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

Appendix 0: Proof of Lemma 1

Applying the Sherman-Morrison-Woodbury formula

$$(A + UDV)^{-1} = A^{-1} - A^{-1}U(D^{-1} + VA^{-1}U)^{-1}VA^{-1},$$

we have

$$r(rI_p + X^TX)^{-1} = I_p - X^T(I_n + \frac{1}{r}XX^T)^{-1}X\frac{1}{r} = I_p - X^T(rI_n + XX^T)^{-1}X.$$

Multiplying X^TY on both sides, we get

$$r(rI_p + X^TX)^{-1}X^TY = X^TY - X^T(rI_n + XX^T)^{-1}XX^TY.$$

The right hand side can be further simplified as

$$\begin{aligned} &X^TY - X^T(rI_n + XX^T)^{-1}XX^TY \\ &\quad = X^TY - X^T(rI_n + XX^T)^{-1}(rI_n + XX^T - rI_n)Y \\ &\quad = X^TY - X^TY + r(rI_n + XX^T)^{-1}Y = rX^T(rI_n + XX^T)^{-1}Y. \end{aligned}$$

Therefore, we have

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

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 three 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 $i \in Q, j \neq i$,*

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

and

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

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

The proof can be found in the Lemma 4 and 5 in Wang and Leng (2015) for elliptical distributions. The special case of Gaussian is also proved in the Lemma 3 of Wang et al. (2015). Notice that the eigenvalue assumption in Wang and Leng (2015) is not used for proving Lemma 4 and 5.

Lemma 3. *Assume x_i follows $EN(L, \Sigma)$. If $E[L^{-2}] < M_1$ for some constant $M_1 > 0$, $var(\epsilon) = \sigma^2$ and $\log p = o(n)$, then for any $0 < \alpha < 1$ we have*

$$P\left(\|\eta\|_\infty \leq \frac{c_1\kappa^{-1}\tau^*}{6}\frac{n}{p}\right) \geq 1 - O\left(\frac{\sigma^2\kappa^4\log p}{\tau^{*2}n^{1-\alpha}}\right),$$

where τ^* is defined as the minimum value for the important signals and $\kappa = cond(\Sigma)$.

To prove Lemma 3 we need the following two propositions.

Proposition 1. *(Lounici, 2008 Lounici (2008); Nemirovski, 2000 Akritas et al. (2014)) 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*

$$E\Big\|\sum_{i=1}^n Y_i\Big\|_k^2 \leq \tilde{C}\min\{k, \log p\}\sum_{i=1}^n E\|Y_i\|_k^2, \tag{3}$$

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

As each row of X can be represented as $X = \bar{L}Z\Sigma^{1/2}$, where $\bar{L} = diag(\sqrt{p}L_1/\|z_1\|_2, \cdots, \sqrt{p}L_n/\|z_n\|_2)$ and Z is a matrix of independent Gaussian entries, i.e., $Z \sim N(0, I_p)$. 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*

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

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

$$P\left(\lambda_{min}(ZZ^T/p) > c\right) \geq 1 - 2\exp(-n/2), \tag{4}$$

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

The proof follows Corollary 5.35 in Vershynin (2010).

Proof of Lemma 3. Let $A = pX^T(XX^T)^{-1}\bar{L}$ and $Z = \bar{L}^{-1}X\Sigma^{-1/2}$. Then $\eta = p^{-1}A\bar{L}^{-1}\epsilon$.

Part 1. Bounding $|A_{ij}|$. 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

$$A = pX^T(XX^T)^{-1}L = p\Sigma^{\frac{1}{2}}Z^TL(LZ\Sigma Z^TL)^{-1}L = p\Sigma^{\frac{1}{2}}UDV^TL(LVDU^T\Sigma UDV^TL)^{-1}L$$
$$= p\Sigma^{\frac{1}{2}}U(U^T\Sigma U)^{-1}D^{-1}V^T = \sqrt{p}\Sigma^{\frac{1}{2}}U(U^T\Sigma U)^{-1}(\frac{D}{\sqrt{p}})^{-1}V^T.$$

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

$$A \overset{(d)}{=} AT. \tag{5}$$

Our first goal is to bound the magnitude of individual entries A_{ij}. Let $v_i = e_i^TAA^Te_i$, which is a function of U and D (see below). From (5), we know that e_i^TA 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

$$e_i^TA \overset{(d)}{=} \sqrt{v_i}\Bigg(\frac{x_1}{\sqrt{\sum_{j=1}^n x_j^2}}, \frac{x_2}{\sqrt{\sum_{j=1}^n x_j^2}}, \cdots, \frac{x_n}{\sqrt{\sum_{j=1}^n x_j^2}}\Bigg), \tag{6}$$

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

$$v_i = e_i^TAA^Te_i = pe_i^T\Sigma^{\frac{1}{2}}U(U^T\Sigma U)^{-1}(\frac{D^2}{p})^{-1}(U^T\Sigma U)^{-1}U^T\Sigma^{\frac{1}{2}}e_i.$$
$$= pe_i^TH(U^T\Sigma U)^{-\frac{1}{2}}(\frac{D^2}{p})^{-1}(U^T\Sigma U)^{-\frac{1}{2}}H^Te_i$$
$$\leq pe_i^THH^Te_i \cdot \lambda_{min}^{-1}(U^T\Sigma U) \cdot \lambda_{min}^{-1}(\frac{D^2}{p})$$

Here $H = \Sigma^{\frac{1}{2}}U(U^T\Sigma U)^{-1/2}$ is defined the same as in Wang and Leng (2015) and can be bounded as $e_i^THH^Te_i \leq c_2n\kappa/p$ with probability $1 - 2\exp(-Cn)$ (see the proof of Lemma 3 in Wang et al. (2015)). Therefore, we have

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

Now applying the tail bound and the concentration inequality to (6) we have for any $t > 0$ and any $C > 0$

$$P(|x_j| > t) \leq 2\exp(-t^2/2) \qquad P\Bigg(\frac{\sum_{j=1}^n x_j^2}{n} \leq c_3\Bigg) \leq \exp(-Cn). \tag{7}$$

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

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

Now according to (4), we can further bound $\lambda_{min}(D^2/p)$ and obtain that

$$P\Bigg(\max_{ij}|A_{ij}| \leq \sqrt{\frac{c_2}{cc_3}}\kappa t\Bigg) \geq 1 - 2np\exp(-t^2/2) - 3p\exp(-Cn) - 2\exp(-n/2). \tag{8}$$

Part 2. Bounding η he second step is to use (8) and Proposition 1 to bound η. The procedure follows similarly as in Lounici's paper. We first note that $\|z_i\|_2^2$ follows a chi-square distribution $\mathcal{X}^2(p)$. We have for any t

$$P\bigg(\frac{\|z_i\|_2^2}{p} \geq 1 + 2\sqrt{\frac{t}{p}} + \frac{2t}{p}\bigg) \leq e^{-t},$$

from which we know

$$P\bigg(\max_i p^{-1}\|z_i\|_2^2 < 5/2\bigg) \geq 1 - pe^{-p/4}. \tag{9}$$

Now define $W_j = (A_{1j}p^{-1/2}\|z_j\|_2 L_j^{-1}\epsilon_j, A_{2j}p^{-1/2}\|z_j\|_2 L_j^{-1}\epsilon_j, \cdots, A_{pj}p^{-1/2}\|z_j\|_2 L_j^{-1}\epsilon_j)$. It's clear that $\eta = \sum_{j=1}^n W_j/p$. Applying Proposition 1 to $W_j's$ with the l_∞ norm and noticing tht L_j is independent of z_j we have

$$E\|\sum_{j=1}^n W_j\|_\infty^2 \leq \log p \sum_{j=1}^n E\|W_j\|_\infty^2 \leq \log p \frac{7c_2}{cc_3}\sigma^2\kappa^2 t^2 \sum_{j=1}^n E[L_j^{-2}] \leq \frac{c_2}{cc_3}\sigma^2\kappa^2 t^2 M_1^2 n \log p.$$

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

$$\begin{aligned} P\bigg(\|\eta\|_\infty \geq \frac{\sqrt{n}r}{p}\bigg) &= P\bigg(\frac{p}{\sqrt{n}}\|\eta\|_\infty \geq r\bigg) \leq \frac{p^2 E\|\eta\|_\infty^2}{nr^2} = \frac{E\|\sum_{j=1}^n W_j\|_\infty^2}{nr^2} \\ &\leq \frac{7c_2\sigma^2\kappa^2 M_1^2 t^2 \log p}{cc_3 r^2}. \end{aligned}$$

To match our previous result, we take $r = c_1\sqrt{n}\tau^*\kappa^{-1}/6$ and $t = n^{(1-\alpha)/2}$ for some small α,

$$\begin{aligned} P\bigg(\|\eta\|_\infty \leq \frac{c_1\kappa^{-1}\tau^*}{6}\frac{n}{p}\bigg) &\geq 1 - \frac{342c_2\sigma^2\kappa^4 M_1}{c_1^2 cc_3\tau^{*2}}\frac{\log p}{n^\alpha} - 2np\exp(-n^{1-\alpha}/2) - 3p\exp(-Cn) - 2\exp(-n/2) \\ &\geq 1 - O\bigg(\frac{\sigma^2\kappa^4\log p}{\tau^{*2}n^\alpha}\bigg). \end{aligned}$$

□

Lemma 4. *Assume $var(Y) \leq M_0$. Define $\Phi = X^T(XX^T)^{-1}X$. If $p > c_0 n$ for some $c_0 > 1$, then we have for any $t > 0$*

$$P\bigg(\max_i \sum_{j\neq i}|\Phi_{ij}\beta_j| \geq c_4\sqrt{M_0}\kappa^{\frac{3}{2}}t\frac{\sqrt{n}}{p}\bigg) \leq 2pe^{-t^2/2} + 5pe^{-Cn}.$$

where c_4, κ are defined in Lemma 2.

Proof of Lemma 4. Following Wang and Leng (2015); Wang et al. (2015), we define $H = X^T(XX^T)^{-\frac{1}{2}}$. When $X \sim N(0, \Sigma)$, H follows the $MACG(\Sigma)$ distribution as indicated in Lemma 3 in Wang et al. (2015) and Theorem 1 in Wang and Leng (2015). For simplicity, we only consider a particular case where $i = 1$.

For vector v with $v_1 = 0$, we define $v' = (v_2, v_3, \cdots, v_p)^T$ and we can always identify a $(p-1)\times(p-1)$ orthogonal matrix T' such that $T'v' = \|v'\|_2 e_1'$ where e_1' is a $(p-1)\times 1$ unit vector with the first coordinate being 1. Now we define a new orthogonal matrix T as

$$T = \begin{pmatrix} 1 & 0 \\ 0 & T' \end{pmatrix}$$

and we have

$$Tv = \begin{pmatrix} 1 & 0 \\ 0 & T' \end{pmatrix} \begin{pmatrix} 0 \\ v' \end{pmatrix} = \begin{pmatrix} 0 \\ \|v\|_2 e_1' \end{pmatrix} = \|v\|_2 e_2. \quad \text{and} \quad e_1^T T^T = e_1^T \begin{pmatrix} 1 & 0 \\ 0 & T'^T \end{pmatrix} = e_1^T$$

Therefore, we have

$$e_1^T H H^T v = e_1^T T^T T H H^T T^T T v = e_1^T T^T H H^T T^T e_2 = \|v\|_2 e_1^T \tilde{H}\tilde{H}^T e_2.$$

Since H follows $MACG(\Sigma)$, $\tilde{H} = T^T H$ follows $MACG(T^T \Sigma T)$ for any fixed T. Therefore, we can apply Lemma 2 again to obtain that

$$\begin{aligned} &P\bigg(|e_1^T X^T (XX^T)^{-1} X v| \geq \|v\|_2 c_4 \kappa t \frac{\sqrt{n}}{p}\bigg) = P\bigg(|e_1^T H H^T v| \geq \|v\|_2 c_4 \kappa t \frac{\sqrt{n}}{p}\bigg) \\ &= P\bigg(\|v\|_2 |e_1^T \tilde{H}\tilde{H}^T e_2| \geq \|v\|_2 c_4 \kappa t \frac{\sqrt{n}}{p}\bigg) = P\bigg(\|v\|_2 |\Phi_{12}| \geq \|v\|_2 c_4 \kappa t \frac{\sqrt{n}}{p}\bigg) \\ &= P\bigg(|\Phi_{12}| \geq c_4 \kappa t \frac{\sqrt{n}}{p}\bigg) \leq 5e^{-Cn} + 2e^{-t^2/2}. \end{aligned}$$

Applying the above result to $v = (0, \beta_*^{(-1)})$ we have

$$\sum_{j \neq 1} |\Phi_{1j}\beta_j| \leq c_4 \kappa t \|\beta\|_2 \frac{\sqrt{n}}{p}$$

with probability at least $1 - 5e^{-Cn} - 2e^{-t^2/2}$.

In addition, we know that $var(Y) = \beta_*^T \Sigma \beta_* + \sigma^2 \leq M_0$ and thus

$$\|\beta\|_2 \leq \sqrt{M_0 \kappa}.$$

Consequently, we have

$$P\bigg(\max_i \sum_{j \neq i} |\Phi_{ij}\beta_j| \geq c_4 \sqrt{M_0} \kappa^{\frac{3}{2}} t \frac{\sqrt{n}}{p}\bigg) \leq 2pe^{-t^2/2} + 5pe^{-Cn}.$$

□

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 we have

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

For the first term, we have

$$|\min_{ii} \beta_i| \geq c_1 \kappa^{-1} \tau^* \frac{n}{p} \quad \forall i \in S^*$$

with probability $1 - |S^*| e^{-Cn}$ and

$$|\min_{ii} \beta_i| \leq c_1 \kappa \tau_* \frac{n}{p} \quad \forall i \in S_*$$

with probability $1-|S_*|e^{-Cn}$. Now, for the second term, using Lemma 4, we have

$$\sum_{j\neq i}|\Phi_{ij}\beta_j| \leq \frac{c_1\kappa^{-1}\tau^*}{6} \quad \forall i=1,2,\cdots,p$$

with probability at least $1-2p\exp\{-\frac{c_1^2\kappa^{-1}\tau^{*2}}{72c_4^2M_0}n\}-5pe^{-Cn}$. Therefore, we have for any $i\in S^*$

$$|\xi_i| \geq c_1\kappa^{-1}\tau^*\frac{n}{p}-\frac{c_1\kappa^{-1}\tau^*}{6}\frac{n}{p} \geq \frac{5c_1\kappa^{-1}\tau^*}{6}\frac{n}{p}.$$

and for $i\in S_*$ we have

$$|\xi_i| \leq c_1\kappa\tau_*\frac{n}{p}+\frac{c_1\kappa^{-1}\tau^*}{6}\frac{n}{p} \leq \frac{7c_1\kappa^{-1}\tau^*}{12}\frac{n}{p},$$

where we use the assumption that $\tau^* > 4\kappa^2\tau_*$. Now combining the result from Lemma 3, we can obtain

$$P\bigg(\min_{i\in S^*}|\hat{\beta}_i| \geq \frac{2c_1\kappa^{-1}\tau^*}{3}\frac{n}{p}\bigg) \geq 1-O\bigg(\frac{\sigma^2\kappa^4\log p}{\tau^{*2}n^{\alpha}}\bigg),$$

and

$$P\bigg(\max_{i\in S_*}|\hat{\beta}_i| \leq \frac{7c_1\kappa^{-1}\tau^*}{12}\frac{n}{p}\bigg) \geq 1-O\bigg(\frac{\sigma^2\kappa^4\log p}{\tau^{*2}n^{\alpha}}\bigg).$$

Taking $\gamma=\frac{2c_1\kappa^{-1}\tau^*}{3}np$, we have

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

□

Proof of Theorem 2 and 3

For the selected submodel $\hat{\mathcal{M}}_d$, we define X_d to be the variables contained in $\hat{\mathcal{M}}_d$ and $X_{d,c}$ to be variables that are excluded from $\hat{\mathcal{M}}_d$. It is clear that

$$\hat{\beta}_d^{(OLS)}=(X_d^TX_d)^{-1}X_d^TY=\beta_d+(X_d^TX_d)^{-1}X_d^T\varepsilon+(X_d^TX_d)^{-1}X_d^TX_{d,c}\beta_{d,c}=\beta_d+\eta_d+\omega.$$

To prove Theorem 2 is essentially to bound η and ω. Thus, we need following three lemmas.

Lemma 5 (Garvesh, Wainwright and Yu. (2010) Raskutti et al. (2010)). *Assume $Z\sim N(0,\Sigma)$. There exists some absolute constant $c',c''>0$ such that*

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

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

$\|\Sigma^{1/2}v\|_2 \geq \lambda_{\min}^{\frac{1}{2}}(\Sigma)\|v\|_2$. *Therefore,*

$$\frac{\|Zv\|_2}{\sqrt{n}} \geq \left(\frac{\lambda_{\min}^{\frac{1}{2}}(\Sigma)}{4} - 9\sqrt{\frac{d\log p}{n}}\right)\|v\|_2, \quad \|v\|_0 \leq d.$$

Thus, as long as $n \geq 6^4\kappa d\log p$, *we have*

$$\min_{|\hat{\mathcal{M}}|\leq d} \lambda_{min}^{1/2}(Z_{\hat{\mathcal{M}}}^T Z_{\hat{\mathcal{M}}}/n) \geq \frac{\lambda_{\min}^{\frac{1}{2}}(\Sigma)}{8}.$$

Lemma 6. *Assume* $E[L^{-12}] \leq M_1$ *and* $e[L^{12}] \leq M_2$. *For any* $\hat{\mathcal{M}}$ *such that* $S^* \subset \hat{\mathcal{M}}$ *and* $|\hat{\mathcal{M}}| \leq d$, *we have for any* $\alpha > 0$

$$P\left(\max_{|\hat{\mathcal{M}}|\leq d} \|\eta_d\|_\infty \leq \sigma\sqrt{\frac{\log p}{n^\alpha}}\right) = 1 - O\left(\frac{\lambda_*^{-2} d\log d}{n^{\frac{1}{3}(1-\alpha)}} + \frac{M_1 + M_2}{n^{\frac{1}{3}(1-4\alpha)}}\right),$$

where $\lambda_* = \lambda_{\min}(\Sigma)$.

Proof of Lemma 6. Define $A = (X_d^T X_d)^{-1}X_d^T$, we have

$$\eta = (X_d^T X_d)^{-1}X_d^T\epsilon = A\epsilon.$$

For A, we can bound its entries as

$$\begin{aligned}\max_{ij}|A_{ij}| &\leq \max_{ij}|e_i^T(X_d^T X_d)^{-1}X_d^T e_j| \leq \max_{ij}\|e_i^T(X_d^T X_d)^{-1}\|_1\|X_d^T e_j\|_\infty \\ &\leq \sqrt{d}\max_{ij}\|e_i^T(X_d^T X_d)^{-1}\|_2 \max_{ij}|X_d^T| \leq \frac{\sqrt{d}}{n}\lambda_{min}^{-1}\left(\frac{X_d^T X_d}{n}\right)\max_{ij}|X_d^T|.\end{aligned}$$

Recall that $X = \bar{L}Z\Sigma^{1/2}$, where $\bar{L} = diag(\sqrt{p}L_1/\|z_1\|_2, \cdots, \sqrt{p}L_n/\|z_n\|_2)$ and thus X_d possesses a representation as $X_d = \bar{L}Z\Sigma_d^{1/2}$, where $\Sigma_d^{1/2}$ is an $p \times d$ matrix formed by the selected d columns of $\Sigma^{1/2}$. We can now further bound $\lambda_{min}^{-1}\left(\frac{X_d^T X_d}{n}\right)$ as

$$\begin{aligned}\lambda_{min}^{-1}\left(\frac{X_d^T X_d}{n}\right) &= \lambda_{min}^{-1}\left(\frac{\Sigma_d^{\frac{T}{2}} Z^T\bar{L}^T\bar{L}Z\Sigma_d^{\frac{1}{2}}}{n}\right) \\ &\leq \left(\lambda_{\min}(\bar{L}^T\bar{L})\lambda_{\min}(\Sigma_d^{\frac{T}{2}}Z^T Z\Sigma_d^{\frac{1}{2}}/n)\right)^{-1}.\end{aligned}$$

Using Lemma 5, it is clear that

$$\min_{|\hat{\mathcal{M}}|\leq d}\lambda_{\min}(\Sigma_d^{\frac{T}{2}}Z^T Z\Sigma_d^{\frac{1}{2}}/n) \geq \frac{\lambda_{\min}(\Sigma)}{64} \geq \frac{\lambda_*}{64},$$

with probability at least $1 - O(e^{-c'n})$. In addition, since $E[L^{-12}] \leq M_1$ and $E[L^{12}] \leq M_2$, we have for any $k_1 > 0, k_2 > 0$

$$P(L^2 \leq k_1) \leq k_1^6 M_1 \quad \text{and} \quad P(L \geq k_2) \leq \frac{M_2}{k_2^{12}}.$$

Combining with equation (9) implies that

$$\lambda_{\min}(\bar{L}^T\bar{L}) \geq \frac{2k_1}{5},$$

with probability at least $1 - pe^{-p/4} - nk_1^6 M_1$. Therefore, we have

$$\max_{|\hat{\mathcal{M}}|\leq d} \lambda_{min}^{-1}\left(\frac{X_d^T X_d}{n}\right) \leq \frac{162}{\lambda_* k_1}.$$

with probability $1 - O(nk_1^6 M_1)$.

For $\max_{ij} |X_d^T|$, we just need to bound $\max_{ij} X_{ij}$. Using the representation $X = \bar{L}Z\Sigma^{1/2}$, we know that

$$X_{ij} = \frac{\sqrt{p}L_i}{\|z_i\|_2} Z_i \Sigma^{1/2} e_j.$$

It is easy to see that $Z_i\Sigma^{1/2}e_j$ is a Gaussian random variable with mean zero and variance 1, thus for any $t > 0$

$$P(|Z_i\Sigma^{1/2}e_j| \geq t) \leq 2e^{-t^2/2}.$$

In addition, $\|z_i\|_2^2/p$ follows a $\mathcal{X}^2(p)$ and we have

$$P\left(\frac{\|z_i\|_2^2}{p} \geq 1 - 2\sqrt{\frac{t}{p}}\right) \geq 1 - e^{-t}.$$

Taking $t = p/4$, we have $\max_i \|z_i\|_2/\sqrt{p} \geq 1/2$ with probability at least $1 - ne^{-p/4}$ and thus

$$P(\max_{ij} |X_{ij}| \leq 4k_2\sqrt{\log p}) \geq 1 - \frac{M_2 n}{k_2^{12}} - 2p^{-1} - ne^{-p/4}.$$

Combining all pieces of results, we obtain that

$$P\left(\min_{|\hat{\mathcal{M}}|\leq d} \max_{ij} |A_{ij}| \leq \frac{648k_2\sqrt{d}\sqrt{\log p}}{\lambda_* k_1 n}\right) \geq 1 - O\left(nk_1^6 M_1 + \frac{nM_2}{k_2^{12}}\right).$$

Following a similar argument in proving Lemma 3, we define $W_j = (A_{1j}\epsilon_j, A_{2j}\epsilon_j, \cdots, A_{dj}\epsilon_j)$ and then

$$\eta = \sum_{j=1}^{n} W_j.$$

Using Proposition 1, we have

$$E\|\eta\|_\infty^2 = E\|\sum_{j=1}^{n} W_j\|_\infty^2 \leq \tilde{C}\log d \sum_{j=1}^{n} E\|W_j\|_\infty^2 \leq O\left(\frac{\sigma^2 k_2^2}{\lambda_*^2 k_1^2}\frac{d\log d\log p}{n}\right).$$

Using the Markov inequality implies that for any $r > 0$

$$P\left(\max_{|\hat{\mathcal{M}}|\leq d} \|\eta\|_\infty > r\right) \leq \frac{\|\eta\|_\infty^2}{r^2} = O\left(\frac{\sigma^2 k_2^2}{\lambda_*^2 k_1^2 r^2}\frac{d\log d\log p}{n}\right) + O\left(nk_1^6 M_1 + \frac{nM_2}{k_2^{12}}\right).$$

Let $r = \sigma\sqrt{\frac{\log p}{n^\alpha}}$, $k_1 = n^{-\frac{2(1-\alpha)}{9}}$ and $k_2 = n^{\frac{1-\alpha}{9}}$, we have

$$P\left(\max_{|\hat{\mathcal{M}}|\le d} \|\eta\|_\infty \le \sigma\sqrt{\frac{\log p}{n^\alpha}}\right) = 1 - O\left(\frac{\lambda_*^{-2} d\log d}{n^{\frac{1}{3}(1-\alpha)}} + \frac{M_1+M_2}{n^{\frac{1}{3}(1-4\alpha)}}\right)$$

□

Lemma 7. *Assume $E[L^{-12}] \le M_1$ and $e[L^{12}] \le M_2$. For any $\hat{\mathcal{M}}$ such that $S^* \subset \hat{\mathcal{M}}$ and $|\hat{\mathcal{M}}| \le d$. Assume that $d - |S^*| \le \tilde{c}$ and $\sum_{i\notin S^*} |\beta_i|^\iota \le R$ for some $\iota \in (0,1)$, then for any $\alpha > 0$, we have*

$$P\left(\max_{|\hat{\mathcal{M}}|\le d} \|w\|_2 \le \sigma\sqrt{\frac{\log p}{n^\alpha}}\right) \ge 1 - O\left(\frac{(M_1+M_2)R^3}{(\log p)^{2\iota} n^{3-4\alpha-2\iota}}\right).$$

Proof of Lemma 7. According to our definition that $\omega = (X_d^T X_d)^{-1} X_d^T X_{d,c}\beta_{d,c}$, we can directly bound the l_2 norm of ω as

$$\|\omega\|_2^2 = \beta_{d,c}^T X_{d,c}^T X_d (X_d^T X_d)^{-2} X_d^T X_{d,c}\beta_{d,c} \le \frac{1}{n}\beta_{d,c}^T X_{d,c}^T X_{d,c}\beta_{d,c}\lambda_{min}^{-1}\left(\frac{X_d^T X_d}{n}\right)$$

where $\lambda_{min}^{-1}\left(\frac{X_d^T X_d}{n}\right)$ has already obtained a bound in Lemma 6 as

$$\max_{|\hat{\mathcal{M}}|\le d} \lambda_{min}^{-1}\left(\frac{X_d^T X_d}{n}\right) \le \frac{162}{\lambda_* k_1}.$$

with probability $1 - O(nk_1^6 M_1)$. Now for $\frac{1}{n}\beta_{d,c}^T X_{d,c}^T X_{d,c}\beta_{d,c}$ we have

$$\frac{1}{n}\beta_{d,c}^T X_{d,c}^T X_{d,c}\beta_{d,c} = \frac{1}{n}\beta_{d,c}^T \Sigma_{d,c}^{T/2} Z^T \bar{L}^T \bar{L} Z \Sigma_{d,c}^{1/2}\beta_{d,c} \le \frac{1}{n}\beta_{d,c}^T \Sigma_{d,c}^{T/2} Z^T Z \Sigma_{d,c}^{1/2}\beta_{d,c} \max_i \frac{pL_i^2}{\|z_i\|_2^2}$$

Since $Z \sim N(0, I_p)$, we can choose an orthogonal matrix Q such that $\beta_{d,c}\Sigma_{d,c}^{1/2} = e_1 Q\|\beta_{d,c}\Sigma_{d,c}^{1/2}\|_2$ and

$$\frac{1}{n}\beta_{d,c}^T \Sigma_{d,c}^{T/2} Z^T Z \Sigma_{d,c}^{1/2}\beta_{d,c} = \|\beta_{d,c}\Sigma_{d,c}^{1/2}\|_2^2 e_1 \tilde{Z}^T \tilde{Z} e_1^T \le \|\beta_{d,c}\|_2^2 \lambda^* e_1 \tilde{Z}^T \tilde{Z} e_1,$$

where $\tilde{Z} \sim N(0, I_p)$. It is easy to see that for any $t > 0$

$$P\left(\frac{e_1^T \tilde{Z}^T \tilde{Z} e_1}{n} \le 1 + 2\sqrt{\frac{t}{n}} + \frac{2t}{n}\right) \ge 1 - e^{-t}.$$

and $\|\beta_{d,c}\|_2^2 \le \tau_*^{2-\iota} R$. Thus, taking $t = (1+\tilde{c})\log p$, we have

$$\max_{|\hat{\mathcal{M}}|\le d} \frac{1}{n}\beta_{d,c}^T \Sigma_{d,c}^{T/2} Z^T Z \Sigma_{d,c}^{1/2}\beta_{d,c} \le 5\tau_*^{2-\iota} R\lambda^*$$

with probability $1 - p^{-1}$ as long as $n \ge (1+\tilde{c})\log p$ where $\tilde{c}$ is the upper bound on $d - |S^*|$. For $\max_i pL_i^2/\|z_i\|_2^2$, we follow the same argument in Lemma 6

$$P\left(\max_i \frac{pL_i^2}{\|z_i\|_2^2} \le 2k_2^2\right) \ge 1 - ne^{-p/4} - \frac{nM_2}{k_2^{12}}.$$

Putting all pieces together, we have

$$\max_{|\hat{\mathcal{M}}|\le d} \|w\|_2 \le 36\tau_*^{1-\frac{\iota}{2}} R^{\frac{1}{2}} \kappa^{\frac{1}{2}} \sqrt{\frac{k_2^2}{k_1}},$$

with probability at least $1 - O\bigg(\frac{nM_2}{k_2^{12}} + nk_1^6 M_1\bigg)$. According to our assumption that $\tau_* \le \frac{\sigma}{\kappa}\sqrt{\frac{\log p}{n}}$ and taking $k_1 = \frac{n^{\iota/4}R^{1/2}}{(\log p)^{\iota/4} n^{(1-\alpha)/2}}$ and $k_2 = 1/\sqrt{k_1}$ we have

$$P\bigg(\max_{|\hat{\mathcal{M}}|\le d} \|w\|_2 \le \sigma\sqrt{\frac{\log p}{n^\alpha}}\bigg) \ge 1 - O\bigg(\frac{(M_1+M_2)R^3}{(\log p)^{2\iota} n^{3-4\alpha-2\iota}}\bigg).$$

□

We are now ready to prove Theorem 2

Proof of Theorem 2. We just need to combine the results of Lemma 6 and 7, i.e.,

$$\hat{\beta}_d^{(OLS)} = \beta_d + \eta + \omega,$$

where

$$P\bigg(\max_{|\hat{\mathcal{M}}|\le d} \|\eta\|_\infty \le \sigma\sqrt{\frac{\log p}{n^\alpha}}\bigg) = 1 - O\bigg(\frac{\lambda_*^{-2} d\log d}{n^{\frac{1}{3}(1-\alpha)}} + \frac{M_1+M_2}{n^{\frac{1}{3}(1-4\alpha)}}\bigg)$$

and

$$P\bigg(\max_{|\hat{\mathcal{M}}|\le d} \|w\|_2 \le \sigma\sqrt{\frac{\log p}{n^\alpha}}\bigg) \ge 1 - O\bigg(\frac{(M_1+M_2)R^3}{(\log p)^{2\iota} n^{3-4\alpha-2\iota}}\bigg).$$

Therefore, we have

$$P\bigg(\max_{|\hat{\mathcal{M}}|\le d, S^*\subset\hat{\mathcal{M}}} \|\hat{\beta}_d^{(OLS)} - \beta_d\|_\infty \le 2\sigma\sqrt{\frac{\log p}{n^\alpha}}\bigg) = 1 - O\bigg(\frac{\lambda_*^{-2} d\log d}{n^{\frac{1}{3}(1-\alpha)}} + \frac{M_1+M_2}{n^{\frac{1}{3}(1-4\alpha)}} + \frac{(M_1+M_2)R^3}{(\log p)^{2\iota} n^{3-4\alpha-2\iota}}\bigg)$$

□

Proof of Theorem 3. Recall that X_d consists of variables contained in $\hat{\mathcal{M}}_d$, the definition of $\hat{\beta}(r)^{(Ridge)}$ becomes

$$\begin{aligned}
\hat{\beta}(r)^{(Ridge)} &= (X_d^T X_d + rI_d)^{-1} X_d^T X_d \beta + (X_d^T X_d + rI_d)^{-1} X_d^T \varepsilon + (X_d^T X_d + rI_d)^{-1} X_d^T X_{d,c}\beta_{d,c} \\
&= \beta - r(X_d^T X_d + rI_d)^{-1}\beta + (X_d^T X_d + rI_d)^{-1} X_d^T \varepsilon + (X_d^T X_d + rI_d)^{-1} X_d^T X_{d,c}\beta_{d,c} \\
&= \beta - \tilde{\xi}(r) + \tilde{\eta}(r) + \tilde{\omega}(r).
\end{aligned}$$

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

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

As proved in Lemma 6, we know that

$$\max_{|\hat{\mathcal{M}}|\le d} \lambda_{min}\bigg(\frac{X_d^T X_d}{n}\bigg) \ge \frac{\lambda_* k_1}{162}.$$

with probability $1 - O(nk_1^6 M_1)$. Adding r/n to the above matrix will only increase the smallest eigenvalue. Thus, we have

$$\|\tilde{\xi}(r)\|_2 \leq r^2 \beta^T (X_d^T X_d + rI_d)^{-2} \beta \leq \frac{162 r\lambda^* M_0}{n\lambda_* k_1} = \frac{162 r\kappa M_0}{nk_1}.$$

Where we used $M_0 \geq var(Y) \geq \|\beta\|_2^2 \lambda_{max}^{-1}(\Sigma)$. Choosing $k_1 = n^{-\frac{2(1-\alpha)}{9}}$, we have

$$P\bigg(\max_{|\hat{\mathcal{M}}| \leq d} \|\tilde{\xi}(r)\|_2 \leq \frac{162 r\kappa M_0}{n^{\frac{1}{9}(7+2\alpha)}}\bigg) = 1 - O\bigg(\frac{M_1}{n^{\frac{1}{3}(1-4\alpha)}}\bigg),$$

which implies that as long as $r \leq \frac{\sigma n^{(7/9 - 5\alpha/18)}\sqrt{\log p}}{162\kappa M_0}$, we have

$$P\bigg(\max_{|\hat{\mathcal{M}}| \leq d} \|\tilde{\xi}(r)\|_2 \leq \sigma\sqrt{\frac{\log p}{n^\alpha}}\bigg) = 1 - O\bigg(\frac{M_1}{n^{\frac{1}{3}(1-4\alpha)}}\bigg).$$

In addition, the proof for $\|\eta\|_\infty$ and $\|\omega\|_2$ shows that the only key quantity that has changed is $\max_{|\hat{\mathcal{M}}| \leq d} \lambda_{min}\left(\frac{X_d^T X_d}{n}\right)$ which is replaced by $\max_{|\hat{\mathcal{M}}| \leq d} \lambda_{min}\left(\frac{X_d^T X_d + rI_d}{n}\right)$ for $\beta^{(ridge)}$. While the latter is trivially lower bounded by the former, we thus have

$$P\bigg(\max_{|\hat{\mathcal{M}}| \leq d} \|\tilde{\eta}(r)\|_\infty \leq \sigma\sqrt{\frac{\log p}{n^\alpha}}\bigg) = 1 - O\bigg(\frac{\lambda_*^{-2} d\log d}{n^{\frac{1}{3}(1-\alpha)}} + \frac{M_1 + M_2}{n^{\frac{1}{3}(1-4\alpha)}}\bigg)$$

and

$$P\bigg(\max_{|\hat{\mathcal{M}}| \leq d} \|\tilde{w}(r)\|_2 \leq \sigma\sqrt{\frac{\log p}{n^\alpha}}\bigg) \geq 1 - O\bigg(\frac{(M_1 + M_2)R^3}{(\log p)^{2\iota} n^{3-4\alpha-2\iota}}\bigg).$$

Consequently, we have

$$P\bigg(\max_{|\hat{\mathcal{M}}| \leq d, S^* \subset \hat{\mathcal{M}}} \|\hat{\beta}_d^{(ridge)} - \beta_d\|_\infty \leq 3\sigma\sqrt{\frac{\log p}{n^\alpha}}\bigg) = 1 - O\bigg(\frac{\lambda_*^{-2} d\log d}{n^{\frac{1}{3}(1-\alpha)}} + \frac{2M_1 + M_2}{n^{\frac{1}{3}(1-4\alpha)}} + \frac{(M_1 + M_2)R^3}{(\log p)^{2\iota} n^{3-4\alpha-2\iota}}\bigg),$$

as long as

$$r \leq \frac{\sigma n^{(7/9 - 5\alpha/18)}\sqrt{\log p}}{162\kappa M_0}.$$

□

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 $i \in \tilde{\mathcal{M}}_d$,

$$\tilde{\eta}_i = e_i^T (X_{\tilde{\mathcal{M}}_d}^T X_{\tilde{\mathcal{M}}_d})^{-1} X_{\tilde{\mathcal{M}}_d}^T \varepsilon \sim N(0, \sigma^2 e_i^T (X_{\tilde{\mathcal{M}}_d}^T X_{\tilde{\mathcal{M}}_d})^{-1} e_i) \overset{(d)}{=} \sigma\sqrt{e_i^T (X_{\tilde{\mathcal{M}}_d}^T X_{\tilde{\mathcal{M}}_d})^{-1} e_i} N(0,1). \quad (10)$$

Likewise in the proof of Lemma 5, 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

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

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

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

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

$$\max_{i\in\tilde{\mathcal{M}}_d}|\tilde{\eta}_i| \le 8\sigma\sqrt{\frac{2\kappa\log(4d/\delta)}{n}}. \tag{11}$$

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 Vershynin (2010), 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 $t\ge 0$ we have,

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

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

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

which implies that

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

Then we know that

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

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

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

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

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

and for $i\in S$ we have

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

$\square$

References

Akritas, M. G., Lahiri, S., and Politis, D. N. (2014). Topics in nonparametric statistics. In *Proceedings of the First Conference of the International Society for Nonparametric Statistics*. Springer.

Lounici, K. (2008). Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electronic Journal of Statistics*, 2:90–102.

Raskutti, G., Wainwright, M. J., and Yu, B. (2010). Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Wang, X. and Leng, C. (2015). High dimensional ordinary least squares projection for screening variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Wang, X., Leng, C., and Dunson, D. B. (2015). On the consistency theory of high dimensional variable screening. *arXiv preprint arXiv:1502.06895*.