Web Supplementary for "Fully efficient robust estimation, outlier detection and variable selection via penalized regression"

Dehan Kong[1], Howard Bondell[2] and Yichao Wu[2]

[1]*University of Toronto and* [2]*North Carolina State University*

In this Web Supplement, we first include additional simulations results for different settings in Section 1. Then, we present important lemmas needed for the main theorems and the corollaries in Section 2. The proof of the lemmas are also in this section. Section 3 presents the proofs of all the theorems. Section 4 proves all of the corollaries.

1 Additional Simulation Results

In this section, we include more simulation results by considering more simulation settings. In particularly, we consider the following settings:

- Setting I: We consider different structure of correlation structure of the design matrix X. In this setting, we consider the compound symmetry structure for the correlation matrix. Specifically, the covariate X_i is generated independently and identically from a multivariate normal distribution with zero mean and covariance matrix Σ, where the jkth element of the matrix $\Sigma_{jk} = 0.5 + 0.5 * 1(j = k)$. For the other settings, they are the same as the simulation settings in the main paper. The results corresponding to Setting I are in Table 1 ($n = 100$) and Table 2 ($n = 200$).

- Setting II: In this case, we consider different γ values, i.e. the γ's are not constant any more. In particular, we contaminate the first cn observations by setting $X_i^* = X_i + L_i$ and $y_i^* = y_i + V_i$ for $1 \le i \le cn$. We consider $(L, V) = (0, 4 + 0.2i)$ and $(L, V) = (4 + 0.2i, 4 + 0.2i)$. For the other settings, they are the same as the simulation settings in the main paper. The results corresponding to Setting II are in Table 3 ($n = 100$) and Table 4 ($n = 200$).

- Setting III: In this case, we consider different combinations of p, q and β values. In particular, we set $p = 25$ and $q = 10$. The true $\beta_0 = (1, \ldots, 1, 0, \ldots, 0)^{\mathrm{T}}$ with $q = 10$ nonzero components and the remaining 15 elements being zero. For the other settings, they are the same as the simulation settings in the main paper. The results corresponding to Setting III are in Table 5 ($n = 100$) and Table 6 ($n = 200$).

From the results, we can see that the findings are similar as those in the main paper.

2 Lemmas and the Proof of Lemmas

In this section, we state the lemmas that are needed for the theorems and corollaries.

Denote $X_{i,j}$ the ijth element of the design matrix X. Without loss of generality, we assume the first s_n points as outliers. Recall that $G = \{s_n + 1, s_n + 2, \ldots, n\}$, which denotes the indices corresponding to the good points, and define $G^c = \{1, \ldots, s_n\}$, which denotes the indices of outliers. Denote $\delta_n = \min_{i \in G^c} |\widetilde{\gamma}_i|$ and $\kappa_n = \max_{i \in G} |\widetilde{\gamma}_i|$. When no outliers exist, we have $G = \{1, 2, \ldots, n\}$, while $G^c = \emptyset$. Thus, only κ_n is properly defined in this case.

When p and q are fixed, we use least trimmed squares as the initial estimates. Under this scenario, we have the first two lemmas, which are used for proofs of Theorems 1 and 3. In particular, Lemma 1 gives the rate of κ_n when no outliers exist. Lemma 2 gives the rate of κ_n and δ_n when outliers exist.

Lemma 1. *Under Condition (A), we have* $\kappa_n = O_P(n^{1/(2k)})$.

Proof. As the least trimmed estimator is root-n consistent (Rousseeuw and Leroy, 1987), we have $||\widetilde{\beta} - \beta_0||_2 = O_P(n^{-1/2})$. By Condition (B2), one has $\max_{1 \le i \le n} ||X_i||_2^2 \le \sum_{j=1}^{p} X_{,j}^{\mathrm{T}} X_{,j} \le pnM_1$, which indicates $\max_{1 \le i \le n} ||X_i||_2 = O_P(n^{1/2})$. Thus,
$\max_{1 \le i \le n} |X_i(\widetilde{\beta} - \beta_0)| \le \max_{1 \le i \le n} ||X_i|| O_P(n^{-1/2}) = O_P(1)$. Notice that $\widetilde{\gamma}_i = \epsilon_i - X_i(\widetilde{\beta} - \beta_0)$, we obtain $\kappa_n = O_P(n^{1/(2k)})$ as $\epsilon_i = O_P(n^{1/(2k)})$ by Condition (A). □

Denote $a_n \equiv O_P(b_n)$ if a_n and b_n have the same order in probability.

Lemma 2. *Under Conditions (A),(C1),(C2), and (C3), one has* $\kappa_n = O_P(n^{1/(2k)})$. *we have* $\delta_n \equiv O_P(\pi_n)$.

Proof. As the least trimmed estimator is root-n consistent (Rousseeuw and Leroy, 1987) under Condition (C2), we have $||\widetilde{\beta} - \beta_0||_2 = O_P(n^{-1/2})$. By Condition (B2), one has $\max_{1 \le i \le n} |X_i(\widetilde{\beta} - \beta_0)| \le \max_{1 \le i \le n} ||X_i|| O_P(n^{-1/2}) = O_P(1)$. For $i \in G$, one has $\widetilde{\gamma}_i = -X_i(\widetilde{\beta} - \beta_0) + \epsilon_i$, by Condition (A), one obtains $\max_{i \in G} |\widetilde{\gamma}_i| = O_P(n^{1/(2k)})$. For $i \in G^c$, one has $\widetilde{\gamma}_i = -X_i(\widetilde{\beta} - \beta_0) + \epsilon_i + \gamma_{i0}$. Recall that $\pi_n = \min_{i=1,\ldots,s_n} |\gamma_{i0}|$, and $\pi_n n^{-1/(2k)} \to \infty$ by Condition (C1), one obtains $\delta_n = \min_{i \in G^c} |\widetilde{\gamma}_i| \equiv O_P(\pi_n)$. □

The next three lemmas are used when p_n diverges at the exponential rate of n and for proofs of Corollaries 1 and 3. Under this scenario, we use the sparse least trimmed squares as the initial estimates. Lemma 3 states that the sparse least trimmed squares estimates have order $O_P(1)$. With a little abuse of the notation, we still use the same definitions that $\delta_n = \min_{i \in G^c} |\widetilde{\gamma}_i|$ and $\kappa_n = \max_{i \in G} |\widetilde{\gamma}_i|$, where $\widetilde{\gamma}_i$ are the residuals obtained from the sparse least trimmed squares. Lemma 4 gives the rate of κ_n when no outliers exist. Lemma 5 gives the rate of κ_n and δ_n when outliers exist.

Lemma 3. *Under Conditions (C2), (D2) and (D3), one has* $||\hat{\beta}_{H_{opt}}||_2 = O_P(1)$.

Proof. Denote $\{(\widetilde{X}_i, \widetilde{y}_i), 1 \le i \le n\}$ as the contaminated sample, and we have $(\widetilde{X}_i, \widetilde{y}_i) = (X_i, y_i)$ for all $i \in G$. As $h \le n - s_n$, one has

$$\min_H Q(H, 0) = \min_H \sum_{i \in H} \widetilde{y}_i^2 \le \min_{H \subset G} \min_H \sum_{i \in H} \widetilde{y}_i^2 = \min_{H \subset G} \min_H \sum_{i \in H} y_i^2 \le h(n - s_n)^{-1} \sum_{i \in G} y_i^2.$$

As $E(y_i^2) < \infty$ for all $i \in G$ by Condition (D3), one has $\min_H Q(H, 0) = O_P(h)$.

Then we will prove that there exists a constant M such that $||\hat{\beta}_{H_{opt}}||_2 \le M$ with probability 1. For any $||\beta|| > M$, one has $\min_{H,\beta} Q(H, \beta) \ge h\lambda ||\beta||_1 \ge h\lambda ||\beta||_2 > h\lambda_s M$. When $\lambda_s \to \infty$ by Condition (D2), one has $\min_H Q(H, \beta) > \min_H Q(H, 0)$ with probability one. Since $Q(H_{opt}, \hat{\beta}_{H_{opt}}) \le \min_H Q(H, 0)$, one has $||\hat{\beta}_{H_{opt}}||_2 \le M$ with probability 1, i.e. $||\hat{\beta}_{H_{opt}}||_2 = O_P(1)$. □

Lemma 4. *Under Conditions (C2), (D1)-(D3), we have* $\kappa_n = O_P(n)$.

Table 1: Simulation Setting I: Simulation results for our methods PM compared with the SLTS, LL, REWLS, SHE, LASSO, ALASSO, ORACLE and AORACLE methods when $n = 100$. The $*$ denotes the values that are not applicable.

(n,p,V,L,c)	method	M	S	JD	FZR	FPR	SR	CR	MSE
(100,15,4,0,0.1)	PM	0(0)	0.01(0)	1	0.01(0.002)	0.21(0.004)	0.09	0.93	0.2(0.002)
	SLTS	0(0)	0.04(0.001)	1	0.02(0.002)	0.38(0.005)	0.01	0.91	0.29(0.002)
	LL	*	*	*	0.05(0.003)	0.38(0.01)	0.1	0.75	0.23(0.002)
	REWLS	0(0)	0.01(0)	1	*	*	*	*	0.22(0.001)
	SHE	0(0)	0.08(0.002)	0.999	*	*	*	*	0.3(0.002)
	LASSO	*	*	*	0.08(0.003)	0.25(0.005)	0.04	0.61	0.57(0.003)
	ALASSO	*	*	*	0.21(0.004)	0.05(0.003)	0.09	0.18	3.47(0.014)
	ORACLE	*	*	*	0.01(0.001)	0.27(0.005)	0.06	0.97	0.17(0.001)
	AORACLE	*	*	*	0.06(0.003)	0.05(0.003)	0.42	0.71	0.16(0.001)
(100,15,4,0,0.2)	PM	0.01(0.002)	0.02(0.001)	0.994	0.02(0.002)	0.23(0.004)	0.07	0.9	0.22(0.003)
	SLTS	0(0)	0.01(0)	0.999	0.01(0.002)	0.36(0.005)	0.02	0.93	0.23(0.002)
	LL	*	*	*	0.03(0.002)	0.71(0.008)	0.01	0.86	0.29(0.003)
	REWLS	0(0)	0(0)	0.995	*	*	*	*	0.23(0.001)
	SHE	0.44(0.016)	0.08(0.003)	0.552	*	*	*	*	0.72(0.01)
	LASSO	*	*	*	0.12(0.004)	0.25(0.005)	0.04	0.47	0.96(0.003)
	ALASSO	*	*	*	0.25(0.004)	0.05(0.003)	0.06	0.1	3.65(0.014)
	ORACLE	*	*	*	0.01(0.001)	0.27(0.005)	0.06	0.95	0.18(0.001)
	AORACLE	*	*	*	0.07(0.003)	0.05(0.003)	0.39	0.66	0.17(0.002)
(100,15,4,4,0.1)	PM	0(0)	0(0)	1	0.01(0.002)	0.2(0.004)	0.1	0.93	0.2(0.002)
	SLTS	0(0)	0.03(0.001)	1	0.02(0.002)	0.38(0.005)	0.01	0.9	0.27(0.002)
	LL	*	*	*	0.03(0.003)	0.86(0.006)	0	0.84	1.15(0.021)
	REWLS	0(0)	0.01(0)	1	*	*	*	*	0.22(0.001)
	SHE	0.7(0.012)	0.07(0.004)	0.141	*	*	*	*	3.05(0.028)
	LASSO	*	*	*	0.61(0.005)	0.09(0.004)	0	0	3.83(0.008)
	ALASSO	*	*	*	0.66(0.006)	0.05(0.003)	0	0.01	6.61(0.065)
	ORACLE	*	*	*	0.01(0.001)	0.27(0.005)	0.06	0.97	0.17(0.001)
	AORACLE	*	*	*	0.06(0.003)	0.05(0.003)	0.42	0.71	0.16(0.001)
(100,15,4,4,0.2)	PM	0(0)	0.01(0)	1	0.02(0.002)	0.21(0.004)	0.1	0.9	0.22(0.002)
	SLTS	0(0)	0.01(0)	1	0.01(0.002)	0.38(0.006)	0.02	0.94	0.23(0.002)
	LL	*	*	*	0.01(0.002)	0.98(0.002)	0	0.95	4.05(0.009)
	REWLS	0(0)	0(0)	1	*	*	*	*	0.23(0.001)
	SHE	0.94(0.004)	0.03(0.002)	0	*	*	*	*	3.92(0.007)
	LASSO	*	*	*	0.65(0.005)	0.12(0.005)	0	0	4.32(0.011)
	ALASSO	*	*	*	0.66(0.006)	0.07(0.004)	0	0.01	6.9(0.128)
	ORACLE	*	*	*	0.01(0.001)	0.27(0.005)	0.06	0.95	0.18(0.001)
	AORACLE	*	*	*	0.07(0.003)	0.05(0.003)	0.39	0.66	0.17(0.002)
(100,15,0,0,0)	PM	*	0(0)	*	0.01(0.001)	0.21(0.004)	0.11	0.95	0.18(0.002)
	SLTS	*	0.07(0.001)	*	0.02(0.002)	0.38(0.005)	0.02	0.89	0.32(0.003)
	LL	*	*	*	0.04(0.003)	0.33(0.009)	0.14	0.78	0.21(0.001)
	REWLS	*	0.02(0.001)	*	*	*	*	*	0.22(0.001)
	SHE	*	0.01(0.001)	*	*	*	*	*	0.2(0.001)
	LASSO	*	*	*	0(0.001)	0.27(0.005)	0.05	0.98	0.16(0.001)
	ALASSO	*	*	*	0.05(0.003)	0.05(0.002)	0.47	0.73	0.15(0.001)

Table 2: Simulation Setting I: Simulation results for our methods PM compared with the SLTS, LL, REWLS, She, LASSO, ALASSO, Oracle and Aoracle methods when $n = 200$. The $*$ denotes the values that are not applicable.

(n,p,V,L,c)	method	M	S	JD	FZR	FPR	SR	CR	MSE
(200,15,4,0,0.1)	PM	0(0)	0.01(0)	1	0(0)	0.21(0.004)	0.12	1	0.14(0.001)
	SLTS	0(0)	0.02(0)	1	0(0.001)	0.36(0.005)	0.02	0.99	0.2(0.002)
	LL	*	*	*	0.03(0.002)	0.21(0.007)	0.28	0.86	0.17(0.001)
	REWLS	0(0)	0(0)	1	*	*	*	*	0.15(0.001)
	She	0(0)	0.06(0.001)	1	*	*	*	*	0.22(0.001)
	LASSO	*	*	*	0.04(0.003)	0.24(0.005)	0.06	0.8	0.5(0.002)
	ALASSO	*	*	*	0.15(0.003)	0.03(0.002)	0.21	0.31	3.34(0.01)
	Oracle	*	*	*	0(0)	0.25(0.005)	0.08	1	0.12(0.001)
	Aoracle	*	*	*	0.01(0.001)	0.03(0.002)	0.75	0.95	0.1(0.001)
(200,15,4,0,0.2)	PM	0.02(0.005)	0.02(0)	0.975	0(0.001)	0.21(0.004)	0.12	0.98	0.19(0.004)
	SLTS	0(0)	0(0)	1	0(0)	0.35(0.005)	0.02	1	0.16(0.001)
	LL	*	*	*	0.01(0.001)	0.63(0.007)	0.01	0.95	0.23(0.002)
	REWLS	0(0)	0(0)	0.996	*	*	*	*	0.16(0.001)
	She	0.38(0.015)	0.07(0.002)	0.623	*	*	*	*	0.57(0.009)
	LASSO	*	*	*	0.07(0.003)	0.24(0.005)	0.07	0.68	0.88(0.002)
	ALASSO	*	*	*	0.2(0.004)	0.03(0.002)	0.11	0.16	3.54(0.011)
	Oracle	*	*	*	0(0)	0.25(0.005)	0.08	1	0.13(0.001)
	Aoracle	*	*	*	0.01(0.002)	0.03(0.002)	0.69	0.93	0.11(0.001)
(200,15,4,4,0.1)	PM	0(0)	0(0)	1	0(0.001)	0.19(0.004)	0.14	0.99	0.15(0.001)
	SLTS	0(0)	0.02(0)	1	0(0.001)	0.35(0.005)	0.02	0.99	0.19(0.002)
	LL	*	*	*	0.05(0.003)	0.71(0.008)	0	0.76	0.76(0.008)
	REWLS	0(0)	0(0)	1	*	*	*	*	0.15(0.001)
	She	0.78(0.011)	0.05(0.003)	0.102	*	*	*	*	3.25(0.025)
	LASSO	*	*	*	0.55(0.005)	0.18(0.006)	0	0.01	3.77(0.006)
	ALASSO	*	*	*	0.55(0.008)	0.1(0.005)	0	0.05	8.06(0.149)
	Oracle	*	*	*	0(0)	0.25(0.005)	0.08	1	0.12(0.001)
	Aoracle	*	*	*	0.01(0.001)	0.03(0.002)	0.75	0.95	0.1(0.001)
(200,15,4,4,0.2)	PM	0(0)	0(0)	1	0(0.001)	0.18(0.004)	0.16	0.99	0.18(0.001)
	SLTS	0(0)	0(0)	1	0(0)	0.37(0.005)	0.02	1	0.15(0.001)
	LL	*	*	*	0.01(0.002)	0.98(0.002)	0	0.94	4.09(0.006)
	REWLS	0(0)	0(0)	1	*	*	*	*	0.16(0.001)
	She	0.96(0.003)	0.02(0.002)	0	*	*	*	*	4.04(0.005)
	LASSO	*	*	*	0.59(0.006)	0.22(0.007)	0	0.01	4.27(0.008)
	ALASSO	*	*	*	0.55(0.008)	0.13(0.005)	0	0.04	8.29(0.174)
	Oracle	*	*	*	0(0)	0.25(0.005)	0.08	1	0.13(0.001)
	Aoracle	*	*	*	0.01(0.002)	0.03(0.002)	0.69	0.93	0.11(0.001)
(200,15,0,0,0)	PM	*	0(0)	*	0(0)	0.2(0.004)	0.14	1	0.13(0.001)
	SLTS	*	0.04(0.001)	*	0(0.001)	0.35(0.005)	0.02	0.98	0.23(0.002)
	LL	*	*	*	0.03(0.002)	0.09(0.005)	0.51	0.83	0.15(0.001)
	REWLS	*	0.01(0)	*	*	*	*	*	0.15(0.001)
	She	*	0(0)	*	*	*	*	*	0.14(0.001)
	LASSO	*	*	*	0(0)	0.25(0.005)	0.07	1	0.11(0.001)
	ALASSO	*	*	*	0.01(0.001)	0.02(0.002)	0.79	0.96	0.09(0.001)

Table 3: Simulation Setting II: Simulation results for our methods PM compared with the SLTS, LL, REWLS, She, LASSO, ALASSO, Oracle and Aoracle methods when $n = 100$. The $*$ denotes the values that are not applicable.

(n,p,V,L,c)	method	M	S	JD	FZR	FPR	SR	CR	MSE
(100,15,4+0.2i,0,0.1)	PM	0(0)	0.01(0)	1	0.02(0.002)	0.04(0.002)	0.59	0.92	0.2(0.002)
	SLTS	0(0)	0.03(0.001)	1	0.02(0.002)	0.3(0.006)	0.1	0.92	0.28(0.003)
	LL	*	*	*	0.04(0.003)	0.37(0.01)	0.14	0.8	0.22(0.002)
	REWLS	0(0)	0.01(0)	1	*	*	*	*	0.22(0.001)
	She	0(0)	0.08(0.002)	1	*	*	*	*	0.3(0.002)
	LASSO	*	*	*	0.12(0.003)	0.09(0.004)	0.18	0.42	0.72(0.003)
	ALASSO	*	*	*	0.22(0.004)	0.04(0.003)	0.06	0.12	3.7(0.009)
	Oracle	*	*	*	0.01(0.001)	0.11(0.004)	0.36	0.97	0.17(0.001)
	Aoracle	*	*	*	0.04(0.003)	0.05(0.003)	0.51	0.78	0.16(0.001)
(100,15,4+0.2i,0,0.2)	PM	0(0)	0.01(0)	1	0.02(0.002)	0.06(0.003)	0.51	0.9	0.21(0.002)
	SLTS	0(0)	0.01(0)	1	0.01(0.001)	0.29(0.007)	0.14	0.94	0.22(0.002)
	LL	*	*	*	0.02(0.002)	0.7(0.008)	0.02	0.89	0.29(0.003)
	REWLS	0(0)	0(0)	1	*	*	*	*	0.23(0.001)
	She	0.13(0.01)	0.12(0.002)	0.865	*	*	*	*	0.63(0.012)
	LASSO	*	*	*	0.2(0.004)	0.07(0.003)	0.1	0.22	1.45(0.004)
	ALASSO	*	*	*	0.31(0.004)	0.04(0.002)	0.02	0.04	3.97(0.01)
	Oracle	*	*	*	0.01(0.001)	0.12(0.004)	0.36	0.95	0.18(0.002)
	Aoracle	*	*	*	0.05(0.003)	0.05(0.003)	0.47	0.76	0.17(0.002)
(100,15,4+0.2i,4+0.2i,0.1)	PM	0(0)	0(0)	1	0.03(0.002)	0.05(0.002)	0.5	0.87	0.22(0.002)
	SLTS	0(0)	0.03(0.001)	1	0.02(0.002)	0.31(0.007)	0.11	0.91	0.27(0.002)
	LL	*	*	*	0.04(0.003)	0.95(0.003)	0	0.85	2.59(0.006)
	REWLS	0(0)	0.01(0)	1	*	*	*	*	0.22(0.001)
	She	0.94(0.004)	0.03(0.002)	0	*	*	*	*	2.45(0.005)
	LASSO	*	*	*	0.32(0.005)	0.75(0.004)	0	0.09	2.51(0.006)
	ALASSO	*	*	*	0.26(0.006)	0.58(0.005)	0	0.2	25.16(6.203)
	Oracle	*	*	*	0.01(0.001)	0.11(0.004)	0.36	0.97	0.17(0.001)
	Aoracle	*	*	*	0.04(0.003)	0.05(0.003)	0.51	0.78	0.16(0.001)
(100,15,4+0.2i,4+0.2i,0.2)	PM	0(0)	0.01(0)	1	0.04(0.003)	0.09(0.003)	0.35	0.8	0.25(0.002)
	SLTS	0(0)	0.01(0)	1	0.01(0.001)	0.28(0.007)	0.15	0.94	0.23(0.002)
	LL	*	*	*	0.03(0.003)	0.96(0.003)	0	0.89	2.71(0.006)
	REWLS	0(0)	0(0)	1	*	*	*	*	0.23(0.001)
	She	0.97(0.003)	0.03(0.002)	0	*	*	*	*	2.55(0.006)
	LASSO	*	*	*	0.34(0.005)	0.76(0.004)	0	0.08	2.62(0.007)
	ALASSO	*	*	*	0.3(0.006)	0.61(0.005)	0	0.15	11.08(0.596)
	Oracle	*	*	*	0.01(0.001)	0.12(0.004)	0.36	0.95	0.18(0.002)
	Aoracle	*	*	*	0.05(0.003)	0.05(0.003)	0.47	0.76	0.17(0.002)
(100,15,0,0,0)	PM	*	0(0)	*	0.01(0.002)	0.03(0.002)	0.68	0.94	0.18(0.002)
	SLTS	*	0.08(0.001)	*	0.03(0.002)	0.28(0.007)	0.11	0.86	0.27(0.003)
	LL	*	*	*	0.04(0.003)	0.3(0.009)	0.2	0.81	0.2(0.002)
	REWLS	*	0.02(0.001)	*	*	*	*	*	0.22(0.001)
	She	*	0.01(0.001)	*	*	*	*	*	0.2(0.001)
	LASSO	*	*	*	0(0.001)	0.11(0.004)	0.41	0.98	0.16(0.001)
	ALASSO	*	*	*	0.04(0.002)	0.05(0.003)	0.54	0.82	0.15(0.001)

Table 4: Simulation Setting II: Simulation results for our methods PM compared with the SLTS, LL, REWLS, She, LASSO, ALASSO, Oracle and Aoracle methods when $n = 200$. The $*$ denotes the values that are not applicable.

(n,p,V,L,c)	method	M	S	JD	FZR	FPR	SR	CR	MSE
(200,15,4+0.2i,0,0.1)	PM	0(0)	0(0)	1	0(0)	0.03(0.002)	0.72	1	0.14(0.001)
	SLTS	0(0)	0.02(0)	1	0(0.001)	0.26(0.005)	0.08	0.99	0.2(0.002)
	LL	*	*	*	0.02(0.002)	0.2(0.007)	0.34	0.91	0.16(0.001)
	REWLS	0(0)	0(0)	1	*	*	*	*	0.15(0.001)
	She	0(0)	0.06(0.001)	1	*	*	*	*	0.22(0.001)
	LASSO	*	*	*	0.1(0.003)	0.06(0.003)	0.28	0.51	0.75(0.002)
	ALASSO	*	*	*	0.18(0.003)	0.03(0.002)	0.12	0.18	3.67(0.006)
	Oracle	*	*	*	0(0)	0.08(0.003)	0.5	1	0.12(0.001)
	Aoracle	*	*	*	0(0.001)	0.02(0.002)	0.79	0.98	0.1(0.001)
(200,15,4+0.2i,0,0.2)	PM	0(0)	0.01(0)	1	0(0.001)	0.04(0.002)	0.67	0.99	0.15(0.001)
	SLTS	0(0)	0(0)	1	0(0)	0.27(0.005)	0.08	1	0.15(0.001)
	LL	*	*	*	0.01(0.001)	0.61(0.007)	0.01	0.97	0.23(0.002)
	REWLS	0(0)	0(0)	1	*	*	*	*	0.16(0.001)
	She	0.01(0.004)	0.11(0.002)	0.985	*	*	*	*	0.39(0.006)
	LASSO	*	*	*	0.2(0.004)	0.06(0.003)	0.1	0.2	1.79(0.003)
	ALASSO	*	*	*	0.31(0.004)	0.03(0.002)	0.02	0.04	4.09(0.007)
	Oracle	*	*	*	0(0)	0.09(0.003)	0.46	1	0.13(0.001)
	Aoracle	*	*	*	0.01(0.001)	0.03(0.002)	0.75	0.96	0.11(0.001)
(200,15,4+0.2i,4+0.2i,0.1)	PM	0(0)	0(0)	1	0(0.001)	0.06(0.003)	0.57	0.98	0.17(0.001)
	SLTS	0(0)	0.02(0)	1	0(0.001)	0.27(0.005)	0.1	0.99	0.18(0.002)
	LL	*	*	*	0.06(0.004)	0.94(0.003)	0	0.76	2.67(0.004)
	REWLS	0(0)	0(0)	1	*	*	*	*	0.15(0.001)
	She	0.98(0.002)	0.01(0.001)	0	*	*	*	*	2.59(0.004)
	LASSO	*	*	*	0.28(0.005)	0.89(0.003)	0	0.1	2.61(0.004)
	ALASSO	*	*	*	0.17(0.005)	0.73(0.004)	0	0.34	25.08(1.941)
	Oracle	*	*	*	0(0)	0.08(0.003)	0.5	1	0.12(0.001)
	Aoracle	*	*	*	0(0.001)	0.02(0.002)	0.79	0.98	0.1(0.001)
(200,15,4+0.2i,4+0.2i,0.2)	PM	0(0)	0.01(0)	1	0.01(0.001)	0.12(0.004)	0.32	0.95	0.22(0.002)
	SLTS	0(0)	0(0)	1	0(0)	0.3(0.006)	0.12	1	0.15(0.001)
	LL	*	*	*	0.06(0.004)	0.94(0.003)	0	0.78	2.77(0.005)
	REWLS	0(0)	0(0)	1	*	*	*	*	0.16(0.001)
	She	0.99(0.002)	0.01(0.001)	0	*	*	*	*	2.68(0.004)
	LASSO	*	*	*	0.28(0.005)	0.89(0.003)	0	0.09	2.7(0.004)
	ALASSO	*	*	*	0.22(0.005)	0.74(0.004)	0	0.22	18.83(3.148)
	Oracle	*	*	*	0(0)	0.09(0.003)	0.46	1	0.13(0.001)
	Aoracle	*	*	*	0.01(0.001)	0.03(0.002)	0.75	0.96	0.11(0.001)
(200,15,0,0,0)	PM	*	0(0)	*	0(0)	0.02(0.002)	0.78	1	0.12(0.001)
	SLTS	*	0.05(0.001)	*	0.01(0.001)	0.22(0.005)	0.14	0.97	0.19(0.002)
	LL	*	*	*	0.02(0.002)	0.07(0.004)	0.6	0.9	0.14(0.001)
	REWLS	*	0.01(0)	*	*	*	*	*	0.15(0.001)
	She	*	0(0)	*	*	*	*	*	0.14(0.001)
	LASSO	*	*	*	0(0)	0.07(0.003)	0.52	1	0.11(0.001)
	ALASSO	*	*	*	0(0.001)	0.02(0.002)	0.83	0.99	0.09(0.001)

Table 5: Simulation Setting III: Simulation results for our methods PM compared with the SLTS, LL, REWLS, She, LASSO, ALASSO, Oracle and Aoracle methods when $n = 100$. The $*$ denotes the values that are not applicable.

(n,p,V,L,c)	method	M	S	JD	FZR	FPR	SR	CR	MSE
(100,25,4,0,0.1)	PM	0(0)	0.01(0.001)	0.999	0(0)	0.07(0.002)	0.41	1	0.27(0.002)
	SLTS	0(0)	0.08(0.001)	0.999	0(0)	0.38(0.007)	0.04	1	0.38(0.003)
	LL	*	*	*	0(0)	0.38(0.009)	0.19	1	0.25(0.002)
	REWLS	0(0)	0.01(0)	0.996	*	*	*	*	0.29(0.002)
	She	0(0)	0.09(0.002)	0.997	*	*	*	*	0.39(0.002)
	LASSO	*	*	*	0(0)	0.14(0.004)	0.18	1	0.68(0.003)
	ALASSO	*	*	*	0(0)	0.06(0.002)	0.52	1	0.67(0.011)
	Oracle	*	*	*	0(0)	0.15(0.004)	0.17	1	0.22(0.001)
	Aoracle	*	*	*	0(0)	0.03(0.002)	0.73	1	0.18(0.001)
(100,25,4,0,0.2)	PM	0.01(0.003)	0.02(0.001)	0.985	0(0)	0.1(0.003)	0.28	1	0.31(0.004)
	SLTS	0(0)	0.02(0.001)	0.999	0(0)	0.33(0.007)	0.05	1	0.31(0.002)
	LL	*	*	*	0(0)	0.66(0.009)	0.05	1	0.34(0.003)
	REWLS	0.06(0.001)	0(0)	0.042	*	*	*	*	0.41(0.003)
	She	0.66(0.014)	0.05(0.003)	0.257	*	*	*	*	1(0.008)
	LASSO	*	*	*	0(0)	0.14(0.004)	0.19	1	1.07(0.003)
	ALASSO	*	*	*	0.01(0.001)	0.07(0.003)	0.38	0.92	1.09(0.01)
	Oracle	*	*	*	0(0)	0.16(0.004)	0.16	1	0.24(0.002)
	Aoracle	*	*	*	0(0)	0.04(0.003)	0.68	1	0.2(0.001)
(100,25,4,4,0.1)	PM	0(0)	0.01(0)	1	0(0)	0.09(0.003)	0.32	1	0.28(0.002)
	SLTS	0(0)	0.08(0.001)	1	0(0)	0.37(0.007)	0.05	1	0.37(0.003)
	LL	*	*	*	0.03(0.002)	0.95(0.003)	0	0.8	2.75(0.007)
	REWLS	0(0)	0.01(0)	1	*	*	*	*	0.29(0.002)
	She	0.94(0.004)	0.04(0.002)	0	*	*	*	*	2.48(0.006)
	LASSO	*	*	*	0.21(0.005)	0.54(0.006)	0	0.14	2.68(0.012)
	ALASSO	*	*	*	0.15(0.005)	0.42(0.005)	0	0.28	10.97(0.526)
	Oracle	*	*	*	0(0)	0.15(0.004)	0.17	1	0.22(0.001)
	Aoracle	*	*	*	0(0)	0.03(0.002)	0.73	1	0.18(0.001)
(100,25,4,4,0.2)	PM	1(0)	0(0)	0	0.75(0.005)	0.03(0.003)	0	0	4.3(0.018)
	SLTS	0(0)	0.02(0.001)	1	0(0)	0.32(0.007)	0.06	1	0.31(0.002)
	LL	*	*	*	0.03(0.002)	0.96(0.003)	0	0.84	2.83(0.007)
	REWLS	0.93(0.002)	0.05(0.001)	0	*	*	*	*	2.79(0.009)
	She	0.96(0.003)	0.03(0.002)	0	*	*	*	*	2.55(0.006)
	LASSO	*	*	*	0.23(0.005)	0.55(0.006)	0	0.12	2.77(0.012)
	ALASSO	*	*	*	0.19(0.006)	0.43(0.006)	0	0.24	8.65(0.242)
	Oracle	*	*	*	0(0)	0.16(0.004)	0.16	1	0.24(0.002)
	Aoracle	*	*	*	0(0)	0.04(0.003)	0.68	1	0.2(0.001)
(100,25,0,0,0)	PM	*	0(0)	*	0(0)	0.05(0.002)	0.49	1	0.23(0.002)
	SLTS	*	0.15(0.001)	*	0(0)	0.43(0.006)	0.02	1	0.4(0.003)
	LL	*	*	*	0(0)	0.4(0.009)	0.2	1	0.23(0.002)
	REWLS	*	0.05(0.001)	*	*	*	*	*	0.32(0.002)
	She	*	0.01(0.001)	*	*	*	*	*	0.25(0.001)
	LASSO	*	*	*	0(0)	0.14(0.004)	0.2	1	0.21(0.001)
	ALASSO	*	*	*	0(0)	0.03(0.002)	0.76	1	0.17(0.001)

Table 6: Simulation Setting III: Simulation results for our methods PM compared with the SLTS, LL, REWLS, She, LASSO, ALASSO, Oracle and Aoracle methods when $n = 200$. The $*$ denotes the values that are not applicable.

(n,p,V,L,c)	method	M	S	JD	FZR	FPR	SR	CR	MSE
(200,25,4,0,0.1)	PM	0(0)	0.01(0)	1	0(0)	0.05(0.002)	0.49	1	0.2(0.001)
	SLTS	0(0)	0.03(0)	1	0(0)	0.35(0.006)	0.04	1	0.26(0.002)
	LL	*	*	*	0(0)	0.16(0.006)	0.42	1	0.17(0.001)
	REWLS	0(0)	0(0)	1	*	*	*	*	0.19(0.001)
	She	0(0)	0.07(0.002)	0.999	*	*	*	*	0.27(0.001)
	LASSO	*	*	*	0(0)	0.11(0.003)	0.27	1	0.56(0.002)
	ALASSO	*	*	*	0(0)	0.02(0.002)	0.76	1	0.51(0.007)
	Oracle	*	*	*	0(0)	0.1(0.003)	0.29	1	0.16(0.001)
	Aoracle	*	*	*	0(0)	0.01(0.001)	0.85	1	0.13(0.001)
(200,25,4,0,0.2)	PM	0.02(0.005)	0.02(0)	0.979	0(0)	0.06(0.002)	0.43	1	0.23(0.004)
	SLTS	0(0)	0.01(0)	1	0(0)	0.33(0.006)	0.05	1	0.21(0.001)
	LL	*	*	*	0(0)	0.58(0.007)	0.03	1	0.24(0.002)
	REWLS	0(0)	0(0)	0.994	*	*	*	*	0.2(0.001)
	She	0.49(0.016)	0.06(0.002)	0.505	*	*	*	*	0.71(0.009)
	LASSO	*	*	*	0(0)	0.1(0.003)	0.29	1	0.95(0.002)
	ALASSO	*	*	*	0(0)	0.03(0.002)	0.68	1	0.9(0.006)
	Oracle	*	*	*	0(0)	0.11(0.003)	0.28	1	0.17(0.001)
	Aoracle	*	*	*	0(0)	0.02(0.001)	0.84	1	0.13(0.001)
(200,25,4,4,0.1)	PM	0(0)	0(0)	1	0(0)	0.08(0.002)	0.36	1	0.22(0.001)
	SLTS	0(0)	0.03(0)	1	0(0)	0.34(0.006)	0.06	1	0.24(0.002)
	LL	*	*	*	0.02(0.002)	0.93(0.004)	0	0.82	2.8(0.004)
	REWLS	0(0)	0(0)	1	*	*	*	*	0.19(0.001)
	She	0.97(0.003)	0.01(0.001)	0	*	*	*	*	2.64(0.004)
	LASSO	*	*	*	0.05(0.002)	0.76(0.004)	0	0.6	2.68(0.005)
	ALASSO	*	*	*	0.03(0.002)	0.63(0.005)	0	0.75	27.65(1.278)
	Oracle	*	*	*	0(0)	0.1(0.003)	0.29	1	0.16(0.001)
	Aoracle	*	*	*	0(0)	0.01(0.001)	0.85	1	0.13(0.001)
(200,25,4,4,0.2)	PM	0.44(0.016)	0.01(0)	0.555	0.18(0.009)	0.21(0.007)	0.12	0.6	1.73(0.054)
	SLTS	0(0)	0.01(0)	1	0(0)	0.32(0.006)	0.05	1	0.2(0.001)
	LL	*	*	*	0.02(0.002)	0.94(0.003)	0	0.85	2.89(0.005)
	REWLS	0.43(0.015)	0.01(0)	0.555	*	*	*	*	1.33(0.04)
	She	0.98(0.002)	0.02(0.002)	0	*	*	*	*	2.73(0.004)
	LASSO	*	*	*	0.06(0.003)	0.76(0.004)	0	0.53	2.76(0.005)
	ALASSO	*	*	*	0.04(0.002)	0.63(0.005)	0	0.69	26.97(6.71)
	Oracle	*	*	*	0(0)	0.11(0.003)	0.28	1	0.17(0.001)
	Aoracle	*	*	*	0(0)	0.02(0.001)	0.84	1	0.13(0.001)
(200,25,0,0,0)	PM	*	0(0)	*	0(0)	0.03(0.002)	0.61	1	0.17(0.001)
	SLTS	*	0.07(0.001)	*	0(0)	0.35(0.005)	0.02	1	0.28(0.002)
	LL	*	*	*	0(0)	0.1(0.005)	0.55	1	0.15(0.001)
	REWLS	*	0.02(0)	*	*	*	*	*	0.19(0.001)
	She	*	0(0)	*	*	*	*	*	0.18(0.001)
	LASSO	*	*	*	0(0)	0.1(0.003)	0.29	1	0.15(0.001)
	ALASSO	*	*	*	0(0)	0.01(0.001)	0.84	1	0.12(0.001)

Proof. Since $\hat{\beta}_{H_{opt}}$ can have at most n nonzero components by standard lasso literature, see Zou and Hastie (2005) for example, $\hat{\beta}_{H_{opt}} - \beta_0$ has at most $n + q_n$ nonzero components. Assume the index set of these nonzero components is S, and we have $S \subset \{1, 2, \dots, p_n\}$ and $|S| \leq n + q_n$. Thus

$$\max_{1\leq i\leq n} |X_i(\hat{\beta}_{H_{opt}} - \beta_0)| = \max_{1\leq i\leq n} |\sum_{j\in S} X_{i,j}(\hat{\beta}_{H_{opt},j} - \beta_{0,j})| \leq \max_{1\leq i\leq n} (\sum_{j\in S} X_{i,j}^2)^{1/2} ||\hat{\beta}_{H_{opt}} - \beta_0||_2.$$

By Lemma 3 and $||\beta_0||_2 < \infty$, one has $||\hat{\beta}_{H_{opt}} - \beta_0||_2 = O_P(1)$. Meanwhile, by Condition (B2), one has $\max_{1\leq i\leq n}(\sum_{j\in S} X_{i,j}^2)^{1/2} \leq \sum_{j\in S} X_{,j}^{\mathrm{T}} X_{,j} \leq (|S| M_1 n)^{1/2} \leq 2^{1/2} M_1 n$. Thus, one has $\max_{1\leq i\leq n} |X_i(\hat{\beta}_{H_{opt}} - \beta_0)| = O_P(n)$. Notice that $\widetilde{\gamma}_i = \epsilon_i - X_i(\widetilde{\beta} - \beta_0)$, we obtain $\kappa_n = O_P(n)$ as $\epsilon_i = O_P((\log n)^{1/2}) = o_P(n)$ by Condition (D1). □

Lemma 5. *Under Conditions (C2), (C5), (D1)-(D3), one has $\kappa_n = O_P(n)$. we have $\delta_n \equiv O_P(\pi_n)$.*

Proof. By Lemma 4, one has $\max_{1\leq i\leq n} |X_i(\hat{\beta}_{H_{opt}} - \beta_0)| = O_P(n)$. For $i \in G$, one has $\widetilde{\gamma}_i = -X_i(\widetilde{\beta} - \beta_0) + \epsilon_i$, by Condition (D1), one obtains $\max_{i\in G} |\widetilde{\gamma}_i| = O_P(n)$. For $i \in G^c$, one has $\widetilde{\gamma}_i = -X_i(\widetilde{\beta} - \beta_0) + \epsilon_i + \gamma_{i0}$. As $\pi_n n^{-1} \to \infty$ by Condition (C5), one obtains $\min_{i\in G^c} |\widetilde{\gamma}_i| \equiv O_P(\pi_n)$. □

3 Proof of Theorems

3.1 Proof of Theorem 1

The proof of Theorem 1 can be divided into two parts.

Part 1:

For any k dimensional vectors a and b, define $ab = (a_1 b_1, \dots, a_k b_k)^{\mathrm{T}}$ and $a/b = (a_1/b_1, \dots, a_k/b_k)^{\mathrm{T}}$.

Let $u = \theta - \theta_0 = (\theta(1)^{\mathrm{T}} - \theta_0(1)^{\mathrm{T}}, \theta(2)^{\mathrm{T}}, \theta(3)^{\mathrm{T}}) = (u(1)^{\mathrm{T}}, u(2)^{\mathrm{T}}, u(3)^{\mathrm{T}})^{\mathrm{T}}$ and $\hat{u} = \hat{\theta} - \theta_0 = (\hat{\theta}(1)^{\mathrm{T}} - \theta_0(1)^{\mathrm{T}}, \hat{\theta}(2)^{\mathrm{T}}, \hat{\theta}(3)^{\mathrm{T}}) = (\hat{u}(1)^{\mathrm{T}}, \hat{u}(2)^{\mathrm{T}}, \hat{u}(3)^{\mathrm{T}})^{\mathrm{T}}$. Also define $V_n(u) = \sum_{i=1}^n [(\epsilon_i - Au)^2 - \epsilon_i^2] + \lambda_n ||u(1) + \theta_0(1)||_1 + \lambda_n ||u(2)||_1 + n^{1/2}\mu_n(||u(3)||_1/|\widetilde{\gamma}|)$, and one has $\hat{u} = \mathrm{argmin}_u V_n(u)$.

The first summation in $V_n(u)$ can be simplified as $-2W(n^{1/2}u) + (n^{1/2}u)^{\mathrm{T}} C(n^{1/2}u)$, where $W = (W(1)^{\mathrm{T}}, W(2)^{\mathrm{T}}, W(3)^{\mathrm{T}})^{\mathrm{T}} = (n^{-1/2} X_{1,q}^{\mathrm{T}}\epsilon, n^{-1/2} X_{q+1,p}^{\mathrm{T}}\epsilon, \epsilon)$. Then by definition,

$$\mathrm{sign}(\hat{\theta}_j) = \mathrm{sign}(\theta_{j0}), \text{for all } j = 1, \dots, q \text{ implies that } \mathrm{sign}(\theta_0(1))\hat{u}(1) > -|\theta_0(1)|.$$

By the Karush-Kuhn-Tucker conditions and the uniqueness of LASSO solutions, if there exists a $\hat{u}$ the following equation and equalities hold,

$$\begin{aligned}
&C_{11}(n^{1/2}\hat{u}(1)) - W(1) = -(4n)^{-1/2}\lambda_n \mathrm{sign}(\theta_0(1)),\\
&|\hat{u}(1)| < |\theta_0(1)|,\\
&-(4n)^{-1/2}\lambda_n 1 \leq C_{21}(n^{1/2}\hat{u}(1)) - W(2) \leq (4n)^{-1/2}\lambda_n 1,\\
&-(\mu_n/2)(1/|\widetilde{\gamma}|) \leq C_{31}(n^{1/2}\hat{u}(1)) - W(3) \leq (\mu_n/2)(1/|\widetilde{\gamma}|),
\end{aligned}$$

then $\mathrm{sign}(\hat{\theta}(1)) = \mathrm{sign}(\theta_0(1))$, $\mathrm{sign}(\hat{\theta}(2)) = \mathrm{sign}(\theta_0(2)) = 0$, $\mathrm{sign}(\hat{\theta}(3)) = \mathrm{sign}(\theta_0(3)) = 0$. Substitute $\hat{u}(1)$ and bound the absolute values, the existence of such $\hat{u}$ is implied by

$$\begin{aligned}
A_n &: |C_{11}^{-1}W(1)| < n^{1/2}\{|\theta_0(1)| - (2n)^{-1}\lambda_n |C_{11}^{-1}\mathrm{sign}(\theta_0(1))|\}\\
B_n &: |C_{21}C_{11}^{-1}W(1) - W(2)| \leq (4n)^{-1/2}\lambda_n(1 - |C_{21}C_{11}^{-1}\mathrm{sign}(\theta_0(1))|)\\
C_n &: |C_{31}C_{11}^{-1}W(1) - W(3)| \leq (\mu_n/2)(1/|\widetilde{\gamma}|) - (4n)^{-1/2}\lambda_n |C_{31}C_{11}^{-1}\mathrm{sign}(\theta_0(1))|.
\end{aligned}$$

As a result, $pr(\hat{\theta} =_s \theta_0) \geq pr(A_n \cap B_n \cap C_n)$.

After we obtain the lower bound of $pr(\hat{\theta} =_s \theta_0)$, we will show that $pr(A_n \cap B_n \cap C_n) \to 1$ when $n \to \infty$. Notice that

$$\begin{aligned} & 1 - pr(A_n \cap B_n \cap C_n) \\ \leq & \sum_{j=1}^{q} pr(|z_j| \geq n^{1/2}\{|\theta_0(1)| - (2n)^{-1}\lambda_n a_j\}) + \sum_{j=1}^{p-q} pr(|\zeta_j| \geq (4n)^{-1/2}\lambda_n b_j) \\ & + \sum_{j=1}^{n} pr(|\xi_j| \geq \mu_n/(2|\widetilde{\gamma}_j|) - (4n)^{-1/2}\lambda_n \phi_j), \end{aligned}$$

where $Z = (z_1, \ldots, z_q)^{\mathrm{T}} = C_{11}^{-1}W(1)$, $\zeta = (\zeta_1, \ldots, \zeta_{p-q})^{\mathrm{T}} = C_{21}C_{11}^{-1}W(1) - W(2)$, $\xi = (\xi_1, \ldots, \xi_n)^{\mathrm{T}} = C_{31}C_{11}^{-1}W(1) - W(3)$, $a = (a_1, \ldots, a_q)^{\mathrm{T}} = |C_{11}^{-1}\mathrm{sign}(\theta_0(1))|$, $b = (b_1, \ldots, b_{p-q})^{\mathrm{T}} = 1 - |C_{21}C_{11}^{-1}\mathrm{sign}(\theta_0(1))|$ and $\phi = (\phi_1, \ldots, \phi_n)^{\mathrm{T}} = |C_{31}C_{11}^{-1}\mathrm{sign}(\theta_0(1))|$.

Now if we write $Z = H_A^{\mathrm{T}}\epsilon$ where $H_A^{\mathrm{T}} = (h_1^a, \ldots, h_q^a)^{\mathrm{T}} = C_{11}^{-1}(n^{-1/2}X_{1,q}^{\mathrm{T}})$, then $H_A^{\mathrm{T}}H_A = C_{11}^{-1}$. Under the Condition (B3), one has $z_j = (h_j^a)^{\mathrm{T}}\epsilon$ with $||h_j^a||^2 \leq M_2^{-1}$ for any $j = 1, \ldots, q$. Similarly if we write $\zeta = H_B^{\mathrm{T}}\epsilon$ where $H_B^{\mathrm{T}} = (h_1^b, \ldots, h_{p-q}^b)^{\mathrm{T}} = C_{21}C_{11}^{-1}(n^{-1/2}X_{1,q}^{\mathrm{T}}) - n^{-1/2}X_{q+1,p}^{\mathrm{T}}$, then

$$H_B^{\mathrm{T}}H_B = n^{-1}X_{q+1,p}^{\mathrm{T}}\{I_n - X_{1,q}(X_{1,q}^{\mathrm{T}}X_{1,q})^{-1}X_{1,q}^{\mathrm{T}}\}X_{q+1,p}.$$

Since $I_n - X_{1,q}(X_{1,q}^{\mathrm{T}}X_{1,q})^{-1}X_{1,q}^{\mathrm{T}}$ has eigenvalues between 0 and 1, under Condition (B2), one obtains $\zeta_j = (h_j^b)^{\mathrm{T}}\epsilon$ with $||h_j^b||^2 \leq M_1$ for any $1 \leq j \leq p-q$. Similarly, we can write $\xi = H_C^{\mathrm{T}}\epsilon$ where $H_C^{\mathrm{T}} = (h_1^c, \ldots, h_n^c)^{\mathrm{T}} = C_{31}C_{11}^{-1}(n^{-1/2}X_{1,q}^{\mathrm{T}}) - I_n$, then $H_C^{\mathrm{T}}H_C = I_n - X_{1,q}(X_{1,q}^{\mathrm{T}}X_{1,q})^{-1}X_{1,q}^{\mathrm{T}}$, which indicates $\xi_j = (h_j^c)^{\mathrm{T}}\epsilon$ with $||h_j^c||^2 \leq 1$ for any $1 \leq j \leq n$.

Part 2:

Thus, under Condition (A), one obtains $E(x_j^{2k}) < \infty$, $E(\zeta_j^{2k}) < \infty$ and $E(\xi_j^{2k}) < \infty$ because for any given constant n-dimensional vector α, $E(\alpha^{\mathrm{T}}\epsilon^{2k}) \leq (2k-1)!!||\alpha||^2 E(\epsilon_i^{2k})$. For random variables Z with bounded $2k$'th moments, the tail probability is bounded by $pr(Z \geq t) = O(t^{-2k})$.

Also notice that

$$n^{-1}\lambda_n a_j \leq n^{-1}\lambda_n||C_{11}^{-1}\mathrm{sign}(\theta_0(1))||_2 \leq M_2^{-1}n^{-1}\lambda_n q^{1/2},$$

under Condition (B4), for $\lambda_n = o(n^{(d+1)/2})$, we obtain

$$\sum_{j=1}^{q} pr(|z_j| \geq n^{1/2}\{|\theta_0(1)| - (2n)^{-1}\lambda_n a_j\}) = O(n^{-kd}) = o(1),$$

Further, under Condition (B1), one has

$$\sum_{j=1}^{p-q} pr(|\zeta_j| \geq (4n)^{-1/2}\lambda_n b_j) = (p-q)O(n^k\lambda_n^{-2k}) = O(n^k\lambda_n^{-2k}) = o(1).$$

Under Conditions (B2) and (C3), one has $|(4n)^{-1/2}\lambda_n\phi_j| \leq q^{1/2}M_1^{1/2}M_2^{-1}(4n)^{-1/2}\lambda_n q^{1/2} = (2M_2)^{-1}M_1^{1/2}n^{-1/2}\lambda_n$. Thus when $\lambda_n n^{-1/2} = o(\mu_n/\kappa_n)$, which is indicated by $\mu_n n^{-1/(2k)-d/2} \to \infty$, by Lemma 1, we have

$$\sum_{j=1}^{n} pr(|\xi_j| \geq \mu_n/(2|\widetilde{\gamma}_j|) - (4n)^{-1/2}\lambda_n\phi_j) = O(n\kappa_n^{2k}\mu_n^{-2k}) = o(1).$$

This completes the proof of Theorem 1.

3.2 Proof of Theorem 2

Suppose we contaminate $m \leq \min\{n-h, \lfloor (n-p)/2 \rfloor - 1\}$ observations, and assume $\widetilde{y}$ and $\widetilde{X}$ are the contaminated data. Define $\widetilde{\beta}$ as the solution using least trimmed squares when subset size is set as h and $\widetilde{\gamma}_i = \widetilde{y}_i - \widetilde{X}_i\widetilde{\beta}$. Since the least trimmed squares with truncation number h have a breakdown point $\min\{(n-h+1)/n, \lfloor (n-p)/2 \rfloor /n\}$ (Alfons et al., 2013), we have $||\widetilde{\beta}||_2 \leq M$ for some $M > 0$. As L_1 norm and the Euclidean norm are topologically equivalent, for any p dimensional vector β, we have $c_1||\beta||_2 \leq ||\beta||_1 \leq c_2||\beta||_2$. Consequently, we obtain $Q_n(\widetilde{\beta}, \widetilde{\gamma}) = \lambda_n||\widetilde{\beta}||_1 + n\mu_n \leq c_2\lambda_n M + n\mu_n = M^*$.

Recall that we solve the following minimization problem

$$\min_{\beta,\gamma} Q_n(\beta,\gamma) = \min_{\beta,\gamma} ||y - X\beta - \gamma||_2^2 + \lambda_n \sum_{j=1}^{p} |\beta_j| + \mu_n \sum_{i=1}^{n} |\gamma_i|/|\widetilde{\gamma}_i|,$$

and suppose the minimizer of $Q_n(\beta,\gamma)$ is $\hat{\beta}$ and $\hat{\gamma}$.

For any $||\beta||_2 \geq (M^*+1)/(\lambda_n c_1)$, we have $Q_n(\beta,\gamma) \geq \lambda_n||\beta||_1 \geq c_1\lambda_n||\beta||_2 \geq M^*+1 > Q_n(\widetilde{\beta},\widetilde{\gamma}) \geq Q_n(\hat{\beta},\hat{\gamma})$. As a result, $||\hat{\beta}||_2 \leq (M^*+1)/(\lambda_n c_1)$, which indicates the breakdown point is at least $\min\{(n-h+1)/n, \lfloor (n-p)/2 \rfloor /n\}$. This completes the proof of Theorem 2.

3.3 Proof of Theorem 3

The proof of Theorem 3 can be divided into two parts.

Part 1:

Let $v = \eta - \eta_0 = (\eta(1)^{\mathrm{T}} - \eta_0(1)^{\mathrm{T}}, \eta(2)^{\mathrm{T}} - \eta_0(2)^{\mathrm{T}}, \eta(3)^{\mathrm{T}}) = (v(1)^{\mathrm{T}}, v(2)^{\mathrm{T}}, v(3)^{\mathrm{T}})^{\mathrm{T}}$ and $\hat{v} = \hat{\eta} - \eta_0 = (\hat{\eta}(1)^{\mathrm{T}} - \eta_0(1)^{\mathrm{T}}, \hat{\eta}(2)^{\mathrm{T}} - \eta_0(2)^{\mathrm{T}}, \hat{\eta}(3)^{\mathrm{T}}) = (\hat{v}(1)^{\mathrm{T}}, \hat{v}(2)^{\mathrm{T}}, \hat{v}(3)^{\mathrm{T}})^{\mathrm{T}}$.

We define $U_n(v) = \sum_{i=1}^n [(\epsilon_i - Bv)^2 - \epsilon_i^2] + n^{1/2}\mu_n(||v(1)+\eta_0(1)||_1/|\widetilde{\gamma}(1)|) + \lambda_n||v(2)+\beta_0)||_1 + n^{1/2}\mu_n(||v(3)||_1/|\widetilde{\gamma}(2)|)$. One has $\hat{v} = \mathrm{argmin}_v U_n(v)$.

The first summation in $U_n(v)$ can be simplified as $-2W^*(n^{1/2}v) + (n^{1/2}v)^{\mathrm{T}} D(n^{1/2}v)$, where $W^* = (W^*(1)^{\mathrm{T}}, W^*(2)^{\mathrm{T}}, W^*(3)^{\mathrm{T}})^{\mathrm{T}} = (n^{-1/2}B_1^{\mathrm{T}}\epsilon, n^{-1/2}B_2^{\mathrm{T}}\epsilon, n^{-1/2}B_3^{\mathrm{T}}\epsilon)$. Then by definition,

$$\mathrm{sign}(\hat{\eta}_j) = \mathrm{sign}(\eta_{j0}), \text{for all } j = 1,\ldots,s_n \quad \text{implies that} \quad \mathrm{sign}(\eta_0(1))\hat{v}(1) > -|\eta_0(1)|.$$

By the Karush-Kuhn-Tucker conditions and the uniqueness of LASSO solutions, if there exists a $\hat{v}$ the following equation and equalities hold,

$$\begin{aligned}
&D_{11}(n^{1/2}\hat{v}(1)) + D_{12}(n^{1/2}\hat{v}(2)) - W^*(1) = \varpi, \\
&|\hat{v}(1)| < |\eta_0(1)|, \\
&-(4n)^{-1/2}\lambda_n 1 \leq D_{21}(n^{1/2}\hat{v}(1)) + D_{22}(n^{1/2}\hat{v}(2)) - W^*(2) \leq (4n)^{-1/2}\lambda_n 1, \\
&-(\mu_n/2)(1/|\widetilde{\gamma}(2)|) \leq D_{31}(n^{1/2}\hat{v}(1)) + D_{32}(n^{1/2}\hat{v}(2)) - W^*(3) \leq (\mu_n/2)(1/|\widetilde{\gamma}(2)|),
\end{aligned}$$

where $\varpi = (-\mu_n/2)\{|\mathrm{sign}(\psi_0(1))|/|\widetilde{\gamma}(1)|\}^{\mathrm{T}}$, then $\mathrm{sign}(\hat{\eta}(1)) = \mathrm{sign}(\eta_0(1))$ and $\mathrm{sign}(\hat{\eta}(3)) = \mathrm{sign}(\eta_0(3)) = 0$. As $D_{11} = I_{s_n}$ and $D_{31} = 0$, substitute $\hat{v}(1)$ and bound the absolute values, the existence of such $\hat{v}$ satisfying $\hat{v}(2) = 0$ is implied by

$$
\begin{aligned}
A_n^* &: |W^*(1)| < n^{1/2}|\eta_0(1)| - |\varpi| \\
B_n^* &: |D_{21}W^*(1) - W^*(2)| \le (4n)^{-1/2}\lambda_n 1 - |D_{21}\varpi| \\
C_n^* &: |W^*(3)| \le (\mu_n/2)(1/|\widetilde{\gamma}(2)|).
\end{aligned}
$$

As a result, $pr(\hat{\gamma} =_s \gamma_0) \ge pr(\hat{\gamma} =_s \gamma_0, \hat{v}(2) = 0) \ge pr(A_n^* \cap B_n^* \cap C_n^*)$.

After we obtain the lower bound of $pr(\hat{\gamma} =_s \gamma_0)$, we will show that $pr(A_n^* \cap B_n^* \cap C_n^*) \to 1$ when $n \to \infty$. Notice that

$$
\begin{aligned}
& 1 - pr(A_n^* \cap B_n^* \cap C_n^*) \\
\le & \sum_{j=1}^{s_n} pr(|z_j^*| \ge n^{1/2}\{|\eta_0(1)|\}_j - |\varpi_j|^*) + \sum_{j=1}^{p} pr(|\zeta_j^*| \ge b_j^*) \\
& + \sum_{j=1}^{n-s_n} pr(|\xi_j^*| \ge \phi_j^*),
\end{aligned}
$$

where $z^* = (z_1^*, \ldots, z_{s_n}^*)^{\mathrm{T}} = W^*(1)$, $\zeta^* = (\zeta_1^*, \ldots, \zeta_p^*)^{\mathrm{T}} = D_{21}W^*(1) - W^*(2)$, $b^* = (b_1^*, \ldots, b_p^*)^{\mathrm{T}} = (4n)^{-1/2}\lambda_n 1 - |D_{21}\varpi|$, $\xi^* = (\xi_1^*, \ldots, \xi_{n-s_n}^*)^{\mathrm{T}} = W^*(3)$ and $\phi^* = (\phi_1^*, \ldots, \phi_{n-s_n}^*)^{\mathrm{T}} = (\mu_n/2)(1/|\widetilde{\gamma}(2)|)$.

For z_j^*, we can write $z_j^* = \epsilon_j$ for $1 \le j \le s_n$ and for ξ_j^*, we can write $\xi_j^* = \epsilon_{j+s_n}$ for any $1 \le j \le n - s_n$. Further, we can write $\zeta = H_E^{\mathrm{T}}\epsilon$ where $H_E^{\mathrm{T}} = (h_1^e, \ldots, h_{p-q}^e)^{\mathrm{T}} = D_{21}(n^{-1/2}B_1^{\mathrm{T}}) - n^{-1/2}B_2^{\mathrm{T}}$, then

$$H_E^{\mathrm{T}} H_E = n^{-1}B_2^{\mathrm{T}}\{I_n - B_1(B_1^{\mathrm{T}}B_1)^{-1}B_1^{\mathrm{T}}\}B_2.$$

Since $I_n - B_1(B_1^{\mathrm{T}}B_1)^{-1}B_1^{\mathrm{T}}$ has eigenvalues between 0 and 1, under Condition (B2), one obtains $\zeta_j^* = (h_j^e)^{\mathrm{T}}\epsilon$ with $||h_j^e||^2 \le M_1$ for any $1 \le j \le p$.

Part 2:

Thus, under Condition (A), one obtains $E((z_j^*)^{2k}) < \infty$, $E((\zeta_j^*)^{2k}) < \infty$ and $E((\xi_j^*)^{2k}) < \infty$ because for any given constant n-dimensional vector α, $E(\alpha^{\mathrm{T}}\epsilon^{2k}) \le (2k-1)!!||\alpha||^2 E(\epsilon_i^{2k})$. For random variables Z with bounded $2k$'th moments, the tail probability is bounded by $pr(Z \ge t) = O(t^{-2k})$.

By Lemma 2, for any $1 \le j \le s_n$,

$$|\varpi_j^*| \le 2^{-1}\mu_n\pi_n^{-1}.$$

Noticing that $n^{1/2}\min_{j=1,\ldots,s_n}|\eta_{0j}(1)| = O_P(\pi_n)$, by Conditions (C1), combining with the fact that $s_n = O(n)$, $\mu_n\pi_n^{-1} = o(\pi_n)$, we obtain

$$\sum_{j=1}^{s_n} pr(|z_j^*| \ge n^{1/2}\{|\eta_0(1)|\}_j - |\varpi_j|^*) = O(n\pi_n^{-2k}) = o(1).$$

Further, as

$$D_{21}\varpi = D_{21}(\mu_n/2)|\mathrm{sign}(\psi_0(1))|/|\widetilde{\gamma}(1)|$$

Thus, under Conditions (B2), combining with Lemma 2, the second term of $|b_j^*|$ can be bounded above by $M_1^{1/2}s_n^{1/2}(\mu_n/2)/\delta_n =$

$O_P(n^{1/2}\mu_n/\pi_n)$. Combining with Condition (C3), when $n^{1/2}\mu_n\pi_n^{-1} = o(n^{-1/2}\lambda_n)$ and $\lambda_n n^{-1/2} \to \infty$, we obtain

$$\sum_{j=1}^{p} pr(|\zeta_j^*| \geq b_j^*) \leq pO(n^k\lambda_n^{-2k}) = o(1).$$

Moreover, recall that $D_{31} = 0$ when $n\kappa_n^{2k}\mu_n^{-2k} = o(1)$, we have

$$\sum_{j=1}^{n-s_n} pr(|\xi_j^*| \geq \phi_j^*) = O(n\kappa_n^{2k}\mu_n^{-2k}) = o(1).$$

As we require $\pi_n n^{-1/2k} \to \infty$, $\mu_n\pi_n^{-1} = o(\pi_n)$, $n\kappa_n^{2k}\mu_n^{-2k} = o(1)$ and $n^{1/2}\mu_n\pi_n^{-1} = o(n^{-1/2}\lambda_n)$, combining with the fact that $\lambda_n n^{-1/2} \to \infty$, to guarantee there exist a μ_n satisfies these conditions, we reed to require $\pi_n n^{-1/2k} \to \infty$, $\mu_n = o(\pi_n^2)$, $\mu_n^k n^{-1} \to \infty$ and $\lambda_n n^{-1}\mu_n^{-1}\pi_n \to \infty$. Thus, we have $pr(\hat{\gamma} =_s \gamma_0) \to 1$.

This completes the proof of Theorem 3.

4 Proof of Corollaries

4.1 Proof of Corollary 1

We will follow the same framework as the proof of Theorem 1. Part 1 is the same as the proof of Theorem 1 under Conditions (B1)-(B4) except that we need to change p into p_n and q and q_n. We change part 2 of the proof of Theorem 1 as follows:

Under Condition (D1), the tail probabilities of z_j, ζ_j and ξ_j are bounded by $pr(Z \geq t) = o(e^{-Ct^2})$ for some constant C.

Also notice that

$$n^{-1}\lambda_n a_j \leq n^{-1}\lambda_n ||C_{11}^{-1}\text{sign}(\theta_0(1))||_2 \leq M_2^{-1}n^{-1}\lambda_n q_n^{1/2},$$

By Condition (B5), for $\lambda_n \equiv O(n^{(c_3+1)/2})$, with $0 < c_3 < d - c_1$, one has $n^{-1}\lambda_n a_j = o(n^{(d-1)/2})$. Thus, we obtain

$$\sum_{j=1}^{q_n} pr(|z_j| \geq n^{1/2}\{|\theta_0(1)| - (2n)^{-1}\lambda_n a_j\}) = q_n o(e^{-Cn^d}) = o(1),$$

Further, by Condition (D1), and $p_n = O(e^{n^{c_2}})$ with $0 < c_2 < c_3$, one has

$$\sum_{j=1}^{p-q} pr(|\zeta_j| \geq (4n)^{-1/2}\lambda_n b_j) = o(p_n e^{-Cn^{c_3}}) = o(1).$$

Under Conditions (B2) and (C3), one has $|(4n)^{-1/2}\lambda_n\phi_j| \leq q_n^{1/2}M_1^{1/2}M_2^{-1}(4n)^{-1/2}\lambda_n q_n^{1/2} = (2M_2)^{-1}M_1^{1/2}n^{-1/2}\lambda_n n^{c_1}$. Thus when $\lambda_n n^{-1/2}n^{c_1} = o(\mu_n/\kappa_n)$, which is indicated by $\mu_n n^{-1-c_1-c_3/2} \to \infty$, by Lemma 4, we have

$$\sum_{j=1}^{n} pr(|\xi_j| \geq \mu_n/(2|\widetilde{\gamma}_j|) - (4n)^{-1/2}\lambda_n\phi_j) = o(1).$$

This completes the proof of Corollary 1.

4.2 Proof of Corollary 2

If we replace the initial least trimmed squares estimator and the residual by the corresponding sparse least trimmed squares estimator and residual, we can prove this corollary similarly as the proof of Theorem 2 because the sparse least trimmed squares estimator has a breakdown point of $(n-h+1)/n$ (Alfons, Croux, and Gelper, 2013).

4.3 Proof of Corollary 3

We will follow the same framework as the proof of Theorem 3. Part 1 is the same as the proof of Theorem 3 except that we need to change p into p_n. We change part 2 of the proof of Theorem 3 as follows:

Under Condition (D1), the tail probability of z_j^*, ζ_j^* and ξ_j^* are bounded by $pr(Z \geq t) = o(e^{-Ct^2})$ for some constant C.

By Lemma 2, for any $1 \leq j \leq s_n$,

$$|\varpi_j^*| \leq 2^{-1}\mu_n \pi_n^{-1}.$$

Noticing that $n^{1/2}\min_{j=1,\ldots,s_n}|\eta_{0j}(1)| = O_P(\pi_n)$, by Conditions (C1) and $\pi_n(\log n)^{-1/2} \to \infty$, combining with the fact that $s_n = O(n)$, $\mu_n\pi_n^{-1} = o(\pi_n)$, we obtain

$$\sum_{j=1}^{s_n} pr(|z_j^*| \geq n^{1/2}\{|\eta_0(1)|\}_j - |\varpi_j|^*) = O(ne^{-C\pi_n^2}) = o(1).$$

Further, as

$$D_{21}\varpi = D_{21}(\mu_n/2)|\mathrm{sign}(\psi_0(1))|/|\widetilde{\gamma}(1)|$$

Thus, under Conditions (B2), combining with Lemma 2, the second term of $|b_j^*|$ can be bounded above by $M_1^{1/2}s_n^{1/2}(\mu_n/2)/\delta_n = O_P(n^{1/2}\mu_n/\pi_n)$. Combining with Condition (C3) and $p_n = o(e^{n^{c_2}})$ with $0 < c_2 < d_1 - c_1$ $(c_1 > 0)$, when $n^{1/2}\mu_n\pi_n^{-1} = o(n^{-1/2}\lambda_n)$ and $\lambda_n n^{-1/2-d_1/2} \to \infty$, we obtain

$$\sum_{j=1}^{p_n} pr(|\zeta_j^*| \geq b_j^*) \leq p_n O(e^{-Cn^{d_1}}) = o(1).$$

Moreover, recall that $D_{31} = 0$ when $\mu_n\kappa_n^{-1}(\log n)^{-1/2} \to \infty$, we have

$$\sum_{j=1}^{n-s_n} pr(|\xi_j^*| \geq \phi_j^*) = O(ne^{-C\mu_n^2\kappa_n^{-2}}) = o(1).$$

As we require $\pi_n(\log n)^{-1/2} \to \infty$, $\mu_n\pi_n^{-1} = o(\pi_n)$, $\mu_n\kappa_n^{-1}(\log n)^{-1/2} \to \infty$ and $n^{1/2}\mu_n\pi_n^{-1} = o(n^{-1/2}\lambda_n)$, to guarantee there exist a μ_n satisfies these conditions, we need to require $\pi_n n^{-1/2}(\log n)^{-1/4} \to \infty$, $\mu_n = o(\pi_n^2)$, $\mu_n n^{-1}(\log n)^{-1/2} \to \infty$ and $\lambda_n n^{-1}\mu_n^{-1}\pi_n \to \infty$. Thus, we have $pr(\hat{\gamma} =_s \gamma_0) \to 1$.

This completes the proof of Corollary 3.

Bibliography

Alfons, A., Croux, C., and Gelper, S. (2013), "Sparse least trimmed squares regression for analyzing high-dimensional large data sets," *The Annals of Applied Statistics*, 7, 226–248.

Rousseeuw, P. J. and Leroy, A. M. (1987), *Robust regression and outlier detection*, New York, NY, USA: John Wiley & Sons, Inc.

Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society. Series B.*, 67, 301–320.