Calibrated zero-norm regularized LS estimator for high-dimensional error-in-variables regression

Ting Tao, Shaohua Pan and Shujun Bi

School of Mathematics, South China University of Technology, Guangzhou.

Supplementary Material

This Supplementary Material includes the proof of Theorems, the additional theoretical results, the implementation of GEP-MSCRA, and the ADMM Algorithm for CoCoLasso Datta and Zou (2017).

S1. Proof of Theorems

In this part, we write $\Delta\beta^k = \beta^k - \beta^*$ and $v^k = e - w^k$ for $k = 1, 2, \ldots$.

S1.1 The proof of Theorem 2

To get the conclusion of Theorem 2, we need the following two lemmas.

Lemma 1. *For any $\beta \in \mathbb{R}^p$, it holds that $\frac{1}{2n}\|\widetilde{Z}\beta\|^2 \geq \frac{1}{2n}\|X\beta\|^2 + \frac{1}{2}\beta^{\mathbb{T}} D\beta$.*

Proof. From $\widetilde{\Sigma} = \frac{1}{n}\widetilde{Z}^{\mathbb{T}}\widetilde{Z}$ and $\widetilde{\Sigma} = \widehat{\epsilon}I + \Pi_{\mathbb{S}_+^p}(\widehat{\Sigma} - \widehat{\epsilon}I)$, for any $\beta \in \mathbb{R}^p$, we get

$$\begin{aligned}\frac{1}{2n}\big\|\widetilde{Z}\beta\big\|^2 &= \frac{1}{2n}\|X\beta\|^2 + \frac{1}{2}\beta^{\mathbb{T}}(\widetilde{\Sigma} - \widehat{\Sigma})\beta + \frac{1}{2}\beta^{\mathbb{T}}(\widehat{\Sigma} - \Sigma)\beta \\ &= \frac{1}{2n}\|X\beta\|^2 + \frac{1}{2}\beta^{\mathbb{T}}\Pi_{\mathbb{S}_+^p}(\widehat{\epsilon}I - \widehat{\Sigma})\beta + \frac{1}{2}\beta^{\mathbb{T}}D\beta \\ &\geq \frac{1}{2n}\|X\beta\|^2 + \frac{1}{2}\beta^{\mathbb{T}}D\beta\end{aligned}$$

where the inequality is by the positive semidefiniteness of $\Pi_{\mathbb{S}_+^p}(\widehat{\epsilon}I - \widehat{\Sigma})$. □

Lemma 2. *Suppose that for some $k \geq 1$ there exists an index set $S^{k-1} \supseteq S^*$ such that $\max_{i\in(S^{k-1})^c} w_i^{k-1} \leq \frac{1}{2}$. Then, whenever $\lambda \geq 8\|\widetilde{\varepsilon}\|_\infty$, it holds that*

$$\begin{aligned}\|\Delta\beta^k_{(S^{k-1})^c}\|_1 &\leq 3\|\Delta\beta^k_{S^{k-1}}\|_1, \\ \frac{1}{2n}\big\|\widetilde{Z}\Delta\beta^k\big\|^2 &\leq \Big(\|\widetilde{\varepsilon}_{S^{k-1}}\| + \lambda\sqrt{\textstyle\sum_{i\in S^*}(v_i^{k-1})^2}\Big)\big\|\Delta\beta^k_{S^{k-1}}\big\|.\end{aligned}$$

Proof. From the optimality of β^k and the feasibility of β^* to (3.1), we have

$$\frac{1}{2n}\big\|\widetilde{Z}\beta^k - \widetilde{y}\big\|^2 + \lambda\sum_{i=1}^p v_i^{k-1}\big|\beta_i^k\big| \leq \frac{1}{2n}\big\|\widetilde{Z}\beta^* - \widetilde{y}\big\|^2 + \lambda\sum_{i=1}^p v_i^{k-1}\big|\beta_i^*\big|$$

which, by $\Delta\beta^k = \beta^k - \beta^*$ and $\widetilde{\varepsilon} = \frac{1}{n}\widetilde{Z}^{\mathbb{T}}(\widetilde{y} - \widetilde{Z}\beta^*)$, can be rearranged as

$$\begin{aligned}
\frac{1}{2n}\big\|\widetilde{Z}\Delta\beta^k\big\|^2 &\le \langle\widetilde{\varepsilon}, \Delta\beta^k\rangle + \lambda\sum_{i\in S^*} v_i^{k-1}\left(|\beta_i^*| - |\beta_i^k|\right) - \lambda\sum_{i\in(S^*)^c} v_i^{k-1}|\beta_i^k| \\
&\le \langle\widetilde{\varepsilon}, \Delta\beta^k\rangle + \lambda\sum_{i\in S^*} v_i^{k-1}|\Delta\beta_i^k| - \lambda\sum_{i\in(S^{k-1})^c} v_i^{k-1}|\Delta\beta_i^k| \qquad \text{(S1.1)} \\
&\le \sum_{i\in S^{k-1}} |\widetilde{\varepsilon}_i||\Delta\beta_i^k| + \sum_{i\in(S^{k-1})^c} |\widetilde{\varepsilon}_i||\Delta\beta_i^k| \\
&\quad + \lambda\sum_{i\in S^*} v_i^{k-1}|\Delta\beta_i^k| - \lambda\sum_{i\in(S^{k-1})^c} v_i^{k-1}|\Delta\beta_i^k| \\
&\le (\lambda + \|\widetilde{\varepsilon}\|_\infty)\big\|\Delta\beta^k_{S^{k-1}}\big\|_1 + \big(\|\widetilde{\varepsilon}\|_\infty - \lambda/2\big)\big\|\Delta\beta^k_{(S^{k-1})^c}\big\|_1
\end{aligned}$$

where the second inequality is using $S^{k-1} \supseteq S^*$, and the last one is due to $v_i^k \le 1$ for $i \in S^*$ and $\min_{i\notin S^{k-1}} v_i^{k-1} \ge \frac{1}{2}$. From $\lambda \ge 8\|\widetilde{\varepsilon}\|_\infty$ and $\frac{1}{2n}\big\|\widetilde{Z}\Delta\beta^k\big\|^2 \ge 0$, we obtain the first inequality. For the second inequality, by using inequality (S1.1) and $\min_{i\notin S^{k-1}} v_i^{k-1} \ge \frac{1}{2}$, it follows that

$$\begin{aligned}
\frac{1}{2n}\big\|\widetilde{Z}\Delta\beta^k\big\|^2 &\le \sum_{i=1}^{p} |\widetilde{\varepsilon}_i||\Delta\beta_i^k| - \frac{1}{2}\lambda\sum_{i\in(S^{k-1})^c} |\Delta\beta_i^k| + \lambda\sum_{i\in S^*} v_i^{k-1}|\Delta\beta_i^k| \\
&\le \sum_{i\in S^{k-1}} |\widetilde{\varepsilon}_i||\Delta\beta_i^k| + \lambda\textstyle\sum_{i\in S^*} v_i^{k-1}|\Delta\beta_i^k| \\
&\le \|\widetilde{\varepsilon}_{S^{k-1}}\|\big\|\Delta\beta^k_{S^{k-1}}\big\| + \lambda\sqrt{\textstyle\sum_{i\in S^*}(v_i^{k-1})^2}\,\big\|\Delta\beta^k_{S^{k-1}}\big\|,
\end{aligned}$$

where the second inequality is due to $\lambda \ge 8\|\widetilde{\varepsilon}\|_\infty$. □

The proof of Theorem 2: Define $S^{k-1} := S^* \cup \{i \notin S^* : w_i^{k-1} > \frac{1}{2}\}$ for each $k \in \mathbb{N}$. We first argue that if $|S^{l-1}| \le 1.5s$ for some $l \in \mathbb{N}$, and consequently the following inequality holds

$$\left\|\Delta\beta^l\right\| \le \frac{2(\|\widetilde{\varepsilon}\|_\infty\sqrt{1.5s}+\lambda\sqrt{s})}{\kappa-24s\|D\|_{\max}} \le \frac{(2+\sqrt{6}/8)\lambda\sqrt{s}}{\kappa-24s\|D\|_{\max}}. \tag{S1.2}$$

Since $S^{l-1}\supseteq S^*$ with $|S^{l-1}|\le 1.5s$ and $\lambda\ge 8\|\widetilde{\varepsilon}\|_\infty$, from Lemma 2 we have

$$\frac{1}{2n}\left\|\widetilde{Z}\Delta\beta^l\right\|^2 \le \left[\|\widetilde{\varepsilon}_{S^{l-1}}\|+\lambda\sqrt{\textstyle\sum_{i\in S^*}(v_i^{l-1})^2}\right]\left\|\Delta\beta^l_{S^{l-1}}\right\|,$$

$$\begin{aligned}\left|(\Delta\beta^l)^{\mathbb{T}}D\Delta\beta^l\right| &\le \|D\|_{\max}\|\Delta\beta^l\|_1^2 = \|D\|_{\max}\big(\|\Delta\beta^l_{S^{l-1}}\|_1+\|\Delta\beta^l_{(S^{l-1})^c}\|_1\big)^2\\ &\le 16\|D\|_{\max}\|\Delta\beta^l_{S^{l-1}}\|_1^2 \le 16|S^{l-1}|\|D\|_{\max}\|\Delta\beta^l_{S^{l-1}}\|^2\\ &\le 24s\|D\|_{\max}\|\Delta\beta^l_{S^{l-1}}\|^2. \end{aligned} \tag{S1.3}$$

By combining the last two inequalities with Lemma 1, it then follows that

$$\frac{1}{2n}\left\|X\Delta\beta^l\right\|^2 - 12s\|D\|_{\max}\|\Delta\beta^l\|^2 \le \left[\|\widetilde{\varepsilon}_{S^{l-1}}\|+\lambda\sqrt{\textstyle\sum_{i\in S^*}(v_i^{l-1})^2}\right]\left\|\Delta\beta^l_{S^{l-1}}\right\|.$$

Notice that $\Delta\beta^l\in\mathcal{C}(S^*)$ since $S^{l-1}\supseteq S^*$ with $|S^{l-1}|\le 1.5s$. Together with the κ-REC of Σ on $\mathcal{C}(S^*)$, it is immediate to obtain

$$\frac{1}{2}\big(\kappa-24s\|D\|_{\max}\big)\|\Delta\beta^l\|^2 \le \left[\|\widetilde{\varepsilon}_{S^{l-1}}\|+\lambda\sqrt{\textstyle\sum_{i\in S^*}(v_i^{l-1})^2}\right]\left\|\Delta\beta^l_{S^{l-1}}\right\| \tag{S1.4}$$

$$\le \left[\|\widetilde{\varepsilon}\|_\infty\sqrt{|S^{l-1}|}+\lambda\sqrt{s}\right]\left\|\Delta\beta^l\right\|$$

$$\le \left[\|\widetilde{\varepsilon}\|_\infty\sqrt{1.5s}+\lambda\sqrt{s}\right]\left\|\Delta\beta^l\right\|.$$

This, by $\|\widetilde{\varepsilon}\|_\infty\le\frac{1}{8}\lambda$, implies that the inequality (S1.2) holds.

Next we show that $|S^{k-1}|\le 1.5s$ for all $k\in\mathbb{N}$. When $k=1$, this inequality holds automatically since $S^0=S^*$ implied by $w^0\le\frac{1}{2}e$. Now assume that $|S^{k-1}|\le 1.5s$ for $k=l$ with $l\ge 1$. From the above argument, we have $\|\beta^l-\beta^*\|\le\frac{(2+\sqrt{6}/8)\lambda\sqrt{s}}{\kappa-24s\|D\|_{\max}}$. Notice that $i\in S^l\backslash S^*$ implies $i\notin S^*$ and $w_i^l\in(\frac{1}{2},1]$. By equation (5.10), the latter implies $\rho_l|\beta_i^l|\ge 1$. Consequently,

$$\begin{aligned}\sqrt{|S^l\backslash S^*|}&\le\sqrt{\textstyle\sum_{i\in S^l\backslash S^*}(\rho_l|\beta_i^l|)^2}\le\rho_l\|\beta^l-\beta^*\|\\&\le\frac{(2+\sqrt{6}/8)\rho_l\lambda\sqrt{s}}{\kappa-24s\|D\|_{\max}}\le\sqrt{0.5s}\end{aligned}\qquad\text{(S1.5)}$$

where the last inequality is by $\rho_l\lambda\le\rho_3\lambda\le\frac{2(\kappa-24s\|D\|_{\max})}{5\sqrt{2}}$. Thus, $|S^l|\le 1.5s$. Hence, $|S^{k-1}|\le 1.5s$ for all $k\in\mathbb{N}$, and the error bound follows from (S1.2).

S1.2 The proof of Theorem 3

To achieve the conclusion of Theorem 3, we need the following lemma.

Lemma 3. *Let F^k and Λ^k be the sets in (4.6). Then, for each $k\in\{0\}\cup\mathbb{N}$,*

$$\sqrt{\textstyle\sum_{i\in S^*}(v_i^k)^2}\le\sqrt{\textstyle\sum_{i\in S^*}\max(\mathbb{I}_{\Lambda^k}(i),\mathbb{I}_{F^k}(i))}.$$

Proof. Fix an arbitrary $i\in S^*$. If $i\in F^k$, from $v_i^k=1-w_i^k\le 1$ we have $v_i^k\le\mathbb{I}_{F^k}(i)$. If $i\notin F^k$, from $v_i^k=1-w_i^k$ and (3.3), it follows that $v_i^k=\max\big(0,\min(1,\frac{2a-(a+1)\rho_k|\beta_i^k|}{2(a-1)})\big)$, and hence $v_i^k\le\mathbb{I}_{\{i:\,\rho_k|\beta_i^k|\le 2a/(a+1)\}}(i)\le\mathbb{I}_{\Lambda^k}(i)$. Thus, for each i, it holds that $(v_i^k)^2\le v_i^k\le\max(\mathbb{I}_{\Lambda^k}(i),\mathbb{I}_{F^k}(i))$. From this, it is immediate to obtain the desired result. □

The proof of Theorem 3: Write $S^{k-1} := S^* \cup \{i \notin S^* : w_i^{k-1} > \frac{1}{2}\}$ for each $k \in \mathbb{N}$. Since the conclusion holds automatically for $k = 1$, it suffices to consider the case $k \geq 2$. From the proof of Theorem 2, we know that $|S^{k-1}| \leq 1.5s$ for all $k \in \mathbb{N}$. Moreover, by using (S1.5) and $\rho_k \geq 1$,

$$\big\|\widetilde{\varepsilon}_{S^{k-1}}\big\| \leq \big\|\widetilde{\varepsilon}_{S^*}\big\| + \sqrt{|S^{k-1}\backslash S^*|}\big\|\widetilde{\varepsilon}\big\|_\infty \leq \big\|\widetilde{\varepsilon}_{S^*}\big\| + \frac{\rho_{k-1}\lambda}{8}\sqrt{|S^{k-1}\backslash S^*|}. \quad \text{(S1.6)}$$

By using inequality (S1.4) and Lemma 3, it follows that

$$\begin{aligned}
\|\beta^k - \beta^*\| &\leq \frac{2}{\kappa - 24s\|D\|_{\max}}\Big[\|\widetilde{\varepsilon}_{S^{k-1}}\| + \lambda\sqrt{\textstyle\sum_{i\in S^*}(v_i^{k-1})^2}\Big] \\
&\leq \frac{2}{\kappa - 24s\|D\|_{\max}}\Big[\|\widetilde{\varepsilon}_{S^{k-1}}\| + \lambda\sqrt{\textstyle\sum_{i\in S^*}\max(\mathbb{I}_{\Lambda^{k-1}}(i), \mathbb{I}_{F^{k-1}}(i))}\Big] \\
&\leq \frac{2}{\kappa - 24s\|D\|_{\max}}\Big[\|\widetilde{\varepsilon}_{S^{k-1}}\| + \lambda\sqrt{\textstyle\sum_{i\in S^*}\max\big(\mathbb{I}_{\Lambda^{k-1}}(i), \big||\beta_i^{k-1}| - |\beta_i^*|\big|^2(\rho_{k-1})^2\big)}\Big] \\
&\leq \frac{2}{\kappa - 24s\|D\|_{\max}}\Big(\|\widetilde{\varepsilon}_{S^{k-1}}\| + \lambda\sqrt{\max\big(\textstyle\sum_{i\in S^*}\mathbb{I}_{\Lambda^{k-1}}(i), (\rho_{k-1})^2\|\Delta\beta^{k-1}\|^2\big)}\Big)
\end{aligned}$$

where the third inequality is by the definition of F^{k-1}. Together with (S1.6),

$$\begin{aligned}
\|\beta^k - \beta^*\| &\leq \frac{2}{\kappa - 24s\|D\|_{\max}}\Big[\big\|\widetilde{\varepsilon}_{S^*}\big\| + \lambda\sqrt{\textstyle\sum_{i\in S^*}\mathbb{I}_{\Lambda^{k-1}}(i)} + \frac{9\rho_{k-1}\lambda}{8}\big\|\Delta\beta^{k-1}\big\|\Big] \\
&\leq \frac{2}{\kappa - 24s\|D\|_{\max}}\Big(\big\|\widetilde{\varepsilon}_{S^*}\big\| + \lambda\sqrt{\textstyle\sum_{i\in S^*}\mathbb{I}_{\Lambda^{k-1}}(i)}\Big) + \frac{1}{\sqrt{2}}\|\beta^{k-1} - \beta^*\|
\end{aligned}$$

where the second inequality is using $\rho_{k-1}\lambda \leq \rho_3\lambda \leq \frac{2(\kappa - 24s\|D\|_{\max})}{5\sqrt{2}}$. The desired result follows by solving this recursion with respect to $\|\beta^k - \beta^*\|$.

S1.3 The proof of Theorem 4

We need the following two lemmas with $\Delta\widehat{\beta}^k = \beta^k - \beta^{\mathrm{LS}}$ for $k = 1, 2, \ldots$.

Lemma 4. *Suppose that for some $k \geq 1$ there exists an index set $S^{k-1} \supseteq S^*$ such that $\max_{i\in(S^{k-1})^c} w_i^{k-1} \leq \frac{1}{2}$. Then, whenever $\lambda \geq 6\|\varepsilon^{\rm LS}\|_\infty$, it holds that*

$$\|\Delta\widehat{\beta}^k_{(S^{k-1})^c}\|_1 \leq 3\|\Delta\widehat{\beta}^k_{S^{k-1}}\|_1.$$

Proof. By the optimality of β^k and the feasibility of $\beta^{\rm LS}$ to (3.1), we have

$$\frac{1}{2n}\big\|\widetilde{Z}\beta^k-\widetilde{y}\big\|^2+\lambda\sum_{i=1}^p v_i^{k-1}\big|\beta_i^k\big| \leq \frac{1}{2n}\big\|\widetilde{Z}\beta^{\rm LS}-\widetilde{y}\big\|^2+\lambda\sum_{i=1}^p v_i^{k-1}\big|\beta_i^{\rm LS}\big|,$$

which, by $\Delta\widehat{\beta}^k=\beta^k-\beta^{\rm LS}$ and $\varepsilon^{\rm LS}=\frac{1}{n}\widetilde{Z}^{\mathbb{T}}(\widetilde{y}-\widetilde{Z}\beta^{\rm LS})$, can be rearranged as

$$\begin{aligned}\frac{1}{2n}\big\|\widetilde{Z}\Delta\widehat{\beta}^k\big\|^2 &\leq \langle\varepsilon^{\rm LS},\Delta\widehat{\beta}^k\rangle+\lambda\sum_{i=1}^p v_i^{k-1}(|\beta_i^{\rm LS}|-|\beta_i^k|)\\ &=\sum_{i\notin S^*}\varepsilon_i^{\rm LS}\Delta\widehat{\beta}_i^k+\lambda\sum_{i\in S^*}v_i^{k-1}\big(\big|\beta_i^{\rm LS}\big|-\big|\beta_i^k\big|\big)-\lambda\sum_{i\notin S^*}v_i^{k-1}\big|\beta_i^k\big|\\ &\leq\sum_{i\notin S^*}\big|\varepsilon_i^{\rm LS}\big|\big|\Delta\widehat{\beta}_i^k\big|+\lambda\sum_{i\in S^*}v_i^{k-1}\big|\Delta\widehat{\beta}_i^k\big|-\lambda\sum_{i\notin S^*}v_i^{k-1}\big|\beta_i^k\big|\end{aligned}$$

where the equality is using $\varepsilon_i^{\rm LS}=0$ for $i\in S^*$ and $\beta_i^{\rm LS}=0$ for all $i\notin S^*$. Now from $S^{k-1}\supseteq S^*$ and $v_i^{k-1}=1-w_i^{k-1}\geq 1/2$ for $i\notin S^{k-1}$, we obtain

$$\begin{aligned}\frac{1}{2n}\big\|\widetilde{Z}\Delta\widehat{\beta}^k\big\|^2 &\leq\sum_{i\notin S^*}\big|\varepsilon_i^{\rm LS}\big|\big|\Delta\widehat{\beta}_i^k\big|+\lambda\sum_{i\in S^*}v_i^{k-1}|\Delta\widehat{\beta}_i^k|-\lambda\sum_{i\notin S^{k-1}}v_i^{k-1}\big|\Delta\widehat{\beta}_i^k\big|\\ &\leq\sum_{i\in S^{k-1}\backslash S^*}\big|\varepsilon_i^{\rm LS}\big|\big|\Delta\widehat{\beta}_i^k\big|+\lambda\sum_{i\in S^*}v_i^{k-1}\big|\Delta\widehat{\beta}_i^k\big|\\ &\qquad+\sum_{i\in(S^{k-1})^c}\big|\varepsilon_i^{\rm LS}\big|\big|\Delta\widehat{\beta}_i^k\big|-\frac{1}{2}\lambda\big\|\Delta\widehat{\beta}^k_{(S^{k-1})^c}\big\|_1 \qquad\text{(S1.7)}\\ &\leq\max\big(\|\varepsilon^{\rm LS}\|_\infty,\lambda\big)\big\|\Delta\widehat{\beta}^k_{S^{k-1}}\big\|_1+\big(\|\varepsilon^{\rm LS}\|_\infty-\frac{1}{2}\lambda\big)\big\|\Delta\widehat{\beta}^k_{(S^{k-1})^c}\big\|_1\end{aligned}$$

which along with the nonnegativity of $\frac{1}{2n}\|\widetilde{Z}\Delta\widehat{\beta}^k\|^2$ implies the result. □

Lemma 5. *Suppose that for some $k \geq 1$ there exists $S^{k-1} \supseteq S^*$ with $|S^{k-1}| \leq 1.5s$ such that $\max_{i\in(S^{k-1})^c} w_i^{k-1} \leq \frac{1}{2}$, and that the matrix Σ satisfies the κ-REC on $\mathcal{C}(S^*)$ with $\kappa > 24s\|D\|_{\max}$. Then, when $\lambda \geq 6\|\varepsilon^{\mathrm{LS}}\|_\infty$,*

$$\big\|\Delta\widehat{\beta}^k\big\| \leq \frac{2}{\kappa - 24s\|D\|_{\max}}\Big(\big\|\varepsilon^{\mathrm{LS}}_{S^{k-1}}\big\| + \lambda\sqrt{\textstyle\sum_{i\in S^*}(v_i^{k-1})^2}\Big).$$

Proof. First of all, from equation (S1.7) and $\lambda \geq 6\|\varepsilon^{\mathrm{LS}}\|_\infty$, it follows that

$$\begin{aligned}\frac{1}{2n}\big\|\widetilde{Z}\Delta\widehat{\beta}^k\big\|^2 &\leq \sum_{i\in S^{k-1}\setminus S^*}\big|\varepsilon_i^{\mathrm{LS}}\big|\big|\Delta\widehat{\beta}_i^k\big| + \lambda\sum_{i\in S^*}v_i^{k-1}\big|\Delta\widehat{\beta}_i^k\big| \\ &\leq \|\varepsilon^{\mathrm{LS}}_{S^{k-1}}\|\|\Delta\widehat{\beta}^k_{S^{k-1}}\| + \lambda\sqrt{\textstyle\sum_{i\in S^*}(v_i^{k-1})^2}\,\|\Delta\widehat{\beta}^k_{S^{k-1}}\|\end{aligned}$$

where the second inequality is using $S^{k-1} \supseteq S^*$. Together with Lemma 1,

$$\frac{1}{2n}\big\|X\Delta\widehat{\beta}^k\big\|^2 \leq \Big[\|\varepsilon^{\mathrm{LS}}_{S^{k-1}}\| + \lambda\sqrt{\textstyle\sum_{i\in S^*}(v_i^{k-1})^2}\Big]\big\|\Delta\widehat{\beta}^k_{S^{k-1}}\big\| - \frac{1}{2}(\Delta\widehat{\beta}^k)^{\mathbb{T}}D\Delta\widehat{\beta}^k.$$

Since $S^{k-1} \supseteq S^*$ with $|S^{k-1}| \leq 1.5s$, using Lemma 4 and the same arguments as for (S1.3) yields that $-(\Delta\widehat{\beta}^k)^{\mathbb{T}}D\Delta\widehat{\beta}^k \leq 24s\|D\|_{\max}\|\Delta\widehat{\beta}^k\|^2$. Then,

$$\frac{1}{2n}\big\|X\Delta\widehat{\beta}^k\big\|^2 - 12s\|D\|_{\max}\|\Delta\widehat{\beta}^k\|^2 \leq \Big[\|\varepsilon^{\mathrm{LS}}_{S^{k-1}}\| + \lambda\sqrt{\textstyle\sum_{i\in S^*}(v_i^{k-1})^2}\Big]\big\|\Delta\widehat{\beta}^k_{S^{k-1}}\big\|.$$

Since Σ satisfies the κ-RSC on the set $\mathcal{C}(S^*)$ with $\kappa > 24s\|D\|_{\max}$, we have

$$\frac{1}{2}(\kappa - 24s\|D\|_{\max})\big\|\Delta\widehat{\beta}^k\big\|^2 \leq \Big[\|\varepsilon^{\mathrm{LS}}_{S^{k-1}}\| + \lambda\sqrt{\textstyle\sum_{i\in S^*}(v_i^{k-1})^2}\Big]\big\|\Delta\widehat{\beta}^k_{S^{k-1}}\big\|.$$

This implies the desired result. The proof is then completed. □

The proof of Theorem 4: Let $S^{k-1}:=S^*\cup\{i\notin S^*: w_i^{k-1}>\frac{1}{2}\}$ for each $k\in\mathbb{N}$. We first prove that the desired inequalities holds by the induction on $k\in\mathbb{N}$. Since $w^0\le\frac{1}{2}e$, we have $S^0=S^*$ and $|S^0|=s$. Notice that Σ satisfies the κ-REC on $\mathcal{C}(S^*)$ with $\kappa>24s\|D\|_{\max}$ and $\lambda\ge 6\|\varepsilon^{\rm LS}\|_\infty$. The conditions of Lemma 5 are satisfied. Along with $\varepsilon^{\rm LS}_{S^*}=0$ and $F^0=S^*$,

$$\begin{aligned}\|\beta^1-\beta^{\rm LS}\| &\le \frac{2}{\gamma}\Big(\|\varepsilon^{\rm LS}_{S^0}\|+\lambda\sqrt{\textstyle\sum_{i\in S^*}(v_i^0)^2}\Big)\\ &\le \frac{2}{\gamma}\Big(\|\varepsilon^{\rm LS}_{S^*}\|+\lambda\sqrt{|F^0|}\Big)\le\frac{2.03\rho_0\lambda\sqrt{|F^0|}}{\gamma}. \end{aligned}\tag{S1.8}$$

Since $|\beta_i^{\rm LS}-\beta_i^*|\le\|\widetilde{\varepsilon}^\dagger\|_\infty$ for $i\in S^*$ by (4.8) and $\rho_1\ge\gamma\lambda^{-1}\|\widetilde{\varepsilon}^\dagger\|_\infty$, we have

$$|\beta_i^{\rm LS}-\beta_i^1|\ge|\beta_i^*-\beta_i^1|-|\beta_i^*-\beta_i^{\rm LS}|\ge\frac{1}{\rho_1}-\frac{\rho_1\lambda}{\gamma}\ge\frac{9\sqrt{3}-4}{9\sqrt{3}\rho_1}\quad\forall i\in F^1$$

where the last inequality is by $1\le\rho_1\le\sqrt{\frac{4\gamma}{9\sqrt{3}\lambda}}$. By the last two equations,

$$\sqrt{|F^1|}=\sqrt{\textstyle\sum_{i=1}^p\mathbb{I}_{F^1}(i)}\le\frac{9\sqrt{3}\rho_1}{9\sqrt{3}-4}\sqrt{\textstyle\sum_{i=1}^p|\beta_i^{\rm LS}-\beta_i^1|^2}\le\frac{18.27\sqrt{3}\rho_1\rho_0\lambda}{(9\sqrt{3}-4)\gamma}\sqrt{|F^0|}.$$

Together with (S1.8) and $1=\rho_0<\rho_1\le\rho_3$, we conclude that the desired inequalities holds for $k=1$. Now, assuming that the conclusion holds for $k\le l-1$ with $l\ge 2$, we prove that the conclusion holds for $k=l$. For this purpose, we first argue $|S^{l-1}|\le 1.5s$. Indeed, for $i\in S^{l-1}\backslash S^*$, we have

$w_i^{l-1} \in (\frac{1}{2}, 1]$, which by (3.3) implies that $\rho_{l-1}|\beta_i^{l-1}| \geq 1$. Then,

$$\begin{aligned}
\sqrt{|S^{l-1}\backslash S^*|} &\leq \sqrt{|F^{l-1}|} \leq \frac{18.27\sqrt{3}\rho_{l-1}\rho_{l-2}\lambda}{(9\sqrt{3}-4)\gamma}\sqrt{|F^{l-2}|} \leq \cdots \\
&\leq \Big(\frac{18.27\sqrt{3}\lambda}{(9\sqrt{3}-4)\gamma}\Big)^{l-1}\rho_{l-1}\rho_{l-2}^2\cdots\rho_2^2\rho_1\sqrt{|F^0|} \\
&\leq \sqrt{\Big(\frac{18.27\sqrt{3}(\rho_3)^2\lambda}{(9\sqrt{3}-4)\gamma}\Big)^{2l-2}|F^0|} \leq \sqrt{\Big(\frac{8.12}{9\sqrt{3}-4}\Big)^{2l-2}|F^0|} \leq \sqrt{0.5s},
\end{aligned}$$

where the first inequality is due to $S^{l-1}\backslash S^* \subseteq F^{l-1}$, the second is since the conclusion holds for $k \leq l-1$ with $l \geq 2$, the next to the last is using $\rho_3 \leq \sqrt{\frac{4\gamma}{9\sqrt{3}\lambda}}$, and the last one is using $2l-2 \geq 2$. The last inequality implies that $|S^{l-1}| \leq 1.5s$. Using Lemma 5 delivers that

$$\begin{aligned}
\|\beta^l - \beta^{\mathrm{LS}}\| &\leq \frac{2}{\gamma}\Big(\|\varepsilon^{\mathrm{LS}}_{S^{l-1}}\| + \lambda\sqrt{\textstyle\sum_{i\in S^*}(v_i^{l-1})^2}\Big) \\
&\leq \frac{2}{\gamma}\Big(\|\varepsilon^{\mathrm{LS}}_{S^{l-1}\backslash S^*}\| + \lambda\sqrt{\textstyle\sum_{i\in S^*}\mathbb{I}_{F^{l-1}}(i)}\Big) \\
&\leq \frac{2}{\gamma}\Big(\|\varepsilon^{\mathrm{LS}}\|_\infty\sqrt{|S^{l-1}\backslash S^*|} + \lambda\sqrt{|F^{l-1}\cap S^*|}\Big) \\
&\leq \frac{2\lambda}{\gamma}\Big(\frac{1}{6}\sqrt{|F^{l-1}\backslash S^*|} + \sqrt{|F^{l-1}\cap S^*|}\Big) \\
&\leq \frac{2\lambda}{\gamma}\sqrt{(1+1/36)|F^{l-1}|} \leq \frac{2.03\rho_{l-1}\lambda}{\gamma}\sqrt{|F^{l-1}|},
\end{aligned}$$

where the second inequality is using $\varepsilon^{\mathrm{LS}}_{S^*} = 0$, Lemma 3 and $\rho_{l-1} \geq \rho_1 > \frac{4a}{(a+1)\min_{i\in S^*}|\beta_i|}$, the fourth one is due to $\lambda \geq 6\|\varepsilon^{\mathrm{LS}}\|_\infty$, and the fifth one is since $\frac{1}{6}a + b \leq \sqrt{(1+\frac{1}{36})(a^2+b^2)}$ for all $a, b \in \mathbb{R}$. Now using the same argument as those for $k = 1$, we have $|\beta_i^l - \beta_i^{\mathrm{LS}}| \geq \frac{9\sqrt{3}-4}{9\sqrt{3}\rho_l}$ for all $i \in F^l$, and hence $\sqrt{|F^l|} \leq \frac{18.27\sqrt{3}\rho_l\rho_{l-1}\lambda}{(9\sqrt{3}-4)\gamma}\sqrt{|F^{l-1}|}$. Thus, we complete the proof of the case $k = l$, and the desired inequalities hold for all k.

Note that $(\rho_3)^2\lambda \leq \frac{4\gamma}{9\sqrt{3}}$ and $\rho_k \leq \rho_3$ for all $k \in \mathbb{N}$. So, it holds that

$$\sqrt{|F^{\overline{k}}|} \leq \frac{18.27\sqrt{3}\rho_{\overline{k}}\rho_{\overline{k}-1}\lambda}{(9\sqrt{3}-4)\gamma}\sqrt{|F^{\overline{k}-1}|} \leq \cdots \leq \Big(\frac{18.27\sqrt{3}(\rho_3)^2\lambda}{(9\sqrt{3}-4)\gamma}\Big)^{\overline{k}}\sqrt{|F^0|} < 1,$$

which implies that $|F^k| = 0$ when $k \geq \overline{k}$. Together with the first inequality obtained, we have $\beta^k = \beta^{\rm LS}$ when $k \geq \overline{k}$. From $\rho_3 \leq \sqrt{\frac{4\gamma}{9\sqrt{3}\lambda}}$ and (4.8),

$$\big||\beta_i^*| - |\beta_i^{\rm LS}|\big| \leq |\beta_i^* - \beta_i^{\rm LS}| \leq \|\widehat{\varepsilon}^{\dagger}\|_\infty \leq \rho_k\lambda\gamma^{-1} \leq \frac{4}{9\sqrt{3}\rho_k} \quad \forall i \in S^*. \tag{S1.9}$$

This, along with $\min_{i\in S^*}|\beta_i^*| \geq \frac{4a}{(a+1)\rho_k} > \frac{4}{9\sqrt{3}\rho_k}$, implies $|\beta_i^{\rm LS}| > 0$ for all $i \in S^*$ (if not, one will obtain $\frac{a}{a+1} \leq \frac{1}{9\sqrt{3}}$, a contradiction to $a > 1$), and hence ${\rm supp}(\beta^{\rm LS}) = S^*$. The last inequality also implies ${\rm sign}(\beta^{\rm LS}) = {\rm sign}(\beta^*)$ (if not, there exists $i_0 \in S^*$ such that ${\rm sign}(\beta_{i_0}^{\rm LS}) = -{\rm sign}(\beta_{i_0}^*)$ and then $|\beta_{i_0}^* - \beta_{i_0}^{\rm LS}| > |\beta_{i_0}^*| \geq \min_{i\in S^*}|\beta_i^*| > \frac{4}{9\sqrt{3}\rho_k}$, a contradiction to (S1.9).) Thus, $\beta^k = \beta^{\rm LS}$ and ${\rm sign}(\beta^k) = {\rm sign}(\beta^*)$ for all $k \geq \overline{k}$. We complete the proof.

S2. Additional Theoretical Results

In this part, we need the following assumption on the noise vector ε.

Assumption 1. Assume that $\varepsilon_i\,(i = 1, \ldots, m)$ are i.i.d. sub-Gaussians, i.e., there is $\sigma > 0$ such that $\mathbb{E}[\exp(t\varepsilon_i)] \leq \exp(\sigma^2t^2/2)$ for all i and $t \in \mathbb{R}$.

S2.1 Additive errors case

In this part, we consider that the matrix X is contaminated by additive measurement errors, i.e., $Z = X + A$, where $A = (a_{ij})$ is the matrix of measurement errors and the rows of A are assumed to be i.i.d. with zero mean, finite covariance Σ_A and sub-Gaussian parameter τ^2. Following the line of Loh (2014), we assume that Σ_A is known. Now the unbiased surrogates of Σ and ξ are given by $\widehat{\Sigma}_{\text{add}} = \frac{1}{n} Z^{\mathbb{T}} Z - \Sigma_A$ and $\widehat{\xi}_{\text{add}} = \frac{1}{n} Z^{\mathbb{T}} y$, respectively. We write $\widetilde{\Sigma}_{\text{add}} := \widehat{\epsilon} I + \Pi_{\mathbb{S}_+^p}(\widehat{\Sigma}_{\text{add}} - \widehat{\epsilon} I)$ and $\widetilde{\varepsilon}_{\text{add}} := \widehat{\xi}_{\text{add}} - \widetilde{\Sigma}_{\text{add}} \beta^*$.

Lemma 6. *Let* $K := 2(\lambda_{\max}(\Sigma_A) + \widehat{\epsilon}) \|\beta^*\|_1$ *and* $\eta = \min\left(1, \frac{\epsilon_0}{\lambda_{\max}(\Sigma_A) + \widehat{\epsilon}}\right)$. *Then, there exist universal positive constants* C *and* c, *and positive function* $\widehat{\zeta}$ *(depending only on* β^*, τ^2, σ^2 *and* $\lambda_{\max}(\Sigma_A)$*) such that*

$$\mathbb{P}\{\|(\widetilde{\Sigma}_{\text{add}} - \Sigma)\beta^*\|_\infty > K\} \le Cp^2 \exp(-cn\widehat{\zeta}^{-1}\eta^2), \tag{S2.1}$$

$$\mathbb{P}\{\|\widetilde{\varepsilon}_{\text{add}}\|_\infty > K\} \le Cp^2 \exp(-cns^{-2}\widehat{\zeta}^{-1}\eta^2). \tag{S2.2}$$

Proof. From the expression of $\widetilde{\Sigma}_{\text{add}}$, it follows that

$$\begin{aligned}
\|(\widetilde{\Sigma}_{\text{add}} - \Sigma)\beta^*\|_\infty &\le \|(\widetilde{\Sigma}_{\text{add}} - \widehat{\Sigma}_{\text{add}})\beta^*\|_\infty + \|(\widehat{\Sigma}_{\text{add}} - \Sigma)\beta^*\|_\infty \\
&= \|\Pi_{\mathbb{S}_+^p}(\widehat{\epsilon} I - \widehat{\Sigma}_{\text{add}})\beta^*\|_\infty + \|(\widehat{\Sigma}_{\text{add}} - \Sigma)\beta^*\|_\infty \\
&\le \|\Pi_{\mathbb{S}_+^p}(\widehat{\epsilon} I - \widehat{\Sigma}_{\text{add}})\|_{\max} \|\beta^*\|_1 + \|(\widehat{\Sigma}_{\text{add}} - \Sigma)\beta^*\|_\infty.
\end{aligned}$$

For a matrix $\Gamma \in \mathbb{S}_+^p$, it is not hard to check that $\lambda_{\max}(\Gamma) \geq \|\Gamma\|_{\max}$. Thus,

$$\begin{aligned}\|(\widetilde{\Sigma}_{\rm add} - \Sigma)\beta^*\|_\infty &\leq \lambda_{\max}\big[\Pi_{\mathbb{S}_+^p}(\widehat{\epsilon} I - \widehat{\Sigma}_{\rm add})\big]\|\beta^*\|_1 + \|(\widehat{\Sigma}_{\rm add} - \Sigma)\beta^*\|_\infty \\ &= \big[\widehat{\epsilon} - \lambda_{\min}(\widehat{\Sigma}_{\rm add})\big]\|\beta^*\|_1 + \|(\widehat{\Sigma}_{\rm add} - \Sigma)\beta^*\|_\infty. \quad \text{(S2.3)}\end{aligned}$$

Notice that $\lambda_{\min}(\widehat{\Sigma}_{\rm add}) \geq \lambda_{\min}(\frac{1}{n}Z^{\mathbb{T}}Z) - \lambda_{\max}(\Sigma_A) \geq -\lambda_{\max}(\Sigma_A)$ implied by (Horn and Johnson, 1990, Theorem 4.3.7). Together with (S2.3),

$$\|(\widetilde{\Sigma}_{\rm add} - \Sigma)\beta^*\|_\infty \leq (\widehat{\epsilon} + \lambda_{\max}(\Sigma_A))\|\beta^*\|_1 + \|(\widehat{\Sigma}_{\rm add} - \Sigma)\beta^*\|_\infty.$$

By this and (Datta and Zou, 2017, Lemma 1) with $\epsilon = \frac{K\eta}{2\|\beta^*\|_1} \leq \epsilon_0$, there exist universal positive constants C, c and positive functions ζ (depending only on β^*, τ^2, σ^2 and $\lambda_{\max}(\Sigma_A)$) such that

$$\begin{aligned}\mathbb{P}\{\|(\widetilde{\Sigma}_{\rm add} - \Sigma)\beta^*\|_\infty > K\} &\leq \mathbb{P}\Big\{\|(\widehat{\Sigma}_{\rm add} - \Sigma)\beta^*\|_\infty > K/2\Big\} \\ &\leq \mathbb{P}\Big\{\|\widehat{\Sigma}_{\rm add} - \Sigma\|_{\max} > \frac{K\eta}{2\|\beta^*\|_1}\Big\} \\ &\leq Cp^2 \exp(-cn\eta^2(\lambda_{\max}(\Sigma_A) + \widehat{\epsilon})^2\zeta^{-1}).\end{aligned}$$

This shows that (S2.1) holds. Recall that $\widetilde{\varepsilon}_{\rm add} = \widehat{\xi}_{\rm add} - \widetilde{\Sigma}_{\rm add}\beta^*$. Hence,

$$\|\widetilde{\varepsilon}_{\rm add}\|_\infty \leq \|\widehat{\xi}_{\rm add} - \xi\|_\infty + \|\xi - \Sigma\beta^*\|_\infty + \|(\widetilde{\Sigma}_{\rm add} - \Sigma)\beta^*\|_\infty.$$

By applying (Datta and Zou, 2017, Lemma 1) with $\epsilon = \frac{K\eta_0}{3} \leq \epsilon_0$ where $\eta_0 = \min(1, \frac{1.5\eta}{\|\beta^*\|_1})$, we obtain

$$\mathbb{P}\Big\{\|\widehat{\xi}_{\rm add} - \xi\|_\infty \geq \frac{K}{3}\Big\} \leq \mathbb{P}\Big\{\|\widehat{\xi}_{\rm add} - \xi\|_\infty \geq \frac{K\eta_0}{3}\Big\} \leq Cp\exp(-ncs^{-2}K^2\eta_0^2\zeta^{-1}),$$

while $\mathbb{P}\{\|\xi - \Sigma\beta^*\|_\infty \geq K/3\} \leq Cp\exp(-nc\sigma^{-2}K^2)$ holds by (Datta and Zou, 2017, Property B.2). Together with the last inequality and inequality (S2.1), we obtain the inequality (S2.2). □

Lemma 6 states that $\|(\widetilde{\Sigma}_{\text{add}} - \Sigma)\beta^*\|_\infty$ and $\|\widehat{\xi}_{\text{add}}\|_\infty$ can be controlled by $\|\beta^*\|_1$. From the proof of (Datta and Zou, 2017, Theorem 1), we know that there also exist universal positive constants C' and c' and positive function $\widehat{\zeta}'$ (depending on $\beta^*_{S^*}, \tau^2$ and σ^2) such that for all $\epsilon \leq \min(\epsilon_0, \frac{\kappa}{64s})$,

$$\mathbb{P}\big\{\|D\|_{\max} \geq \kappa/(64s)\big\} \leq C'p^2\exp(-nc'\epsilon^2(\widehat{\zeta}')^{-1}). \tag{S2.4}$$

Combining with Lemma 6 and Theorem 3, we have the following result.

Corollary 1. *Suppose that Σ satisfies the κ-REC on $\mathcal{C}(S^*)$. If λ and ρ_3 in Algorithm 1 are chosen such that $\lambda \geq 8K$ and $\rho_3 \leq \frac{\kappa}{4\sqrt{2}\lambda}$ where K is the constant same as in Lemma 6, then for all $k \in \mathbb{N}$ the following inequality*

$$\|\beta^k - \beta^*\| \leq \frac{4\sqrt{s}\,\lambda}{\kappa} \tag{S2.5}$$

holds w.p. at least $1 - p^2C\exp(-cns^{-2}\zeta^{-1})$, where C and c are universal positive constants and ζ is a positive function on $\beta^*, \tau^2, \sigma^2, \kappa$ and $\lambda_{\max}(\Sigma_A)$.

Write $\widetilde{G}_{\text{add}} := [\widetilde{\Sigma}_{\text{add}}]_{(S^*)^cS^*}[\widetilde{\Sigma}_{\text{add}}]^{-1}_{S^*S^*}$. By recalling $\varepsilon^{\text{LS}} = \frac{1}{n}\widetilde{Z}^{\mathbb{T}}(\widetilde{y} - \widetilde{Z}\beta^{\text{LS}})$ and using the equality (4.8), it is not difficult to obtain the inequalities

$$\|\varepsilon^{\text{LS}}\|_\infty \leq \max(2, 1{+}s\|\widetilde{G}_{\text{add}}\|_{\max})\|\widetilde{\varepsilon}_{\text{add}}\|_\infty,\ \|\widetilde{\varepsilon}^\dagger\|_\infty \leq s\|[\widetilde{\Sigma}_{\text{add}}]^{-1}_{S^*S^*}\|_{\max}\|\widetilde{\varepsilon}_{\text{add}}\|_\infty.$$

Along with Lemma 6, Theorem 4 and (S2.4), we obtain the following result.

Corollary 2. *Suppose that Σ satisfies the κ-REC on the set $\mathcal{C}(S^*)$. Write $K' = K\max(2, 1+s\|\widetilde{G}_{\rm add}\|_{\max})$ and $K'' = Ks\|[\widetilde{\Sigma}_{\rm add}]^{-1}_{S^*S^*}\|_{\max}$ where the constant K is same as the one in Lemma 6. If λ, ρ_1 and ρ_3 are chosen such that $\lambda \geq 6K'$, $\rho_1 > \max\big(\frac{4a}{(a+1)\min_{i\in S^*}|\beta_i^*|}, \frac{5\kappa K''}{8\lambda}\big)$ and $\rho_3 \leq \sqrt{\frac{5\kappa}{18\sqrt{3}\lambda}}$, then $\beta^k = \beta^{\rm LS}$ and ${\rm sign}(\beta^k) = {\rm sign}(\beta^*)$ for $k \geq \widehat{k} = \lceil\frac{0.5\ln(s)}{\ln[(9\sqrt{3}-4)5\kappa\lambda^{-1}]-\ln[147\sqrt{3}(\rho_3)^2]}\rceil$ w.p. at least $1-Cp^2\exp(-cns^{-2}\zeta^{-1})$, where C, c are universal positive constants and ζ is a positive function depending on $\beta^*, \tau^2, \sigma^2, \kappa$ and $\lambda_{\max}(\Sigma_A)$.*

As remarked in the beginning of this subsection, when X is from the Σ_x-Gaussian ensemble, with high probability there exists a constant $\kappa > 0$ such that Σ satisfies the REC on $\mathcal{C}(S^*)$. We see that if κ has a small value, there is a great possibility for the choice range of ρ_3 to be empty, and it is impossible to achieve the sign consistency; and when κ is not too small, say, $\frac{5\kappa}{108\sqrt{3}K'} > 1$, after $k \geq \widehat{k} \geq \lceil\frac{0.5\ln(s)}{\ln(1.42)}\rceil$ the iterate β^k is sign-consistent.

S2.2 Multiplicative errors and missing data

In this part, we consider that the matrix X is contaminated by multiplicative measurement errors, i.e. $Z = X \circ M$, where $M = (m_{ij})$ is the matrix of measurement errors and the rows of M are assumed to be i.i.d. with mean μ_M, covariance Σ_M and sub-Gaussian parameter τ^2. Similar to Datta and

Zou (2017), in the sequel we need the following conditions

$$\max_{i,j} |X_{ij}| \le c_X,\ \max_{i,j} |M_{ij}| \le c_M,\ \min_{i,j} (\Sigma_M)_{ij} > 0,\ (\mu_M)_{\min} > 0 \qquad \text{(S2.6)}$$

where c_X and c_M are universal positive constants. From Loh and Wainwright (2012), $\widehat{\Sigma}_{\mathrm{mul}} = \frac{1}{n} Z^{\mathbb{T}} Z \oslash (\Sigma_M + \mu_M \mu_M^{\mathbb{T}})$ and $\widehat{\xi}_{\mathrm{mul}} = \frac{1}{n} Z^{\mathbb{T}} y \oslash \mu_M$ are the unbiased surrogates of Σ and ξ, where $\oslash$ denotes the elementwise division operator. Let $\widetilde{\Sigma}_{\mathrm{mul}} := \widehat{\epsilon} I + \Pi_{\mathbb{S}_+^p}(\widehat{\Sigma}_{\mathrm{mul}} - \widehat{\epsilon} I)$ and $\widetilde{\varepsilon}_{\mathrm{mul}} := \widehat{\xi}_{\mathrm{mul}} - \widetilde{\Sigma}_{\mathrm{mul}} \beta^*$.

Lemma 7. *Let* $\widetilde{K} := 2\big[\widehat{\epsilon} - \min(\lambda_{\min}(\Sigma_M^\dagger), 0) c_M^2\big] \|\beta^*\|_1$ *with* $\Sigma_M^\dagger = E \oslash (\Sigma_M + \mu_M \mu_M^{\mathbb{T}})$ *where* E *is the matrix of all ones and* $\widetilde{\eta} = \min\big(1, \frac{\epsilon_0}{\widehat{\epsilon} - \min(\lambda_{\min}(\Sigma_M^\dagger), 0) c_M^2}\big)$. *Then, there exist universal positive constants* $\widetilde{C}, \widetilde{c}$ *and positive function* $\widetilde{\zeta}$ *(depending on* $\beta^*, \tau^2, \sigma^2, \lambda_{\min}(\Sigma_M^\dagger)$ *and the constants in* (S2.6)*) such that*

$$\mathbb{P}\{\|(\widetilde{\Sigma}_{\mathrm{mul}} - \Sigma)\beta^*\|_\infty > \widetilde{K}\} \le \widetilde{C} p^2 \exp(-\widetilde{c} n \widetilde{\zeta}^{-1} \widetilde{\eta}^2), \qquad \text{(S2.7)}$$

$$\mathbb{P}\{\|\widetilde{\varepsilon}_{\mathrm{mul}}\|_\infty > \widetilde{K}\} \le \widetilde{C} p^2 \exp(-\widetilde{c} n s^{-2} \widetilde{\zeta}^{-1} \widetilde{\eta}^2). \qquad \text{(S2.8)}$$

Proof. From the expression of $\widetilde{\Sigma}_{\mathrm{mul}}$ and the proof of Lemma 6, we have

$$\|(\widetilde{\Sigma}_{\mathrm{mul}} - \Sigma)\beta^*\|_\infty \le \big[\widehat{\epsilon} - \lambda_{\min}(\widehat{\Sigma}_{\mathrm{mul}})\big] \|\beta^*\|_1 + \|(\widehat{\Sigma}_{\mathrm{mul}} - \Sigma)\beta^*\|_\infty. \qquad \text{(S2.9)}$$

Next we provide a lower bound for $\lambda_{\min}(\widehat{\Sigma}_{\rm mul})$. Write $\Sigma_Z = \frac{1}{n}Z^{\mathbb{T}}Z$. Then,

$$\begin{aligned}
\lambda_{\min}(\widehat{\Sigma}_{\rm mul}) &= \lambda_{\min}\big[\Sigma_Z \circ (\Sigma_M^{\dagger} - \lambda_{\min}(\Sigma_M^{\dagger})I) + (\Sigma_Z \circ \lambda_{\min}(\Sigma_M^{\dagger})I)\big]\\
&\geq \lambda_{\min}\big[\Sigma_Z \circ (\Sigma_M^{\dagger} - \lambda_{\min}(\Sigma_M^{\dagger})I)\big] + \lambda_{\min}[\Sigma_Z \circ \lambda_{\min}(\Sigma_M^{\dagger})I]\\
&\geq \lambda_{\min}(\Sigma_Z)\lambda_{\min}(\Sigma_M^{\dagger} - \lambda_{\min}(\Sigma_M^{\dagger})I) + \lambda_{\min}[\Sigma_Z \circ \lambda_{\min}(\Sigma_M^{\dagger})I]\\
&\geq \lambda_{\min}[\Sigma_Z \circ \lambda_{\min}(\Sigma_M^{\dagger})I] \geq \min(\lambda_{\min}(\Sigma_M^{\dagger}), 0) \max_{1\leq j\leq p}(Z_j^{\mathbb{T}}Z_j/n)\\
&\geq \min(\lambda_{\min}(\Sigma_M^{\dagger}), 0)c_M^2
\end{aligned}$$

where the first inequality is using (Horn and Johnson, 1990, Theorem 4.3.1), the second one is due to $\Sigma_M^{\dagger} - \lambda_{\min}(\Sigma_M^{\dagger})I \succeq 0$ and (Horn and Johnson, 1991, Theorem 5.3.1), the fourth one is using the positive semidefiniteness of Σ_Z, and the last one is due to $Z = X \circ M$ and the first two relations in (S2.6). Together with (S2.9) and the definition of $\widetilde{K}$,

$$\|(\widetilde{\Sigma}_{\rm mul} - \Sigma)\beta^*\|_\infty \leq (\widetilde{K}/2) + \|(\widehat{\Sigma}_{\rm mul} - \Sigma)\beta^*\|_\infty.$$

By (Datta and Zou, 2017, Lemma 2) for $\epsilon = \frac{\widetilde{K}\widetilde{\eta}}{2\|\beta^*\|_1} \leq \epsilon_0$, there are universal positive constants C, c and positive functions ζ (depending on $\beta^*, \tau^2, \sigma^2$) and the constants in (S2.6) such that

$$\begin{aligned}
&\mathbb{P}\{\|(\widetilde{\Sigma}_{\rm mul} - \Sigma)\beta^*\|_\infty > \widetilde{K}\} \leq \mathbb{P}\Big\{\|(\widehat{\Sigma}_{\rm mul} - \Sigma)\beta^*\|_\infty > \frac{\widetilde{K}}{2}\Big\}\\
&\leq \mathbb{P}\Big\{\|(\widehat{\Sigma}_{\rm mul} - \Sigma)\beta^*\|_\infty > \frac{\widetilde{K}\widetilde{\eta}}{2}\Big\} \leq \mathbb{P}\Big\{\|\widehat{\Sigma}_{\rm mul} - \Sigma\|_{\max} > \frac{\widetilde{K}\widetilde{\eta}}{2\|\beta^*\|_1}\Big\}\\
&\leq Cp^2 \exp\big(-cn(\widehat{\epsilon} - \min(\lambda_{\min}(\Sigma_M^{\dagger}), 0)c_M^2)^2\widetilde{\eta}^2\zeta^{-1}\big).
\end{aligned}$$

Thus, we get (S2.7). From (Datta and Zou, 2017, Property B.2) and $\|\widetilde{\varepsilon}_{\mathrm{mul}}\|_\infty \le \|\widehat{\xi}_{\mathrm{mul}} - \xi\|_\infty + \|\xi - \Sigma\beta^*\|_\infty + \|(\widetilde{\Sigma}_{\mathrm{mul}} - \Sigma)\beta^*\|_\infty$, it follows that $\mathbb{P}\{\|\xi - \Sigma\beta^*\|_\infty \ge \widetilde{K}/3\} \le Cp\exp(-nc\sigma^{-2}\widetilde{K}^2)$. Together with (Datta and Zou, 2017, Lemma 2) and the inequality (S2.7), we obtain (S2.8). □

By using Lemma 7 and the same arguments as those for Corollary 1 and 2, the following conclusions hold where $\widetilde{G}_{\mathrm{mul}} := [\widetilde{\Sigma}_{\mathrm{mul}}]_{(S^*)^c S^*}[\widetilde{\Sigma}_{\mathrm{mul}}]^{-1}_{S^* S^*}$.

Corollary 3. *Suppose that Σ satisfies the κ-REC on the set $\mathcal{C}(S^*)$. If λ and ρ_3 are chosen such that $\lambda \ge 8\widetilde{K}$ and $\rho_3 \le \frac{\kappa}{4\sqrt{2}\lambda}$ where $\widetilde{K}$ is the constant in Lemma 7, then for all $k \in \mathbb{N}$ the inequality* (S2.5) *holds w.p. at least $1 - Cp^2\exp(-cns^{-2}\zeta^{-1})$ where C, c are universal positive constants and ζ is a positive function on $\beta^*, \tau^2, \sigma^2, \kappa, \lambda_{\min}(\Sigma_M^\dagger)$ and the constants in* (S2.6).

Corollary 4. *Suppose that Σ satisfies the κ-REC one the set $\mathcal{C}(S^*)$. Write $\widetilde{K}' = \widetilde{K}\max(2, 1+s\|\widetilde{G}_{\mathrm{mul}}\|_{\max})$ and $\widetilde{K}'' = \widetilde{K}s\|[\widetilde{\Sigma}_{\mathrm{mul}}]^{-1}_{S^*S^*}\|_{\max}$ where $\widetilde{K}$ is same as in Lemma 7. If the parameters λ, ρ_1 and ρ_3 in Algorithm 1 are chosen such that $\lambda \ge 6\widetilde{K}'$, $\rho_1 > \max\big(\frac{4a}{(a+1)\min_{i\in S^*}|\beta_i^*|}, \frac{5\kappa\widetilde{K}''}{8\lambda}\big)$ and $\rho_3 \le \sqrt{\frac{5\kappa}{18\sqrt{3}\lambda}}$, then the result of Corollary 2 holds w.p. at least $1 - Cp^2\exp(-cns^{-2}\zeta^{-1})$, where C and c are universal positive constants and ζ is a positive function depending on $\beta^*, \tau^2, \sigma^2, \kappa, \lambda_{\min}(\Sigma_M^\dagger)$ and the constants in* (S2.6).

S3. Implementation of GEP-MSCRA

In this part we pay our attention to the implementation of GEP-MSCRA. We know that GEP-MSCRA consists of solving a sequence of weighted ℓ_1-regularized LS, which can be equivalently written as

$$\min_{\beta,u\in\mathbb{R}^p}\Big\{\frac{1}{2}\|u\|^2+\textstyle\sum_{i=1}^{m}\omega_i|\beta_i|:\ \widetilde{Z}\beta-u=\widetilde{y}\Big\}, \tag{S3.1}$$

where $\omega_i=n\lambda(1-w_i^k)$ for $i=1,\ldots,p$ are the weights. There are some solvers developed for (S3.1); for example, the **SLEP** developed by Liu, Ji and Ye (2011) with the accelerated proximal gradient method in Nesterov (2013), and the semismooth Newton ALM developed by Li, Sun and Toh (2018). Motivated by the performance of the semismooth Newton ALM of Li, Sun and Toh (2018), we apply it for solving the dual of (S3.1), i.e.,

$$\min_{\zeta,\eta\in\mathbb{R}^p}\left\{\frac{1}{2}\|\zeta\|^2+\langle\widetilde{y},\zeta\rangle+\delta_\Lambda(\eta):\ \widetilde{Z}^{\mathbb{T}}\zeta-\eta=0\right\}\ \text{ with }\ \Lambda=[-\omega,\omega]. \tag{S3.2}$$

For a given $\mu>0$, define the augmented Lagrangian function of (S3.2) by

$$L_\mu(\zeta,\eta;\beta):=\frac{1}{2}\|\zeta\|^2+\langle\widetilde{y},\zeta\rangle+\delta_\Lambda(\eta)+\langle\beta,\widetilde{Z}^{\mathbb{T}}\zeta-\eta\rangle+\frac{\mu}{2}\|\widetilde{Z}^{\mathbb{T}}\zeta-\eta\|^2.$$

The iteration steps of the ALM for solving (S3.2) are described as follows.

Next we focus on the solution of the subproblem (S3.3). For any $\zeta\in\mathbb{R}^p$, define $\Phi_j(\zeta):=\min_{\eta\in\mathbb{R}^p}L_{\mu_j}(\zeta,\eta;\beta^j)$. After an elementary calculation,

$$\Phi_j(\zeta)=\frac{\mu_j}{2}\big\|\Pi_\Lambda\big(\widetilde{Z}^{\mathbb{T}}\zeta+\beta^j/\mu_j\big)-\big(\widetilde{Z}^{\mathbb{T}}\zeta+\beta^j/\mu_j\big)\big\|^2+\frac{1}{2}\|\zeta\|^2+\langle\widetilde{y},\zeta\rangle.$$

Algorithm 2 An inexact ALM for the dual problem (S3.2)

Initialization: Choose $\mu_0 > 0$ and a starting point $(\zeta^0, \eta^0, \beta^0)$. Set $j = 0$.

while the stopping conditions are not satisfied **do**

1. Solve the following nonsmooth convex minimization inexactly

$$(\zeta^{j+1}, \eta^{j+1}) \approx \operatorname*{arg\,min}_{\zeta,\eta\in\mathbb{R}^p} L_{\mu_j}(\zeta, \eta; \beta^j). \tag{S3.3}$$

2. Update the multiplier by the formula $\beta^{j+1} = \beta^j + \mu_j(\widetilde{Z}^{\mathbb{T}}\zeta^{j+1} - \eta^{j+1})$.

3. Update $\mu_{j+1} \uparrow \mu_\infty \leq \infty$. Set $j \leftarrow j + 1$, and then go to Step 1.

end while

It is easy to verify that $(\zeta^{j+1}, \eta^{j+1})$ is an optimal solution of (S3.3) iff

$$\zeta^{j+1} = \operatorname*{arg\,min}_{\zeta\in\mathbb{R}^p} \Phi_j(\zeta) \quad \text{and} \quad \eta^{j+1} = \Pi_\Lambda\big(\widetilde{Z}^{\mathbb{T}}\zeta^{j+1} + \beta^j/\mu_j\big).$$

By the strong convexity of Φ_j, $\zeta^{j+1} = \arg\min_{\zeta\in\mathbb{R}^p} \Phi_j(\zeta)$ iff ζ^{j+1} satisfies

$$\nabla\Phi_j(\zeta) = \widetilde{y} + \zeta + \mu_j\widetilde{Z}\left[\left(\widetilde{Z}^{\mathbb{T}}\zeta + \beta^j/\mu_j\right) - \Pi_\Lambda\left(\widetilde{Z}^{\mathbb{T}}\zeta + \beta^j/\mu_j\right)\right] = 0. \tag{S3.4}$$

The system (S3.4) is strongly semismooth (see the related discussion in Mifflin (1977); Qi and Sun (1993)), and we apply the semismooth Newton method for solving it. Write $h := \widetilde{Z}^{\mathbb{T}}\zeta + \beta^j/\mu_j$. By (Clarke, 1983, Proposition 2.3.3 and Theorem 2.6.6), the Clarke Jacobian $\partial\nabla\Phi_j$ satisfies

$$\partial(\nabla\Phi_j)(\zeta) \subseteq \widehat{\partial}^2\Phi_j(\zeta) := I + \mu_j\widetilde{Z}\big(I - \partial\Pi_\Lambda(h)\big)\widetilde{Z}^{\mathbb{T}} \tag{S3.5}$$

where $\widehat{\partial}^2\Phi_j$ is the generalized Hessian of Φ_j at ζ. Since the exact characterization of $\partial\nabla\Phi_j$ is difficult to obtain, we replace $\partial\nabla\Phi_j$ with $\widehat{\partial}^2\Phi_j$ in the solution of (S3.4). Let $W \in \partial\Pi_\Lambda(h)$. By (Clarke, 1983, Theorem 2.6.6), $W = \mathrm{Diag}(\varpi_1,\ldots,\varpi_p)$ with $\varpi_i \in \partial\Pi_{\Lambda_i}(h_i)$ where

$$\partial\Pi_{\Lambda_i}(h_i) = \begin{cases} \{1\} & \text{if } |h_i| < \omega_i; \\ [0,1] & \text{if } |h_i| = \omega_i; \\ \{0\} & \text{if } |h_i| > \omega_i. \end{cases}$$

From the last two equations, each element in $\widehat{\partial}^2\Phi_j(\zeta)$ is positive definite, which by Qi and Sun (1993) implies that the following semismooth Newton method has a fast convergence rate.

It is worthwhile to point out that due to the special structure of V^l, the computation work of solving the linear system (S3.6) is tiny; see the discussion in (Li, Sun and Toh, 2018, Section 3.3). During the implementation of the semismooth Newton ALM, we terminated the iterates of Algorithm 2 when $\max\{\epsilon^j_{\rm pinf},\epsilon^j_{\rm dinf},\epsilon^j_{\rm gap}\} \le \epsilon^j$, where $\epsilon^j_{\rm gap}$ is the primal-dual gap, i.e., the sum of the objective values of (S3.1) and (S3.2) at (β^j,ζ^j,η^j), and $\epsilon^j_{\rm pinf}$ and $\epsilon^j_{\rm dinf}$ are the primal and dual infeasibility measure at (β^j,ζ^j,η^j). By comparing the optimality condition of (S3.3) with that of (S3.2), we defined

$$\epsilon^j_{\rm pinf} := \frac{\|\nabla\Phi_j(\zeta^j)\|}{1+\|\widetilde{y}\|} \ \text{ and } \ \epsilon^j_{\rm dinf} := \frac{\|\beta^j-\beta^{j-1}\|}{\mu_{j-1}(1+\|\widetilde{y}\|)}.$$

Algorithm 3 A semismooth Newton-CG algorithm for (S3.4)

Initialization: Choose $\vartheta, \varsigma, \delta \in (0,1)$, $\varrho \in (0, \frac{1}{2})$ and $\zeta^0 \in \mathbb{R}^p$. Set $l = 0$.

while the stopping conditions are not satisfied **do**

1. Choose a matrix $V^l \in \widehat{\partial}^2 \Phi_j(\zeta^l)$. Solve the following linear system

$$V^l d = -\nabla \Phi_j(\zeta^l) \tag{S3.6}$$

with the conjugate gradient (CG) algorithm to find d^l such that

$$\|V^l d^l + \nabla \Phi_j(\zeta^l)\| \leq \min(\vartheta, \|\nabla \Phi_j(\zeta^l)\|^{1+\varsigma}).$$

2. Set $\alpha_l = \delta^{m_l}$, where m_l is the first nonnegative integer m for which

$$\Phi_j(\zeta^l + \delta^m d^l) \leq \Phi_j(\zeta^l) + \varrho \delta^m \langle \nabla \Phi_j(\zeta^l), d^l \rangle.$$

3. Set $\zeta^{l+1} = \zeta^l + \alpha_l d^l$ and $l \leftarrow l+1$, and then go to Step 1.

end while

We adopted a stopping criteria similar to those in Li, Sun and Toh (2018):

$$\|\nabla \Phi_j(\zeta^{j+1})\| \leq \delta_j \min\left(0.1, \max(\epsilon^j_{\text{dinf}}, \epsilon^j_{\text{gap}})\right) \quad \text{with} \quad \textstyle\sum_{j=0}^{\infty} \delta_j < \infty.$$

S4. ADMM Algorithm for CoCoLasso

This part includes our implementation for CoCoLasso (a convex conditioned Lasso of Datta and Zou (2017)). They first solved the following PSD optimization problem

$$\overline{\Sigma} \in \mathop{\arg\min}_{W \succeq \widehat{\epsilon} I} \|W - \widehat{\Sigma}\|_{\max} \ \text{ for some } \widehat{\epsilon} > 0. \tag{S4.1}$$

When the optimal solution $\overline{\Sigma}$ of (S4.1) is available, one may apply the semismooth Newton ALM in Section S3 for solving

$$\overline{\beta} = \mathop{\arg\min}_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|\overline{y} - \overline{Z}\beta\|^2 + \lambda \|\beta\|_1 \right\} \tag{S4.2}$$

with the Cholesky factor $\overline{Z}/\sqrt{n}$ of $\overline{\Sigma}$ and the vector $\overline{y}$ satisfying $\overline{Z}^{\mathbb{T}}\overline{y} = Z^{\mathbb{T}}y$. Therefore, we here focus on the computation of $\overline{\Sigma}$. The problem (S4.1) can be equivalently written as

$$\min_{W, B \in \mathbb{S}^p} \left\{ \|B\|_{\max} \colon W - B = \widehat{\Sigma},\, W \succeq \widehat{\epsilon} I \right\}, \tag{S4.3}$$

whose dual, after an elementary calculation, takes the following form

$$\min_{Y \in \mathbb{S}^p_+ \cap \mathbb{B}} \left\langle Y, \widehat{\Sigma} - \widehat{\epsilon} I \right\rangle \ \text{ with } \ \mathbb{B} := \left\{ Y \in \mathbb{S}^p \colon \|Y\|_1 \le 1 \right\}. \tag{S4.4}$$

Here, $\|Y\|_1$ means the elementwise ℓ_1-norm of Y. Different from Datta and Zou (2017), we use the ADMM with a large step-size $\tau \in (1, \frac{\sqrt{5}+1}{2})$ instead of the unit one to solve (S4.3). From the numerical results in Sun, Yang and Toh (2016), the ADMM with a larger step-size has better performance. For a given $\mu > 0$, define the augmented Lagrangian function of (S4.3) by

$$L_\mu(W,B;\Gamma) := \|B\|_{\max} + \langle W - B - \widehat{\Sigma}, \Gamma\rangle + (\mu/2)\|W - B - \widehat{\Sigma}\|_F^2.$$

The iterations of the ADMM for (S4.3) with a step-size are as follows.

Algorithm 4 ADMM for solving the problem (S4.3)

Initialization: Choose $\mu > 0, \tau \in (1, \frac{\sqrt{5}+1}{2})$ and (W^0, B^0, Γ^0). Set $k = 0$.

while the stopping conditions are not satisfied **do**

1. Compute the following strongly convex minimization problem

$$W^{k+1} = \operatorname*{arg\,min}_{W \succeq \widehat{\epsilon} I} L_\mu(W, B^k; \Gamma^k). \tag{S4.5}$$

2. Compute the following strongly convex minimization problem

$$B^{k+1} = \operatorname*{arg\,min}_{B \in \mathbb{S}^p} L_\mu(W^{k+1}, B; \Gamma^k). \tag{S4.6}$$

3. Update the multiplier by the formula

$$\Gamma^{k+1} = \Gamma^k + \tau\mu(W^{k+1} - B^{k+1} - \widehat{\Sigma}).$$

4. Set $k \leftarrow k + 1$, and then go to Step 1.

end while

Due to the speciality of the constraint $W - B = \widehat{\Sigma}$, the convergence of Algorithm 4 can be directly obtained from (Fazel et al., 2013, Theorem B.1) with $S=T=0$. By the expression of $L_{\mu}(W,B;\Gamma)$, it holds that

$$W^{k+1}=\widehat{\epsilon}I+\Pi_{\mathbb{S}_{+}^{n}}\big(B^{k}-\mu^{-1}\Gamma^{k}+\widehat{\Sigma}-\widehat{\epsilon}I\big),$$
$$B^{k+1}=(W^{k+1}+\mu^{-1}\Gamma^{k}-\widehat{\Sigma})-\Pi_{\mu^{-1}\mathbb{B}}\big(W^{k+1}+\mu^{-1}\Gamma^{k}-\widehat{\Sigma}\big) \qquad \text{(S4.7)}$$

where the equality (S4.7) is obtained from $\text{prox}_{f^*}(G)+\text{prox}_{f}(G)=G$ with $\text{prox}_{f}(G):=\arg\min_{B\in\mathbb{S}^p}\big\{\frac{1}{2}\|B-G\|_F^2+f(B)\big\}$ for $f(B):=\mu^{-1}\|B\|_{\max}$. Just like Datta and Zou (2017), we use the algorithm proposed in Duchi et al. (2008) to compute the projection involved in (S4.7).

During our implementation of Algorithm 4, we adjust μ dynamically by the ratio of the primal and dual infeasibility. By the optimality conditions of (S4.3) and (S4.5)-(S4.6), we measure the primal and dual infeasibility and the dual gap at $(W^{k+1},B^{k+1},\Gamma^{k+1})$ in terms of $\epsilon_{\rm pinf}^k,\epsilon_{\rm dinf}^k$ and $\epsilon_{\rm gap}^k$, where

$$\epsilon_{\rm pinf}^k:=\frac{\|\mu(B^{k+1}-B^k)+(\tau^{-1}-1)(\Gamma^{k+1}-\Gamma^k)\|_F}{1+\|\widehat{\Sigma}\|_F},$$
$$\epsilon_{\rm dinf}^k:=\frac{\|\Gamma^{k+1}-\Gamma^k\|_F}{\tau\mu(1+\|\widehat{\Sigma}\|_F)}\ \ \text{and}\ \ \epsilon_{\rm gap}^k:=\frac{|\|B^{k+1}\|_{\max}+\langle\Gamma^{k+1},\widehat{\Sigma}-\widehat{\epsilon}I\rangle|}{\max(1,0.5(|\Gamma^{k+1}|+|\langle\Gamma^{k+1},\widehat{\Sigma}-\widehat{\epsilon}I\rangle|))}.$$

References

Clarke, F. H. (1983). *Optimization and Nonsmooth Analysis.* New York: John Wiley and Sons.

Datta, A. and Zou, H. (2017). *CoCoLASSO for high-dimensional error-in-variables regression. The Annals of Statistics 45*, pp. 2400–2426.

Duchi, J., Shalev-Shwartz, S., Singer, Y. and Chandra T. (2008). Efficient projections onto the L_1-ball for learning in high-dimensions. *In Proceedings of the 25th International Conference on Machine Learning*, pp. 272–279.

Fazel, M., Pong, T. K., Sun, D. F. and Tseng, P. (2013). Hankel matrix rank minimization with applications in system identification and realization. *SIAM Journal on Matrix Analysis and Applications 34*, pp. 946–977.

Horn, R. A. and Johnson, C. R. (1990). *Matrix Analysis* (2 ed.). New York: Cambridge University Press.

Horn, R. A. and Johnson, C. R. (1991). *Topics in Matrix Analysis*. New York: Cambridge University Press.

Li, X. D., Sun, D. F. and Toh, K.-C. (2018). A highly efficient semismooth Newton augmented Lagrangian method for solving Lasso problems. *SIAM Journal on Optimization 28*, pp. 433–458.

Liu, J., Ji, S. W. and Ye, J. P. (2011). SLEP: Sparse Learning with Efficient Projections. *Arizona State University. URL: http://www.public.asu.edu/jye02/Software/SLEP*.

Loh, P. L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics 40*, pp. 1637–1664.

Loh, P. L. (2014). High-dimensional statistics with systematically corrupted data. *University of California, PhD thesis, http://escholarship.org/uc/item/8j49c5n4*.

Mifflin, R. (1977). Semismooth and semiconvex functions in constrained optimization. *SIAM Journal on Control and Optimization 15*, pp. 959–972.

Nesterov, Y. (2013). Gradient methods for minimizing composite objective function. *Mathematical Programming 140*, pp. 125–161.

Qi, L. and Sun, J. (1993). A nonsmooth version of Newton's method. *Mathematical Programming 58*, pp. 353–367.

Sun, D. F., Yang, L. Q. and Toh, K.-C. (2016). An efficient inexact ABCD method for least squares semidefinite programming. *SIAM Journal on Optimization 26*, pp. 1072–1100.