Supplementary of "Model Selection for High Dimensional Quadratic Regression via Regularization"

Supplementary A: Theorem 2

In this supplementary to our paper, we show a generalized version of Theorem 1 without Gaussian assumption. Similar as in our paper, constants C_1, C_2,... and c_1, c_2,... are locally defined and may take different values in different sections. We start with a brief review of definition of a subgaussian random variable and its properties.

A random variable X is called b-subgaussian if for some $b > 0$, $E(e^{tX}) \leq e^{b^2t^2/2}$ for all $t \in \mathbb{R}$. The set of all subgaussian random variables is closed under linear operation by the following proposition.

Proposition 1 *Let X_i be b_i-subgaussian for $i = 1, ..., n$. Then $a_1X_1 + ... + a_nX_n$ is B-subgaussian with $B = \sum_{i=1}^n |a_i|b_i$. Moreover, if X_1,...,X_n are independent, $a_1X_1 + ... + a_nX_n$ is B-subgaussian with $B = (\sum_{i=1}^n a_i^2 b_i^2)^{\frac{1}{2}}$.*

Moreover, the tail probability of a subgaussian variable can be well controlled.

Proposition 2 *If X is b-subgaussian, then $\mathbf{P}(|X| > t) \leq 2e^{-\frac{t^2}{2b^2}}$ for all $t > 0$. Moreover, there exists a positive constant, say $a = 1/6b^2$, such that $Ee^{aX^2} \leq 2$.*

These well-known results can be found, e.g., in Rivasplata (2012).

Condition (SG) $\{\mathbf{x}_i\}_{i=1}^n$ are IID random vectors from an elliptical distribution with marginal b-subgaussian distribution. Moreover, $\{\varepsilon_i\}_{i=1}^n$ are IID with b-subgaussian distribution.

We still use Σ and $\Sigma_{\mathcal{AB}}$ denote the covariance matrix of $\mathbf{x}_i$ and its submatrix corresponding to index sets $\mathcal{A}$ and $\mathcal{B}$. $\mathbf{B} = (B_{jk})$ is the coefficient matrix for interaction effects with $B_{jk} = \beta_{j,k}/2$, $(j \neq k)$ and $B_{jj} = \beta_{j,j}$. $\Lambda_{\min}(\mathbf{A})$ and $\Lambda_{\max}(\mathbf{A})$ denote the smallest and largest eigenvalues of a matrix $\mathbf{A}$. We need the following technical conditions:

(C1) (Irrepresentable Condition) $\|\Sigma_{\mathcal{S}^c\mathcal{S}}(\Sigma_{\mathcal{SS}})^{-1}\|_\infty \leq 1-\gamma$, $\gamma \in (0,1]$.

(C2) (Eigenvalue Condition) $\Lambda_{\min}(\Sigma_{\mathcal{SS}}) \geq C_{\min} > 0$.

(C3) (Dimensionality and Sparsity) $s\log p = o(n)$ and $s(\log s)^{\frac{1}{2}} = o(n^{\frac{1}{3}})$.

(C4) (Coefficient Matrix) $\mathbf{B}$ is sparse and supported in a submatrix $\mathbf{B}_{\mathcal{SS}}$. $\Lambda_{\max}(\mathbf{B}^2) = \Lambda_{\max}(\mathbf{B}^2_{\mathcal{SS}}) \leq C^2_{\mathbf{B}}$ for a positive constant $C_{\mathbf{B}}$.

Condition (C3) is employed to replace (6) in Theorem 1. Similar conditions are standard in the literature. Condition (C4) on $\mathbf{B}$ is used to control the overall interaction effect, which is treated as noise in stage one. $\Lambda_{\max}(\mathbf{B})$ can be bounded, e.g., by $\|\boldsymbol{\beta}_{\mathcal{I}}\|_1$.

Theorem 2 *Suppose that conditions (SG), (C1)-(C4) hold. For* $\lambda_n \gg \tau\,(\log p/n)^{\frac{1}{2}}$, *with probability tending to 1, the LASSO has a unique solution* $\hat{\boldsymbol{\beta}}_L$ *with support contained within* $\mathcal{S}$. *Moreover, if* $\beta_{\min} = \min_{j\in\mathcal{S}}|\beta_j| > 2(s^{-\frac{1}{2}} + \|\boldsymbol{\beta}_{\mathcal{I}}\|_2/s + \lambda_n s^{\frac{1}{2}})/C_{\min}$, *then* $\mathrm{sign}(\hat{\boldsymbol{\beta}}_L) = \mathrm{sign}(\boldsymbol{\beta}_{\mathcal{M}})$.

Note that $\|\boldsymbol{\beta}_{\mathcal{I}}\|_2 = \mathrm{tr}(B^2) \leq sC^2_{\mathbf{B}}$, so $\|\boldsymbol{\beta}_{\mathcal{I}}\|_2/s \leq C_{\mathbf{B}}s^{-\frac{1}{2}}$.

Supplementary B: Proof of Theorem 2

Recall that we use $(W1)$, $(W2)$,... to denote the formula (1), (2),... in Wainwright (2009). The n-vector $\boldsymbol{\omega}$ is the imaginary noise at Stage 1, which is the sum of the subgaussian noise $\boldsymbol{\varepsilon}$ and the interaction effects $(\mathbf{u}_1^\top\boldsymbol{\beta}_{\mathcal{I}}, ..., \mathbf{u}_n^\top\boldsymbol{\beta}_{\mathcal{I}})^\top$.

Part I: Verifying strict dual feasibility.

We show that inequality $|Z_j| < 1$ holds for each $j \in \mathcal{S}^c$, with overwhelming probability, where Z_j is defined in (W10). For every $j \in \mathcal{S}^c$, conditional on $\mathbf{X}_{\mathcal{S}}$, (W37) gives a decomposition $Z_j = A_j + B_j$ where

$$\begin{aligned} A_j &= \mathbf{E}_j^\top \left\{ \mathbf{X}_\mathcal{S}(\mathbf{X}_\mathcal{S}^\top \mathbf{X}_\mathcal{S})^{-1}\check{\mathbf{z}}_\mathcal{S} + \Pi_{\mathbf{X}_\mathcal{S}^\perp}\left(\frac{\boldsymbol{\omega}}{\lambda_n n}\right)\right\} \\ B_j &= \Sigma_{j\mathcal{S}}(\Sigma_{\mathcal{S}\mathcal{S}})^{-1}\check{\mathbf{z}}_\mathcal{S}, \end{aligned}$$

where $\mathbf{E}_j^\top = \mathbf{X}_j^\top - \Sigma_{j\mathcal{S}}(\Sigma_{\mathcal{S}\mathcal{S}})^{-1}\mathbf{X}_\mathcal{S}^\top \in \mathbb{R}^n$ with entries E_{ij} that is $2b$-subgaussian by Proposition 1 and condition (C1).

Condition (C1) implies

$$\max_{j\in\mathcal{S}^c} |B_j| \le 1-\gamma.$$

Conditional on $\mathbf{X}_\mathcal{S}$ and $\boldsymbol{\omega}$, A_j is $2bM_n^{\frac{1}{2}}$-subgaussian, where

$$M_n = \frac{1}{n}\check{\mathbf{z}}_\mathcal{S}^\top \left(\frac{\mathbf{X}_\mathcal{S}^\top \mathbf{X}_\mathcal{S}}{n}\right)^{-1} \check{\mathbf{z}}_\mathcal{S} + \left\|\Pi_{\mathbf{X}_\mathcal{S}^\perp}\left(\frac{\boldsymbol{\omega}}{\lambda_n n}\right)\right\|_2^2.$$

We need the following lemma that is proved in Supplementary C.

Lemma 2 *For any $\epsilon \in (0, \frac{1}{2})$, define the event $\overline{\mathcal{T}}(\epsilon) = \{M_n > \overline{M}_n(\epsilon)\}$, where*

$$\overline{M}_n(\epsilon) = \frac{2s}{C_{\min} n} + \frac{4(\sigma^2+\tau^2)}{\lambda_n^2 n}.$$

Then $\mathbf{P}(\overline{\mathcal{T}}(\epsilon)) \le C_1 s^2 \exp(-C_2 n^{\frac{1}{2}}\epsilon^2)$ for some C_1, $C_2 > 0$.

By Lemma 2,

$$\begin{aligned} \mathbf{P}\left(\max_{j\in\mathcal{S}^c}|Z_j| \ge 1\right) &\le \mathbf{P}\left(\max_{j\in\mathcal{S}^c}|A_j| \ge \gamma\right) \\ &\le \mathbf{P}\left(\max_{j\in\mathcal{S}^c}|A_j| \ge \gamma \mid \overline{\mathcal{T}}^c(\epsilon)\right) + C_1 s^2 \exp(-C_2 n^{\frac{1}{2}}\epsilon^2). \qquad (20) \end{aligned}$$

Conditional on $\overline{\mathcal{T}}^c(\epsilon)$, A_j is $2b\overline{M}_n^{\frac{1}{2}}(\epsilon)$-subgaussian, so by Proposition 2

$$\mathbf{P}\left(\max_{j\in\mathcal{S}^c}|A_j| \ge \gamma \mid \overline{\mathcal{T}}^c(\epsilon)\right) \le 2(p-s)\exp\left(-\frac{\gamma^2}{8b^2\overline{M}_n(\epsilon)}\right),$$

where the right hand side goes to 0 by condition (C3). Therefore, $\max_{j\in\mathcal{S}^c}|Z_j| < 1$ holds with probability tending to 1.

Part II: Sign consistency.

In order to show sign consistency, by Lemma 3 in Wainwright (2009) it is sufficient to show

$$\text{sign}(\beta_j + \Delta_j) = \text{sign}(\beta_j), \quad \text{for all } j \in \mathcal{S}, \tag{21}$$

where

$$\Delta_j = \mathbf{e}_j^\top \left(\frac{\mathbf{X}_\mathcal{S}^\top \mathbf{X}_\mathcal{S}}{n}\right)^{-1} \left[\frac{1}{n}\mathbf{X}_\mathcal{S}^\top \boldsymbol{\omega} - \lambda_n \text{sign}(\boldsymbol{\beta}_\mathcal{S})\right].$$

It is straightforward that

$$\begin{aligned}
\max_{j\in\mathcal{S}} |\Delta_j| &\leq \left\| \left(\frac{\mathbf{X}_\mathcal{S}^\top \mathbf{X}_\mathcal{S}}{n}\right)^{-1} \right\|_2 \left\| \frac{1}{n}\mathbf{X}_\mathcal{S}^\top \boldsymbol{\omega} - \lambda_n \text{sign}(\boldsymbol{\beta}_\mathcal{S}) \right\|_2 \\
&\leq \left\| \left(\frac{\mathbf{X}_\mathcal{S}^\top \mathbf{X}_\mathcal{S}}{n}\right)^{-1} \right\|_2 \left(\left\| \frac{1}{n}\mathbf{X}_\mathcal{S}^\top \boldsymbol{\varepsilon} \right\|_2 + \left\| \frac{1}{n}\mathbf{X}_\mathcal{S}^\top \mathbf{y}_\mathcal{I} \right\|_2 + \|\lambda_n \text{sign}(\boldsymbol{\beta}_\mathcal{S})\|_2 \right).
\end{aligned}$$

By Lemma 3,

$$\left\| \left(\frac{\mathbf{X}_\mathcal{S}^\top \mathbf{X}_\mathcal{S}}{n}\right)^{-1} \right\|_2 < 2/C_{\min},$$

with probability at least $1 - s^2 C_3 \exp(-C_4 n/s^2)$. Moreover,

$$\|\lambda_n \text{sign}(\boldsymbol{\beta}_\mathcal{S})\|_2 \leq \lambda_n s^{\frac{1}{2}}.$$

$$\left\| \frac{1}{n}\mathbf{X}_\mathcal{S}^\top \mathbf{y}_\mathcal{I} \right\|_2 \leq \|\boldsymbol{\beta}_\mathcal{I}\|_2 \max_{j,k,\ell\in\mathcal{S}} \left\{ \left| \frac{1}{n}\mathbf{X}_j^\top (\mathbf{X}_k \star \mathbf{X}_\ell) \right| \right\},$$

where $\frac{1}{n}\mathbf{X}_j^\top(\mathbf{X}_k \star \mathbf{X}_\ell)$ is a sample third moment. By Remark B.2 and Lemma B.5 in Hao & Zhang (2014),

$$\mathbf{P}\left(\left| \frac{1}{n}\mathbf{X}_j^\top (\mathbf{X}_k \star \mathbf{X}_\ell) \right| > \epsilon \right) \leq c_1 \exp(-c_2 n^{\frac{2}{3}} \epsilon^2).$$

Because $|\mathcal{S}| = s$, we have

$$\mathbf{P}\left(\left\|\frac{1}{n}\mathbf{X}_{\mathcal{S}}^{\top}\mathbf{y}_{\mathcal{I}}\right\|_2 \geq \|\boldsymbol{\beta}_{\mathcal{I}}\|_2\epsilon\right) \leq s^3 c_1 \exp(-c_2 n^{\frac{2}{3}}\epsilon^2).$$

which, with $\epsilon = 1/s$ leads to

$$\mathbf{P}\left(\left\|\frac{1}{n}\mathbf{X}_{\mathcal{S}}^{\top}\mathbf{y}_{\mathcal{I}}\right\|_2 \geq \|\boldsymbol{\beta}_{\mathcal{I}}\|_2 s^{-1}\right) \leq s^3 c_1 \exp(-c_2 n^{\frac{2}{3}}/s^2).$$

Similarly,

$$\mathbf{P}\left(\left\|\frac{1}{n}\mathbf{X}_{\mathcal{S}}^{\top}\boldsymbol{\varepsilon}\right\|_2 > s^{\frac{1}{2}}\epsilon\right) < s c_3 \exp(-c_4 n \epsilon^2),$$

which, with $\epsilon = 1/s$ leads to

$$\mathbf{P}\left(\left\|\frac{1}{n}\mathbf{X}_{\mathcal{S}}^{\top}\boldsymbol{\varepsilon}\right\|_2 > s^{-\frac{1}{2}}\right) < s c_3 \exp(-c_4 n/s^2).$$

Overall, with probability greater than $1 - c_5 s^3 \exp(-c_6 n^{\frac{2}{3}}/s^2)$,

$$\max_{j\in\mathcal{S}} |\Delta_j| \leq 2\left(s^{-\frac{1}{2}} + \|\boldsymbol{\beta}_{\mathcal{I}}\|_2 s^{-1} + \lambda_n s^{\frac{1}{2}}\right)/C_{\min} = g(\lambda_n).$$

Therefore (21) holds when $\beta_{\min} > g(\lambda_n)$. $\square$

Supplementary C: Proof of Lemma 2.

The first summand of M_n can be bounded as

$$\frac{1}{n}\check{\mathbf{z}}_{\mathcal{S}}^{\top}\left(\frac{\mathbf{X}_{\mathcal{S}}^{\top}\mathbf{X}_{\mathcal{S}}}{n}\right)^{-1}\check{\mathbf{z}}_{\mathcal{S}} \leq \frac{2s}{nC_{\min}}$$

with probability at least $1 - s^2 C_3 \exp(-C_4 n/s^2)$, where C_3, C_4 are positive constants. It directly follows the fact $\|\check{\mathbf{z}}_{\mathcal{S}}\|_2^2 \leq s$ and Lemma 3 in Supplementary D, which says the largest eigenvalue of $\left(\frac{\mathbf{X}_{\mathcal{S}}^{\top}\mathbf{X}_{\mathcal{S}}}{n}\right)^{-1}$ can be controlled by $2/C_{\min}$.

For the second summand, because $\Pi_{\mathbf{X}_{\mathcal{S}}^{\perp}}$ is an orthogonal projection matrix and $\boldsymbol{\omega} = \boldsymbol{\varepsilon} + \mathbf{y}_{\mathcal{I}}$, we have

$$\left\|\Pi_{\mathbf{X}_{\mathcal{S}}^{\perp}}\left(\frac{\boldsymbol{\omega}}{\lambda_n n}\right)\right\|_2^2 \leq \frac{\|\boldsymbol{\omega}\|_2^2}{\lambda_n^2 n^2} \leq \frac{2}{\lambda_n^2 n}\frac{\|\boldsymbol{\varepsilon}\|_2^2 + \|\mathbf{y}_{\mathcal{I}}\|_2^2}{n}.$$

As $\{\boldsymbol{\varepsilon}_i\}_{i=1}^n$ are IID subgaussian, by Proposition 2, and Lemma B.4 in Hao & Zhang (2014), we have

$$\mathbf{P}\left(\frac{\|\boldsymbol{\varepsilon}\|_2^2}{n} \leq (1+\epsilon)\sigma^2\right) \leq c_1 \exp\left(-c_2 n\epsilon^2\right). \tag{22}$$

On the other hand,

$$\|\mathbf{y}_{\mathcal{I}}\|_2^2 - n\tau^2 = \sum_{i=1}^n (\mathbf{u}_i^\top \boldsymbol{\beta}_{\mathcal{I}})^2 - \tau^2,$$

is a sum of mean zero independent random variables.

Define $W_i = \frac{(\mathbf{u}_i^\top \boldsymbol{\beta}_{\mathcal{I}})^2}{\tau^2} - 1$, then $\mathrm{E}(W_i) = 0$. By condition (C4),

$$\mathbf{u}_i^\top \boldsymbol{\beta}_{\mathcal{I}} = \mathbf{x}_i^\top \mathbf{B}\mathbf{x}_i - \mathrm{E}(\mathbf{x}_i^\top \mathbf{B}\mathbf{x}_i) = (\mathbf{x}_i)_{\mathcal{S}}^\top \mathbf{B}_{\mathcal{S}\mathcal{S}}(\mathbf{x}_i)_{\mathcal{S}} - \mathrm{E}\left((\mathbf{x}_i)_{\mathcal{S}}^\top \mathbf{B}_{\mathcal{S}\mathcal{S}}(\mathbf{x}_i)_{\mathcal{S}}\right).$$

So W_i is a degree 4 polynomial of subgaussian variables dominated by $[C_{\mathbf{B}}(\mathbf{x}_i)_{\mathcal{S}}^\top(\mathbf{x}_i)_{\mathcal{S}}]^2$, which is, up to the constant $C_{\mathbf{B}}^2$, a summation of at most s^2 degree 4 monomials of subgaussian variables. The tail probability of each of these monomials can be bounded as in Lemma B.5 in Hao & Zhang (2014). Therefore, we have

$$\mathbf{P}\left(\left|\sum_{i=1}^n W_i\right| > n\epsilon\right) \leq c_3 s^2 \exp(-c_4 n^{\frac{1}{2}}\epsilon^2),$$

for some positive constants c_3, c_4. That is

$$\mathbf{P}\left(\left|\|\mathbf{y}_{\mathcal{I}}\|_2^2 - n\tau^2\right| \geq \tau^2 n\epsilon\right) \leq c_3 s^2 \exp(-c_4 n^{\frac{1}{2}}\epsilon^2),$$

which implies

$$\mathbf{P}\left(\frac{\|\mathbf{y}_{\mathcal{I}}\|_2^2}{n} \leq (1+\epsilon)\tau^2\right) \leq c_3 s^2 \exp\left(-c_4 n^{\frac{1}{2}}\epsilon^2\right). \tag{23}$$

(22) and (23) imply

$$\mathbf{P}\left(\left\|\Pi_{\mathbf{X}_{\mathcal{S}}^{\perp}}\left(\frac{\boldsymbol{\omega}}{\lambda_n n}\right)\right\|_2^2 \geq (1+\epsilon)\frac{2(\sigma^2+\tau^2)}{\lambda_n^2 n}\right) \leq c_5 s^2 \exp\left(-c_6 n^{\frac{1}{2}}\epsilon^2\right),$$

for some positive constants c_5, c_6. With $\epsilon = 1$, the conclusion of Lemma 2 follows. □

Supplementary D: Lemma 3 and its proof.

Lemma 3 *Under conditions (SG) and (C3), we have*

$$\mathbf{P}\left(\Lambda_{\min}\left(\frac{\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}}{n}\right) > C_{\min}/2\right) > 1 - s^2 C_3 \exp(-C_4 n/s^2) \to 1,$$

where $C_{\min} = \Lambda_{\min}(\Sigma_{\mathcal{SS}})$*,* $C_3 > 0$*,* $C_4 > 0$*.*

Proof. We need bound

$$\mathbf{P}\left(\sup_{\|\mathbf{v}\|_2=1} |\mathbf{v}^\top(\Sigma_{\mathcal{SS}} - \mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}/n)\mathbf{v}| > \epsilon\right). \tag{24}$$

For easy presentation, we assume that the s-vector $\mathbf{v}$ is indexed by $\mathcal{S}$. Then

$$\begin{aligned}
& |\mathbf{v}^\top(\Sigma_{\mathcal{SS}} - \mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}/n)\mathbf{v}| \\
\leq & \sum_{j,k\in\mathcal{S}} |v_j v_k||\Sigma_{jk} - \mathbf{X}_j^\top \mathbf{X}_k/n| \\
\leq & \|\mathbf{v}\|_1^2 \max_{j,k\in\mathcal{S}} |\Sigma_{jk} - \mathbf{X}_j^\top \mathbf{X}_k/n| \\
\leq & s \max_{j,k\in\mathcal{S}} |\Sigma_{jk} - \mathbf{X}_j^\top \mathbf{X}_k/n|
\end{aligned}$$

So (24) is bounded from above by

$$\mathbf{P}\left(\max_{j,k\in\mathcal{S}} |\Sigma_{jk} - \mathbf{X}_j^\top \mathbf{X}_k/n| > \epsilon/s\right) \tag{25}$$

Following Remark B.2 and Lemma B.5 in Hao & Zhang (2014), it is easy to derive

$$\mathbf{P}\left(|\Sigma_{jk} - \mathbf{X}_j^\top \mathbf{X}_k/n| > \epsilon\right) < C_3 \exp(-C_5 n \epsilon^2),$$

for constants $C_3 > 0$, $C_5 > 0$ under subgaussian assumption. Therefore, (25) is further bounded by $s^2 C_3 \exp(-C_5 n\epsilon^2/s^2)$. Take $\epsilon = \min\{C_{\min}/2, 1/2\}$, we have

$$\mathbf{P}\left(\Lambda_{\min}\left(\frac{\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}}{n}\right) > C_{\min}/2\right) > 1 - s^2 C_3 \exp(-C_4 n/s^2) \to 1,$$

by condition (C3), where $C_4 = C_5(\min\{C_{\min}/2, 1/2\})^2$.

References

HAO, N. & ZHANG, H. H. (2014). Interaction screening for ultra-high dimensional data. *Journal of the American Statistical Association* **109**, 1285–1301.

RIVASPLATA, O. (2012). Subgaussian random variables: An expository note .

WAINWRIGHT, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso). *Information Theory, IEEE Transactions on* **55**, 2183–2202.