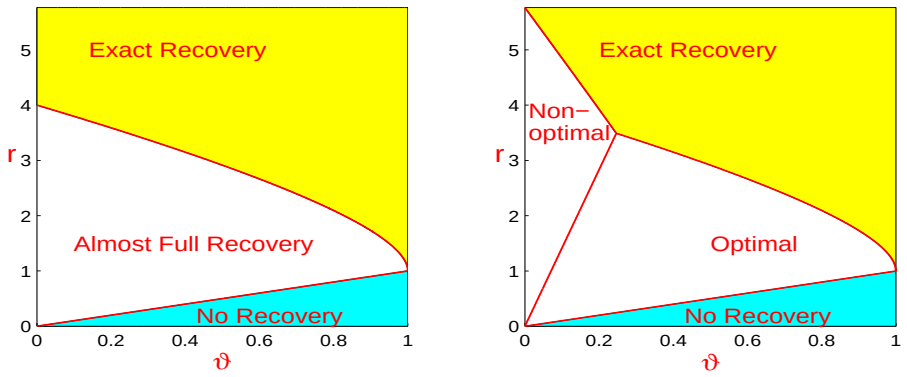

Figure 3.2: Left: a re-display of the left panel of Fig 3.1. Right: partition of the phase space by the subset selection in the tridiagonal model (3.1.11)-(3.1.12) ($a = 0.4$). The subset selection is not rate optimal in the Non-optimal region. The Exact Recovery region by the subset selection is substantially smaller than that of the optimal procedure, displayed on the left.

s when the true signals appear in clusters. These falsely killed signals account for the non-optimality. See Section 3.4.2 for details.

3.1.9 Connection to recent literature

This work is related to recent literature on oracle property [33, 23], but is different in important ways. A procedure has the oracle property if it yields exact recovery. However, exact recovery is rarely seen in applications, especially when $p \gg n$. In many applications (e.g. genomics), a large p usually means that signals are sparse or rare, and a small n usually means signals are weak. For rare and weak signals, exact recovery is usually impossible. Therefore, it is both scientifically more relevant and technically more challenging to compare error rates of different procedures than to investigate when they satisfy the oracle property.

The work is also related to [5, 31] on asymptotic minimaxity, where the lasso was shown to be asymptotic rate optimal in the worst-case scenario. While their results seem to contradict with that in this paper, the difference can be easily reconciled. In the minimax approach, the asymptotic least favorable distribution of β is given by $\beta_j \overset{iid}{\sim} (1-\epsilon_p)\nu_0 + \epsilon_p \nu_{\tau_p}$, where $\epsilon_p = p^{-\vartheta}$, $\tau_p = \sqrt{2r\log p}$ and notably $\vartheta = r$, which corresponds the boundary line of the Region of No Recovery in the phase space (e.g. [31, Page 18-19], [1, Section 3]). This suggests that the minimax approach has limitations: it reduces the analysis to the worst-case scenario, but the worst-case scenario may be outside the range of interest. In our approach, we let (ϑ, r) range freely, and evaluate a procedure based on how it partitions the phase space. Our approach has a similar spirit to that in [10].

The work is also related to the adaptive lasso [33]. 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. Philosophically, we can view the adaptive lasso as a Screen and Clean method. Still, the proposed approach is different from the adaptive lasso in important ways. First, Zou [33] suggested weight choices by the least squares estimate, which is only feasible when p is small. In fact, when $p \gg n$, our results suggest that feasible weights should be very sparse, while the weights suggested by 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-penalty. As pointed out before, the L^1-penalty in the clean step is too loose on fake signals, which prohibits the procedure from being rate optimal.

The work is also related to other multi-stage methods, e.g., the threshold lasso [32] or the LOL [22]. These methods first use the lasso and the OLS for variable selection, respectively, followed by an additional thresholding step. However, by similar argument as in Sections 3.1.7-3.1.8, it is not hard to see that these procedures do not partition the phase diagram optimally.

3.1.10 Contents

In summary, we propose the UPS as a two-stage method for variable selection. 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. On the other hand, the lasso and even the subset selection do not partition the phase space optimally.

The remaining sections are organized as follows. Section 3.2 discusses the UPS procedure and the upper bound for the 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 3.3 discusses a refinement of the UPS for moderately large p. Section 3.4 discusses the behavior of the lasso and the subset selection. Section 3.5 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).

The corresponding paper [19] is to appear in *Annals of Statitics* with the supplementary material for proofs [18].

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

3.2 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. We then discuss the U-step and its Sure Screening and SAS properties. Next, we 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 separately in different subsections, we summarize them as follows. The model we consider is

$$Y = X\beta + z, \qquad z \sim N(0, I_n), \tag{3.2.1}$$

where

$$X_i \overset{iid}{\sim} N(0, \frac{1}{n}\Omega), \quad \beta_j \overset{iid}{\sim} (1-\epsilon_p)\nu_0 + \epsilon_p\pi_p, \quad 1 \le i \le n, \quad 1 \le j \le p. \tag{3.2.2}$$

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

$$\epsilon_p = p^{-\vartheta}, \qquad \tau_p = \sqrt{2r\log p}, \qquad n_p = p^{\theta}, \tag{3.2.3}$$

assuming that

$$\theta < (1 - \vartheta). \tag{3.2.4}$$

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,

$$\vartheta < r < (1+\sqrt{1-\vartheta})^2. \tag{3.2.5}$$

In the phase space, this corresponds to the Region of Almost Full Recovery. The case $r < \vartheta$ corresponds to the Region of No Recovery and is studied in Theorem 3.1.1. The case $r > (1+\sqrt{1-\vartheta})^2$ corresponds to the Region of Exact Recovery. The discussion in this case is similar but is much easier, so we omit it.

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

$$\mathcal{M}_p(\gamma, A) = \{\Omega\text{: } p\times p \text{ correlation matrix, } \sum_{j=1}^{p} |\Omega(i,j)|^\gamma \le A,\ \forall\, 1\le i\le p\}.$$

For any Ω, let $U = U(\Omega)$ be the $p\times p$ matrix satisfying $U(i,j) = \Omega(i,j)1\{i<j\}$, and let $d(\Omega) = \max\{\|U(\Omega)\|_1,\ \|U(\Omega)\|_\infty\}$. Fixing $\omega_0 \in (0,1/2)$, introduce $\mathcal{M}_p^*(\omega_0,\gamma,A) = \{\Omega\in\mathcal{M}_p(\gamma,A)\text{: } d(\Omega)\le\omega_0\}$, and a subset of $\mathcal{M}_p^*(\omega_0,\gamma,A)$,

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

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).

Last, introduce a constant $\eta = \eta(\vartheta, r, \omega_0)$ by

$$\eta = \frac{\sqrt{\vartheta r}}{(\vartheta+r)\sqrt{1+2\omega_0}}\min\Big\{\frac{2\vartheta}{r}, 1-\frac{\vartheta}{r}, \sqrt{2(1-\omega_0)}-1+\frac{\vartheta}{r}\Big\}. \tag{3.2.6}$$

We suppose the support of signal distribution π_p is contained in

$$[\tau_p, (1+\eta)\tau_p], \tag{3.2.7}$$

where $\tau_p = \sqrt{2r\log(p)}$ as in (3.1.7). This assumption is only needed for proving the main lemma of the *P*-step (Lemma 3.6.5), 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 with two-sided signals.

We now discuss the U-step. As mentioned before, the benefits of the U-step are threefold: 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.

3.2.1 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. Fixing a constant $q \in (0, (\vartheta + r)^2/(4r))$, we set the threshold t in the U-step

$$t_p^* = t_p^*(q) = \sqrt{2q \log(p)}. \tag{3.2.8}$$

Lemma 3.2.1. *(Sure Screening). In Model (3.2.1)-(3.2.2), suppose (3.2.3)-(3.2.7) hold, and t_p^* is as in* (3.2.8)*. For sufficiently large p, if $\Omega^{(p)} \in \mathcal{M}_p^+(\omega_0, \gamma, A)$, 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 true if alternatively $\Omega^{(p)} \in \mathcal{M}_p^*(\omega_0, \gamma, A)$ but $r/\vartheta \leq 3 + 2\sqrt{2}$.*

This says that the Hamming errors we make in the U-step are not substantially larger than the optimal rate of convergence, and thus negligible.

3.2.2 The SAS property of the U-step

We need some terminology in graph theory (e.g. [7]). A graph $G = (V, E)$ consists of two finite sets V and E, where V is the set of *nodes*, and E is the set of *edges*. A *component* $\mathcal{I}_0$ of V is a maximal connected subgraph, denoted by $\mathcal{I}_0 \lhd V$. For any node $v \in V$, there is a unique component $\mathcal{I}_0$ such that $v \in \mathcal{I}_0 \lhd V$.

Fix a $p \times p$ 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 graph $G = (V_0, \Omega_0)$. Fix $t > 0$. Recall that $\mathcal{U}_p(t)$ is the set of surviving indices in the U-step:

$$\mathcal{U}_p(t) = \mathcal{U}_p(t, Y, X) = \{j : \ (x_j, Y) \geq t, \ 1 \leq j \leq p\}. \tag{3.2.9}$$

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

Definition 3.2.1. *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.*

Note that if $\mathcal{U}_p(t)$ 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).

Return to Model (3.2.1)-(3.2.2). 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 (3.2.2). However, the SAS property makes more sense if Ω_0 is sparse and known, while Ω is neither. In light of this, we take Ω_0 to be regularized empirical covariance matrix.

In detail, let $\hat{\Omega} = X'X$ be the empirical covariance matrix. Recall that $X = (X_1, X_2, \ldots, X_n)'$ and $X_i \sim N(0, \frac{1}{n}\Omega)$. It is known [4] that there is a constant $C > 0$ such that with probability $1 - o(1/p^2)$, for all $1 \le i, j \le p$,

$$|\hat{\Omega}(i,j) - \Omega(i,j)| \le C\sqrt{\log(p)}/\sqrt{n}. \tag{3.2.10}$$

For large p, $\hat{\Omega}$ is a noisy estimate for Ω, so we regularize it by

$$\Omega^*(i,j) = \hat{\Omega}(i,j)1_{\{|\hat{\Omega}(i,j)| \ge \log^{-1}(p)\}}. \tag{3.2.11}$$

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 (3.2.10); we omit the proof.

Lemma 3.2.2. *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 $\|\Omega^* - \Omega\|_\infty \le C(\log(p))^{-(1-\gamma)}$.*

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

Lemma 3.2.3. *(SAS). Consider Model (3.2.1)-(3.2.2) where (3.2.3)-(3.2.7) hold. Set t_p^* as* (3.2.8)*. As $p \to \infty$, there is a constant K 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).*

3.2.3 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 $\mathcal{J}_0 \lhd \mathcal{U}_p(t)$,

$$\Omega^*(i,j) = 0, \qquad \forall\ i \in \mathcal{I}_0,\ \ j \in \mathcal{J}_0. \tag{3.2.12}$$

Recall that the regression model (3.1.1) 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

$$(\Omega\beta)^{\mathcal{I}_0} \approx \Omega^{\mathcal{I}_0,\mathcal{I}_0}\beta^{\mathcal{I}_0}. \tag{3.2.13}$$

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

$$(\Omega\beta)^{\mathcal{I}_0} - \Omega^{\mathcal{I}_0,\mathcal{I}_0}\beta^{\mathcal{I}_0} = (\Omega^*)^{\mathcal{I}_0,\mathcal{I}_0^c}\beta^{\mathcal{I}_0^c} + (\Omega - \Omega^*)^{\mathcal{I}_0,\mathcal{I}_0^c}\beta^{\mathcal{I}_0^c} = I + II. \tag{3.2.14}$$

First, by Lemma 3.2.2, $|II| \le C\|\Omega - \Omega^*\|_\infty \|\beta\|_\infty = o(\sqrt{\log(p)})$ coordinate-wise, hence II is negligible. Second, by the Sure Screening property, signals that are falsely screened out in the U-step are fewer than $L_p p^{1-(\vartheta+r)^2/(4r)}$, 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 3.6.4 for formal treatment). This, with (3.2.12), implies that $I = 0$, and (3.2.13) follows.

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

$$(X'Y)^{\mathcal{I}_0} \approx N(\Omega^{\mathcal{I}_0,\mathcal{I}_0}\beta^{\mathcal{I}_0},\ \Omega^{\mathcal{I}_0,\mathcal{I}_0}) \tag{3.2.15}$$

that 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.

3.2.4 *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 (3.2.15) with an error rate $\leq L_p p^{-(\vartheta+r)^2/(4r)}$. 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 $\sim \epsilon_p^k$. This motivates us to fit Model (3.2.15) by maximizing the likelihood function $\epsilon_p^k \cdot \exp[-\frac{1}{2}[(X'Y)^{\mathcal{I}_0} - A\mu]' A^{-1} [(X'Y)^{\mathcal{I}_0} - A\mu]]$, subject to $\|\mu\|_0 = k$. Recalling $A = (X'X)^{\mathcal{I}_0,\mathcal{I}_0} \approx \Omega^{\mathcal{I}_0,\mathcal{I}_0}$, this is proportional to the density of $(X'Y)^{\mathcal{I}_0}$ in (3.2.15), hence the name of Penalized MLE. Recalling $\epsilon_p = p^{-\vartheta}$ and $\lambda_p^{ups} = \sqrt{2\vartheta \log p}$, it is equivalent to minimizing

$$[(X'Y)^{\mathcal{I}_0} - A\mu]' A^{-1} [(X'Y)^{\mathcal{I}_0} - A\mu] + (\lambda_p^{ups})^2 \cdot \|\mu\|_0. \tag{3.2.16}$$

Unfortunately, (3.2.16) 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 3.1. In Theorem 3.2.1 below, we show that this procedure obtains the desired rate of convergence provided that u^{ups} is properly set.

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, as the main goal is to separate nonzero coordinates of $\beta^{\mathcal{I}_0}$ from the zero ones.

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 (3.2.15) 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 (e.g. [29]); a noisy estimate of π_p may hurt rather than help in fitting Model (3.2.15).

3.2.5 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 $\mathcal{U}_p(t) = \{j : 1 \le j \le p,\ (x_j, Y) \ge t\}$.
- Obtain Ω^* as in (3.2.11), 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 (3.2.16), 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 $1 \le j \le p$, 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 $\hat{\beta}(Y, X; t, \lambda^{ups}, u^{ups})$. We have the following theorem.

Theorem 3.2.1. *Consider Model (3.2.1)-(3.2.2) where (3.2.3)-(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 $p \to \infty$, $\mathrm{Hamm}_p(\hat{\beta}^{ups}(Y, X; t_p^*, \lambda_p^{ups}, u_p^{ups}), \vartheta, r, \Omega^{(p)}) \leq L_p \cdot s_p \cdot p^{-\frac{(r-\vartheta)^2}{4r}}$. *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 3.1.1. Therefore, both bounds are tight and the UPS is rate optimal.

3.2.6 Tuning parameters of the UPS

The UPS uses three tuning parameters $(t_p^*, \lambda_p^{ups}, u_p^{ups})$. In this section, we show that under certain conditions, the parameters $(\lambda_p^{ups}, u_p^{ups})$ can be estimated from the data.

In detail, recall that $\tilde{Y} = X'Y$. For $t > 0$, introduce $\bar{F}_p(t) = \frac{1}{p}\sum_{j=1}^p 1\{\tilde{Y}_j > t\}$ and $\mu_p(t) = \frac{1}{p}\sum_{j=1}^p \tilde{Y}_j \cdot 1\{\tilde{Y}_j > t\}$. 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

$$2\delta_0(1+\eta) - 1 \leq \vartheta/r, \qquad \text{so that} \qquad \delta_0^2(1+\eta)^2 r < \frac{(\vartheta + r)^2}{4r}. \tag{3.2.17}$$

Let $\mu_p^*(\pi_p)$ be the mean of π_p. The following is proved later.

Lemma 3.2.4. *Fix q such that $\max\{\delta_0^2(1+\eta)^2 r, \vartheta\} < q \le (\vartheta + r)^2/(4r)$, and let $t_p^* = \sqrt{2q\log p}$. Suppose the conditions in Theorem 3.2.1 hold. As $p \to \infty$, with probability of $1 - o(1/p)$,*

$$|[\bar{F}_p(t_p^*)/\epsilon_p] - 1| = o(1), \quad \text{and} \quad |[\mu_p(t_p^*)/(\epsilon_p \mu_p^*(\pi_p))] - 1| = o(1). \tag{3.2.18}$$

Motivated by Lemma 3.2.18, we propose to estimate (λ^{ups}, u^{ups}) by

$$\hat{\lambda}_p^{ups} = \hat{\lambda}_p^{ups}(q) = \sqrt{-2\log(\bar{F}_p(t_p^*))},\ \hat{u}_p^{ups} = \hat{u}_p^{ups}(q) = \mu_p(t_p^*)/\bar{F}_p(t_p^*). \tag{3.2.19}$$

Theorem 3.2.2. *Fix q such that $\max\{\delta_0^2(1+\eta)^2 r, \vartheta\} < q \le (\vartheta + r)^2/(4r)$, and let $t_p^* = \sqrt{2q\log p}$. Suppose the conditions of Theorem 3.2.1 hold. As $p \to \infty$, if additionally $\mu_p^*(\pi_p) \le (1 + o(1))\tau_p$, then $\mathrm{Hamm}_p(\hat{\beta}^{ups}) \le L_p \cdot s_p \cdot p^{-(r-\vartheta)^2/(4r)}$.*

As a result, t_p^* is the only tuning parameter needed by the UPS. By Theorem 3.2.2, 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 3.5 confirm this for finite p. The numerical study also suggests that the lasso is comparably more sensitive to its tuning parameter λ^{lasso}.

3.2.7 Discussions

While the conditions in Theorems 3.2.1-3.2.2 are relatively strong, the key idea of the paper applies to much broader settings. The success of UPS attributes to the interaction of the signal sparsity and graph sparsity, which can be found in many applications (e.g. Compressive Sensing, Genome-wide Association Study (GWAS)).

In the forthcoming papers [11, 20, 21], we revisit the key idea of this paper, and extend our results to more general settings. However, the current paper is different from [11, 20, 21] in important ways. First, the focus of [11] is on ill-posed regression models and change-point problems, and the focus of [21] is on Ising model and network data. Second, the current paper uses the so-called "phase diagram" as a new criterion for optimality (e.g., [10]), and [20] uses the more traditional "asymptotic minimaxity" as the criterion for optimality. Due to the complexity of the problem, one type of optimality usually does not imply the other. The current paper and [20] have very different targets, objectives, and underlying mathematical techniques, and the results in either one can not be deduced from the other.

The current paper is new in at least two aspects. First, given that marginal regression is a widely used method but is not well justified, this paper shows that marginal regression can actually work, provided that an additional cleaning stage is performed. Second, it shows that L^0-penalization method—the target of many relaxation methods—is non-optimal, even in very simple settings and even when the tuning parameter is ideally set.

3.3 A refinement for moderately large p

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 Model (3.1.1) is closely related to the following model:

$$X'Y = X'X\beta + X'z, \qquad z \sim N(0, I_n), \tag{3.3.1}$$

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

$$X'Y \approx \Omega\beta + N(0, \Omega) \qquad \Longleftrightarrow \qquad \Omega^{-1}X'Y \approx N(\beta, \Omega^{-1}). \tag{3.3.2}$$

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. [29]). When Ω is not diagonal, Stein's normal means model is intrinsically a regression problem. To see how close Models (3.3.1) and (3.3.2) are, write

$$X'Y = [\Omega\beta + \frac{\sqrt{n}}{\|z\|}X'z] + [(X'X - \Omega)\beta + (\frac{\|z\|}{\sqrt{n}} - 1)\frac{\sqrt{n}}{\|z\|}X'z] = I + II. \tag{3.3.3}$$

First, note that $I \sim N(\Omega\beta, \Omega)$. For II, we have the following lemma.

Lemma 3.3.1. *Consider Model (3.2.1)-(3.2.2) where (3.2.2)-(3.2.4) hold. As $p \to \infty$, there is a constant $C > 0$ such that except for a probability of $o(1/p)$,*

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

It follows that $|II| \leq C\sqrt{2\log(p)} \cdot p^{-[\theta-(1-\vartheta)]/2}$ coordinate-wise. Therefore, *asymptotically*, Models (3.3.1) and (3.3.2) have negligible difference. However, when p is moderately large, the difference between Models (3.3.1) and (3.3.2) may be non-negligible. In Table 3.1, we tabulate the values of $\sqrt{2\log(p)} \cdot p^{-[\theta-(1-\vartheta)]/2}$, which are relatively large for moderately large p.

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

This says that, for moderately large p, the random design model is much noisier than Stein's normal means model. As a result, 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, and it is hard to clean all of them in the P-step. To see how the problem can be fixed, we write

$$X'X\beta = (X'X - \Omega^*)\beta + \Omega^*\beta. \tag{3.3.4}$$

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 $p \times 1$ 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: (1) Run the UPS and obtain an estimate of β, say, $\hat{\beta}$. Let $W^{(0)} = X'Y$ and $\hat{\beta}^{(0)} = \hat{\beta}$. (2) 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 3.5 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.

3.4 Understanding the lasso and the subset selection

In this section, we show that there is a region in the phase space where the lasso is rate non-optimal (similarly for subset selection). We use a Stein's normal

means model instead of the random design model (as the goal is to understand the non-optimality of these methods, focusing on a simpler model enjoys mathematical convenience, yet is also sufficient; see Section 3.1.7).

To recap, the model we consider in this section is $\tilde{Y} \sim N(\Omega\beta, \Omega)$, where $\tilde{Y}$ is the counterpart of $X'Y$ in the random design model. Fix $a \in (-1/2, 1/2)$. As in Section 3.1.7, we let Ω be the tridiagonal matrix as in (3.1.12), and π_p be the point mass at $\tau_p = \sqrt{2r\log p}$. In other words,

$$\beta_j \overset{iid}{\sim} (1-\epsilon_p)\nu_0 + \epsilon_p \nu_{\tau_p}, \qquad \epsilon_p = p^{-\vartheta}, \qquad \tau_p = \sqrt{2r\log p}. \tag{3.4.1}$$

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 neither the lasso nor the subset selection is optimal. To shed light, we first take a heuristic approach below. Formal statements are given later.

3.4.1 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 (3.4.1), true signals may appear as singletons, pairs, triplets, etc., but singletons are the most common and therefore have the major effect. For each signal singleton, since Ω is tridiagonal, we have two fake signals, one to the left and one to the right. Given a site j, $1 \le j \le p$, 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 [29], where the expected sum of Type I and Type II errors is

$$p[(1-\epsilon_p)\bar{\Phi}(\lambda^{lasso}) + \epsilon_p \Phi(\lambda^{lasso} - \tau_p)]. \tag{3.4.2}$$

Here, $\bar{\Phi} = 1 - \Phi$ is the survival function of $N(0,1)$. In (3.4.2), fixing $0 < q < 1$ and taking $\lambda^{lasso} = \lambda_p^{lasso} = \sqrt{2q\log(p)}$, the expected sum of errors is

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

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

Return 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 $\frac{\vartheta+r}{2r}\tau_p$. This is confirmed in Lemma 3.4.2 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 depending on p that has the following effects: (a) If $\tilde{Y}_j$ is a pure noise, and there is no signal within a distance of K from it, then either $\hat{\beta}_j^{lasso} = 0$, or $\hat{\beta}_j^{lasso} \neq 0$ but $\hat{\beta}_{j\pm1}^{lasso} = 0$, and (b) If $\tilde{Y}_j$ is a signal singleton, and there is no other signal within a distance of K from it, 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. These heuristics are justified in [] (we use such heuristics to provide insight, but not for proving results below).

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, $(\hat{\beta}^{lasso})^{\mathcal{I}_0}$, the restriction of $\hat{\beta}^{lasso}$ to $\mathcal{I}_0$, is the solution of the following small-size minimization problem:

$$\frac{1}{2}\mu'(\Omega^{\mathcal{I}_0,\mathcal{I}_0})\mu - \mu'\tilde{Y}^{\mathcal{I}_0} + \lambda^{lasso}\|\mu\|_1, \qquad \text{where } \mu \text{ is a } k \times 1 \text{ vector.} \tag{3.4.3}$$

See Definition 1.1. Two special cases are noteworthy. First, $\mathcal{I}_0 = \{j\}$, and the solution of (3.4.3) is given by $\hat{\beta}_j^{lasso} = \mathrm{sgn}(\tilde{Y}_j)(|\tilde{Y}_j| - \lambda^{lasso})^+$, which is the soft-thresholding [29]. Second, $\mathcal{I}_0 = \{j-1, j\}$. We call the solution of (3.4.3) in this case the *bivariate lasso*. We have the following lemma, where all regions *I-IIId* are illustrated in Figure 3.3 (x-axis is $\tilde{Y}_{j-1}$, y-axis is $\tilde{Y}_j$).

Lemma 3.4.1. *Denote $\lambda = \lambda^{lasso}$. The solution of the bivariate lasso $(\hat{\beta}_{j-1}^{lasso}, \hat{\beta}_j^{lasso})$ is given by $(\hat{\beta}_{j-1}^{lasso}, \hat{\beta}_j^{lasso}) = (\mathrm{sgn}(\tilde{Y}_{j-1})(|\tilde{Y}_{j-1}| - \lambda)^+,\ \mathrm{sgn}(\tilde{Y}_j)(|\tilde{Y}_j| - \lambda)^+)$ if $(\tilde{Y}_{j-1}, \tilde{Y}_j)$ is in Regions I, IIa-IId, and $(\hat{\beta}_{j-1}^{lasso}, \hat{\beta}_j^{lasso}) = \frac{1}{1-a^2}(Z_{j-1} - aZ_j,\ Z_j - aZ_{j-1})$ if $(\tilde{Y}_{j-1}, \tilde{Y}_j)$ is in Regions IIIa-IIId. Here, $Z_{j-1} = \tilde{Y}_{j-1} - \lambda$ if $(\tilde{Y}_{j-1}, \tilde{Y}_j)$ is in Regions IIIa, IIId, and $Z_{j-1} = \tilde{Y}_{j-1} + \lambda$ otherwise; $Z_j = \tilde{Y}_j - \lambda$ if $(\tilde{Y}_{j-1}, \tilde{Y}_j)$ is in Regions IIIa, IIIb, and $Z_j = \tilde{Y}_j + \lambda$ otherwise.*

In the white region of Figure 3.3, 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 3.4.1 is proved later.

As a result, 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 (univariate) 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).

Noting that $\tilde{Y}_j \sim N(0,1)$ if it is a pure noise and $\tilde{Y}_j \sim N(\tau_p, 1)$ if it is a signal singleton, the sum of Type I and Type II errors is $L_p p[P(N(0,1) \geq \lambda_p^{lasso}) + \epsilon_p P(N(\tau_p, 1) < \lambda_p^{lasso})] = L_p p[\bar{\Phi}(\lambda_p^{lasso}) + \epsilon_p \Phi(\lambda_p^{lasso} - \tau_p)]$. Also, 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 $L_p p \cdot P(\beta_{j-1} = 0, \beta_j = \tau_p, (\tilde{Y}_{j-1}, \tilde{Y}_j)' \in$ Regions *IIIa* or *IIIb*$) \sim L_p p\epsilon_p \cdot P(W \in$ Regions *IIIa* or *IIIb*$)$. Therefore, the sum of three types of errors is

$$L_p p \cdot [\bar{\Phi}(\lambda_p^{lasso}) + \epsilon_p \Phi(\lambda_p^{lasso} - \tau_p) + \epsilon_p P(W \in \text{Regions } IIIa \text{ or } IIIb)], \tag{3.4.4}$$

which can be conveniently evaluated. Note that the sum of Type I and Type II errors in the correlated case is the same as that in the uncorrelated case, which 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. Unfortunately, in certain regions of the phase space, the Type III error can be significantly larger than the optimal rate. 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, which is the culprit for its non-optimality.

For short, write $\mathrm{Hamm}_p(\hat{\beta}^{lasso}(\lambda_p^{lasso})) = \mathrm{Hamm}(\hat{\beta}^{lasso}(\lambda_p^{lasso}); \epsilon_p, \tau_p, a)$. The following lemma confirms the above heuristics.

Lemma 3.4.2. *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 $p \to \infty$,*

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

The exponent on the right-hand side is minimized at $q = (\vartheta + r)^2/(4r)$ when $r < [(1 + \sqrt{1-a^2})/|a|]\vartheta$ and $q = (1 + |a|)(1 - \sqrt{1-a^2})r/(2a^2)$ when $r > [(1 + \sqrt{1-a^2})/|a|]\vartheta$, where we note that $r < [(1 + \sqrt{1-a^2})/|a|]\vartheta$ and $r > [(1 + \sqrt{1-a^2})/|a|]\vartheta$ 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). Combining this with Lemma 3.4.2 gives the following theorem, the proof of which is omitted.

Theorem 3.4.1. *Set $\lambda_p^{lasso} = \sqrt{2q \log p}$. For all choices of $q > 0$, the error rate of the lasso satisfies $\mathrm{Hamm}_p(\hat{\beta}^{lasso}(\lambda_p^{lasso})) \geq L_p \cdot s_p \cdot p^{-\frac{(\vartheta-r)^2}{4r}}$ when $r/\vartheta < (1 + \sqrt{1-a^2})/|a|$ and*

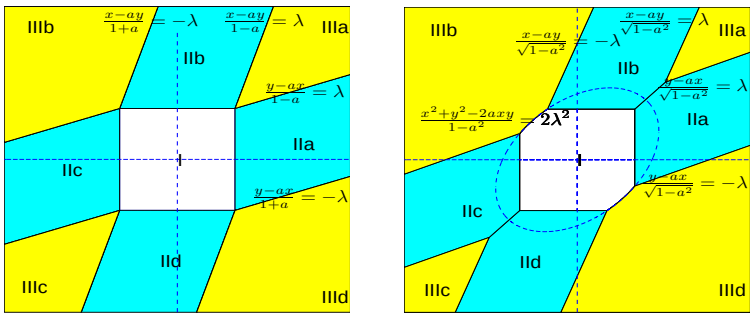

Figure 3.3: Partition of regions as in Lemma 3.4.1 (left) and in Lemma 3.4.3 (right).

$$\text{Hamm}_p(\hat{\beta}^{lasso}(\lambda_p^{lasso})) \geq L_p \cdot s_p \cdot p^{\vartheta - \frac{(1-|a|)(1-\sqrt{1-a^2})}{2a^2} r} \text{ when } r/\vartheta > (1 + \sqrt{1-a^2})/|a|.$$

In [] we show that when $r/\vartheta \leq 3 + 2\sqrt{2}$, the lower bound in Theorem 3.4.1 is tight. The proofs are relatively long, so we leave the details to [].

3.4.2 Understanding subset selection

The discussion is similar, so we keep it brief. Fix $1 \leq j \leq p$. The major errors that subset selection makes are the following (Type III is defined differently from that in the preceding section):

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

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, $\hat{\beta}^{lasso}_{j\pm1} = 0$, and the subset selection acts on site j as hard thresholding [29]: $\hat{\beta}^{ss}_j = \tilde{Y}_j \cdot 1\{|\tilde{Y}_j| \geq \lambda^{ss}\}$. Recall that $\tilde{Y}_j \sim N(0,1)$ if it is a pure noise, and $\tilde{Y}_j \sim N(\tau_p, 1)$ if it is a signal singleton. Take $\lambda^{ss} = \lambda^{ss}_p = \sqrt{2q\log p}$ as before. Similarly, the expected sum of Type I and Type II errors is

$$L_p p[\bar{\Phi}(\lambda^{ss}_p) + p^{-\vartheta}\Phi(\lambda^{ss}_p - \tau_p)] = \begin{cases} L_p(p^{1-q} + p^{1-\vartheta-(\sqrt{q}-\sqrt{r})^2}), & \text{if } 0 < q < r, \\ L_p(p^{1-q} + p^{1-\vartheta}), & \text{if } q > r. \end{cases} \tag{3.4.5}$$

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, $(\hat{\beta}^{ss}_{j-1}, \hat{\beta}^{ss}_j)'$ is the minimizer of the functional $\frac{1}{2}\beta^2_{j-1} + \frac{1}{2}\beta^2_j + a\beta_{j-1}\beta_j - (\tilde{Y}_{j-1}\beta_{j-1} + \tilde{Y}_j\beta_j) + \frac{(\lambda^{ss}_p)^2}{2}\Big(I\{\beta_{j-1} \neq 0\} + I\{\beta_j \neq 0\}\Big)$. We call the resultant procedure *bivariate subset selection*. The following lemma is proved later, with the regions illustrated in Figure 3.3.

Lemma 3.4.3. *The solution of the bivariate subset selection is given by* $(\hat{\beta}^{ss}_{j-1}, \hat{\beta}^{ss}_j) = (0,0)$ *if* $(\tilde{Y}_{j-1}, \tilde{Y}_j)$ *is in Region I,* $(\hat{\beta}^{ss}_{j-1}, \hat{\beta}^{ss}_j) = (\tilde{Y}_{j-1}, 0)$ *if* $(\tilde{Y}_{j-1}, \tilde{Y}_j)$ *is in Regions IIa, IIc,* $(\hat{\beta}^{ss}_{j-1}, \hat{\beta}^{ss}_j) = (0, \tilde{Y}_j)$ *if* $(\tilde{Y}_{j-1}, \tilde{Y}_j)$ *is in Regions IIb, IId, and* $(\hat{\beta}^{ss}_{j-1}, \hat{\beta}^{ss}_j) = (\frac{\tilde{Y}_{j-1} - a\tilde{Y}_j}{1-a^2}, \frac{\tilde{Y}_j - a\tilde{Y}_{j-1}}{1-a^2})$ *if* $(\tilde{Y}_{j-1}, \tilde{Y}_j)$ *is in Regions IIIa-IIId.*

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 there are $O(p\epsilon^2_p)$ 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 $L_p p^{1-(2\vartheta+\min\left\{[(\sqrt{r(1-a^2)}-\sqrt{q})^+]^2, 2[(\sqrt{r(1+a)}-\sqrt{q})^+]^2\right\}}$. Combining with (3.4.5) and Mills' ratio gives the sum of all three types of errors. Formally, writing for short

$\text{Hamm}_p(\hat{\beta}^{ss}(\lambda_p^{ss})) = \text{Hamm}_p(\hat{\beta}^{ss}(\lambda_p^{ss}); \epsilon_p, \tau_p, a)$, we have the following lemma proved later.

Lemma 3.4.4. *Set the tuning parameter* $\lambda_p^{ss} = \sqrt{2q \log p}$. *The Hamming error for the subset selection* $\text{Hamm}_p(\hat{\beta}^{ss}(\lambda_p^{ss}))$ *is at least*

$$\begin{cases} L_p \cdot s_p \cdot p^{-\min\{q-\vartheta,\, \vartheta+[(\sqrt{r(1-a^2)}-\sqrt{q})^+]^2\}}, & \text{if } 0 < q < \frac{(\vartheta+r)^2}{4r}, \\ L_p \cdot s_p \cdot p^{-\min\{(\sqrt{r}-\sqrt{q})^2,\, \vartheta+[(\sqrt{r(1-a^2)}-\sqrt{q})^+]^2\}}, & \text{if } \frac{(\vartheta+r)^2}{4r} < q < r, \\ s_p \cdot (1+o(1)), & \text{if } q > r. \end{cases}$$

The exponents on the right-hand side are minimized at $q = (\vartheta + r)^2/(4r)$ if $r/\vartheta < [2 - \sqrt{1-a^2}]/[\sqrt{1-a^2}(1 - \sqrt{1-a^2})]$, and at $q = [2\vartheta + r(1-a^2)]^2/[4r(1-a^2)]$ if $r/\vartheta > [2 - \sqrt{1-a^2}]/[\sqrt{1-a^2}(1 - \sqrt{1-a^2})]$. As a result, we have the following theorem, the proof of which is omitted.

Theorem 3.4.2. *Set the tuning parameter* $\lambda_p^{ss} = \sqrt{2q \log p}$. *Then for all* $q > 0$, *the Hamming error of the subset selection satisfies*

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

This gives the phase diagram in Figure 3.2, where (ϑ, r) satisfying $r/\vartheta < [2 - \sqrt{1-a^2}]/[\sqrt{1-a^2}(1 - \sqrt{1-a^2})]$ defines the optimal region, and (ϑ, r) with $r/\vartheta > [2 - \sqrt{1-a^2}]/[\sqrt{1-a^2}(1 - \sqrt{1-a^2})]$ defines the non-optimal region. 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 []we further show that in certain regions of the phase space, the lower bound in Theorem 3.4.1 is tight.

3.5 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 $p \times 1$ vector β by $\beta_j \overset{iid}{\sim} (1-\epsilon_p)\nu_0 + \epsilon_p \pi_p$, and an $n_p \times 1$ vector $z \sim N(0, I_{n_p})$. (2) Generate an $n_p \times p$ 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.* [14] (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 in the study for it is the idealized version of the random design model.

Experiment 1. In this experiment, we use 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 (3.1.12) 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 3.2 and Section 3.4, 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 3.2, 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 τ_p exceeds $8, 7$ or 7, respectively, but that of the lasso won't fall below 1 until τ_p exceeds $12, 8$ or 7, respectively. In Section 3.1, 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.

	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 3.2: Hamming errors (Experiment 1). UPS needs weaker signals for exact recovery.

Experiment 2. We use a random design model where $(p, \vartheta, \theta) = (10^4, 0.65, 0.91)$, and $\tau_p \in \{1, 2, \ldots, 7\}$. The experiment contains three parts, 2a–2c. In 2a, we take Ω to be the penta-diagonal matrix $\Omega(i, j) = 1\{i = j\} + 0.4 \cdot 1\{|i - j| = 1\} + 0.1 \cdot 1\{|i - j| = 2\}$. Also, for each τ_p, we set π_p as Uniform$(\tau_p - 0.5, \tau_p + 0.5)$. In 2b, we generate Ω in a way such that it has 4 nonzero off-diagonal elements on average in each row and each column, at locations randomly chosen. Also, for each τ_p, we take π_p to be Uniform$(\tau_p - 1, \tau_p + 1)$. In 2c, we use a non-Gaussian design for X. In detail, first, we generate an $n \times p$ matrix M the coordinates of which are iid samples from Uniform$(-\sqrt{3}, \sqrt{3})$. Second, we generate Ω as in 2b. Last, we let $X = (1/\sqrt{n})M\Omega^{1/2}$. Also, for each τ_p, we take π_p to be the mixture of two uniform distributions $\frac{1}{2}$Uniform$(\tau_p - 0.5, \tau_p + 0.5) + \frac{1}{2}$Uniform$(-\tau_p - 0.5, -\tau_p + 0.5)$. In all

these experiments, the tuning parameters are set the same way as in Experiment 1. The results are reported in Table 3.3, suggesting that the UPS outperforms the lasso almost over the whole range of τ_p.

τ_p	1	2	3	4	5	6	7
2a	**1.01** 1.02	**.96** 1.04	**.82** .97	**.51** .64	**.24** .28	**.09** .10	**.04** .04
2b	**1.00** 1.00	**.98** 1.04	**.84** .96	**.55** .67	**.26** .32	**.10** .12	**.05** .05
2c	**.94** .95	**.90** .91	**.89** .95	**.48** .60	**.18** .27	**.05** .11	**.01** .03

Table 3.3: Ratios between Hamming errors and $p\epsilon_p$ (Experiment 2a-2c). Bold: UPS. Plain: lasso.

Experiment 3. The goal of this experiment is twofold. 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 3.3. Fix $q > 0$. 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 (3.2.19). Theorem 3.2.2 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.

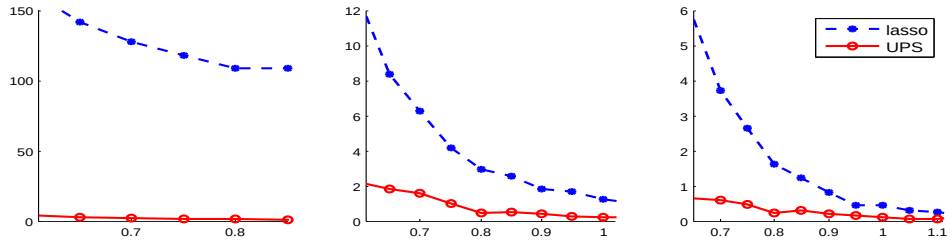

Figure 3.4: Experiment 3a. x-axis: q. y-axis: Hamming error. Left to right: $\vartheta = 0.2, 0.5, 0.65$.

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 $\vartheta = 0.65$, $(\max\{\vartheta, \delta_0^2(1+\eta)^2 r\}, (\vartheta + r)^2/(4r)) = (0.65, 1)$ (similarly for other ϑ), so we let $q \in \{0.7, 0.8, \ldots, 1.1\}$.

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.

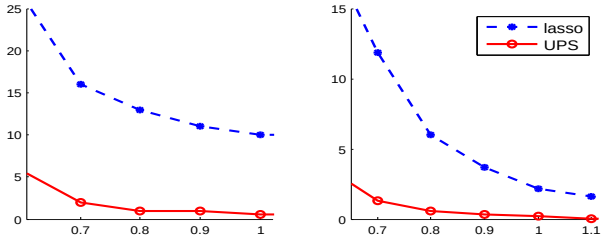

Figure 3.5: Experiment 3b. x-axis: q. y-axis: Hamming error. Left: $\vartheta = 0.5$. Right: $\vartheta = 0.65$.

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 of 3a–3c are reported in Figures 3.4-3.6, correspondingly. These results suggest that, first, the UPS consistently outperforms the lasso, and, second, the UPS is relatively less sensitive to different choices of q.

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 left part of Table 3.4, where the second line displays the ratios between the Hamming errors by the lasso and that by the UPS. Theoretic results (Sections 3.1.7 and 3.4) 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.

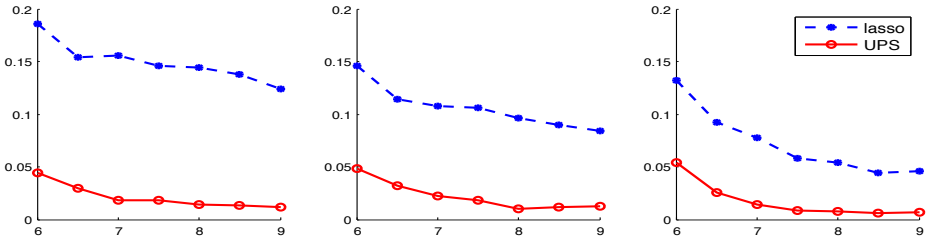

Figure 3.6: Experiment 3c. The x-axis is τ_p, and the y-axis is the ratio between the Hamming error and $p\epsilon_p$. Left to right: $\vartheta = 0.65, 0.5, 0.2$.

In 4b, we illustrate that in a random design model, if we fix p and let 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 also take a Stein's normal means model with the same $(p, \vartheta, r, \Omega, \pi_p)$. The performance of the UPS in both models is reported in the right part of Table 3.4, where the last line is 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	n	300	900	2700	8100	24000
	2.43	5.81	6.25	8.80	10.37		479.25	54.04	12.66	1.08	1.01

Table 3.4: Left: Ratios between the Hamming errors by the UPS and that by the lasso (Experiment 4a). Right: Ratios between the Hamming errors by the UPS for the random design model and that for Stein's normal means model (Experiment 4b).

3.6 Proofs

3.6.1 Proof of Theorem 1.1

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

$$P(\mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j)) \ge P(\beta_j = 0, \hat{\beta}_j \neq 0) + P(\beta_j \neq 0, \hat{\beta}_j = 0). \tag{3.6.1}$$

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,

$$P(\beta_j = 0, \hat{\beta}_j \neq 0) + P(\beta_j \neq 0, \hat{\beta}_j = 0) \ge \frac{1}{2}[1 - \|(1-\epsilon_p)f_0 - \epsilon_p f_1\|_1], \tag{3.6.2}$$

where $\|\cdot\|_1$ denotes the L^1 distance. Combining (3.6.1) and (3.6.2) gives

$$P(\mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j)) \ge \frac{1}{2}[1 - \|(1-\epsilon_p)f_0 - \epsilon_p f_1\|_1]. \tag{3.6.3}$$

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

$$h(y; \tilde{\beta}, \alpha) = h(y, \tilde{\beta}, 0) \cdot e^{\alpha x_j'(y - X\tilde{\beta}) - \alpha^2 x_j' x_j/2}, \tag{3.6.4}$$

and that

$$f_0(y) = \int h(y; \tilde{\beta}, 0) dF(\tilde{\beta}), \qquad f_1(y) = \int h(y; \tilde{\beta}, \alpha) d\pi_p(\alpha) dF(\tilde{\beta}), \tag{3.6.5}$$

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

$$\begin{aligned}
\|(1-\epsilon_p)f_0 - \epsilon_p f_1\|_1 &= \int \Big| \int \Big((1-\epsilon_p)h(y,\tilde{\beta},0) - \epsilon_p h(y,\tilde{\beta},\alpha) \Big) d\pi_p(\alpha) dF(\tilde{\beta}) \Big| dy \\
&\leq \int \int \big| (1-\epsilon_p)h(y,\tilde{\beta},0) - \epsilon_p h(y,\tilde{\beta},\alpha) \big| d\pi_p(\alpha) dF(\tilde{\beta}) dy \\
&= \int \Big[\int \big| (1-\epsilon_p)h(y,\tilde{\beta},0) - \epsilon_p h(y,\tilde{\beta},\alpha) \big| dy \Big] d\pi_p(\alpha) dF(\tilde{\beta}) \\
&= \int H(\tilde{\beta},\alpha) d\pi_p(\alpha) dF(\tilde{\beta}), \qquad (3.6.6)
\end{aligned}$$

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 $\alpha \in [-\tau_p, 0) \cup (0, \tau_p]$,

$$H(\tilde{\beta},\alpha) \leq H(\tilde{\beta},\tau_p). \qquad (3.6.7)$$

Recall that the support of π_p is contained in $[-\tau_p, 0) \cup (0, \tau_p]$. Inserting (3.6.7) into (3.6.6) gives

$$\|(1-\epsilon_p)f_0 - \epsilon_p f_1\|_1 \leq \int H(\tilde{\beta},\tau_p) dF(\tilde{\beta}). \qquad (3.6.8)$$

The following lemma is proved in Section 3.6.1.

Lemma 3.6.1. *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* $\lambda_p = \lambda_p(\epsilon_p,\tau_p) = \frac{1}{\tau_p}\big[\log(\frac{1-\epsilon_p}{\epsilon_p}) + \frac{\tau_p^2}{2}\big]$.

Using Lemma 3.6.1, it follows from (3.6.8) and definitions that

$$\frac{1}{2}\big[1 - \|(1-\epsilon_p)f_0 - \epsilon_p f_1\|_1\big] \geq (1-\epsilon_p)\bar{\Phi}(\lambda_p) + \epsilon_p \Phi(\lambda_p - \tau_p). \qquad (3.6.9)$$

Inserting (3.6.9) into (3.6.3) 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 [29] give that as $p \to \infty$,

$$\frac{1-\epsilon_p}{\epsilon_p}\bar{\Phi}(\lambda_p) = L_p p^{-\frac{(r-\vartheta)^2}{4r}}, \qquad \Phi(\lambda_p - \tau_p) = \begin{cases} L_p p^{-\frac{(r-\vartheta)^2}{4r}}, & r > \vartheta, \\ (1+o(1)), & r < \vartheta, \end{cases} \tag{3.6.10}$$

and the second claim follows. □

Proof of Lemma 3.6.1

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 (3.6.4), $y \in D_p(\tilde{\beta})$ if and only if $\epsilon_p h(y, \tilde{\beta}, \tau_p) > (1-\epsilon_p)h(x, \tilde{\beta}, 0)$. It follows that

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

At the same time,

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

Combining these gives

$$\frac{1}{2}[1 - \int \big|(1-\epsilon_p)h(y,\tilde{\beta},0) - \epsilon_p h(y,\tilde{\beta},\tau_p)\big| dy] = (1-\epsilon_p)\int_{D_p(\tilde{\beta})} h(y,\tilde{\beta},0)dy + \epsilon_p \int_{D_p^c(\tilde{\beta})} h(y,\tilde{\beta},\tau_p)dy. \tag{3.6.11}$$

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

$$\int_{D_p(\tilde{\beta})} h(y,\tilde{\beta},0)dy = P_0(W_j > \lambda_p), \qquad \int_{D_p^c(\tilde{\beta})} h(y,\tilde{\beta},\tau_p)dy = P_1(W_j \le \lambda_p), \tag{3.6.12}$$

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 (3.6.12) gives

$$\int_{D_p(\tilde{\beta})} h(y, \tilde{\beta}, 0)dy = \bar{\Phi}(\lambda_p), \qquad \int_{D_p^c(\tilde{\beta})} h(y, \tilde{\beta}, \tau_p)dy = \Phi(\lambda_p - \tau_p). \tag{3.6.13}$$

The claim follows by inserting (3.6.13) into (3.6.11). □

3.6.2 Proof of Lemma 2.1

Let D_p be the event

$$\{\|(X'X - \Omega)\beta\|_\infty \leq C\|\Omega\| \sqrt{\log(p)}\, p^{-(\theta-(1-\vartheta))/2}, \, |\frac{\|z\|}{\sqrt{n}} - 1| \leq C\sqrt{\log(p)} p^{-\theta/2}\}. \tag{3.6.14}$$

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 (i.e. $O(p^{-c})$ for some constant c). 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 [29].

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

$$P(\beta_j \neq 0, B_j^c) \leq \sum_{k \in S_j, k \neq j} P(\beta_j \neq 0, \beta_k \neq 0) \leq 2\log(p)\epsilon_p^2 = 2\log(p)p^{-2\vartheta}. \tag{3.6.15}$$

Since $r/\vartheta \leq 3 + 2\sqrt{2}$, $2\vartheta \geq (\vartheta + r)^2/(4r)$. Compare (3.6.15) with the desired claim, it is sufficient to show

$$P(x_j'Y < t_p^*, \beta_j \neq 0, B_j) \leq L_p p^{-(\vartheta+r)^2/(4r)}. \tag{3.6.16}$$

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, (3.6.16) follows. □

3.6.3 Proof of Lemma 2.2

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

$$\sum_{j=1}^{p} |\Omega(i,j) - \Omega^*(i,j)| = I + II, \tag{3.6.17}$$

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,

$$I \leq 2\log(p) \max_{1\leq i,j\leq p} \{|\hat{\Omega}(i,j) - \Omega(i,j)|\} \leq L_p\, p^{-\theta/2}. \tag{3.6.18}$$

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

$$II \leq \sum_{\{j:\,|\Omega(i,j)|\leq 2\delta_p\}} |\Omega(i,j)| \leq \sum_{\{j:\,|\Omega(i,j)|\leq 2\delta_p\}} (|\Omega(i,j)|^{\gamma})(|\Omega(i,j)|^{1-\gamma}), \tag{3.6.19}$$

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 (3.6.18)-(3.6.19) into (3.6.17) gives the claim. □

3.6.4 Proof of Lemma 2.3

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

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

The following lemma is proved in Frieze and Molloy [15].

Lemma 3.6.2. *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,

$$P(\mathcal{U}_p(t_p^*) \text{ contains a CG with size } \ell_0) \leq o(1/p). \tag{3.6.20}$$

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 (3.6.20), it is sufficient to show

$$P(\mathcal{U}_p(t_p^*) \text{ contains a CG with size } \ell_0,\ D_p) \leq o(1/p). \tag{3.6.21}$$

Recall that by Lemma 2.2, each row of Ω^* has at most $2\log(p)$ nonzero coordinates. Using Lemma 3.6.2, there are at most $p(2e\log(p))^{\ell_0}$ CG with size ℓ_0. So to show (3.6.21), it is sufficient to show for any fixed CG of size ℓ_0, say $\mathcal{I}_0 = \{i_1, i_2, \ldots, i_{\ell_0}\}$,

$$P(\mathcal{I}_0 \subset \mathcal{U}_p(t_p^*), D_p) \leq o(1/p^2). \tag{3.6.22}$$

We now show (3.6.22). 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

$$\eta'\Omega_0^{-1}\eta \sim \chi^2(\ell_0). \tag{3.6.23}$$

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

$$\|\eta\|^2 \geq \frac{1}{2}\|W\|^2 - \|M\beta\|^2 - rem, \qquad \text{over the event } D_p, \tag{3.6.24}$$

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,

$$\eta' \Omega_0^{-1} \eta \geq \frac{1}{2} \|\eta\|^2. \tag{3.6.25}$$

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

$$\|W\|^2 \geq \ell_0 t_p^{*2} \geq 2q\ell_0 \log(p). \tag{3.6.26}$$

Combining (3.6.24)-(3.6.26) gives that over the event $\{\mathcal{I}_0 \subset \mathcal{U}_p(t_p^*)\} \cap D_p$,

$$\eta' \Omega_0^{-1} \eta \geq \frac{1}{2} \|\eta\|^2 \geq \frac{1}{2} [q\ell_0 \log(p) - \|M\beta\|^2 - rem]. \tag{3.6.27}$$

The following lemma is proved in Section 3.6.4.

Lemma 3.6.3. *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 3.6.3 and (3.6.27),

$$P(D_p \cap \tilde{D}_p) \leq L_p p^{-\vartheta k_0}, \qquad \text{and} \qquad \eta' \Omega_0^{-1} \eta \gtrsim \tfrac{1}{4} q\ell_0 \log(p) \text{ over } D_p \cap \tilde{D}_p^c. \tag{3.6.28}$$

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 (3.6.23) and (3.6.28), it follows from basic statistics that

$$P(\mathcal{I}_0 \subset \mathcal{U}_p(t_p^*), D_p) \leq L_p(p^{-\frac{1}{8} q\ell_0} + p^{-\vartheta k_0}). \tag{3.6.29}$$

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 (3.6.29) $\leq o(1/p^2)$ and then gives (3.6.22). □

Proof of Lemma 3.6.3

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

$$|S| \le 2\ell_0 \log(p). \tag{3.6.30}$$

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 2.2 and assumptions, $\|\Omega^* - \Omega\|_\infty \le C(\log(p))^{-(1-\gamma)}$ and $\|\beta\|_\infty \le (1+\eta)\tau_p$. Therefore, $\|M\beta - M_1\xi\|_\infty \le C(1+\eta)(\log(p))^{-(1-\gamma)}\tau_p$, and

$$\|M\beta - M_1\xi\|^2 \le C(1+\eta)^2\ell_0(\log(p))^{-2(1-\gamma)}\tau_p^2. \tag{3.6.31}$$

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 $\|\xi\|_\infty \le \|\beta\|_\infty \le (1+\eta)\tau_p$,

$$\|M_1\xi\|^2 \le 4\|\xi\|^2 \le 4\|\xi\|_0(1+\eta)^2\tau_p^2. \tag{3.6.32}$$

Combining (3.6.31)–(3.6.32) gives

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

Recall that $\epsilon_p = p^{-\vartheta}$ and $\|\xi\|_0$ is distributed as Binomial$(|S|, \epsilon_p)$ (see (2.2)). Using (3.6.30),

$$P(\|\xi\|_0 \ge k) = \sum_{j=k}^{|S|}\binom{|S|}{j}\epsilon_p^j(1-\epsilon_p)^{|S|-j} \le \sum_{j=k}^{|S|}(2\ell_0\log(p))^j p^{-\vartheta j} \le 2(2\ell_0\log(p))^k p^{-\vartheta k}. \tag{3.6.33}$$

Combining (3.6.33)-(3.6.32), the claim follows by recalling $\tau_p = \sqrt{2r\log p}$. □

3.6.5 Proof of Theorem 2.1

By (2.2), with probability at least $1 - o(\frac{1}{p})$,

$$|(X'X)(i,j) - \Omega(i,j)| \le L_p p^{-\theta/2}, \qquad \forall\, 1 \le i,j \le n. \tag{3.6.34}$$

Fix $K \geq 1$. It is seen that for all connected subgraph $\mathcal{I}_0$ of size $\ell \leq K$,

$$\|(X'X)^{\mathcal{I}_0,\mathcal{I}_0} - \Omega^{\mathcal{I}_0,\mathcal{I}_0}\|_\infty \leq L_p p^{-\theta/2}. \tag{3.6.35}$$

Write for short $\hat{\beta} = \hat{\beta}^{ups}(Y, X; t_p^*, \lambda_p^{ups}, u_p^{ups})$. By definitions, $\mathrm{Hamm}_p(\hat{\beta}, \beta) = E[h_p(\hat{\beta}|X)]$, where $h_p(\hat{\beta}|X) \leq p$ for all X. So the event where X does not satisfy either (3.6.34) or (3.6.35) only has a negligible effect on the claim. All we need to show is that, for any X satisfying (3.6.34)-(3.6.35), $h_p(\hat{\beta}|X) \leq L_p p^{1-(\vartheta+r)^2/(4r)}$, where the right-hand side does not depend on X.

We now show the last inequality. Given X satisfying (3.6.34) and (3.6.35), write $h_p(\hat{\beta}|X) = \sum_{j=1}^p P(\mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j)|X) = I + II$, where $I = \sum_{j=1}^p P(\mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j), j \notin \mathcal{U}_p(t_p^*)|X)$ and $II = \sum_{j=1}^p P(\mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j), j \in \mathcal{U}_p(t_p^*)|X)$. 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 $II \leq L_p p^{1-(\vartheta+r)^2/(4r)}$.

By Lemma 2.3, 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. It is sufficient to show that for all $1 \leq j \leq p$, $P(\mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j), j \in \mathcal{U}_p(t_p^*), A_p) \leq L_p p^{-(\vartheta+r)^2/(4r)}$. 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, $P(\mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j), j \in \mathcal{U}_p(t_p^*), A_p) \leq \sum_{\mathcal{I}_0} P(\mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j), j \in \mathcal{I}_0 \lhd \mathcal{U}_p(t_p^*), A_p)$, 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 2.2, each row of Ω^* has no more than $2\log(p)$ nonzero coordinates. It follows from Lemma 3.6.2 that there are at most $C(2e\log(p))^K$ of such $\mathcal{I}_0$. It remains to show for any fixed connected subgraph $\mathcal{I}_0$ of (V_0, Ω^*) that contains j, $P(\mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j), j \in \mathcal{I}_0 \lhd \mathcal{U}_p(t_p^*)) \leq L_p p^{-(\vartheta+r)^2/(4r)}$.

Introduce the event $B_p(\mathcal{I}_0) = B_p(\mathcal{I}_0, \beta; X, j)$ through its complement $B_p^c(\mathcal{I}_0) = \{$There are indices $i \notin \mathcal{I}_0$ and $k \in \mathcal{I}_0$ 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 $P(j \in \mathcal{I}_0 \lhd \mathcal{U}_p(t_p^*), B_p^c \cap A_p) \leq \sum_i P(\beta_i \neq 0, x_i'Y < t_p^*)$, 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 gives that $P(j \in \mathcal{I}_0 \lhd \mathcal{U}_p(t_p^*), B_p^c \cap A_p) \leq L_p p^{-(\vartheta+r)^2/(4r)}$. Comparing this with what remains, it is sufficient to show $P(\mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j), j \in \mathcal{I}_0 \lhd \mathcal{U}_p(t_p^*), B_p \cap A_p) \leq L_p p^{-(\vartheta+r)^2/(4r)}$.

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 3.6.5.

Lemma 3.6.4. *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 $Q_p = \{\mathrm{sgn}(\beta_j) \neq \mathrm{sgn}(\hat{\beta}_j), j \in \mathcal{I}_0 \lhd \mathcal{U}_p(t_p^*)\} \cap B_p \cap A_p$ to the P-step. 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 the Penalized MLE

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

where the coordinates of μ take values from $\{0, u_p^{ups}\}$, $\lambda_p^{ups} = \sqrt{2\vartheta\log p}$, and $u_p^{ups} = \tau_p = \sqrt{2r\log p}$. By the definition of the UPS, the event Q_p is contained in the

event $\{\text{sgn}(\hat{\mu}(\mathcal{I}_0)) \neq \text{sgn}(\beta^{\mathcal{I}_0}), j \in \mathcal{I}_0 \lhd \mathcal{U}_p(t_p^*), B_p \cap A_p\}$, where $\text{sgn}(\beta)$ is the vector of signs of β. The claim follows from the following lemma, which is proved in Section 3.6.5.

Lemma 3.6.5. *Suppose the conditions of Theorem 2.1 hold. For the event* $Q_p^* = \{\text{sgn}(\hat{\mu}(\mathcal{I}_0; Y, X, t_p^*, \lambda_p^{ups}, u_p^{ups}, p)) \neq \text{sgn}(\beta^{\mathcal{I}_0}), j \in \mathcal{I}_0 \lhd \mathcal{U}_p(t_p^*)\} \cap B_p \cap A_p$. *Fix* $1 \leq j \leq p$. *As* $p \to \infty$, *for any fixed* $\mathcal{I}_0$ *with size* $\leq K$ *that contains* j, $P(Q_p^*) \leq L_p p^{-(\vartheta+r)^2/(4r)} + p^{-2\vartheta}$. *If furthermore all coordinates of* $\Omega^{\mathcal{I}_0,\mathcal{I}_0}$ *are non-negative, then* $P(Q_p^*) \leq L_p p^{-(\vartheta+r)^2/(4r)}$.

□

Proof of Lemma 3.6.4

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

$$(\Omega\beta)^{\mathcal{I}_0} - \Omega^{\mathcal{I}_0,\mathcal{I}_0}\beta^{\mathcal{I}_0} = \Omega^{\mathcal{I}_0,\mathcal{I}_0}\beta^{\mathcal{I}_0} + \Omega^{\mathcal{I}_0,\mathcal{I}_0^c}\beta^{\mathcal{I}_0^c} - \Omega^{\mathcal{I}_0,\mathcal{I}_0}\beta^{\mathcal{I}_0} = \Omega^{\mathcal{I}_0,\mathcal{I}_0^c}\beta^{\mathcal{I}_0^c}. \tag{3.6.36}$$

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

$$(\Omega^*)^{\mathcal{I}_0,\mathcal{I}_0^c}\beta^{\mathcal{I}_0^c} = 0. \tag{3.6.37}$$

Combining (3.6.36)-(3.6.37) 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 2.2,

$$\|(\Omega - \Omega^*)^{\mathcal{I}_0,\mathcal{I}_0^c}\beta^{\mathcal{I}_0^c}\|_\infty \leq \|(\Omega - \Omega^*)^{\mathcal{I}_0,\mathcal{I}_0^c}\|_\infty \cdot \|\beta^{\mathcal{I}_0^c}\|_\infty \leq C\tau_p(\log(p))^{-(1-\gamma)}. \tag{3.6.38}$$

The claim follows. □

Proof of Lemma 3.6.5

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

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

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

$$P(\operatorname{sgn}(\hat{\mu}(\mathcal{I}_0)) \ne \operatorname{sgn}(\beta^{\mathcal{I}_0}), j \in \mathcal{I}_0 \lhd \mathcal{U}_p(t_p^*), B_p \cap A_p \cap \tilde{D}_p) \le L_p p^{-(\vartheta+r)^2/(4r)}. \quad (3.6.39)$$

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 $\hat{\mu}(\mathcal{I}_0)$,

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

By (3.6.35), $\|(X'X)^{\mathcal{I}_0,\mathcal{I}_0} - \Omega^{\mathcal{I}_0,\mathcal{I}_0}\|_\infty$ is algebraically small. So up to a negligible effect,

$$\begin{aligned}&\frac{1}{2}(\tilde{Y}^{\mathcal{I}_0} - \Omega^{\mathcal{I}_0,\mathcal{I}_0}\hat{\mu}(\mathcal{I}_0))'(\Omega^{\mathcal{I}_0,\mathcal{I}_0})^{-1}(\tilde{Y}^{\mathcal{I}_0} - \Omega^{\mathcal{I}_0,\mathcal{I}_0}\hat{\mu}(\mathcal{I}_0)) + \frac{\lambda^2}{2}\|\hat{\mu}(\mathcal{I}_0)\|_0\\ \le&\frac{1}{2}(\tilde{Y}^{\mathcal{I}_0} - \Omega^{\mathcal{I}_0,\mathcal{I}_0}(\beta^*)^{\mathcal{I}_0})'(\Omega^{\mathcal{I}_0,\mathcal{I}_0})^{-1}(\tilde{Y}^{\mathcal{I}_0} - \Omega^{\mathcal{I}_0,\mathcal{I}_0}(\beta^*)^{\mathcal{I}_0}) + \frac{\lambda^2}{2}\|(\beta^*)^{\mathcal{I}_0}\|_0. \quad (3.6.40)\end{aligned}$$

Denote $d = d(\mathcal{I}_0) = \|(\beta^*)^{\mathcal{I}_0}\|_0 - \|\hat{\mu}(\mathcal{I}_0)\|_0$. Reorganizing, it follows from (3.6.40) that

$$((\beta^*)^{\mathcal{I}_0} - \hat{\mu}(\mathcal{I}_0))'\tilde{Y}^{\mathcal{I}_0} \le \frac{1}{2}[d\lambda^2 + ((\beta^*)^{\mathcal{I}_0})'\Omega^{\mathcal{I}_0,\mathcal{I}_0}(\beta^*)^{\mathcal{I}_0} - \hat{\mu}'(\mathcal{I}_0)\Omega^{\mathcal{I}_0,\mathcal{I}_0}\hat{\mu}(\mathcal{I}_0)], \quad (3.6.41)$$

where by Lemma 3.6.4, there is an $|\mathcal{I}_0| \times 1$ vector $\tilde{z} \sim N(0, \Omega^{\mathcal{I}_0,\mathcal{I}_0})$ independent of $\beta^{\mathcal{I}_0}$ such that

$$\tilde{Y}^{\mathcal{I}_0} = \Omega^{\mathcal{I}_0,\mathcal{I}_0}\beta^{\mathcal{I}_0} + \tilde{z} + rem, \qquad \|rem\|_\infty \le o(\sqrt{\log p}). \quad (3.6.42)$$

Now, for notational simplicity, we drop $\mathcal{I}_0$ everywhere in (3.6.40)–(3.6.42). This is a slight misuse of the notations. Note that β and Ω below are low-dimensional. Write

$$\beta - \hat{\mu} = \tau_p(\Delta_1 + \Delta_2), \qquad \text{where} \qquad \Delta_1 = \frac{1}{\tau_p}(\beta^* - \hat{\mu}), \qquad \Delta_2 = \frac{1}{\tau_p}(\beta - \beta^*). \tag{3.6.43}$$

Plug (3.6.42)-(3.6.43) into (3.6.41) and reorganize. We conclude that over the event (3.6.39),

$$-\frac{\Delta_1'\tilde{z}}{\sqrt{\Delta_1'\Omega\Delta_1}} \geq \frac{1}{2\sqrt{\Delta_1'\Omega\Delta_1}}(-d(\vartheta/r)+2\Delta_1'\Omega\Delta_2+\Delta_1'\Omega\Delta_1)\sqrt{2r\log p}+o(\sqrt{\log p}), \tag{3.6.44}$$

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 (3.6.39), $B_{ns} + B_{sn} \geq 1$. 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 (3.6.44), to show (3.6.39), it is sufficient to show

$$\begin{aligned} &p^{-\vartheta(B_{sn}+B_{ss})}\bar{\Phi}\Big(\frac{(-d(\vartheta/r)+2\Delta_1'\Omega\Delta_2+\Delta_1'\Omega\Delta_1)}{2\sqrt{\Delta_1'\Omega\Delta_1}}\sqrt{2r\log p}\Big) \\ &\leq \begin{cases} L_p p^{-\frac{(\vartheta+r)^2}{4r}}, & \text{if } \Omega \text{ only has non-negative coordinates,} \\ L_p p^{-\frac{(\vartheta+r)^2}{4r}} + p^{-2\vartheta}, & \text{if } \Omega \text{ may have negative coordinates,} \end{cases} \end{aligned} \tag{3.6.45}$$

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

First, we consider (3.6.45) 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 (3.6.45) 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 [29]. 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 [29], 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

$$1 \leq B_{sn} + B_{ss} \leq (\vartheta + r)^2/(4\vartheta r), \qquad \text{and} \qquad \text{when } B_{ns} = 0,\ B_{ss} + B_{sn} \geq 2. \quad (3.6.46)$$

The following lemma is proved in Section 3.6.5.

Lemma 3.6.6. *Fix $\omega_0 \in [0, 1/2)$. Suppose that Ω has unit diagonals and only non-negative coordinates, and that*

$$\max\{\|U(\Omega)\|_\infty, \|U(\Omega)\|_1\} \leq \omega_0. \quad (3.6.47)$$

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-Schwarz inequality,

$$|\Delta_1'\Omega\Delta_2| \leq \sqrt{\Delta_1'\Omega\Delta_1}\sqrt{\Delta_2'\Omega\Delta_2}. \tag{3.6.48}$$

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

$$\Delta_2'\Omega\Delta_2 \leq (1+2\omega_0)\|\Delta_2\|_2^2. \tag{3.6.49}$$

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

$$\Delta_2'\Omega\Delta_2 \leq (1+2\omega_0)\|\Delta_2\|_2^2 \leq (1+2\omega_0)(B_{ss}+B_{sn})\eta^2. \tag{3.6.50}$$

Recall that $B_{sn}+B_{ss} \leq (\vartheta+r)^2/(4r\vartheta)$. Combining this with (3.6.48)-(3.6.50) gives

$$|\Delta_1'\Omega\Delta_2| \leq \sqrt{(1+2\omega_0)\frac{(\vartheta+r)^2}{4\vartheta r}\eta^2}\cdot\sqrt{\Delta_1'\Omega\Delta_1}. \tag{3.6.51}$$

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)),

$$2\sqrt{c} \leq \min\{\frac{2\vartheta}{r}, 1-\frac{\vartheta}{r}, \sqrt{2-2\omega_0}-1+\frac{\vartheta}{r}\}. \tag{3.6.52}$$

Combining these with (3.6.51) gives

$$\frac{-d(\vartheta/r)+2\Delta_1'\Omega\Delta_2+\Delta_1'\Omega\Delta_1}{2\sqrt{\Delta_1'\Omega\Delta_1}} \geq \frac{-d(\vartheta/r)+\Delta_1'\Omega\Delta_1}{2\sqrt{\Delta_1'\Omega\Delta_1}} - \sqrt{c}. \tag{3.6.53}$$

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 3.6.6, $\Delta_1'\Omega\Delta_1 \geq 2B_{sn}(1-2\omega_0)+2\omega_0 \geq 2-2\omega_0$. It follows that

$$\frac{-d(\vartheta/r)+\Delta_1'\Omega\Delta_1}{2\sqrt{\Delta_1'\Omega\Delta_1}} - \sqrt{c} = \frac{1}{2}\sqrt{\Delta_1'\Omega\Delta_1} - \sqrt{c} \geq \frac{1}{2}(\sqrt{2-2\omega_0}-2\sqrt{c}). \tag{3.6.54}$$

By (3.6.52),

$$2\sqrt{c} \leq \sqrt{2-2\omega_0}-1+(\vartheta/r). \tag{3.6.55}$$

Combining (3.6.53)-(3.6.55) 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 (3.6.45) and noting $B_{ss} + B_{sn} \geq 1$, the claim follows by Mills' ratio [29].

Consider (b). In this case, $B_{ns} > B_{sn}$ and so $d \leq -1$. First, by (3.6.52), $\sqrt{c} \leq \vartheta/r$. 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 3.6.6, $\Delta_1'\Omega\Delta_1 \geq 1$. 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 (3.6.46), $B_{sn} + B_{ss} \geq 1$. Inserting these into (3.6.45), the claim follows by Mills' ratio [29].

Consider (c). In this case, $B_{ns} < B_{sn}$. We have either $B_{ns} = 0$ or $B_{ns} \geq 1$. By (3.6.46), 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 3.6.6, $\Delta_1'\Omega\Delta_1 \geq B_{sn} - B_{ns}$. 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 (3.6.52), $\sqrt{c} \leq \theta/r$. 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 (3.6.45) 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 [29]. This proves (3.6.45) for the case where Ω has only non-negative coordinates.

Next, consider (3.6.45) 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 3.6.6,

$$\Delta_1' \Omega \Delta_1 \geq 1. \tag{3.6.56}$$

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 (3.6.56) 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 (3.6.56) 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 (3.6.56) 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. □

Proof of Lemma 3.6.6

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 $k \times k$ 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, $1 \le i \le k$. Let $\tilde{\Omega} = \Lambda'\Omega\Lambda$. Note that $|\tilde{\Omega}(i,j)| = |\Omega(i,j)|$ for all $1 \le i, j \le k$, and so $\max\{\|U(\tilde{\Omega})\|_\infty, \|U(\tilde{\Omega})\|_1\} \le \omega_0$. It is seen that

$$\Delta'\Omega\Delta = 1'\Lambda'\Omega\Lambda 1 = 1'\tilde{\Omega}1, \tag{3.6.57}$$

where 1 is the $k \times 1$ vector of ones. We discuss the case $A_1 = A_2 \ge 1$ and the case $A_1 \ne A_2$ separately.

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

$$1'\tilde{\Omega}1 = \sum_{i=1}^{k} \tilde{\Omega}(i,i) + 2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} \tilde{\Omega}(i,j) \ge k - 2\sum_{i=1}^{k-1} \omega_0 \ge k - 2(k-1)\omega_0. \tag{3.6.58}$$

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

$$1'\tilde{\Omega}1 = \sum_{i=1}^{k} \tilde{\Omega}(i,i) + \sum_{i \ne j} \tilde{\Omega}(i,j) \ge k + (I + II), \tag{3.6.59}$$

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 $j \in S_2$, $\sum_{i \in S_1} |\tilde{\Omega}(i,j)| \le 2\omega_0$. Similarly, for each fixed $i \in S_2$, $\sum_{j \in S_1} |\tilde{\Omega}(i,j)| \le 2\omega_0$. Inserting these into (3.6.59) gives $1'\tilde{\Omega}1 \ge (A_1 + A_2) - 4\omega_0 A_2$, and the claim follows. □

3.6.6 Proof of Lemma 2.4

In this section and Sections 3.6.6 and 3.6.6, 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 $o(1/p)$,

$$\|W - \tilde{Y}\|_\infty \leq b_p \cdot \sqrt{2\log(p)}. \tag{3.6.60}$$

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

$$\|W - \tilde{Y}\|_\infty \leq \|W - \tilde{W}\|_\infty + \|\tilde{W} - \tilde{Y}\|_\infty. \tag{3.6.61}$$

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

$$\|\tilde{Y} - \tilde{W}\|_\infty \leq C\sqrt{\log(p)}(p^{-(\theta-(1-\vartheta))/2} + p^{-\theta/2}). \tag{3.6.62}$$

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 $o(1/p)$,

$$\|W - \tilde{W}\|_\infty \leq o(\sqrt{2\log(p)}). \tag{3.6.63}$$

Inserting (3.6.62)-(3.6.63) into (3.6.61) gives (3.6.60).

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 3.6.6.

Lemma 3.6.7. *Under the conditions of Lemma 2.4, there is a constant $c = c(\vartheta, r) > 0$ such that, with probability $1 - o(1/p)$,*

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

Proof of Lemma 3.6.7

Let $\tilde{\Omega}$ be defined as above. The following lemma is proved below in Section 3.6.6.

Lemma 3.6.8. *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 3.6.7. 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 3.6.8, for any fixed $c_0 > 0$, as $p \to \infty$, except for a probability of $o(1/p)$,

$$I \leq p^{c_0} \sum_{j=1}^{p} P(\tilde{z}_j \geq t) = p^{1+c_0} \bar{\Phi}(t). \tag{3.6.64}$$

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

$$\sum_{\{(k,\ell) \in H\}} 1_{\{\beta_k \neq 0, \beta_\ell \neq 0\}} \leq p^{c_0} E[\sum_{\{(k,\ell) \in H\}} 1_{\{\beta_k \neq 0, \beta_\ell \neq 0\}}] \leq 4\log^2(p) p^{1+c_0} \epsilon_p^2, \tag{3.6.65}$$

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

$$II \leq \sum_{j=1}^{p} 1_{\{A_{2j}\}} \leq 4\log^2(p)p^{1+c_0}\epsilon_p^2. \tag{3.6.66}$$

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 $j_1 < j_2 < \cdots < j_\ell$ 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 3.6.8, for any $c_0 > 0$, as $p \to \infty$, except for a probability of $o(1/p)$,

$$\sum_{k=1}^{\ell} 1_{\{\tilde{z}_{j_k} \geq t - \delta_0(1+\eta)\tau_p\}} \leq p^{c_0} \sum_{k=1}^{\ell} P(\tilde{z}_{j_k} \geq t - \delta_0(1+\eta)\tau_p) \leq p^{c_0}\ell\bar{\Phi}(t - \delta_0(1+\eta)\tau_p). \tag{3.6.67}$$

Since (3.6.67) 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

$$IIIa \leq 2p^{1+c_0}\epsilon_p\bar{\Phi}(t - \delta_0(1+\eta)\tau_p). \tag{3.6.68}$$

Consider $IIIb$. Write

$$IIIb = \sum_{j=1}^{p} 1_{\{\beta_j \neq 0\}} - \sum_{j=1}^{p} 1_{\{\beta_j \neq 0\}} 1_{\{A_2 j\}}, \tag{3.6.69}$$

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 (3.6.66), $\sum_{j=1}^{p} 1_{\{\beta_j \neq 0\}} 1_{\{A_2 j\}} \leq \sum_{j=1}^{p} 1_{\{A_{2j}\}} \leq 4\log^2(p)p^{1+c_0}\epsilon_p^2$. It follows that except for a probability of $o(1/p)$,

$$|IIIb - p\epsilon_p| \leq C/\sqrt{p\epsilon_p} + 4\log^2(p)p^{1+c_0}\epsilon_p^2. \tag{3.6.70}$$

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 $o(1/p)$,

$$IIIc \leq \sum_{j=1}^{p} 1_{\{\tilde{z} < t - \tau_p\}} 1_{\{\tilde{\beta}_j \neq 0\}} \leq 4\log(p) p^{1+c_0} \epsilon_p \bar{\Phi}(\tau_p - t). \tag{3.6.71}$$

Combining (3.6.68), (3.6.70), and (3.6.71) gives that except for a probability of $o(1/p)$,

$$|III - p\epsilon_p| \leq C\log^2(p)\Big[p^{1+c_0}\epsilon_p^2 \bar{\Phi}(t - \delta_0(1+\eta)\tau_p) + p^{1+c_0}\epsilon_p \bar{\Phi}(\tau_p - t) + \sqrt{\frac{1}{p\epsilon_p}} + p^{1+c_0}\epsilon_p^2\Big]. \tag{3.6.72}$$

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 (3.6.64), (3.6.66), and (3.6.72), the claim follows by Mill's ratio [29].

Proof of Lemma 3.6.8

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,

$$\begin{aligned} E[S_p^m(t)] &= \sum_{k=1}^{m} \sum_{a_1 + \ldots a_k = m} \sum_{i_1 < \ldots < i_k} E\Big[(1_{\{Y_{i_1} \geq t\}}) \ldots (1_{\{Y_{i_k} \geq t\}})\Big] \\ &\leq \sum_{k=1}^{m} \binom{m-1}{k-1} \sum_{i_1 < \ldots < i_k} E\Big[(1_{\{Y_{i_1} \geq t\}}) \ldots (1_{\{Y_{i_k} \geq t\}})\Big]. \end{aligned} \tag{3.6.73}$$

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 3.6.2 and basic combinatorics, $|\mathcal{M}(\ell; k)| \leq$

$\binom{p}{\ell}\binom{k-1}{\ell-1}(2e\log(p))^k \le C(m)p^\ell(2e\log(p))^k$. Second, note that for any $\{i_1,\dots,i_k\} \in \mathcal{M}(\ell;k)$, $E[(1_{\{Y_{i_1}\ge t\}})\dots(1_{\{Y_{i_k}\ge t\}})] \le (\bar\Phi(t))^\ell$. Combining these gives that for each $1 \le k \le m$,

$$\sum_{i_1<\dots<i_k} E[(1_{\{Y_{i_1}\ge t\}})\dots(1_{\{Y_{i_k}\ge t\}})] = \sum_{\ell=1}^{k} \sum_{\{\{i_1,\dots,i_k\}\in\mathcal{M}(\ell;k)\}} E[(1_{\{Y_{i_1}\ge t\}})\dots(1_{\{Y_{i_k}\ge t\}})]$$
$$\le \sum_{\ell=1}^{k}(2e\log(p))^k \sum_{\ell=1}^{k}(p\bar\Phi(t))^\ell \le k(2e\log(p)\bar\Phi(t))^k.$$

Inserting this into (3.6.73) gives the claim. □

3.6.7 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 2.4, there is a positive number $\delta_p = o(1)$ such that except for a probability of $o(1/p)$,

$$(1-\delta_p)\lambda_p \le \hat\lambda_p \le (1+\delta_p)\lambda_p, \qquad (1-\delta_p)u_p \le \hat u_p \le (1+\delta_p)u_p. \tag{3.6.74}$$

Note that Lemma 3.6.5 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 (3.6.74) and a close investigation of the proof of Lemma 3.6.5. □

3.6.8 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 $1 \le j \le p$, except for a probability of $o(1/p^2)$, $|e_j'(X'X - \Omega)\beta| \le C\|\Omega\| \sqrt{\log p}\ p^{-[\theta-(1-\vartheta)]/2}$. 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 $o(1/p^2)$,

$$|\sum_{i=1}^p a_i\beta_i| \le C\|\Omega\| \sqrt{\log p}\ p^{-[\theta-(1-\vartheta)]/2}. \tag{3.6.75}$$

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

$$p\mu_p \asymp \epsilon_p \sqrt{\log p} \sum_{i=1}^p a_i, \qquad p\sigma_p^2 \asymp \epsilon_p \log(p) \sum_{i=1}^p a_i^2. \tag{3.6.76}$$

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 $\|\eta\|^2 = \frac{1}{\|\Omega\|}(\frac{1}{p}1_p'\Omega 1_p) \le 1$. It follows that

$$\sum_{i=1}^p a_i = e_1'(X'X - \Omega)1_p = \sqrt{p\|\Omega\|}(\xi'Z'Z\eta - \xi'\eta). \tag{3.6.77}$$

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 $1 \le i \le n$, $(\sqrt{n}\xi'Z_i, \sqrt{n}\eta'Z_i)'$ 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 $o(1/p^2)$,

$$|\xi'Z'Z\eta - \xi'\eta| \le C\sqrt{\log(n)}/\sqrt{n} \le Cp^{-\theta/2}\sqrt{\log(p)}. \tag{3.6.78}$$

Combining (3.6.76)-(3.6.78), we have that except for a probability of $o(1/p^2)$,

$$p\mu_p \le C\epsilon_p \log(p)\sqrt{p\|\Omega\|}p^{-\theta/2} \le C\log(p)\sqrt{\|\Omega\|}p^{-\frac{\vartheta}{2}-[\theta-(1-\vartheta)]/2}. \tag{3.6.79}$$

Second, write

$$\sum_{i=1}^{p} a_i^2 = e_1'(X'X - \Omega)(X'X - \Omega)e_1 = \xi'(Z'Z - I_p)\Omega(Z'Z - I_p)\xi. \tag{3.6.80}$$

It is known [27] 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 $\|\xi\| = 1$,

$$\xi'(Z'Z - I_p)\Omega(Z'Z - I_p)\xi \le \|\Omega\|\xi'(Z'Z - I_p)(Z'Z - I_p)\xi \le C\|\Omega\|(p/n). \tag{3.6.81}$$

Combining (3.6.76), (3.6.80) and (3.6.81) gives

$$p\sigma_p^2 \le C\|\Omega\|\log(p)\epsilon_p p/n \le C\|\Omega\|\log(p)p^{-[\theta-(1-\vartheta)]}. \tag{3.6.82}$$

Last, since $\beta_i \le C\sqrt{\log p}$, using Bennett's lemma [25], for any $\lambda > 0$,

$$P(\sum_{i=1}^{p} a_i\beta_i \ge p\mu_p + \sqrt{p}\lambda) \le \exp(-\frac{\lambda^2}{2\sigma_p^2}\psi(\frac{\lambda C\sqrt{\log p}}{\sigma_p^2\sqrt{p}})), \tag{3.6.83}$$

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 (3.6.82), it follows from (3.6.83) that

$$P(\sum_{i=1}^{p} a_i\beta_i \ge p\mu_p + \sqrt{p}\lambda) = o(1/p^2). \tag{3.6.84}$$

Combining (3.6.84) with (3.6.79) and (3.6.82) gives (3.6.75). □

3.6.9 Proof of Lemma 4.1

For notational 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 [28], $(\hat{\beta}_1, \hat{\beta}_2)'$ minimizes the functional if

and only if there is a sub-gradient $\alpha = (\alpha_1, \alpha_2)'$ such that

$$\begin{pmatrix} 1 & a \\ a & 1 \end{pmatrix}\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} - \begin{pmatrix} \tilde{y}_1 \\ \tilde{y}_2 \end{pmatrix} + \lambda\alpha = 0, \qquad \text{and} \qquad \begin{cases} \alpha_i = \operatorname{sgn}(\hat{\beta}_i), & \text{if } \hat{\beta}_i \neq 0, \\ |\alpha_i| \leq 1, & \text{otherwise.} \end{cases} \tag{3.6.85}$$

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 (3.6.85) 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 (3.6.85) 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 (3.6.85). 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 (3.6.85). 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 (3.6.85). Moreover, by the definition of Region *IIIa*, $\hat{\beta}_1 > 0$ and $\hat{\beta}_2 > 0$, so $\alpha_i = \operatorname{sgn}(\hat{\beta}_i)$ and the second requirement of (3.6.85) is also satisfied. Combining these gives the claim. □

3.6.10 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 $A_{0j} = \{\beta_k = 0,\ j-2 \le k \le j+1\}$, $A_{1j} = \{\beta_{j-2} = \beta_{j-1} = \beta_{j+1} = 0, \beta_j = \tau_p\}$, $B_{0j} = \{\hat{\beta}_{j-2} = \hat{\beta}_{j-1} = \hat{\beta}_j = \hat{\beta}_{j+1} = 0\}$, 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

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

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

$$\sum_{j=1}^{p} P(\mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j)) \ge \frac{1}{7}\sum_{j=3}^{p-1}[P(A_{0j} \cap B_{0j}^c) + P(A_{1j} \cap B_{1j}^c)]. \tag{3.6.86}$$

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

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

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

$$B_{0j}^c \supseteq \{|\tilde{Y}_j| > \lambda_p\}, \qquad B_{1j}^c \supseteq (D_{1j} \cup \tilde{D}_{1j}). \tag{3.6.87}$$

This is equivalent to show that

$$B_{0j} \cap D_{0j} = \emptyset, \qquad B_{1j} \cap (D_{1j} \cup \tilde{D}_{1j}) = \emptyset. \tag{3.6.88}$$

Towards this end, we note that by the KKT condition [28],

$$\Omega\hat{\beta} = \tilde{Y} - \lambda_p\alpha, \tag{3.6.89}$$

where α is the vector of sub-gradients (i.e. $\alpha_j = \mathrm{sgn}(\hat{\beta}_j)$ if $\hat{\beta}_j \neq 0$ and $|\alpha_j| \leq 1$ otherwise). Consider the first claim in (3.6.88). Recall that Ω is a tridiagonal matrix. When B_{0j} happens, it follows from (3.6.89) 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 (3.6.88). 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 (3.6.88) 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 (3.6.86) and (3.6.87) gives

$$\sum_{j=1}^{p} P(\mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j)) \geq \frac{1}{7}\sum_{j=3}^{p-1}[P(A_{0j} \cap D_{0j}) + P(A_{1j} \cap D_{1j}) + P(A_{1j} \cap \tilde{D}_{1j})]. \tag{3.6.90}$$

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

$$P(A_{0j} \cap D_{0j}) = (1-\epsilon_p)^4 P(N(0,1) \geq \lambda_p) = L_p p^{-q}, \tag{3.6.91}$$

and that

$$P(A_{1j} \cap D_{1j}) = (1-\epsilon_p)^3\epsilon_p P(N(\tau_p, 1) \leq \lambda_p) = \begin{cases} L_p p^{-[\vartheta + (\sqrt{q} - \sqrt{r})^2]}, & q < r, \\ p^{-\vartheta}(1 + o(1)), & q > r. \end{cases} \tag{3.6.92}$$

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

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

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 $P((\tilde{Y}_{j-1}, \tilde{Y}_j)' \in R | A_{1j})$ gives

$$P(A_{1j} \cap \tilde{D}_{1j}) = \begin{cases} L_p p^{-\vartheta - \frac{1-|a|}{1+|a|}q}, & 0 < q < r, \\ L_p p^{-\vartheta \frac{1}{1+|a|}(2q+(1+|a|)r-2(1+|a|)\sqrt{qr})}, & r < q. \end{cases} \tag{3.6.93}$$

Inserting (3.6.91)-(3.6.93) into (3.6.90) gives the claim. □

3.6.11 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

$$\begin{cases} 0, & \text{if } \beta_1 = 0 \,\&\, \beta_2 = 0, \\ (\lambda_p^2 - \tilde{y}_1^2)/2, & \text{if } \beta_1 \neq 0 \,\&\, \beta_2 = 0, \\ (\lambda_p^2 - \tilde{y}_2^2)/2, & \text{if } \beta_1 = 0 \,\&\, \beta_2 \neq 0, \\ \lambda_p^2 - (\tilde{y}_1^2 + \tilde{y}_2^2 - 2a\tilde{y}_1\tilde{y}_2)/(2(1-a^2)), & \text{if } \beta_1 \neq 0 \,\&\, \beta_2 \neq 0, \end{cases} \tag{3.6.94}$$

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 (3.6.94), 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. □

3.6.12 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 $A_{0j} = \{\beta_{j-2} = \beta_{j-1} = \beta_j = \beta_{j+1} = 0\}$, $A_{1j} = \{\beta_{j-2} = \beta_{j-1} = \beta_{j+1} = 0, \beta_j = \tau_p\}$, $A_{2j} = \{\beta_{j-2} = \beta_{j+1} = 0, \beta_{j-1} = \beta_j = \tau_p\}$, $B_{0j} = \{\hat{\beta}_{j-2} = \hat{\beta}_{j-1} = \hat{\beta}_j = \hat{\beta}_{j+1} = 0\}$, $B_{1j} = \{\hat{\beta}_{j-2} = \hat{\beta}_{j-1} = \hat{\beta}_{j+1} = 0, \hat{\beta}_j \neq 0\}$, and $B_{2j} = \{\hat{\beta}_{j=1} = \hat{\beta}_{j+1} = 0, \hat{\beta}_{j-1} \neq 0, \hat{\beta}_j \neq 0\}$. The Hamming distance is

$$\begin{aligned}\sum_{j=1}^{p} P(\mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j)) &\geq \sum_{j=3}^{p-1} P(\hat{\beta}_j \neq 0, \beta_j = 0) + P(\hat{\beta}_j = 0, \beta_j \neq 0) \\ &\geq \frac{1}{9}\sum_{j=3}^{p-2}(I_j + II_j + III_j),\end{aligned}$$

where

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

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

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

$$\sum_{j=1}^{p} P(\mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j)) \geq \frac{1}{9}\sum_{j=3}^{p-1}[P(A_{0j} \cap B_{0j}^c) + P(A_{1j} \cap B_{1j}^c) + P(A_{2j} \cap B_{2j}^c)]. \quad (3.6.95)$$

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

$$B_{0j}^c \supseteq D_{0j}, \qquad B_{1j}^c \supseteq D_{1j}, \qquad B_{2j}^c \supseteq D_{2j}, \quad (3.6.96)$$

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

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

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 (3.6.95) into (3.6.96) gives

$$\sum_{j=1}^{p} P(\mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j)) \geq \frac{1}{3}\sum_{j=3}^{p-1}[P(A_{0j} \cap D_{0j}) + P(A_{1j} \cap D_{1j}) + P(A_{2j} \cap D_{2j})]. \quad (3.6.97)$$

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

$$P(A_{0j} \cap D_{0j}) = (1-\epsilon_p)^4 P(N(0,1) \geq \lambda_p) = L_p p^{-q}, \quad (3.6.98)$$

and that

$$P(A_{1j} \cap D_{1j}) = (1-\epsilon_p)^3\epsilon_p P(N(\tau_p, 1) \leq \lambda_p) = \begin{cases} L_p p^{-[\vartheta + (\sqrt{q}-\sqrt{r})^2]}, & q < r, \\ p^{-\vartheta}(1+o(1)), & q > r. \end{cases} \quad (3.6.99)$$

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,

$$P(A_{2j} \cap D_{2j}) = (1-\epsilon_p)^2 \epsilon_p^2 P((\tilde{Y}_{j-1}, \tilde{Y}_j) \in R) = L_p p^{-2\vartheta - \min\{[(\sqrt{r(1-a^2)} - \sqrt{q})^+]^2, 2[(\sqrt{r(1+a)} - \sqrt{q})^+]^2\}}. \tag{3.6.100}$$

Inserting (3.6.98)-(3.6.100) into (3.6.97) gives the claim. □

Bibliography

[1] ABRAMOVICH, F., BENJAMINI, Y., DONOHO, D. L. and JOHNSTONE, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.*, **34** 584–653. URL http://dx.doi.org/10.1214/009053606000000074.

[2] AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control*, **19** 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. J. and LEVINA, E. (2008). Covariance regularization by thresholding. *Ann. Statist.*, **36** 2577–2604. URL http://dx.doi.org/10.1214/08-AOS600.

[5] CANDÈS, E. J. and PLAN, Y. (2009). Near-ideal model selection by ℓ_1 minimization. *Ann. Statist.*, **37** 2145–2177. URL http://dx.doi.org/10.1214/08-AOS653.

[6] CHEN, S. S., DONOHO, D. L. and SAUNDERS, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, **20** 33–61. URL http://dx.doi.org/10.1137/S1064827596304010.

[7] DIESTEL, R. (2005). *Graph Theory*, vol. 173 of *Graduate Texts in Mathematics*. 3rd ed. Springer, Berlin.

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

[9] DONOHO, D. L. (2006). Compressed sensing. *IEEE Trans. Inform. Theory*, **52** 1289–1306. URL http://dx.doi.org/10.1109/TIT.2006.871582.

[10] DONOHO, D. L. and TANNER, J. (2005). Sparse nonnegative solution of underdetermined linear equations by linear programming. *Proc. Natl. Acad. Sci. U.S.A.*, **102** 9446–9451.

[11] FAN, J., JIN, J. and KE, Z. (2011). Optimal procedure for variable selection in the presence of strong dependence. *Unpublished Manuscript*.

[12] FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B*, **70** 849–911. URL http://dx.doi.org/10.1111/j.1467-9868.2008.00674.x.

[13] FOSTER, D. P. and GEORGE, E. I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.*, **22** 1947–1975. URL http://dx.doi.org/10.1214/aos/1176325766.

[14] FRIEDMAN, J. H., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33** 1–22. Http://cran.r-project.org/web/packages/glmnet/index.html, URL http://www.jstatsoft.org/v33/i01.

[15] FRIEZE, A. M. and MOLLOY, M. (1999). Splitting an expander graph. *J. Algorithms*, **33** 166–172. URL http://dx.doi.org/10.1006/jagm.1999.1023.

[16] GENOVESE, C., JIN, J. and WASSERMAN, L. (2011). Revisiting marginal regression. *Unpublished Manuscript*.

[17] HALL, P. and JIN, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *Ann. Statist.*, **38** 1686–1732. URL http://dx.doi.org/10.1214/09-AOS764.

[18] JI, P. and JIN, J. (2011). Supplementary material for "ups delivers optimal phase diagram in high dimensional variable selection".

[19] JI, P. and JIN, J. (2011). Ups delivers optimal phase diagram in high dimensional variable selection. Technical Report, available at arXiv:1010.5028.

[20] JIN, J. and ZHANG, C.-H. (2011). Adaptive optimality of ups in high dimensional variable selection. *Unpublished Manuscript*.

[21] JIN, J. and ZHANG, Q. (2011). Optimal selection of variable when signals come from an ising model. *Unpublished Manuscript*.

[22] KERKYACHARIAN, G., MOUGEOT, M., PICARD, D. and TRIBOULEY, K. (2009). Learning out of leaders. In *Multiscale, Nonlinear and Adaptive Approximation*. Springer, 295–324.

[23] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, **34** 1436–1462. URL http://dx.doi.org/10.1214/009053606000000281.

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

[25] SHORACK, G. R. and WELLNER, J. A. (1986). *Empirical Processes with Applications to Statistics*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, Wiley, New York.

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

[27] VERSHYNIN, R. (2010). *Introduction to the Non-asymptotic Analysis of Random Matrices*. Lecture notes, Department of Mathematics, University of Michigan.

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

[29] WASSERMAN, L. (2006). *All of Nonparametric Statistics*. Springer Texts in Statistics, Springer, New York.

[30] WASSERMAN, L. and ROEDER, K. (2009). High-dimensional variable selection. *Ann. Statist.*, **37** 2178–2201. URL http://dx.doi.org/10.1214/08-AOS646.

[31] YE, F. and ZHANG, C.-H. (2009). Rate minimaxity of the lasso and dantzig estimators. Technical Report, Department of Statistics and Biostatistics, Rutgers University.

[32] ZHOU, S. (2010). Thresholded lasso for high dimensional variable selection and statistical estimation. arXiv:1002.1583.

[33] ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, **101** 1418–1429. URL http://dx.doi.org/10.1198/016214506000000735.