SUPPLEMENT TO "Penalized Interaction Estimation for Ultrahigh Dimensional Quadratic Regression"

Cheng Wang, Binyan Jiang and Liping Zhu

Shanghai Jiao Tong University, Hong Kong Polytechnic University, Renmin University of China

This supplementary material provides additional simulations and technical proofs of the main results in the main text.

Appendix

S1.1 Simulation results for models (3.14) and (3.16) in Section 3.1

The simulation results for models (3.14) and (3.16) are presented in Table 1 of this Supplement.

S1.2 Ultrahigh dimensional covariates

Our algorithm is very efficient with cheap computation complexity and low computer memory. In this section, we demonstrate the performance of our proposal under ultrahigh dimension settings. In addition to the three

Table 1: The averages (and standard deviations) of the support recovery rate ("rate"), the Frobenius loss ("loss"), the model size ("size") and the exact support recovery rate ("exact") for models (3.14) and (3.16).

p		PIEy	PIEr	RAMPs	RAMPw	all-pairs-LASSO	Oracle
model (3.14) where the strong heredity condition is satisfied							
100	rate	99.00(5.71)	100.00(0.00)	89.67(29.47)	99.67(3.33)	100.00(0.00)	100.00(0.00)
	size	5.21(3.17)	4.22(2.47)	2.90(0.92)	3.20(0.53)	10.80(6.32)	3.00(0.00)
	loss	0.40(0.24)	0.26(0.14)	0.34(0.69)	0.11(0.11)	0.41(0.12)	0.09(0.04)
	exact	0.36(0.48)	0.58(0.50)	0.88(0.33)	0.82(0.39)	0.06(0.24)	1.00(0.00)
200	rate	99.00(5.71)	99.33(4.69)	80.67(39.41)	97.67(11.85)	100.00(0.00)	100.00(0.00)
	size	5.26(3.01)	4.01(2.48)	2.70(1.40)	3.38(1.41)	10.36(7.07)	3.00(0.00)
	loss	0.42(0.24)	0.30(0.19)	0.52(0.89)	0.16(0.33)	0.46(0.12)	0.09(0.04)
	exact	0.36(0.48)	0.60(0.49)	0.80(0.40)	0.78(0.42)	0.04(0.20)	1.00(0.00)
model (3.16) where the heredity condition is violated							
100	rate	99.00(5.71)	99.67(3.33)	22.00(35.52)	51.00(29.00)	100.00(0.00)	100.00(0.00)
	size	4.76(2.80)	3.82(1.86)	1.19(1.78)	4.46(4.98)	7.23(4.26)	3.00(0.00)
	loss	0.34(0.23)	0.19(0.13)	1.92(0.63)	1.39(0.75)	0.41(0.10)	0.09(0.04)
	exact	0.42(0.49)	0.64(0.48)	0.08(0.27)	0.20(0.40)	0.11(0.31)	1.00(0.00)
200	rate	99.67(3.33)	100.00(0.00)	22.33(33.18)	48.00(32.24)	100.00(0.00)	100.00(0.00)
	size	4.40(2.48)	3.62(1.36)	1.21(1.82)	3.02(3.64)	7.35(5.05)	3.00(0.00)
	loss	0.29(0.20)	0.18(0.11)	1.97(0.48)	1.42(0.79)	0.43(0.09)	0.09(0.04)
	exact	0.54(0.50)	0.71(0.46)	0.01(0.10)	0.20(0.40)	0.17(0.38)	1.00(0.00)

criteria considered in Section S1.1, we also compare the computation time to illustrate the computational efficiency of our method. The parameter settings are the same as those in Section 3.1 except that the data dimension p is now set to be 500, 1000 or 2000, and the sample size n is set to be 400 or 800. To save space, we only report the results for model (3.15) where the weak heredity condition holds.

Table 2 summaries the simulations results including the "rate", "loss", "size" and the computation time in seconds (denoted as "time"). All methods are implemented with a PC with a 3.3 GHz Intel Core i7 CPU and 16GB memory. Overall, the patterns of the estimation accuracy are similar to those in Table 1. Our proposals are much more efficient than other methods in terms of "time". It seems that the computation time of our proposals increases linearly in n and p^2, which is consistent with the overall computation complexity $O\{\min(n, p)p^2\}$ of our algorithms. The computation time of RAMP is not very sensitive to the sample size or data dimension since it used the structure information of heredity conditions. For the all-pairs-LASSO, we tried to implement LARS (Efron et al., 2004). It turns out to be very slow. Therefore, we implement instead the "glmnet" (Friedman et al., 2010). In our view, "glmnet" (Friedman et al., 2010) is perhaps the state of art algorithm for LASSO problems and the package was further

accelerated by strong rules (Tibshirani et al., 2012). Table 2 indicates that the computation time seems to be increasing linearly in n and quadratically in p. However, the all-pairs-LASSO uses more memory since the number of covariates is of order $O(p^2)$ and will break down when $p = 2000$ due to out of memory in R. In summary, our proposal are more efficient than the all-pairs-LASSO in both computation complexity and computation memory.

S1.3 Non-normal covariates

In this section we investigate the performance of our proposal when the covariates are non-normal, and the factor model assumptions are violated. Let $\mathbf{x} = \mathbf{\Sigma}^{1/2}\mathbf{z}$, $\mathbf{\Sigma} = (0.5^{|i-j|})_{100\times 100}$ and $\mathbf{z} = (Z_1, \cdots, Z_p)^{\mathrm{T}}$. We draw Z_ks independently from (i) uniform distribution on the interval $[-\sqrt{3}, \sqrt{3}]$ where $\Delta = 1.8$, (ii) Student's t-distribution $t(5)\sqrt{3/5}$ where $\Delta = 9$ and (iii) Laplace distribution $\text{Laplace}(0, 1)/\sqrt{2}$ where $\Delta = 6$. In all scenarios, the Z_k's are symmetric and have unit variance.

Tables 3 and 4 report the support recovery rate ("rate") and the number of interactions that are estimated as nonzero ("size") and the Frobenius loss ("loss") of $\widehat{\mathbf{\Omega}}$ where $n = 400$ and $p = 100$. Both tables indicate that that PIEy and PIEr are very effective when the covariates are non-normal, and

Table 2: Simulation results for model (3.15) in ultrahigh dimensions

p		PIEy	PIEr	RAMPs	RAMPw	all-pairs-LASSO	Oracle
				$n = 400$			
500	rate	100.00(0.00)	100.00(0.00)	38.00(13.42)	99.00(10.00)	100.00(0.00)	100.00(0.00)
	loss	0.14(0.08)	0.11(0.07)	1.94(0.27)	0.09(0.24)	0.31(0.06)	0.06(0.02)
	size	3.56(1.29)	3.15(0.58)	1.27(0.74)	3.24(1.91)	5.11(4.17)	3.00(0.00)
	time	3.90(0.41)	3.75(0.33)	28.71(8.54)	26.37(5.39)	32.90(3.43)	0.02(0.00)
1000	rate	100.00(0.00)	100.00(0.00)	37.00(15.64)	95.33(20.66)	100.00(0.00)	100.00(0.00)
	loss	0.13(0.08)	0.09(0.05)	1.93(0.33)	0.17(0.52)	0.34(0.06)	0.06(0.03)
	size	3.60(1.62)	3.20(0.95)	1.38(1.20)	3.76(3.85)	4.02(2.09)	3.00(0.00)
	time	12.70(0.49)	12.56(0.55)	48.26(8.88)	50.84(10.54)	126.66(0.55)	0.04(0.01)
2000	rate	100.00(0.00)	100.00(0.00)	31.67(11.96)	88.00(32.66)	-	100.00(0.00)
	loss	0.15(0.10)	0.13(0.09)	2.02(0.14)	0.34(0.78)	-	0.06(0.02)
	size	3.83(2.00)	3.29(0.82)	1.19(0.92)	4.67(6.38)	-	3.00(0.00)
	time	58.46(6.15)	59.33(6.21)	34.61(4.62)	92.41(21.77)	-	0.19(0.03)
				$n = 800$			
500	rate	100.00(0.00)	100.00(0.00)	38.67(13.99)	100.00(0.00)	100.00(0.00)	100.00(0.00)
	loss	0.09(0.04)	0.06(0.04)	1.92(0.33)	0.04(0.02)	0.22(0.04)	0.04(0.02)
	size	3.20(0.64)	3.05(0.26)	1.98(3.08)	3.04(0.20)	3.50(1.45)	3.00(0.00)
	time	4.24(0.55)	4.13(0.58)	102.20(15.51)	132.41(15.95)	62.85(0.72)	0.02(0.00)
1000	rate	100.00(0.00)	100.00(0.00)	38.33(11.96)	100.00(0.00)	100.00(0.00)	100.00(0.00)
	loss	0.09(0.05)	0.06(0.03)	1.94(0.21)	0.04(0.02)	0.23(0.03)	0.04(0.01)
	size	3.28(0.98)	3.06(0.37)	1.77(2.24)	3.01(0.10)	3.41(0.78)	3.00(0.00)
	time	25.95(2.65)	25.64(2.42)	116.46(23.30)	131.51(25.39)	261.54(9.76)	0.06(0.01)
2000	rate	100.00(0.00)	100.00(0.00)	37.00(12.44)	100.00(0.00)	-	100.00(0.00)
	loss	0.09(0.06)	0.06(0.04)	1.94(0.31)	0.04(0.02)	-	0.04(0.02)
	size	3.52(1.73)	3.08(0.37)	1.37(1.32)	3.01(0.10)	-	3.00(0.00)
	time	90.72(6.72)	96.52(7.18)	243.33(72.01)	249.13(46.48)	-	0.24(0.02)

– out of memory in R

the performance comparing with other methods are similar to those we have observed under normal assumptions, indicating that our proposal is practically robust to the violation of the theoretical assumptions. This is not very surprising because we use PIE to find the model and the coefficients are estimated by a further lease square estimation on the support of $\hat{\mathbf{\Omega}}$.

S1.4 Example 3 of Hao et al. (2018)

We conduct simulations using Example 3 of Hao et al. (2018). The underlying true model is of the form

$$Y = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{x}^{\mathrm{T}}\mathbf{\Omega}\mathbf{x} + \varepsilon,$$

where $\boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_5 = 3,\ \boldsymbol{\beta}_6 = \cdots = \boldsymbol{\beta}_{10} = 2$ and

$$\mathbf{\Omega}_{1,2} = \mathbf{\Omega}_{1,13} = \mathbf{\Omega}_{2,3} = \mathbf{\Omega}_{2,15} = \mathbf{\Omega}_{3,4} = 2$$

$$\mathbf{\Omega}_{6,10} = \mathbf{\Omega}_{6,18} = \mathbf{\Omega}_{7,9} = \mathbf{\Omega}_{7,18} = \mathbf{\Omega}_{10,19} = 1.$$

All other entries in $\boldsymbol{\beta}$ and $\mathbf{\Omega}$ are identically zero. The covariates $\mathbf{x}$ are generated from $\mathcal{N}\big(\mathbf{0}_{p\times 1}, \mathbf{\Sigma}\big)$, where $\mathbf{\Sigma} = (0.5^{|k-l|})_{p\times p}$, and ε is drawn from $\mathcal{N}(0, 2^2)$. The simulation results based on 100 replicates are reported in Table 5. In this example, the weak heredity assumption is satisfied and the signals of main effects are pretty strong. We further conduct simulations where the main effects are weakened to $\boldsymbol{\beta}/5$ from $\boldsymbol{\beta}$. The simulation re-

Table 3: Simulation results for models (3.14) and (3.15) with non-normal covariates.

		PIEy	PIEr	RAMPs	RAMPw	all-pairs-LASSO	Oracle
model (3.14) where the strong heredity condition is satisfied							
Unif	rate	99.33(4.69)	99.67(3.33)	100.00(0.00)	100.00(0.00)	100.00(0.00)	100.00(0.00)
	size	3.86(1.73)	3.19(0.72)	3.00(0.00)	3.03(0.17)	5.96(3.94)	3.00(0.00)
	loss	0.22(0.18)	0.13(0.12)	0.06(0.02)	0.06(0.03)	0.26(0.05)	0.06(0.02)
t(5)	rate	93.33(17.08)	95.33(14.23)	93.00(25.64)	99.00(5.71)	100.00(0.00)	100.00(0.00)
	size	6.12(3.35)	5.99(5.80)	3.40(2.27)	3.59(1.93)	7.61(4.96)	3.00(0.00)
	loss	0.47(0.55)	0.33(0.50)	0.22(0.62)	0.11(0.27)	0.25(0.06)	0.06(0.03)
Lap	rate	100.00(0.00)	100.00(0.00)	90.67(28.85)	98.67(6.56)	100.00(0.00)	100.00(0.00)
	size	5.87(4.12)	4.90(3.61)	2.93(1.27)	3.75(3.15)	7.10(5.27)	3.00(0.00)
	loss	0.25(0.12)	0.15(0.06)	0.27(0.66)	0.11(0.24)	0.23(0.06)	0.06(0.03)
model (3.15) where the weak heredity condition is satisfied							
Unif	rate	99.67(3.33)	99.67(3.33)	46.67(20.65)	100.00(0.00)	100.00(0.00)	100.00(0.00)
	size	3.19(0.61)	3.14(0.62)	2.17(2.28)	3.01(0.10)	4.30(2.53)	3.00(0.00)
	loss	0.13(0.11)	0.11(0.11)	1.75(0.53)	0.07(0.04)	0.28(0.06)	0.07(0.03)
t(5)	rate	94.33(15.75)	95.00(14.51)	51.33(27.80)	98.00(14.07)	100.00(0.00)	100.00(0.00)
	size	6.17(4.74)	6.14(6.50)	2.99(2.85)	3.39(1.98)	5.20(3.39)	3.00(0.00)
	loss	0.36(0.55)	0.31(0.53)	1.61(0.69)	0.11(0.36)	0.25(0.05)	0.06(0.03)
Lap	rate	100.00(0.00)	100.00(0.00)	48.67(28.59)	94.00(23.87)	100.00(0.00)	100.00(0.00)
	size	5.37(3.72)	5.21(4.17)	2.54(2.72)	3.56(3.07)	4.87(2.41)	3.00(0.00)
	loss	0.17(0.08)	0.13(0.08)	1.63(0.69)	0.20(0.57)	0.25(0.06)	0.05(0.02)

Table 4: Simulation results for models (3.16) and (3.17) with non-normal covariates.

		PIEy	PIEr	RAMPs	RAMPw	all-pairs-LASSO	Oracle
model (3.16) where the heredity conditions is violated							
Unif	rate	99.67(3.33)	99.67(3.33)	14.00(29.66)	46.33(25.47)	100.00(0.00)	100.00(0.00)
	size	3.95(1.83)	3.13(0.44)	1.19(2.91)	4.27(5.99)	5.26(3.89)	3.00(0.00)
	loss	0.22(0.16)	0.12(0.12)	2.05(0.50)	1.46(0.66)	0.28(0.06)	0.06(0.02)
t(5)	rate	94.00(15.98)	94.33(15.02)	37.00(41.81)	68.33(30.84)	100.00(0.00)	100.00(0.00)
	size	5.93(3.25)	5.24(3.01)	2.99(4.30)	5.07(5.39)	6.26(4.24)	3.00(0.00)
	loss	0.40(0.54)	0.33(0.55)	1.65(0.85)	0.98(0.87)	0.25(0.07)	0.06(0.03)
Lap	rate	99.67(3.33)	100.00(0.00)	42.00(41.47)	62.00(31.79)	100.00(0.00)	100.00(0.00)
	size	5.80(4.08)	5.08(4.07)	2.76(3.16)	6.02(6.95)	5.86(3.35)	3.00(0.00)
	loss	0.21(0.17)	0.11(0.06)	1.59(0.85)	1.12(0.87)	0.24(0.06)	0.06(0.03)
model (3.17) is a pure interaction model where the heredity conditions are violated							
Unif	rate	99.67(3.33)	99.67(3.33)	6.67(17.08)	15.67(32.29)	100.00(0.00)	100.00(0.00)
	size	3.08(0.53)	3.06(0.34)	0.48(1.42)	3.28(6.52)	3.82(1.50)	3.00(0.00)
	loss	0.08(0.11)	0.09(0.11)	2.18(0.17)	1.98(0.71)	0.31(0.06)	0.06(0.03)
t(5)	rate	94.33(15.75)	94.67(14.77)	26.00(33.02)	49.67(46.30)	100.00(0.00)	100.00(0.00)
	size	6.00(6.19)	5.97(6.15)	1.81(2.92)	5.66(7.17)	5.10(4.03)	3.00(0.00)
	loss	0.29(0.57)	0.29(0.55)	1.92(0.61)	1.28(1.09)	0.27(0.07)	0.06(0.03)
Lap	rate	100.00(0.00)	100.00(0.00)	27.67(30.72)	52.00(45.52)	100.00(0.00)	100.00(0.00)
	size	5.11(4.34)	5.07(4.43)	2.78(4.39)	5.55(6.93)	4.41(2.79)	3.00(0.00)
	loss	0.07(0.04)	0.08(0.05)	1.95(0.50)	1.25(1.09)	0.26(0.06)	0.06(0.02)

sults are charted in Table 5, from which it can be clearly seen that, except for the oracle estimate, our proposals, particularly the PIEr method, are the most computationally efficient ("time"), with a moderate model size ("size") and comparatively high recovery rate ("rate"). Such advantages are more obvious when the main effects are weakened.

Table 5: Simulation results for Example 3 of Hao et al. (2018) with $n = 400$ and $p = 100$.

	PIEy	PIEr	RAMPs	RAMPw	pairs-LASSO	Oracle
			Example 3 with $\boldsymbol{\beta}$			
rate	29.30(13.20)	78.00(11.63)	45.50(5.75)	79.10(8.18)	100.00(0.00)	100.00(0.00)
size	15.81(6.55)	15.17(5.39)	5.49(1.18)	8.61(0.63)	27.75(13.16)	10.00(0.00)
loss	4.91(0.62)	1.27(0.35)	2.48(0.11)	1.13(0.28)	0.76(0.10)	0.24(0.05)
exact	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00(0.00)
time	0.68(0.06)	0.60(0.06)	15.71(2.24)	21.30(2.45)	1.73(0.11)	0.00(0.00)
			Example 3 with $\boldsymbol{\beta}/5$			
rate	75.50(12.09)	78.70(11.95)	8.60(10.54)	69.00(20.57)	100.00(0.00)	100.00(0.00)
size	15.32(6.05)	15.57(5.32)	3.93(2.54)	9.67(2.57)	27.21(12.86)	10.00(0.00)
loss	1.44(0.39)	1.24(0.39)	3.86(0.45)	1.78(0.86)	0.76(0.10)	0.24(0.05)
exact	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.00(0.00)
time	0.61(0.07)	0.61(0.07)	15.16(2.51)	19.83(1.90)	1.82(0.14)	0.00(0.01)

S1.5 Appendix A: Some Useful Lemmas

We first show that the ADMM algorithm to minimize (2.8) converges linearly.

Lemma 1. Given $\widehat{\mathbf{\Sigma}}$ and $\widehat{\mathbf{\Lambda}}$. Suppose that the ADMM algorithm (2.10)-(2.12) generates a sequence of solutions $\{(\mathbf{B}^k, \mathbf{\Psi}^k, \mathbf{L}^k), k = 1, \ldots\}$. Then $\{\mathbf{B}^k, \mathbf{\Psi}^k\}$ converges linearly to the minimizer of (2.8), and $\|\mathbf{B}^k - \mathbf{\Psi}^k\|_F$ converges linearly to zero.

Proof. The objective function in the minimization problem (2.8) can be decomposed into two components: $f(\mathbf{B}, \mathbf{\Psi}) = f_1(\mathbf{B}) + f_2(\mathbf{\Psi})$, where $f_1(\mathbf{B}) \stackrel{\text{def}}{=} \text{tr}\{(\mathbf{B}\widehat{\mathbf{\Sigma}})^2\} - \text{tr}(\mathbf{B}\widehat{\mathbf{\Lambda}})$ and $f_2(\mathbf{\Psi}) \stackrel{\text{def}}{=} \lambda_n \|\mathbf{\Psi}\|_1$. Rewrite $\text{tr}\{(\mathbf{B}\widehat{\mathbf{\Sigma}})^2\} = \text{vec}(\mathbf{B})^{\text{T}}(\widehat{\mathbf{\Sigma}} \otimes \widehat{\mathbf{\Sigma}})\text{vec}(\mathbf{B})$. Denote $\widehat{\mathbf{\Sigma}} \otimes \widehat{\mathbf{\Sigma}} = \mathbf{U}^{\text{T}}\mathbf{\Lambda}\mathbf{U}$ and $\mathbf{A}_1 = \mathbf{U}^{\text{T}}\mathbf{\Lambda}^{1/2}\mathbf{U}$. Let $g_1(\mathbf{x}) \stackrel{\text{def}}{=} \|\mathbf{x}\|_F^2$ be a function defined on $\mathbb{R}^{p^2} \mapsto \mathbb{R}$, and $h_1(\mathbf{x}) \stackrel{\text{def}}{=} \text{tr}(\widehat{\mathbf{\Lambda}}\mathbf{x})$, $h_2(\mathbf{x}) \stackrel{\text{def}}{=} \lambda_n \|\mathbf{x}\|_1$ be two functions defined on $\mathbb{R}^{p^2} \mapsto \mathbb{R}$. Then $f_1(\mathbf{B}) = g_1\{\mathbf{A}_1\text{vec}(\mathbf{B})\} + h_1\{\text{vec}(\mathbf{B})\}$ and $f_2(\mathbf{\Psi}) = h_2\{\text{vec}(\mathbf{\Psi})\}$. Given $\widehat{\mathbf{\Sigma}}$, $\widehat{\mathbf{\Lambda}}$ and λ_n, the gradient of g_1 is uniformly Lipschitz continuous and h_1 and h_2 are polyhedral. Lemma 1 thus follows immediately from Theorem 3.1 of Hong and Luo (2017). □

Next we present some useful lemmas for the proofs of the main theorems. Without loss of generality, in what follows we assume that $\text{E}(\mathbf{x}) = \mathbf{0}$ and $\text{E}(Y) = 0$.

Lemma 2. Let $W_1, \cdots, W_n$ be independent variables and $\mathrm{E}\{\exp(c_1|W_i|^{\alpha_0})\} < A_0$ for some $0 < \alpha_0 \leq 1, c_1 > 0, , A_0 > 0$. Then for $0 < t \leq 1$, there exist constants $c_2, c_3 > 0$ such that

$$\mathrm{pr}\Big\{\Big|n^{-1}\sum_{i=1}^{n}(W_i - \mathrm{E}W_i)\Big| > t\Big\} \leq c_2 \exp(-c_3 n^{\alpha_0} t^2).$$

Proof of Lemma 2: For $\mathrm{E}W_i = 0$, see Lemma B.4 of Hao and Zhang (2014). Here, we only need to show $\mathrm{E}\{\exp(c_1|W_i - \mathrm{E}W_i|^{\alpha_0})\} < A_1$ for some $A_1 > 0$. By the integral identity of the expectation, we have

$$\begin{aligned}
\mathrm{E}|W_i| &= \int_0^\infty \mathrm{pr}\left\{\exp(c_1|W_i|^{\alpha_0}) > \exp(c_1 t^{\alpha_0})\right\} dt \\
&\leq \int_0^\infty \mathrm{E}\{\exp(c_1|W_i|^{\alpha_0})\}\exp(-c_1 t^{\alpha_0}) dt \leq A_0 \int_0^\infty \exp(-c_1 t^{\alpha_0}) dt \stackrel{\mathrm{def}}{=} c.
\end{aligned}$$

Consequently, $\mathrm{E}\{\exp(c_1|W_i - \mathrm{E}W_i|^{\alpha_0})\} \leq \mathrm{E}\{\exp(2c_1|W_i|^{\alpha_0} + 2c_1|\mathrm{E}W_i|^{\alpha_0})\} \leq A_0 \exp(2c_1 c^{\alpha_0}) \stackrel{\mathrm{def}}{=} A_1$. The proof is completed. □

Lemma 3. Let W_1 and W_2 be two variables such that $\mathrm{E}\{\exp(c_1|W_1|^{\alpha_1})\} \leq A_1$ and $\mathrm{E}\{\exp(c_2|W_2|^{\alpha_2})\} \leq A_2$, where $c_1, c_2, \alpha_1, \alpha_2, A_1, A_2 > 0$. We have

$$\mathrm{E}\big\{\exp\big(\min(c_1, c_2)|W_1 W_2|^{\alpha_1\alpha_2/(\alpha_1+\alpha_2)}\big)\big\} < \max(A_1, A_2).$$

Proof of Lemma 3: By Holder's or Young's inequality,

$$\begin{aligned}
&\mathrm{E}\big\{\exp\big(\min(c_1, c_2)|W_1 W_2|^{\alpha_1\alpha_2/(\alpha_1+\alpha_2)}\big)\big\} \\
\leq\ & \min(c_1, c_2)\mathrm{E}\big[\exp\{|W_1|^{\alpha_1}\alpha_2/(\alpha_1+\alpha_2) + |W_2|^{\alpha_2}\alpha_1/(\alpha_1+\alpha_2)\}\big] \\
\leq\ & A_1\alpha_2/(\alpha_1+\alpha_2) + A_2\alpha_1(\alpha_1+\alpha_2) \leq \max(A_1, A_2).
\end{aligned}$$

The proof is completed. □

Lemma 4. Under condition (A2), we have there exists a constant $C > 0$,

$$\mathrm{pr}\left(\|\bar{\mathbf{x}}\bar{\mathbf{x}}^{\mathrm{T}}\|_\infty \geq C\log(p)/n\right) = O(p^{-1}), \quad \text{and} \tag{S1.2}$$

$$\mathrm{pr}\{\|n^{-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} - \boldsymbol{\Sigma}\|_\infty \geq C\{\log(p)/n\}^{1/2}\} = O(p^{-1}). \tag{S1.3}$$

Proof of Lemma 4: Writing $\mathbf{e}_k$ as the unit-length p-vector with its k-th entry being one, we have $\|\bar{\mathbf{x}}\bar{\mathbf{x}}^{\mathrm{T}}\|_\infty = \max_{k,l}|\mathbf{e}_k^{\mathrm{T}}\bar{\mathbf{x}}\bar{\mathbf{x}}^{\mathrm{T}}\mathbf{e}_l|$. Note that $\mathbf{e}_k^{\mathrm{T}}\mathbf{x}_1, \cdots, \mathbf{e}_k^{\mathrm{T}}\mathbf{x}_n$ are independent centered sub-Gaussian variables. By Hoeffding's inequality (Vershynin, 2017, Theorem 2.6.3), $\mathrm{pr}(|\mathbf{e}_k^{\mathrm{T}}\bar{\mathbf{x}}| \geq t) \leq 2\exp(-cnt^2)$, for any $t \geq 0$, and then $\mathrm{pr}(|\mathbf{e}_k^{\mathrm{T}}\bar{\mathbf{x}}\bar{\mathbf{x}}^{\mathrm{T}}\mathbf{e}_l| \geq t) \leq \mathrm{pr}(|\mathbf{e}_k^{\mathrm{T}}\bar{\mathbf{x}}| \geq \sqrt{t}) + \mathrm{pr}(|\mathbf{e}_l^{\mathrm{T}}\bar{\mathbf{x}}| \geq \sqrt{t}) \leq 4\exp(-cnt)$. Therefore,

$$\mathrm{pr}\left\{\|\bar{\mathbf{x}}\bar{\mathbf{x}}^{\mathrm{T}}\|_\infty \geq t\right\} \leq \sum_{k,l}\mathrm{pr}(|\mathbf{e}_k^{\mathrm{T}}\bar{\mathbf{x}}\bar{\mathbf{x}}^{\mathrm{T}}\mathbf{e}_l| \geq t) \leq 4p^2 \cdot \exp(-cnt).$$

Set $t = c^{-1}C\log(p)/n$ for large enough C, which yields the conclusion (S1.2).

Similarly, $\mathbf{e}_k^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{e}_l - \mathbf{e}_k^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{e}_l,\ i = 1, \cdots, n$ are independent centered sub-exponential variables. By Bernstein's inequality (Vershynin, 2017, Theorem 2.8.2), we get

$$\mathrm{pr}\left\{\left|\mathbf{e}_k^{\mathrm{T}}\Big(n^{-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} - \boldsymbol{\Sigma}\Big)\mathbf{e}_l\right| \geq t\right\} \leq 2\exp\left\{-n\min(c_1t^2, c_2t)\right\}, \quad \text{and}$$

$$\mathrm{pr}\left(\left\|n^{-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} - \Sigma\right\|_\infty \geq t\right) \leq 2p^2 \cdot \exp\left\{-n\min(c_1t^2, c_2t)\right\}.$$

Choose $t = C\{\log(p)/n\}^{1/2}$ with a sufficiently large C to complete proof of (S1.3). □

Lemma 5. Under conditions (A2) and (A3), there exists a constant $C > 0$ such that,

$$\text{pr}\left[\left\|n^{-1}\sum_{i=1}^{n} Y_i\mathbf{x}_i - \text{E}Y\mathbf{x}\right\|_\infty \geq C\{n^{-\alpha/(\alpha+1)}\log(p)\}^{1/2}\right] = O(p^{-1}), \text{ and} \tag{S1.4}$$

$$\text{pr}\left[\left\|n^{-1}\sum_{i=1}^{n} Y_i\mathbf{x}_i\mathbf{x}_i^{\text{T}} - \text{E}Y\mathbf{x}\mathbf{x}^{\text{T}}\right\|_\infty \geq C\{n^{-\alpha/(\alpha+1)}\log(p)\}^{1/2}\right] = O(p^{-1}). \tag{S1.5}$$

Proof of Lemma 5: We prove (S1.6) only in what follows and (S1.5) can be proved using similar arguments. For $\mathbf{e}_k, \mathbf{e}_l$,

$$\mathbf{e}_k^{\text{T}}\Big(n^{-1}\sum_{i=1}^{n} Y_i\mathbf{x}_i\mathbf{x}_i^{\text{T}} - \text{E}Y\mathbf{x}\mathbf{x}^{\text{T}}\Big)\mathbf{e}_l = n^{-1}\sum_{i=1}^{n} Y_i(\mathbf{e}_k^{\text{T}}\mathbf{x}_i)(\mathbf{e}_l^{\text{T}}\mathbf{x}_i) - \mathbf{e}_k^{\text{T}}(\text{E}Y\mathbf{x}\mathbf{x}^{\text{T}})\mathbf{e}_l.$$

By condition (A2), there exist constants c_0 and C_0 such that

$$\text{E}\{\exp(c_0|\mathbf{e}_k^{\text{T}}\mathbf{x}_i\mathbf{e}_l^{\text{T}}\mathbf{x}_i|)\} \leq \text{E}\{\exp(c_0|\mathbf{e}_k^{\text{T}}\mathbf{x}_i|^2)\} + \text{E}\{\exp(c_0|\mathbf{e}_l^{\text{T}}\mathbf{x}_i|^2)\} \leq 2C_0.$$

By condition (A3) and Lemma 3, we have there exist constants c_2, C_2 such that $\text{E}\Big\{\exp\Big(c_2|Y_i(\mathbf{e}_k^{\text{T}}\mathbf{x}_i)(\mathbf{e}_l^{\text{T}}\mathbf{x}_i)|^{\alpha/(\alpha+1)}\Big)\Big\} \leq C_2$. By Lemma 2, we have

$$\text{pr}\Big\{\Big|\mathbf{e}^{\text{T}}(n^{-1}\sum_{i=1}^{n} Y_i\mathbf{x}_i\mathbf{x}_i^{\text{T}} - \text{E}Y\mathbf{x}\mathbf{x}^{\text{T}})^{\text{T}}\tilde{\mathbf{e}}\Big| \geq t\Big\} \leq c_2\exp(-c_3n^{\alpha/(\alpha+1)}t^2).$$

Using the similar arguments as in the proofs of Lemma 4, we can show

$$\mathrm{pr}\Big[\Big\|n^{-1}\sum_{i=1}^{n} Y_i\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} - \mathrm{E}Y\mathbf{x}\mathbf{x}^{\mathrm{T}}\Big\|_\infty \geq C\{n^{-\alpha/(\alpha+1)}\log(p)\}^{1/2}\Big] = O(p^{-1}).$$

The proof is completed. □

S1.6 Appendix B: The ℓ_1-Penalized Estimation

Let $\mathbf{A} \in \mathbb{R}^{q\times q}$, $\mathbf{a} \in \mathbb{R}^q$ be unknown parameters and $\mathbf{A}$ is a positive definite symmetric matrix. To estimate $\mathbf{b}^* \stackrel{\mathrm{def}}{=} \mathbf{A}^{-1}\mathbf{a}$, we consider the ℓ_1-penalized approach:

$$\widehat{\mathbf{b}} = \arg\min_{\mathbf{b}\in\mathbb{R}^q} \mathbf{b}^{\mathrm{T}}\widehat{\mathbf{A}}\mathbf{b}/2 - \widehat{\mathbf{a}}^{\mathrm{T}}\mathbf{b} + \lambda\|\mathbf{b}\|_1, \tag{S1.6}$$

where λ is the tuning parameter and $\hat{\mathbf{A}}$ and $\hat{\mathbf{a}}$ are the empirical estimators of $\mathbf{A}$ and $\mathbf{a}$, respectively. In the sequel, we establish theoretical results for solving (S1.6). These general results will then be used to prove the main theorems in our paper.

Lemma 6. Denote $\Delta = \|\widehat{\mathbf{a}}-\mathbf{a}\|_\infty + \|(\widehat{\mathbf{A}}-\mathbf{A})\mathbf{b}^*\|_\infty$ and let $\mathcal{S} = \{i : \mathbf{b}_i^* \neq 0\}$ be the support of $\mathbf{b}^*$. Assume that $\|\mathbf{A}_{\mathcal{S}^c,\mathcal{S}}\mathbf{A}_{\mathcal{S},\mathcal{S}}^{-1}\|_L + 2\|\mathbf{b}^*\|_0\|\mathbf{A}_{\mathcal{S},\mathcal{S}}^{-1}\|_L\|\widehat{\mathbf{A}} - \mathbf{A}\|_\infty < 1$, and $\lambda > 2(1 - \|\mathbf{A}_{\mathcal{S}^c,\mathcal{S}}\mathbf{A}_{\mathcal{S},\mathcal{S}}^{-1}\|_L - 2\|\mathbf{b}^*\|_0\|\mathbf{A}_{\mathcal{S},\mathcal{S}}^{-1}\|_L\|\widehat{\mathbf{A}} - \mathbf{A}\|_\infty)^{-1}\Delta$, we have

(i) $\widehat{\mathbf{b}}_{\mathcal{S}^c} = \mathbf{0}$;

(ii) $\|\widehat{\mathbf{b}} - \mathbf{b}^*\|_\infty \leq 2\lambda(1 - \|\mathbf{b}^*\|_0\|\mathbf{A}_{\mathcal{S},\mathcal{S}}^{-1}\|_L\|\widehat{\mathbf{A}} - \mathbf{A}\|_\infty)^{-1}\|\mathbf{A}_{\mathcal{S},\mathcal{S}}^{-1}\|_L$.

Proof of Lemma 6: Given the true support $\mathcal{S}$, we consider the estimation

$$\begin{aligned}\widehat{\mathbf{b}}^0 &= \underset{\mathbf{b}\in\mathbb{R}^q,\ \mathbf{b}_{\mathcal{S}^c}=0}{\arg\min}\ \mathbf{b}^{\mathrm{T}}\widehat{\mathbf{A}}\mathbf{b}/2 - \widehat{\mathbf{a}}^{\mathrm{T}}\mathbf{b} + \lambda\|\mathbf{b}\|_1 \\ &= \underset{\mathbf{b}\in\mathbb{R}^q,\ \mathbf{b}_{\mathcal{S}^c}=0}{\arg\min}\ \mathbf{b}_{\mathcal{S}}^{\mathrm{T}}\widehat{\mathbf{A}}_{\mathcal{S},\mathcal{S}}\mathbf{b}_{\mathcal{S}}/2 - \widehat{\mathbf{a}}_{\mathcal{S}}^{\mathrm{T}}\mathbf{b}_{\mathcal{S}} + \lambda\|\mathbf{b}_{\mathcal{S}}\|_1.\end{aligned}$$

By the Karush-Kuhn-Tucker (KKT) condition, we have

$$\widehat{\mathbf{A}}_{\mathcal{S},\mathcal{S}}\widehat{\mathbf{b}}_{\mathcal{S}}^0 - \widehat{\mathbf{a}}_{\mathcal{S}} = -\lambda\mathbf{Z}, \tag{S1.7}$$

where $\mathbf{Z}$ is the sub-gradient of $\|\mathbf{b}_{\mathcal{S}}\|_1$. By the definition of $\mathbf{b}^* = \mathbf{A}^{-1}\mathbf{a}$, we have

$$\begin{pmatrix}\mathbf{a}_{\mathcal{S}} \\ \mathbf{a}_{\mathcal{S}^c}\end{pmatrix} = \begin{pmatrix}\mathbf{A}_{\mathcal{S},\mathcal{S}} & \mathbf{A}_{\mathcal{S},\mathcal{S}^c} \\ \mathbf{A}_{\mathcal{S}^c,\mathcal{S}} & \mathbf{A}_{\mathcal{S}^c,\mathcal{S}^c}\end{pmatrix}\begin{pmatrix}\mathbf{b}_{\mathcal{S}}^* \\ \mathbf{0}\end{pmatrix} = \begin{pmatrix}\mathbf{A}_{\mathcal{S},\mathcal{S}}\mathbf{b}_{\mathcal{S}}^* \\ \mathbf{A}_{\mathcal{S}^c,\mathcal{S}}\mathbf{b}_{\mathcal{S}}^*\end{pmatrix},$$

and hence we have $\widehat{\mathbf{A}}_{\mathcal{S},\mathcal{S}}\widehat{\mathbf{b}}_{\mathcal{S}}^0 - \mathbf{A}_{\mathcal{S},\mathcal{S}}\mathbf{b}_{\mathcal{S}}^* + \mathbf{a}_{\mathcal{S}} - \widehat{\mathbf{a}}_{\mathcal{S}} = -\lambda\mathbf{Z}$. Consequently, we obtain,

$$\widehat{\mathbf{b}}_{\mathcal{S}}^0 - \mathbf{b}_{\mathcal{S}}^* = -\mathbf{A}_{\mathcal{S},\mathcal{S}}^{-1}\Big\{\lambda\mathbf{Z} + (\widehat{\mathbf{A}}_{\mathcal{S},\mathcal{S}} - \mathbf{A}_{\mathcal{S},\mathcal{S}})\widehat{\mathbf{b}}_{\mathcal{S}}^0 + (\mathbf{a}_{\mathcal{S}} - \widehat{\mathbf{a}}_{\mathcal{S}})\Big\}. \tag{S1.8}$$

Using the triangle inequality, we can show that,

$$\begin{aligned}&\|\widehat{\mathbf{b}}_{\mathcal{S}}^0 - \mathbf{b}_{\mathcal{S}}^*\|_\infty \\ \leq\ &\|\mathbf{A}_{\mathcal{S},\mathcal{S}}^{-1}\|_L\Big\{\lambda\|\mathbf{Z}\|_\infty + \|(\widehat{\mathbf{A}}_{\mathcal{S},\mathcal{S}} - \mathbf{A}_{\mathcal{S},\mathcal{S}})(\widehat{\mathbf{b}}_{\mathcal{S}}^0 - \mathbf{b}_{\mathcal{S}}^*)\|_\infty \\ &+\|(\widehat{\mathbf{A}}_{\mathcal{S},\mathcal{S}} - \mathbf{A}_{\mathcal{S},\mathcal{S}})\mathbf{b}_{\mathcal{S}}^* + \mathbf{a}_{\mathcal{S}} - \widehat{\mathbf{a}}_{\mathcal{S}}\|_\infty\Big\} \\ \leq\ &\|\mathbf{A}_{\mathcal{S},\mathcal{S}}^{-1}\|_L\Big\{\lambda + \|\mathbf{b}^*\|_0\|\widehat{\mathbf{A}} - \mathbf{A}\|_\infty\|\widehat{\mathbf{b}}_{\mathcal{S}}^0 - \mathbf{b}_{\mathcal{S}}^*\|_\infty + \|(\widehat{\mathbf{A}} - \mathbf{A})\mathbf{b}^* + \mathbf{a} - \widehat{\mathbf{a}}\|_\infty\Big\},\end{aligned}$$

which implies that

$$\begin{aligned} & \|\widehat{\mathbf{b}}^0_{\mathcal{S}} - \mathbf{b}^*_{\mathcal{S}}\|_\infty \qquad\qquad (S1.9) \\ \leq\ & (1 - \|\mathbf{b}^*\|_0 \|\mathbf{A}^{-1}_{\mathcal{S},\mathcal{S}}\|_L \|\widehat{\mathbf{A}} - \mathbf{A}\|_\infty)^{-1} \|\mathbf{A}^{-1}_{\mathcal{S},\mathcal{S}}\|_L (\lambda + \Delta). \end{aligned}$$

Next, we show that $\widehat{\mathbf{b}}^0$ is exactly the minimizer to $\min_{\mathbf{b}\in\mathbb{R}^q} \mathbf{b}^{\mathrm{T}}\widehat{\mathbf{A}}\mathbf{b}/2 - \widehat{\mathbf{a}}^{\mathrm{T}}\mathbf{b} + \lambda\|\mathbf{b}\|_1$. By the KKT condition, it is sufficient to prove

$$\|(\widehat{\mathbf{A}}\widehat{\mathbf{b}}^0 - \widehat{\mathbf{a}})_{\mathcal{S}}\|_\infty \leq \lambda, \text{ and} \qquad (S1.10)$$

$$\|(\widehat{\mathbf{A}}\widehat{\mathbf{b}}^0 - \widehat{\mathbf{a}})_{\mathcal{S}^c}\|_\infty < \lambda. \qquad (S1.11)$$

Since $(\widehat{\mathbf{A}}\widehat{\mathbf{b}}^0 - \widehat{\mathbf{a}})_{\mathcal{S}} = \widehat{\mathbf{A}}_{\mathcal{S},\mathcal{S}}\widehat{\mathbf{b}}^0_{\mathcal{S}} - \widehat{\mathbf{a}}_{\mathcal{S}}$, (S1.10) is true by (S1.7). For (S1.11), we have

$$\begin{aligned} (\widehat{\mathbf{A}}\widehat{\mathbf{b}}^0 - \widehat{\mathbf{a}})_{\mathcal{S}^c} &= \widehat{\mathbf{A}}_{\mathcal{S}^c,\mathcal{S}}\widehat{\mathbf{b}}^0_{\mathcal{S}} - \widehat{\mathbf{a}}_{\mathcal{S}^c} = \widehat{\mathbf{A}}_{\mathcal{S}^c,\mathcal{S}}\widehat{\mathbf{b}}^0_{\mathcal{S}} - \mathbf{A}_{\mathcal{S}^c,\mathcal{S}}\mathbf{b}^*_{\mathcal{S}} + \mathbf{a}_{\mathcal{S}^c} - \widehat{\mathbf{a}}_{\mathcal{S}^c} \\ &= \widehat{\mathbf{A}}_{\mathcal{S}^c,\mathcal{S}}(\widehat{\mathbf{b}}^0_{\mathcal{S}} - \mathbf{b}^*_{\mathcal{S}}) + (\widehat{\mathbf{A}}_{\mathcal{S}^c,\mathcal{S}} - \mathbf{A}_{\mathcal{S}^c,\mathcal{S}})\mathbf{b}^*_{\mathcal{S}} + \mathbf{a}_{\mathcal{S}^c} - \widehat{\mathbf{a}}_{\mathcal{S}^c} \\ &= (\widehat{\mathbf{A}}_{\mathcal{S}^c,\mathcal{S}} - \mathbf{A}_{\mathcal{S}^c,\mathcal{S}})(\widehat{\mathbf{b}}^0_{\mathcal{S}} - \mathbf{b}^*_{\mathcal{S}}) + \mathbf{A}_{\mathcal{S}^c,\mathcal{S}}\mathbf{A}^{-1}_{\mathcal{S},\mathcal{S}}\{\mathbf{A}_{\mathcal{S},\mathcal{S}}(\widehat{\mathbf{b}}^0_{\mathcal{S}} - \mathbf{b}^*_{\mathcal{S}})\} \\ &\quad + \{(\widehat{\mathbf{A}} - \mathbf{A})\mathbf{b}^* + \mathbf{a} - \widehat{\mathbf{a}}\}_{\mathcal{S}^c}. \end{aligned}$$

Thus, it follows from (S1.8) and (S1.9) that $\|(\widehat{\mathbf{A}}\widehat{\mathbf{b}}^0 - \widehat{\mathbf{a}})_{\mathcal{S}^c}\|_\infty$ is less than or

equal to

$$
\begin{aligned}
& \|\mathbf{b}^*\|_0\|\widehat{\mathbf{A}}-\mathbf{A}\|_\infty\|\widehat{\mathbf{b}}^0_{\mathcal{S}}-\mathbf{b}^*_{\mathcal{S}}\|_\infty \\
& +\|\mathbf{A}_{\mathcal{S}^c,\mathcal{S}}\mathbf{A}^{-1}_{\mathcal{S},\mathcal{S}}\|_L(\lambda+\Delta+\|\mathbf{b}^*\|_0\|\widehat{\mathbf{A}}-\mathbf{A}\|_\infty\|\widehat{\mathbf{b}}^0_{\mathcal{S}}-\mathbf{b}^*_{\mathcal{S}}\|_\infty)+\Delta \\
\leq\ & \frac{(1+|\mathbf{A}_{\mathcal{S}^c,\mathcal{S}}\mathbf{A}^{-1}_{\mathcal{S},\mathcal{S}}\|_L)(\lambda+\Delta)}{1-\|\mathbf{b}^*\|_0\|\mathbf{A}^{-1}_{\mathcal{S},\mathcal{S}}\|_L\|\widehat{\mathbf{A}}-\mathbf{A}\|_\infty}-\lambda \\
=\ & \lambda+\left\{\Delta-\frac{1-|\mathbf{A}_{\mathcal{S}^c,\mathcal{S}}\mathbf{A}^{-1}_{\mathcal{S},\mathcal{S}}\|_L-2\|\mathbf{b}^*\|_0\|\mathbf{A}^{-1}_{\mathcal{S},\mathcal{S}}\|_L\|\widehat{\mathbf{A}}-\mathbf{A}\|_\infty}{1+|\mathbf{A}_{\mathcal{S}^c,\mathcal{S}}\mathbf{A}^{-1}_{\mathcal{S},\mathcal{S}}\|_L}\lambda\right\} \\
& \left\{\frac{1+\|\mathbf{A}_{\mathcal{S}^c,\mathcal{S}}\mathbf{A}^{-1}_{\mathcal{S},\mathcal{S}}\|_L}{1-\|\mathbf{b}^*\|_0\|\mathbf{A}^{-1}_{\mathcal{S},\mathcal{S}}\|_L\|\widehat{\mathbf{A}}-\mathbf{A}\|_\infty}\right\}.
\end{aligned}
$$

When $\lambda > 2(1-\|\mathbf{A}_{\mathcal{S}^c,\mathcal{S}}\mathbf{A}^{-1}_{\mathcal{S},\mathcal{S}}\|_L-2\|\mathbf{b}^*\|_0\|\mathbf{A}^{-1}_{\mathcal{S},\mathcal{S}}\|_L\|\widehat{\mathbf{A}}-\mathbf{A}\|_\infty)^{-1}\Delta$, we have $\|(\widehat{\mathbf{A}}\widehat{\mathbf{b}}^0-\widehat{\mathbf{a}})_{\mathcal{S}^c}\|_\infty < \lambda$. Consequently, $\widehat{\mathbf{b}}=\widehat{\mathbf{b}}^0$ and (S1.11) is an immediate result of (S1.9) by noting $\Delta \leq \lambda$. The proof is completed. □

S1.7 Appendix C: Proof of Proposition 1

Recall that $\mathrm{E}(Y \mid \mathbf{x}) = \alpha+(\mathbf{x}-\mathbf{u})^{\mathrm{T}}\boldsymbol{\beta}+(\mathbf{x}-\mathbf{u})^{\mathrm{T}}\boldsymbol{\Omega}(\mathbf{x}-\mathbf{u})$. Direct calculations show

$$
\begin{aligned}
\mathrm{cov}(\mathbf{x},Y) &= E\Big[\{Y-E(Y)\}(\mathbf{x}-\mathbf{u})\Big] \\
&= E\Big[\{(\mathbf{x}-\mathbf{u})^{\mathrm{T}}\boldsymbol{\beta}+(\mathbf{x}-\mathbf{u})^{\mathrm{T}}\boldsymbol{\Omega}(\mathbf{x}-\mathbf{u})-\mathrm{tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma})\}(\mathbf{x}-\mathbf{u})\Big] \\
&= E\{(\mathbf{x}-\mathbf{u})(\mathbf{x}-\mathbf{u})^{\mathrm{T}}\boldsymbol{\beta}+(\mathbf{z}^{\mathrm{T}}\boldsymbol{\Gamma}_0^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\Gamma}_0\mathbf{z})\boldsymbol{\Gamma}_0\mathbf{z}\} = \boldsymbol{\Sigma}\boldsymbol{\beta}.
\end{aligned}
$$

The proof of the first part is completed. Next we prove the second part.

$$\begin{aligned}
\boldsymbol{\Lambda}_y &= E\Big[\left\{(\mathbf{x}-\mathbf{u})^{\mathrm{T}}\boldsymbol{\beta}+(\mathbf{x}-\mathbf{u})^{\mathrm{T}}\boldsymbol{\Omega}(\mathbf{x}-\mathbf{u})-\mathrm{tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma})\right\}(\mathbf{x}-\mathbf{u})(\mathbf{x}-\mathbf{u})^{\mathrm{T}}\Big] \\
&= E(\mathbf{x}-\mathbf{u})(\mathbf{x}-\mathbf{u})^{\mathrm{T}}\boldsymbol{\Omega}(\mathbf{x}-\mathbf{u})(\mathbf{x}-\mathbf{u})^{\mathrm{T}}-\mathrm{tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma})\boldsymbol{\Sigma} \\
&= E\left\{\boldsymbol{\Gamma}_0\mathbf{z}\mathbf{z}^{\mathrm{T}}(\boldsymbol{\Gamma}_0^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\Gamma}_0)\mathbf{z}\mathbf{z}^{\mathrm{T}}\boldsymbol{\Gamma}_0^{\mathrm{T}}\right\}-\mathrm{tr}(\boldsymbol{\Gamma}_0^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\Gamma}_0)\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^{\mathrm{T}} \\
&= \boldsymbol{\Gamma}_0\Big[E\left\{\mathbf{z}\mathbf{z}^{\mathrm{T}}(\boldsymbol{\Gamma}_0^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\Gamma}_0)\mathbf{z}\mathbf{z}^{\mathrm{T}}\right\}-\mathrm{tr}(\boldsymbol{\Gamma}_0^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\Gamma}_0)\mathbf{I}_p\Big]\boldsymbol{\Gamma}_0^{\mathrm{T}} \\
&= \boldsymbol{\Gamma}_0\{2\boldsymbol{\Gamma}_0^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\Gamma}_0-(\Delta-3)\mathrm{diag}(\boldsymbol{\Gamma}_0^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\Gamma}_0)\}\boldsymbol{\Gamma}_0^{\mathrm{T}} \\
&= 2\boldsymbol{\Sigma}\boldsymbol{\Omega}\boldsymbol{\Sigma}-(\Delta-3)\boldsymbol{\Gamma}_0\mathrm{diag}(\boldsymbol{\Gamma}_0^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\Gamma}_0)\boldsymbol{\Gamma}_0^{\mathrm{T}}. \qquad \text{(S1.12)}
\end{aligned}$$

Thus, $\boldsymbol{\Omega}=\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}/2$ when $\Delta=3$ or $\mathrm{diag}(\boldsymbol{\Gamma}_0^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\Gamma}_0)=0$. The proof is completed. □

S1.8 Appendix D: Proof of Theorem 1

We provide proofs for (i) and (iii) in what follows because (ii) is an immediate result of (i) and (iii) and (iv) can be obtained analog to (iii). For the target parameter matrix $2\boldsymbol{\Omega}=\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}$, we consider its vectorization

$$2\mathrm{vec}(\boldsymbol{\Omega})=\mathrm{vec}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1})=(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})\mathrm{vec}(\boldsymbol{\Lambda})=\boldsymbol{\Gamma}^{-1}\mathrm{vec}(\boldsymbol{\Lambda}), \qquad \text{(S1.13)}$$

where $\boldsymbol{\Gamma}=\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma}$ is a positive and symmetric matrix. For the estimation,

$$\widehat{\boldsymbol{\Omega}}_y=\underset{\mathbf{B}\in\mathbb{R}^{p\times p}}{\arg\min}\,\mathrm{tr}\{(\mathbf{B}\widehat{\boldsymbol{\Sigma}})^2\}-\mathrm{tr}(\mathbf{B}\widehat{\boldsymbol{\Lambda}}_y)+\lambda_{1n}\|\mathbf{B}\|_1.$$

Equivalently, we have

$$\mathrm{vec}(\widehat{\boldsymbol{\Omega}}_y) = \underset{\mathbf{B}\in\mathbb{R}^{p\times p}}{\arg\min}\, \mathrm{vec}(\mathbf{B})^{\mathrm{T}}\widehat{\boldsymbol{\Gamma}}\mathrm{vec}(\mathbf{B}) - \mathrm{vec}(\widehat{\boldsymbol{\Lambda}}_y)^{\mathrm{T}}\mathrm{vec}(\mathbf{B}) + \lambda_{1n}\|\mathrm{vec}(\mathbf{B})\|_1,$$

where $\widehat{\boldsymbol{\Gamma}} \overset{\text{def}}{=} \widehat{\boldsymbol{\Sigma}}\otimes\widehat{\boldsymbol{\Sigma}}$. Therefore, we can use Lemma 6 to derive the theoretical properties by letting $\mathbf{A} = 2\boldsymbol{\Gamma},\ \mathbf{a} = \mathrm{vec}(\boldsymbol{\Lambda}_y),\ \widehat{\mathbf{A}} = 2\widehat{\boldsymbol{\Gamma}}$ and $\widehat{\mathbf{a}} = \mathrm{vec}(\widehat{\boldsymbol{\Lambda}}_y)$.

Recall the definition of $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{\Lambda}}_y$.

$$\begin{aligned}
\widehat{\boldsymbol{\Sigma}} &= n^{-1}\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^{\mathrm{T}} = n^{-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} - \bar{\mathbf{x}}\bar{\mathbf{x}}^{\mathrm{T}}, \quad \text{and}\\
\widehat{\boldsymbol{\Lambda}}_y &= n^{-1}\sum_{i=1}^{n}Y_i\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} - n^{-1}\sum_{i=1}^{n}Y_i(\bar{\mathbf{x}}\mathbf{x}_i^{\mathrm{T}} + \mathbf{x}_i\bar{\mathbf{x}}^{\mathrm{T}}) - n^{-1}\bar{Y}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} + 2\bar{Y}\bar{\mathbf{x}}\bar{\mathbf{x}}^{\mathrm{T}}.
\end{aligned}$$

Lemmas 4 and 5 ensure that there exists a constant $C > 0$ such that with probability greater than $1 - O(p^{-1})$, $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_\infty \le C(n^{-1}\log p)^{1/2}$ and $\|\widehat{\boldsymbol{\Lambda}}_y - \boldsymbol{\Lambda}_y\|_\infty \le C\{n^{-\alpha/(\alpha+1)}\log(p)\}^{1/2}$. Note that $\|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}\|_\infty = \|\widehat{\boldsymbol{\Sigma}}\otimes(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}) + (\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\otimes\boldsymbol{\Sigma}\|_\infty \le (\|\widehat{\boldsymbol{\Sigma}}\|_\infty + \|\boldsymbol{\Sigma}\|_\infty)\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_\infty$, with probability greater than $1 - O(p^{-1})$, we have, $\|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}\|_\infty \le C_1(n^{-1}\log p)^{1/2}$, for some constant $C_1 > 0$ and $\|\boldsymbol{\Gamma}_{\mathcal{S}^c,\mathcal{S}}\boldsymbol{\Gamma}_{\mathcal{S},\mathcal{S}}^{-1}\|_L + 2\|\boldsymbol{\Omega}\|_0\|\boldsymbol{\Gamma}_{\mathcal{S},\mathcal{S}}^{-1}\|_L\|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}\|_\infty \le 1 - \kappa + 2C_1Ms_p(n^{-1}\log p)^{1/2} = 1-\kappa+o(1) < 1$. Next, we consider $\Delta_1 \overset{\text{def}}{=} \|\mathrm{vec}(\widehat{\boldsymbol{\Lambda}}_y) - \mathrm{vec}(\boldsymbol{\Lambda}_y)\|_\infty + 2\|(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma})\mathrm{vec}(\boldsymbol{\Omega})\|_\infty$. Note that

$$\begin{aligned}
\|(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma})\mathrm{vec}(\boldsymbol{\Omega})\|_\infty &= \|(\widehat{\boldsymbol{\Sigma}}\otimes\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})\mathrm{vec}(\boldsymbol{\Omega})\|_\infty\\
&= \|\mathrm{vec}(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Omega}\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\boldsymbol{\Omega}\boldsymbol{\Sigma})\|_\infty = \|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Omega}\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\boldsymbol{\Omega}\boldsymbol{\Sigma}\|_\infty\\
&\le \|(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{\Omega}(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\|_\infty + 2\|\boldsymbol{\Sigma}\boldsymbol{\Omega}(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\|_\infty.
\end{aligned}$$

Under the conditions of the Proposition 1,

$$\mathrm{var}\{E(Y \mid \mathbf{x})\} = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta} + 2\mathrm{tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma}\boldsymbol{\Omega}\boldsymbol{\Sigma}) \leq EY^2 < \infty. \tag{S1.14}$$

We thus conclude $\|\boldsymbol{\Omega}\|_\infty < \infty$ and $\|\boldsymbol{\Omega}\boldsymbol{\Sigma}\| < \infty$. Then, $\|(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{\Omega}(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\|_\infty \leq s_p\|\boldsymbol{\Omega}\|_\infty\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_\infty^2 = o(1)(n^{-1}\log p)^{1/2}$,and $\mathrm{pr}\{\|\boldsymbol{\Sigma}\boldsymbol{\Omega}(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\|_\infty \geq C\{\log(p)/n\}^{1/2}\} = O(p^{-1})$ by invoking Lemma 4 and the fact $\|\boldsymbol{\Sigma}\boldsymbol{\Omega}(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\|_\infty = \max_{i,j} |\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\Omega}(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\mathbf{e}_j|$. Consequently, there exist a constant $C_2 > 0$ such that $\Delta_1 \leq C_2\{n^{-\alpha/(\alpha+1)}\log(p)\}^{1/2}$ with probability larger than $1 - O(p^{-1})$. Set $\lambda_{1n} = 3\kappa^{-1}C_2\{n^{-\alpha/(\alpha+1)}\log(p)\}^{1/2}$ and by Lemma 6. We can conclude that with probability larger than $1 - O(p^{-1})$, $\{\widehat{\boldsymbol{\Omega}}_y\}_{\mathcal{S}} = \mathbf{0}$, and $\|\widehat{\boldsymbol{\Omega}}_y - \boldsymbol{\Omega}\|_\infty \leq 4\kappa^{-1}C_2M\{n^{-\alpha/(\alpha+1)}\log(p)\}^{1/2}$. The proof is now completed.

□

S1.9 Appendix E: Proof of Theorem 2

Given $\widehat{\boldsymbol{\beta}}$,

$$\begin{aligned}\widehat{\boldsymbol{\Lambda}}_r &= n^{-1}\sum_{i=1}^n\{(Y_i - \bar{Y}) - (\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}}\boldsymbol{\beta}\}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}} \\ &\quad + n^{-1}\sum_{i=1}^n(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})\cdot(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}} \stackrel{\mathrm{def}}{=} \mathbf{A}_1 + \mathbf{A}_2.\end{aligned}$$

Given true $\boldsymbol{\beta}$, (S1.14) ensures that $\boldsymbol{\beta}\boldsymbol{\Sigma}\boldsymbol{\beta} \leq EY^2 < \infty$, indicating that $\|\boldsymbol{\beta}\| < C$ for some constant C. Thus, $E\{\exp(c_1|Y - \mathbf{b}^{\mathrm{T}}\mathbf{x}|^\alpha)\} \leq C_1 < \infty$

and with probability greater than $1 - O(p^{-1})$,

$$\|\mathbf{A}_1 - \mathbf{\Lambda}\|_\infty \leq C_1\{n^{-\alpha/(\alpha+1)} \log(p)\}^{1/2}. \tag{S1.15}$$

Writing $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\eta_1, \cdots, \eta_p)^{\mathrm{T}} = \sum_{k=1}^{p} \eta_k \mathbf{e}_k$, we have,

$$\begin{aligned}
\Big\|\frac{1}{n}\sum_{i=1}^{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}}\mathbf{x}_i(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}})\Big\|_\infty &= \Big\|\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{p}\eta_t\mathbf{e}_k\right)^{\mathrm{T}}\mathbf{x}_i(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}})\Big\|_\infty \\
&\leq \sum_{k=1}^{p}|\eta_k| \cdot \Big\|n^{-1}\sum_{i=1}^{n}\mathbf{e}_k^{\mathrm{T}}\mathbf{x}_i(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}})\Big\|_\infty,
\end{aligned}$$

For $\mathbf{e}_k$, $\mathrm{E}\{(\mathbf{e}_k^{\mathrm{T}}\mathbf{x})(\mathbf{x}\mathbf{x}^{\mathrm{T}})\} = 0$. By Lemma 5, there exists a large constant C_2 such that,

$$\mathrm{pr}\Big\{\Big\|n^{-1}\sum_{i=1}^{n}\mathbf{e}_k^{\mathrm{T}}\mathbf{x}_i(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}})\Big\|_\infty \geq C_2(n^{-2/3}\log(p))^{1/2}\Big\} \leq p^{-2},$$

which implies

$$\begin{aligned}
&\mathrm{pr}\Big\{\Big\|n^{-1}\sum_{i=1}^{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}}\mathbf{x}_i(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}})\Big\|_\infty \geq C_2\sum_{k=1}^{p}|\eta_k|(n^{-2/3}\log(p))^{1/2}\Big\} \\
\leq\ &\sum_{k=1}^{p}\mathrm{pr}\Big\{\Big\|n^{-1}\sum_{i=1}^{n}\mathbf{e}_k^{\mathrm{T}}\mathbf{x}_i(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}})\Big\|_\infty \geq C_2(n^{-2/3}\log(p))^{1/2}\Big\} \leq p^{-1}.
\end{aligned}$$

Note that $\sum_{k=1}^{p}|\eta_k| = \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1$. With probability greater than $1 - p^{-1}$,

$$\Big\|n^{-1}\sum_{i=1}^{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}}\mathbf{x}_i(\mathbf{x}_i\mathbf{x}_i)^{\mathrm{T}}\Big\|_\infty \leq C_2\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1\{n^{-2/3}\log(p)\}^{1/2},$$

which together with Lemma 4 yields

$$\|\mathbf{A}_2\|_\infty \leq C_3\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1\{n^{-2/3}\log(p)\}^{1/2}. \tag{S1.16}$$

Combing (S1.15) and (S1.16), with probability greater than $1 - O(p^{-1})$,

$$
\begin{aligned}
& \|\widehat{\boldsymbol{\Lambda}}_r - \boldsymbol{\Lambda}\|_\infty && \text{(S1.17)} \\
\leq \; & \|\widehat{\boldsymbol{\Lambda}}_r - \boldsymbol{\Lambda}\|_\infty C_4 \{n^{-\alpha/(\alpha+1)} \log(p)\}^{1/2} + C_5 \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 \{n^{-2/3} \log(p)\}^{1/2}.
\end{aligned}
$$

Similarly to the proof of the Theorem 1, we can set

$$
\lambda_{2n} = C_6 \{n^{-\alpha/(\alpha+1)} \log(p)\}^{1/2} + C_7 \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 \{n^{-2/3} \log(p)\}^{1/2}
$$

and conclude that with probability lager than $1 - O(p^{-1})$, $\{\widehat{\boldsymbol{\Omega}}_r\}_{\mathcal{S}} = \mathbf{0}$ and $\|\widehat{\boldsymbol{\Omega}}_r - \boldsymbol{\Omega}\|_\infty \leq C_8 M \lambda_{2n}$, for some constant C_8.

References

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of Statistics 32*(2), 407–499.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software 33*(1), 1.

Hao, N., Y. Feng, and H. H. Zhang (2018). Model selection for high-dimensional quadratic regression via regularization. *Journal of the American Statistical Association 113*(522), 615–625.

Hao, N. and H. H. Zhang (2014). Interaction screening for ultrahigh-dimensional data. *Journal of the American Statistical Association 109*(507), 1285–1301.

Hong, M. and Z.-Q. Luo (2017). On the linear convergence of the alternating direction method of multipliers. *Mathematical Programming 162*(1-2), 165–199.

Tibshirani, R., J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R. J. Tibshirani (2012). Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*(2), 245–266.

Vershynin, R. (2017). *High Dimensional Probability.* In press.

School of Mathematical Sciences, MOE-LSC, Shanghai Jiao Tong University, Shanghai 200240, China.

E-mail: chengwang@sjtu.edu.cn

Department of Applied Mathematics, Hong Kong Polytechnic University, Hong Kong.

E-mail: by.jiang@polyu.edu.hk

Institute of Statistics and Big Data and Center for Applied Statistics, Renmin University of China, Beijing 100872, China.

E-mail: zhu.liping@ruc.edu.cn