# Efficient and Robust Recovery of Sparse Signal and Image Using Generalized Nonconvex Regularization

Fei Wen, *Member, IEEE*, Ling Pei, *Member, IEEE*, Yuan Yang, Wenxian Yu, and Peilin Liu, *Member, IEEE*

***Abstract*—This paper addresses the robust reconstruction problem of a sparse signal from compressed measurements. We propose a robust formulation for sparse reconstruction that employs the $\ell_1$-norm as the loss function for the residual error and utilizes a generalized nonconvex penalty for sparsity inducing. The $\ell_1$-loss is less sensitive to outliers in the measurements than the popular $\ell_2$-loss, while the nonconvex penalty has the capability of ameliorating the bias problem of the popular convex LASSO penalty and thus can yield more accurate recovery. To solve this nonconvex and nonsmooth minimization formulation efficiently, we propose a first-order algorithm based on alternating direction method of multipliers. A smoothing strategy on the $\ell_1$-loss function has been used in deriving the new algorithm to make it convergent. Further, a sufficient condition for the convergence of the new algorithm has been provided for generalized nonconvex regularization. In comparison with several state-of-the-art algorithms, the new algorithm showed better performance in numerical experiments in recovering sparse signals and compressible images. The new algorithm scales well for large-scale problems, as often encountered in image processing.**

***Index Terms*—Alternating direction method, compressive sensing, impulsive noise, nonconvex regularization, robust sparse recovery.**

## I. Introduction

COMPRESSIVE sensing (CS) allows us to acquire sparse signals at a significantly lower rate than the classical Nyquist sampling [1], [2], which has attracted much attention in recent years and found wide applications in radar [3], communications [4], and speech processing [5]. Particularly, the CS theory is relevant in some applications in image processing, such as magnetic resonant imaging (MRI) [6], image super-resolution and denoising [7]–[9], and hyper-spectral imaging [10]. The CS theory states that, if a signal $\mathbf{x} \in \mathbb{R}^n$ is sparse, or can be sparsely represented on a basis, it can be recovered from a small number of linear measurements $\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{R}^m$ with $m < n$, where $\mathbf{A} \in \mathbb{R}^{m\times n}$ is the sensing matrix. With the consideration of measurement noise, the compressed measurements can be modeled as

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n} \tag{1}$$

where $\mathbf{n} \in \mathbb{R}^m$ denotes the measurement noise.

An intuitive method to reconstruct the sparse vector $\mathbf{x}$ consists in the following $\ell_0$-minimization problem

$$\min_{\mathbf{x}} \|\mathbf{x}\|_0 \ \text{ subject to } \ \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 \le \epsilon \tag{2}$$

where $\|\mathbf{x}\|_0$ is the $\ell_0$-norm, which counts the number of nonzero elements in the vector $\mathbf{x}$, $\epsilon > 0$ constrains the strength of the residual error. Generally, the nonconvex $\ell_0$-minimization problem (2) is difficult to solve, known to be NP-hard. To address this problem, convex relaxation methods have been proposed, such as basis-pursuit denoising (BPDN) [11]

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 \ \text{ subject to } \ \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 \le \epsilon \tag{3}$$

which relaxes the $\ell_0$-norm regularization into the $\ell_1$-norm regularization. The problem (3) can be equivalently converted into an unconstrained formulation (also called LASSO [12])

$$\min_{\mathbf{x}} \left\{ \frac{1}{\mu}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \|\mathbf{x}\|_1 \right\} \tag{4}$$

where $\mu > 0$ is a regularization parameter that balances the fidelity and sparsity of the solution.

The properties of $\ell_1$-regularization have been well studied in the context of CS. It has been demonstrated that the sparse signal $\mathbf{x}$ can be reliably recovered by $\ell_1$-regularized methods under some conditions of $\mathbf{A}$ [2]. However, as a relaxation of the $\ell_0$-regularization, the performance of the $\ell_1$-regularization is limited in two aspects. First, it would produce biased estimates for large coefficients. Second, it cannot recover a signal with the least measurements [13]. As a result, the estimate given by an $\ell_1$-regularized method is not sparse enough in some situations. A simple example of such a case can be found in [14].

To address this limitation, many improved methods employing $\ell_q$-regularization have been proposed, such as the $\ell_q$-regularized least-squares ($\ell_q$-LS) formulation

$$\min_{\mathbf{x}} \left\{ \frac{1}{\mu}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \|\mathbf{x}\|_q^q \right\} \tag{5}$$

with $0 \le q < 1$, where $\|\mathbf{x}\|_q^q$ is the nonconvex $\ell_q$ quasi-norm defined as $\|\mathbf{x}\|_q^q = \sum_i |x_i|^q$. Compared with $\ell_1$-regularization,

Manuscript received August 25, 2016; revised May 5, 2017 and August 7, 2017; accepted August 18, 2017. Date of publication August 25, 2017; date of current version November 6, 2017. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61401501 and Grant 61472442. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Laura Waller. *(Corresponding author: Fei Wen.)*

F. Wen, L. Pei, W. Yu, and P. Liu are with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: wenfei@sjtu.edu.cn; ling.pei@sjtu.edu.cn; wxyu@sjtu.edu.cn; liupeilin@sjtu.edu.cn).

Y. Yang is with the Air Control and Navigation Institution, Air Force Engineering University, Xian 710000, China (e-mail: yangyuankgd@126.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCI.2017.2744626

$\ell_q$-regularization has shown significantly better recovery performance in many applications [15]–[23]. Under certain conditions of the sensing matrix, $\ell_q$-regularized methods require fewer measurements to achieve reliable reconstruction than $\ell_1$-regularized methods [13]. Meanwhile, the sufficient conditions for reliable reconstruction for $\ell_q$-regularized methods are weaker than those for $\ell_1$-regularized methods [15], [23].

As in (2)–(5), many existing sparse recovery methods use the $\ell_2$-norm loss function. It is reasonable when the measurement noise is Gaussian distributed, since $\ell_2$-norm data fitting is optimal for Gaussian noise. However, the noise in practical applications often exhibits non-Gaussian properties. One important class of non-Gaussian noises arises in numerous practical situations is impulsive noise. Impulsive noise is well suited to model large outliers [24], which is frequently encountered in image processing [25]–[27]. Impulsive corruption may come from buffer overflow [28], missing data in the measurement process, bit errors in transmission [29], [30], and unreliable memory [31]. In these cases, the performance of $\ell_2$-loss based methods may severely degrade.

To achieve robust sparse recovery in the presence of impulsive noise, the Lorentzian-norm has been used as the loss function in [32], [33]. Meanwhile, in [34], the $\ell_1$-norm has been employed as the metric for the residual error to obtain the $\ell_1$-regularized least-absolute ($\ell_1$-LA) formulation

$$\min_{\mathbf{x}} \left\{ \frac{1}{\mu} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_1 + \|\mathbf{x}\|_1 \right\}. \tag{6}$$

Then, more efficient ADMM algorithms for this $\ell_1$-LA problem have been developed in [35]. Meanwhile, the Huber penalty function has be considered in [36]. In [37], ADMM and fast iterative shrinkage/thresholding algorithm (FISTA) have been used to efficiently solve the Huber penalty based formulation. Moreover, the $\ell_p$-norm loss with $0 \leq p < 2$ has been considered in [38], [39]. Notably, due to its simultaneous convexity and robustness, the $\ell_1$-loss has been wide used in designing robust methods, such as in sparse representation based face recognition [40] and channel estimation [41].

In this paper, we consider the following $P(\cdot)$-regularized least-absolute formulation for sparse recovery

$$\min_{\mathbf{x}} \left\{ \frac{1}{\mu} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_1 + P(\mathbf{x}) \right\}. \tag{7}$$

where $P(\cdot)$ is a generalized nonconvex penalty for sparsity promotion, such as the hard-thresholding, smoothly clipped absolute deviation (SCAD), or $\ell_q$-norm penalty. On the one hand, like the works [34], [35], [40], [41], we use the $\ell_1$-loss function as it is less sensitive to outliers compared with the quadratic function. On the other hand, unlike all of the existing robust methods [32]–[41] employing the $\ell_1$-regularization for sparsity inducing, we use a generalized nonconvex regularization in the new formulation. It is expected that, compared with the $\ell_1$-LA formulation (6), the new formulation retains the same robustness against outliers while can yield more accurate recovery via nonconvex regularization.

### A. Contributions

Generally, the problem (7) is difficult to solve, since in addition to the nonconvexity of the regularization term, both terms in the objective are nonsmooth. The main contributions of this work are as follows.

First, we propose an efficient first-order algorithm for the problem (7) based on ADMM. The standard ADMM algorithm can be directly used to solve (7), but it is not convergent for a nonconvex $P(\cdot)$ as the loss term is nonsmooth. To derive a convergent algorithm for generalized nonconvex $P(\cdot)$, a smoothing strategy of the $\ell_1$-loss has been adopted. The new algorithm scales well for high-dimensional problems, as often encountered in image processing.

Second, a convergence condition of the new algorithm has been derived for a generalized nonconvex regularization penalty. Finally, we have evaluated the new algorithm via reconstruction experiments on both simulated vector-signals and images. The results showed that the new algorithm is more robust than $\ell_2$-loss based methods while be more accurate than $\ell_1$-regularization based methods.

Matlab codes for the proposed algorithm and for reproducing the results in this work are available online at https://github.com/FWen/LqLA-Sparse-Recovery.git.

### B. Outline and Notations

The rest of this paper is organized as follows. Section II introduces the proximity operator for several generalized nonconvex penalty functions. In Section III, the new algorithm is presented. Section IV contains convergence analysis of the new algorithm. Section V contains experimental results. Finally, Section VI ends the paper with concluding remarks.

*Notations:* For a vector $\mathbf{v}$, $\mathrm{diag}(\mathbf{v})$ represents a diagonal matrix with diagonal elements be $\mathbf{v}$. $\mathcal{N}(0, \sigma^2)$ denotes a Gaussian distribution with zero-mean and variance $\sigma^2$. $E(\cdot)$, $\langle \cdot, \cdot \rangle$ and $(\cdot)^T$ stand for the expectation, inner product and transpose, respectively. $\nabla f(\cdot)$ and $\partial f(\cdot)$ stand for the gradient and subdifferential of the function $f$, respectively. $\mathrm{sign}(\cdot)$ denotes the sign of a quantity with $\mathrm{sign}(0) = 0$. $\lambda_{\max}(\cdot)$ denotes the maximal eigenvalue of a matrix. $I(\cdot)$ denotes the indicator function. $\mathbf{I}_n$ stands for an $n \times n$ identity matrix. $\|\cdot\|_q$ with $q \geq 0$ denotes the $\ell_q$-norm defined as $\|\mathbf{x}\|_q = (\sum_{i=1} |x_i|^q)^{1/q}$. $\mathrm{dist}(\mathbf{x}, S) := \inf\{\|\mathbf{y} - \mathbf{x}\|_2 : \mathbf{y} \in S\}$ denotes the distance from a point $\mathbf{x} \in \mathbb{R}^n$ to a subset $S \subset \mathbb{R}^n$. For a matrix $\mathbf{X}$, $\mathbf{X} \succeq \mathbf{0}$ means that it is positive-semidefinite.

## II. Proximity Operator for Sparsity Inducing Penalties

This section revisits sparsity inducing penalties and the corresponding proximity operators, including the hard-thresholding, soft-thresholding, $\ell_q$-thresholding, SCAD, and minimax concave (MC) penalties, which are the most commonly used penalties for sparsity inducing. For many nonsmooth problems employing such regularization penalties, proximity operator plays a central role in developing highly-efficient first-order algorithms which scale well to high-dimensional problems.

For a proper and lower semicontinuous function $P(\cdot)$, its proximity operator with penalty $\eta$ ($\eta > 0$) is defined as

$$\text{prox}_{P,\eta}(t) = \arg\min_x \left\{ P(x) + \frac{\eta}{2}(x-t)^2 \right\}. \tag{8}$$

1) Hard-thresholding. The penalty is given by [42]

$$P(x) = 2 - (|x| - \sqrt{2})^2 I(|x| < \sqrt{2})$$

and the corresponding thresholding function is

$$\text{prox}_{P,\eta}(t) = tI(|t| > \sqrt{2/\eta}). \tag{9}$$

Note that, the $\ell_0$-norm penalty $P(x) = |x|_0$ also results in (9).

2) Soft-thresholding, $P(x) = |x|$. The corresponding thresholding function is

$$\text{prox}_{P,\eta}(t) = S_{1/\eta}(t) = \text{sign}(t)\max\{|t| - 1/\eta, 0\} \tag{10}$$

where $S_\alpha$ is well-known as the soft-thresholding/shrinkage operator.

3) $\ell_q$-norm ($0 < q < 1$), $P(x) = |x|^q$. In this case, the proximity operator (8) does not have a closed-form solution except for the two special cases of $q = \frac{1}{2}$ and $q = \frac{2}{3}$ [43], and it can be solved as [44]

$$\text{prox}_{P,\eta}(t) = \begin{cases} 0, & |t| < \tau \\ \{0, \text{sign}(t)\beta\}, & |t| = \tau \\ \text{sign}(t)y^*, & |t| > \tau \end{cases} \tag{11}$$

where $\beta = [2(1-q)/\eta]^{\frac{1}{2-q}}$, $\tau = \beta + q\beta^{q-1}/\eta$, $y^*$ is the solution of $h(y) = qy^{q-1} + \eta y - \eta|t| = 0$ over the region $(\beta, |t|)$. The function $h(y)$ is convex, thus, when $|t| > \tau$, $y^*$ can be iteratively computed by a Newton's method.

4) SCAD. The penalty is given by

$$P(x;\lambda) = \begin{cases} \lambda|x|, & |x| < \lambda \\ \dfrac{2a\lambda|x| - x^2 - \lambda^2}{2(a-1)}, & \lambda \le |x| < a\lambda \\ (a+1)\lambda^2/2, & |x| \ge a\lambda \end{cases}$$

for some $a > 2$, where $\lambda > 0$ is a threshold parameter. The corresponding thresholding function is [45]

$$\text{prox}_{P,\eta}(t) = \begin{cases} \text{sign}(t)\max\{|t| - \lambda, 0\}, & |t| \le 2\lambda \\ \dfrac{(a-1)t - \text{sign}(t)a\lambda}{a-2}, & 2\lambda < |t| \le a\lambda \\ t, & |t| > a\lambda \end{cases}. \tag{12}$$

5) MC penalty. Similar to the hard, $\ell_q$, and SCAD penalties, MC can also ameliorate the bias problem of LASSO [46], and it has been widely used for penalized variable selection in high-dimensional linear regression. MC has a parametric formulation as

$$P(x;\lambda) = \lambda \int_0^{|x|} \max(1 - t/(\gamma\lambda), 0)dt$$

with $\gamma > 1$. The corresponding thresholding function is

$$\text{prox}_{P,\eta}(t) = \begin{cases} 0, & |t| \le \lambda/\eta \\ \dfrac{\text{sign}(t)(|t| - \lambda/\eta)}{1 - 1/\gamma}, & \lambda/\eta < |t| \le \gamma\lambda/\eta \\ t, & |t| > \gamma\lambda/\eta \end{cases}.$$

For each $\lambda > 0$, we can obtain a continuum of penalties and threshold operators by varying $\gamma$ in the range $(0, +\infty)$.

## III. Proposed Algorithm

ADMM is a simple but powerful framework, which is well suited to distributed optimization and meanwhile is flexible to solve many high-dimensional optimization problems [47]. ADMM uses a variable-splitting scheme, which separates coupled components in the cost function via introducing auxiliary constraint variables. This procedure naturally decouples the variables and transforms the original problem into an equivalent problem that can be effectively solved in an alternating minimization manner.

### A. Standard ADMM Algorithm Without Smoothing

Using an auxiliary variable $\mathbf{v} \in \mathbb{R}^m$, the formulation (7) can be rewritten as

$$\min_{\mathbf{x},\mathbf{v}} \left\{ \frac{1}{\mu}\|\mathbf{v}\|_1 + P(\mathbf{x}) \right\} \text{ subject to } \mathbf{A}\mathbf{x} - \mathbf{y} = \mathbf{v}. \tag{13}$$

The augmented Lagrangian of the problem is

$$\mathcal{L}(\mathbf{v},\mathbf{x},\mathbf{w}) = \frac{1}{\mu}\|\mathbf{v}\|_1 + P(\mathbf{x}) - \langle \mathbf{w}, \mathbf{A}\mathbf{x} - \mathbf{y} - \mathbf{v}\rangle + \frac{\rho}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y} - \mathbf{v}\|_2^2$$

where $\mathbf{w} \in \mathbb{R}^m$ is the Lagrangian multiplier, $\rho > 0$ is a penalty parameter. Then, ADMM consists of the following three steps

$$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}} \left( P(\mathbf{x}) + \frac{\rho}{2}\left\| \mathbf{A}\mathbf{x} - \mathbf{y} - \mathbf{v}^k - \frac{\mathbf{w}^k}{\rho} \right\|_2^2 \right) \tag{14}$$

$$\mathbf{v}^{k+1} = \arg\min_{\mathbf{v}} \left( \frac{1}{\mu}\|\mathbf{v}\|_1 + \frac{\rho}{2}\left\| \mathbf{A}\mathbf{x}^{k+1} - \mathbf{y} - \mathbf{v} - \frac{\mathbf{w}^k}{\rho} \right\|_2^2 \right) \tag{15}$$

$$\mathbf{w}^{k+1} = \mathbf{w}^k - \rho\left(\mathbf{A}\mathbf{x}^{k+1} - \mathbf{y} - \mathbf{v}^{k+1}\right). \tag{16}$$

The $\mathbf{x}$-update step (14) in fact solves a penalized LS problem. We use a standard trick for speeding up ADMM that solve this subproblem approximately. Specifically, let $\mathbf{u}^k = \mathbf{y} + \mathbf{v}^k + \mathbf{w}^k/\rho$, we linearize the quadratic term in the objective

function of (14) at a point $\mathbf{x}^k$ as

$$\begin{aligned}&\frac{1}{2}\left\|\mathbf{A}\mathbf{x}-\mathbf{u}^k\right\|_2^2\\&\approx\frac{1}{2}\left\|\mathbf{A}\mathbf{x}^k-\mathbf{u}^k\right\|_2^2+\left\langle\mathbf{x}-\mathbf{x}^k,d_1(\mathbf{x}^k)\right\rangle+\frac{1}{2\tau_1}\left\|\mathbf{x}-\mathbf{x}^k\right\|_2^2\\&=\frac{1}{2}\left\|\mathbf{A}\mathbf{x}^k-\mathbf{u}^k\right\|_2^2+\frac{1}{2\tau_1}\left\|\mathbf{x}-\mathbf{x}^k+\tau_1 d_1(\mathbf{x}^k)\right\|_2^2\\&\qquad-\frac{\tau_1}{2}\left\|d_1(\mathbf{x}^k)\right\|_2^2\end{aligned}$$

where $d_1(\mathbf{x}^k)=\mathbf{A}^T(\mathbf{A}\mathbf{x}^k-\mathbf{u}^k)$ is the gradient of the quadratic term at $\mathbf{x}^k$, $\tau_1>0$ is a proximal parameter. Based on this approximation, the $\mathbf{x}$-update step becomes easy to solve since it can be computed element-wise as the proximity operator (8)

$$\mathbf{x}^{k+1}=\mathrm{prox}_{P,\rho}(\mathbf{b}^k)\tag{17}$$

with $\mathbf{b}^k=\mathbf{x}^k-\tau_1\mathbf{A}^T(\mathbf{A}\mathbf{x}^k-\mathbf{u}^k)$. As will be shown in Lemma 1 in Section IV, for a generalized nonconvex penalty if $1/\tau_1$ is selected to be a Lipschitz constant of $d_1(\mathbf{x})$, i.e., $1/\tau_1>\lambda_{\max}(\mathbf{A}^T\mathbf{A})$, the augmented Lagrangian function is guaranteed nonincreasing when the $\mathbf{x}$-update step is approximately solved by (17).

The $\mathbf{v}$-upadte step (15) has an closed-form solution as

$$\mathbf{v}^{k+1}=S_{1/(\mu\rho)}\left(\mathbf{A}\mathbf{x}^{k+1}-\mathbf{y}-\frac{\mathbf{w}^k}{\rho}\right).\tag{18}$$

When $P(\cdot)$ is the $\ell_1$-norm penalty, the ADMM algorithm using the update steps (17), (18) and (16) reduces to *Your ALgorithm for L1* (YALL1) [34], and it is guaranteed to converge to the global minimizer of the problem (13) if $\tau_1<1/\lambda_{\max}(\mathbf{A}^T\mathbf{A})$ [34]. However, for a nonconvex penalty, e.g., MC, SCAD or $\ell_q$-norm with $q<1$, the convergence of this ADMM algorithm is not guaranteed. Empirical studies show that it always fails to converge in this case.

### *B. Proposed ADMM Algorithm With Smoothed $\ell_1$-Loss*

To develop a convergent algorithm for the nonconvex case with a nonconvex penalty $P(\cdot)$, we consider a smoothed $\ell_1$-loss function and propose a smoothed formulation of the problem (7) as

$$\min_{\mathbf{x}}\left\{\frac{1}{\mu}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|_{1,\varepsilon}+P(\mathbf{x})\right\}\tag{19}$$

where the smoothed $\ell_1$-norm is defined as

$$\|\mathbf{v}\|_{1,\varepsilon}=\sum\nolimits_i(v_i^2+\varepsilon^2)^{\frac{1}{2}}$$

with $\varepsilon>0$ be an approximation parameter. Since $\lim_{\varepsilon\to0}\|\mathbf{v}\|_{1,\varepsilon}=\|\mathbf{v}\|_1$, $\|\mathbf{v}\|_{1,\varepsilon}$ accurately approximates $\|\mathbf{v}\|_1$ when $\varepsilon$ is sufficiently small. Similar to (13), the problem (19) can be equivalently expressed as

$$\min_{\mathbf{x},\mathbf{v}}\left\{\frac{1}{\mu}\|\mathbf{v}\|_{1,\varepsilon}+P(\mathbf{x})\right\}\ \text{subject to}\ \mathbf{A}\mathbf{x}-\mathbf{y}=\mathbf{v}.\tag{20}$$

The augmented Lagrangian of the problem is

$$\begin{aligned}\mathcal{L}_\varepsilon(\mathbf{v},\mathbf{x},\mathbf{w})=&\frac{1}{\mu}\|\mathbf{v}\|_{1,\varepsilon}+P(\mathbf{x})-\langle\mathbf{w},\mathbf{A}\mathbf{x}-\mathbf{y}-\mathbf{v}\rangle\\&+\frac{\rho}{2}\|\mathbf{A}\mathbf{x}-\mathbf{y}-\mathbf{v}\|_2^2.\end{aligned}\tag{21}$$

Using the smoothed $\ell_1$-loss, the $\mathbf{v}$-update step becomes

$$\mathbf{v}^{k+1}=\arg\min_{\mathbf{v}}\left(\frac{1}{\mu}\|\mathbf{v}\|_{1,\varepsilon}+\frac{\rho}{2}\left\|\mathbf{A}\mathbf{x}^{k+1}-\mathbf{y}-\mathbf{v}-\frac{\mathbf{w}^k}{\rho}\right\|_2^2\right).\tag{22}$$

As the objective function in (22) is smooth, the subproblem (22) can be solved by a standard iterative method, such as the gradient descent method, conjugate gradient method, or quasi-Newton method. However, using such an iterative method, the overall algorithm has double loops and therefore is inefficient. To improve the overall efficiency of the algorithm, we adopt the standard strategy for accelerating ADMM again, which bypasses the inner loop in this step via solving (22) approximately. Specifically, we approximate the term $\|\mathbf{v}\|_{1,\varepsilon}$ in the objective function of (22) by

$$\|\mathbf{v}\|_{1,\varepsilon}\approx\left\|\mathbf{v}^k\right\|_{1,\varepsilon}+\left\langle\mathbf{v}-\mathbf{v}^k,d_2(\mathbf{v}^k)\right\rangle+\frac{1}{2\tau_2}\left\|\mathbf{v}-\mathbf{v}^k\right\|_2^2$$

where $d_2(\mathbf{v}^k)=\nabla\left\|\mathbf{v}^k\right\|_{1,\varepsilon}$ with $d_2(\mathbf{v}^k)_i=v_i(v_i^2+\varepsilon^2)^{-1/2}$, $\tau_2>0$ is an approximation parameter. Using this linearization, the solution of the problem is explicitly given by

$$\begin{aligned}\mathbf{v}^{k+1}=&\frac{\tau_2}{\rho\mu\tau_2+1}\left[\frac{1}{\tau_2}\mathbf{v}^k-d_2(\mathbf{v}^k)\right.\\&\left.+\rho\mu\left(\mathbf{A}\mathbf{x}^{k+1}-\mathbf{y}-\frac{\mathbf{w}^k}{\rho}\right)\right].\end{aligned}\tag{23}$$

The main consideration of using such a smoothing strategy is that, the gradient of $\|\mathbf{v}\|_{1,\varepsilon}$ is Lipschitz continuous when $\varepsilon>0$, e.g., $\nabla^2\|\mathbf{v}\|_{1,\varepsilon}\preceq\frac{1}{\varepsilon}\mathbf{I}_n$. This smoothness property is crucial for the convergence of the new algorithm in the case of a nonconvex $P(\cdot)$. As will be shown in the convergence analysis in Appendix C, using the smoothed $\ell_1$-loss, the changes in the dual iterates can be bounded by the changes in the primal iterates. This is the key point to show the descent property of the augmented Lagrangian function which leads to establishment of convergence. Moreover, for the proposed algorithm, the dominant computational load in each iteration is matrix-vector multiplication with complexity $O(mn)$. Thus, it scales well for high-dimension problems.

## IV. Convergence Analysis

This section analyzes the convergence property of the new algorithm for a generalized nonconvex penalty. While the convergence issue of ADMM has been well addressed for the convex case [47], [53], there have been only a few works reported very recently on the convergence issue for the nonconvex case [48]–[50]. The convergence theory in [50] cannot be applied to the proposed algorithm since it is restricted to the case of that, the regularization term is a smooth function or a convex nonsmooth function. Meanwhile, since the proposed formulation

involves a smoothing parameter, the convergence condition of the proposed ADMM algorithm cannot be directly derived from the results in [48] and [49]. The following results are derived following similarly the line in [48], [49]. We first give some definitions and lemmas in the proof of the main result. All the proofs are given in Appendix.

*Definition 1* (*lower semicontinuous function):* An extended real-valued function $f$ is lower semicontinuous at a point $x_0$ if $f(x_0) \leq \liminf_{x \to x_0} f(x)$. If a function is lower semicontinuous at every point of its domain of definition, then it is a lower semicontinuous function.

*Definition 2* (*Kurdyka-Lojasiewicz (KL) function):* A proper function $f : \mathbb{R}^n \to \mathbb{R}$ is said to have the KL property at $x_0 \in \mathrm{dom}\partial f$ if there exists $\eta > 0$, a neighborhood $V$ of $x_0$ and a continuous concave function $\varphi : [0, \eta) \to \mathbb{R}_+$ such that: (i) $\varphi(0) = 0$ and $\varphi$ is continuously differentiable on $(0, \eta)$ with positive derivatives; (ii) for all $x \in V$ satisfying $f(x_0) < f(x) < f(x_0) + \eta$, it holds that $\varphi'(f(x) - f(x_0))\mathrm{dist}(0, \partial f(x)) \geq 1$. A proper closed function $f$ satisfying the KL property at all points in $\mathrm{dom}\partial f$ is called a KL function.

The KL property allows, through a "uniformization" result inspired by [55] to considerably simplify the main arguments of the convergence analysis and avoid involved induction reasoning.

*Lemma 1:* Suppose that $P(\cdot)$ is a closed, proper, lower semicontinuous function, for any $\mathbf{x}^k \in \mathbb{R}^n$, the minimizer $\mathbf{x}^{k+1}$ given by (17) satisfies

$$\mathcal{L}_\varepsilon(\mathbf{v}^k, \mathbf{x}^{k+1}, \mathbf{w}^k) \leq \mathcal{L}_\varepsilon(\mathbf{v}^k, \mathbf{x}^k, \mathbf{w}^k) - c_0 \left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|_2^2$$

where

$$c_0 = \frac{\rho}{2}\left(\frac{1}{\tau_1} - \lambda_{\max}(\mathbf{A}^T\mathbf{A})\right).$$

*Lemma 2:* For any $\mathbf{v}^k \in \mathbb{R}^m$, the minimizer $\mathbf{v}^{k+1}$ given by (23) satisfies

$$\mathcal{L}_\varepsilon(\mathbf{v}^{k+1}, \mathbf{x}^{k+1}, \mathbf{w}^k) \leq \mathcal{L}_\varepsilon(\mathbf{v}^k, \mathbf{x}^{k+1}, \mathbf{w}^k) - c_1 \left\|\mathbf{v}^{k+1} - \mathbf{v}^k\right\|_2^2$$

where

$$c_1 = \frac{1}{\mu\tau_2} + \frac{\rho}{2} - \frac{1}{2\mu\varepsilon}.$$

Lemmas 1 and 2 establish the descent properties for the $\mathbf{x}$- and $\mathbf{v}$-subproblems, respectively.

*Lemma 3:* Suppose that $P(\cdot)$ is a closed, proper, lower semicontinuous function, let $\tilde{\mathcal{L}}(\mathbf{v}, \mathbf{x}, \mathbf{w}, \tilde{\mathbf{v}}) := \mathcal{L}_\varepsilon(\mathbf{v}, \mathbf{x}, \mathbf{w}) + c_2\|\mathbf{v} - \tilde{\mathbf{v}}\|_2^2$, for $(\mathbf{v}^k, \mathbf{x}^k, \mathbf{w}^k)$ generated via (17), (23) and (16), if $\varepsilon > 0$ and (24) holds, then

$$\tilde{\mathcal{L}}(\mathbf{v}^k, \mathbf{x}^k, \mathbf{w}^k, \mathbf{x}^{k-1}) \geq \tilde{\mathcal{L}}(\mathbf{v}^{k+1}, \mathbf{x}^{k+1}, \mathbf{w}^{k+1}, \mathbf{x}^k) + c_0 \left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|_2^2 + c_3\left\|\mathbf{v}^{k+1} - \mathbf{v}^k\right\|_2^2$$

where $c_2, c_3 > 0$ are given by

$$c_2 = \frac{2}{\rho\mu^2}\left(\frac{1}{\varepsilon} + \frac{1}{\tau_2}\right)^2$$

$$c_3 = \frac{1}{2}\rho - \frac{2}{\rho\mu^2}\left[\frac{2}{{\tau_2}^2} + \frac{2}{\tau_2\varepsilon} + \frac{1}{\varepsilon^2}\right] + \frac{2\varepsilon - \tau_2}{2\mu\tau_2\varepsilon}.$$

Lemma 3 establishes the sufficient decrease property for the auxiliary function $\tilde{\mathcal{L}}$, which indicates that $\tilde{\mathcal{L}}$ is nonincreasing and thus is convergent as it is lower semicontinuous.

*Lemma 4:* Suppose that $P(\cdot)$ is a closed, proper, lower semicontinuous function, let $\mathbf{z}^k := (\mathbf{v}^k, \mathbf{x}^k, \mathbf{w}^k)$ with $(\mathbf{v}^k, \mathbf{x}^k, \mathbf{w}^k)$ generated via (17), (23) and (16), suppose that $\varepsilon > 0$, $\tau_1 < 1/\lambda_{\max}(\mathbf{A}^T\mathbf{A})$, and (24) holds, then

$$\lim_{k\to\infty} \left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\|_2^2 = 0.$$

In particular, any cluster point of $\{\mathbf{z}^k\}$ is a stationary point of $\mathcal{L}_\varepsilon$.

*Lemma 5:* Suppose that $P(\cdot)$ is a closed, proper, lower semicontinuous function, let $\tilde{\mathcal{L}}(\mathbf{v}, \mathbf{x}, \mathbf{w}, \tilde{\mathbf{v}}) := \mathcal{L}_\varepsilon(\mathbf{v}, \mathbf{x}, \mathbf{w}) + c_2\|\mathbf{v} - \tilde{\mathbf{v}}\|^2$ with $c_2$ defined in Lemma 3, suppose that $\varepsilon > 0$, $\tau_1 < 1/\lambda_{\max}(\mathbf{A}^T\mathbf{A})$ and (24) holds, for $(\mathbf{v}^k, \mathbf{x}^k, \mathbf{w}^k)$ generated via (17), (23) and (16), there exists a constant $c_4 > 0$ such that

$$\mathrm{dist}(0, \partial\tilde{\mathcal{L}}(\mathbf{v}^{k+1}, \mathbf{x}^{k+1}, \mathbf{w}^{k+1}, \mathbf{v}^k)) \leq c_4\left(\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|_2 + \left\|\mathbf{v}^{k+1} - \mathbf{v}^k\right\|_2 + \left\|\mathbf{v}^k - \mathbf{v}^{k-1}\right\|_2\right).$$

Lemma 5 establishes a subgradient lower bound for the iterate gap, which together with Lemma 4 implies that $\mathrm{dist}(0, \partial\tilde{\mathcal{L}}(\mathbf{v}^{k+1}, \mathbf{x}^{k+1}, \mathbf{w}^{k+1}, \mathbf{v}^k)) \to 0$ as $k \to \infty$.

*Theorem 1:* Suppose that $P(\cdot)$ is a closed, proper, lower semicontinuous, Kurdyka-Lojasiewicz function, $\varepsilon > 0$ and $\tau_1 < 1/\lambda_{\max}(\mathbf{A}^T\mathbf{A})$, then, if

$$\rho > \frac{\sqrt{36\varepsilon^2 + 28\tau_2\varepsilon + 17{\tau_2}^2} + \tau_2 - 2\varepsilon}{2\mu\tau_2\varepsilon} \tag{24}$$

the sequence $\{(\mathbf{v}^k, \mathbf{x}^k, \mathbf{w}^k)\}$ generated by the ADMM algorithm via the three steps (17), (23) and (16) converges to a stationary point of the problem (20).

In the conditions in Theorem 1, there is no restriction on the proximal parameter $\tau_2$. That is due to the fact that if (24) is satisfied, the sufficient decrease property of the $\mathbf{v}$-subproblem is guaranteed since $c_1$ in Lemma 2 is positive in this case. However, the value of $\tau_2$ would affect the convergence speed of the algorithm. Intensive numerical studies show that selecting a value of the same order as $\varepsilon$ for $\tau_2$ can yield satisfactory convergence rate.

When $\varepsilon \to 0$, the problem (20) reduces to the original problem (13) and thus the solution of (20) accurately approximates that of (13). However, from the convergence condition in Theorem 1, the penalty parameter should be chosen to be $\rho \to \infty$ in this case. Generally, with a very large value of $\rho$, the ADMM algorithm would be very slow and impractical. In practical applications, selecting a moderate value of $\varepsilon$ suffices to achieve satisfactory performance. Moreover, a standard trick to speed up the algorithm is to adopt a continuation process for the penalty parameter. Specifically, we can use a properly small starting value of the penalty parameter and gradually increase it by iteration until reaching the target value, e.g., $0 < \rho_0 \leq \rho_1 \leq \cdots \leq \rho_K = \rho_{K+1} = \cdots = \rho$. In this case, Theorem 1 still applies as the value of the penalty parameter turns into fixed at $\rho$ within finite iterations. Furthermore, with an initialization which is usually used for nonconvex algorithms,

the new algorithm often converges quickly even in the case of a large $\rho$.

When $P(\cdot)$ is nonconvex, the formulation (19) is nonconvex and the proposed algorithm may converge to one of its many local minimizers. In this case, a good initialization is crucial for the new algorithm to achieve satisfactory performance. Since a standard CS method (e.g., BPDN or LASSO) may break down in highly impulsive noise, it is more appropriate to employ a robust method for initialization such as $\ell_1$-LA (6). The $\ell_1$-LA problem can be solved via the ADMM update steps (17), (18) and (16), which is guaranteed to converge to the global minimizer if $\tau_1 < 1/\lambda_{\max}(\mathbf{A}^T\mathbf{A})$ [45].

## V. Numerical Experiments

We evaluate the new method in comparison with L1LS-FISTA [52], LqLS-ADMM [48], and YALL1 [34]. L1LS-FISTA solves the $\ell_1$-LS formulation (4). For this standard CS formulation, there exist a number of solvers can find the global minimizer of (4) and achieve the same accuracy, but with different computational complexity. Among these solvers, ADMM and FISTA are two of the most computational efficient. LqLS-ADMM solves the $\ell_q$-LS formulation (5) based on ADMM. LqLS-ADMM is run with $q = 0.5$ and it is guaranteed to converge when the penalty parameter is properly chosen [48]. YALL1 solves the robust $\ell_1$-LA formulation (6) using an ADMM scheme. We conduct mainly two reconstruction experiments on simulated vector-signals and images, respectively.

For the proposed method, we use the $\ell_q$-norm penalty as it has a flexible parametric form that adapts to different thresholding functions while includes the hard- and soft-thresholding as special cases, which is termed as LqLA-ADMM in the following. It is run with $\tau_1 = 0.99/\lambda_{\max}(\mathbf{A}^T\mathbf{A})$, $\varepsilon = 10^{-3}$, $\tau_2 = \varepsilon$ and $\rho = \frac{3.2}{\mu\varepsilon}$, and the $\mathbf{v}$-subproblem is updated via (23). Different values of $q$, $q \in \{0.2, 0.5, 0.7\}$, are examined for LqLA-ADMM. We use a stopping tolerance parameter of $10^{-7}$ for LqLA-ADMM. Moreover, a continuation process is used for the penalty parameter as $\rho_k = 1.02\rho_{k-1}$ if $\rho_k < \rho$ and $\rho_k = \rho$ otherwise. The two noncnvex algorithms, LqLS-ADMM and LqLA-ADMM, are initialized by the solution of YALL1. Note that LqLA-ADMM with $q = 1, \varepsilon = 0$ and updated via the steps (17), (18) and (16) reduces to YALL1.

We consider two types of impulsive noise. *1) Gaussian mixture noise:* we consider a typical two-term Gaussian mixture model with probability density function (pdf) given by

$$(1-\xi)\mathcal{N}(0,\sigma^2) + \xi\mathcal{N}(0,\kappa\sigma^2)$$

where $0 \le \xi < 1$ and $\kappa > 1$. This model is an approximation to Middleton's Class A noise model, where the two parameters $\xi$ and $\kappa > 1$ respectively control the ratio and the strength of outliers in the noise. In this model, the first term stands for the nominal background noise, e.g., Gaussian thermal noise, while the second term describes the impulsive behavior of the noise. *2) Symmetric $\alpha$-stable ($\mathrm{S}\alpha\mathrm{S}$) noise:* except for a few known cases, the $\mathrm{S}\alpha\mathrm{S}$ distributions do not have analytical formulations. The characteristic function of a zero-location $\mathrm{S}\alpha\mathrm{S}$ distribution can be expressed as

$$\varphi(\omega) = \exp\left(ja\omega - \gamma^\alpha|\omega|^\alpha\right)$$

where $0 < \alpha \le 2$ is the characteristic exponent and $\gamma > 0$ is the scale parameter or dispersion. The characteristic exponent measures the thickness of the tail of the distribution. The smaller the value of $\alpha$, the heavier the tail of the distribution and the more impulsive the noise is. When $\alpha = 2$, the $\mathrm{S}\alpha\mathrm{S}$ distribution becomes the Gaussian distribution with variance $2\gamma^2$. When $\alpha = 1$, the $\mathrm{S}\alpha\mathrm{S}$ distribution reduces to the Cauchy distribution.

For Gaussian and Gaussian mixture noise, we use the signal-to-noise ratio (SNR) to quantify the strength of noise, which is defined by

$$\mathrm{SNR} = 20\log_{10}\left(\frac{\|\mathbf{A}\mathbf{x}^o - E\{\mathbf{A}\mathbf{x}^o\}\|_2}{\|\mathbf{n}\|_2}\right)$$

where $\mathbf{x}^o$ denotes the true signal. Since an $\mathrm{S}\alpha\mathrm{S}$ distribution with $\alpha < 2$ has infinite variance, the strength of $\mathrm{S}\alpha\mathrm{S}$ noise is quantified by the dispersion $\gamma$.

All the compared methods require the selection of the regularization parameter $\mu$, which balances the fidelity and sparsity of the solution and is closely related to the performance of these methods. A popular approach is to compute the recovery along the regularization path (a set of $\mu$), and select the optimal value based on the statistical information of the noise. More specifically, for the $\ell_1$-loss based formulations, the optimal $\mu$ can be selected as the maximum value of $\mu$ such that the bound constraint on the residual is met, e.g., $\|\mathbf{A}\hat{\mathbf{x}} - \mathbf{y}\|_1 \le \delta$, where $\delta$ is the estimated first-order moment of the noise. The approach also applies to the new method for sufficiently small $\varepsilon$. However, this approach cannot be used in the case of $\mathrm{S}\alpha\mathrm{S}$ impulsive noise with $\alpha \le 1$, since the first-order moment of such noise is infinite. Another effective approach is to learn a value of $\mu$ via cross-validation [51]. In our experiments, to compare the methods fairly, the regularization parameter in each method is chosen by providing the best performance in terms of relative error of recovery.

### A. Recovery of Simulated Sparse Signals

In the first experiment, we evaluate the compared methods using simulated sparse signals in various noise conditions. We use a simulated $K$-sparse signal of length $n = 512$, in which the positions of the $K$ nonzeros are uniformly randomly chosen while the amplitude of each nonzero entry is generated according to the Gaussian distribution $\mathcal{N}(0,1)$. The signal is normalized to have a unit energy value. The $m \times n$ sensing matrix $\mathbf{A}$ is chosen to be an orthonormal Gaussian random matrix with $m = 200$. A recovery $\hat{\mathbf{x}}$ is regarded as successful if the relative error satisfies

$$\frac{\|\hat{\mathbf{x}} - \mathbf{x}^o\|_2}{\|\mathbf{x}^o\|_2} \le 10^{-2}.$$

Each provided result is an average over 200 independent Monte Carlo runs. Three noise conditions are considered, Gaussian noise with SNR = 30 dB, Gaussian mixture noise with $\xi = 0.1$, $\kappa = 1000$ and SNR = 30 dB, and $\mathrm{S}\alpha\mathrm{S}$ noise with $\alpha = 1$ (Cauchy noise) and $\gamma = 10^{-4}$.

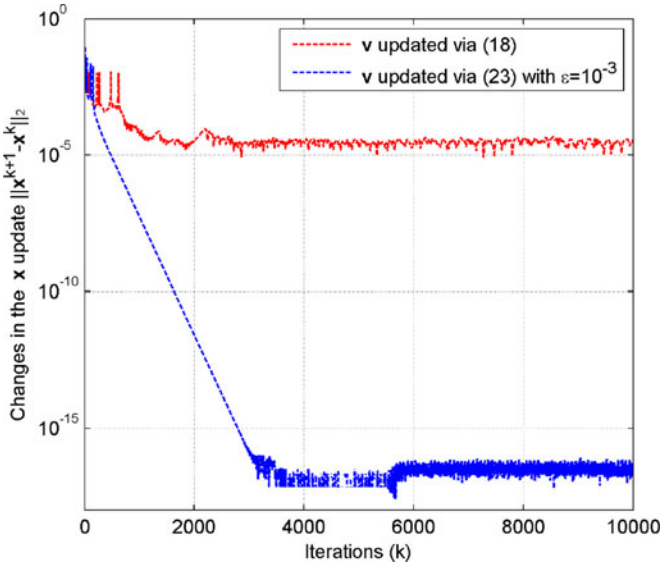


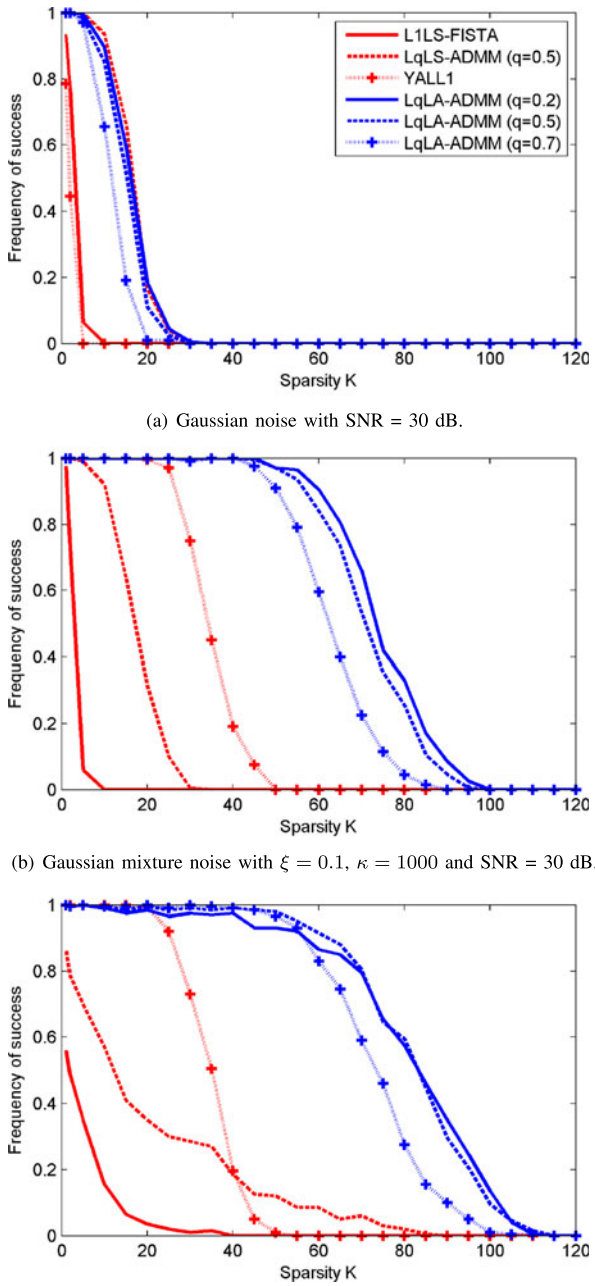

Fig. 1. Typical convergence behavior of LqLA-ADMM with $q = 0.5$ (Gaussian mixture noise with $\xi = 0.1$, $\kappa = 1000$ and SNR $= 30$ dB).

Fig. 1 shows the typical convergence behavior of LqLA-ADMM with $q = 0.5$ in two conditions with the $\mathbf{v}$-subproblem be solved by (18) and (23), respectively. The sparsity of the signal is $K = 30$. It can be seen that LqLA-ADMM does not converge when the $\mathbf{v}$-subproblem is updated via (18).

Fig. 2 presents the successful rate of recovery of the compared algorithms versus sparsity $K$ in the three noise conditions. It is clear that in Gaussian noise, L1LS-FISTA and LqLS-ADMM respectively slightly outperform YALL1 and LqLA-ADMM. That is in Gaussian noise, the $\ell_1$-loss does not lead to considerable performance degradation relative to the $\ell_2$-one which is optimal in a maximum likelihood sense in this case. Moreover, LqLS-ADMM and LqLA-ADMM significantly outperform L1LS-FISTA and YALL1, which demonstrates the superiority of the $\ell_q$-regularization over the $\ell_1$-regularization.

In the two impulsive noise conditions, the $\ell_1$-loss based YALL1 and LqLA-ADMM algorithms outperform the $\ell_2$-loss based L1LS-FISTA and LqLS-ADMM algorithms in most cases. That demonstrates the robustness of $\ell_1$-loss against impulsive corruptions in the measurements. Meanwhile, in impulsive noise, the advantage of $\ell_q$-regularization over $\ell_1$-regularization remains considerable. For example, LqLA-ADMM significantly outperforms YALL1 while LqLS-ADMM significantly outperforms L1LS-FISTA. In the S$\alpha$S noise condition, LqLA-ADMM can achieve a rate of successful recovery greater than $80\%$ when $K \leq 70$, while YALL1 achieves such a rate only when $K \leq 30$. Among the three tested values of $q$ ($q \in \{0.2, 0.5, 0.7\}$) for LqLA-ADMM, $q = 0.2$ and $q = 0.5$ yield better performance than $q = 0.7$.

Moreover, it can be seen from Fig. 2(a) and (b) that, with fixed SNR of 30 dB, the $\ell_2$-loss yields comparable performance in the two noise conditions (Gaussian and Gaussian mixture). This is due to the fact that, the recovery error of the $\ell_2$-loss based formulation is bounded by the noise variance [54], which does not




(a) Gaussian noise with SNR = 30 dB.

(b) Gaussian mixture noise with $\xi = 0.1$, $\kappa = 1000$ and SNR = 30 dB.

(c) $S\alpha S$ noise with $\alpha = 1$ and $\gamma = 10^{-4}$.

Fig. 2. Recovery performance versus sparsity for the compared methods in different noise conditions (a) Gaussian noise, (b) Gaussian mixture noise, (c) $S\alpha S$ noise. .

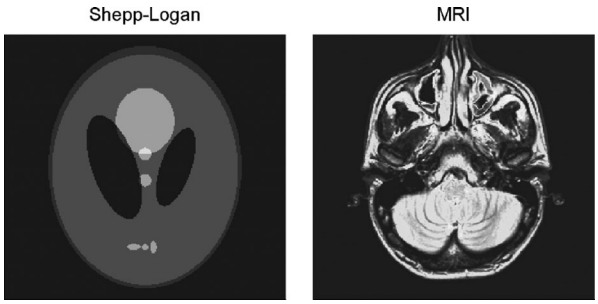



Fig. 3. The two $256 \times 256$ images used for performance evaluation.

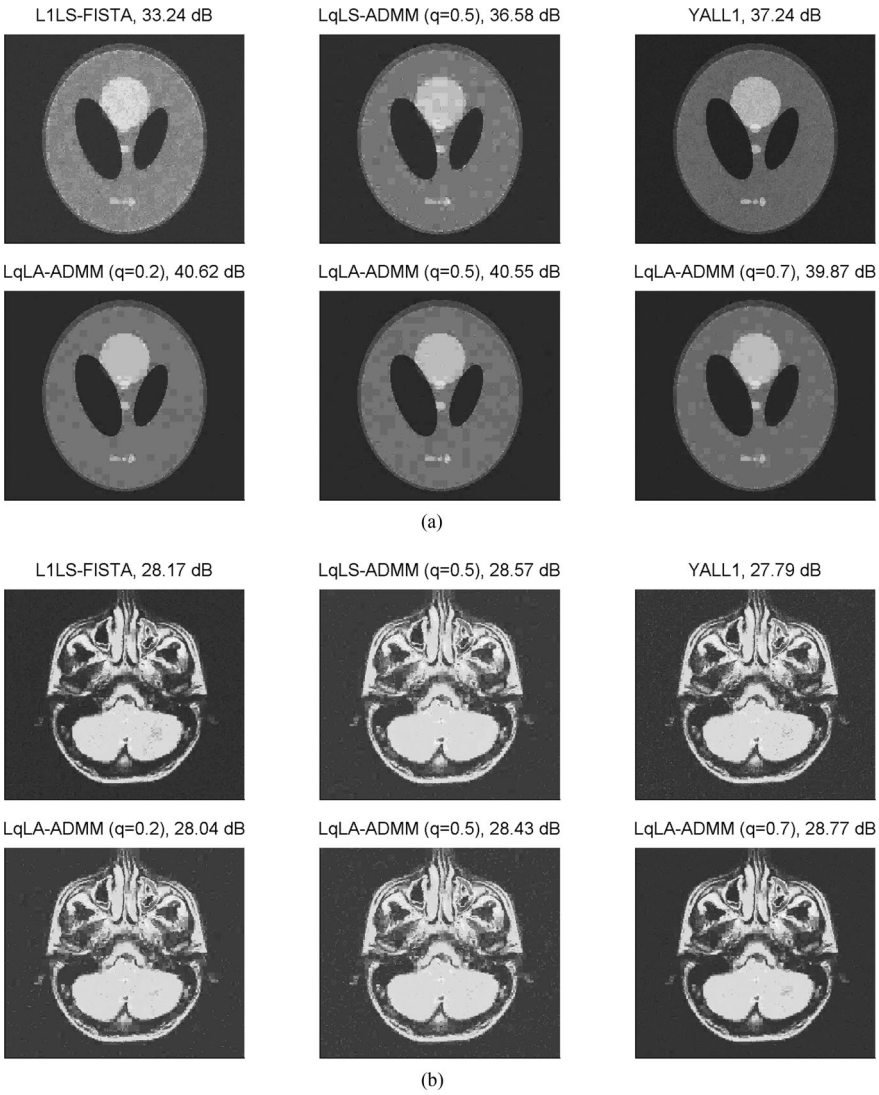



Fig. 4. Recovery performance of the compared methods on two $256 \times 256$ images in Gaussian mixture noise with $\xi = 0.1$, $\kappa = 1000$ and SNR $= 20$ dB. (a) Shepp-Logan. (b) MRI.

change in different impulsive conditions when the noise power is fixed. Meanwhile, the $\ell_1$-loss can yield significantly better performance in Gaussian mixture noise than that in Gaussian noise. This is due to the fact that, the recovery error of the $\ell_1$-loss based formulation is bounded by the first-order moment of the noise (this property can be derived via directly extending [32, Th. 1]), which decreases significantly when the noise gets more impulsive. Note that, these results hold only when the noise has finite variance, e.g., Gaussian or Gaussian mixture noise, and they break down when the noise has infinite variance, e.g., $\mathrm{S}\alpha\mathrm{S}$ impulsive noise as shown in Fig. 2(c).

### *B. Recovery of Images*

This experiment evaluates the algorithms on image recovery. The used images include a synthetic image, "Shepp-Logan", and a magnetic resonance imaging (MRI) image, as shown in Fig. 3. Each image has a size $256 \times 256$ ($n = 65536$), and the measurement number is set to $m = \mathrm{round}(0.4n)$. We employ a partial discrete cosine transformation (DCT) matrix as the sensing matrix $\mathbf{A}$, which is obtained by randomly selecting $m$ out of $n$ rows of the full DCT matrix. We use an implicit representation of this matrix since it is hardly explicitly available in high-dimensional conditions. Another advantage of using such a sensing matrix is that the multiplication of $\mathbf{A}$ (or $\mathbf{A}^T$) with a vector can be rapidly obtained via picking the discrete cosine transform of the vector. We use the Haar wavelets as the basis functions and consider two impulsive noise conditions, Gaussian mixture noise with $\xi = 0.1$, $\kappa = 1000$ and SNR $=$ 20 dB, and $\mathrm{S}\alpha\mathrm{S}$ noise with $\alpha = 1$ and $\gamma = 10^{-4}$. The recovery performance is evaluated in terms of peak-signal noise ratio (PSNR).

Fig. 4 shows the recovery performance of the compared algorithms in Gaussian mixture noise. The PSNR results are shown in Table I. It can be seen that each algorithm can achieve much higher PSNR in recovering the synthetic image than that in recovering the MRI image. This is due the nature that, the Haar wavelet coefficients of the synthetic image "Shepp-Logan" are truly sparse (approximately $3.2\%$ nonzeros), while the wavelet

TABLE I
PSNR OF IMAGE RECOVERY BY THE COMPARED ALGORITHMS IN TWO NOISE CONDITIONS

| Method | | L1LS-FISTA | LqLS-ADMM ($q = 0.5$) | YALL1 | LqLA-ADMM ($q = 0.2$) | LqLA-ADMM ($q = 0.5$) | LqLA-ADMM ($q = 0.7$) |
|---|---|---|---|---|---|---|---|
| Gaussian mixture | Logan | 33.24 | 36.58 | 37.24 | **40.62** | 40.55 | 39.87 |
| | MRI | 28.17 | 28.57 | 27.79 | 28.04 | 28.43 | **28.77** |
| S$\alpha$S | Logan | 12.33 | 12.25 | 28.37 | 29.76 | 33.73 | **35.96** |
| | MRI | 14.60 | 14.53 | 24.56 | 25.51 | 27.05 | **27.29** |

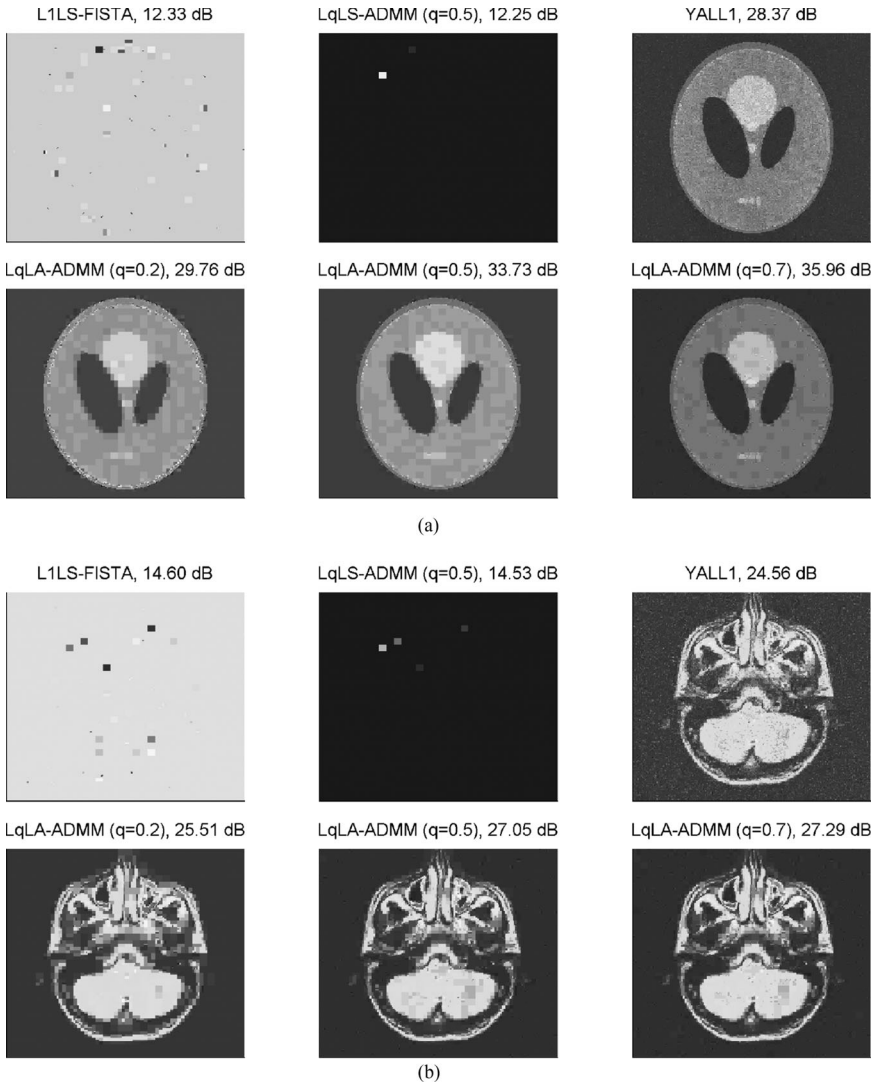



Fig. 5. Recovery performance of the compared methods on two $256 \times 256$ images in $S\alpha S$ noise with $\alpha = 1$ and $\gamma = 10^{-4}$. (a) Shepp-Logan. (b) MRI.

coefficients of a real-life image are not sparse but rather approximately follow an exponential decay, which is referred to as compressible. Moreover, LqLA-ADMM significantly outperforms the other algorithms in recovering "Shepp-Logan", e.g., the improvements attained by LqLA-ADMM (with $q = 0.2$) over L1LS-FISTA, LqLS-ADMM and YALL1 are 7.38, 4.04 and 3.38 dB, respectively. However, this advantage decreases in recovering the MRI image, e.g., the improvements attained by LqLA-ADMM (with $q = 0.7$) over L1LS-FISTA, LqLS-ADMM and YALL1 are 0.6, 0.2 and 0.98 dB, respectively. The results indicate that the advantage of an $\ell_q$-regularization based algorithm over an $\ell_1$-regularization based algorithm generally decreases as the compressibility of the image decreases.

Fig. 5 presents the recovery performance of the compared algorithms in the S$\alpha$S noise condition. The PSNR results are shown in Table I. The considered S$\alpha$S noise with $\alpha = 1$ contains extremely large outliers and is more impulsive than the Gaussian mixture noise. It can be seen in Fig. 5 that the $\ell_2$-loss based algorithms, L1LS-FISTA and LqLS-ADMM, break down, while the $\ell_1$-loss based algorithms, YALL1 and LqLA-ADMM,

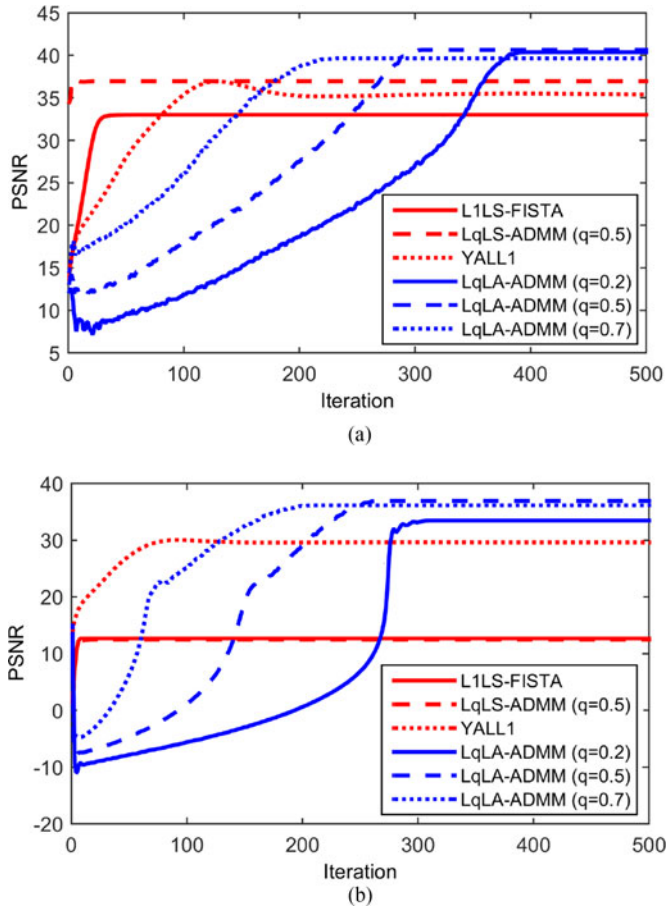



Fig. 6. PSNR versus optimization iterations for the compared algorithms in recovering the Shepp-Logan image. (a) Gaussian mixture noise. (b) $\mathrm{S}\alpha\mathrm{S}$ noise.

work well. LqLA-ADMM again achieves the best performance, and its advantage over YALL1 is more significant in this noise condition than that in the Gaussian mixture noise condition in recovering the MRI image. For example, in recovering the MRI image, the improvement attained by LqLA-ADMM (with $q = 0.7$) over YALL1 in Gaussian mixture noise is 0.98 dB, while that in $\mathrm{S}\alpha\mathrm{S}$ noise is 2.73 dB.

Moreover, the results show that in recovering the MRI image, for LqLA-ADMM, $q = 0.5$ and $q = 0.7$ yield better performance than $q = 0.2$, which is different form the results in the previous experiment, where $q = 0.2$ and $q = 0.5$ generate significantly better performance than $q = 0.7$ in recovering simulated sparse signals. This is due to the nature that, real-life images are not strictly sparse as simulated sparse signals but rather compressible, e.g., with wavelet coefficients approximately follow an exponential decay.

Fig. 6 presents a typical plot of PSNR against optimization iterations for the compared algorithms in recovering the Shepp-Logan image. It can be observed that the proposed algorithm needs more iterations to converge than L1LS-FISTA, LqLS-ADMM, and YALL1, especially for small $q$. Similar to L1LS-FISTA, LqLS-ADMM, and YALL1, the proposed LqLA-ADMM algorithm is also a first-order algorithm and scales well for large-scale problems, as the dominant computational load in each iteration is matrix-vector multiplication with complexity $O(mn)$.

## VI. Conclusion

This work introduced a robust formulation for sparse recovery, which improves the $\ell_1$-LA formulation via replacing the $\ell_1$-regularization by a generalized nonconvex regularization. A first-order algorithm based on ADMM has been developed to efficiently solve the nonconvex and nonsmooth minimization problem. In developing the new algorithm, a smoothing strategy on the $\ell_1$-loss function has been used to make it convergent. Moreover, a sufficient condition for the convergence of the new algorithm has been derived for a generalized nonconvex penalty. Simulation results on recovering both sparse vector-valued signals and images demonstrated that, in impulsive noise, the new method offers considerable performance gain over the methods which solve the $\ell_1$-LS, $\ell_q$-LS, and $\ell_1$-LA formulations.

## Appendix A
## Proof of Lemma 1

Let $h_1(\mathbf{x}) = \frac{\rho}{2}\left\|\mathbf{A}\mathbf{x} - \mathbf{y} - \mathbf{v}^k - \mathbf{w}^k/\rho\right\|_2^2$, the $\mathbf{x}$-subproblem in fact minimizes the following approximated objective

$$Q_{\mathbf{x}^k}(\mathbf{x}) = P(\mathbf{x}) + \left\langle \mathbf{x} - \mathbf{x}^k, \nabla h_1(\mathbf{x}^k)\right\rangle + \frac{\rho}{2\tau_1}\left\|\mathbf{x} - \mathbf{x}^k\right\|_2^2.$$

From the definition of $\mathbf{x}^{k+1}$ as a minimizer of $Q_{\mathbf{x}^k}(\mathbf{x})$, we have

$$\begin{aligned} &Q_{\mathbf{x}^k}(\mathbf{x}^{k+1}) \\ &= P(\mathbf{x}^{k+1}) + \left\langle \mathbf{x}^{k+1} - \mathbf{x}^k, \nabla h_1(\mathbf{x}^k)\right\rangle + \frac{\rho}{2\tau_1}\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|_2^2 \\ &\le Q_{\mathbf{x}^k}(\mathbf{x}^k) = P(\mathbf{x}^k). \end{aligned} \tag{25}$$

Further, the Hessian of $h_1(\mathbf{x})$ is

$$\nabla^2 h_1(\mathbf{x}) = \rho\mathbf{A}^T\mathbf{A}$$

which implies that $\nabla h_1(\mathbf{x})$ is $\rho\lambda_{\max}(\mathbf{A}^T\mathbf{A})$-Lipschitz continuous. Thus, for any $\mathbf{x}^k, \mathbf{x}^{k+1} \in \mathbb{R}^n$ we have

$$\begin{aligned} h_1(\mathbf{x}^{k+1}) \le\ & h_1(\mathbf{x}^k) + \left\langle \mathbf{x}^{k+1} - \mathbf{x}^k, \nabla h_1(\mathbf{x}^k)\right\rangle \\ &+ \frac{\rho\lambda_{\max}(\mathbf{A}^T\mathbf{A})}{2}\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|_2^2. \end{aligned} \tag{26}$$

It follows from (25) and (26) that

$$\begin{aligned} &P(\mathbf{x}^{k+1}) + h_1(\mathbf{x}^{k+1}) \\ &\le P(\mathbf{x}^{k+1}) + h_1(\mathbf{x}^k) + \left\langle \mathbf{x}^{k+1} - \mathbf{x}^k, \nabla h_1(\mathbf{x}^k)\right\rangle \\ &\quad + \frac{\rho\lambda_{\max}(\mathbf{A}^T\mathbf{A})}{2}\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|_2^2 \\ &\le P(\mathbf{x}^k) + h_1(\mathbf{x}^k) - c_0\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|_2^2 \end{aligned}$$

which results in Lemma 1.

## APPENDIX B
## PROOF OF LEMMA 2

First, the Hessian of $\|\mathbf{v}\|_{1,\varepsilon}$ is

$$\nabla^2\|\mathbf{v}\|_{1,\varepsilon} = \varepsilon^2 \mathrm{diag}\left\{(v_1^2+\varepsilon^2)^{-\frac{3}{2}}, \cdots, (v_N^2+\varepsilon^2)^{-\frac{3}{2}}\right\} \preceq \frac{1}{\varepsilon}\mathbf{I}_n \tag{27}$$

which implies that $\nabla\|\mathbf{v}\|_{1,\varepsilon}$ is $\frac{1}{\varepsilon}$-Lipschitz continuous, thus, for any $\mathbf{v}^k, \mathbf{v}^{k+1} \in \mathbb{R}^m$ we have

$$\begin{aligned}\left\|\mathbf{v}^{k+1}\right\|_{1,\varepsilon} \leq \left\|\mathbf{v}^k\right\|_{1,\varepsilon} + \langle \mathbf{v}^{k+1}-\mathbf{v}^k, \nabla\|\mathbf{v}^k\|_{1,\varepsilon}\rangle \\ + \frac{1}{2\varepsilon}\left\|\mathbf{v}^{k+1}-\mathbf{v}^k\right\|_2^2.\end{aligned} \tag{28}$$

Let $h_2(\mathbf{v}) = \frac{\rho}{2}\left\|\mathbf{A}\mathbf{x}^{k+1}-\mathbf{y}-\mathbf{v}-\mathbf{w}^k/\rho\right\|_2^2$, the $\mathbf{v}$-subproblem actually minimizes the following approximated objective

$$G_{\mathbf{v}^k}(\mathbf{v}) = \frac{1}{\mu}\left\langle \mathbf{v}-\mathbf{v}^k, \nabla\left\|\mathbf{v}^k\right\|_{1,\varepsilon}\right\rangle + \frac{1}{2\mu\tau_2}\left\|\mathbf{v}-\mathbf{v}^k\right\|_2^2 + h_2(\mathbf{v}). \tag{29}$$

Since $G_{\mathbf{v}^k}(\mathbf{v})$ is $(\frac{1}{\mu\tau_2}+\rho)$-strongly convex, for any $\mathbf{v}^k \in \mathbb{R}^m$ we have

$$\begin{aligned}G_{\mathbf{v}^k}(\mathbf{v}^k) \geq G_{\mathbf{v}^k}(\mathbf{v}^{k+1}) + \left\langle \mathbf{v}^k-\mathbf{v}^{k+1}, \nabla G_{\mathbf{v}^k}(\mathbf{v}^{k+1})\right\rangle \\ + \frac{1}{2}\left(\frac{1}{\mu\tau_2}+\rho\right)\left\|\mathbf{v}^k-\mathbf{v}^{k+1}\right\|_2^2.\end{aligned} \tag{30}$$

From the definition of $\mathbf{v}^{k+1}$ as a minimizer of $G_{\mathbf{v}^k}(\mathbf{v})$, we have $\nabla G_{\mathbf{v}^k}(\mathbf{v}^{k+1}) = 0$. Further, since $G_{\mathbf{v}^k}(\mathbf{v}^k) = h_2(\mathbf{v}^k)$, it follows from (29) and (30) that

$$\begin{aligned}&\frac{1}{\mu}\left\langle \mathbf{v}^{k+1}-\mathbf{v}^k, \nabla\left\|\mathbf{v}^k\right\|_{1,\varepsilon}\right\rangle + h_2(\mathbf{v}^{k+1}) \\ &\leq h_2(\mathbf{v}^k) - \left(\frac{1}{\mu\tau_2}+\frac{\rho}{2}\right)\left\|\mathbf{v}^{k+1}-\mathbf{v}^k\right\|_2^2\end{aligned}$$

which together with (28) yields

$$\begin{aligned}&\frac{1}{\mu}\left\|\mathbf{v}^{k+1}\right\|_{1,\varepsilon} + h_2(\mathbf{v}^{k+1}) \\ &\leq \frac{1}{\mu}\left\|\mathbf{v}^k\right\|_{1,\varepsilon} + h_2(\mathbf{v}^k) - c_1\left\|\mathbf{v}^{k+1}-\mathbf{v}^k\right\|_2^2\end{aligned}$$

which finally results in Lemma 2.

## APPENDIX C
## PROOF OF LEMMA 3

First, we show that the changes in the dual iterates can be bounded by the changes in the primal iterates. Observe that the approximated $\mathbf{v}$-subproblem actually minimizes the objective $G_{\mathbf{v}^k}(\mathbf{v})$ given in (29), whose minimizer $\mathbf{v}^{k+1}$ satisfies

$$\begin{aligned}\nabla\left\|\mathbf{v}^k\right\|_{1,\varepsilon} + \frac{1}{\tau_2}\left(\mathbf{v}^{k+1}-\mathbf{v}^k\right) \\ + \mu\rho\left(\mathbf{A}\mathbf{x}^{k+1}-\mathbf{y}-\mathbf{v}^{k+1}-\mathbf{w}^k/\rho\right) = \mathbf{0}.\end{aligned} \tag{31}$$

Substituting (16) into (31) yields

$$\mathbf{w}^{k+1} = \frac{1}{\mu}\nabla\left\|\mathbf{v}^k\right\|_{1,\varepsilon} + \frac{1}{\mu\tau_2}(\mathbf{v}^{k+1}-\mathbf{v}^k). \tag{32}$$

Then, it follows that

$$\begin{aligned}&\left\|\mathbf{w}^{k+1}-\mathbf{w}^k\right\|_2^2 \\ &\leq \frac{1}{\mu^2}\left(\left\|\nabla\left\|\mathbf{v}^k\right\|_{1,\varepsilon} - \nabla\left\|\mathbf{v}^{k-1}\right\|_{1,\varepsilon}\right\|_2 + \frac{1}{\tau_2}\left\|\mathbf{v}^{k+1}-\mathbf{v}^k\right\|_2 \right. \\ &\qquad \left. + \frac{1}{\tau_2}\left\|\mathbf{v}^k-\mathbf{v}^{k-1}\right\|_2\right)^2 \\ &\leq \frac{1}{\mu^2}\left(\frac{1}{\tau_2}\left\|\mathbf{v}^{k+1}-\mathbf{v}^k\right\|_2 + \left(\frac{1}{\varepsilon}+\frac{1}{\tau_2}\right)\left\|\mathbf{v}^k-\mathbf{v}^{k-1}\right\|_2\right)^2 \\ &\leq \frac{2}{\mu^2\tau_2^2}\left\|\mathbf{v}^{k+1}-\mathbf{v}^k\right\|_2^2 + \frac{2}{\mu^2}\left(\frac{1}{\varepsilon}+\frac{1}{\tau_2}\right)^2\left\|\mathbf{v}^k-\mathbf{v}^{k-1}\right\|_2^2\end{aligned} \tag{33}$$

where the second inequality follows from (27).

From (16) and the definition of $\mathcal{L}_\varepsilon$, we have

$$\begin{aligned}&\mathcal{L}_\varepsilon(\mathbf{v}^{k+1},\mathbf{x}^{k+1},\mathbf{w}^{k+1}) - \mathcal{L}_\varepsilon(\mathbf{v}^{k+1},\mathbf{x}^{k+1},\mathbf{w}^k) \\ &\qquad = \frac{1}{\rho}\left\|\mathbf{w}^{k+1}-\mathbf{w}^k\right\|_2^2.\end{aligned} \tag{34}$$

Then, with the use of (33), it follows from Lemma 1, Lemma 2 and (34) that

$$\begin{aligned}&\mathcal{L}_\varepsilon(\mathbf{v}^{k+1},\mathbf{x}^{k+1},\mathbf{w}^{k+1}) - \mathcal{L}_\varepsilon(\mathbf{v}^k,\mathbf{x}^k,\mathbf{w}^k) \\ &\leq -c_0\left\|\mathbf{x}^{k+1}-\mathbf{x}^k\right\|_2^2 - \left(c_1 - \frac{2}{\rho\mu^2\tau_2^2}\right)\left\|\mathbf{v}^{k+1}-\mathbf{v}^k\right\|_2^2 \\ &\qquad + \frac{2}{\rho\mu^2}\left(\frac{1}{\varepsilon}+\frac{1}{\tau_2}\right)^2\left\|\mathbf{v}^k-\mathbf{v}^{k-1}\right\|_2^2\end{aligned}$$

which consequently results in Lemma 3, where $c_3$ is positive when (24) holds. Moreover, it is easy to see that, when (24) is satisfied, $c_1$ in Lemma 2 is also positive, which implies the sufficient decrease of $\mathcal{L}_\varepsilon$ by the $\mathbf{v}$-subproblem updated via (23).

## APPENDIX D
## PROOF OF LEMMA 4

First, we show the sequence $\{\mathbf{z}^k\}$ generated via (17), (23) and (16) is bounded. From (32), we have

$$\begin{aligned}\left\|\mathbf{w}^k\right\|_2^2 &\leq \frac{1}{\mu}\left(\left\|\nabla\left\|\mathbf{v}^{k-1}\right\|_{1,\varepsilon}\right\|_2 + \frac{1}{\tau_2}\left\|\mathbf{v}^k-\mathbf{v}^{k-1}\right\|_2\right)^2 \\ &\leq \frac{2}{\mu^2}\left\|\nabla\left\|\mathbf{v}^{k-1}\right\|_{1,\varepsilon}\right\|_2^2 + \frac{2}{\mu^2\tau_2^2}\left\|\mathbf{v}^k-\mathbf{v}^{k-1}\right\|_2^2 \\ &\leq \frac{2n}{\mu^2} + \frac{2}{\mu^2\tau_2^2}\left\|\mathbf{v}^k-\mathbf{v}^{k-1}\right\|_2^2\end{aligned} \tag{35}$$

where the last inequality follows from $\|\nabla\|\mathbf{v}^k\|_{1,\varepsilon}\|_2^2 \leq n$ when $\varepsilon > 0$. Define $\tilde{\mathbf{z}}^k := (\mathbf{v}^k,\mathbf{x}^k,\mathbf{w}^k,\mathbf{x}^{k-1})$, under the assumption that $\tilde{\mathcal{L}}(\tilde{\mathbf{z}}^k)$ is lower semicontinuous, it is bounded from below. Further, when (24) holds, $\tilde{\mathcal{L}}(\tilde{\mathbf{z}}^k)$ is nonincreasing by Lemma 3,

thus it is convergent. Then, form the definition of $\tilde{\mathcal{L}}$, we have

$$\begin{aligned}
&\tilde{\mathcal{L}}(\tilde{\mathbf{z}}^1) \geq \tilde{\mathcal{L}}(\tilde{\mathbf{z}}^k)\\
&= \frac{1}{\mu}\left\|\mathbf{v}^k\right\|_{1,\varepsilon} + P(\mathbf{x}^k) + \frac{\rho}{2}\left\|\mathbf{A}\mathbf{x}^k - \mathbf{y} - \mathbf{v}^k - \frac{\mathbf{w}^k}{\rho}\right\|_2^2\\
&\qquad - \frac{1}{2\rho}\left\|\mathbf{w}^k\right\|_2^2 + c_2\left\|\mathbf{v}^k - \mathbf{v}^{k-1}\right\|_2^2\\
&\geq \frac{1}{\mu}\left\|\mathbf{v}^k\right\|_{1,\varepsilon} + P(\mathbf{x}^k) + \frac{\rho}{2}\left\|\mathbf{A}\mathbf{x}^k - \mathbf{y} - \mathbf{v}^k - \frac{\mathbf{w}^k}{\rho}\right\|_2^2\\
&\qquad - \frac{n}{\rho\mu^2} + \left(c_2 - \frac{1}{\rho\mu^2\tau_2^2}\right)\left\|\mathbf{v}^k - \mathbf{v}^{k-1}\right\|_2^2
\end{aligned}$$

where the last inequality follows from (35). Since $c_2 > \frac{1}{\rho\mu^2\tau_2^2}$, when $P(\cdot)$ is coercive (e.g., for the hard-thresholding, soft-thresholding, SCAD, MC, and $\ell_q$-norm penalties), and by (35), it is easy to see that $\mathbf{v}^k$, $\mathbf{x}^k$ and $\mathbf{w}^k$ are bounded.

Since $\tilde{\mathbf{z}}^k$ is bounded, there exists a convergent subsequence $\tilde{\mathbf{z}}^{k_j}$ which converges to a cluster point $\tilde{\mathbf{z}}^*$. Moreover, $\tilde{\mathcal{L}}(\tilde{\mathbf{z}}^k)$ is convergent and $\tilde{\mathcal{L}}(\tilde{\mathbf{z}}^k) \geq \tilde{\mathcal{L}}(\tilde{\mathbf{z}}^*)$ for any $k$ if $c_3 > 0$. Then, it follows from Lemma 3 that

$$\begin{aligned}
&c_0\sum_{k=1}^{N}\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|_2^2 + c_3\sum_{k=1}^{N}\left\|\mathbf{v}^{k+1} - \mathbf{v}^k\right\|_2^2\\
&\leq \sum_{k=1}^{N}\left[\tilde{\mathcal{L}}(\tilde{\mathbf{z}}^k) - \tilde{\mathcal{L}}(\tilde{\mathbf{z}}^{k+1})\right]\\
&= \tilde{\mathcal{L}}(\tilde{\mathbf{z}}^1) - \tilde{\mathcal{L}}(\tilde{\mathbf{z}}^{k+1})\\
&\leq \tilde{\mathcal{L}}(\tilde{\mathbf{z}}^1) - \tilde{\mathcal{L}}(\tilde{\mathbf{z}}^*) < \infty.
\end{aligned}$$

Let $N \to \infty$, since $c_0 > 0$ and $c_3 > 0$ when $\tau_1 < 1/\lambda_{\max}(\mathbf{A}^T\mathbf{A})$ and (24) are satisfied, we have

$$\sum_{k=1}^{\infty}\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|_2^2 < \infty$$

$$\sum_{k=1}^{\infty}\left\|\mathbf{v}^{k+1} - \mathbf{v}^k\right\|_2^2 < \infty$$

which together with (33) implies

$$\sum_{k=1}^{\infty}\left\|\mathbf{w}^{k+1} - \mathbf{w}^k\right\|_2^2 < \infty.$$

Thus, we have $\lim_{k\to\infty}\left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\|_2^2 = 0$.

Next, we show that any cluster point of the sequence $\{\mathbf{z}^k\}$ generated via (17), (23) and (16) is a stationary point of (21). From the optimality conditions, the sequence generated via (17), (23) and (16) satisfies

$$\begin{cases}
\mathbf{0} \in \partial P(\mathbf{x}^{k+1}) - \mathbf{A}^T\mathbf{w}^{k+1} + \rho\mathbf{A}^T(\mathbf{v}^{k+1} - \mathbf{v}^k)\\
\qquad\qquad + \dfrac{\rho}{\tau_1}(\mathbf{x}^{k+1} - \mathbf{x}^k),\\
\mathbf{0} = \dfrac{1}{\mu}\nabla\left\|\mathbf{v}^{k+1}\right\|_{1,\varepsilon} + \mathbf{w}^{k+1} + \dfrac{1}{\mu\tau_2}(\mathbf{v}^{k+1} - \mathbf{v}^k),\\
\mathbf{w}^{k+1} = \mathbf{w}^k - \rho(\mathbf{A}\mathbf{x}^{k+1} - \mathbf{y} - \mathbf{v}^{k+1}).
\end{cases} \tag{36}$$

Let $\{\mathbf{z}^{k_j}\}$ be a convergent subsequence of $\{\mathbf{z}^k\}$, since $\lim_{k\to\infty}\left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\|_2^2 = 0$, $\mathbf{z}^{k_j}$ and $\mathbf{z}^{k_j+1}$ have the same limit point $\mathbf{z}^* := (\mathbf{v}^*, \mathbf{x}^*, \mathbf{w}^*)$. Moreover, since $\tilde{\mathcal{L}}(\tilde{\mathbf{z}}^k)$ is convergent, $P(\mathbf{x}^k)$ is also convergent. Then, passing to the limit in (36) along the subsequence $\{\mathbf{z}^{k_j}\}$ yields

$$\mathbf{A}^T\mathbf{w}^* \in \partial P(\mathbf{x}^*),\ -\mathbf{w}^* = \frac{1}{\mu}\nabla\|\mathbf{v}^*\|_{1,\varepsilon},\ \mathbf{A}\mathbf{x}^* - \mathbf{y} = \mathbf{v}^*.$$

In particular, $\mathbf{z}^*$ is a stationary point of $\mathcal{L}_\varepsilon$.

## Appendix E
## Proof of Lemma 5

Let $\tilde{\mathbf{z}}^k := (\mathbf{v}^k, \mathbf{x}^k, \mathbf{w}^k, \mathbf{x}^{k-1})$, from the definition of $\tilde{\mathcal{L}}(\tilde{\mathbf{z}}^{k+1})$, we have

$$\partial_{\mathbf{x}}\tilde{\mathcal{L}}(\tilde{\mathbf{z}}^{k+1}) = \partial P(\mathbf{x}^{k+1}) - \mathbf{A}^T\mathbf{w}^{k+1} + \mathbf{A}^T(\mathbf{w}^k - \mathbf{w}^{k+1})$$

which together with the first relation in (36) yields

$$\begin{aligned}
&\rho\mathbf{A}^T(\mathbf{v}^k - \mathbf{v}^{k+1}) + \frac{\rho}{\tau_1}(\mathbf{x}^k - \mathbf{x}^{k+1})\\
&\qquad + \mathbf{A}^T(\mathbf{w}^k - \mathbf{w}^{k+1}) \in \partial_{\mathbf{x}}\tilde{\mathcal{L}}(\tilde{\mathbf{z}}^{k+1}).
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
&\nabla_{\mathbf{v}}\tilde{\mathcal{L}}(\tilde{\mathbf{z}}^{k+1})\\
&= \frac{1}{\mu}\nabla\left\|\mathbf{v}^{k+1}\right\|_{1,\varepsilon} + \mathbf{w}^{k+1} - \rho(\mathbf{w}^k - \mathbf{w}^{k+1}) + 2c_2(\mathbf{v}^{k+1} - \mathbf{v}^k)\\
&= \rho(\mathbf{w}^{k+1} - \mathbf{w}^k) + \left(2c_2 - \frac{1}{\mu\tau_2}\right)(\mathbf{v}^{k+1} - \mathbf{v}^k)
\end{aligned}$$

where the second equality follows from the second relation in (36). Similarly,

$$\nabla_{\tilde{\mathbf{v}}}\tilde{\mathcal{L}}(\tilde{\mathbf{z}}^{k+1}) = 2c_2(\mathbf{v}^{k+1} - \mathbf{v}^k),$$

$$\nabla_{\mathbf{w}}\tilde{\mathcal{L}}(\tilde{\mathbf{z}}^{k+1}) = \mathbf{A}\mathbf{x}^{k+1} - \mathbf{y} - \mathbf{v}^{k+1} = \frac{1}{\rho}(\mathbf{w}^k - \mathbf{w}^{k+1}).$$

Thus, we can find a constant $c_5 > 0$ such that

$$\begin{aligned}
&\operatorname{dist}(0, \partial\tilde{\mathcal{L}}(\tilde{\mathbf{z}}^{k+1}))\\
&\leq c_5(\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|_2 + \left\|\mathbf{v}^{k+1} - \mathbf{v}^k\right\|_2 + \left\|\mathbf{w}^k - \mathbf{w}^{k+1}\right\|_2)
\end{aligned}$$

which together with (33) consequently results in Lemma 5.

## Appendix F
## Proof of Theorem 1

Let $\mathbf{z}^k := (\mathbf{v}^k, \mathbf{x}^k, \mathbf{w}^k)$, based on the above lemmas, the rest proof of Theorem 1 is to show that the sequence $\{\mathbf{z}^k\}$ has finite length, i.e.,

$$\sum_{k=0}^{\infty}\left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\|_2 < \infty \tag{37}$$

which implies that $\{\mathbf{z}^k\}$ is a Cauchy sequence and thus is convergent. Finally, the property (37) together with Lemma 4 implies that the sequence $\{\mathbf{z}^k\}$ converges to a stationary point of $\mathcal{L}_\varepsilon$. The derivation of (37) relies heavily on the KL property of $\tilde{\mathcal{L}}$, which holds if the penalty $P(\cdot)$ is a KL function. This is the case of the

hard-thresholding, soft-thresholding, SCAD, MC and $\ell_q$-norm penalties with $0 \leq q \leq 1$. With the above lemmas, the proof of (37) follows similarly the proof of [48, Th. 3] with some minor changes, thus is omitted here for succinctness.

## REFERENCES


[1] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[2] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[3] L. C. Potter, E. Ertin, J. T. Parker, and M. Cetin, "Sparsity and compressed sensing in radar imaging," *IEEE Proc.*, vol. 98, no. 6, pp. 1006–1020, Jun. 2010.

[4] C. R. Berger, Z. Wang, J. Huang, and S. Zhou, "Application of compressive sensing to sparse channel estimation," *IEEE Commun. Mag.*, vol. 48, no. 11, pp. 164–174, Nov. 2010.

[5] X. Jiang, R. Ying, F. Wen *et al.*, "An improved sparse reconstruction algorithm for speech compressive sensing using structured priors," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, 2016, pp. 1–6.

[6] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Magn. Reson. Med.*, vol. 58, no. 6, pp. 1182–1195, 2007.

[7] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[8] J. Tan, Y. Ma, and D. Baron, "Compressive imaging via approximate message passing with image denoising," *IEEE Trans. Signal Process.*, vol. 63, no. 8, pp. 2085–2092, Apr. 2015.

[9] T. Köhler, X. Huang, F. Schebesch, A. Aichert, A. Maier, and J. Hornegger, "Robust multiframe super-resolution employing iteratively reweighted minimization," *IEEE Trans. Comput. Imaging*, vol. 2 no. 1, pp. 42–58, Mar. 2016.

[10] C. Li, T. Sun, K. F. Kelly, and Y. Zhang, "A compressive sensing and unmixing scheme for hyperspectral data processing," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1200–1210, Mar. 2012.

[11] E. J. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, Aug. 2006.

[12] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. R. Statist. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.

[13] R. Chartrand and V. Staneva, "Restricted isometry properties and nonconvex compressive sensing," *Inverse Probl.*, vol. 24, 2008, Art. no. 035020.

[14] E. J. Candès, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted $\ell_1$ minimization," *J. Fourier Anal. Appl.*, vol. 14, pp. 877–905, 2008.

[15] S. Foucart and M.–J. Lai, "Sparsest solutions of underdetermined linear systems via $\ell_q$–minimization for $0 < q \leq 1$," *Appl. Comput. Harmon. Anal.*, vol. 26, no. 3, pp. 395–407, May 2009.

[16] H. Mohimani, M. Babie–Zadeh, and C. Jutten, "A fast approach for overcomplete sparse decomposition based on smoothed $\ell_0$–norm," *IEEE Trans. Signal Process.*, vol. 57, no. 1, pp. 289–301, Jan. 2009.

[17] I. Daubechies *et al.*, "Iteratively reweighted least squares minimization for sparse recovery," *Commun. Pure Appl. Math.*, vol. 63, no. 1, pp. 1–38, 2010.

[18] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *Proc. IEEE Int. Conf. Acoust, Speech, Signal Process.*, 2008, pp. 3869–3872.

[19] F. Wen, L. Adhikari, L. Pei, R. F. Marcia, P. Liu, and R.C. Qiu, "Nonconvex regularization based sparse recovery and demixing with application to color image inpainting," *IEEE Access*, vol. 5, pp. 11513–11527, 2017.

[20] F. Wen, Y. Yang, P. Liu, and R. C. Qiu, "Positive definite estimation of large covariance matrix using generalized nonconvex penalties," *IEEE Access*, vol. 4, pp. 4168–4182, 2016.

[21] M.–J. Lai, Y. Xu, and W. Yin, "Improved iteratively reweighted least squares for unconstrained smoothed $\ell_q$ minimization," *SIAM J. Numer. Anal.*, vol. 51, no. 2, pp. 927–957, 2013.

[22] J. K. Pant, W. Lu, and A. Antoniou, "New improved algorithms for compressive sensing based on $\ell_p$-norm," *IEEE Trans. Circuits Syst. II, Express Briefs*, vol. 61, no. 3, pp. 198–202, Mar. 2014.

[23] Q. Sun, "Recovery of sparsest signals via $\ell_q$-minimization," *Appl. Comput. Harmon. Anal.*, vol. 32, no. 3, pp. 329–341, 2012.

[24] P. J. Huber, *Robust Statistics*. New York, NY, USA: Wiley, 1981.

[25] L. Bar, A. Brook, N. Sochen, and N. Kiryati, "Deblurring of color images corrupted by impulsive noise," *IEEE Trans. Image Process.*, vol. 16, no. 4, pp. 1101–1111, Apr. 2007.

[26] P. Windyga, "Fast impulsive noise removal," *IEEE Trans. Image Process.*, vol. 10, no. 1, pp. 173–179, Jan. 2001.

[27] P. Civicioglu, "Using uncorrupted neighborhoods of the pixels for impulsive noise suppression with ANFIS," *IEEE Trans. Image Process.*, vol. 16, no. 3, pp. 759–773, Mar. 2007.

[28] T. Hashimoto, "Bounds on a probability for the heavy tailed distribution and the probability of deficient decoding in sequential decoding," *IEEE Trans. Inf. Theory*, vol. 51, no. 3, pp. 990–1002, Mar. 2005.

[29] E. J. Candès and P. A. Randall, "Highly robust error correction by convex programming," *IEEE Trans. Inf. Theory*, vol. 54, no. 7, pp. 2829–2840, Jul. 2008.

[30] B. Popilka, S. Setzer, and G. Steidl, "Signal recovery from incomplete measurements in the presence of outliers," *Inverse Probl. Imag.*, vol. 1, no. 4, pp. 661–672, Nov. 2007.

[31] R. Chan, C.-W. Ho, and M. Nikolova, "Salt-and-pepper noise removal by median-type noise detectors and de tail-preserving regularization," *IEEE Trans. Image Process.*, vol. 14, no. 10, pp. 1479–1485, Oct. 2005.

[32] R. E. Carrillo, K. E. Barner, and T. C. Aysal, "Robust sampling and reconstruction methods for sparse signals in the presence of impulsive noise," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 392–408, Apr. 2010.

[33] R. E. Carrillo and K. E. Barner, "Lorentzian iterative hard thresholding: Robust compressed sensing with prior information," *IEEE Trans. Signal Process.*, vol. 61, no. 19, pp. 4822–4833, Oct. 2013.

[34] J. F. Yang and Y. Zhang, "Alternating direction algorithms for $\ell_1$-problems in compressive sensing," *SIAM J. Sci. Comput.*, vol. 33, pp. 250–278, 2011.

[35] Y. Xiao, H. Zhu, and S.-Y. Wu, "Primal and dual alternating direction algorithms for $\ell_1$-$\ell_1$-norm minimization problems in compressive sensing," *Comput. Optim. Appl.*, vol. 54, no. 2, pp. 441–459, 2013.

[36] D. S. Pham and S. Venkatesh, "Improved image recovery from compressed data contaminated with impulsive noise," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 397–405, Jan. 2012.

[37] D. S. Pham and S. Venkatesh, "Efficient algorithms for robust recovery of images from compressed data," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 4724–4737, Dec. 2013.

[38] F. Wen, P. Liu . Y. Liu, R. C. Qiu, and W. Yu, "Robust sparse recovery for compressive sensing in impulsive noise using Lp-norm model fitting," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 4643–4647.

[39] F. Wen, P. Liu, Y. Liu, R. C. Qiu, and W. Yu, "Robust sparse recovery in impulsive noise via $\ell_p$-$\ell_1$ optimization," *IEEE Trans. Signal Process.*, vol. 65, no. 1, pp. 105–118, Jan. 2017.

[40] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma, "Toward a practical face recognition system: Robust alignment and illumination by sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 2, pp. 372–386, Feb. 2012.

[41] X. Jiang, T. Kirubarajan, and W.-J. Zeng, "Robust sparse channel estimation and equalization in impulsive noise using linear programming," *Signal Process.*, vol. 93, no. 5, pp. 1095–1105, 2013.

[42] A. Antoniadis, "Wavelets in statistics: A review," *J. Italian Stat. Assoc.*, vol. 6, pp. 97–144, 1997.

[43] Z. Xu, X. Chang, F. Xu, and H. Zhang, "L1/2 regularization: A thresholding representation theory and a fast solver," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 7, pp. 1013–1027, Jul. 2012.

[44] G. Marjanovic and V. Solo, "On $\ell_q$ optimization and matrix completion," *IEEE Trans. Signal Process.*, vol. 60, no. 11, pp. 5714–5724, Nov. 2012.

[45] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *J. Amer. Stat. Assoc.*, vol. 96, 1348–1360, 2001.

[46] C. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *Ann. Stat.*, vol. 38, pp. 894–942, 2010.

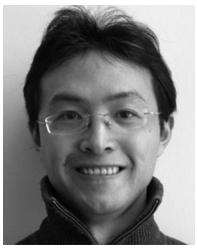
[47] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[48] G. Li and T. K. Pong, "Global convergence of splitting methods for nonconvex composite optimization," *SIAM J. Optim.*, vol. 25, no. 4, pp. 2434–2460, Jul. 2015.

[49] F. Wang, Z. Xu, and H.-K. Xu, "Convergence of Bregman alternating direction method with multipliers for nonconvex composite problems," arXiv:1410.8625, Dec. 2014.

[50] M. Hong, Z. Q. Luo, and M. Razaviyayn, "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems," *SIAM J. Optim.*, vol. 26, no. 1, pp. 337–364, 2016.
[51] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction* (Series in Statistics). New York, NY, USA: Springer, 2001.
[52] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.
[53] S. Ramani and J. A. Fessler, "A splitting-based iterative algorithm for accelerated statistical X-ray CT reconstruction," *IEEE Trans. Med. Imaging*, vol. 31, no. 3, pp. 677–688, Mar. 2012.
[54] E. J. Candes, "The restricted isometry property and its implications for compressed sensing," *Comptes. Rendus. Mathematique*, vol. 346, no. 9, pp. 589–592, 2008.
[55] H. Attouch and J. Bolte, "On the convergence of the proximal algorithm for nonsmooth functions involving analytic features," *Math. Program.*, vol. 116, pp. 5–16, 2009.

**Fei Wen** (M'15) received the B.S. degree and the Ph.D. degree in communications and information engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2006 and 2013, respectively. Since December 2012, he has been a Lecturer at the Air Force Engineering University. He is currently a Research Fellow of the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China. His main research interests are statistical signal processing and nonconvex optimization.



**Ling Pei** (M'14) received the Ph.D. degree from Southeast University, Nanjing, China, in 2007. He is an Associate Professor in the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China. From 2007 to 2013, he was a specialist research scientist in the Finnish Geospatial Research Institute. He has authored or coauthored more than 70 scientific papers. He is also an inventor of 14 patents and pending patents. His main research interests include indoor/outdoor seamless positioning, ubiquitous computing, wireless positioning, mobile computing, context-aware applications, location-based services, and navigation of unmanned systems. He was granted the Shanghai Pujiang Talent in 2014.

**Yuan Yang** received the B.S. degree from Air Force Early Warning Academy, Wuhan, China, in 2004, and the MA.SC. and Ph.D. degrees in communications and information engineering from Air Force Engineering University, Xi'an, China, in 2007 and 2010, respectively. He is currently an Associate Professor in Air Force Engineering University. His main research interests include statistical signal and information processing.

**Wenxian Yu** received the B.Sc., M.Sc., and Ph.D. degrees from the National University of Defense Technology, Changsha, China, in 1985, 1988, and 1993, respectively. From 1996 to 2008, he was a Professor with the College of Electronic Science and Engineering, National University of Defense Technology, where he served as the Deputy Head of the College and an Assistant Director of the National Key Laboratory of Automatic Target Recognition. He is currently with the School of Electronic, Information, and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China, where he is a Yangtze River Scholar Distinguished Professor and the Head of the research part of the school, and he was the Executive Dean of the School from 2009 to 2011.

**Peilin Liu** (M'99) received the Ph.D. degree from the University of Tokyo majoring in electronic engineering in 1998 and worked there as a Research Fellow in 1999. From 1999 to 2003, she worked as a Senior Researcher for Central Research Institute of Fujitsu, Tokyo, Japan. Her research mainly focuses on signal processing, low-power computing architecture, and application-oriented SoC design and Verification. She is now a Professor in the Department of Electronic Engineering, Shanghai Jiao Tong University, the Executive Director of Shanghai Key Laboratory of Navigation and Location Based Service, and responsible for a series of important projects, such as BDSSoC platform development, low power and high-performance communication DSP. He is the Chair of Shanghai Chapter of the IEEE Circuit and System.