Statistica Sinica

A proximal dual semismooth Newton method for zero-norm penalized quantile regression estimator

Dongdong Zhang, Shaohua Pan and Shujun Bi

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

Supplementary Materials

Appendix A

This part includes some preliminary knowledge on generalized subdifferentials and Clarke Jacobian, and some lemmas used in Section 2-5. First, we recall from (Rockafellar and Wets, 1998, Definition 8.3) the notion of the subdifferential of an extended real-valued function.

Definition 1. Consider a function $f\colon \mathbb{R}^p \to (-\infty, +\infty]$ and $x \in \mathrm{dom} f$. The regular subdifferential of f at x, denoted by $\widehat{\partial} f(x)$, is defined as

$$\widehat{\partial} f(x) := \Big\{ v \in \mathbb{R}^p \mid \liminf_{\substack{x' \to x \\ x' \neq x}} \frac{f(x') - f(x) - \langle v, x' - x\rangle}{\|x' - x\|} \geq 0 \Big\};$$

and the (limiting) subdifferential of f at x, denoted by $\partial f(x)$, is defined as

$$\partial f(x) := \Big\{ v \in \mathbb{R}^p \mid \exists\, x^k \to x \text{ with } f(x^k) \to f(x) \text{ and } v^k \in \widehat{\partial} f(x^k) \text{ with } v^k \to v \Big\}.$$

Remark 1. At each $x \in \mathrm{dom} f$, $\widehat{\partial} f(x)$ and $\partial f(x)$ are closed and satisfy $\widehat{\partial} f(x) \subseteq \partial f(x)$, and the set $\widehat{\partial} f(x)$ is convex but $\partial f(x)$ is generally nonconvex. When f is convex, $\widehat{\partial} f(x) = \partial f(x)$ and is precisely the subdifferential of f at x in the sense of convex analysis Rockafellar (1970).

Definition 2. (see Clarke (1983)) Let $H\colon \Omega \to \mathbb{R}^n$ be a locally Lipschitz continuous mapping defined on an open set $\Omega \subseteq \mathbb{R}^p$. Denote by $D_H \subseteq \Omega$ the set of points where H is differentiable and by $H'(z) \in \mathbb{R}^{n\times p}$ the Jacobian of H at $z \in D_H$. The Clarke Jacobian of H at $\overline{z} \in \Omega$ is

$$\partial_C H(\overline{z}) := \mathrm{conv}\Big\{ \lim_{k\to\infty} H'(z^k) \mid \{z^k\} \subseteq D_H \text{ with } \lim_{k\to\infty} z^k = \overline{z}\Big\}.$$

Generally, it is not easy to characterize the Clarke Jacobian of a locally Lipschitz mapping. The following lemmas provide such a characterization for the proximal mappings of the weighted ℓ_1-norm and the check loss function.

Lemma 1. *For a given $\omega \in \mathbb{R}^p_+$, let $h(x) := \|\omega \circ x\|_1$ for $x \in \mathbb{R}^p$. Then,*

$$\mathcal{P}_{\gamma^{-1}h}(z) = \mathrm{sign}(z)\max\big(|z| - \gamma^{-1}w, 0\big) \quad \forall z \in \mathbb{R}^p,$$

$$\partial_C(\mathcal{P}_{\gamma^{-1}h})(z) = \big\{\mathrm{Diag}(v_1,\ldots,v_n) \,|\, v_i = 1 \text{ if } |\gamma z_i| > \omega_i, \text{ otherwise } v_i \in [0,1]\big\}.$$

Lemma 2. *For any given $\tau \in (0,1)$, let θ_τ and f_τ be the function defined as in (2.2). Then, for any given $\gamma > 0$ and $z \in \mathbb{R}^p$, it holds that*

$$\big[\mathcal{P}_{\gamma^{-1}f_\tau}(z)\big]_i = \max\Big(\max\Big(z_i - \frac{\tau}{n\gamma}, 0\Big), \frac{\tau-1}{n\gamma} - z_i\Big) \quad \text{for } i = 1,2,\ldots,p$$

and $\partial_C(\mathcal{P}_{\gamma^{-1}f_\tau})(y) = \big\{\mathrm{Diag}(v_1,\ldots,v_n) \mid v_i \in \partial_C\big[\mathcal{P}_{\gamma^{-1}}(n^{-1}\theta_\tau)\big](z_i)\big\}$ with

$$\partial_C\big[\mathcal{P}_{\gamma^{-1}}(n^{-1}\theta_\tau)\big](t) = \begin{cases} \{1\} & \text{if } t > \frac{\tau}{n\gamma} \text{ or } t < \frac{\tau-1}{n\gamma}; \\ [0,1] & \text{if } t = \frac{\tau}{n\gamma} \text{ or } \frac{\tau-1}{n\gamma}; \\ \{0\} & \text{if } \frac{\tau-1}{n\gamma} < t < \frac{\tau}{n\gamma}. \end{cases} \tag{01}$$

To close this part, we show that under a mild condition, the zero-norm regularized composite problem has a nonempty global optimal solution set.

Lemma 3. *Let $A \in \mathbb{R}^{n\times p}$ and $b \in \mathbb{R}^n$ be given, and let $g\colon \mathbb{R}^n \to \mathbb{R}$ be an lsc coercive function with $\inf_{z\in\mathbb{R}^n} g(z) > -\infty$. Then, for any given $\nu > 0$, the zero-norm composite problem*

$$\min_{x\in\mathbb{R}^p} \Big\{\nu g(b-Ax) + \|x\|_0\Big\} \tag{02}$$

has a nonempty global optimal solution set.

Proof. Notice that the objective function of (02) is lower bounded. So, it has an infimum, say α^*. Then there exists a sequence $\{x^k\} \subset \mathbb{R}^p$ such that

$$\nu g(b-Ax^k) + \|x^k\|_0 \le \alpha^* + 1/k \ \text{ for each } k. \tag{03}$$

If $\{x^k\}$ is bounded, then by letting $\overline{x}$ be an arbitrary limit point of $\{x^k\}$ and using the lsc of $x \mapsto g(b-A\,x)$ and $\|\cdot\|_0$, we have $\nu g(b-A\overline{x}) + \|\overline{x}\|_0 \le \alpha^*$. This shows that $\overline{x}$ is a global optimal solution of the problem (02). Next we consider the case that $\{x^k\}$ is unbounded. Define

$$J := \big\{i \in \{1,\ldots,p\} \mid \{x_i^k\} \text{ is unbounded}\big\} \ \text{ and } \ \overline{J} := \{1,\ldots,p\}\backslash J.$$

Along with (03), it immediately follows that for all sufficiently large k,

$$\nu g(b-Ax^k) + |J| + \|x^k_{\overline{J}}\|_0 \le \alpha^* + 1/k. \tag{04}$$

This, by the coerciveness of g, means that there is a bounded sequence $\{z^k\} \subset \mathbb{R}^n$ such that $z^k = b - Ax^k$. Clearly, $A_J x^k_J = b - z^k - A_{\overline{J}} x^k_{\overline{J}}$. Notice that $\{z^k\}$ and $\{x^k_{\overline{J}}\}$ are bounded. We may assume (taking a subsequence if necessary) that $\{z^k\}$ and $\{x^k_{\overline{J}}\}$ are convergent, say, $z^k \to z^*$ and $x^k_{\overline{J}} \to \xi^*$. Notice that for each k, x^k_J is a solution of the system $A_J y = b - z^k - A_{\overline{J}} x^k_{\overline{J}}$, that is, $\{b - z^k - A_{\overline{J}} x^k_{\overline{J}}\} \subset A_J(\mathbb{R}^{|J|})$. Together with the closedness of the set $A_J(\mathbb{R}^{|J|})$, it follows that $b - z^* - A_{\overline{J}}\xi^* \in A_J(\mathbb{R}^{|J|})$. So, there exists $u^* \in \mathbb{R}^{|J|}$ such that $A_J u^* = b - z^* - A_{\overline{J}}\xi^*$, i.e., $A_J u^* + A_{\overline{J}}\xi^* - z^* = b$. Taking the limit to the both sides of (04) and using $b - Ax^k = z^k$ gives

$$\nu g(z^*) + |J| + \|\xi^*\|_0 \le \alpha^*.$$

Together with $\nu g(b - A_J u^* - A_{\overline{J}}\xi^*) + \|u^*\|_0 + \|\xi^*\|_0 \le \nu g(z^*) + |J| + \|\xi^*\|_0$, we conclude that $(u^*;\xi^*)$ is a global optimal solution of the zero-norm composite problem (02). □

Appendix B

In this part, for each $k \in \mathbb{N}$ we write $v^k := e - w^k$ and $z^k := y - X\beta^k$. To present the proof of Theorem 2, we need the following technical lemma.

Lemma 4. *Suppose that Assumption 1 holds and for some $k \ge 1$ there exists $S^{k-1} \supseteq S^*$ with $\max_{i\in(S^{k-1})^c} w_i^{k-1} \le \frac{1}{2}$. Then, when $\lambda \ge 16\overline{\tau}n^{-1}\|X\|_1 + 8r_k$,*

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

Proof. By the approximate optimality of β^k to (3.1) and Remark 1(iv),

$$f_\tau(y - X\beta^*) + \lambda\langle v^{k-1}, |\beta^*|\rangle \ge f_\tau(y - X\beta^k) + \lambda\langle v^{k-1}, |\beta^k|\rangle + \langle \delta^k, \beta^* - \beta^k\rangle$$

which, after a suitable rearrangement, takes the following form

$$f_\tau(y - X\beta^k) - f_\tau(y - X\beta^*) + \langle \delta^k, \beta^* - \beta^k\rangle \le \lambda\langle v^{k-1}, |\beta^*| - |\beta^k|\rangle. \tag{05}$$

Recall that $\varepsilon = y - X\beta^*$ and $\|\varepsilon\|_\infty > 0$. We define the following index sets

$$\mathcal{I} := \big\{i \in \{1,\ldots,n\} \colon \varepsilon_i \ne 0\big\} \ \text{ and } \ \mathcal{J}_k := \big\{i \notin \mathcal{I} \colon z_i^k \ne 0\big\}. \tag{06}$$

By the expression of f_τ and $\theta_\tau(0) = 0$, with the index sets $\mathcal{I}$ and $\mathcal{J}_k$,

$$\begin{aligned} f_\tau(y - X\beta^k) - f_\tau(y - X\beta^*) &= \frac{1}{n}\sum_{i=1}^n [\theta_\tau(z_i^k) - \theta_\tau(\varepsilon_i)] \\ &= \frac{1}{n}\Big[\sum_{i\in\mathcal{J}_k} \frac{\theta_\tau^2(z_i^k) - \theta_\tau^2(\varepsilon_i)}{\theta_\tau(z_i^k) + \theta_\tau(\varepsilon_i)} + \sum_{i\in\mathcal{I}} \frac{\theta_\tau^2(z_i^k) - \theta_\tau^2(\varepsilon_i)}{\theta_\tau(z_i^k) + \theta_\tau(\varepsilon_i)}\Big] \\ &\ge \frac{1}{n}\Big[\sum_{i\in\mathcal{J}_k} \frac{\theta_\tau^2(z_i^k) - \theta_\tau^2(\varepsilon_i)}{\overline{\tau}\|z^k\|_\infty} + \sum_{i\in\mathcal{I}} \frac{\theta_\tau^2(z_i^k) - \theta_\tau^2(\varepsilon_i)}{\theta_\tau(z_i^k) + \theta_\tau(\varepsilon_i)}\Big]. \end{aligned} \tag{07}$$

Notice that θ_τ^2 is smooth and strongly convex of modulus $2\underline{\tau}^2$. For each i,

$$\theta_\tau^2(z_i^k)-\theta_\tau^2(\varepsilon_i)\geq 2(\tau-\mathbb{I}_{\mathbb{R}_-}(\varepsilon_i))^2\varepsilon_i(z_i^k-\varepsilon_i)+\underline{\tau}^2(z_i^k-\varepsilon_i)^2. \tag{08}$$

This implies that $\theta_\tau^2(z_i^k)-\theta_\tau^2(\varepsilon_i)\geq\underline{\tau}^2(z_i^k-\varepsilon_i)^2$ for each $i\in\mathcal{J}_k$, and then

$$\sum_{i\in\mathcal{J}_k}\frac{\theta_\tau^2(z_i^k)-\theta_\tau^2(\varepsilon_i)}{\overline{\tau}\|z^k\|_\infty}\geq\frac{\underline{\tau}^2}{\overline{\tau}}\sum_{i\in\mathcal{J}_k}\frac{(z_i^k-\varepsilon_i)^2}{\|z^k\|_\infty}. \tag{09}$$

For each $i\in\mathcal{I}$, write $\widetilde{z}_i^k:=\frac{2(\tau-\mathbb{I}_{\mathbb{R}_-}(\varepsilon_i))^2\varepsilon_i}{\theta_\tau(z_i^k)+\theta_\tau(\varepsilon_i)}$. From (08), it follows that

$$\begin{aligned}\sum_{i\in\mathcal{I}}\frac{\theta_\tau^2(z_i^k)-\theta_\tau^2(\varepsilon_i)}{\theta_\tau(z_i^k)+\theta_\tau(\varepsilon_i)}&\geq\sum_{i\in\mathcal{I}}\widetilde{z}_i^k(z_i^k-\varepsilon_i)+\underline{\tau}^2\sum_{i\in\mathcal{I}}\frac{(z_i^k-\varepsilon_i)^2}{\theta_\tau(z_i^k)+\theta_\tau(\varepsilon_i)}\\&\geq-\|\widetilde{z}^k\|_\infty\|X(\beta^k-\beta^*)\|_1+\underline{\tau}^2\sum_{i\in\mathcal{I}}\frac{(z_i^k-\varepsilon_i)^2}{\overline{\tau}(\|z^k\|_\infty+\|\varepsilon\|_\infty)}\\&\geq-2\overline{\tau}\big\|X(\beta^k-\beta^*)\big\|_1+\frac{\underline{\tau}^2}{\overline{\tau}}\sum_{i\in\mathcal{I}}\frac{(z_i^k-\varepsilon_i)^2}{\|z^k\|_\infty+\|\varepsilon\|_\infty}\end{aligned} \tag{010}$$

where the second inequality is by $\theta_\tau(z_i^k)\leq\overline{\tau}\|z^k\|_\infty$ for $i\in\mathcal{I}$, and the last one is since $|\widetilde{z}_i^k|\leq\frac{2(\tau-\mathbb{I}_{\mathbb{R}_-}(\varepsilon_i))^2|\varepsilon_i|}{\theta_\tau(\varepsilon_i)}\leq 2\overline{\tau}$ for each $i\in\mathcal{I}$. Substituting the inequalities (09)-(010) into (07), we obtain that

$$\begin{aligned}f_\tau(y-X\beta^k)-f_\tau(y-X\beta^*)&\geq\frac{\underline{\tau}^2}{n\overline{\tau}}\sum_{i\in\mathcal{J}_k\cup\mathcal{I}}\frac{(z_i^k-\varepsilon_i)^2}{\|z^k\|_\infty+\|\varepsilon\|_\infty}-\frac{2\overline{\tau}}{n}\|X(\beta^k-\beta^*)\|_1\\&=\frac{\underline{\tau}^2\|X(\beta^k-\beta^*)\|^2}{n\overline{\tau}(\|z^k\|_\infty+\|\varepsilon\|_\infty)}-\frac{2\overline{\tau}}{n}\|X(\beta^k-\beta^*)\|_1.\end{aligned}$$

Combining this inequality and (05) and recalling that $\|\delta^k\|\leq r_k$, we get

$$\begin{aligned}\frac{\underline{\tau}^2\|X(\beta^k-\beta^*)\|^2}{n\overline{\tau}(\|z^k\|_\infty+\|\varepsilon\|_\infty)}&\leq\lambda\langle v^{k-1},|\beta^*|-|\beta^k|\rangle+\frac{2\overline{\tau}}{n}\big\|X(\beta^k-\beta^*)\big\|_1+\langle\delta^k,\beta^k-\beta^*\rangle\\&\leq\lambda\Big(\textstyle\sum_{i\in S^*}v_i^{k-1}|\Delta\beta_i^k|-\sum_{i\in(S^{k-1})^c}v_i^{k-1}|\Delta\beta_i^k|\Big)\\&\quad+\big(2n^{-1}\overline{\tau}\|X\|_1+r_k\big)\|\beta^k-\beta^*\|_1\\&=\lambda\Big(\textstyle\sum_{i\in S^*}v_i^{k-1}|\Delta\beta_i^k|-\sum_{i\in(S^{k-1})^c}v_i^{k-1}|\Delta\beta_i^k|\Big)\\&\quad+\big(2n^{-1}\overline{\tau}\|X\|_1+r_k\big)\big(\|\Delta\beta_{S^{k-1}}^k\|_1+\|\Delta\beta_{(S^{k-1})^c}^k\|_1\big)\end{aligned} \tag{011}$$

Since $S^{k-1} \supset S^*$ and $v_i^{k-1} \in [0.5, 1]$ for $i \in (S^{k-1})^c$, from the last inequality,

$$\begin{aligned}\frac{\underline{\tau}^2\|X(\beta^k-\beta^*)\|^2}{n\overline{\tau}(\|z^k\|_\infty+\|\varepsilon\|_\infty)} &\le \textstyle\sum_{i\in S^{k-1}}\big(\lambda v_i^{k-1}+2n^{-1}\overline{\tau}\|X\|_1+r_k\big)\big|\Delta\beta_i^k\big| \\ &\quad+\textstyle\sum_{i\in(S^{k-1})^c}\big(2n^{-1}\overline{\tau}\|X\|_1+r_k-\lambda/2\big)\big|\Delta\beta_i^k\big| \\ &\le \big(\lambda+2n^{-1}\overline{\tau}\|X\|_1+r_k\big)\big\|\Delta\beta_{S^{k-1}}^k\big\|_1 \\ &\quad+\big(2n^{-1}\overline{\tau}\|X\|_1+r_k-\lambda/2\big)\big\|\Delta\beta_{(S^{k-1})^c}^k\big\|_1.\end{aligned}$$

By the nonnegativity of the left hand side and the given assumption on λ,

$$\big\|\Delta\beta_{(S^{k-1})^c}^k\big\|_1 \le \frac{\lambda+2n^{-1}\overline{\tau}\|X\|_1+r_k}{0.5\lambda-2n^{-1}\overline{\tau}\|X\|_1-r_k}\big\|\Delta\beta_{S^{k-1}}^k\big\|_1 \le 3\big\|\Delta\beta_{S^{k-1}}^k\big\|_1.$$

The desired result follows. The proof is then completed. □

Lemma 5. *Suppose that Assumption 1 holds, that X satisfies the κ-RSC over $\mathcal{C}(S^*)$, and that for some $k \ge 1$ there exists an index set S^{k-1} with $|S^{k-1}| \le 1.5s^*$ such that $S^{k-1} \supseteq S^*$ and $\max_{i\in(S^{k-1})^c} w_i^{k-1} \le \frac{1}{2}$. Then, when $16\overline{\tau}n^{-1}\|X\|_1+8r_k \le \lambda < \frac{\underline{\tau}^2\kappa-2\overline{\tau}\|X\|_{\max}(2n^{-1}\overline{\tau}\|X\|_1+r_k)|S^{k-1}|}{2\overline{\tau}\|X\|_{\max}\|v_{S^*}^{k-1}\|_\infty|S^{k-1}|}$,*

$$\big\|\Delta\beta^k\big\| \le \frac{\overline{\tau}\big(\lambda\|v_{S^*}^{k-1}\|_\infty+2n^{-1}\overline{\tau}\|X\|_1+r_k\big)\sqrt{|S^{k-1}|}\|\varepsilon\|_\infty}{\underline{\tau}^2\kappa-2\overline{\tau}\|X\|_{\max}\big(\lambda\|v_{S^*}^{k-1}\|_\infty+2n^{-1}\overline{\tau}\|X\|_1+r_k\big)|S^{k-1}|}.$$

Proof. Notice that $\|z^k\|_\infty+\|\varepsilon\|_\infty \le \|X\Delta\beta^k\|_\infty+2\|\varepsilon\|_\infty$. So, we have

$$\frac{\underline{\tau}^2\|X(\beta^k-\beta^*)\|^2}{n\overline{\tau}(\|z^k\|_\infty+\|\varepsilon\|_\infty)} \ge \frac{\underline{\tau}^2\|X\Delta\beta^k\|^2}{n\overline{\tau}(\|X\Delta\beta^k\|+2\|\varepsilon\|_\infty)}.$$

Together with (011) and $v_i^{k-1} \in [0.5, 1]$ for $i \in (S^{k-1})^c$, it follows that

$$\begin{aligned}\frac{\underline{\tau}^2\|X\Delta\beta^k\|^2}{n\overline{\tau}(\|X\Delta\beta^k\|_\infty+2\|\varepsilon\|_\infty)} &\le \lambda\sum_{i\in S^*}v_i^{k-1}|\Delta\beta_i^k|-\frac{\lambda}{2}\sum_{i\in(S^{k-1})^c}|\Delta\beta_i^k| \\ &\quad+\big(2n^{-1}\overline{\tau}\|X\|_1+r_k\big)\big(\|\Delta\beta_{S^{k-1}}^k\|_1+\|\Delta\beta_{(S^{k-1})^c}^k\|_1\big) \\ &\le \Big(\lambda\|v_{S^*}^{k-1}\|_\infty+2n^{-1}\overline{\tau}\|X\|_1+r_k\Big)\|\Delta\beta_{S^{k-1}}^k\|_1\end{aligned}$$

where the last inequality is due to $\lambda > 16n^{-1}\overline{\tau}\|X\|_1 + 8r_k$. By Lemma 4, $\|\Delta\beta^k_{(S^{k-1})^c}\|_1 \leq 3\|\Delta\beta^k_{S^{k-1}}\|_1$. By the given assumption, $\Delta\beta^k \in \mathcal{C}(S^*)$. From the κ-RSC property of X on $\mathcal{C}(S^*)$, it follows that $\|X\Delta\beta^k\|^2 \geq 2n\kappa\|\Delta\beta^k\|^2$. Then, we obtain

$$\frac{2\underline{\tau}^2\kappa\|\Delta\beta^k\|^2}{\overline{\tau}\big(\|X\Delta\beta^k\|_\infty + 2\|\varepsilon\|_\infty\big)} \leq \Big(\lambda\|v^{k-1}_{S^*}\|_\infty + \frac{2\overline{\tau}\|X\|_1}{n} + r_k\Big)\big\|\Delta\beta^k_{S^{k-1}}\big\|_1.$$

Multiplying this inequality with $\overline{\tau}\big(\|X\Delta\beta^k\|_\infty + 2\|\varepsilon\|_\infty\big)$ yields that

$$\begin{aligned}
2\underline{\tau}^2\kappa\|\Delta\beta^k\|^2 &\leq \overline{\tau}\big(\|X\Delta\beta^k\|_\infty + 2\|\varepsilon\|_\infty\big)\Big(\lambda\|v^{k-1}_{S^*}\|_\infty + \frac{2\overline{\tau}\|X\|_1}{n} + r_k\Big)\big\|\Delta\beta^k_{S^{k-1}}\big\|_1 \\
&\leq \overline{\tau}\|X\Delta\beta^k\|_\infty\Big(\lambda\|v^{k-1}_{S^*}\|_\infty + 2n^{-1}\overline{\tau}\|X\|_1 + r_k\Big)\big\|\Delta\beta^k_{S^{k-1}}\big\|_1 \\
&\quad + 2\overline{\tau}\|\varepsilon\|_\infty\Big(\lambda\|v^{k-1}_{S^*}\|_\infty + 2n^{-1}\overline{\tau}\|X\|_1 + r_k\Big)\big\|\Delta\beta^k_{S^{k-1}}\big\|_1.
\end{aligned}$$

Since $\|X\Delta\beta^k\|_\infty \leq \|X\|_{\max}\|\Delta\beta^k\|_1$, along with $\|\Delta\beta^k_{(S^{k-1})^c}\|_1 \leq 3\|\Delta\beta^k_{S^{k-1}}\|_1$, we have $\|X\Delta\beta^k\|_\infty \leq 4\|X\|_{\max}\|\Delta\beta^k_{S^{k-1}}\|_1$. Thus, from the last inequality,

$$\begin{aligned}
2\underline{\tau}^2\kappa\|\Delta\beta^k\|^2 &\leq 4\overline{\tau}\|X\|_{\max}\Big(\lambda\|v^{k-1}_{S^*}\|_\infty + 2n^{-1}\overline{\tau}\|X\|_1 + r_k\Big)\big\|\Delta\beta^k_{S^{k-1}}\big\|_1^2 \\
&\quad + 2\overline{\tau}\Big(\lambda\|v^{k-1}_{S^*}\|_\infty + 2n^{-1}\overline{\tau}\|X\|_1 + r_k\Big)\big\|\Delta\beta^k_{S^{k-1}}\big\|_1\|\varepsilon\|_\infty \\
&\leq 4\overline{\tau}\|X\|_{\max}\Big(\lambda\|v^{k-1}_{S^*}\|_\infty + \frac{2\overline{\tau}\|X\|_1}{n} + r_k\Big)|S^{k-1}|\big\|\Delta\beta^k_{S^{k-1}}\big\|^2 \\
&\quad + 2\overline{\tau}\Big(\lambda\|v^{k-1}_{S^*}\|_\infty + 2n^{-1}\overline{\tau}\|X\|_1 + r_k\Big)\sqrt{|S^{k-1}|}\big\|\Delta\beta^k_{S^{k-1}}\big\|\|\varepsilon\|_\infty \\
&\leq 4|S^{k-1}|\overline{\tau}\|X\|_{\max}\Big(\lambda\|v^{k-1}_{S^*}\|_\infty + \frac{2\overline{\tau}\|X\|_1}{n} + r_k\Big)\big\|\Delta\beta^k\big\|^2 \\
&\quad + 2\overline{\tau}\Big(\lambda\|v^{k-1}_{S^*}\|_\infty + \frac{2\overline{\tau}\|X\|_1}{n} + r_k\Big)\sqrt{|S^{k-1}|}\big\|\Delta\beta^k_{S^{k-1}}\big\|\|\varepsilon\|_\infty.
\end{aligned}$$

After a suitable rearrangement, this inequality is equivalent to saying that

$$\begin{aligned}
&\Big[2\underline{\tau}^2\kappa - 4\overline{\tau}\|X\|_{\max}\big(\lambda\|v^{k-1}_{S^*}\|_\infty + 2n^{-1}\overline{\tau}\|X\|_1 + r_k\big)|S^{k-1}|\Big]\|\Delta\beta^k\|^2 \\
&\leq 2\overline{\tau}\Big(\lambda\|v^{k-1}_{S^*}\|_\infty + 2n^{-1}\overline{\tau}\|X\|_1 + r_k\Big)\sqrt{|S^{k-1}|}\big\|\Delta\beta^k\big\|\|\varepsilon\|_\infty,
\end{aligned}$$

which by $\lambda < \frac{\underline{\tau}^2\kappa - 2\overline{\tau}\|X\|_{\max}(2n^{-1}\overline{\tau}\|X\|_1 + r_k)|S^{k-1}|}{2\overline{\tau}\|X\|_{\max}\|v^{k-1}_{S^*}\|_\infty|S^{k-1}|}$ implies the result. □

Proof of Theorem 2

Proof. For each $k \in \mathbb{N}$, let $S^{k-1} := S^* \cup \{i \notin S^* : w_i^{k-1} > \frac{1}{2}\}$. If $|S^{k-1}| \le 1.5s^*$, by invoking Lemma 5 and using the given assumption, we have

$$\begin{aligned}\big\|\beta^k - \beta^*\big\| &\le \frac{\overline{\tau}\big(\lambda\|v_{S^*}^{k-1}\|_\infty + 2n^{-1}\overline{\tau}\|X\|_1 + r_k\big)\sqrt{|S^{k-1}|}\|\varepsilon\|_\infty}{\underline{\tau}^2\kappa - 2\overline{\tau}\|X\|_{\max}\big(\lambda\|v_{S^*}^{k-1}\|_\infty + 2n^{-1}\overline{\tau}\|X\|_1 + r_k\big)|S^{k-1}|}\\ &\le \frac{\overline{\tau}\big(\lambda\|v_{S^*}^{k-1}\|_\infty + 2n^{-1}\overline{\tau}\|X\|_1 + r_k\big)\sqrt{|S^{k-1}|}\|\varepsilon\|_\infty}{\underline{\tau}^2\kappa - 3\overline{\tau}\|X\|_{\max}\big(\lambda + 2n^{-1}\overline{\tau}\|X\|_1 + \epsilon\big)s^*}\\ &\le c\overline{\tau}\big(\lambda\|v_{S^*}^{k-1}\|_\infty + 2n^{-1}\overline{\tau}\|X\|_1 + r_k\big)\sqrt{|S^{k-1}|}\|\varepsilon\|_\infty \qquad (012)\end{aligned}$$

where the second inequality is by the nondecreasing of $t \mapsto \frac{c_2+t}{c_1-t}$ for constants $c_1, c_2 > 0$, and the last one is by the restriction on λ. Since $2n^{-1}\overline{\tau}\|X\|_1 + r_k \le \frac{\lambda}{8}$ and $\|v_{S^*}^{k-1}\|_\infty \le 1$, it follows that $\big\|\beta^k - \beta^*\big\| \le \frac{9c\overline{\tau}\lambda\|\varepsilon\|_\infty}{8n}\sqrt{1.5s^*}$, and the desired result holds. So, it suffices to argue that $|S^{k-1}| \le 1.5s^*$ for all $k \in \mathbb{N}$. When $k = 1$, the statement holds trivially since $w^0 = 0$ implies $S^0 = S^*$. Assuming that $|S^{k-1}| \le 1.5s^*$ holds for $k = l$ with $l \ge 1$, we prove that it holds for $k = l+1$. Indeed, since $S^l \setminus S^* = \big\{i \notin S^* : w_i^l > \frac{1}{2}\big\}$, we have $w_i^l \in (\frac{1}{2}, 1]$ for $i \in S^l \setminus S^*$. Together with formula (3.3), we deduce that $\rho_l|\beta_i^l| \ge 1$, and hence the following inequality holds:

$$\sqrt{|S^l \setminus S^*|} \le \sqrt{\sum_{i\in S^l\setminus S^*}\rho_l^2|\beta_i^l|^2} = \sqrt{\sum_{i\in S^l\setminus S^*}\rho_l^2|\beta_i^l - \beta_i^*|^2}.$$

Since the statement holds for $k = l$, we get $\|\beta^l - \beta^*\| \le \frac{9c\overline{\tau}\lambda\|\varepsilon\|_\infty\sqrt{1.5s^*}}{8}$. So, it holds that

$$\sqrt{|S^l \setminus S^*|} \le \rho_l\|\beta^l - \beta^*\| \le \frac{9c\overline{\tau}\rho_l\lambda\|\varepsilon\|_\infty}{8}\sqrt{1.5s^*} \le \sqrt{0.5s^*} \qquad (013)$$

where the last inequality is due to $\rho_l\lambda \le \rho_3\lambda \le \frac{8}{9\sqrt{3}c\overline{\tau}\|\varepsilon\|_\infty}$. The inequality (013) implies $|S^l| \le 1.5s^*$. This shows that the statement follows. □

To present the proof of Theorem 3, we need the following lemma which upper bounds $\|v_{S^*}^k\|_\infty$, whose proof is given in Lemma 3 of Tao et al. (2018).

Lemma 6. *Let F^k and Λ^k be the index sets defined by (4.9). Then,*

$$\|v^k_{S^*}\|_\infty \le \max_{i\in S^*} \mathbb{I}_{\Lambda^k}(i) + \max_{i\in S^*} \mathbb{I}_{F^k}(i) \quad \text{for each } k \in \{0\}\cup\mathbb{N}.$$

Proof of Theorem 3:

Proof. For each $k \in \mathbb{N}$, define $S^{k-1} := S^* \cup \{i \notin S^* : w_i^{k-1} > \frac{1}{2}\}$. Since the conclusion holds for $k = 1$, it suffices to consider $k \ge 2$. By the proof of Theorem 2, $|S^{k-1}| \le 1.5s^*$ for all $k \in \mathbb{N}$. Moreover, by (013) and $\rho_k \ge 1$,

$$\begin{aligned}\sqrt{|S^{k-1}|} &= \sqrt{|S^*| + |S^{k-1}\setminus S^*|} \le \sqrt{s^*} + \sqrt{|S^{k-1}\setminus S^*|}\\ &\le \sqrt{s^*} + \big(2n^{-1}\overline{\tau}\|X\|_1 + r_k\big)^{-1}\frac{\lambda\rho_{k-1}}{8}\big\|\beta^{k-1} - \beta^*\big\| \end{aligned} \tag{014}$$

where the first inequality is due to $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $a, b \ge 0$, the last one is due to $\lambda \ge 16n^{-1}\overline{\tau}\|X\|_1 + 8r_k$. From (012) and Lemma 6, we have

$$\begin{aligned}\|\beta^k - \beta^*\| &\le c\overline{\tau}\|\varepsilon\|_\infty\sqrt{|S^{k-1}|}\big[\lambda\big(\max_{i\in S^*}\mathbb{I}_{\Lambda^{k-1}}(i) + \max_{i\in S^*}\mathbb{I}_{F^{k-1}}(i)\big)\big]\\ &\quad + c\overline{\tau}\|\varepsilon\|_\infty\sqrt{|S^{k-1}|}\big[2n^{-1}\overline{\tau}\|X\|_1 + r_k\big]\\ &\le c\overline{\tau}\|\varepsilon\|_\infty\Big[\lambda\sqrt{1.5s^*}\max_{i\in S^*}\mathbb{I}_{\Lambda^0}(i) + \lambda\sqrt{1.5s^*}\rho_{k-1}\|\beta^{k-1} - \beta^*\|\\ &\quad + \big(2n^{-1}\overline{\tau}\|X\|_1 + r_k\big)\sqrt{|S^{k-1}|}\Big]\end{aligned}$$

where the last inequality is since $\max_{i\in S^*}\mathbb{I}_{F^{k-1}}(i) \le \max_{i\in S^*}\rho_{k-1}\big||\beta_i^{k-1}| - |\beta_i^*|\big| \le \rho_{k-1}\|\beta^{k-1} - \beta^*\|$. Substituting (014) into this inequality yields

$$\begin{aligned}\|\Delta\beta^k\| &\le c\overline{\tau}\|\varepsilon\|_\infty\sqrt{s^*}\big(2n^{-1}\overline{\tau}\|X\|_1 + r_k\big) + c\overline{\tau}\lambda\|\varepsilon\|_\infty\sqrt{1.5s^*}\max_{i\in S^*}\mathbb{I}_{\Lambda^0}(i)\\ &\quad + c\overline{\tau}\|\varepsilon\|_\infty\rho_{k-1}\lambda(\sqrt{1.5s^*} + 1/8)\|\beta^{k-1} - \beta^*\|\\ &\le 2cn^{-1}\overline{\tau}^2\|\varepsilon\|_\infty\sqrt{s^*}\|X\|_1 + c\overline{\tau}\|\varepsilon\|_\infty\sqrt{s^*}r_k\\ &\quad + c\overline{\tau}\lambda\|\varepsilon\|_\infty\sqrt{1.5s^*}\max_{i\in S^*}\mathbb{I}_{\Lambda^0}(i) + \frac{\sqrt{3}}{3}\|\Delta\beta^{k-1}\|\end{aligned}$$

where the relation $\rho_{k-1}\lambda \leq \rho_3\lambda \leq [\sqrt{3}c\overline{\tau}\|\varepsilon\|_\infty(\sqrt{1.5s^*}+1/8)]^{-1}$ is used. The desired result follows by using the last recursion inequality. □

Appendix C

We describe the iterates of the semismooth Newton method and those of the semi-proximal ADMM in Gu and Zou (2016). The iterates of the semismooth Newton method are as follows.

Algorithm 1 A semismooth Newton method

Initialization: Fix k and j. Choose $0 < c_1 < c_2 < 1, \mu = 10^{-5}$ and $u^0 = 0$.

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

1. Choose $U^l \in \mathcal{U}_j(u^l), V^l \in \mathcal{V}_j(u^l)$ and set $W^l = \gamma_{2,j}^{-1}U^l + \gamma_{1,j}^{-1}XV^lX^{\mathbb{T}}$. Then, seek a solution $d^l \in \mathbb{R}^n$ to the following linear system

$$(W^l + \mu I)d = -\Phi_{k,j}(u^l). \tag{015}$$

2. Search the step-size α_l in the direction d^l to satisfy

$$\Psi_{k,j}(u^l + \alpha_l d^l) \leq \Psi_{k,j}(u^l) + c_1\alpha_l\langle\nabla\Psi_{k,j}(u^l), d^l\rangle,$$

$$|\langle\nabla\Psi_{k,j}(u^l + \alpha_l d^l), d^l\rangle| \leq c_2|\langle\nabla\Psi_{k,j}(u^l), d^l\rangle|.$$

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

end while

Notice that the subproblem (3.1) can be equivalently written as

$$\min_{\beta\in\mathbb{R}^p, z\in\mathbb{R}^n}\Big\{f_\tau(z) + \|\omega^{k-1}\circ\beta\|_1 \ \text{ s.t. } \ X\beta + z - y = 0\Big\} \tag{016}$$

whose dual problem, after an elementary calculation, takes the form of

$$\min_{u\in\mathbb{R}^n}\Big\{f_\tau^*(u)+\langle u,y\rangle \ \ \text{s.t.}\ \ |(X^{\mathbb{T}}u)_i|\le\omega_i^{k-1},\ i=1,\ldots,p\Big\}. \tag{017}$$

For a given $\sigma>0$, the augmented Lagrangian function of (016) is given by

$$L_\sigma(\beta,z,u):=f_\tau(z)+\|\omega^{k-1}\circ\beta\|_1+\langle u,X\beta+z-y\rangle+\frac{\sigma}{2}\|X\beta+z-y\|^2.$$

The iterate steps of the semi-proximal ADMM in Gu et al. (2018) are described as follows.

Algorithm 2 Semi-proximal ADMM for solving (016)

Initialization: Choose $\sigma>0,\gamma=\sigma\|X^{\mathbb{T}}X\|$ and $\varrho\in(1,\frac{\sqrt{5}+1}{2})$, and an initial point $(\beta^0,z^0,u^0)\in\mathbb{R}^p\times\mathbb{R}^n\times\mathbb{R}^n$ with $\beta^0=\beta^{k-1}$. Set $j=0$.

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

1. Compute the following convex minimization problem

$$\beta^{j+1}=\mathop{\arg\min}_{\beta\in\mathbb{R}^p}L_\sigma(\beta,z^j,u^j)+\frac{1}{2}\|\beta-\beta^j\|^2_{\gamma I-\sigma X^{\mathbb{T}}X}, \tag{018a}$$

$$z^{j+1}=\mathop{\arg\min}_{z\in\mathbb{R}^n}L_\sigma(\beta^{j+1},z,u^j). \tag{018b}$$

2. Update the multiplier by $u^{j+1}=u^j+\varrho\sigma(X\beta^{j+1}+z^{j+1}-y)$.

3. Set $j\leftarrow j+1$, and then go to Step 1.

end while

Remark 2. (i) Algorithm 2 has a little difference from Algorithm 1 of Gu et al. (2018) since here the semi-proximal term $\frac{1}{2}\|\beta-\beta^j\|^2_{\gamma I-\sigma X^{\mathbb{T}}X}$, rather than $\frac{1}{2}\|\beta-\beta^j\|^2_{\sigma(\gamma I-X^{\mathbb{T}}X)}$, is used. Let

$h^j = \gamma\beta^j + \sigma X^{\mathbb{T}}(X\beta^j + z^j - y + u^j/\sigma)$. Problems (018a) and (018b) have a closed form solution:

$$\beta^{j+1} = \mathrm{sign}\big(\gamma^{-1}h^j\big)\max\big(|\gamma^{-1}h^j| - \gamma^{-1}\omega^{k-1}, 0\big),$$

$$z^{j+1} = \mathcal{P}_{\sigma^{-1}} f_\tau(y - X\beta^{j+1} - \sigma^{-1}u^j).$$

(ii) During our implementation of Algorithm 2, we adjust σ dynamically by the ratio of the primal and dual infeasibility. By comparing the first-order optimality conditions of (018a) and (018b) with those of (016) and using the multiplier updating step, we measure the primal and infeasibility and the dual gap at (β^j, z^j, u^j) in terms of $\epsilon^j_{\mathrm{pinf}}, \epsilon^j_{\mathrm{dinf}}$ and $\epsilon^j_{\mathrm{gap}}$, respectively:

$$\epsilon^j_{\mathrm{dinf}} := \frac{\sqrt{\|\zeta^j\|^2 + \|(\varrho^{-1}-1)(u^j - u^{j-1})\|^2}}{1+\|y\|}, \tag{019a}$$

$$\epsilon^j_{\mathrm{pinf}} := \frac{\|u^j - u^{j-1}\|}{\varrho\sigma(1+\|y\|)}, \quad \epsilon^j_{\mathrm{gap}} := \frac{|\omega^j_{\mathrm{prim}} + \omega^j_{\mathrm{dual}}|}{\max\big(1, 0.5(\omega^j_{\mathrm{prim}} + \omega^j_{\mathrm{dual}})\big)} \tag{019b}$$

where $\zeta^j := X^{\mathbb{T}}(u^j - u^{j-1} - \sigma(X\beta^{j-1} - y + z^{j-1})) - \gamma(\beta^j - \beta^{j-1})$, and ω^j_{prim} and ω^j_{dual} are the objective values of (016) and (017) at (β^j, z^j, u^j). Different from Gu et al. (2018), when $\max(\epsilon^j_{\mathrm{pinf}}, \epsilon^j_{\mathrm{dinf}}, \epsilon^j_{\mathrm{gap}}) \le \epsilon_{\mathrm{ADMM}}$ or $j > j_{\max}$, we terminate Algorithm 2. By comparing with the optimality conditions of (018a)-(018b) with those of (016), such a stopping criterion ensures that the obtained (β^j, z^j, u^j) is an approximate primal-dual solution pair.

Appendix D

D.1. Performance comparisons of three solvers

We shall test the performance of MSCRA_IPM, MSCRA_ADMM and MSCRA_PPA for computing the estimator $\widehat{\beta}$ in the same setting as in Fan et al. (2014a) and Gu et al. (2018). Specifically, with $\beta^* = (2,\ 0,\ 1.5,\ 0,\ 0.8,\ 0,\ 0,\ 1,\ 0,\ 1.75,\ 0,\ 0,\ 0.75,\ 0,\ 0,\ 0.3,\ \mathbf{0}^{\mathbb{T}}_{p-16})^{\mathbb{T}}$ for $(p,n) = (1000, 200)$, we obtain n observations from (2.1), where the noise ε comes from the distributions in Gu et al. (2018), including **(1)** the normal distribution $N(0,2)$; **(2)** the mixture normal distribution $0.9N(0,1) + 0.1N(0,25)$, denoted by MN_1; **(3)** the mixture normal distribution

$N(0,\sigma^2)$ with $\sigma\sim\text{Unif}(1,5)$, denoted by MN_2; **(4)** the Laplace distribution with density $d(u)=0.5\exp(-|u|)$; **(5)** the scaled Student's t-distribution with 4 degrees of freedom $\sqrt{2}\times t_4$; and **(6)** the Cauchy distribution with density $d(u)=\frac{1}{\pi(1+u^2)}$. For the covariance matrix Σ_x, we also consider those scenarios from Gu et al. (2018), including $\Sigma_x=I$; $\Sigma_x=(0.5^{|i-j|})_{ij}$ and $(0.8^{|i-j|})_{ij}$, denoted by $\text{AR}_{0.5}$ and $\text{AR}_{0.8}$; and $\Sigma_x=(\alpha+(1-\alpha)\mathbb{I}_{\{i=j\}})$ with $\alpha=0.5$ and 0.8, denoted by $\text{CS}_{0.5}$ and $\text{CS}_{0.8}$. We test the estimation and selection performance of the estimators computed with the solvers under each scenario in terms of the ℓ_2**-error**, the CPU time, and the number of false positives (**FP**) and negatives (**FN**).

As mentioned by Fan et al. (2014a), the cross-validation is not suitable for choosing the best $\nu=\lambda^{-1}$ due to the instability of ℓ_2-error under heavy tails. We choose the best λ by $\lambda_i=\max\big(0.01,\gamma_i\|X\|_1/n\big)$ with $\gamma_i=\gamma_{\min}+((i-1)/49)(\gamma_{\max}-\gamma_{\min})$ by seeking the constant γ optimally. Inspired by the choice strategy of λ in Fan et al. (2014a), we choose γ based on 100 validation data-sets. Specifically, for each of data-sets, we ran a grid search to find the best γ and then the best λ (with the lowest ℓ_2-error of β^f) for the particular setting. The optimal γ was recorded for each of the 100 validation data-sets. We denote by γ_{opt} the median of the 100 optimal γ, and use $\lambda=\max\big(0.01,\gamma_{\text{opt}}\|X\|_1/n\big)$ for the simulation studies. The best γ is searched from $\gamma_1,\ldots,\gamma_{51}$ for $\gamma_{\min}=0.08$ and $\gamma_{\max}=0.38$. Such $\gamma_{\max}$ is such that $N_{\text{nz}}(\beta^f)$ attains or is close to 0.

Table 1-5 report the average ℓ_2**-error**, **FP** and **FN** for $\tau=0.5$ and 0.75 based on 100 simulations. For almost all test problems, MSCRA_PPA requires only one-fifteenth of the CPU time of MSCRA_ADMM and MSCRA_IPM, and its ℓ_2**-error** is comparable with that of MSCRA_ADMM and MSCRA_IPM. In addition, for all test problems, the **FP** of MSCRA_PPA are lower than that of MSCRA_ADMM and MSCRA_IPM though its **FN** is a little higher than that of the latter two methods.

Table 1: Estimation and selection performance of three solvers for $\Sigma_x = I$

ε	Method	γ_{opt}	L_2-error	FP	FN	Time(s)	γ_{opt}	L_2-error	FP	FN	Time(s)
		$\tau = 0.5$					$\tau = 0.75$				
$\mathcal{N}(0,2)$	IPM	0.104	0.444(0.107)	5.100(2.057)	0.730(0.468)	4.221	0.110	0.523(0.157)	7.840(3.034)	0.670(0.514)	5.613
	ADMM	0.104	0.446(0.106)	5.100(2.028)	0.730(0.468)	3.033	0.110	0.523(0.158)	7.760(3.079)	0.670(0.514)	3.847
	PPA	0.116	0.446(0.119)	1.920(1.228)	0.800(0.426)	0.138	0.119	0.557(0.188)	3.810(1.937)	0.840(0.420)	0.202
MN_1	IPM	0.104	0.345(0.066)	5.030(2.007)	0.410(0.494)	3.566	0.110	0.377(0.078)	6.860(2.741)	0.490(0.502)	4.168
	ADMM	0.104	0.345(0.067)	5.150(2.110)	0.410(0.494)	2.601	0.110	0.377(0.078)	6.890(2.723)	0.480(0.502)	3.062
	PPA	0.110	0.347(0.066)	3.260(1.779)	0.510(0.502)	0.131	0.116	0.375(0.061)	5.050(2.333)	0.590(0.494)	0.191
MN_2	IPM	0.104	1.425(0.361)	6.750(2.955)	1.860(0.921)	5.558	0.122	1.764(0.501)	4.220(2.377)	2.660(1.085)	5.568
	ADMM	0.104	1.427(0.356)	6.760(3.114)	1.880(0.902)	3.829	0.122	1.749(0.512)	4.270(2.432)	2.670(1.064)	3.825
	PPA	0.116	1.347(0.343)	2.480(1.823)	2.320(0.994)	0.133	0.134	1.742(0.537)	1.790(1.690)	3.260(1.050)	0.151
Laplace	IPM	0.098	0.324(0.071)	7.410(2.775)	0.220(0.416)	3.835	0.110	0.364(0.089)	6.550(2.484)	0.410(0.494)	3.789
	ADMM	0.098	0.324(0.070)	7.450(2.797)	0.220(0.416)	2.709	0.110	0.365(0.089)	6.580(2.458)	0.400(0.492)	2.761
	PPA	0.104	0.326(0.073)	4.700(2.209)	0.280(0.451)	0.144	0.116	0.382(0.094)	4.970(2.158)	0.480(0.502)	0.204
$\sqrt{2} \times \text{t}_4$	IPM	0.104	0.487(0.139)	5.330(2.301)	0.760(0.474)	4.677	0.110	0.649(0.238)	7.300(2.880)	0.840(0.507)	4.907
	ADMM	0.104	0.487(0.138)	5.360(2.325)	0.760(0.474)	3.214	0.110	0.647(0.239)	7.360(2.812)	0.840(0.507)	3.340
	PPA	0.110	0.502(0.180)	3.160(1.587)	0.790(0.478)	0.157	0.122	0.684(0.286)	2.970(1.861)	1.010(0.643)	0.239
Cauchy	IPM	0.098	0.536(0.217)	8.340(3.019)	0.670(0.533)	4.954	0.110	0.730(0.364)	6.740(2.493)	1.000(0.765)	5.488
	ADMM	0.098	0.531(0.216)	8.340(2.879)	0.680(0.530)	2.989	0.110	0.729(0.360)	6.720(2.551)	1.010(0.759)	3.404
	PPA	0.116	0.560(0.274)	1.780(1.203)	0.910(0.637)	0.166	0.125	0.816(0.381)	2.760(1.837)	1.280(0.792)	0.243

Table 2: Estimation and selection performance of three solvers for $\text{AR}_{0.5}$

ε	Method	γ_{opt}	L_2-error	FP	FN	Time(s)	γ_{opt}	L_2-error	FP	FN	Time(s)
		$\tau = 0.5$					$\tau = 0.75$				
$\mathcal{N}(0,2)$	IPM	0.104	0.467(0.119)	4.650(2.148)	0.710(0.456)	3.744	0.110	0.609(0.222)	6.830(2.843)	0.800(0.512)	4.312
	ADMM	0.104	0.474(0.120)	4.620(2.112)	0.730(0.446)	2.553	0.110	0.606(0.214)	6.860(2.853)	0.800(0.512)	3.143
	PPA	0.110	0.491(0.145)	2.810(1.594)	0.760(0.474)	0.133	0.122	0.591(0.199)	3.020(1.664)	0.870(0.442)	0.201
MN_1	IPM	0.098	0.365(0.074)	7.020(2.515)	0.410(0.494)	3.661	0.110	0.399(0.076)	6.450(2.679)	0.570(0.498)	3.729
	ADMM	0.098	0.367(0.073)	7.070(2.536)	0.400(0.492)	2.746	0.110	0.399(0.076)	6.500(2.676)	0.570(0.498)	2.819
	PPA	0.098	0.366(0.073)	7.060(2.566)	0.410(0.494)	0.139	0.122	0.423(0.127)	3.390(1.959)	0.630(0.485)	0.180
MN_2	IPM	0.104	1.383(0.394)	4.990(2.472)	2.060(0.930)	5.168	0.122	1.665(0.434)	3.640(2.013)	2.610(0.920)	5.339
	ADMM	0.104	1.379(0.384)	5.220(2.747)	2.010(0.937)	3.446	0.122	1.679(0.420)	3.670(2.080)	2.590(0.911)	3.764
	PPA	0.119	1.365(0.420)	1.590(1.436)	2.490(0.937)	0.101	0.131	1.705(0.512)	2.100(1.755)	3.010(0.959)	0.167
Laplace	IPM	0.098	0.349(0.089)	7.250(2.564)	0.360(0.482)	3.818	0.110	0.381(0.099)	6.320(2.624)	0.580(0.496)	4.513
	ADMM	0.098	0.349(0.089)	7.250(2.591)	0.360(0.482)	2.851	0.110	0.381(0.099)	6.380(2.666)	0.570(0.498)	3.130
	PPA	0.104	0.352(0.088)	4.600(2.079)	0.410(0.494)	0.125	0.116	0.408(0.154)	4.610(2.188)	0.480(0.522)	0.209
$\sqrt{2} \times \text{t}_4$	IPM	0.104	0.534(0.165)	4.580(2.142)	0.830(0.473)	4.341	0.110	0.734(0.291)	6.920(2.990)	1.070(0.573)	5.785
	ADMM	0.104	0.533(0.165)	4.590(2.109)	0.830(0.473)	3.179	0.110	0.736(0.288)	6.860(3.052)	1.070(0.573)	3.891
	PPA	0.110	0.542(0.180)	3.020(1.723)	0.860(0.472)	0.129	0.122	0.710(0.283)	3.240(1.782)	1.150(0.575)	0.209
Cauchy	IPM	0.101	0.544(0.245)	6.130(2.232)	0.820(0.539)	4.912	0.104	0.695(0.343)	9.450(3.105)	0.980(0.681)	5.948
	ADMM	0.104	0.538(0.258)	4.890(2.136)	0.860(0.513)	2.952	0.104	0.693(0.335)	9.530(2.883)	0.950(0.672)	3.686
	PPA	0.116	0.561(0.280)	1.740(1.292)	0.980(0.603)	0.169	0.122	0.879(0.473)	3.270(1.814)	1.430(0.956)	0.233

D.2. Performance on a real data example

Now we test the performance of MSCRA_PPA on a real data set from `https://www.ncbi.nlm.nih.gov`, which is used by Scheetz et al. (2006) to illustrate the gene regulation in mam-

Table 3: Estimation and selection performance of three solvers for $\mathrm{AR}_{0.8}$

ε	Method	γ_{opt}	L_2-error	FP	FN	Time(s)	γ_{opt}	L_2-error	FP	FN	Time(s)
		$\tau = 0.5$					$\tau = 0.75$				
	IPM	0.095	0.852(0.361)	7.050(2.504)	1.260(0.733)	4.117	0.098	0.986(0.408)	10.740(3.852)	1.400(0.804)	6.170
$\mathcal{N}(0,2)$	ADMM	0.092	0.835(0.336)	8.800(2.723)	1.240(0.698)	3.306	0.098	0.996(0.404)	10.940(3.961)	1.400(0.816)	4.721
	PPA	0.110	0.910(0.404)	2.390(1.550)	1.520(0.731)	0.111	0.110	0.965(0.387)	5.140(2.454)	1.440(0.701)	0.193
	IPM	0.098	0.530(0.208)	5.300(2.368)	0.780(0.504)	3.683	0.098	0.622(0.254)	9.510(4.036)	0.850(0.557)	5.205
MN_1	ADMM	0.092	0.519(0.184)	8.460(2.844)	0.770(0.489)	2.933	0.098	0.625(0.261)	9.630(4.099)	0.850(0.557)	3.851
	PPA	0.104	0.550(0.227)	3.550(1.977)	0.800(0.512)	0.132	0.110	0.644(0.321)	5.120(2.363)	1.000(0.682)	0.184
	IPM	0.104	1.742(0.616)	4.350(2.086)	2.590(0.889)	4.362	0.122	2.113(0.641)	3.120(1.981)	3.020(0.995)	5.187
MN_2	ADMM	0.104	1.713(0.642)	4.560(2.203)	2.500(0.959)	3.187	0.116	2.139(0.629)	4.230(2.155)	2.970(0.958)	4.269
	PPA	0.140	1.809(0.649)	0.820(0.936)	2.920(0.929)	0.085	0.152	2.125(0.721)	0.940(0.886)	3.290(0.868)	0.126
	IPM	0.098	0.520(0.257)	5.810(2.639)	0.720(0.637)	3.767	0.104	0.650(0.375)	6.980(3.291)	0.980(0.710)	3.990
Laplace	ADMM	0.098	0.510(0.242)	5.880(2.626)	0.710(0.608)	2.864	0.104	0.645(0.370)	7.140(3.333)	0.970(0.703)	3.180
	PPA	0.104	0.543(0.267)	3.780(2.177)	0.840(0.615)	0.124	0.116	0.679(0.386)	3.710(2.176)	1.150(0.716)	0.167
	IPM	0.095	0.955(0.412)	7.180(2.754)	1.470(0.658)	4.517	0.098	1.135(0.465)	10.250(4.029)	1.660(0.831)	5.201
$\sqrt{2} \times \mathrm{t}_4$	ADMM	0.092	0.934(0.407)	8.700(3.125)	1.410(0.653)	3.236	0.098	1.135(0.485)	10.400(3.929)	1.660(0.867)	3.641
	PPA	0.110	1.009(0.400)	2.570(1.736)	1.630(0.646)	0.118	0.110	1.190(0.542)	5.450(2.516)	1.870(0.939)	0.194
	IPM	0.104	0.891(0.452)	3.440(2.134)	1.420(0.684)	3.853	0.110	1.168(0.573)	4.970(2.676)	1.790(0.946)	4.842
Cauchy	ADMM	0.098	0.850(0.435)	5.590(2.586)	1.320(0.723)	2.672	0.110	1.153(0.549)	4.950(2.668)	1.770(0.908)	2.901
	PPA	0.116	0.962(0.452)	1.380(1.237)	1.570(0.700)	0.157	0.122	1.138(0.570)	2.920(1.895)	1.800(0.921)	0.205

Table 4: Estimation and selection performance of three solvers for $\mathrm{CS}_{0.5}$

ε	Method	γ_{opt}	L_2-error	FP	FN	Time(s)	γ_{opt}	L_2-error	FP	FN	Time(s)
		$\tau = 0.5$					$\tau = 0.75$				
	IPM	0.092	0.683(0.266)	1.710(1.597)	1.130(0.464)	3.819	0.092	0.943(0.366)	3.810(2.759)	1.340(0.685)	4.533
$\mathcal{N}(0,2)$	ADMM	0.092	0.700(0.272)	1.750(1.459)	1.140(0.472)	3.336	0.098	0.962(0.388)	2.780(2.245)	1.450(0.757)	3.761
	PPA	0.104	0.744(0.282)	0.650(0.880)	1.260(0.543)	0.195	0.116	0.934(0.347)	1.020(1.163)	1.580(0.684)	0.227
	IPM	0.092	0.437(0.093)	1.300(1.243)	0.810(0.394)	3.366	0.098	0.505(0.157)	2.070(1.816)	0.840(0.368)	3.687
MN_1	ADMM	0.098	0.441(0.097)	0.730(0.777)	0.820(0.386)	2.981	0.098	0.506(0.148)	2.030(1.702)	0.840(0.368)	3.475
	PPA	0.104	0.448(0.107)	0.350(0.557)	0.930(0.293)	0.178	0.116	0.523(0.192)	0.420(0.867)	1.020(0.200)	0.235
	IPM	0.110	1.919(0.526)	2.320(1.999)	3.090(0.877)	3.447	0.122	2.253(0.492)	2.690(1.813)	3.550(0.744)	3.224
MN_2	ADMM	0.122	1.977(0.490)	3.210(2.271)	3.100(0.882)	3.088	0.143	2.268(0.451)	3.800(2.094)	3.530(0.745)	3.241
	PPA	0.152	2.016(0.545)	1.650(1.480)	3.410(0.866)	0.117	0.155	2.444(0.579)	2.600(1.717)	3.830(0.842)	0.170
	IPM	0.086	0.445(0.140)	2.390(2.117)	0.810(0.394)	3.926	0.098	0.568(0.253)	2.290(2.027)	1.010(0.414)	3.868
Laplace	ADMM	0.086	0.445(0.139)	2.520(2.134)	0.800(0.402)	3.773	0.092	0.559(0.212)	3.480(2.552)	0.920(0.442)	3.889
	PPA	0.098	0.469(0.167)	0.930(1.380)	0.910(0.379)	0.181	0.104	0.586(0.279)	1.570(2.171)	1.110(0.510)	0.250
	IPM	0.092	0.874(0.352)	1.960(1.780)	1.400(0.651)	4.345	0.092	1.206(0.486)	4.150(2.724)	1.710(0.868)	4.657
$\sqrt{2} \times \mathrm{t}_4$	ADMM	0.086	0.905(0.339)	3.600(2.229)	1.310(0.598)	4.071	0.095	1.259(0.448)	3.760(2.527)	1.800(0.791)	3.875
	PPA	0.110	0.966(0.347)	0.910(1.215)	1.610(0.680)	0.165	0.116	1.172(0.429)	1.290(1.241)	1.980(0.816)	0.216
	IPM	0.086	0.803(0.377)	3.050(2.208)	1.330(0.620)	5.123	0.092	1.239(0.575)	3.910(2.016)	1.900(0.859)	5.142
Cauchy	ADMM	0.092	0.896(0.436)	2.270(1.869)	1.480(0.674)	3.599	0.095	1.392(0.592)	4.190(2.608)	2.040(0.887)	3.471
	PPA	0.101	0.880(0.415)	1.200(1.198)	1.460(0.658)	0.278	0.113	1.237(0.502)	1.470(1.540)	2.030(0.834)	0.333

malian eyes and to gain insight into genetic variation related to human eyes. This microarray data comprises gene expression levels of $31,042$ probes on 120 twelve-week-old laboratory rats. For the $31,042$ probes, as suggested by Scheetz et al. (2006), we first carry out the preprocessing

Table 5: Estimation and selection performance of three solvers for $CS_{0.8}$

ε	Method	γ_{opt}	L_2-error	FP	FN	Time(s)	γ_{opt}	L_2-error	FP	FN	Time(s)
		$\tau = 0.5$					$\tau = 0.75$				
$\mathcal{N}(0,2)$	IPM	0.092	1.572(0.411)	1.020(1.263)	2.630(0.761)	2.879	0.098	1.803(0.469)	1.480(1.337)	2.890(0.840)	2.907
	ADMM	0.131	1.683(0.365)	2.050(1.617)	2.820(0.796)	2.979	0.116	1.923(0.462)	3.050(2.057)	2.950(0.903)	3.077
	PPA	0.140	1.709(0.423)	0.650(1.029)	3.010(0.759)	0.229	0.140	1.939(0.460)	1.210(1.233)	3.220(0.773)	0.177
MN_1	IPM	0.086	0.971(0.339)	0.330(0.604)	1.750(0.657)	3.269	0.086	1.118(0.405)	0.700(0.835)	1.840(0.762)	3.355
	ADMM	0.086	0.952(0.363)	0.910(1.173)	1.600(0.696)	3.178	0.098	1.249(0.365)	1.620(1.523)	1.980(0.738)	3.230
	PPA	0.110	1.128(0.336)	0.110(0.314)	2.070(0.655)	0.202	0.110	1.283(0.392)	0.460(0.784)	2.270(0.777)	0.150
MN_2	IPM	0.134	3.087(0.643)	3.890(2.331)	4.510(0.893)	2.683	0.125	3.371(0.602)	4.780(2.729)	4.910(0.911)	2.739
	ADMM	0.137	2.897(0.496)	7.840(3.589)	4.250(0.903)	3.432	0.134	3.197(0.477)	8.640(3.586)	4.600(0.964)	3.491
	PPA	0.158	3.161(0.681)	3.910(2.708)	4.680(0.898)	0.146	0.149	3.507(0.625)	4.710(2.467)	5.120(0.868)	0.117
Laplace	IPM	0.086	1.066(0.409)	0.380(0.708)	1.910(0.753)	3.352	0.086	1.372(0.493)	1.130(1.284)	2.350(0.903)	3.417
	ADMM	0.098	1.177(0.441)	1.350(1.591)	2.010(0.745)	3.248	0.104	1.540(0.494)	2.510(2.267)	2.510(0.904)	3.223
	PPA	0.110	1.254(0.427)	0.220(0.561)	2.350(0.783)	0.192	0.128	1.558(0.496)	0.710(0.977)	2.800(0.829)	0.157
$\sqrt{2} \times t_4$	IPM	0.101	1.795(0.435)	1.300(1.314)	2.940(0.789)	2.923	0.104	2.160(0.517)	2.280(1.735)	3.230(0.827)	2.980
	ADMM	0.128	1.889(0.409)	3.320(2.344)	2.920(0.813)	3.215	0.110	2.210(0.462)	5.180(3.439)	3.250(0.833)	3.345
	PPA	0.146	1.923(0.454)	1.150(1.507)	3.200(0.816)	0.166	0.152	2.261(0.547)	1.580(1.505)	3.570(0.807)	0.137
Cauchy	IPM	0.095	1.986(0.618)	1.560(1.486)	3.230(0.874)	3.267	0.113	2.498(0.734)	2.390(1.933)	3.850(1.019)	3.122
	ADMM	0.128	2.181(0.564)	4.210(2.552)	3.440(0.903)	2.870	0.116	2.417(0.587)	5.240(3.108)	3.630(1.012)	2.881
	PPA	0.158	2.357(0.700)	1.460(1.374)	3.800(0.888)	0.212	0.134	2.667(0.805)	2.650(2.167)	4.160(1.080)	0.178

to obtain $18,986$ probes. Among those probes, there is one probe, 1389163_at, corresponding to gene TRIM32, that was found to be associated with the Bardet-Biedl syndrome (see Chiang (2006)). We are interested in how the expression of this gene depends on the expressions of all other 18,985 genes. To achieve this goal, we select 3,000 probes with the largest variances and then standardize the selected 3,000 probes such that they have mean 0 and standard deviation 1, as Gu and Zou (2016) and Wang et al. (2012) did. Thus, we obtain an $n \times p$ sample matrix X' with $n = 120$ and $p = 3000$, and set $X = [e \;\; X'] \in \mathbb{R}^{n\times(p+1)}$.

Since the previous numerical tests show that MSCRA_IPM and MSCRA_ADMM have very similar performance, we use MSCRA_PPA and MSCRA_ADMM with $\tau = 0.25, 0.5$ and 0.75 to analyze the data on all 120 rats. The parameter $\nu = \lambda^{-1}$ is used with $\lambda = \max\big(0.01, \gamma\|X\|_1/n\big)$, where γ is selected via five-fold cross-validation. The results are reported on the third and fourth columns of Table 6. We also conduct 50 random partitions on the data, each of which has 80 rats in the training set and 40 rats in the validation set. We apply MSCRA_ADMM and MSCRA_PPA to the training set with λ chosen as above and evaluate its prediction error

on the validation set by calculating $\frac{1}{40}\sum_{i\in\text{validation}}\theta_\tau(y_i-\beta_0-x_i^{\mathbb{T}}\widehat{\beta}^f)$, where $x_i^{\mathbb{T}}$ means the ith row of X'. The average number of selected genes, prediction errors and times over the 50 partitions are listed in the last three columns of Table 6. We see that the average number of the genes selected by MSCRA_PPA is less than that of the genes selected by MSCRA_ADMM, the average prediction error of the former is lower than that of the latter, and the average CPU time of the former is about one-fifteenth of the latter.

Table 6: Analysis of the microarray data by MSCRA_PPA and MSCRA_ADMM

Method	τ	All data		Random partition		
		#genes	Time(s)	Ave.#genes	Pre_error	Time(s)
ADMM	0.25	17	3.843	17.200(1.807)	0.050(0.009)	4.686(0.804)
	0.5	27	4.141	20.960(4.323)	0.029(0.005)	3.555(0.496)
	0.75	19	4.314	21.280(2.611)	0.040(0.005)	3.534(0.405)
PPA	0.25	20	0.208	16.440(3.721)	0.023(0.006)	0.235(0.056)
	0.5	27	0.226	20.740(4.237)	0.029(0.005)	0.247(0.136)
	0.75	17	0.181	12.500(3.032)	0.024(0.004)	0.352(0.068)

References

Chiang, A. P. (2006). *Homozygosity mapping with SNP arrays identifies Trim32, an e3 Ubiquitin Ligase, as a Bardet-Biedl Syndrome Gene (BBS11). Proceedings of the National Academy of Sciences 103*, pp. 6287–6292.

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

Fan, J., Fan, Y. Y. and Barut, E. (2014). Adaptive robust variable selection. *The Annals of Statistics 42*, pp. 324–351.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*, pp. 1–22.

Gu, Y. W., and Zou, H. (2016). High-dimensional generalizations of asymmetric least squares regression and their applications. *The Annals of Statistics 44*, pp. 2661–2694.

Gu, Y. W., Fan, J., Kong, L. C., Ma, S. Q. and Zou, H. (2018). ADMM for high-dimensional sparse penalized quantile regression. *Technometrics 60*, pp. 319–331.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton, NJ: Princeton University Press.

Rockafellar, R. T. and Wets, R. J-B. (1998). *Variational Analysis*. Springer.

Scheetz, T. E., Kim, K. Y. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J., Casavant, T. L., Shefield, V. C. and Stone, E. M. (2006). Regulation of Gene Expression in theMammalian Eye and Its Relevance to Eye Disease. *Proceedings of the National Academy of Sciences 103*, pp. 14429–14434.

REFERENCES

Tao, T., Pan, S. H. and Bi, S. J. (2018). Calibrated zero-norm regularized LS estimator for high-dimensional error-in-variables regression. accpeted by Statistica Sinica.

Wang, L., Wu, Y. C. and Li, R. Z. (2012). Quantile regression for analyzing heterogeneity in ultra high dimension. *Journal of the American Statistical Association 107*, pp. 214–222.

School of Mathematics, South China University of Technology

E-mail: (*mathzdd@mail.scut.edu.cn)*

School of Mathematics, South China University of Technology

E-mail: (shhpan@scut.edu.cn)

School of Mathematics, South China University of Technology

E-mail: (bishj@scut.edu.cn)