No penalty no tears: Least squares in high-dimensional linear models

Xiangyu Wang, David Dusnon and Chenlei Leng

Abstract

Ordinary least squares (OLS) is the default method for fitting linear models, but is not applicable for problems with dimensionality larger than the sample size. For these problems, we advocate the use of a generalized version of OLS motivated by ridge regression, and propose two novel three-step algorithms involving least squares fitting and hard thresholding. The algorithms are methodologically simple to understand intuitively, computationally easy to implement efficiently, and theoretically appealing for choosing models consistently. Numerical exercises comparing our methods with penalization-based approaches in simulations and data analyses illustrate the great potential of the proposed algorithms.

1 Introduction

Long known for its consistency, simplicity and optimality under mild conditions, ordinary least squares (OLS) is the most widely used technique for fitting linear models. Developed originally for fitting fixed dimensional linear models, unfortunately, classical OLS fails in high dimensional linear models where the number of predictors p far exceeds the number of observations n. To deal with this problem, Tibshirani[1] proposed ℓ_1-penalized regression, a.k.a, *lasso*, which triggered the recent overwhelming exploration in both theory and methodology of penalization-based methods. These methods usually assume that only a small number of coefficients are nonzero (known as the *sparsity* assumption), and minimize the same least squares loss function as OLS by including an additional penalty on the coefficients, with the typical choice being the ℓ_1 norm. Such "penalization" constrains the solution space to certain directions favoring sparsity of the solution, and thus overcomes the non-unique issue with OLS. It yields a sparse solution and achieves model selection consistency and estimation consistency under certain conditions [2, 3, 4, 5].

Despite the success of the methods based on regularization, there are important issues that can not be easily neglected. On the one hand, methods using convex penalties, such as *lasso*, usually require strong conditions for model selection consistency[2, 6]. On the other hand, methods using non-convex penalties[3, 4] that can achieve model selection consistency under mild conditions often require huge computational expense. These concerns have limited the practical use of regularized methods, motivating alternative strategies such as direct hard thresholding [7].

In this article, we aim to solve the problem of fitting high-dimensional sparse linear models by reconsidering OLS and answering the following simple question: Can ordinary least squares consistently fit these models with some suitable algorithms? Our result provides an affirmative answer to this question under fairly general settings. In particular, we give a generalized form of OLS in high dimensional linear regression, and develop two algorithms that can consistently estimate the coefficients and recover the support. These algorithms involve least squares type of fitting and hard thresholding, and are non-iterative in nature. Extensive empirical experiments are provided in Section 4 to compare the proposed estimators to many existing penalization methods. The performance of the new estimators is very competitive under various setups in terms of model selection, parameter estimation and computational time.

Related works The work that is most closely related to ours is [8], in which the authors proposed an algorithm based on OLS and the ridge regression. However, both their methodology and theory are still within the ℓ_1 regularization framework, and their conditions (especially their C-Ridge and C-OLS conditions) are overly strong and can be easily violated in practice. [7] proposed an iterative hard thresholding algorithm for sparse regression, which shares a similar spirit of hard thresholding as our algorithm. Nevertheless, their motivation is completely different, their algorithm lacks theoretical guarantees for consistent support recovery, and they require an iterative estimation procedure.

Our contributions We provide a generalized form of OLS for fitting high dimensional data motivated by ridge regression, and develop two algorithms that can consistently fit a sparse linear model and recover its support. We summarize the advantages of our new algorithms in three points. First, our algorithms work for highly correlated features under random designs. The consistency of the algorithms only needs a conditional number constraint, as opposed to the strong irrepresentable condition[2, 9] required by *lasso*. Second, our algorithms can achieve consistent support recovery for general noise (with finite second-order moment) in the ultra-high dimension setting where $\log p = o(n)$. This is remarkable as most methods

(c.f. [4, 8, 10, 9, 11, 12]) that work for $\log p = o(n)$ case rely on a sub-Gaussian tail/bounded error assumption, which might fail to hold for general noise. [6] proved that *lasso* also works for a second-order condition similar to ours, but requires two additional strong assumptions. Third, the algorithms are simple, efficient and scale well for large p. In particular, the matrix operations are fully parallelizable with very few communications for very large p, while regularization methods are either hard to be computed in parallel in the feature space, or the parallelization requires a large amount of machine communications.

The remainder of this article is organized as follows. In Section 2 we generalize the ordinary least squares estimator for high dimensional problems where $p > n$, and propose two three-step algorithms consisting only of least squares fitting and hard thresholding in a loose sense. Section 3 provides consistency theory for the algorithms. Section 4 evaluates the empirical performance. We conclude and discuss further implications of our algorithms in the last section. All the proofs are provided in the supplementary materials.

2 High dimensional ordinary least squares

Consider the usual linear model

$$Y = X\beta + \varepsilon,$$

where X is the $n \times p$ design matrix, Y is the $n \times 1$ response vector and β is the coefficient. As is common in the high dimensional literature, we assume that most β_i's are zero except for a small subset $S = supp(\beta)$ with cardinality s; i.e., $S = \{i|\beta_i \neq 0\}$ is the support of β and $s = card(S)$.

To carefully tailor the low-dimensional OLS estimator for a high dimensional scenario, one needs to answer the following two questions. i) What is the correct form of OLS in the high dimensional setting? ii) How to correctly use this estimator? To answer these, we reconsider OLS from a different perspective. In fact, OLS can be viewed as the limit of the ridge estimator when the ridge parameter goes to zero, i.e.,

$$(X^T X)^{-1} X^T Y = \lim_{r \to 0} \ (X^T X + r I_p)^{-1} X^T Y.$$

One nice property of the ridge estimator is that it exists regardless of the relationship between p and n. A keen observation[12] reveals the following relationship immediately.

Lemma 1. *For any $p, n, r > 0$, we have*

$$(X^T X + rI_p)^{-1} X^T Y = X^T (XX^T + rI_n)^{-1} Y. \tag{1}$$

Notice that the right hand side of (1) exists when $p > n$ and $r = 0$. Consequently, we can naturally extend the classical OLS to the high dimensional scenario by letting r tend to zero in (1). Denote this high dimensional version of the OLS as

$$\hat{\beta}^{(HD)} = \lim_{r \to 0} X^T (XX^T + rI_n)^{-1} Y = X^T (XX^T)^{-1} Y.$$

The above equation indicates that $\hat{\beta}^{(HD)}$ is essentially an orthogonal projection of β onto the row space of X. Unfortunately, this (low dimensional) projection does not have good general performance in estimating sparse vectors in high-dimensional cases. Instead of directly estimating β as $\hat{\beta}^{HD}$, however, this new estimator of β may be used for dimension reduction by observing $\hat{\beta}^{(HD)} = X^T (XX^T)^{-1} X\beta + X^T (XX^T)^{-1} \varepsilon = \Phi\beta + \eta$ [12]. Since η is stochastically small, if Φ is close to a diagonally dominant matrix and β is sparse, then the zero and non-zero coefficients can be separated by simply thresholding the small entries of $\hat{\beta}^{(HD)}$. The exact meaning of this statement will be discussed in next section. Some simple examples demonstrating the diagonal dominance of $X^T (XX^T)^{-1} X$ are illustrated immediately in Figure 1, where the rows of X in the left two plots are drawn from $N(0, \Sigma)$ with $\sigma_{ij} = 0.6$ or $\sigma_{ij} = 0.99^{|i-j|}$. The sample size and data dimension are chosen as $(n, p) = (50, 1000)$. The right plot takes the standardized design matrix directly from the real data in Section 4 with $(n, p) = (120, 5000)$. A clear diagonal dominance pattern is visible in each plot.

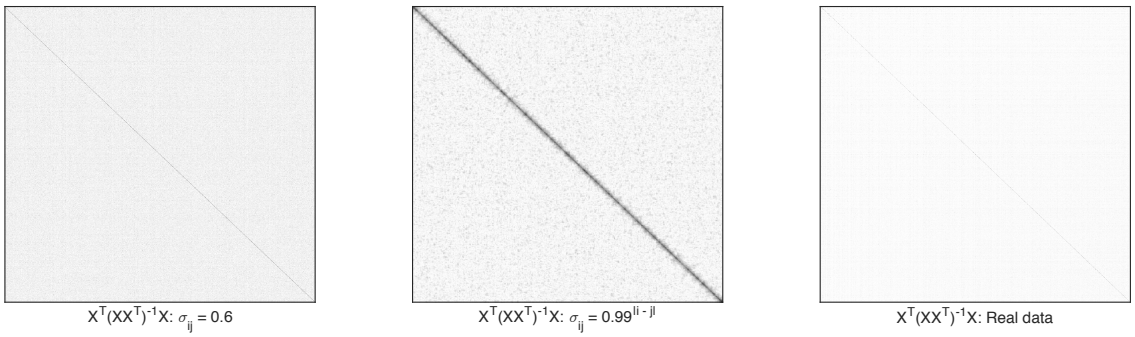

Figure 1: Examples for $X^T (XX^T)^{-1} X$. Left: $X \sim N(0, \Sigma)$ with $\sigma_{ij} = 0.6$ and $\sigma_{ii} = 1$; Middle: $X \sim N(0, \Sigma)$ with $\sigma_{ij} = 0.9^{|i-j|}$; Right: Real data from Section 4.

This ability to separate zero and non-zero coefficients allows us to first obtain a smaller model with size d such that $s < d < p$ which includes all the nonzero variables in S. Once d is below n, one can directly apply the usual OLS to obtain an estimator, which will be thresholded further to obtain a more refined model. The final estimator will then be obtained

by an OLS fit on the refined model. This three-stage non-iterative algorithm is termed *Least-squares adaptive thresholding (LAT)* and the concrete procedure is described in Algorithm 1.

Algorithm 1 *The Least-squares Adaptive Thresholding Algorithm (LAT)*

Initialization:

Input $(Y, X), d, \delta$

 # where X, Y are standardized data, n is the sample size, p is the number of features, d is the number of variables selected at stage 1 and $\delta \in (0, 1)$ is a tuning parameter determining the selection confidence

Stage 1 : Pre-selection

Compute $\hat{\beta}^{(HD)} = X^T(XX^T)^{-1}Y$. Rank the importance of the variables by $|\hat{\beta}_i^{(HD)}|$;

Denote the model corresponding to the d largest $|\hat{\beta}_i^{(HD)}|$ as $\tilde{\mathcal{M}}_d$. Alternatively use eBIC in [13] in conjunction with the obtained variable importance to select the best submodel.

Stage 2 : Hard thresholding

$\hat{\beta}^{(OLS)} = (X^T_{\tilde{\mathcal{M}}_d} X_{\tilde{\mathcal{M}}_d})^{-1} X^T_{\tilde{\mathcal{M}}_d} Y$;

$\hat{\sigma}^2 = \sum_{i=1}^n (y - \hat{y})^2/(n - d)$;

$\bar{C} = (X^T_{\tilde{\mathcal{M}}_d} X_{\tilde{\mathcal{M}}_d})^{-1}$;

Hard threshold $\hat{\beta}^{(OLS)}$ by $\textsc{mean}(\sqrt{2\hat{\sigma}^2 \bar{C}_{ii} \log(4d/\delta)})$ or use BIC to select the best submodel. Denote the chosen model as $\hat{\mathcal{M}}$.

Stage 3 : Refinement

$\hat{\beta}_{\hat{\mathcal{M}}} = (X^T_{\hat{\mathcal{M}}} X_{\hat{\mathcal{M}}})^{-1} X^T_{\hat{\mathcal{M}}} Y$;

$\hat{\beta}_i = 0, \forall i \notin \hat{\mathcal{M}}$;

return $\hat{\beta}$

The $\bar{C}$ in Stage 2 can be replaced by its ridge version $(X^T_{\tilde{\mathcal{M}}_d} X_{\tilde{\mathcal{M}}_d} + rI_d)^{-1}$ to stabilize numerical computation. This variant of the algorithm is referred to as the *Ridge Adaptive Thresholding (RAT)* algorithm.

3 Theory

In this section, we prove the consistency of Algorithm 1 in selecting the true model and provide concrete forms for all the values needed for the algorithm to work. Recall the linear model $Y = X\beta + \varepsilon$. We consider the random design where the rows of X are drawn from a multivariate Gaussian distribution $N(0, \Sigma)$. This random design allows for various correlation structures among predictors and is widely used to illustrate methods that rely on the restricted eigenvalue conditions [14, 15]. The noise ε, as mentioned earlier, is only assumed to have the second-order moment, i.e., $var(\varepsilon) = \sigma^2 < \infty$, in contrast to the sub-Gaussian/bounded error assumption seen in most high dimension literature [4, 8, 10, 9, 11]. This relaxation is similar to [6]; however we do not require any further assumptions needed by [6]. In Algorithm 1, we also propose to use extended BIC and BIC for parameter

tuning. However, the corresponding details will not be pursued here, as their consistency is straightforwardly implied by the results from this section and the existing literature on extended BIC and BIC [13].

Define $\kappa = cond(\Sigma)$ and $\tau = \min_{i\in S} |\beta_i|$. We state our result in three theorems.

Theorem 1. *Assume $p > c_0 n$ for some $c_0 > 1$ and $var(Y) \leq M_0$. If $s \log p = O(n^\nu)$ for some $\nu < 1$, $n > 4c_0/(c_0-1)^2$, and γ is chosen to be $\gamma = \frac{c_1 \kappa^{-1} \tau}{2} \frac{n}{p}$, where c_1 is some absolute constant specified in Lemma 2 in the supplementary materials, then for any $\delta \in (0,1)$ we have*

$$P\bigg(\max_{i\notin S} |\hat{\beta}_i^{(HD)}| \leq \gamma \leq \min_{i\in S} |\hat{\beta}_i^{(HD)}| \bigg) = 1 - O\bigg(\frac{\sigma^2 \kappa^4 \log p}{\tau^2 n^{1-\delta}} \bigg).$$

Theorem 1 guarantees the model selection consistency of the first stage of Algorithm 1. The proof of Theorem 1 relies on the diagonal dominance of matrix $\Phi = X^T(XX^T)^{-1}X$. In particular, it is shown that the diagonal terms of Φ are $O(\frac{n}{p})$ while the off-diagonal terms are $O(\frac{\sqrt{n}}{p})$ [16]. Thus, with an appropriate signal-to-noise ratio and true model size, $\Phi\beta$ is likely to preserve a correct magnitude order of zero and nonzero coefficients, which can then be separated by a threshold γ. As γ is not easily computable based on data, we propose to rank the $|\hat{\beta}_i|'s$ and select d largest coefficients. Alternatively, we can construct a series of nested models formed by ranking the largest n coefficients and adopt the extended BIC [13] to select the best submodel. Once the submodel $\tilde{\mathcal{M}}_d$ is obtained, we proceed to the second stage by obtaining an estimate via ordinary least squares $\hat{\beta}^{(OLS)}$ corresponding to $\tilde{\mathcal{M}}_d$. From Theorem 1, if $d > s$, we have that with probability tending to one, $\mathcal{M}^* \subseteq \tilde{\mathcal{M}}_d$, where $\mathcal{M}^*$ is the true model. Then for $\hat{\beta}^{(OLS)}$ we have the following result.

Theorem 2. *Assume $n \geq 64\kappa d \log p$, $\log p = O(n^\nu)$ and $d - s \leq \tilde{c}$ for some $\nu < 1$ and $\tilde{c} > 0$. If there exists some $\delta \in (0,1)$ such that $\tau \geq \frac{2\sigma}{n^{\delta/2}}$, then by choosing $\gamma' = \frac{\sigma}{n^{\delta/2}}$ we have*

$$P\bigg(\max_{i\notin S} |\hat{\beta}_i^{(OLS)}| \leq \gamma' \leq \min_{i\in S} |\hat{\beta}_i^{(OLS)}| \bigg) = 1 - O\bigg(\frac{\kappa \log p \log d}{n^{1-\delta}} \bigg).$$

Theorem 2 states that if $\tau = \min_{i\in S} |\beta_i| \geq \gamma'$, where $\gamma' = \sigma/n^{\delta/2}$, then by thresholding $\hat{\beta}^{(OLS)}$ at γ', we can identify the exact model with probability tending to 1. In fact, we have a similar result for ridge regression.

Theorem 3 (Ridge regression)**.** *Assume the conditions in Theorem 2. If there exists some*

$\delta \in (0,1)$ *such that* $\tau \geq \frac{4\sigma}{n^{\delta/2}}$*, then if the ridge parameter* r *satisfies that*

$$r \leq O\Big\{ \min\Big(\frac{\sqrt{n}}{\kappa},\ \frac{\sigma^{\frac{1}{2}} n^{1-\delta/4}}{8^2 M_0^{\frac{1}{2}} \kappa^{\frac{3}{2}}} \Big) \Big\},$$

where M_0 *is defined in Theorem 1, then by choosing* $\gamma' = \frac{2\sigma}{n^{\delta/2}}$ *we have*

$$P\Big(\max_{i \notin S} |\hat{\beta}_i^{(Ridge)}(r)| \leq \gamma' \leq \min_{i \in S} |\hat{\beta}_i^{(Ridge)}(r)| \Big) = 1 - O\Big(\frac{\kappa \log p \log d}{n^{1-\delta}} \Big).$$

Note that the ridge parameter r can be chosen as a constant, bypassing the need to specify r at least in theory. When the noise follows a Gaussian distribution, we can obtain a more explicit form of the threshold γ', as the following Corollary shows.

Corollary 1 (Gaussian noise)**.** *Assume* $\varepsilon \sim N(0, \sigma^2)$*. For any* $\delta \in (0,1)$*, define* $\gamma' = 8\sqrt{2}\hat{\sigma}\sqrt{\frac{2\kappa \log(4d/\delta)}{n}}$*, where* $\hat{\sigma}$ *is the estimated standard error as* $\hat{\sigma}^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2/(n-d)$*. For sufficiently large* n*, if* $d \leq n - 4K^2 \log(2/\delta)/c$ *for some absolute constants* c*,* K *and* $\tau \geq 24\sigma\sqrt{\frac{2\kappa \log(4d/\delta)}{n}}$*, then with probability at least* $1 - 2\delta$*, we have*

$$|\hat{\beta}_i^{(OLS)}| \geq \gamma' \quad \forall i \in S \quad \text{and} \quad |\hat{\beta}_i^{(OLS)}| \leq \gamma' \quad \forall i \notin S.$$

Write $\bar{C} = (X_{\tilde{\mathcal{M}}_d}^T X_{\tilde{\mathcal{M}}_d})^{-1}$ as in Algorithm 1. In practice, we propose to use $\gamma' = mean(\sqrt{2\hat{\sigma}^2 \bar{C}_{ii} \log(4d/\delta)})$ as the threshold (see Algorithm 1), because the estimation error takes a form of $\sqrt{\sigma^2 \bar{C}_{ii} \log(4d/\delta)}$. Alternatively, instead of identifying an explicit form of the threshold value (as is hard for general noise), one may also use BIC on nested models formed by ranking $|\hat{\beta}^{(OLS)}|$ to search for the true model. Once the final model is obtained, as in Stage 3 of Algorithm 1, we refit it again using ordinary least squares. The final output will have the same output as if we knew the true model *a priori* with probability tending to 1, i.e., we have the following result.

Theorem 4. *Let* $\hat{\mathcal{M}}$ *and* $\hat{\beta}$ *be the final output from LAT or RAT. Assume all conditions in Theorem 1, 2 and 3. Then with probability at least* $1 - O\Big(\frac{\sigma^2 \kappa^4 \log p}{\tau^2 n^{1-\delta}} + \frac{\kappa \log p \log d}{n^{1-\delta}} \Big)$ *we have*

$$\hat{\mathcal{M}} = \mathcal{M}^*, \quad \|\hat{\beta} - \beta\|_2^2 \leq \frac{2s\sigma^2}{n^\delta}, \quad \text{and} \quad \|\hat{\beta} - \beta\|_\infty \leq \frac{2\sigma}{n^{\delta/2}}.$$

As implied by Theorem 1 – 4, *LAT* and *RAT* achieve consistent support recovery in the ultra-high dimensional ($\log p = o(n)$) setting only with two assumptions: $\tau = O(\sqrt{(\log p)/n})$ and $var(\varepsilon) < \infty$, in contrast to most existing methods that require $\varepsilon \sim N(0, \sigma^2)$ or $\|\varepsilon\|_\infty < \infty$.

4 Experiments

In this section, we provide extensive numerical experiments for assessing the performance of *LAT* and *RAT*. In particular, we compare the two methods to existing penalized methods including *lasso*, elastic net (*enet* [5]), *scad* [3] and *mc+* [4]. As it is well-known that the *lasso* estimator is biased, we also consider two variations of it by combining *lasso* with Stage 2 and 3 of our *LAT* and *RAT* algorithms, denoted as *lasLAT* (*las1* in Figures) and *lasRAT* (*las2* in Figures) respectively. We code *LAT* and *RAT* in *Matlab*, use `glmnet`[17] for *enet* and *lasso*, and `SparseReg`[18, 19] for *scad* and *mc+*.

4.1 Synthetic datasets

The model used in this section for comparison is the linear model $Y = X\beta + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$ and $X \sim N(0, \Sigma)$. To control the signal-to-noise ratio, we define $r = \|\beta\|_2/\sigma$, which is chosen to be 2.3 for all experiments. The sample size and the data dimension are chosen to be $(n, p) = (200, 1000)$ or $(n, p) = (500, 10000)$ for all experiments. For evaluation purposes, we consider four different structures of Σ below.

(i) *Independent predictors*. The support is set as $S = \{1, 2, 3, 4, 5\}$. We generate X_i from a standard multivariate normal distribution with independent components. The coefficients are specified as

$$\beta_i = (-1)^{u_i}(|N(0, 1)| + 1), \text{ where } u_i \sim Ber(0.5) \text{ for } i \in S \text{ and } \beta_i = 0 \text{ for } i \notin S.$$

(ii) *Compound symmetry*. All predictors are equally correlated with correlation $\rho = 0.6$. The coefficients are set to be $\beta_i = 3$ for $i = 1, ..., 5$ and $\beta_i = 0$ otherwise.

(iii) *Group structure*. This example is Example 4 in [5], for which we allocate the 15 true variables into three groups. Specifically, the predictors are generated as

$$x_{1+3m} = z_1 + N(0, 0.01),\ x_{2+3m} = z_2 + N(0, 0.01),\ x_{3+3m} = z_3 + N(0, 0.01),$$

where $m = 0, 1, 2, 3, 4$ and $z_i \sim N(0, 1)$ are independent. The coefficients are set as

$$\beta_i = 3,\ i = 1, 2, \cdots, 15;\ \beta_i = 0,\ i = 16, \cdots, p.$$

(iv) *Factor models*. This model is also considered in [20] and [21]. Let $\phi_j, j = 1, 2, \cdots, k$ be independent standard normal variables. We set predictors as $x_i = \sum_{j=1}^{k} \phi_j f_{ij} + \eta_i$, where f_{ij} and η_i are generated from independent standard normal distributions. The number of

factors is chosen as $k = 5$ in the simulation while the coefficients are specified the same as in Example (ii).

To compare the performance of all methods, we simulate 200 synthetic datasets for $(n, p) = (200, 1000)$ and 100 for $(n, p) = (500, 10000)$ for each example, and record i) the **root mean squared error (RMSE):** $\|\hat{\beta} - \beta\|_2$, ii) the **false negatives (# FN)**, iii) the **false positives (# FP)** and iv) the actual **runtime** (in milliseconds). We use the extended BIC [13] to choose the parameters for any regularized algorithm. Due to the huge computation expense for *scad* and *mc+*, we only find the first $\lceil\sqrt{p}\rceil$ predictors on the solution path (because we know $s << \sqrt{p}$). For *RAT* and *LAT*, d is set to $0.3 \times n$. For *RAT* and *larsRidge*, we adopt a 10-fold cross-validation procedure to tune the ridge parameter r for a better finite-sample performance, although the theory allows r to be fixed as a constant. For all hard-thresholding steps, we fix $\delta = 0.5$. The results for $(n, p) = (200, 1000)$ are plotted in Figure 2, 3, 4 and 5 and more comprehensive results (average values for **RMSE, # FPs, # FNs, runtime**) are summarized in Table 1 and 2.

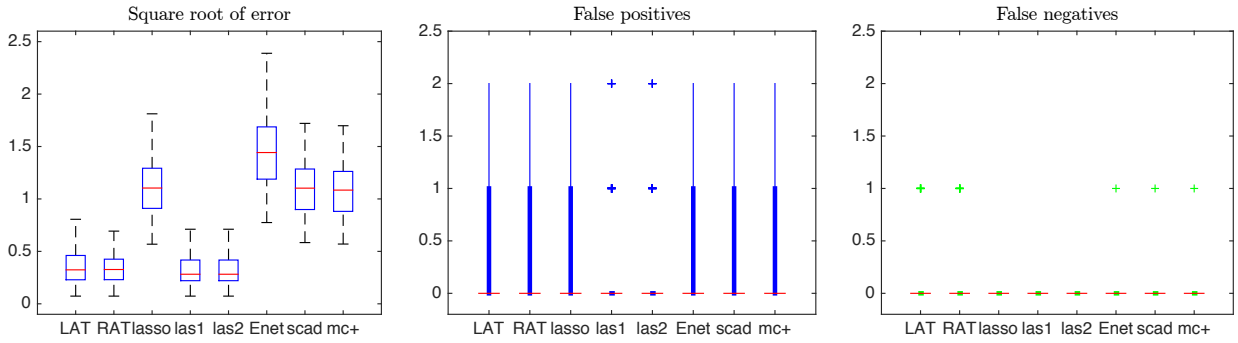

Figure 2: The boxplots for Example (i). Left: Estimation error; Middle: False positives; Right: False negatives

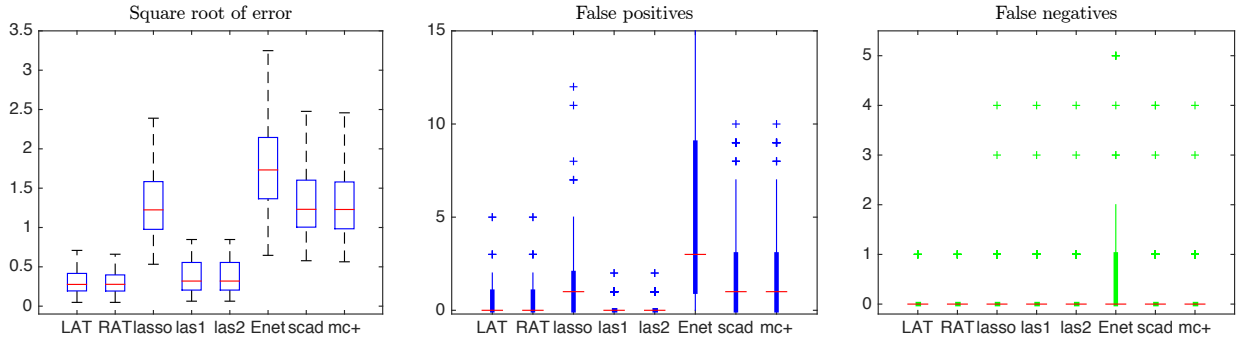

Figure 3: The boxplots for Example (ii). Left: Estimation error; Middle: False positives; Right: False negatives

As can be seen from both the plots and the tables, the performance of *LAT* and *RAT* are on par with *lasLAT* for Example (i), (ii) and (iv), and are often among the best of

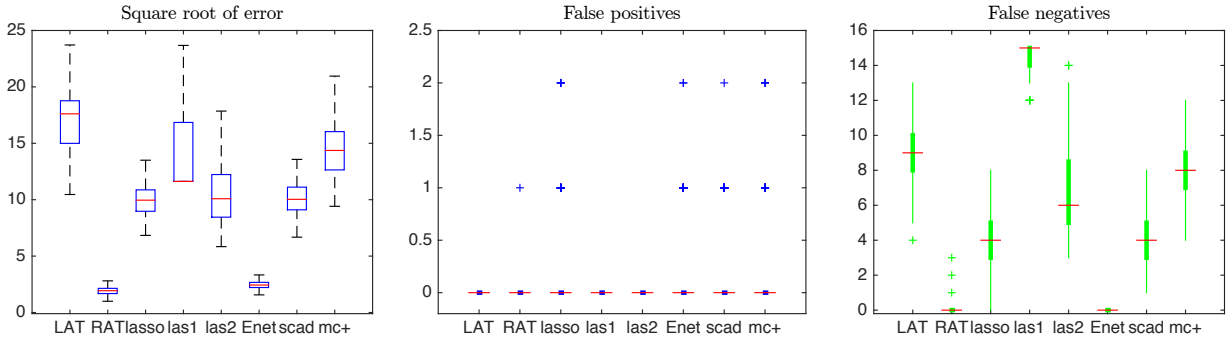

Figure 4: The boxplots for Example (iii). Left: Estimation error; Middle: False positives; Right: False negatives

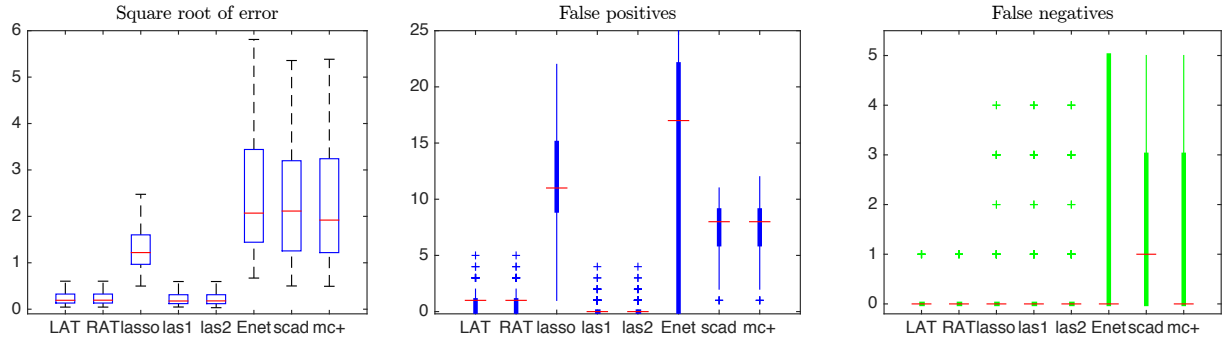

Figure 5: The boxplots for Example (iv). Left: Estimation error; Middle: False positives; Right: False negatives

all methods. For Example (iii), *RAT* and *enet* achieve the best performance while all the other methods fail to work. In addition, the runtime of *LAT* and *RAT* are also competitive compared to that of *lasso* and *enet*. We thus conclude that *LAT* and *RAT* achieve similar or even better performance compared to the usual regularized methods.

4.2 Real data

This dataset, taken from [22], was collected to study mammalian eye diseases, with gene expression for the eye tissues of 120 twelve-week-old male F2 rats recorded. One gene coded as TRIM32 responsible for causing Bardet-Biedl syndrome is of particular interest, and is the response of interest.

Following the method in [22], 18976 probes were selected as they exhibited sufficient signal for reliable analysis and at least 2-fold variation in expressions. Because TRIM32 is believed to be only linked to a small number of genes, we confine our attention to the top 5000 genes with the highest sample variance. The eight methods used in the simulation study are compared, where the performance is assessed via 10-fold cross validation. Because extended BIC does not offer a competitive prediction accuracy (It focuses on ensuring a good

Table 1: Results for $(n, p) = (200, 1000)$

Example		*LAT*	*RAT*	*lasso*	*lasLAT*	*lasRAT*	*enet*	*scad*	*mc+*
Ex. (i)	RMSE	0.398	0.397	1.117	0.329	0.329	1.476	1.110	1.089
	# FPs	0.425	0.450	0.330	0.270	0.270	0.620	0.320	0.325
	# FNs	0.075	0.075	0.000	0.000	0.000	0.005	0.005	0.005
	Time	9.9	46.4	40.4	40.6	54.3	40.2	326.6	289.1
Ex. (ii)	RMSE	0.348	0.352	1.323	0.539	0.541	1.861	1.346	1.321
	# FPs	0.440	0.405	1.470	0.240	0.245	6.535	2.020	1.930
	# FNs	0.040	0.055	0.200	0.200	0.200	0.445	0.215	0.190
	Time	8.4	44.6	40.8	41.0	54.7	46.7	356.9	317.2
Ex. (iii)	RMSE	17.338	2.115	9.960	14.632	11.151	2.453	10.129	14.416
	# FPs	0.000	0.005	0.125	0.000	0.000	0.150	0.140	0.140
	# FNs	8.920	0.030	3.900	14.305	6.910	0.000	4.385	7.695
	Time	8.9	47.0	42.4	42.7	58.1	36.5	2025.9	1133.5
Ex. (iv)	RMSE	0.255	0.260	1.396	0.475	0.475	2.438	2.300	2.260
	# FPs	0.855	0.855	11.850	0.245	0.245	14.165	7.380	7.170
	# FNs	0.030	0.035	0.265	0.270	0.270	1.715	1.540	1.515
	Time	8.0	42.3	40.0	40.3	55.5	46.1	680.2	671.8

Table 2: Results for $(n, p) = (500, 10000)$

Example		*LAT*	*RAT*	*lasso*	*lasLAT*	*lasRAT*	*enet*	*scad*	*mc+*
Ex.(i)	RMSE	0.263	0.264	0.781	0.214	0.214	1.039	0.762	0.755
	# FPs	0.550	0.580	0.190	0.190	0.190	0.470	0.280	0.280
	# FNs	0.010	0.010	0.000	0.000	0.000	0.000	0.000	0.000
	Time	36.1	41.8	72.7	72.7	74.1	71.8	1107.5	1003.2
Ex. (ii)	RMSE	0.204	0.204	0.979	0.260	0.260	1.363	0.967	0.959
	# FPs	0.480	0.480	1.500	0.350	0.350	10.820	2.470	2.400
	# FNs	0.000	0.000	0.040	0.040	0.040	0.040	0.020	0.020
	Time	34.8	40.8	76.1	76.1	77.5	82.0	1557.6	1456.1
Ex. (iii)	RMSE	9.738	1.347	7.326	17.621	3.837	1.843	7.285	8.462
	# FPs	0.000	0.000	0.060	0.000	0.000	0.120	0.120	0.090
	# FNs	4.640	0.000	1.440	13.360	1.450	0.000	1.800	2.780
	Time	35.0	41.6	75.6	75.6	77.5	74.4	6304.4	4613.8
Ex. (iv)	RMSE	0.168	0.168	1.175	0.256	0.256	1.780	0.389	0.368
	# FPs	0.920	0.920	21.710	0.260	0.260	37.210	6.360	6.270
	# FNs	0.010	0.010	0.140	0.140	0.140	0.450	0.000	0.000
	Time	34.5	41.1	78.7	78.7	80.8	81.4	1895.6	1937.1

variable selection performance) for regularized methods, for a fair comparison, we apply the conventional BIC instead of the extended BIC to all regularization methods, and record the means and the standard errors of the cross-validation. As a reference, we also report these values for the null model.

Table 3: Analysis of the eye disease data via different methods

methods	CV mean	CV standard error	average model size	total runtime (sec)
LAT	0.015	0.0157	2.6	0.29
RAT	0.014	0.0100	1.5	0.40
lasso	0.012	0.0100	76.8	1.20
lasLAT	0.019	0.0265	18.3	1.21
lasRAT	0.014	0.0064	12.7	2.33
enet	0.011	0.0109	62.2	1.38
scad	0.017	0.0245	12.4	73.12
mc+	0.017	0.0252	10.1	55.13
null	0.022	0.0257	0	—

It can be seen that *enet* and *lasso* achieve the smallest cross-validation errors overall, followed by *RAT* and *LAT*. One caveat for the good performance of *enet* or *lasso* is the large number of variables it selected. If a more parsimonious model for interpretability is preferred, one might want to trade-off some accuracy by obtaining a model with a fewer number of variables given by *LAT* or *RAT*.

5 Conclusion

We have proposed two novel algorithms *Lat* and *Rat* that only rely on least-squares type of fitting and hard thresholding, based on a high-dimensional generalization of OLS. The two methods are simple, easily implementable, and can consistently fit a high dimensional linear model and recover its support. The performance of the two methods are competitive compared to existing regularization methods. It is of great interest to further extend this framework to other models such as generalized linear models and models for survival analysis.

References

[1] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 267–288, 1996.

[2] Peng Zhao and Bin Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

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

[4] Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

[5] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[6] Karim Lounici. Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electronic Journal of Statistics*, 2:90–102, 2008.

[7] Prateek Jain, Ambuj Tewari, and Purushottam Kar. On iterative hard thresholding methods for high-dimensional m-estimation. In *Advances in Neural Information Processing Systems*, pages 685–693, 2014.

[8] Eunho Yang, Aurelie Lozano, and Pradeep Ravikumar. Elementary estimators for high-dimensional linear regression. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 388–396, 2014.

[9] Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.

[10] T Tony Cai and Lie Wang. Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Transactions on Information Theory*, 57(7):4680–4688, 2011.

[11] Cun-Hui Zhang and Jian Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567–1594, 2008.

[12] Xiangyu Wang and Chenlei Leng. High-dimensional ordinary least-squares projection for screening variables. `https://stat.duke.edu/~xw56/holp-paper.pdf`, 2015.

[13] Jiahua Chen and Zehua Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

[14] Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[15] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.

[16] Xiangyu Wang, Chenlei Leng, and David B Dunson. On the consistency theory of high dimensional variable screening. *arXiv preprint arXiv:1502.06895*, 2015.

[17] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

[18] Hua Zhou, Artin Armagan, and David B Dunson. Path following and empirical bayes model selection for sparse regression. *arXiv preprint arXiv:1201.3528*, 2012.

[19] Hua Zhou and Kenneth Lange. A path algorithm for constrained estimation. *Journal of Computational and Graphical Statistics*, 22(2):261–283, 2013.

[20] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

[21] Haeran Cho and Piotr Fryzlewicz. High dimensional variable selection via tilting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):593–622, 2012.

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

[23] Michael G Akritas, SN Lahiri, and Dimitris N Politis. Topics in nonparametric statistics. In *Proceedings of the First Conference of the International Society for Nonparametric Statistics*. Springer.

[24] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Appendix A: Proof of Theorem 1

Recall the estimator $\hat{\beta}^{(HD)} = X^T(XX^T)^{-1}Y = X^T(XX^T)^{-1}X\beta + X^T(XX^T)^{-1}\varepsilon = \xi + \eta$. The following two lemmas will be used to bound ξ and η respectively.

Lemma 2. *Let $\Phi = X^T(XX^T)^{-1}X$. Assume $p > c_0 n$ for some $c_0 > 1$, then for any $C > 0$ there exists some $0 < c_1 < 1 < c_2$ and $c_3 > 0$ such that for any $t > 0$ and any $i \in Q, j \neq i$,*

$$P\bigg(|\Phi_{ii}| < c_1\kappa^{-1}\frac{n}{p}\bigg) \leq 2e^{-Cn}, \quad P(|\Phi_{ii}| > c_2\kappa\frac{n}{p}\bigg) \leq 2e^{-Cn} \tag{2}$$

and

$$P\bigg(|\Phi_{ij}| > c_4\kappa t\frac{\sqrt{n}}{p}\bigg) \leq 5e^{-Cn} + 2e^{-t^2/2}, \tag{3}$$

where $c_4 = \frac{\sqrt{c_2(c_0-c_1)}}{\sqrt{c_3(c_0-1)}}$.

This is exactly the Lemma 3 in [16].

Lemma 3. *Assume X follows $N(0, \Sigma)$. If $var(\epsilon) = \sigma^2$ and $\log p = o(n)$, then for any $0 < \delta < 1$ we have*

$$P\bigg(\|\eta\|_\infty \leq \frac{c_1\kappa^{-1}\tau}{4}\frac{n}{p}\bigg) \geq 1 - O\bigg(\frac{\sigma^2\kappa^4\log p}{\tau^2 n^{1-\delta}}\bigg),$$

where $\tau = \min_{i\in S}|\beta_i|$ and $\kappa = cond(\Sigma)$.

To prove Lemma 3 we need the following two propositions.

Proposition 1. *(Lounici, 2008 [6]; Nemirovski, 2000 [23]) Let $Y_i \in \mathcal{R}^p$ be random vectors with zero means and finite variances. Then we have for any k norm with $k \in [2, \infty]$ and $p \geq 3$, we have*

$$E\big\|\sum_{i=1}^n Y_i\big\|_k^2 \leq \tilde{C}\min\{k, \log p\}\sum_{i=1}^n E\|Y_i\|_k^2, \tag{4}$$

where $\tilde{C}$ is some absolute constant.

As each row of X is an iid draw from $N(0, \Sigma)$, we define $Z = X\Sigma^{-1/2}$, then $Z \sim N(0, I_p)$. For Z, we have the following result.

Proposition 2. *Let $Z \sim N(0, I_p)$, then we have the minimum eigenvalue of ZZ^T/p satisfies that*

$$P\bigg(\lambda_{min}(ZZ^T/p) > (1 - \frac{n}{p} - \frac{t}{p})^2\bigg) \geq 1 - 2\exp(-t^2/2)$$

for any $t > 0$. Assume $p > c_0 n$ for $c_0 > 1$ and take $t = \sqrt{n}$. When $n > 4c_0^2/(c_0 - 1)^2$, we have

$$P\bigg(\lambda_{min}(ZZ^T/p) > c\bigg) \geq 1 - 2\exp(-n/2), \tag{5}$$

where $c = \frac{(c_0-1)^2}{4c_0^2}$.

The proof follows Corollary 5.35 in [24].

Proof of Lemma 3. Let $A = pX^T(XX^T)^{-1}$ and define $Z = X\Sigma^{-1/2}$. Consider the standard SVD on Z as $Z = VDU^T$, where V and D are $n \times n$ matrices and U is a $p \times n$ matrix. Because Z is a matrix of iid Gaussian variables, its distribution is invariant under both left and right orthogonal transformation. In particular, for any $T \in \mathcal{O}(n)$, we have

$$TVDU^T \overset{(d)}{=} VDU^T,$$

i.e., V is uniformly distributed on $\mathcal{O}(n)$ conditional on U and D (they are in fact independent, but we don't need such a strong condition). Therefore, we have

$$\begin{aligned} A &= pX^T(XX^T)^{-1} = p\Sigma^{\frac{1}{2}}Z^T(Z\Sigma Z^T)^{-1} = p\Sigma^{\frac{1}{2}}UDV^T(VDU^T\Sigma UDV^T)^{-1} \\ &= p\Sigma^{\frac{1}{2}}U(U^T\Sigma U)^{-1}D^{-1}V^T = \sqrt{p}\Sigma^{\frac{1}{2}}U(U^T\Sigma U)^{-1}\big(\frac{D}{\sqrt{p}}\big)^{-1}V^T. \end{aligned}$$

Because V is uniformly distributed conditional on U and D, the distribution of A is also invariant under right orthogonal transformation conditional on U and D, i.e., for any $T \in \mathcal{O}(n)$, we have

$$A \overset{(d)}{=} AT. \tag{6}$$

Our first goal is to bound the magnitude of individual entries A_{ij}. Let $v_i = e_i^T AA^T e_i$, which is a function of U and D (see below). From (6), we know that $e_i^T A$ is uniformly distributed

on the sphere $S^{n-1}(\sqrt{v_i})$ if conditional on v_i (i.e., conditional on U, D), which implies that

$$e_i^T A \overset{(d)}{=} \sqrt{v_i}\left(\frac{x_1}{\sqrt{\sum_{j=1}^n x_j^2}}, \frac{x_2}{\sqrt{\sum_{j=1}^n x_j^2}}, \cdots, \frac{x_n}{\sqrt{\sum_{j=1}^n x_j^2}}\right), \tag{7}$$

where $x_j's$ are iid standard Gaussian variables. Thus, A_{ij} can be bounded easily if we can bound v_i. Notice that for v_i we have

$$\begin{aligned} v_i &= e_i^T A A^T e_i = p e_i^T \Sigma^{\frac{1}{2}} U (U^T \Sigma U)^{-1} \big(\frac{D^2}{p}\big)^{-1} (U^T \Sigma U)^{-1} U^T \Sigma^{\frac{1}{2}} e_i. \\ &= p e_i^T H (U^T \Sigma U)^{-\frac{1}{2}} \big(\frac{D^2}{p}\big)^{-1} (U^T \Sigma U)^{-\frac{1}{2}} H^T e_i \\ &\leq p e_i^T H H^T e_i \cdot \lambda_{min}^{-1}(U^T \Sigma U) \cdot \lambda_{min}^{-1}\big(\frac{D^2}{p}\big) \end{aligned}$$

Here $H = \Sigma^{\frac{1}{2}} U (U^T \Sigma U)^{-1/2}$ is defined the same as in [12] and can be bounded as $e_i^T H H^T e_i \leq c_2 n\kappa/p$ with probability $1 - 2\exp(-Cn)$ (see the proof of Lemma 3 in [16]). Therefore, we have

$$P\left(v_i \leq c_2 \kappa^2 \lambda_{min}^{-1}\big(\frac{D^2}{p}\big) n\right) \geq 1 - 2\exp(-Cn)$$

Now applying the tail bound and the concentration inequality to (7) we have for any $t > 0$ and any $C > 0$

$$P(|x_j| > t) \leq 2\exp(-t^2/2) \qquad P\left(\frac{\sum_{j=1}^n x_j^2}{n} \leq c_3\right) \leq \exp(-Cn). \tag{8}$$

Putting the pieces all together, we have for any $t > 0$ and any $C > 0$ that

$$P\left(\max_{ij} |A_{ij}| \leq \kappa t \sqrt{\frac{c_2}{c_3}} \lambda_{min}^{-\frac{1}{2}}\big(\frac{D^2}{p}\big)\right) \geq 1 - 2np\exp(-t^2/2) - 3p\exp(-Cn).$$

Now according to (5), we can further bound $\lambda_{min}(D^2/p)$ and obtain that

$$P\left(\max_{ij} |A_{ij}| \leq \sqrt{\frac{c_2}{c c_3}} \kappa t\right) \geq 1 - 2np\exp(-t^2/2) - 3p\exp(-Cn) - 2\exp(-n/2). \tag{9}$$

The second step is to use (9) and Proposition 1 to bound η. The procedure follows almost the same as in Lounici's paper. Define $Z_j = (A_{1j}\epsilon_j, A_{2j}\epsilon_j, \cdots, A_{pj}\epsilon_j)$. It's clear that

$\eta = \sum_{j=1}^n Z_j/p$. Applying Proposition 1 to $Z_j's$ and choosing the l_∞ norm, we have

$$E\Big\|\sum_{j=1}^n Z_j\Big\|_\infty^2 \le \log p \sum_{j=1}^n E\|Z_j\|_\infty^2 \le \frac{c_2}{cc_3}\sigma^2\kappa^2 t^2 n \log p.$$

Using the Markov inequality on η, we have for any $r > 0$

$$\begin{aligned} P\bigg(\|\eta\|_\infty \ge \frac{\sqrt{n}r}{p}\bigg) = P\bigg(\frac{p}{\sqrt{n}}\|\eta\|_\infty \ge r\bigg) &\le \frac{p^2 E\|\eta\|_\infty^2}{nr^2} = \frac{E\|\sum_{j=1}^n Z_j\|_\infty^2}{nr^2} \\ &\le \frac{c_2\sigma^2\kappa^2 t^2 \log p}{cc_3 r^2}. \end{aligned}$$

To match our previous result, we take $r = c_1\sqrt{n}\tau\kappa^{-1}/4$ and $t = n^{\delta/2}$ for some small δ,

$$\begin{aligned} P\bigg(\|\eta\|_\infty \le \frac{c_1\kappa^{-1}\tau}{4}\frac{n}{p}\bigg) &\ge 1 - \frac{c_2\sigma^2\kappa^4}{c_1^2 cc_3\tau^2}\frac{\log p}{n^{1-\delta}} - 2np\exp(-n^\delta/2) - 3p\exp(-Cn) - 2\exp(-n/2) \\ &\ge 1 - O\bigg(\frac{\sigma^2\kappa^4\log p}{\tau^2 n^{1-\delta}}\bigg). \end{aligned}$$

□

Now we are ready to prove Theorem 1

Proof of Theorem 1. Recall the definition of ξ as $\xi = X^T(XX^T)^{-1}X\beta$. For any $i \in S$ we have

$$\xi_i = e_i^T X^T(XX^T)^{-1}X\beta = \sum_{j\in S}\Phi_{ii}\beta_i + \sum_{j\ne i, j\in S}\Phi_{ij}\beta_j,$$

and for $i \notin S$,

$$\xi_i = e_i^T X^T(XX^T)^{-1}X\beta = \sum_{j\in S}\Phi_{ij}\beta_j.$$

According to our assumption we have $\min_{i\in S}|\beta_i| \ge \tau$ and $var(Y) = var(X\beta) = \beta^T\Sigma\beta \le M_0$ for some M_0. The latter one imples that

$$M_0 \ge \beta^T\Sigma\beta \ge \lambda_{min}(\Sigma)\|\beta\|_2^2.$$

Therefore, we have for any $i \in S$

$$|\xi_i| \ge c_1\kappa^{-1}\tau\frac{n}{p} - \|\beta\|_2\sqrt{\sum_{j\ne i, j\in S}\Phi_{ij}^2} \ge c_1\kappa^{-1}\tau\frac{n}{p} - \frac{c_4\kappa\sqrt{sM_0}t}{\lambda_{min}^{\frac{1}{2}}(\Sigma)}\frac{\sqrt{n}}{p} = \frac{3c_1\kappa^{-1}\tau}{4}\frac{n}{p},$$

if t is taken to be $t = \frac{c_1 \lambda_{min}^{\frac{1}{2}}(\Sigma)\kappa^{-2}\tau\sqrt{n}}{4c_4\sqrt{M_0 s}} \geq \frac{c_1\kappa^{-\frac{5}{2}}\tau\sqrt{n}}{4c_4\sqrt{M_0 s}}$. Hence, one can compute the probability to be greater than $1 - 7\exp(-Cn) - 2\exp\big(-\frac{c_1^2\kappa^{-5}\tau^2}{32c_4^2 M_0 s}n\big)$. Similarly, with the same t we can show that for $i \notin S$

$$|\xi_i| \leq \|\beta\|_2 \sqrt{\sum_{j\neq i, j\in S} \Phi_{ij}^2} \leq \frac{c_1\kappa^{-1}\tau}{4}\frac{n}{p},$$

with probability greater than $1 - 7\exp(-Cn) - 2\exp\big(-\frac{c_1^2\kappa^{-5}\tau^2}{32c_4^2 M_0 s}n\big)$. Next, using the result from Lemma 3, we can obtain

$$P\bigg(\min_{i\in S}|\hat{\beta}_i| \geq \frac{c_1\kappa^{-1}\tau}{2}\frac{n}{p}\bigg) \geq 1 - O\bigg(\frac{\sigma^2\kappa^4\log p}{\tau^2 n^{1-\delta}}\bigg),$$

and

$$P\bigg(\max_{i\notin S}|\hat{\beta}_i| \leq \frac{c_1\kappa^{-1}\tau}{2}\frac{n}{p}\bigg) \geq 1 - O\bigg(\frac{\sigma^2\kappa^4\log p}{\tau^2 n^{1-\delta}}\bigg).$$

Taking $\gamma = \frac{c_1\kappa^{-1}\tau}{2}np$, we have

$$P\bigg(\min_{i\in S}|\hat{\beta}_i| \geq \gamma \geq \max_{i\notin S}|\hat{\beta}_i|\bigg) \geq 1 - O\bigg(\frac{\sigma^2\kappa^4\log p}{\tau^2 n^{1-\delta}}\bigg).$$

□

Proof of Theorem 2 and 3

Lemma 4. *Let $\tilde{\mathcal{M}}_d$ be a submodel that contains the true model $\mathcal{M}^*$ and has a size of d. Define $A = n(X_{\tilde{\mathcal{M}}_d}^T X_{\tilde{\mathcal{M}}_d})^{-1}X_{\tilde{\mathcal{M}}_d}^T$ where $X_{\tilde{\mathcal{M}}_d}$ is the principal submatrix indexed by $\tilde{\mathcal{M}}_d$. Then for any $t > 0$ and $C > 0$, there exists some $c_3 > 0$ such that*

$$P\bigg(\max_{|\tilde{\mathcal{M}}_d|=d, \mathcal{M}^*\subseteq\tilde{\mathcal{M}}_d}\max_{ij}|A_{ij}| \leq \frac{t}{\sqrt{c_3\lambda_0}}\bigg) \geq 1 - 2dn(p-s)^{d-s}\exp\bigg(-\frac{t^2}{2}\bigg) - d(p-s)^{d-s}\exp(-Cn),$$

where $\lambda_0 = \min_{|\tilde{\mathcal{M}}_d|=d, \mathcal{M}^*\subseteq\tilde{\mathcal{M}}_d}\lambda_{min}(X_{\tilde{\mathcal{M}}_d}^T X_{\tilde{\mathcal{M}}_d}/n)$.

Proof of Lemma 4. The proof is similar to the argument in Lemma 3. For a given $\tilde{\mathcal{M}}_d$, $X_{\tilde{\mathcal{M}}_d}$ follows $N(0, \Sigma_{\tilde{\mathcal{M}}_d})$. Similarly, defining $Z = X\Sigma_{\tilde{\mathcal{M}}_d}^{-1/2}$, then $Z \sim N(0, I)$. Assuming the singular value decomposition of Z is $Z = VDU^T$ where V is a $n \times d$ matrix and D, U are $d \times d$ matrices, and conditional on U, D, V is uniformly distributed on $V_{n,d}$. Therefore, we

have

$$A = n(X^T_{\tilde{\mathcal{M}}_d} X_{\tilde{\mathcal{M}}_d})^{-1} X^T_{\tilde{\mathcal{M}}_d} = n\Sigma^{1/2}_{\tilde{\mathcal{M}}_d}(Z^T Z)^{-1} Z^T = n\Sigma^{1/2}_{\tilde{\mathcal{M}}_d} U D^{-1} V^T.$$

We observe that

$$\|e_i^T A\|_2^2 = n^2 \Sigma^{1/2}_{\tilde{\mathcal{M}}_d} U D^{-2} U^T \Sigma^{1/2}_{\tilde{\mathcal{M}}_d} = n^2 (X^T_{\tilde{\mathcal{M}}_d} X_{\tilde{\mathcal{M}}_d})^{-1} \leq \frac{n}{\lambda_{min}(X^T_{\tilde{\mathcal{M}}_d} X_{\tilde{\mathcal{M}}_d}/n)}.$$

Next, following exactly the same argument in Lemma 3, we know that the distribution of A is invariant under the right orthogonal transformation and conditional on $v_i = \|e_i^T A\|_2$, $e_i^T A$ is uniformly distributed on $\mathcal{S}^{n-1}(v_i)$. Using the same inequality in (8), we have

$$P\bigg(\max_{ij} |A_{ij}| \leq \frac{t}{\sqrt{c_3 \lambda_{min}(X^T_{\tilde{\mathcal{M}}_d} X_{\tilde{\mathcal{M}}_d}/n)}} \bigg) \geq 1 - 2dn \exp(-t^2/2) - d\exp(-Cn).$$

Now the total number of possible $\tilde{\mathcal{M}}_d$ is bounded by $(p-s)\times(p-s-1)\times\cdots\times(p-d+1) \leq (p-s)^{(d-s)}$. Therefore, we have

$$P\bigg(\max_{|\tilde{\mathcal{M}}_d|=d, \mathcal{M}^* \subseteq \tilde{\mathcal{M}}_d} \max_{ij} |A_{ij}| \leq \frac{t}{\sqrt{c_3 \lambda_0}} \bigg) \geq 1 - 2dn(p-s)^{d-s} \exp\bigg(-\frac{t^2}{2} \bigg) - d(p-s)^{d-s} \exp(-Cn),$$

where $\lambda_0 = \min_{|\tilde{\mathcal{M}}_d|=d, \mathcal{M}^* \subseteq \tilde{\mathcal{M}}_d} \lambda_{min}(X^T_{\tilde{\mathcal{M}}_d} X_{\tilde{\mathcal{M}}_d}/n)$. □

Lemma 5 (Garvesh, Wainwright and Yu. (2010) [15])**.** *There exists some absolute constant $c', c'' > 0$ such that*

$$\frac{\|Xv\|_2}{\sqrt{n}} \geq \frac{1}{4}\|\Sigma^{\frac{1}{2}} v\|_2 - 9\rho(\Sigma)\sqrt{\frac{\log p}{n}}\|v\|_1, \quad \forall v \in \mathcal{R}^p,$$

with probability at least $1 - c''\exp(-c'n)$, where $\rho(\Sigma) = \max_{i=1,2,\cdots,p} \Sigma_{ii}$.

In our case, for any v with d nonzero coordinates, we have $\|v\|_1 \leq \sqrt{d}\|v\|_2$, $\rho(\Sigma) = 1$ and $\|\Sigma^{1/2} v\|_2 \geq \kappa^{-\frac{1}{2}}\|v\|_2$. Therefore,

$$\frac{\|Xv\|_2}{\sqrt{n}} \geq \bigg(\frac{\kappa^{-1/2}}{4} - 9\sqrt{\frac{d\log p}{n}} \bigg)\|v\|_2, \quad \|v\|_0 \leq d.$$

Proof of Theorem 2. Lemma 5 essentially states that for any $d \times d$ principal submatrix of X, we can bound its smallest eigenvalue. Therefore, for any selected submodel $\tilde{\mathcal{M}}_d$ from

the first stage, we have with probability at least $1 - O(\exp(-c'n))$

$$\min_{|\tilde{\mathcal{M}}_d|=d} \lambda_{min}^{\frac{1}{2}}(X_{\tilde{\mathcal{M}}_d}^T X_{\tilde{\mathcal{M}}_d}/n) \geq \frac{\kappa^{-1/2}}{4} - 9\sqrt{\frac{d\log p}{n}} \geq \frac{\kappa^{-1/2}}{8},$$

as long as $n \geq 6^4 \kappa d \log p$, i.e., $\lambda_0 \geq \frac{\kappa^{-1}}{64}$, where λ_0 is defined in Lemma 4.

A direct calculation shows that $\hat{\beta}^{(OLS)} = \beta + (X_{\tilde{\mathcal{M}}_d}^T X_{\tilde{\mathcal{M}}_d})^{-1} X_{\tilde{\mathcal{M}}_d}^T \varepsilon$. Therefore, we want to bound the error

$$\tilde{\eta} = (X_{\tilde{\mathcal{M}}_d}^T X_{\tilde{\mathcal{M}}_d})^{-1} X_{\tilde{\mathcal{M}}_d}^T \varepsilon = A\varepsilon/n.$$

Following the same argument as Lemma 3, we define $Z_j = (A_{1j}\varepsilon_j, \cdots, A_{dj}\varepsilon_j)$ and $\tilde{\eta} = \sum_{j=1}^n Z_j/n$. Using Proposition 1 and Lemma 4 we have with probability at least $1 - 2d(p-s)^{d-s}\exp(-t^2/2) - d(p-s)^{d-s}\exp(-Cn)$

$$E\Big\|\sum_{j=1}^n Z_j\Big\|_\infty^2 \leq \log d \sum_{j=1}^n E\|Z_j\|_\infty^2 \leq \frac{\sigma^2 n t^2 \log d}{c_3 \lambda_0} \leq 64 c_3^{-1} \kappa \sigma^2 t^2 n \log d. \tag{10}$$

Thus, for any $r > 0$

$$P\bigg(\|\tilde{\eta}\|_\infty \geq \frac{r}{n}\bigg) = P\bigg(\Big\|\sum_{j=1}^n Z_j\Big\|_\infty \geq r\bigg) \leq \frac{E\big\|\sum_{j=1}^n Z_j\big\|_\infty^2}{r^2} \leq \frac{64\kappa n\sigma^2 t^2 \log d}{c_3 r^2}.$$

If we take $t = \sqrt{2(\tilde{c}+3)\log p}$ for any $\delta \in (0,1)$, then it is ensured that

$$\begin{aligned}
&1 - 2dn(p-s)^{d-s}\exp\bigg(-\frac{t^2}{2}\bigg) - d(p-s)^{d-s}\exp(-Cn) \\
&\geq 1 - 2\exp\Big((\tilde{c}+2)\log p - (\tilde{c}+3)\log p\Big) - \exp\Big((\tilde{c}+1)\log p - Cn\Big) \\
&= 1 - O\bigg(\frac{1}{p}\bigg) \geq 1 - O\bigg(\frac{1}{n}\bigg).
\end{aligned}$$

Now taking $r = \sigma n^{1-\delta/2}$ for any $\delta \in (0,1)$ we have

$$P\bigg(\|\tilde{\eta}\|_\infty \leq \frac{\sigma}{n^{\delta/2}}\bigg) \geq 1 - O\bigg(\frac{\kappa \log p \log d}{n^{1-\delta}}\bigg). \tag{11}$$

Consequently, for any $\delta > 0$ we have

$$\|\hat{\beta}^{(OLS)} - \beta_{\tilde{\mathcal{M}}_d}\|_\infty \leq \frac{\sigma}{n^{\delta/2}}, \tag{12}$$

with probability at least $1 - O\Big(\frac{\kappa \log p \log d}{n^{1-\delta}}\Big)$. So if $\tau \geq \frac{2\sigma}{n^{\delta/2}}$, then by choosing $\gamma' = \frac{\sigma}{n^{\delta/2}}$ we

have

$$\min_{i\in S}|\hat{\beta}_i^{(OLS)}| \geq \gamma' \geq \max_{i\notin S}|\hat{\beta}_i^{(OLS)}|.$$

□

Proof of Theorem 3. Denoting $X_{\tilde{\mathcal{M}}_d}$ by X, the definition of $\hat{\beta}(r)^{(Ridge)}$ becomes

$$\begin{aligned}\hat{\beta}(r)^{(Ridge)} &= (X^TX + rI_d)^{-1}X^TX\beta + (X^TX + rI_d)^{-1}X^T\varepsilon \\ &= \beta - r(X^TX + rI_d)^{-1}\beta + (X^TX + rI_d)^{-1}X^T\varepsilon \\ &= \beta - \tilde{\xi}(r) + \tilde{\eta}(r).\end{aligned}$$

For $\tilde{\xi}(r)$ we have

$$\max|\tilde{\xi}(r)| \leq r^2\beta^T(X^TX + rI_d)^{-2}\beta \leq \frac{r^2\|\beta\|_2^2}{n^2\lambda_{min}^2(X^TX/n + r/n)} \leq \frac{8^4r^2\kappa^3M_0}{n^2}$$

with probability $1 - c''\exp(-c'n)$ if $n \geq 6^4\kappa d\log p$. This result is because of Lemma 5 and $M_0 \geq var(Y) \geq \|\beta\|_2^2\lambda_{max}(\Sigma)$.

For $\tilde{\eta}(r)$, we follow the same technique in the proof of Theorem 2. Basically, one just needs to show a similar result as Lemma 4 exists. Let $A = n(X^TX)^{-1}X^T$, which is the key quantity in Lemma 4, and $\tilde{A} = n(X^TX + rI_d)^{-1}X^T$. If we can show that $\tilde{A}$ does not differ too much from A, then the proof is completed. Consider the singular value decomposition directly on X as $X = VDU^T$ (not on Z), where V is a $n \times d$ matrix and D and U are $d \times d$ matrices. We then have

$$A = n(UD^2U^T)^{-1}UDV^T = nUD^{-1}V^T,$$

and

$$\tilde{A} = n(UD^2U^T + rI_d)^{-1}UDV^T = nUD^{-1}\left\{I_d + \frac{r}{n}\left(\frac{D}{\sqrt{n}}\right)^{-2}\right\}^{-1}V^T.$$

When $r \le n\lambda_{min}(X^TX/n)/2$, we can apply Taylor expansion on the inverse. Thus

$$\begin{aligned}\tilde{A} &= nUD^{-1}\Big\{I_d + \sum_{k=1}^{\infty}\Big(\frac{r}{n}\Big)^k\Big(\frac{D}{\sqrt{n}}\Big)^{-2k}\Big\}V^T \\ &= A + rUD^{-1}\Big(\frac{D}{\sqrt{n}}\Big)^{-2}V^T + nUD^{-1}\Big\{\sum_{k=2}^{\infty}\Big(\frac{r}{n}\Big)^k\Big(\frac{D}{\sqrt{n}}\Big)^{-2k}\Big\}V^T \\ &= A + \frac{rU(D/\sqrt{n})^{-3}V^T}{n^{1/2}} + nUD^{-1}\Big\{\sum_{k=2}^{\infty}\Big(\frac{r}{n}\Big)^k\Big(\frac{D}{\sqrt{n}}\Big)^{-2k}\Big\}V^T.\end{aligned}$$

Clearly, we have

$$\lambda_{max}\Big(\frac{rU(D/\sqrt{n})^{-3}V^T}{n^{1/2}}\Big) \le \frac{8^3 r\kappa^{3/2}}{\sqrt{n}},$$

and

$$\begin{aligned}\lambda_{max}\Big[nUD^{-1}\Big\{\sum_{k=2}^{\infty}\Big(\frac{r}{n}\Big)^k\Big(\frac{D}{\sqrt{n}}\Big)^{-2k}\Big\}V^T\Big] &\le \sqrt{n}\lambda_{min}^{-1}\Big(\frac{D}{\sqrt{n}}\Big)\sum_{k=2}^{\infty}\frac{r^k}{n^k}\lambda_{min}^{-k}\Big(\frac{D^2}{n}\Big) \\ &\le \sqrt{n}(8\kappa^{\frac{1}{2}})\sum_{k=2}^{\infty}\Big(\frac{8^2 r\kappa}{n}\Big)^k \le \frac{\sqrt{n}(8\kappa^{\frac{1}{2}})(\frac{8^2 r\kappa}{n})^2}{1-\frac{8^2 r\kappa}{n}} \\ &\le \frac{2\cdot 8^5\kappa^{\frac{5}{2}}r^2}{n^{3/2}}.\end{aligned}$$

The last inequality is because we assume $r \le n\lambda_{min}(X^TX/n)/2$. Together, we have

$$\|\tilde{A}\|_\infty \le \|A\|_\infty + \frac{8^3 r\kappa^{3/2}}{\sqrt{n}} + \frac{2\cdot 8^5\kappa^{\frac{5}{2}}r^2}{n^{3/2}},$$

with probability at least $1 - c''\exp(-c'n)$ if $n \ge 6^4\kappa d\log p$ and $r \le \frac{n}{128\kappa}$. In the proof of Theorem 2, the value of t in Lemma 4 is chosen to be $O(\log p)$. Thus, as long as $r \le O(\kappa^{-1}\sqrt{n})$, (10) and (11) hold for $\tilde{\eta}(r)$ as well, i.e., for any $\delta \in (0,1)$ we have

$$P\Big(\|\tilde{\eta}(r)\|_\infty \le \frac{\sigma}{n^{\delta/2}}\Big) \ge 1 - O\Big(\frac{\kappa\log p\log d}{n^{1-\delta}}\Big).$$

On the other hand, if we require $r \le 8^{-2}M_0^{-1/2}\kappa^{-3/2}\sigma^{1/2}n^{1-\delta/4}$, then we have

$$\max|\tilde{\xi}(r)| \le \frac{8^4 r^2\kappa^3 M_0}{n^2} \le \frac{\sigma}{n^{\delta/2}}.$$

Consequently, if the tuning parameter satisfies that

$$r \leq O\bigg\{ \min\bigg(\frac{\sqrt{n}}{\kappa},\ \frac{\sigma^{\frac{1}{2}} n^{1-\delta/4}}{8^2 M_0^{\frac{1}{2}} \kappa^{\frac{3}{2}}} \bigg) \bigg\},$$

and $n \geq 6^4 \kappa d \log p$, then we have

$$P\bigg(\|\hat{\beta}^{(Ridge)}(r) - \beta_{\tilde{\mathcal{M}}_d}\|_\infty \leq \frac{\sigma}{n^{\delta/2}} \bigg) \geq 1 - O\bigg(\frac{\kappa \log p \log d}{n^{1-\delta}} \bigg). \tag{13}$$

Therefore, if $\tau \geq \frac{4\sigma}{n^{\delta/2}}$, then by choosing $\gamma'(r) = \frac{2\sigma}{n^{\delta/2}}$ we have

$$\min_{i \in S} |\hat{\beta}_i^{(Ridge)}(r)| \geq \gamma' \geq \max_{i \notin S} |\hat{\beta}_i^{(Ridge)}(r)|.$$

□

Proof of Corollary 1. As mentioned before, we have $\hat{\beta}^{(OLS)} = \beta_{\tilde{\mathcal{M}}_d} + (X_{\tilde{\mathcal{M}}_d}^T X_{\tilde{\mathcal{M}}_d})^{-1} X_{\tilde{\mathcal{M}}_d} \varepsilon$. Because $\varepsilon_i \sim N(0, \sigma^2)$ for $i = 1, 2, \cdots, n$, we have for any $i \in \tilde{\mathcal{M}}_d$,

$$\tilde{\eta}_i = e_i^T (X_{\tilde{\mathcal{M}}_d}^T X_{\tilde{\mathcal{M}}_d})^{-1} X_{\tilde{\mathcal{M}}_d}^T \varepsilon \sim N(0, \sigma^2 e_i^T (X_{\tilde{\mathcal{M}}_d}^T X_{\tilde{\mathcal{M}}_d})^{-1} e_i) \overset{(d)}{=} \sigma \sqrt{e_i^T (X_{\tilde{\mathcal{M}}_d}^T X_{\tilde{\mathcal{M}}_d})^{-1} e_i} N(0,1). \tag{14}$$

Likewise in the proof of Lemma 4, we know that as long as $n \geq 64 \kappa d \log p$

$$\lambda_{min}(X_{\tilde{\mathcal{M}}_d}^T X_{\tilde{\mathcal{M}}_d}/n) \geq \frac{1}{64\kappa}.$$

Thus, we have

$$\max_{i \in \tilde{\mathcal{M}}_d} e_i^T (X_{\tilde{\mathcal{M}}_d}^T X_{\tilde{\mathcal{M}}_d})^{-1} e_i \leq 64\kappa/n.$$

Therefore, for any $t > 0$ and $i \in \tilde{\mathcal{M}}_d$, with probability at least $1 - c'' \exp(-c'n) - 2\exp(-t^2/2)$ we have

$$|\tilde{\eta}_i| \leq \sigma t \sqrt{e_i^T (X_{\tilde{\mathcal{M}}_d}^T X_{\tilde{\mathcal{M}}_d})^{-1} e_i} \leq \frac{8\kappa^{\frac{1}{2}} \sigma t}{\sqrt{n}}.$$

Then for any $\delta > 0$, if $n > \log(2c''/\delta)/c'$, then with probability at least $1 - \delta$ we have

$$\max_{i \in \tilde{\mathcal{M}}_d} |\tilde{\eta}_i| \leq 8\sigma \sqrt{\frac{2\kappa \log(4d/\delta)}{n}}. \tag{15}$$

Because σ needs to estimated from the data, we need to obtain a bound as well. Notice that

$\hat{\sigma}^2$ is an unbiased estimator for σ, and

$$\hat{\sigma}^2 = \sigma^2 \epsilon^T (I_n - X_{\tilde{\mathcal{M}}_d} (X_{\tilde{\mathcal{M}}_d}^T X_{\tilde{\mathcal{M}}_d})^{-1} X_{\tilde{\mathcal{M}}_d}) \epsilon \sim \frac{\sigma^2 \mathcal{X}^2(n-d)}{n-d},$$

where $\mathcal{X}^2(k)$ denotes a chi-square random variable with degree of freedom k. Using Proposition 5.16 in [24], we can bound $\hat{\sigma}^2$ as follows. Let $K = \|\mathcal{X}^2(1) - 1\|_{\psi_1}$. There exists some $c_5 > 0$ such that for any $t \geq 0$ we have,

$$P\left(\left| \frac{\mathcal{X}^2(n-d)}{n-d} - 1 \right| \geq t \right) \leq 2 \exp\left\{ - c_5 \min\left(\frac{t^2(n-d)}{K^2}, \frac{t(n-d)}{K} \right) \right\}.$$

Hence for any $\delta > 0$, if $n > d + 4K^2 \log(2/\delta)/c_5$, then with probability at least $1 - \delta$ we have,

$$|\hat{\sigma}^2 - \sigma^2| \leq \sigma^2/2,$$

which implies that

$$\frac{1}{2}\sigma^2 \leq \hat{\sigma}^2 \leq \frac{3}{2}\sigma^2.$$

Then we know that

$$\max_{i \in \tilde{\mathcal{M}}_d} |\tilde{\eta}_i| \leq 8\sigma \sqrt{\frac{2\kappa \log(4d/\delta)}{n}} \leq 8\sqrt{2}\hat{\sigma} \sqrt{\frac{2\kappa \log(4d/\delta)}{n}} \leq 8\sqrt{3}\sigma \sqrt{\frac{2\kappa \log(4d/\delta)}{n}}.$$

Now define $\gamma' = 8\sqrt{2}\hat{\sigma}\sqrt{\frac{2\kappa \log(4d/\delta)}{n}}$. If the signal $\tau = \min_{i \in S} |\beta_i|$ satisfies that

$$\tau \geq 24\sigma \sqrt{\frac{2\kappa \log(4d/\delta)}{n}},$$

then with probability at least $1 - 2\delta$, for any $i \notin S$

$$|\hat{\beta}_i| = |\tilde{\eta}_i| \leq 8\sigma \sqrt{\frac{2\kappa \log(4d/\delta)}{n}} \leq \gamma',$$

and for $i \in S$ we have

$$|\hat{\beta}_i| \geq \tau - \max_{i \in \tilde{\mathcal{M}}_d} |\tilde{\eta}_i| \geq 16\sigma \sqrt{\frac{2\kappa \log(4d/\delta)}{n}} \geq \gamma'.$$

□

Proof of Theorem 4

The result of Theorem 4 can be immediately implied from Theorem 1, 2, 3, (12) and (13).