<table>
<tr><th colspan="2">Statistica Sinica Preprint No: SS-2019-0415</th></tr>
<tr><td><b>Title</b></td><td>A proximal dual semismooth Newton method for zero-norm penalized quantile regression estimator</td></tr>
<tr><td><b>Manuscript ID</b></td><td>SS-2019-0415</td></tr>
<tr><td><b>URL</b></td><td>http://www.stat.sinica.edu.tw/statistica/</td></tr>
<tr><td><b>DOI</b></td><td>10.5705/ss.202019.0415</td></tr>
<tr><td><b>Complete List of Authors</b></td><td>Dongdong Zhang,
Shaohua Pan and
Shujun Bi</td></tr>
<tr><td><b>Corresponding Author</b></td><td>Shujun Bi</td></tr>
<tr><td><b>E-mail</b></td><td>bishj@scut.edu.cn</td></tr>
<tr><td colspan="2">Notice: Accepted version subject to English editing.</td></tr>
</table>

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.

Abstract: This paper is concerned with the computation of the high-dimensional zero-norm penalized quantile regression estimator, defined as a global minimizer of the zero-norm penalized check loss function. To seek a desirable approximation to the estimator, we reformulate this NP-hard problem as an equivalent augmented Lipschitz optimization problem, and exploit its coupled structure to propose a multi-stage convex relaxation approach (MSCRA_PPA), each step of which solves inexactly a weighted ℓ_1-regularized check loss minimization problem with a proximal dual semismooth Newton method. Under a restricted strong convexity condition, we provide the theoretical guarantee for the MSCRA_PPA by establishing the error bound of each iterate to the true estimator and the rate of linear convergence in a statistical sense. Numerical comparisons on some synthetic and real data show that MSCRA_PPA not only has comparable even better estimation performance, but also requires much less CPU time.

Key words and phrases: High-dimension, Zero-norm penalized quantile regression, Variable selection, Proximal dual semismooth Newton method.

1. Introduction

Sparse penalized regression has become a popular approach for high-dimensional data analysis. In the past two decades, many classes of sparse penalized regressions have been developed by imposing a suitable penalty term on the least squares loss such as the bridge penalty in Frank and Friedman (1993), Lasso in Tibshirani (1996), SCAD in Fan and Li (2001), elastic net in Zou and Hastie (2005), adaptive lasso by Zou (2006), and so on. We refer to the survey papers by Bickel and Li (2006) and Fan and Lv (2010) for the references. These penalties, as a convex surrogate (say, ℓ_1-norm) or a nonconvex approximation (say, the bridge penalty) to the zero-norm, essentially try to capture the performance of the zero-norm, first used in the best subsect selection by Breiman (1996). The sparse least squares regression approach is useful, but it only focuses on the central tendency of the conditional distribution. It is known that a certain covariate may not have significant influence on the mean value of the response but may have a strong effect on the upper quantile of the conditional distribution due to the heterogeneity of data. It is likely that a covariate has different effects at different segments of the conditional distribution. As illustrated by Koenker and Bassett (1978), for non-Gaussian error distributions, the least squares regression is substantially out-performed by the quantile regression (QR).

Inspired by this, many researchers recently have considered the QR introduced by Koenker and Bassett (1978) for high-dimensional data analysis, owing to its robustness to outliers and its ability to offer unique insights into the relation between the response variable and the covariates; see, e.g., Wu and Liu (2009); Belloni and Chernozhukov (2011); Wang et al. (2012); Wang (2013); Fan et al. (2014a,b). Belloni and Chernozhukov (2011) focused on the theory of the ℓ_1-penalized QR and showed that this estimator is consistent at the near-oracle rate and provided the conditions under which the selected model includes the true model; Wang (2013) studied the ℓ_1-penalized least absolute derivation (LAD) regression and verified that the estimator has near oracle performance with a high probability; and Fan et al. (2014a) studied the weighted ℓ_1-penalized QR and established the model selection oracle property and the asymptotic normality for this estimator. For nonconvex penalty-type QRs, Wu and Liu (2009) under mild conditions achieved the asymptotic oracle property of the SCAD and adaptive-Lasso penalized QRs, and Wang et al. (2012) showed that with probability approaching one, the oracle estimator is a local optimal solution to the SCAD or MCP penalized QRs of ultra-high dimensionality. We notice that the above results are all established for the asymptotic case $n \to \infty$.

Besides the above theoretical works, there are some works concerned with the computation of (weighted) ℓ_1-penalized QR estimators which, compared to the (weighted) ℓ_1-least-squares estimator, requires more sophisticated algorithms due to the piecewise linearity of the check loss function. Although the ℓ_1-penalized QR model can be transformed into a linear program (LP) by introducing additional variables and one may use the interior point method (IPM) softwares such as SeDuMi in Sturm (1999) to solve it, this is limited to the small or medium scale case; see Figure 1-2 in Section 5. Inspired by this, Wu and Lange (2008) proposed a greedy coordinate descent algorithm for the ℓ_1-penalized LAD regression, Yi and Huang (2017) proposed a semismooth Newton coordinate descent algorithm for the elastic-net penalized QR, and Gu et al. (2018) recently developed a semi-proximal alternating direction method of multipliers (sPADMM) and a combined version of ADMM and coordinate descent method (which is actually an inexact ADMM) for solving the weighted ℓ_1-penalized QR. In addition, for nonconvex penalized QRs, Peng and Wang (2015) developed an iterative coordinate descent algorithm and established the convergence of any subsequence to a stationary point, and Fan et al. (2014b) provided a systematic study for folded concave penalized regressions, including the SCAD and MCP penalized QRs as special cases, and showed that with high probability the oracle estimator can be obtained within two iterations of the local linear approximation (LLA) approach proposed by Zou and Li (2008). We find that Peng and Wang (2015) and Fan et al. (2014b) did not establish the error bound of the iterates to the true solution.

This work is interested in the computation of the high-dimensional zero-norm penalized QR estimator, a global minimizer of the zero-norm regularized check loss. To seek a high-quality approximation to this estimator, we reformulate this NP-hard problem as a mathematical program with an equilibrium constraint (MPEC), and obtain an equivalent augmented Lipschitz optimization problem from the global exact penalty of the MPEC. This augmented problem not only has a favorable coupled structure but also implies an equivalent DC (difference of convex) surrogate for the zero-norm regularized check loss minimization; see Section 2. By solving the augmented Lipschitz problem in an alternating way, we propose in Section 3 an MSCRA to compute a desirable surrogate for the zero-norm penalized QR estimator. Similar to the LLA method owing to Zou and Li (2008), the MSCRA solves in each step a weighted ℓ_1-regularized check loss minimization, but the subproblems are allowed to be solved inexactly. Under a mild restricted strong convexity condition, we provide its theoretical guarantee in Section 4 by establishing the error bound of each iterate to the true estimator and the rate of linear convergence in a statistical sense.

Motivated by the recent work Tang et al. (2019), we also develop a proximal dual semismooth Newton method (PDSN) in Section 5 for solving the subproblems involved in the MSCRA. Different from the semismooth Newton method by Yi and Huang (2017), this is a proximal point algorithm (PPA) with the subproblems solved by applying the semismooth Newton method to their duals, rather than to a smooth approximation to the elastic-net penalized check loss minimization problem. Numerical comparisons are made on some synthetic and real data for MSCRA_PPA, MSCRA_IPM and MSCRA_ADMM, which are the MSCRA with the subproblems solved by PDSN, SeDuMi in Sturm (1999) and semi-proximal ADMM in Gu et al. (2018), respectively. We find that MSCRA_IPM and MSCRA_ADMM have very similar performance, while MSCRA_PPA not only has a comparable estimation performance with the two methods but also requires only one-fifteenth of the CPU time required by MSCRA_ADMM and MSCRA_IPM.

Throughout this paper, I and e denote an identity matrix and a vector of all ones, whose dimensions are known from the context. For an $x \in \mathbb{R}^p$, write $|x| := (|x_1|, \ldots, |x_p|)^{\mathbb{T}}$ and $\text{sign}(x) := (\text{sign}(x_1), \ldots, \text{sign}(x_p))^{\mathbb{T}}$, and denote by $\|x\|_1, \|x\|$ and $\|x\|_\infty$ the l_1-norm, l_2-norm and l_∞-norm of x, respectively. For a matrix $A \in \mathbb{R}^{n\times p}$, $\|A\|, \|A\|_{\max}$ and $\|A\|_1$ respectively denote the spectral norm, element-wise maximum norm, and maximum column sum norm of A. For a set S, $\mathbb{I}_S$ means the characteristic function on S, i.e., $\mathbb{I}_S(z) = 1$ if $z \in S$, otherwise $\mathbb{I}_S(z) = 0$. For given $a, b \in \mathbb{R}^p$ with $a_i \leq b_i$ for $i = 1, \ldots, p$, $[a, b]$ means the box set. For an extended real-valued function $f : \mathbb{R}^p \to (-\infty, +\infty]$, write $\operatorname{dom} f := \{x \in \mathbb{R}^p \mid f(x) < \infty\}$, and denote $\mathcal{P}_\gamma f$ and $e_\gamma f$ for a given $\gamma > 0$ by the proximal mapping and Moreau envelope of f, defined as $\mathcal{P}_\gamma f(x) := \arg\min_{z\in\mathbb{R}^p} \big\{f(z) + \frac{1}{2\gamma}\|z - x\|^2\big\}$ and $e_\gamma f(x) := \min_{z\in\mathbb{R}^p} \big\{f(z) + \frac{1}{2\gamma}\|z - x\|^2\big\}$. In the sequel, we write $\mathcal{P}f$ for $\mathcal{P}_1 f$. When f is convex, $\mathcal{P}_\gamma f : \mathbb{R}^p \to \mathbb{R}^p$ is a Lipschitz mapping with modulus 1, and $e_\gamma f$ is a smooth convex function with $\nabla e_\gamma f(x) = \gamma^{-1}(x - \mathcal{P}_\gamma f(x))$.

2. Zero-norm penalized quantile regression and equivalent difference of convex model

Quantile regression is a popular method for studying the influence of a set of covariates on the conditional distribution of a response variable, and has been widely used to handle heteroscedasticity; see Koenker and Bassett (1982) and Wang et al. (2012). For a univariate response $\mathbf{Y}$ and a vector of covariates $\mathbf{X} \in \mathbb{R}^p$, the conditional cumulative distribution function of $\mathbf{Y}$ is defined as $F_{\mathbf{Y}}(t|x) := \Pr(\mathbf{Y} \leq t \mid \mathbf{X} = x)$, and the τth conditional quantile of $\mathbf{Y}$ is given by $Q_{\mathbf{Y}}(\tau|x) := \inf\big\{t \colon F_{\mathbf{Y}}(t|x) \geq \tau\big\}$. Let $X = [x_1 \ \cdots \ x_n]^{\mathbb{T}}$ be an $n \times p$ design matrix on $\mathbf{X}$. Consider the linear quantile regression

$$y = X\beta^* + \varepsilon \tag{2.1}$$

where $y = (y_1, \ldots, y_n)^{\mathbb{T}} \in \mathbb{R}^n$ is the response vector, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathbb{T}}$ is the noise vector whose components are independently distributed and satisfy $\Pr(\varepsilon_i \leq 0|x_i) = \tau$ for some known constant $\tau \in (0,1)$, and $\beta^* \in \mathbb{R}^p$ is the true but unknown coefficient vector. This quantile regression model actually assumes that $Q_{\mathbf{Y}}(\tau|x_i) = x_i^{\mathbb{T}}\beta^*$ for $i = 1, \ldots, n$. We are interested in the high-dimensional case where $p > n$ and the sparse model in the sense that only $s^*(\ll p)$ components of the unknown true β^* are nonzero.

For $\tau \in (0,1)$, let $f_\tau : \mathbb{R}^n \to \mathbb{R}$ be the check loss function of (2.1), i.e.,

$$f_\tau(z) := n^{-1}\textstyle\sum_{i=1}^n \theta_\tau(z_i) \ \ \text{with} \ \ \theta_\tau(u) := (\tau - \mathbb{I}_{\{u \leq 0\}})u \tag{2.2}$$

which was first introduced by Koenker and Bassett (1978). To estimate the unknown true β^* in (2.1), we consider the zero-norm regularized problem

$$\widehat{\beta}(\tau) \in \mathop{\arg\min}_{\beta \in \mathbb{R}^p} \Big\{ \nu f_\tau(y - X\beta) + \|\beta\|_0 \Big\} \tag{2.3}$$

where $\nu > 0$ is the regularization parameter, and $\|\beta\|_0$ denotes the zero-norm of β (i.e., the number of nonzero entries of β). By the expression of f_τ, f_τ is nonnegative and coercive (i.e., $f_\tau(\beta^k) \to +\infty$ whenever $\|\beta^k\| \to \infty$). By Lemma 3 in Appendix A, the estimator $\widehat{\beta}(\tau)$ is well defined. Since $\widehat{\beta}(\tau)$

depends on τ, there is a great possibility for model (2.3) to monitor different "locations" of the conditional distribution, and then the heteroscedasticity of the data, when existing, can be inspected by solving (2.3) with different $\tau \in (0,1)$. For the simplicity, in the sequel we use $\widehat{\beta}$ to replace $\widehat{\beta}(\tau)$, and for a given $\tau \in (0,1)$, write $\underline{\tau} := \min(\tau, 1-\tau)$ and $\overline{\tau} := \max(\tau, 1-\tau)$.

Due to the combination of the zero-norm, the computation of $\widehat{\beta}$ is NP-hard. To design an algorithm in the next section for seeking a high-quality approximation to $\widehat{\beta}$, we next derive an equivalent augmented Lipschitz optimization problem from a primal-dual viewpoint, and to demonstrate that such a mechanism provides a unified way to yield equivalent DC surrogates for the zero-norm regularized problem (2.3), we introduce a family of proper lsc convex functions on $\mathbb{R}$, denoted by $\mathscr{L}$, satisfying the conditions:

$$\mathrm{int}(\mathrm{dom}\,\phi) \supseteq [0,1],\ \ t^* := \mathop{\arg\min}_{0\le t\le 1} \phi(t),\ \ \phi(t^*) = 0\ \ \text{and}\ \ \phi(1) = 1. \tag{2.4}$$

With a $\phi \in \mathscr{L}$, clearly, the zero-norm $\|z\|_0$ is the optimal value function of

$$\min_{w\in\mathbb{R}^p} \Big\{ \textstyle\sum_{i=1}^{p} \phi(w_i) \quad \text{s.t.}\ \ \langle e-w, |z|\rangle = 0,\ 0 \le w \le e \Big\}.$$

This characterization of zero-norm shows that model (2.3) is equivalent to

$$\min_{\beta\in\mathbb{R}^p, w\in\mathbb{R}^p} \Big\{ \nu f_\tau(y - X\beta) + \sum_{i=1}^{p} \phi(w_i) \quad \text{s.t.}\ \ \langle e-w, |\beta|\rangle = 0,\ 0 \le w \le e \Big\} \tag{2.5}$$

in the following sense: if $\overline{\beta}$ is globally optimal to (2.3), then $(\overline{\beta}, \mathrm{sign}(|\overline{\beta}|))$ is a global optimal solution of problem (2.5), and conversely, if $(\overline{\beta}, \overline{w})$ is a global optimal solution of (2.5), then $\overline{\beta}$ is globally optimal to (2.3). Problem (2.5) is a mathematical program with an equilibrium constraint $e-w \geq 0, |\beta| \geq 0, \langle e-w, |\beta|\rangle = 0$ (abbreviated as MPEC). The equivalence between (2.3) and (2.5) shows that the difficulty of model (2.3) arises from the hidden equilibrium constraint. It is well known that the handling of nonconvex constraints is much harder than that of nonconvex objective functions. Then it is natural to consider the penalized version of problem (2.5)

$$\min_{\beta\in\mathbb{R}^p, w\in[0,e]} \Big\{\nu f_\tau(y-X\beta) + \big[\textstyle\sum_{i=1}^p \phi(w_i) + \rho\langle e-w, |\beta|\rangle\big]\Big\} \tag{2.6}$$

where $\rho > 0$ is the penalty parameter. Since $\beta \mapsto f_\tau(y-X\beta)$ is Lipschitz continuous, the following conclusion holds by Section 3.2 of Liu et al. (2018).

Theorem 1. *The problem* (2.6) *associated to each* $\rho > \overline{\rho} := \frac{\phi'_{-}(1)(1-t^*)\overline{\tau}\nu\|X\|}{1-t_0}$ *has the same global optimal solution set as the MPEC* (2.5) *does, where* t^0 *is the minimum element in* $[t^*, 1)$ *such that* $\frac{1}{1-t^*} \in \partial\phi(t_0)$.

Theorem 1 states that problem (2.6) is a global exact penalty of (2.5) in the sense that there is a threshold $\overline{\rho} > 0$ such that the former associated to every $\rho > \overline{\rho}$ has the same global optimal solution set as the latter does. Together with the equivalence between (2.3) and (2.5), model (2.3) is equivalent to problem (2.6). Notice that the objective function of (2.6) is globally Lipschitz continuous over its feasible set and its nonconvexity is owing to the coupled term $\langle e-w, |\beta|\rangle$ rather than the combination. So, problem (2.6) provides an equivalent augmented Lipschitz reformulation for the zero-norm problem (2.3). In fact, problem (2.6) associated to every $\rho > \overline{\rho}$ implies an equivalent DC surrogate for (2.3). To illustrate this, let $\psi(t) = \phi(t)$ if $t \in [0,1]$ and otherwise $\phi(t) = +\infty$. Then, with the conjugate $\psi^*(s) := \sup_{t\in\mathbb{R}}\{st - \psi(t)\}$ of ψ, one may check that (2.6) is equivalent to

$$\min_{\beta\in\mathbb{R}^p} \Big\{\Theta_{\nu,\rho}(\beta) := f_\tau(y - X\beta) + \nu^{-1}\textstyle\sum_{i=1}^p \big[\rho|\beta_i| - \psi^*(\rho|\beta_i|)\big]\Big\}. \tag{2.7}$$

Since ψ^* is a nondecreasing finite convex function on $\mathbb{R}$, the function $s \mapsto \psi^*(\rho|s|)$ is convex, and problem (2.7) is a DC program. To sum up the above discussions, problem (2.7) associated to every $\rho > \overline{\rho}$ provides an equivalent DC surrogate for (2.3). Moreover, $H_\rho(\beta) := \sum_{i=1}^p h_\rho(\beta_i)$ with $h_\rho(t) := \rho|t| - \psi^*(\rho|t|)$ for $t \in \mathbb{R}$ is a DC surrogate for the zero-norm. To close this section, we present some examples of $\phi \in \mathscr{L}$.

Example 1. Let $\phi(t) = t$ for $t \in \mathbb{R}$. After a simple computation, we have

$$\psi^*(s) = \begin{cases} 0 & \text{if } s \le 1, \\ s-1 & \text{if } s > 1 \end{cases} \quad \text{and} \quad h_\rho(t) = \begin{cases} \rho|t| & \text{if } |t| \le \frac{1}{\rho}, \\ 1 & \text{if } |t| > \frac{1}{\rho}. \end{cases}$$

It is immediate to see that the function $\nu^{-1}h_\rho(t)$ will reduce to the capped ℓ_1-function $t \mapsto \lambda\min(|t|, \alpha)$ in Zhang (2010) with $\nu = \rho/\lambda$ and $\rho = \alpha^{-1}$.

Example 2. Let $\phi(t) := \frac{a-1}{a+1}t^2 + \frac{2}{a+1}t$ $(a > 1)$ for $t \in \mathbb{R}$. One can calculate

$$\psi^*(s) = \begin{cases} 0 & \text{if } s \leq \frac{2}{a+1}, \\ \frac{((a+1)s-2)^2}{4(a^2-1)} & \text{if } \frac{2}{a+1} < s \leq \frac{2a}{a+1}, \\ s-1 & \text{if } s > \frac{2a}{a+1}; \end{cases} \tag{2.8}$$

$$h_\rho(t) = \begin{cases} \rho|t| & \text{if } |t| \leq \frac{2}{(a+1)\rho}, \\ \rho|t| - \frac{((a+1)\rho|t|-2)^2}{4(a^2-1)} & \text{if } \frac{2}{(a+1)\rho} < |t| \leq \frac{2a}{(a+1)\rho}, \\ 1 & \text{if } |t| > \frac{2a}{(a+1)\rho}. \end{cases}$$

It is not hard to check that $\nu^{-1}h_\rho(t)$ will reduces to the SCAD function $\rho_\lambda(t)$ in Fan and Li (2001) when $\nu = \frac{2}{(a+1)\lambda^2}$ and $\rho = \frac{2}{(a+1)\lambda}$.

Example 3. Let $\phi(t) := \frac{a^2}{4}t^2 - \frac{a^2}{2}t + at + \frac{(a-2)^2}{4}$ $(a > 2)$ for $t \in \mathbb{R}$. We have

$$\psi^*(s) = \begin{cases} -\frac{(a-2)^2}{4} & \text{if } s \leq a - a^2/2, \\ \frac{1}{a^2}(\frac{a(a-2)}{2} + s)^2 - \frac{(a-2)^2}{4} & \text{if } a - a^2/2 < s \leq a, \\ s-1 & \text{if } s > a; \end{cases}$$

$$h_\rho(t) = \begin{cases} \rho|t| - \frac{1}{a^2}(\frac{a(a-2)}{2} + \rho|t|)^2 + \frac{(a-2)^2}{4} & \text{if } |t| \leq a/\rho, \\ 1 & \text{if } |t| > a/\rho. \end{cases}$$

The $\nu^{-1}h_\rho(t)$ will reduce to the MCP in Zhang (2010) if $\nu = \frac{2}{a\lambda^2}, \rho = \frac{1}{\lambda}$.

3. Multi-stage convex relaxation approach

From the last section, to compute the estimator $\widehat{\beta}$, we only need to solve a single penalty problem (2.6) that is much easier than the zero-norm problem

(2.3) because its nonconvexity only arises from the coupled term $\langle w, |\beta|\rangle$. Observe that (2.6) becomes a convex program when either of w and β is fixed. So, we solve it in an alternating way and propose the following multi-stage convex relaxation approach (MSCRA) with ϕ in Example 2.

```
Algorithm 1 (MSCRA for computing $\widehat{\beta}$)
Initialization: Choose $\tau \in (0,1), \nu > 0, \rho_0 = 1, w^0 \in [0, \frac{1}{2}e]$. Set $\lambda = \frac{\rho_0}{\nu}$.
for $k = 1, 2, \ldots$.
    1. Seek an inexact solution to the weighted $\ell_1$-regularized problem
       $\beta^k \approx \mathop{\arg\min}_{\beta \in \mathbb{R}^p} \Big\{ f_\tau(y - X\beta) + \lambda \sum_{i=1}^{p} (1 - w_i^{k-1}) |\beta_i| \Big\}.$   (3.1)
    2. When $k = 1$, select a suitable $\rho_1 \geq \rho_0$ in terms of $\|\beta^1\|_\infty$. If $k = 2, 3$,
       select $\rho_k$ such that $\rho_k \geq \rho_{k-1}$; otherwise, set $\rho_k = \rho_{k-1}$.
    3. For $i = 1, 2, \ldots, p$, compute the following minimization problem
       $w_i^k = \mathop{\arg\min}_{0 \leq w_i \leq 1} \big\{ \phi(w_i) - \rho_k w_i |\beta_i^k| \big\}.$   (3.2)
end for
```

Remark 1. **(i)** Step 1 of Algorithm 1 is solving problem (2.6) with w fixed to be w^{k-1}, while Step 3 is solving this problem with β fixed to be β^k; that is, Algorithm 1 is solving the nonconvex penalty problem (2.6) in an alternating way. In the first stage, since there is no any information on estimating the nonzero entries of β^*, it is reasonable to impose an unbiased weight on each component of β. Motivated by this, we restrict the initial w^0 in $[0, 0.5e]$, a subset of the feasible set of w. When $w^0 = 0$, the first stage is precisely the minimization of the ℓ_1-penalized check loss function. Although the threshold $\overline{\rho}$ is known when the parameter ν in (2.3) is given, we select a varying ρ for (3.2) since it is just a relaxation of (2.6).

(ii) By the optimality condition of (3.2), $\rho_k|\beta_i^k| \in \partial\psi(w_i^k)$ for each i, which by Theorem 23.5 in Rockafellar (1970) and (2.8) is equivalent to saying

$$w_i^k = \min\left[1, \max\left(0, \frac{(a+1)\rho_k|\beta_i^k| - 2}{2(a-1)}\right)\right] \;\text{ for } i = 1, \ldots, p. \tag{3.3}$$

Clearly, when $\rho_k|\beta_i^k|$ is close to 0, $(1-w_i^k)$ in (3.3) may not equal 1 though close to 1; when $\rho_k|\beta_i^k|$ is very larger, $(1-w_i^k)$ in (3.3) may not equal 0 though close to 0. To achieve a high-quality solution with Algorithm 1, the last term of (3.1) implies that a smaller $(1-w_i^{k-1})$ but not 0 is expected for those larger $|\beta_i|$, and a larger $(1-w_i^{k-1})$ instead of 1 is expected for those smaller $|\beta_i|$. Thus, the function ϕ in Example 2 is desirable especially for those problems whose solutions have small nonzero entries. The weight w^k associated to the function ϕ in Example 3 has a similar performance, but the weight w^k associated to the function ϕ in Example 1 is different since $w_i^k = 0$ if $\rho_k|\beta_i^k| < 1$, $w_i^k = 1$ if $\rho_k|\beta_i^k| > 1$, otherwise $w_i^k \in [0, 1]$.

(iii) Algorithm 1 is actually an inexact majorization-minimization (MM) method (see Lange et al. (2000)) for solving the equivalent DC surrogate

(2.7) with a special starting point. Indeed, for a given $\beta' \in \mathbb{R}^p$, the convexity and smoothness of ψ^* implies that with $w_i = (\psi^*)'(\rho|\beta_i'|)$ for $i = 1, \ldots, p$,

$$\sum_{i=1}^{p} \psi^*(\rho|\beta_i|) \geq \sum_{i=1}^{p} \psi^*(\rho|\beta_i'|) + \rho\langle w, |\beta| - |\beta'|\rangle \quad \forall \beta \in \mathbb{R}^p. \tag{3.4}$$

Notice that each $w_i \in [0, 1]$ by the expression of ψ^*. Hence, the function

$$f_\tau(y - X\beta) + \lambda\big\|(e - w^{k-1}) \circ \beta\big\|_1 - \lambda\big[\sum_{i=1}^{p} \psi^*(\rho|\beta_i^{k-1}|) + \rho\langle w^{k-1}, |\beta^{k-1}|\rangle\big]$$

is a majorization of $\Theta_{\lambda,\rho}$ at β^{k-1} and the subproblem (3.1) is the inexact minimization of this majorization function. Also, for any given $\rho_0 > 0$, when $\|\beta^0\|_\infty \leq \frac{2}{(a+1)\rho_0}$, we have $w_i^0 = (\psi^*)'(\rho_0|\beta_i^0|) = 0$ by (2.8). Thus, the first stage of Algorithm 1 with $w^0 = 0$ is precisely the inexact MM method for (2.7) with β^0 satisfying $\|\beta^0\|_\infty \leq \frac{2}{(a+1)\rho_0}$. In addition, Algorithm 1 can be regarded as an inexact inversion of the LLA method proposed by Zou and Li (2008) for (2.7), but it is different from the DC algorithm by Wu and Liu (2009) since the latter depends on the majorization of $\beta \mapsto \sum_{i=1}^{p} \psi^*(\rho|\beta_i|)$ at β^k and the obtained approximation is lack of symmetry.

(iv) Considering that practical computation always involves deviation, we allow the problem in (3.1) to be solved inexactly with the accuracy measured in the following way: $\exists \delta^k \in \mathbb{R}^p$ and $r_k \geq 0$ with $\|\delta^k\| \leq r_k$ such that

$$\begin{aligned} \delta^k &\in \partial\big[f_\tau(y - X\beta) + \lambda\|(e - w^{k-1}) \circ \beta\|_1\big]_{\beta=\beta^k} \\ &= -X^{\mathbb{T}}\partial f_\tau(y - X\beta^k) + \lambda\big[(1 - w_1^{k-1})\partial|\beta_1^k| \times \cdots \times (1 - w_p^{k-1})\partial|\beta_p^k|\big] \end{aligned} \tag{3.5}$$

where the equality is by Theorem 23.8 in Rockafellar (1970). Notice that the first-order optimality conditions of (2.6) take the following form

$$u \in \partial f_\tau(z);\;\; \rho|\beta_i| \in \partial\psi(w_i) \text{ for } i = 1,\ldots,p;\;\; y - X\beta - z = 0;$$

$$X^{\mathbb{T}}u \in \lambda\big[(1-w_1)\partial|\beta_1| \times \cdots \times (1-w_p)\partial|\beta_p|\big],$$

where $u \in \mathbb{R}^n$ is the Lagrange multiplier associated to $y - X\beta - z = 0$. By Step 2 of Algorithm 1, $\rho_k|\beta^k| \in \partial\psi(w_1^k) \times \cdots \times \partial\psi(w_p^k)$. In view of this, we measure the KKT residual of (2.6) associated to ρ_k at (β^k, z^k, u^k) by

$$\mathbf{Err}_k := \frac{\sqrt{\|\Delta_1\|^2 + \|\Delta_2^k\|^2 + \|y - X\beta^k - z^k\|^2}}{1 + \|y\|} \le \text{tol} \tag{3.6}$$

where $\Delta_1^k := z^k - \mathcal{P}f_\tau(z^k + u^k)$ and $\Delta_2^k := X^{\mathbb{T}}u^k - \mathcal{P}h_k(X^{\mathbb{T}}u^k + \beta^k)$ with

$$h_k(\beta) := \|\lambda(e - w^k) \circ \beta\|_1 \;\; \text{for } \beta \in \mathbb{R}^p. \tag{3.7}$$

4. Theoretical guarantees of Algorithm 1

We denote by S^* the support of the true vector β^*, and define the set

$$\mathcal{C}(S^*) := \bigcup_{S^* \subset S, |S| \le 1.5s^*} \Big\{\beta \in \mathbb{R}^p \colon \|\beta_{S^c}\|_1 \le 3\|\beta_S\|_1\Big\}.$$

The matrix X is said to have the κ-restricted strong convexity on $\mathcal{C}(S^*)$ if

$$\kappa > 0 \;\text{ and }\; \frac{1}{2n}\|X\Delta\beta\|^2 \ge \kappa\|\Delta\beta\|^2 \quad \text{for all } \Delta\beta \in \mathcal{C}(S^*). \tag{4.8}$$

The RSC is equivalent to the restricted eigenvalue condition of the Gram matrix $\frac{1}{2n}X^{\mathbb{T}}X$ due to van de Geer and Bühlmann (2009) and Bickel et al.

(2009). Notice that $\mathcal{C}(S^*) \supseteq \left\{\beta \in \mathbb{R}^p : \|\beta_{(S^*)^c}\|_1 \leq 3\|\beta_{S^*}\|_1\right\}$. This RSC is a little stronger than the one used by Negahban et al. (2012) for the ℓ_1-regularized smooth loss minimization. In this section, we shall provide the deterministic theoretical guarantees for Algorithm 1 under this RSC, including the error bound of the iterate β^k to the true β^* and the decrease analysis of the error sequence. The proofs are all included in Appendix B. We need the following assumption on the optimality tolerance r_k of β^k:

Assumption 1. There exists $\epsilon > 0$ such that for each $k \in \mathbb{N}$, $r_k \leq \epsilon$.

First, by Lemma 4 in Appendix B, we have the following error bound.

Theorem 2. *Suppose that Assumption 1 holds, that X has the κ-RSC over $\mathcal{C}(S^*)$, and that the noise vector ε is nonzero. If ρ_3 and λ are chosen such that $\rho_3 \leq \frac{8}{9\sqrt{3}c\overline{\tau}\lambda\|\varepsilon\|_\infty}$ and $\lambda \in \left[\frac{16\overline{\tau}\|X\|_1}{n} + 8\epsilon, \frac{\underline{\tau}^2\kappa - c^{-1} - 3\overline{\tau}\|X\|_{\max}(2n^{-1}\overline{\tau}\|X\|_1+\epsilon)s^*}{3\overline{\tau}\|X\|_{\max}s^*}\right]$ for some constant $c \geq \frac{1}{\underline{\tau}^2\kappa - 27\overline{\tau}\|X\|_{\max}(2n^{-1}\overline{\tau}\|X\|_1+\epsilon)s^*}$, then for every $k \in \mathbb{N}$*

$$\|\beta^k - \beta^*\| \leq \frac{9c\overline{\tau}\lambda\sqrt{1.5s^*}}{8}\|\varepsilon\|_\infty.$$

Remark 2. (i) For the ℓ_1-regularized least squares smooth loss estimator $\beta^{\rm LS} \in \arg\min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{2n}\|y - X\beta\|^2 + \lambda_n\|\beta\|_1\right\}$, the error bound $\|\beta^{\rm LS} - \beta^*\| = O(\sigma\sqrt{s^*\log p/n})$ was obtained in Corollary 2 of Negahban et al. (2012) by taking $\lambda_n = \sqrt{\log p/n}$, where $\sigma > 0$ represents the variance of the noise. By comparing with this error bound, the error bound in Theorem 2 involves the infinite norm $\|\varepsilon\|_\infty$ of noise ε rather than its variance, and moreover, it still has the same order $O(\sqrt{s^* \log p/n})$ when the parameter $\lambda = O(1)$ in our model is rescaled to be λ_n.

(ii) For the following ℓ_1-regularized square-root nonsmooth loss estimator $\beta^{\text{sr}} \in \arg\min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{\sqrt{n}}\|y - X\beta\| + \frac{\lambda'}{n}\|\beta\|_1\right\}$, the error bound $\|\beta^{\text{sr}} - \beta^*\| = O\big(\frac{\sigma\sqrt{s^*}\lambda'\varpi}{n}\big)$ with $\varpi \geq \frac{1}{\sqrt{n}}\|\varepsilon\|$ was achieved in Theorem 1 of Belloni et al. (2011) by setting $\lambda' = O(n)$. By considering that $f_\tau(y - X\beta) = O(\sqrt{n}\|y - X\beta\|)$, the parameter λ in our model corresponds to λ'/n. Thus, the error bound in Theorem 2 corresponds to $O(\frac{\sqrt{s^*}\lambda'\|\varepsilon\|_\infty}{n})$, which has the same order as $O\big(\frac{\sigma\sqrt{s^*}\lambda'\varpi}{n}\big)$ since $\|\varepsilon\|_\infty = O(\frac{1}{\sqrt{n}}\|\varepsilon\|)$.

(iii) To ensure that the constant $c > 0$ exists, the constant κ needs to satisfy $\kappa > \frac{54\bar{\tau}^2 s^* \|X\|_{\max}\|X\|_1}{n\underline{\tau}^2}$ and the inexact accuracy ϵ of β^k needs to satisfy $0 \leq \epsilon < \frac{n\underline{\tau}^2\kappa - 54\bar{\tau}^2 s^* \|X\|_{\max}\|X\|_1}{27n\bar{\tau}s^*}$. Since $\|X\|_1 = O(n)$, it is necessary to solve the subproblem (3.1) with a very small inexact accuracy ϵ.

Theorem 2 establishes an error bound for every iterate β^k, but it does not tell us if the error bound of the current β^k is better than that of the previous β^{k-1}. In order to seek the answer, we study the decrease of the error bound sequence by bounding $\max_{i\in S^*}(1 - w_i^k)$. For this purpose, write

$F^0 := S^*$ and $\Lambda^0 := \{i\colon |\beta^*_i| \le \frac{4a}{(a+1)\rho_0}\}$, and for each $k \in \mathbb{N}$ define

$$F^k := \Big\{i\colon \big||\beta^k_i| - |\beta^*_i|\big| \ge \frac{1}{\rho_k}\Big\} \text{ and } \Lambda^k := \Big\{i\colon |\beta^*_i| \le \frac{4a}{(a+1)\rho_k}\Big\}. \tag{4.9}$$

From Lemma 6 in Appendix B, the value $\max_{i\in S^*}(1-w^k_i)$ is upper bounded by $\max_{i\in S^*}\max(\mathbb{I}_{\Lambda^k}(i), \mathbb{I}_{F^k}(i))$. By this, we have the following conclusion.

Theorem 3. *Suppose that Assumption 1 holds, that X has the κ-RSC over $\mathcal{C}(S^*)$, and that the noise ε is nonzero. If λ is chosen as in Theorem 2 and the parameter ρ_3 satisfies $\rho_3 \le \frac{1}{c\overline{\tau}\lambda\|\varepsilon\|_\infty(\sqrt{4.5s^*}+\sqrt{3}/8)}$, then for each $k \in \mathbb{N}$*

$$\begin{aligned}\|\beta^k - \beta^*\| \le \frac{(3+\sqrt{3})c\overline{\tau}^2\sqrt{s^*}\|X\|_1\|\varepsilon\|_\infty}{n} + \frac{(3+3\sqrt{3})c\overline{\tau}\lambda\sqrt{s^*}\|\varepsilon\|_\infty}{2\sqrt{2}}\max_{i\in S^*}\mathbb{I}_{\Lambda^0}(i)\\ + c\overline{\tau}\|\varepsilon\|_\infty\sqrt{s^*}\sum_{j=0}^{k-2} r_{k-j}\Big(\frac{1}{\sqrt{3}}\Big)^j + \Big(\frac{1}{\sqrt{3}}\Big)^{k-1}\|\beta^1 - \beta^*\|\end{aligned} \tag{4.10}$$

where we stipulate that $\sum_{j=0}^{k-2} r_{k-j}(\frac{1}{\sqrt{3}})^j = 0$ for $k = 1$.

Remark 3. **(i)** The error bound in (4.10) consists of the statistical error due to the noise, the identification error $\max_{i\in S^*}\mathbb{I}_{\Lambda^0}(i)$ related to the choice of a and ρ_0, and the computation errors $\sum_{j=0}^{k-2} r_{k-j}(\frac{1}{\sqrt{3}})^j$ and $(\frac{1}{\sqrt{3}})^{k-1}\|\beta^1 - \beta^*\|$. By the definition of Λ^0, when ρ_0 and a are such that $\frac{(a+1)\rho_0}{4a} > \frac{1}{\min_{i\in S^*}|\beta^*_i|}$, the identification error becomes zero. If $\min_{i\in S^*}|\beta^*_i|$ is not too small, it would be easy to choose such ρ_0. Clearly, when ρ_0 and a are chosen to be larger, the identification error is smaller. However, when ρ_0 and a are larger, ρ_1 becomes larger and each component of w^1 is close to 1 by (3.3).

Consequently, it will become very conservative to cut those smaller entries of β^2 when solving the second subproblem. Hence, there is a trade-off between the choice of a and ρ_0 and the computation speed of Algorithm 1. **(ii)** If the subproblem (3.1) could be solved exactly, the computation error $\sum_{j=0}^{k-2} r_{k-j}(\frac{1}{\sqrt{3}})^j$ vanishes. If the subproblem (3.1) is solved with the accuracy r_k satisfying $r_k \leq (\frac{1}{\sqrt{3}})^k \frac{1}{k^\nu}$ for $\nu > 1$, this computation error will tend to 0 as $k \to +\infty$. Since the third term on the right hand side of (4.10) is the combination of the noise and $\sum_{j=0}^{k-2} r_{k-j}(\frac{1}{\sqrt{3}})^j$, it is strongly suggested that the subproblem (3.1) is solved as well as possible.

For the RSC assumption in Theorem 2-3, from Raskutti et al. (2010) we know that if X is from the Σ_x-Gaussian ensemble (i.e., X is formed by independently sampling each row $x_i^{\mathbb{T}} \sim N(0, \Sigma_x)$, there exists a constant $\kappa > 0$ (depending on Σ_x) such that the RSC holds on $\mathcal{C}(S^*)$ with probability greater than $1-c_1 \exp(-c_2 n)$ as long as $n > c_0 s^* \log p$, where c_0, c_1 and c_2 are absolutely positive constants. From Banerjee et al. (2015), for some sub-Gaussian X, the RSC holds on $\mathcal{C}(S^*)$ with a high probability when n is over a threshold depending on the Gaussian width of $\mathcal{C}(S^*)$.

5. Proximal dual semismooth Newton method

By Remark 1 (iv), the pivotal part of Algorithm 1 is the exact solution of

$$\min_{\beta\in\mathbb{R}^p}\left\{f_\tau(y-X\beta)+h_{k-1}(\beta)-\langle\delta^k,\beta-\beta^{k-1}\rangle\right\}\tag{5.1}$$

where, for each $k\in\mathbb{N}$, h_k is the function defined in (3.7). In this section, we develop a proximal dual semismooth Newton method (PDSN) for (5.1), which is a proximal point algorithm (PPA) with the subproblems solved by applying the semismooth Newton method to their dual problems.

```
Algorithm 2  PPA for solving problem (5.1)
Initialization: Fix $k$. Choose $\gamma_{1,0},\gamma_{2,0},\underline{\gamma}>0,\varrho\in(0,1)$. Let $\beta^0=\beta^{k-1}$.
for $j=0,1,2,\ldots$
    1. Seek the unique minimizer $\beta^{j+1}$ to the following convex program
       $\min_{\beta\in\mathbb{R}^p}\Big\{f_\tau(y-X\beta)+h_{k-1}(\beta)-\langle\delta^k,\beta-\beta^{k-1}\rangle+\frac{\gamma_{1,j}}{2}\|\beta-\beta^j\|^2+\frac{\gamma_{2,j}}{2}\|X(\beta-\beta^j)\|^2\Big\}.$
    2. If $\beta^{j+1}$ satisfies the stopping rule, then stop. Otherwise, update $\gamma_{1,j}$
       and $\gamma_{2,j}$ by $\gamma_{1,j+1}=\max(\underline{\gamma},\varrho\gamma_{1,j})$ and $\gamma_{2,j+1}=\max(\underline{\gamma},\varrho\gamma_{2,j})$.
end for
```

Remark 4. **(i)** Since $f_\tau(y-X\cdot)$ and h_{k-1} are convex but nondifferentiable, we follow the same line as in Tang et al. (2019) to introduce a key proximal term $\frac{\gamma_{2,j}}{2}\|X\beta - X\beta^j\|^2$ except the common $\frac{\gamma_{1,j}}{2}\|\beta-\beta^j\|^2$. As will be shown later, this provides an effective way to handle the nonsmooth $f_\tau(y - X\cdot)$.

(ii) The first-order optimality conditions for (5.1) have the following form $u \in \partial f_\tau(z)$, $X^{\mathbb{T}}u + \delta^k \in \partial h_{k-1}(\beta)$, $y - X\beta - z = 0$, where $u \in \mathbb{R}^n$ is the multiplier vector associated to $y - X\beta - z = 0$. Hence, the KKT residual of problem (5.1) at (β^j, z^j, u^j) can be measured by

$$\mathbf{Err}_{\rm PPA}^j := \frac{\sqrt{\|z^j - \mathcal{P}f_\tau(z^j + u^j)\|^2 + \|\beta^j - \mathcal{P}h_{k-1}(X^{\mathbb{T}}u^j + \delta^k)\|^2 + \|y - X\beta^j - z^j\|^2}}{1 + \|y\|}.$$

So, we suggest $\mathbf{Err}_{\rm PPA}^j \le \epsilon_{\rm PPA}^j$ as the stopping condition of Algorithm 2.

The efficiency of Algorithm 2 depends on the solution of its subproblem, which by introducing a variable $z \in \mathbb{R}^n$ is equivalently written as

$$\min_{\beta\in\mathbb{R}^p, z\in\mathbb{R}^n} \Big\{ f_\tau(z) + h_{k-1}(\beta) - \langle \delta^k, \beta - \beta^{k-1}\rangle + \frac{\gamma_{1,j}}{2}\|\beta - \beta^j\|^2 + \frac{\gamma_{2,j}}{2}\|z - z^j\|^2 \Big\}$$
$$\text{s.t.} \quad X\beta + z - y = 0 \quad \text{with} \quad z^j = y - X\beta^j. \qquad (5.2)$$

After an elementary calculation, the dual of (5.2) takes the following form

$$\min_{u\in\mathbb{R}^n} \Big\{ \Psi_{k,j}(u) := \frac{\|u\|^2}{2\gamma_{2,j}} - e_{\gamma_{2,j}^{-1}} f_\tau\Big(z^j - \frac{u}{\gamma_{2,j}}\Big) - e_{\gamma_{1,j}^{-1}} h_{k-1}\Big(\beta^j - \frac{X^{\mathbb{T}}u + \delta^k}{\gamma_{1,j}}\Big) + \frac{\|X^{\mathbb{T}}u\|^2}{2\gamma_{1,j}} \Big\}.$$

Since $\Psi_{k,j}$ is a smooth convex function, seeking an optimal solution of the last dual problem is equivalent to finding a root to the system

$$\Phi_{k,j}(u) := -\mathcal{P}_{\gamma_{2,j}^{-1}} f_\tau\Big(z^j - \frac{u}{\gamma_{2,j}}\Big) - X\mathcal{P}_{\gamma_{1,j}^{-1}} h_{k-1}\Big(\beta^j - \frac{X^{\mathbb{T}}u + \delta^k}{\gamma_{1,j}}\Big) + y = 0. \quad (5.3)$$

Since $\mathcal{P}_{\gamma_{2,j}^{-1}} f_\tau$ and $\mathcal{P}_{\gamma_{1,j}^{-1}} h_{k-1}$ are strongly semismooth by Appendix A and the composition of strongly semismooth mappings is strongly semismooth by Facchinei and Pang (2003), the mapping $\Phi_{k,j}$ is strongly semismooth. Inspired by this, we use the semismooth Newton method to seek a root to system (5.3), which by Qi and Sun (1993) is expected to have a superlinear even quadratic convergence rate. By Proposition 2.3.3 and Theorem 2.6.6 of Clarke (1983), the Clarke Jacobian $\partial_C \Phi_{k,j}(u)$ of $\Phi_{k,j}$ at u is included in

$$\gamma_{2,j}^{-1}\partial_C\big[\mathcal{P}_{\gamma_{2,j}^{-1}} f_\tau\big]\Big(z^j - \frac{u}{\gamma_{2,j}}\Big) + \gamma_{1,j}^{-1} X \partial_C\big[\mathcal{P}_{\gamma_{1,j}^{-1}} h_{k-1}\big]\Big(\beta^j - \frac{X^{\mathbb{T}}u + \delta^k}{\gamma_{1,j}}\Big) X^{\mathbb{T}}$$
$$= \gamma_{2,j}^{-1}\mathcal{U}_j(u) + \gamma_{1,j}^{-1} X \mathcal{V}_j(u) X^{\mathbb{T}} \ \forall u \in \mathbb{R}^n \tag{5.4}$$

where (5.4) is due to Lemma 1-2 in Appendix A, and $\mathcal{U}_j(u)$ and $\mathcal{V}_j(u)$ are

$$\mathcal{U}_j(u) := \Big\{ \mathrm{Diag}(v_1, \ldots, v_n) \mid v_i \in \partial_C\big[\mathcal{P}_{\gamma_{2,j}^{-1}}(n^{-1}\theta_\tau)\big](z_i^j - \gamma_{2,j}^{-1} u_i) \Big\},$$
$$\mathcal{V}_j(u) := \Big\{ \mathrm{Diag}(v) \mid v_i = 1 \text{ if } |(\gamma_{1,j}\beta^j - X^{\mathbb{T}}u - \delta^k)_i| > \omega_i^k, \text{otherwise } v_i \in [0,1] \Big\}.$$

For each $U^j \in \mathcal{U}_j(u)$ and $V^j \in \mathcal{V}_j(u)$, the matrix $\gamma_{2,j}^{-1}U^j + \gamma_{1,j}^{-1} X V^j X^{\mathbb{T}}$ is semidefinite, and positive definite when $\{i \mid \frac{\tau-1}{n\gamma} \le z_i^j - \gamma_{2,j}^{-1} u_i \le \frac{\tau}{n\gamma}\} = \emptyset$ or the matrix X_J has full row rank with $J = \{i \mid |(\gamma_{1,j}\beta^j - X^{\mathbb{T}}u - \delta^k)_i| > \omega_i^k\}$. To ensure that each iterate of the semismooth Newton method works, or each element of Clarke Jacobian $\partial_C \Phi_{k,j}(u)$ is nonsingular, we add a small positive definite perturbation μI to $\gamma_{2,j}^{-1}U^j + \gamma_{1,j}^{-1} X V^j X^{\mathbb{T}}$. The detailed iterates of the semismooth Newton method is provided in Appendix C.

6. Numerical experiments

We shall test the performance of Algorithm 1 with the subproblems solved by PDSN, SeDuMi and sPADMM, respectively, on synthetic and real data, and call the three solvers MSCRA_PPA, MSCRA_IPM and MSCRA_ADMM, respectively. Among others, SeDuMi is solving the equivalent LP of (3.1):

$$\begin{aligned}\min_{(\beta^+,\beta^-)\in\mathbb{R}^{2p}_+,(\zeta^+,\zeta^-)\in\mathbb{R}^{2n}_+}\ & \langle\omega^k,\beta^+\rangle+\langle\omega^k,\beta^-\rangle+\frac{\tau}{n}\langle\zeta^+,e\rangle+\frac{1-\tau}{n}\langle\zeta^-,e\rangle\\ \text{s.t.}\ \ & X\beta^+-X\beta^-+\zeta^+-\zeta^-=y,\end{aligned}\tag{6.5}$$

and the iterates of sPADMM are described in Appendix C. All numerical results are computed by a laptop computer running on 64-bit Windows System with an Intel(R) Core(TM) i7-8565 CPU 1.8GHz and 8 GB RAM.

For SeDuMi, we adopt the default setting, and for sPADMM we choose the step-size $\varrho = 1.618$ and the initial $\sigma = 1$, and adopt the stopping criterion in Appendix C with $j_{\max} = 3000$ and $\epsilon_{\rm ADMM} = 10^{-6}$. For PDSN, we choose $\underline{\gamma} = 10^{-8}$, $\varrho = 5/7$ and $\gamma_{1,0} = \gamma_{2,0} = \min(0.1, R_0)$ where R_0 is the relative KKT residual at the initial (β^0, z^0, u^0), and adopt the stopping criterion in Remark 4(ii) with $\epsilon_{\rm PPA}^{j+1} = \max(10^{-8}, 0.1\epsilon_{\rm PPA}^{j})$ for $\epsilon_{\rm PPA}^0 = 10^{-6}$ and the stopping rule $\frac{\|\Phi_{k,j}(u^l)\|}{1+\|y\|} \le 0.1\epsilon_{\rm PPA}^{j}$ for Algorithm 1 in Appendix C.

For MSCRA_IPM, MSCRA_ADMM and MSCRA_PPA, we use $w^0 = 0$, and terminate them at β^k when $k > 10$, or $N_{\rm nz}(\beta^k) = \cdots = N_{\rm nz}(\beta^{k-3})$ and

$\mathbf{Err}_k \leq 10^{-5}$, or $N_{\rm nz}(\beta^k) = \cdots = N_{\rm nz}(\beta^{k-2})$ and $|\mathbf{Err}_k - \mathbf{Err}_{k-2}| \leq 10^{-6}$, where $N_{\rm nz}(\beta^k) := \sum_{i=1}^{p} \mathbb{I}\{|\beta_i^k| > 10^{-6} \max(1, \|\beta^k\|_\infty)\}$ denotes the number of nonzero entries of β^k, and $\mathbf{Err}_k$ is the KKT residual at the kth step defined in (3.6). We update ρ_k by $\rho_1 = \max\big(1, \frac{1}{3\|\beta^1\|_\infty}\big)$ and $\rho_k = \min\big(\frac{5}{4}\rho_{k-1}, \frac{10^8}{\|\beta^k\|_\infty}\big)$ for $k = 2, 3$. In addition, during the implementation of three solvers, we run SeDuMi, sPADMM and PSDN to solve the kth subproblem with the optimal solution of the $(k-1)$th subproblem yielded by them as the starting point. When $k = 1$, we choose $\beta^0 = 0$ to be the starting point of MSCRA_IPM and MSCRA_ADMM, and use $\beta^0 = 0$ to run Algorithm 2.

6.1. Comparisons of three solvers for the subproblem

We make numerical comparisons among SeDuMi, sPADMM and PDSN by applying them to the problem (3.1) for $k = 1$, i.e., the ℓ_1-regularized check loss minimization problem. Inspired by the work owing to Gu et al. (2018), we consider the simulation model $y_i = x_i^{\mathbb{T}}\beta^* + \kappa\varepsilon_i$ for $i = 1, \ldots, n$ in Friedman et al. (2010) to generate data, where $x_i^{\mathbb{T}} \sim N(0, \Sigma)$ for $i = 1, \ldots, n$ with $\Sigma = (\alpha + (1-\alpha)\mathbb{I}_{\{i=j\}})_{p\times p}$, $\beta_j^* = (-1)^j \exp(-\frac{2j-1}{20})$, $\varepsilon \sim N(0, \Sigma)$, and κ is chosen such that the signal-noise ratio of the data is 3.0. We focus on the high-dimensional situation with $(p, n) = (5000, 500)$ and $\alpha = 0$ and 0.95. Figure 1-2 show the optimal values yielded by three solvers and their CPU time (in seconds) on solving (3.1) with $k = 1$ and the same sequence of 50 values of λ. By the results in Section 4, we select the 50 values of λ by

$$\lambda_i = \max\big(0.01, \gamma_i\|X\|_1/n\big) \;\text{ with }\; \gamma_i = \gamma_{\min}+((i-1)/49)(\gamma_{\max}-\gamma_{\min}) \quad (6.6)$$

for $i = 1, 2, \ldots, 50$, where $\gamma_{\min} = 0.02$, and $\gamma_{\max} = 0.25$ and 0.38 respectively for $\alpha = 0$ and 0.95. Such $\gamma_{\max}$ is such that $N_{\rm nz}(\beta^f)$ attains the value 0, where β^f represents the final output of a solver.

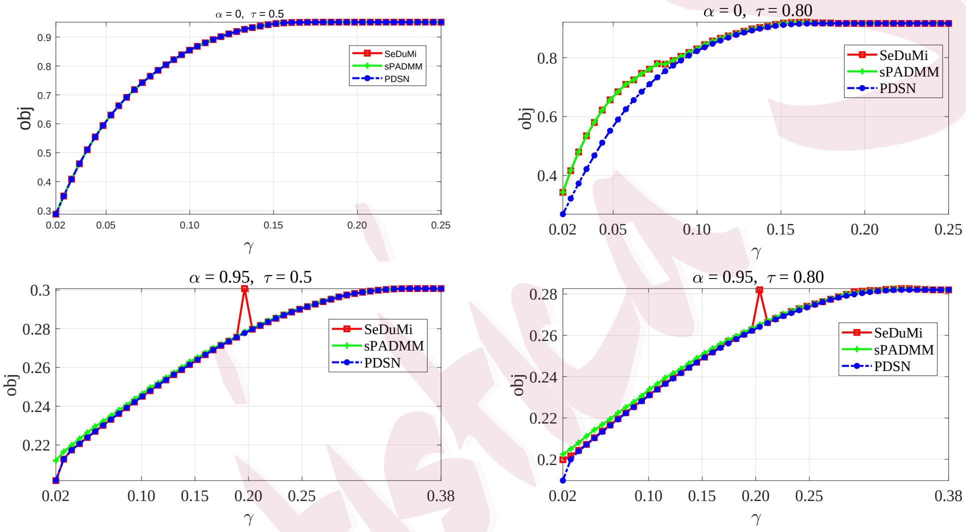

Figure 1: Optimal values of three solvers for the sample size $n = 500$

Figure 1 shows that the three solvers yield comparable optimal values, and the optimal values given by PDSN are a little better than those given by SeDuMi and sPADMM. Figure 2 shows that PDSN requires much less CPU time than SeDuMi and sPADMM do, and for $\alpha = 0.95$ the CPU time of the former is on average about 0.03 and 0.09 times that of SeDuMi and sPADMM, respectively, but for $\alpha = 0, \tau = 0.5$, when $\lambda < \lambda_3$, PDSN

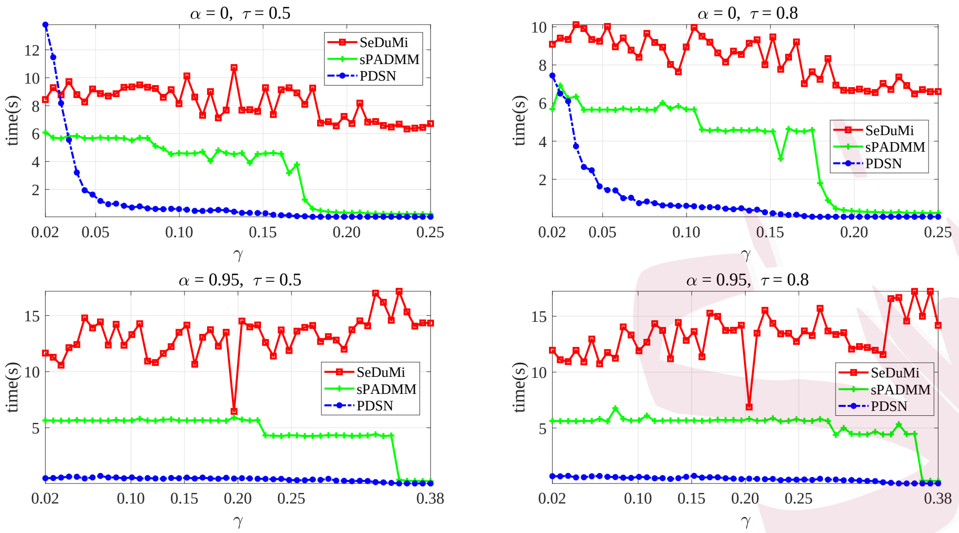

Figure 2: CPU times of three solvers for the sample size $n = 500$

requires more CPU time since the Clarke Jacobians are close to singularity. This shows that if the parameter λ in the model is not too small (a common setting for sparsity), PDSN is superior to SeDuMi and sPADMM in terms of the optimal value and CPU time. We find that sPADMM always attains the maximum number of iterations 3000 for all test problems (it even attains the maximum number of iterations if $j_{\max} = 10000$). Since $j_{\max} = 3000$ is used here, its CPU time is less than that of SeDuMi.

6.2. Numerical performance of Algorithm 1

We first apply MSCRA_PPA to the example in Section 3.1 of Wang et al. (2012), i.e., solve (2.6) with $\nu = \lambda^{-1}$ for $\lambda = \max(0.01, 0.1\|X\|_1/n)$, for which the scalar response is generated according to the heteroscedastic location-scale model $Y = X_6+X_{12}+X_{15}+X_{20}+0.7X_1\varepsilon$, where $\varepsilon \sim N(0,1)$ is independent of the covariates. Table 1 reports its identification performance for $\tau = 0.3, 0.5$ and 0.7 under different sample size, where **Size**, **AE**, P_1 and P_2 have the same meaning as in Wang et al. (2012). We see that, for $\tau = 0.5$, P_2 always equals 0. So, the check loss with $\tau = 0.5$ can not identify X_1, but the check loss with $\tau = 0.3$ and 0.7 can identify X_1 and the proportion of identifying X_1 increases as n becomes large.

Next we use a synthetic example to show that MSCRA_PPA can solve efficiently a series of zero-norm regularized problems (2.3) with different τ but a fixed λ. We generate an i.i.d. standard normal random vector $\beta^*_{S^*}$ with $s^* = \lfloor 0.5\sqrt{p} \rfloor$ entries of S^* chosen randomly from $\{1,\ldots,p\}$ for $p = 15000$, and then obtain the response vector y from model (2.1), where $x_i^{\mathbb{T}} \sim N(0,\Sigma)$ for $i = 1,\ldots,n$ with $\Sigma = 0.6E + 0.4I$ and $n = \lfloor 2s^* \log p \rfloor$, and the noise ε_i is from the Laplace distribution with density $d(u) = 0.5\exp(-|u|)$. Here, E is a $p\times p$ matrix of all ones. Figure 3 describes the average absolute ℓ_2-error $\|\widehat{\beta}^f - \beta^*\|$ and time when applying MSCRA_PPA to 10 test problems for $\tau \in \{0.05, 0.1, 0.15, \ldots, 0.95\}$ with $\nu = \lambda^{-1}$ and $\lambda = 37.5/n$. We see that MSCRA_PPA yields better ℓ_2-errors for τ close to 0.5, and worse ℓ_2-errors for τ close to 0 or 1. So, for this class of noises, the check loss with τ close to 0.5 is suitable. The MSCRA_PPA yields a desired solution for all test

Table 1: Identification performance of MSCRA_PPA

		$n = 250$	$n = 300$	$n = 400$	$n = 500$
$\tau = 0.3$	Size	11.800(4.369)	9.320(3.146)	6.290(1.472)	5.330(0.697)
	P_1	0.81	0.83	0.93	0.91
	P_2	0.81	0.83	0.93	0.91
	AE	0.197(0.174)	0.170(0.165)	0.176(0.155)	0.145(0.127)
$\tau = 0.5$	Size	10.960(3.075)	7.910(2.060)	5.270(1.171)	4.370(0.597)
	P_1	1.00	1.00	1.00	1.00
	P_2	0.00	0.00	0.00	0.00
	AE	0.034(0.014)	0.027(0.011)	0.021(0.010)	0.018(0.008)
$\tau = 0.7$	Size	12.590(4.356)	8.320(2.169)	6.310(1.308)	5.380(0.693)
	P_1	0.79	0.88	0.91	0.93
	P_2	0.79	0.88	0.91	0.93
	AE	0.183(0.175)	0.220(0.180)	0.151(0.146)	0.162(0.142)

problems in 40 seconds, and the CPU time for τ close to 0 or 1 is about 1.5 times that of τ close to 0.5. This means that it is an efficient solver for a series of zero-norm regularized problems in (2.3).

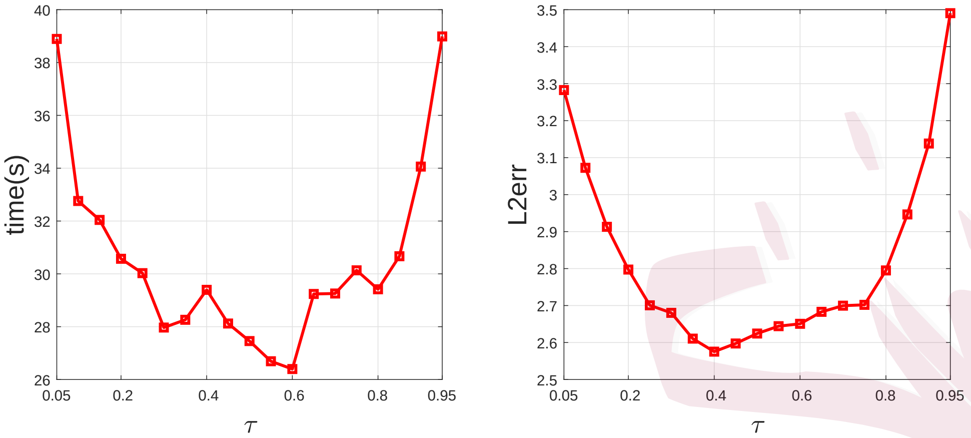

Figure 3: Performance of MSCRA_PPA under different quantile level τ

7. Conclusions

We have proposed a multi-stage convex relaxation approach, MSCRA_PPA, for computing a desirable approximation to the zero-norm penalized QR, which is defined as a global minimizer of an NP-hard problem. Under the common RSC condition and a mild restriction on the noises, we established the error bound of every iterate to the true estimator and the linear rate of convergence of the iterate sequence in a statistical sense. Numerical comparisons with MSCRA_IPM and MSCRA_ADMM show that MSCRA_PPA yields a comparable estimation performance within much less time.

Supplementary Materials

The online supplementary material consists of five parts. Appendix A includes some preliminary knowledge on generalized subdifferentials and

Clarke Jacobian, and some lemmas used in Section 2-5; Appendix B includes the proof of Theorem 2 and Theorem 3; Appendix C introduces the semismooth Newton method and the semi-proximal ADMM in Gu and Zou (2016); Appendix D includes performance comparisons of MSCRA_IPM, MSCRA_ADMM and MSCRA_PPA on some synthetic data and real data.

Acknowledgements

The authors would like to give their sincere thanks to two anonymous reviewers for their helpful comments. The authors would like to express their sincere thanks to Professor Kim-Chuan Toh from National University of Singapore for giving them some help on the implementation of Algorithm 2 when he visited SCUT. This work is supported by the National Natural Science Foundation of China under project Nos. 11971177 and 11701186.

References

Belloni, A. and Chernozhukov, V. (2011). ℓ_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics 39*, pp. 82–130.

Belloni, A., Chernozhukov, V. and Wang, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika 4*, pp. 791–806.

Bickel, P. and Li, B. (2006). Regularization in Statistics. *Sociedad de Estadística e Investigación Operativa Test 15*, pp. 271–344.

Bickel, P., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics 37*, pp. 1705–1732.

Banerjee, A., Chen, S., Fazayeli, F. and Sivakumar, V. (2015). Estimation with norm regularization. *Advances in Neural Information Processing Systems 2*, pp. 1556–1564.

Breiman, L. (1996). *Heuristics of instability and stabilization in model selection. The Annals of Statistics 24*, pp. 2350–2383.

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.

Facchinei, F. and Pang, J. S. (2003). Finite-dimensional Variational Inequalities and Complementarity Problems. Springer, New York, 2003.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica 20*, pp. 101–148.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistics Association 96*, pp. 1348–1360.

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

Fan, J., Xue, L. Z. and Zou, H. (2014). Strong oracle optimality of folded concave penalized estimation. *The Annals of Statistics 42*, pp. 819–849.

Frank, L. E., and Friedman, J. H. (1993). A statistical view of some chemometrices regression tools. *Technometrics 35*, pp. 109–135.

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.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society 46*, pp. 33–50.

Koenker, R. and Bassett, G. (1982). Robust tests for hereroscedasticity based on regression quantiles. *Econometrica: Journal of the Econometric Society 50*, pp. 43–61.

Lemarechal, C. and Sagastizsábal, C. (1997). Practical aspects of the Moreau-Yosida regularization: theoretical preliminaries. *SIAM Journal on Optimization 7*, pp. 367–385.

Lange, K., Hunter, D. R. and Yang, I. (2000). Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics 9*, pp. 1–20.

Liu, Y. L., Bi, S. J. and Pan, S. H. (2018). Equivalent Lipschitz surrogates for zero-norm and rank optimization problems. *Journal of Global Optimization 72*, pp. 679-704.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics 38*, pp. 894-942.

Zhang, T. (2010). Analysis of Multi-stage Convex Relaxation for Sparse Regularization. *Journal of Machine Learning Research 11*, pp. 1081-1107.

Negahban, S., Ravikumar P., Wainwright, M. J. and Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science 27*, pp. 538–557.

Peng, B. and Wang, L. (2015). An iterative coordinate descent algorithm for high-dimensional nonconvex penalized quantile regression. *Journal of Computational and Graphical Statistics 24*, pp. 676–694.

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

Raskutti, G., Wainwright, M. J. and Yu, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *Journal of Machine Learning Research 11*, pp. 2241–2259.

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.

Sturm, J. F. (1999). Using SeDuMi 1.02, a Matlab toolbox for optimization over symmetric cones. *Optimization Methods and Software 11*, pp. 625–653.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B 58*, pp. 267–288.

Tang, P. P., Wang, C. J., Sun, D. F. and Toh, K. C. (2019). A sparse semismooth Newton based proximal majorization-minimization algorithm for nonconvex square-root-loss regression problems. *arXiv:1903.11460v1*.

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.

van de Geer, S. A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics 3*, pp. 1360–1392.

Wu, T. T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statiscs 2*, pp. 224–244.

Wu, Y. C. and Liu, Y. F. (2009). Variable selection in quantile regression. *Statistica Sinica 19*, pp. 801–817.

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.

Wang, L. (2013). The L_1 penalized LAD estimator for high dimensional linear regression.

Journal of Multivariate Analysis 120, pp. 135–151.

Yi, C. R. and Huang, J. (2017). Semismooth Newton coordinate descent algorithm for elastic-net penalized Huber loss regression and quantile regression. *Journal of Computational and Graphical Statistics 26*, pp. 547–557.

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

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association 101*, pp. 1418–1429.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics 36*, pp. 1509–1533.

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)