ADMM for Sparse-Penalized Quantile Regression with Non-Convex Penalties

Reza Mirzaeifard*, Naveen K. D. Venkategowda§, Vinay Chakravarthi Gogineni*, Stefan Werner*
*Dept. of Electronic Systems, Norwegian University of Science and Technology-NTNU, Norway
§Department of Science and Technology, Linköping University, Sweden
E-mails: {reza.mirzaeifard, vinay.gogineni, stefan.werner}@ntnu.no, naveen.venkategowda@liu.se

***Abstract*—This paper studies quantile regression with non-convex and non-smooth sparse-penalties, such as minimax concave penalty (MCP) and smoothly clipped absolute deviation (SCAD). Although iterative coordinate descent and local linear approximation techniques can solve quantile regression problem, convergence is slow for MCP and SCAD penalties. However, alternating direction method of multipliers (ADMM) can be exploited to enhance the convergence speed. Hence, this paper proposes a new ADMM algorithm with an increasing penalty parameter, called IAD, to handle sparse-penalized quantile regression. We first investigate the convergence of the proposed algorithm and establish the conditions for convergence. Then, we present numerical results to demonstrate the efficacy of the proposed algorithm. Our results show that the proposed IAD algorithm can handle sparse-penalized quantile regression more effectively than the state-of-the-art methods.**

***Index Terms*—Quantile regression, non-smooth and non-convex penalties, ADMM, sparse learning.**

I. Introduction

Most regression algorithms aim to estimate the conditional mean of a response variable associated with a set of observations [1]. However, mean-based regression is sensitive to outliers and cannot relate the response variable to another point, or range, of the conditional distribution, e.g., the median or a certain percentile. Alternatively, quantile regression, which provides more comprehensive regression relationships based on quantiles, can be used in such situations [2]. Consequently, it has been found useful in various applications, such as predicting regional wind power [3], estimating uncertainty in electricity smart meter data [4], and forecasting load in smart grids [5].

Real-world applications, such as quantitative traits in genetic [6], and gene selection for microarray gene expression [7], require estimation of models that tend to be sparse. By using the *a priori* information on sparsity, one can achieve better results compared to the conventional quantile regression methods. Hence, sparse-penalized quantile regression has attracted substantial research interest [8], [9]. Quantile regression with l_1-penalty performs well when estimating highly sparse models but suffers when estimating moderately- or non-sparse models as the l_1-penalty uniformly shrinks all coefficients of the model toward zero. Therefore, l_1-penalized quantile regression offer poor performance and bias as the model sparsity decreases. To overcome this problem, we require a sparse penalty that can distinguish between zero and non-zero coefficients of the model. Minimax concave penalty (MCP) [10] and smoothly clipped absolute deviation (SCAD) [11] can accomplish these requirements. MCP and SCAD penalties shrink the model coefficients selectively towards zero, i.e., the shrinkage is limited to zero coefficients. Despite encouraging sparse solutions, these penalties alleviate the bias effect of the l_1-penalty.

Conventionally, linear programming [12], [13] and sub-gradient methods [14] have been exploited to solve l_1-penalized quantile regression. However, these algorithms can not be employed when the penalty functions are non-convex. For solving folded concave penalized regression, a general framework based on local linear approximation (LLA) algorithm was proposed in [15]. For a similar problem setup, iterative coordinate descent algorithm (QICD) was proposed and its convergence was established in [16]. However, the aforementioned approaches are computationally intensive and have a slow convergence rate.

To ease the computational burden associated with the sparse-penalized quantile regression, two algorithms based on alternating direction method of multiplier (ADMM), namely, sparse coordinate descent ADMM (scdADMM) and proximal ADMM (pADMM), have been proposed in [17]. The non-convex penalties were addressed through the LLA framework. A similar framework was also used to deal with non-convex penalties in sparse-penalized composite quantile regression [18]. Since LLA is an iterative method, approaches in [17], [18] exhibit slow convergence. On the other hand, recent works have studied ADMM-based non-convex optimization [19]–[22], but they require that the objective function satisfy Lipschitz differentiability conditions. Thus, ADMM-based optimization methods still need to be improved to work effectively in non-smooth and non-convex settings without assuming Lipschitz differentiability.

In this work, we propose an ADMM based algorithm with time-increasing penalty parameters (IAD) to solve the quantile regression problem with non-convex and non-smooth sparse penalties such as MCP and SCAD. With a time increasing ADMM penalty parameter, we show that the accumulation of ascent changes in the augmented Lagrangian from the dual update step and ADMM penalty parameter update step can be capped by a constant value. This enables to prove that the convergence can be guaranteed. Furthermore, we validate our

This work was supported by the Research Council of Norway.

theoretical claims through numerical simulations. Simulation results demonstrate that the proposed algorithm exhibits better accuracy than state-of-the-art methods such as QICD [16] and the LLA framework with scdADMM (LSCD) or pADMM (LPA) [17].

Mathematical Notations: Bold letters $\mathbf{a}$ and $\mathbf{A}$ are used to represent vectors and matrices respectively. The transpose of $\mathbf{A}$ is denoted as $\mathbf{A}^{\mathrm{T}}$. The jth column of a matrix $\mathbf{A}$ is denoted as $\mathbf{A}_{:,j}$, and the jth element of a vector $\mathbf{x}$ is denoted as x_j. In addition, we let $\mathbf{A}_{<s}\mathbf{x}_{<s} := \sum_{i<s} \mathbf{A}_{:,i}x_i$ and, in a similar fashion, $\mathbf{A}_{>s}\mathbf{x}_{>s} := \sum_{i>s} \mathbf{A}_{:,i}x_i$. Moreover, for a function $h : \mathbb{R}^n \to \mathbb{R}$ and penalty parameter $\gamma > 0$, the proximal function is defined as: $\mathrm{Prox}_h(w;\gamma) = \arg\min_x \left\{h(x) + \frac{1}{2\gamma}\|x-w\|^2\right\}$. Furthermore, for a scalar variable u, and penalty parameter α, $\mathrm{Shrink}(u,\alpha) = \frac{u}{|u|}\max\{0, |u| - \alpha\}$. Finally, $\partial f(u)$ represents the sub-gradient of $f(.)$ at u.

II. Sparse quantile regression

For a scalar random variable Y and a $P \times 1$ vector of covariates $\boldsymbol{\chi}$, let us consider $F_Y(y|\mathbf{x}) = P(Y \leq y|\boldsymbol{\chi} = \mathbf{x})$ as the conditional cumulative distribution function and $Q_Y(\tau|\mathbf{x}) = \inf\{y : F_Y(y|\mathbf{x}) \geq \tau\}$ as the τth conditional quantile for $\tau \in (0,1)$. The linear quantile regression model relates $Q_Y(\tau|\mathbf{x})$ and $\mathbf{x} \in \mathbb{R}^P$ as [23]

$$Q_Y(\tau|\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_\tau, \tag{1}$$

where $\boldsymbol{\beta}_\tau \in \mathbb{R}^P$ is the regression model parameters, which must be estimated. Given the data pairs $\{\mathbf{x}_i, y_i\}_{i=1}^n$ and a specific value of τ, the unknown model parameter can be estimated by solving the following optimization problem [23]:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\mathbf{w}), \tag{2}$$

where $\mathbf{w} = \boldsymbol{\beta}_\tau$, and $\rho_\tau(\mathbf{u}) = \frac{1}{2}(\|\mathbf{u}\|_1 + (2\tau-1)\mathbf{1}^{\mathrm{T}}\mathbf{u})$ which is also known as check loss function.

By penalizing the quantile regression loss function appropriately, one can take advantage of *a priori* information about the model coefficients and thereby enhance the inference quality. After incorporating the penalty $P_{\lambda,\gamma}(\mathbf{w})$, the optimization problem (2) takes the form

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\mathbf{w}) + P_{\lambda,\gamma}(\mathbf{w}). \tag{3}$$

By defining an auxiliary variable $\mathbf{z}$, (3) can be rewritten as

$$\begin{aligned} \min_{\{\mathbf{w},\mathbf{z}\}} \quad & \frac{1}{2}(\|\mathbf{z}\|_1 + (2\tau-1)\mathbf{1}_n^{\mathrm{T}}\mathbf{z}) + n\, P_{\lambda,\gamma}(\mathbf{w}), \\ \text{subject to} \quad & \mathbf{z} + \mathbf{X}\mathbf{w} = \mathbf{y}, \end{aligned} \tag{4}$$

where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times P}$.

Although, LASSO [24], [25] is a popular choice as penalty function, it leads to estimation bias. To overcome this limitation, MCP and SCAD penalties can be used $P_{\lambda,\gamma}(\mathbf{w}) = \sum_{p=1}^{P} g_{\lambda,\gamma}(w_p)$, [26]. The definitions of MCP [10] and SCAD [11] are given by:

$$g_{\lambda,\gamma}^{\text{MCP}}(w_p) = \begin{cases} \lambda|w_p| - \frac{w_p^2}{2\gamma}, & |w_p| \leq \gamma\lambda \\ \frac{\gamma\lambda^2}{2}, & |w_p| > \gamma\lambda \end{cases} \quad \text{for } \gamma \geq 1, \tag{5}$$

and

$$g_{\lambda,\gamma}^{\text{SCAD}}(w_p) = \begin{cases} \lambda|w_p|, & |w_p| \leq \lambda \\ -\frac{|w_p|^2 - 2a\lambda|w_p| + \lambda^2}{2(\gamma-1)}, & \lambda < |w_p| \leq \gamma\lambda \\ \frac{(\gamma+1)\lambda^2}{2}, & |w_p| > \gamma\lambda \end{cases} \text{ for } \gamma \geq 2. \tag{6}$$

Further, MCP and SCAD are weakly convex for $\rho \geq \frac{1}{\gamma}$ and $\rho \geq \frac{1}{\gamma-1}$, respectively [26].

The LLA framework solves this sparse penalized-quantile regression (4) in an iterative procedure by obtaining the sub-gradients of these penalties, i.e., it solves LASSO quantile regression in each iteration. In the next section, we present an ADMM-based algorithm that alleviates the outer loop that appears in the LLA framework.

III. ADMM for sparse-penalized quantile regression

In order to employ the ADMM for solving (4), we write the associated augmented Lagrangian function as

$$\begin{aligned} \mathcal{L}_{\rho_\lambda}(\mathbf{w},\mathbf{z},\boldsymbol{\lambda}) = \frac{1}{2}(\|\mathbf{z}\|_1 + (2\tau-1)\mathbf{1}_n^{\mathrm{T}}\mathbf{z}) + nP_{\lambda,\gamma}(\mathbf{w}) \\ + \boldsymbol{\lambda}^{\mathrm{T}}(\mathbf{z} + \mathbf{X}\mathbf{w} - \mathbf{y}) + \frac{\rho_\lambda}{2}\|\mathbf{z} + \mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2, \end{aligned} \tag{7}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^n$ is Lagrange multiplier, and ρ_λ is the penalty parameter. ADMM tries to find a saddle point for the augmented Lagrangian function in an iterative procedure [27]. The $(k+1)$th iteration of a standard two-block ADMM algorithm can be expressed as [27]

$$\mathbf{w}^{(k+1)} = \arg\min_{\mathbf{w}} \mathcal{L}_{\rho_\lambda}(\mathbf{w}, \mathbf{z}^{(k)}, \boldsymbol{\lambda}^{(k)}), \tag{8a}$$

$$\mathbf{z}^{(k+1)} = \arg\min_{\mathbf{z}} \mathcal{L}_{\rho_\lambda}(\mathbf{w}^{(k+1)}, \mathbf{z}, \boldsymbol{\lambda}^{(k)}), \tag{8b}$$

$$\boldsymbol{\lambda}^{(k+1)} = \boldsymbol{\lambda}^{(k)} + \rho_\lambda(\mathbf{z}^{(k+1)} + \mathbf{X}\mathbf{w}^{(k+1)} - \mathbf{y}). \tag{8c}$$

It can be seen that for optimization in (4), the primal update can be written as

$$\begin{aligned} \mathbf{w}^{(k+1)} = \arg\min_{\mathbf{w}} n\sum_{p=1}^{P} g_{\lambda,\gamma}(w_p) + (\boldsymbol{\lambda}^{(k)})^{\mathrm{T}}\mathbf{X}\mathbf{w} \\ + \rho_\lambda(\mathbf{z}^{(k)} - \mathbf{y})^{\mathrm{T}}(\mathbf{X}\mathbf{w}) + \frac{\rho_\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w}. \end{aligned} \tag{9}$$

Although each basic function $g_{\lambda,\gamma}(w_p)$ has a closed-form proximal function, for a general design matrix $\mathbf{X}$, there is no simple closed-form solution to obtain $\mathbf{w}$. This update can be carried out by block coordinate descent iteratively [27],or by a variety of ADMM methods that are capable of looking at each element separately. However, to the best of our knowledge, any existing ADMM based approaches [19]–[22] cannot guarantee convergence in the absence of Lipschitz differentiability and

convexity of the objective function, as is the case with our problem.

To solve the challenge mentioned above, we propose an ADMM-based algorithm, which is referred as IAD, with an time-increasing penalty parameter. First, by having $\beta > 1$ in each iteration, the ρ_λ can be updated as:

$$\rho_\lambda^{(k+1)} = \beta \rho_\lambda^{(k)}. \tag{10}$$

Moreover, the update of $\mathbf{w}$ is split into P steps. The pth element of $\mathbf{w}$ is updated in the pth iteration as follows:

$$w_p^{(k+1)} = \arg\min_{w_p} \mathcal{L}_{\rho_\lambda^{(k+1)}}(\mathbf{w}_{<p}^{(k+1)}, w_p, \mathbf{w}_{>p}^{(k)}, \mathbf{z}^{(k)}, \boldsymbol{\lambda}^{(k)}). \tag{11}$$

After several simplifications, we can see that the update of the pth element of $\mathbf{w}$ is given by

$$\begin{aligned} w_p^{(k+1)} &= \arg\min_{w_p} n g_{\lambda,\gamma}(w_p) + \frac{\rho_\lambda^{(k+1)} \|\mathbf{X}_{:,p}\|_2^2}{2} \big\| w_p - a_p \big\|_2^2, \\ &= \text{Prox}_{g_{\lambda,\gamma}}(a_p; \frac{n}{\rho_\lambda^{(k+1)} \|\mathbf{X}_{:,p}\|_2^2}), \end{aligned} \tag{12}$$

where $a_p = \frac{-\mathbf{X}_{:,p}^{\mathrm{T}} \mathbf{X}_{<p} \mathbf{w}_{<p}^{(k+1)} - \mathbf{X}_{:,p}^{\mathrm{T}} \mathbf{X}_{>p} \mathbf{w}_{>p}^{(k)} - (\frac{\boldsymbol{\lambda}^{(k)}}{\rho_\lambda^{(k+1)}} + \mathbf{y} - \mathbf{z})^{\mathrm{T}} \mathbf{X}_p}{\|\mathbf{X}_{:,p}\|_2^2}$. Both MCP and SCAD admit closed-form solutions of the proximal operator [28]. Next, the update of $\mathbf{z}$ can be formulated as:

$$\begin{aligned} \mathbf{z}^{(k+1)} &= \arg\min_{\mathbf{z}} \mathcal{L}_{\rho_\lambda^{(k+1)}}(\mathbf{w}^{(k+1)}, \mathbf{z}, \boldsymbol{\lambda}^{(k)}) \\ &= \arg\min_{\mathbf{z}} \frac{1}{2}(\|\mathbf{z}\|_1 + (2\tau - 1)\mathbf{1}_n^{\mathrm{T}} \mathbf{z}) + (\boldsymbol{\lambda}^{(k)})^{\mathrm{T}} \mathbf{z} \\ &\qquad + \frac{\rho_\lambda^{(k+1)}}{2} \|\mathbf{z} + \mathbf{X}\mathbf{w}^{(k+1)} - \mathbf{y}\|_2^2. \end{aligned} \tag{13}$$

It can shown that the update step of $\mathbf{z}$ in ADMM has a closed-form solution. By merging $(\tau - \frac{1}{2})\mathbf{1}_n^{\mathrm{T}} \mathbf{z}$, $(\boldsymbol{\lambda}^{(k)})^{\mathrm{T}} \mathbf{z}$, and $\|\mathbf{z} + \mathbf{X}\mathbf{w}^{(k+1)} - \mathbf{y}\|_2^2$ together, a component-wise solution can be obtained as

$$\begin{aligned} \mathbf{z}^{(k+1)} &= \arg\min_{\mathbf{z}} \frac{\rho_\lambda^{(k+1)}}{2} \|\mathbf{z}\|_1 + \frac{1}{2} \|\mathbf{z} - \boldsymbol{\alpha}\|_2^2 \\ &= \text{Shrink}\big(\alpha_i, \frac{\rho_\lambda^{(k+1)}}{2}\big)_{i=1}^n, \end{aligned} \tag{14}$$

where $\boldsymbol{\alpha} = (\mathbf{y} - \mathbf{X}\mathbf{w}^{(k+1)}) - \frac{\boldsymbol{\lambda}^{(k)} + (\tau - \frac{1}{2})\mathbf{1}_n}{\rho_\lambda^{(k+1)}}$. Finally, the update of dual variable $\boldsymbol{\lambda}$ is given by

$$\boldsymbol{\lambda}^{(k+1)} = \boldsymbol{\lambda}^{(k)} + \rho_\lambda^{(k+1)}(\mathbf{z}^{(k+1)} + \mathbf{X}\mathbf{w}^{(k+1)} - \mathbf{y}). \tag{15}$$

The proposed ADMM based method for solving the sparse-penalized quantile regression is summarized in Algorithm 1.

It is worth mentioning that the stopping criterion in [27] can be adapted to our problem as:

$$\begin{aligned} &\|\mathbf{z}^{(k+1)} + \mathbf{X}\mathbf{w}^{(k+1)} - \mathbf{y}\|_2 \le \sqrt{n}\epsilon_1 \\ &\quad + \epsilon_2 \max\{\|\mathbf{X}\mathbf{w}^{(k+1)}\|_2, \|\mathbf{z}^{(k+1)}\|_2, \|\mathbf{y}^{(k+1)}\|_2\} \end{aligned} \tag{16a}$$

Algorithm 1: ADMM with time-increasing penalty parameter (IAD) for sparse-penalized quantile regression

Initialize $\mathbf{w}^{(0)}$, $\mathbf{z}^{(0)}$, $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\lambda}^{(0)}$ to zero vectors and $\beta > 1$;
repeat
  Update $\rho_\lambda^{(k+1)}$ by (10);
  for $p = 1, \ldots, P$ **do**
    Update $w_p^{(k+1)}$ by (12);
  end
  Update $\mathbf{z}^{(k+1)}$ by (14);
  Update $\boldsymbol{\lambda}^{(k+1)}$ by (15);
until *the convergence criterion in* (16) *is met*;

$$\begin{aligned} &\rho_\lambda^{(k+1)} P \max_p \|\mathbf{X}_{:,p}\|_2^2 \|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|_2 \\ &+ \rho_\lambda^{(k+1)} \|\mathbf{X}^{\mathrm{T}}(\mathbf{z}^{(k+1)} - \mathbf{z}^{(k)})\|_2 \le \sqrt{P}\epsilon_1 + \epsilon_2 \|\mathbf{X}^{\mathrm{T}} \boldsymbol{\lambda}^{(k+1)}\|_2, \end{aligned} \tag{16b}$$

where a typical choice for ϵ_1 and ϵ_2 is 10^{-3}. Alternatively, the algorithm can be terminated when the number of iterations exceeds a certain number.

IV. Convergence proof

The convergence proof is established by corroborating that the accumulation of ascents changes in the augmented Lagrangian in all iterations throughout the $\boldsymbol{\lambda}$ update step, and penalty parameter update step, can be upper bounded by a constant value. Therefore, since the augmented Lagrangian is lower bounded, the weak convexity of MCP, SCAD, and $\rho_\tau(\cdot)$ guarantees the convergence.

Lemma 1. *The augmented Lagrangian* $\mathcal{L}_{\rho_{(k+1)}}(\mathbf{w}^{(k+1)}, \mathbf{z}^{(k+1)}, \boldsymbol{\lambda}^{(k+1)})$ *is lower bounded.*

Proof. Since $\rho_\tau(\mathbf{z})$, and $P_{\lambda,\gamma}(\mathbf{w})$ are non-negative functions, and $\boldsymbol{\lambda}^{(k+1)} \in \partial \rho_\tau(\mathbf{z}^{(k+1)}) \subseteq [\tau - 1, \tau]^n$, the augmented Lagrangian function is lower bounded. □

Lemma 2. *By* $\rho = \frac{1}{\gamma}$ *or* $\rho = \frac{1}{\gamma - 1}$, *for MCP or SCAD respectively, since* $\rho_\lambda^{(k)} > \frac{2n\rho}{\min_p \|\mathbf{X}_{:,p}\|_2^2}$, *by having* $\Omega^{(k)} = \rho_\lambda^{(k)} - \frac{n\rho}{\min_p \|\mathbf{X}_{:,p}\|_2^2}$ *the following inequality holds:*

$$\begin{aligned} &\mathcal{L}_{\rho_\lambda^{(k)}}(\mathbf{w}^{(k)}, \mathbf{z}^{(k)}, \boldsymbol{\lambda}^{(k)}) - \mathcal{L}_{\rho_\lambda^{(k-1)}}(\mathbf{w}^{(k-1)}, \mathbf{z}^{(k-1)}, \boldsymbol{\lambda}^{(k-1)}) \le \\ &\quad - \frac{\Omega^{(k)}}{2} \|\mathbf{w}^{(k)} - \mathbf{w}^{(k-1)}\|_2^2 - \frac{\rho_\lambda^{(k)}}{2} \|\mathbf{z}^{(k)} - \mathbf{z}^{(k-1)}\|_2^2 \\ &\qquad + \frac{P}{\rho_\lambda^{(k)}} + \frac{P(\beta - 1)}{2\rho_\lambda^{(k-1)}}. \end{aligned} \tag{17}$$

Proof. The second term of (17) comes from the weak convexity of MCP or SCAD, and $\rho_\tau(\cdot)$. Also, as long as $\lambda^{(k)} \in [\tau - 1, \tau]^n$, $\forall k \ge 1$, $\|\boldsymbol{\lambda}^{(k)} - \boldsymbol{\lambda}^{(k-1)}\|_2 \le P$; therefore, the the augmented Lagrangian's change in the $\boldsymbol{\lambda}$ update step can be bounded by $\frac{P}{\rho_\lambda^{(k)}}$. Moreover, by utilizing (10) to update ρ_λ, the augmented Lagrangian's change in this step can be bounded

by $\frac{P(\beta-1)}{2\rho_\lambda^{(k-1)}}$. Therefore, by considering the update steps of $\boldsymbol{\lambda}$ and ρ_λ together, the third term of (17) can be obtained. □

Theorem 1. *For suitable values if* $\rho_\lambda^{(0)}$, *satisfying* $\rho_\lambda^{(0)} > \frac{n\rho}{\min_p \|\mathbf{X}_{:,p}\|_2^2}$, *where* $\rho > \frac{1}{\gamma}$ *or* $\rho > \frac{1}{\gamma-1}$ *for MCP or SCAD, and* $\beta > 1$ *the following holds for the IAD algorithm:* $\lim_{k\to\infty} c^{(k)} = \lim_{k\to\infty} \|\mathbf{z}^{(k)} + \mathbf{X}\mathbf{w}^{(k)} - \mathbf{y}\|_2^2 + \rho_\lambda^{(k)} P^2 \|\mathbf{X}_{:,p}\|_2^4 \|\mathbf{w}^{(k)} - \mathbf{w}^{(k-1)}\|_2^2 + \rho_\lambda^{(k)} \|\mathbf{X}^T(\mathbf{z}^{(k)} - \mathbf{z}^{(k-1)})\|_2^2 \to 0$.

Proof. By considering $\Omega^{(k)} = \rho_\lambda^{(k)} - \frac{n\rho}{\min_p \|\mathbf{X}_{:,p}\|_2^2}$, we get:

$$\mathcal{L}_{\rho_\lambda^{(K)}}(\mathbf{w}^{(K)}, \mathbf{z}^{(K)}, \boldsymbol{\lambda}^{K}) - \mathcal{L}_{\rho_\lambda^{(0)}}(\mathbf{w}^{(0)}, \mathbf{z}^{(0)}, \boldsymbol{\lambda}^{(0)}) \leq$$

$$\sum_{k=1}^{K} \Big(\frac{P}{\rho_\lambda^{(k)}} + \frac{P(\beta-1)}{2\rho_\lambda^{(k)}} - \frac{\Omega^{(k)}}{2} \|\mathbf{w}^{(k)} - \mathbf{w}^{(k-1)}\|_2^2$$

$$- \frac{\rho_\lambda^{(k)}}{2} \|\mathbf{z}^{(k)} - \mathbf{z}^{(k-1)}\|_2^2 \Big) \leq \frac{P\rho_\lambda^{(1)}\beta}{\beta-1} + \frac{P\rho_\lambda^{(0)}\beta}{2} - S_K,$$

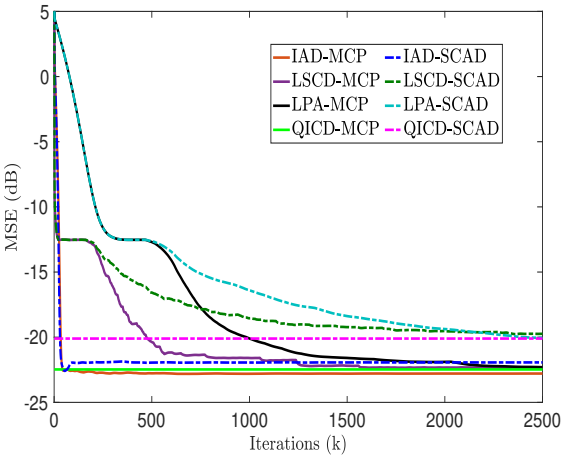
where $S_K = \sum_{k=1}^{K} \Big(\frac{\rho_\lambda^{(k)}}{2} \|\mathbf{z}^{(k)} - \mathbf{z}^{(k-1)}\|_2^2 + \frac{\Omega^{(k)}}{2} \|\mathbf{w}^{(k)} - \mathbf{w}^{(k-1)}\|_2^2 \Big)$. Since the augmented Lagrangian is lower bounded and each element in the sum S_K is non-negative, we have $\lim_{K\to\infty} S_K < \infty$. Therefore, $\lim_{k\to\infty} \frac{\Omega_\lambda^{(k)}}{2} \|\mathbf{w}^{(k)} - \mathbf{w}^{(k-1)}\|_2^2 = 0$, and $\lim_{k\to\infty} \frac{\rho_\lambda^{(k)}}{2} \|\mathbf{z}^{(k)} - \mathbf{z}^{(k-1)}\|_2^2 = 0$. Moreover, $\|\mathbf{z}^{(k)} + \mathbf{X}\mathbf{w}^{(k)} - \mathbf{y}\|_2^2 = \frac{\|\boldsymbol{\lambda}^{(k)} - \boldsymbol{\lambda}^{(k-1)}\|_2^2}{(\rho_\lambda^{(k-1)})^2}$ goes to zero when k goes to ∞. Therefore, it can be guaranteed that the summation of residuals defined as $c^{(k)} \|\mathbf{z}^{(k)} + \mathbf{X}\mathbf{w}^{(k)} - \mathbf{y}\|_2^2 + \rho_\lambda^{(k)} P^2 \|\mathbf{X}_{:,p}\|_2^4 \|\mathbf{w}^{(k)} - \mathbf{w}^{(k-1)}\|_2^2 + \rho_\lambda^{(k)} \|\mathbf{X}^{\mathrm{T}}(\mathbf{z}^{(k)} - \mathbf{z}^{(k-1)})\|_2^2$ converges to 0. □

V. Simulation Results

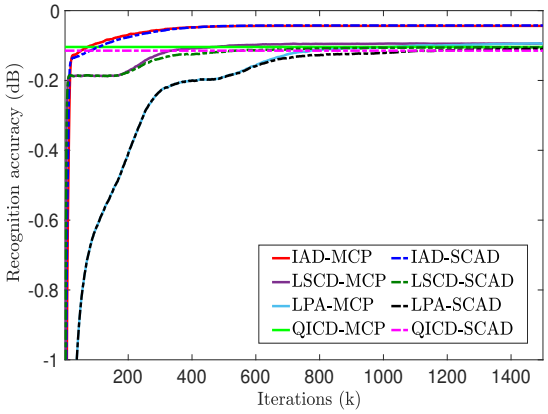
Here, we conduct experiments in the context of sparse quantile regression to demonstrate the effectiveness of the proposed IAD algorithm. For this we considered the following observation model:

$$\mathbf{y} = \sum_{p=1}^{P} \xi_p \mathbf{x}_p + \mathbf{x}_6 + \mathbf{x}_{12} + \mathbf{x}_{15} + \mathbf{x}_{20} + 0.7\boldsymbol{\epsilon}\mathbf{x}_1, \qquad (19)$$

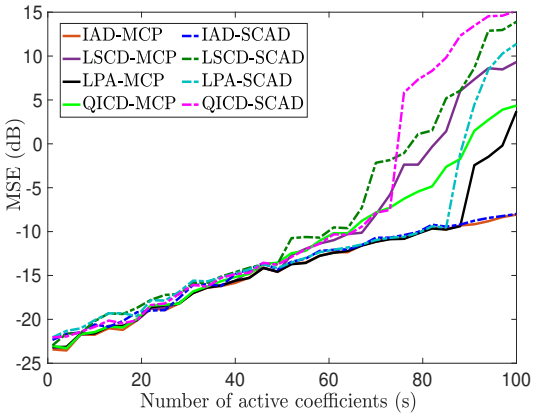
where $\boldsymbol{\epsilon} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1)$, and $\xi_p \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 10^{-6})$. We set $\mathbf{x}_1 = \Phi(\tilde{\mathbf{x}}_1)$ and $\mathbf{x}_p = \tilde{\mathbf{x}}_p$, otherwise. Here, $\Phi(\cdot)$ being the cumulative distribution function of $\mathcal{N}(0,1)$ and $(\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_P)^{\mathrm{T}} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}_{pq} = 0.5^{|p-q|}$. Under these settings the model to be estimated will be a compressible system [29]. The τth conditional quantile linear function can be achieved by $\sum_{p=1}^{P} \xi_p \mathbf{x}_p + \mathbf{x}_6 + \mathbf{x}_{12} + \mathbf{x}_{15} + \mathbf{x}_{20} + 0.7\Phi(\tau)^{-1}\mathbf{x}_1$. We simulate the proposed IAD algorithm to perform the quantile regression. In our case, $\gamma_{\text{SCAD}} = 4.1$, $\gamma_{\text{MCP}} = 2.1$, $\rho_\lambda^{(0)} = 3$, $\beta = 1.001$ and $\lambda = 0.055$. For comparative assessment, we also simulated QICD [16], LPA [17], and LSCD [17] to carry out the same quantile regression task. The mean square error (MSE): $\mathrm{E}\|\hat{\mathbf{w}} - \mathbf{w}\|_2^2$ was considered to be a performance metric.



(a)



(b)



(c)

Fig. 1: Performance Comparison: (a) MSE versus ADMM iterations k, (b) Accuracy of correctly recognizing active and non-active coefficients, (c) MSE versus the number of active coefficients s in model parameter $\boldsymbol{\beta}_\tau \in \mathbb{R}^P$.

In the first scenario, we compare the convergence rate, and the efficiency of these algorithms in terms of MSE. The simulation results were obtained by averaging over 100 independent trials for $(n, P) = (300, 100)$ and $\tau = 0.7$. The learning curves (MSE versus iterations) of the algorithms are shown in the Fig. 1a. Fig. 1a shows that the proposed IAD is

able to achieve a lower MSE than other existing approaches. Furthermore, the IAD algorithm exhibited faster convergence rate over LPA and LSCD.

In the second scenario, these algorithms were compared on the accuracy of recognizing active and non-active coefficients correctly. The accuracy measure is defined as the ratio of the number of active and non-active coefficients correctly identified to the total number of coefficients. Fig. 1b shows the accuracy versus iterations of these algorithms. Fig. 1b illustrates that the IAD can distinguish active and non-active coefficients more accurately compared to other methods. Thus, IAD was able to achieve lower MSE compared to other approaches.

Finally, in the third scenario, we compared the robustness of the algorithms under different levels of sparsity (i.e., number active coefficients increses from 1 to $P-1$). For this, we generated observation as $\mathbf{y} = \sum_{p=1}^{P} \xi_p \mathbf{x}_p + \sum_{i \in \mathcal{M}} \mathbf{x}_i + 0.7\boldsymbol{\epsilon}\mathbf{x}_1$, with $\mathcal{M} \in \{2, \ldots, P\}$, $\boldsymbol{\epsilon} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1)$, and $\xi_p \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 10^{-6})$. The iterations number was set large enough to ensure a fair comparison. Fig. 1c illustrates the MSE vs number of active coefficients for all algorithms. From Fig. 1c, it can be observed that the proposed IAD performs consistently against all sparsity-levels, varying from highly-sparse to non-sparse. Whereas other state-of-the-art approaches exhibit poor performance when the sparsity-level varies from moderately-sparse to non-sparse.

VI. CONCLUSIONS

In this paper, an ADMM based algorithm with time-increasing penalty parameters for the quantile regression penalized with non-convex and non-smooth sparse-penalties has been proposed. With our novel analysis, the convergence proof for the proposed algorithm has been conducted. The simulation results demonstrated that this single loop ADMM algorithm could achieve better MSE than the QICD method and the LLA framework. Also, this algorithm performs consistently against all sparsity-levels, especially in moderately-sparse or non-sparse, where other algorithms had shown worse results.

REFERENCES

[1] G. A. Seber and A. J. Lee, *Linear Regression Analysis*. John Wiley & Sons, 2012, vol. 329.

[2] R. Koenker and K. F. Hallock, "Quantile regression," *Journal of Economic Perspectives*, vol. 15, no. 4, pp. 143–156, Dec. 2001.

[3] Y. Yu, X. Han, M. Yang, and J. Yang, "Probabilistic prediction of regional wind power based on spatiotemporal quantile regression," *IEEE Transactions on Industry Applications*, vol. 56, no. 6, pp. 6117–6127, Dec. 2020.

[4] S. Ben Taieb, R. Huser, R. J. Hyndman, and M. G. Genton, "Forecasting uncertainty in electricity smart meter data by boosting additive quantile regression," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2448–2455, Mar. 2016.

[5] H. Aprillia, H.-T. Yang, and C.-M. Huang, "Statistical load forecasting using optimal quantile regression random forest and risk assessment index," *IEEE Transactions on Smart Grid*, vol. 12, no. 2, pp. 1467–1480, Oct. 2021.

[6] Q. He, L. Kong, Y. Wang, S. Wang, T. A. Chan, and E. Holland, "Regularized quantile regression under heterogeneous sparsity with application to quantitative genetic traits," *Computational Statistics & Data Analysis*, vol. 95, pp. 222–239, Mar. 2016.

[7] Z. Y. Algamal, R. Alhamzawi, and H. T. M. Ali, "Gene selection for microarray gene expression classification using bayesian lasso quantile regression," *Computers in biology and medicine*, vol. 97, pp. 145–152, June 2018.

[8] Y. Wu and Y. Liu, "Variable selection in quantile regression," *Statistica Sinica*, vol. 19, no. 2, pp. 801–817, Apr. 2009.

[9] L. Xue, S. Ma, and H. Zou, "Positive-definite l1-penalized estimation of large covariance matrices," *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1480–1491, Dec. 2012.

[10] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American statistical Association*, vol. 96, no. 456, pp. 1348–1360, Dec. 2001.

[11] C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *The Annals of statistics*, vol. 38, no. 2, pp. 894–942, Apr. 2010.

[12] A. Belloni and V. Chernozhukov, "l_1-penalized quantile regression in high-dimensional sparse models," *The Annals of Statistics*, vol. 39, no. 1, pp. 82–130, Feb. 2011.

[13] R. Koenker and P. Ng, "A frisch-newton algorithm for sparse quantile regression," *Acta Mathematicae Applicatae Sinica*, vol. 21, no. 2, pp. 225–236, May 2005.

[14] H. Wang and C. Li, "Distributed quantile regression over sensor networks," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 4, no. 2, pp. 338–348, June 2017.

[15] H. Zou and R. Li, "One-step sparse estimates in nonconcave penalized likelihood models," *Annals of Statistics*, vol. 36, no. 4, p. 1509, Aug 2008.

[16] B. Peng and L. Wang, "An iterative coordinate descent algorithm for high-dimensional nonconvex penalized quantile regression," *Journal of Computational and Graphical Statistics*, vol. 24, no. 3, pp. 676–694, July 2015.

[17] Y. Gu, J. Fan, L. Kong, S. Ma, and H. Zou, "Admm for high-dimensional sparse penalized quantile regression," *Technometrics*, vol. 60, no. 3, pp. 319–331, July 2018.

[18] Y. Gu and H. Zou, "Sparse composite quantile regression in ultrahigh dimensions with tuning parameter calibration," *IEEE Transactions on Information Theory*, vol. 66, no. 11, pp. 7132–7154, June 2020.

[19] Y. Wang, W. Yin, and J. Zeng, "Global convergence of admm in nonconvex nonsmooth optimization," *Journal of Scientific Computing*, vol. 78, no. 1, pp. 29–63, Jan. 2019.

[20] M. Hong, Z.-Q. Luo, and M. Razaviyayn, "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Apr. 2015, pp. 3836–3840.

[21] M. Yashtini, "Convergence analysis of a variable metric proximal linearized ADMM with over-relaxation parameter in nonconvex nonsmooth optimization," *arXiv preprint arXiv:2009.05361*, Sep. 2020.

[22] A. Themelis and P. Patrinos, "Douglas–Rachford splitting and ADMM for nonconvex optimization: Tight convergence results," *SIAM Journal on Optimization*, vol. 30, no. 1, pp. 149–181, Jan. 2020.

[23] R. Koenker and G. Bassett Jr, "Robust tests for heteroscedasticity based on regression quantiles," *Econometrica: Journal of the Econometric Society*, pp. 43–61, Jan. 1982.

[24] P. Bühlmann and S. Van De Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media, 2011.

[25] Y. Li and J. Zhu, "l_1-norm quantile regression," *Journal of Computational and Graphical Statistics*, vol. 17, no. 1, pp. 163–185, Mar. 2008.

[26] R. Varma, H. Lee, J. Kovačević, and Y. Chi, "Vector-valued graph trend filtering with non-convex penalties," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 6, pp. 48–62, Dec. 2019.

[27] S. Boyd, N. Parikh, and E. Chu, *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*. Now Publishers Inc, 2011.

[28] J. Huang, P. Breheny, and S. Ma, "A selective review of group selection in high-dimensional models," *Statistical Science*, vol. 27, no. 4, Oct. 2012.

[29] M. V. Lima, T. N. Ferreira, W. A. Martins, and P. S. Diniz, "Sparsity-aware data-selective adaptive filters," *IEEE Transactions on Signal Processing*, vol. 62, no. 17, pp. 4557–4572, Sep. 2014.