Title

Robust Linear Regression via ℓ_0 Regularization

Journal

IEEE Transactions on Signal Processing, 66(3)

ISSN

1053-587X

Authors

Liu, Jing
Cosman, Pamela C
Rao, Bhaskar D

Publication Date

2018

DOI

10.1109/tsp.2017.2771720

Peer reviewed

Robust Linear Regression via ℓ_0 Regularization

Jing Liu, *Student Member, IEEE*, Pamela C. Cosman, *Fellow, IEEE*, and Bhaskar D. Rao, *Fellow, IEEE*

***Abstract*—Linear regression in the presence of outliers is an important problem and is challenging as the support of outliers is not known beforehand. Many robust estimators solve this problem via explicitly or implicitly assuming that outliers are sparse and result in large observation errors. We propose an algorithm for robust outlier support identification (AROSI) utilizing a novel objective function with ℓ_0-"norm" regularization which models the sparsity of outliers. The optimization procedure naturally utilizes the large observation error assumption of outliers and directly operates on the ℓ_0-"norm" and is guaranteed to converge. When only sparse outliers are present (no dense inlier noise), we show that, under certain model and algorithm parameter settings, AROSI can recover the solution exactly. In the case, where both dense inlier noise and sparse outliers are present, we prove that the estimation error is bounded. Extensive empirical comparisons with state-of-the-art methods demonstrate the advantage of the proposed method.**

***Index Terms*—ℓ_0 regularization, algorithm for robust outlier support identification, robust linear regression, sparse recovery.**

I. Introduction

In a linear regression setting, the goal is to estimate the linear relationship between two variables: $a \in \mathbb{R}^n$ (explanatory variable) and $y \in \mathbb{R}$ (response variable), from m pairs of training samples $\{(y_i, a_i), i = 1, \ldots, m\}$, where $m > n$. The following model is commonly assumed:

$$y_i = {a_i}^T x + \mu_i, \quad i = 1, \ldots, m \tag{1}$$

or in matrix form: $y = Ax + \mu$, where measurements $y = (y_1, \ldots, y_m)^T$, and matrix $A = [a_1, \ldots, a_m]^T$ are known. $x \in \mathbb{R}^n$ is the model parameter to be estimated, and $\mu = (\mu_1, \ldots, \mu_m)^T$ is the observation error. It is also commonly assumed that A has full column rank. In many linear regression data sets, there are some observations y_i known as *outliers* that have been corrupted by large observation errors [1]. Such outliers often lead to the failure of Ordinary Least Square (OLS) estimation [2]. The goal of robust linear regression is to accurately estimate the model parameter in the presence of these troublesome outliers. Many robust estimators [3]–[5] have been developed in the spirit of *Robust Statistics*. Recently, this problem has received considerable interest from the signal processing community due to its underlying connections with the rapidly developing *Sparse Signal Recovery (SSR)* framework, which aims to recover a sparse solution from an under-determined system of linear equations. The SSR formulation often splits the observation error μ into two terms: $\mu = \eta + e$, where $\eta \in \mathbb{R}^m$ is small magnitude bounded inlier noise, and $e \in \mathbb{R}^m$ represents the large error component that captures outliers. So model (1) becomes:

$$y = Ax + \eta + e. \tag{2}$$

Additional prior information or assumptions are needed in order to solve the problem. We make the following two reasonable and common assumptions about outliers:

1. Outlier entries often have significantly larger observation errors than inlier entries have, and $\min\{|e_i| : e_i \neq 0\} > \|\eta\|_\infty$.
2. The fraction of outliers in the whole dataset is usually small, so the outlier corruptions vector e is *sparse*, i.e., most entries in e are zero.

In Robust Statistics, many robust regression estimators aim to limit the influence of large error entries under the first assumption. The most popular family of these methods is the M-estimators [5]. For the second assumption, it is often utilized under the principle of fitting the majority of the data. Least Median of Squares (LMedS) [6], Least Trimmed Squares (LTS) [3], [4], and Random Sample Consensus (RANSAC) [7] are representative methods. LMedS was introduced by Rousseeuw [6]; it minimizes the median of squared residuals instead of the mean (or equivalently, sum). To improve estimation efficiency, Rousseeuw further introduced LTS [3], [4], which aims to minimize $\sum_{i=1}^{h} r^2_{(i)}$, where $r^2_{(1)} \leq r^2_{(2)} \cdots \leq r^2_{(m)}$ are the ordered squared residuals, and the value of h is set between $\frac{m}{2}$ and m. RANSAC [7] uses random sampling to calculate possible model parameters and pick the best among them which can fit most of the data. However, due to the combinatorial nature, all of these algorithms are impractical for solving high dimensional problems.

In contrast to the robust statistics approach, most SSR methods merely use the first assumption in the final reprojection step via thresholding, e.g., [8]. One exception is [9], [10], which developed a general thresholding function based iterative procedure and [9] was shown to be equivalent to a special class of M-estimators. For the second assumption, the SSR methods explicitly model the sparsity of outliers. Recently many works [11]–[14] address the outliers in the SSR framework, where x is also sparse (in the typically overcomplete dictionary A), and the corruptions may also admit a sparse representation in another general dictionary [15], [16]. Here we focus on the

Manuscript received February 28, 2017; revised September 20, 2017; accepted October 25, 2017. Date of publication November 8, 2017; date of current version December 26, 2017. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gonzalo Mateos. This work was supported by NSF Award IIS-1522125. *(Corresponding author: Jing Liu.)*

The authors are with the Department of Electrical and Computer Engineering, University of California, San Diego, CA 92093 USA (e-mail: jil292@ucsd.edu; pcosman@ucsd.edu; brao@ucsd.edu).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author. This includes a PDF file containing more experiments and additional proofs. The material is 2.18 MB in size.

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/TSP.2017.2771720

traditional linear regression problem, where x is general, A is over-determined and we have no freedom to design A. Under this setting, the existing SSR methods deal with outliers in two major ways, Projection Approach [17] and Joint Approach [18]. Let V denote the subspace spanned by the columns of A, and let $F \in \mathbb{R}^{(m-n)\times m}$ be a matrix whose rows form an orthobasis of $V^{\perp}$. Then we have $FA = 0$. The Projection Approach applies F to the measurements and from (2) we obtain

$$b \triangleq Fy = FAx + Fe + F\eta = Fe + F\eta. \tag{3}$$

The original problem is transferred to the recovery of a sparse vector e, given the under-determined measurement matrix F and noisy measurements b. Various SSR methods can be directly applied to solve this problem, such as BSRR [19] [20] which is based on Sparse Bayesian Learning (SBL) [21], [22], and Second-Order Cone Programming (SOCP) [8] which is based on ℓ_1 minimization [23]–[25]. Note that the traditional ℓ_1 estimator ($\arg\min_x \|y - Ax\|_1$) was shown to be equivalent to the SOCP case of no dense inlier noise [17]. The Joint Approach reformulates the original model into $y = [A\ I_{m\times m}]\left[\begin{smallmatrix}x\\e\end{smallmatrix}\right] + \eta$, where $[A\ I_{m\times m}]$ is under-determined and the lower part of $\left[\begin{smallmatrix}x\\e\end{smallmatrix}\right]$ is sparse. Many existing SSR methods can be extended to deal with this formulation via restricting the lower part of $\left[\begin{smallmatrix}x\\e\end{smallmatrix}\right]$ to be sparse, e.g., BPRR which is based on ℓ_1 minimization [20], $\ell_p (0 < p \leq 1)$ regularization which assumes a super-Gaussian prior for e to encourage sparsity [18], [26], Giannakis's algorithm for robust sensing [27] that utilizes a log-sum penalty function [28]–[31], Jin's empirical Bayesian inference-based algorithm which is extended from SBL [26], and GARD [1] which is based on Orthogonal Matching Pursuit (OMP) [32], [33]. An important finding in sparse recovery theory is that although finding the sparsest solution from under-determined linear equations is also of a combinatorial nature, some polynomial-time sparse recovery methods are guaranteed to find the sparsest solution under certain conditions on the sparsity of e and conditioning of matrix F [34], [35]. It was shown in [20] that BSRR outperforms LMedS and RANSAC.

The key to successful sparse recovery lies in identifying the support (nonzero entries), as one can simply add a reprojection step to estimate magnitude later. We propose a novel objective function and corresponding algorithm to help identify the *support of outliers*. The method is developed under the paradigm of the Joint Approach, but there is a fundamental difference with existing SSR methods. The existing methods often tackle the ℓ_0-'norm' of e implicitly (e.g., via OMP or SBL), or through the use of surrogate measures for the ℓ_0-'norm', such as the log-sum function or the ℓ_p-norm ($0 < p \leq 1$). Besides these methods, the hard thresholding based iterative method [9] shows its equivalence with a family (infinitely many) of nonconvex penalties for e (plus the ℓ_2-norm on the noise term), thus promoting the sparsity of e (the author noted that this method relies on a preliminary robust fit). In contrast to all these methods, we explicitly model and operate on the ℓ_0-'norm' of e, and the optimization procedure naturally utilizes the large observation error prior, and does not need a preliminary robust fit. Theoretical guarantees regarding exact recovery or error bounds are derived to support the efficacy of the method. The overall best performance in terms of the quality of recovery and lower complexity (over competing methods) further demonstrates the notable benefits of the proposed method.

TABLE I
ALGORITHM FOR ROBUST OUTLIER SUPPORT IDENTIFICATION (AROSI)

Input: $y, A, \alpha > 0$
Initialization: $k = 0, e^{(0)} = 0, \boldsymbol{S}_0 = \{1, \ldots, m\}$
While $J(x, e)$ not converged DO:
  Iteration $k+1$
  Step 1 (update x): $x^{(k+1)} = \arg\min_x \left\|y_{\boldsymbol{S}_k} - A_{\boldsymbol{S}_k} x\right\|_1$;
    If $\left\|y_{\boldsymbol{S}_k} - A_{\boldsymbol{S}_k} x^{(k+1)}\right\|_1 = \left\|y_{\boldsymbol{S}_k} - A_{\boldsymbol{S}_k} x^{(k)}\right\|_1$,
    further update $x^{(k+1)} = x^{(k)}$.
  Step 2 (update e and $\boldsymbol{S}$): $e_i^{(k+1)} = \begin{cases} 0, & \left|(y - Ax^{(k+1)})_i\right| \leq \alpha \\ (y - Ax^{(k+1)})_i, & otherwise \end{cases}$
    $\boldsymbol{S}_{k+1} := \{i : e_i^{(k+1)} = 0\}$
  $k := k+1$
End While
Output: solution $\tilde{x}$

The remainder of the paper is organized as follows: In Section II, we introduce the nonconvex objective function and the associated optimization procedure to help identify the support of outliers to be used in the reprojection step. Section III gives theoretical results regarding its convergence, exact recovery or recovery error. We empirically study the performance of the proposed method and compare with other state-of-the-art methods in Section IV. Conclusions are made in Section V.

Notation: Capital letters denote matrices, e.g., A, while lowercase letters denote vectors, e.g., e. The ith row of matrix A is denoted by ${a_i}^T$, while the ith element of vector e is denoted by e_i. The ℓ_0-'norm'[1] of e, i.e., $\|e\|_0$, counts the number of nonzero elements of e. Bold capital letters are reserved for sets, e.g., $\boldsymbol{S}$, where $\boldsymbol{S}^{\boldsymbol{c}}$ and $|\boldsymbol{S}|$ denote the complement and the cardinality of $\boldsymbol{S}$ respectively, and $\boldsymbol{S}_k$ denotes the set $\boldsymbol{S}$ obtained from the kth iteration. We use $A_{\boldsymbol{S}}$ to denote the $|\boldsymbol{S}| \times n$ submatrix of A containing the rows indexed by $\boldsymbol{S}$. Similarly, $e_{\boldsymbol{S}}$ denotes the subvector of e containing the entries indexed by $\boldsymbol{S}$. The indicator function is denoted as $I(\cdot)$.

II. ROBUST LINEAR REGRESSION VIA ℓ_0 REGULARIZATION

We propose minimizing the following objective function to help identify the support of outliers.

$$J(x, e) = \|y - Ax - e\|_1 + \alpha \|e\|_0 \tag{4}$$

In the second term, we directly use the ℓ_0-'norm' to enforce the sparsity in the outlier corruptions e, rather than relaxing it to the ℓ_p-norm ($0 < p \leq 1$).

We use the alternating minimization "like" approach to minimize the nonconvex objective function in (4). The detailed procedure is summarized in Table I, where $x^{(k+1)}$ and $e^{(k+1)}$ denote the updated x and e at the $(k+1)$st iteration. $\boldsymbol{S}_k$ is the complementary set of the support of $e^{(k)}$, which is the index set for "valid" entries of y that are estimated to be free of outliers in the kth iteration. Here the convergence of $J(x, e)$ means $J(x^{(k+1)}, e^{(k+1)}) = J(x^{(k)}, e^{(k)})$, and $\boldsymbol{S}_k = \boldsymbol{S}_{k-1}$ is a sufficient condition for convergence (see Appendix A).

At first glance, it seems more reasonable to use the ℓ_2-norm rather than the ℓ_1-norm in the first term of the objective function

[1] ℓ_0-'norm' is not a norm as it does not satisfy the axioms of a norm.

(4) and in Step 1, especially for Gaussian noise. Here we emphasize that the minimizer of the objective function (4) is not our final solution; it will be followed by a reprojection step described later. In Step 1 of each iteration, we only use our estimated "valid" outlier free entries/rows indicated by $\boldsymbol{S}$ to estimate x. However, we do not expect that all the outliers are identified by the previous iteration; it is very likely that some outliers have not been removed. So it is safer to use the ℓ_1-norm in Step 1, as the ℓ_1 estimator is more robust to outliers than OLS. In case there are multiple solutions[2] [36] for $\min_x \|y_{\boldsymbol{S}_k} - A_{\boldsymbol{S}_k} x\|_1$ and $x^{(k)}$ happens to be one of these solutions, we set $x^{(k+1)} = x^{(k)}$ to make the algorithm more stable.

At the beginning, we have no information about the positions of outliers except that they are sparse. So we simply initialize $e^{(0)} = 0$, and index set $\boldsymbol{S}_0 := \{i : e_i^{(0)} = 0\} = \{1, \ldots, m\}$. So in Step 1 of the first iteration, all the data will be used and it is equivalent to the ℓ_1 estimator, which has been justified by many authors (e.g., [17], [37]).

In Step 2, when x is fixed, define $r \overset{\Delta}{=} y - Ax$,

$$\min_e \left(\|y - Ax - e\|_1 + \alpha\|e\|_0\right) = \min_e \left(\|r - e\|_1 + \alpha\|e\|_0\right)$$

$$= \min_e \sum_{i=1}^{m} \left(|r_i - e_i| + \alpha I\left(e_i \neq 0\right)\right)$$

$$= \sum_{i=1}^{m} \min_{e_i} \left(|r_i - e_i| + \alpha I\left(e_i \neq 0\right)\right)$$

$$\hat{e}_i := \begin{cases} 0, & |r_i| \leq \alpha \\ r_i, & \text{otherwise} \end{cases} \in \arg\min_{e_i} \left(|r_i - e_i| + \alpha I\left(e_i \neq 0\right)\right), \tag{5}$$

and $\min_{e_i}(|r_i - e_i| + \alpha I(e_i \neq 0))$

$$= \min\left(|r_i|, \alpha\right) = \begin{cases} |r_i|, & |r_i| \leq \alpha \\ \alpha, & \text{otherwise} \end{cases}. \tag{6}$$

We can see from (5) that Step 2 directly promotes the sparsity of e via hard thresholding. Any entry of $|y - Ax|$ larger than α will be considered an outlier corrupted entry. In general, α should be set at least larger than the inlier noise level. Our analysis shows that under certain reasonable conditions on the model parameters, if α is greater than some certain threshold, we can guarantee that all the inliers are kept in *every* iteration. Conservatively, one may use a very large α, aiming to keep most of the inliers while safely removing some large outliers. Alternatively, one may use a small α (e.g., 4σ), aiming to get rid of more outliers, with the possibility one may also lose more inliers. If there is no prior knowledge of σ, it can be estimated from the residuals of the ℓ_1 estimation (which is also Step 1 of our first iteration) [4]: $\hat{\sigma} = \frac{1}{0.675} median(\ |r_i^{(1)}| \mid r_i^{(1)} \neq 0)$.

Reprojection Step for the Joint Approach: Our theoretical results in Section III show that AROSI can guarantee the exact support recovery of outliers. This motivates us to add a reprojection step in the end. The reprojection step [38] is widely used in sparse recovery methods; it often improves the estimation of the magnitudes of the nonzero entries. In the Projection Approach, as the original problem is transferred to the conventional sparse recovery problem form, it is straightforward to use reprojection (e.g., [8]). Here we present the reprojection step for the Joint Approach. Recall that the original model (2) is reformulated as $y = [A\ I_{m\times m}]\left[\begin{smallmatrix} x \\ e \end{smallmatrix}\right] + \eta$, where the lower part of $\left[\begin{smallmatrix} x \\ e \end{smallmatrix}\right]$ is sparse. With estimated $\tilde{x}$ or $\tilde{e}$ by some Joint Approach algorithm, the reprojection step is as follows:

1. Estimate the support $\boldsymbol{E}$ of e by thresholding $|\tilde{e}|$ or $|y - A\tilde{x}|$, e.g., $\hat{\boldsymbol{E}} := \{i : |\tilde{e}_i| > p\sigma\}$, where σ is the standard deviation of the inlier noise, and p is a scaling factor.
2. Regress y onto the selected columns of $[A\ I_{m\times m}]$, i.e., $[A\ (I_{\hat{\boldsymbol{E}}})^T]$ by least squares:

$$\hat{z} = \arg\min_z \left\| y - \left[A\ (I_{\hat{\boldsymbol{E}}})^T\right] z \right\|_2 \tag{7}$$

3. Finally, obtain $\hat{x} = \hat{z}_{\{1,\ldots,n\}}$, and $\hat{e}_{\hat{\boldsymbol{E}}} = \hat{z}_{\{n+1,\ldots,end\}}$, which is the estimated outlier corruption values corresponding to $\hat{\boldsymbol{E}}$.

In general, setting p is a tradeoff between false alarms and false negatives in identifying outliers, and so a relatively small p is recommended to have fewer false negatives. If it is known that the magnitudes of outliers are much larger than inlier noise (or if we are less concerned about the noise level outliers), a slightly larger p can be employed to decrease false alarms. When thresholding $|y - A\tilde{x}|$, since the inlier noise is present in this residual, the scaling factor p should be greater than 2. While when thresholding $|\tilde{e}|$, since e is already separated from the inlier noise in the model, a small p can be employed, e.g., [8] uses $p = 1$ in their Projection Approach.

A sufficient condition for $[A\ (I_{\hat{\boldsymbol{E}}})^T]$ to be full column rank is $|\hat{\boldsymbol{E}}| \leq \max(2\boldsymbol{m}(A) - 1,\ 0)$ (defined in Definition 1, guaranteed by Theorem 2). When $p \to \infty$, $|\hat{\boldsymbol{E}}| \to 0$. In case the generated $[A\ (I_{\hat{\boldsymbol{E}}})^T]$ is under-determined or not full column rank, we can always increase the scaling factor p to make $[A\ (I_{\hat{\boldsymbol{E}}})^T]$ full column rank, thus (7) has a unique solution.

The major difference with the reprojection step in the Projection Approach is the alternative way to estimate the support of e, i.e., via thresholding $|y - A\tilde{x}|$, if we have more confidence in estimated $\tilde{x}$ than $\tilde{e}$. In AROSI, we are more confident about the estimated $\tilde{x}$, as it is less sensitive to the parameter α than $\tilde{e}$. So, to estimate the support of e, we threshold $|y - A\tilde{x}|$, i.e., $\hat{\boldsymbol{E}} := \{i : |(y - A\tilde{x})_i| > p\sigma\}$.

Complexity: AROSI alternates between ℓ_1 estimation (Step 1) and entrywise thresholding (Step 2). So the main computational step (complexity) is ℓ_1 estimation in each iteration, which can be recast as Linear Programming. If AROSI converges in K iterations (usually a few iterations), the worst run time estimate will be K times the run time of the ℓ_1 estimator. In fact, the total run time is often less than that. This is not only because some entries are pruned in Step 1, but also because the result of the previous iteration is used as the initial point for the current iteration (*a.k.a.* warm-start). This is usually a good initial point and improves the speed of ℓ_1.

[2]In practice, when $A_{\boldsymbol{S}_k}$ is full column rank, this rarely happens, and we have not experienced this in our numerical experiments.

III. Theoretical Analysis

In this section, we analyze AROSI (without adding the reprojection step unless otherwise noted) and establish some theoretical guarantees which support its robustness and effectiveness. The theoretical results depend on the matrix A, the bounds for the inlier noise, and the sparsity of the outlier component. The exact conditions are included as part of the theorem statements. The main results include the following:

1) Exact recovery of x under *any* parameter setting (i.e., any $\alpha > 0$) in the presence of outliers only, i.e., absence of dense inlier noise (Theorem 3).
2) The estimation error is bounded in the noisy case (Theorem 6).
3) Exact support recovery of outliers in both no dense inlier noise case (Theorem 3) and noisy case (Theorem 6d)).
4) The ability to keep *all* the inliers and remove *significant* outliers in *every* iteration (Theorems 3 and 6–7, and Remark 2).
5) Even if the number of outliers is greater than the regression breakdown point of the ℓ_1 estimator, AROSI can still guarantee exact recovery (no dense inlier noise case, Remarks 1 and 2) or bounded estimation error (noisy case, Remark 5 and Theorem 7).

A. Convergence Property

Note that Step 1 of the algorithm deviates from the standard alternating minimization approach. Thus, the convergence of the algorithm is not assured based on the alternating minimization framework and needs to be established.

Theorem 1: AROSI converges in a finite number of iterations to a fixed point, which is a local optimum. Moreover, the objective function is strictly decreasing before convergence.

The proof of the theorem is in Appendix A.

B. Characterization of AROSI When Only Outliers Present

Here we discuss the case when there are only sparse outliers present and no dense inlier noise. Our model in (2) degenerates to $y = \mathrm{A}x + e$. The analysis benefits greatly from the analysis of the ℓ_1 estimator in [37], which is equivalent to the Step 1 of our first iteration. We further build and extend the work to understand AROSI, based on an important property stated in Lemma 1. We first introduce some definitions and properties regarding the leverage constants and their related quantity $\boldsymbol{m}(A)$ for matrix A that are important to the analysis.

Definition 1 (from [37]): Define $\boldsymbol{M} = \{1, \ldots, m\}$ as the index set of all the observations. Define for every $q \in \{1, \ldots, m\}$ the leverage constants c_q of A as $c_q(A) = \min_{\substack{E \subset \boldsymbol{M} \\ |\boldsymbol{E}| = q}} \min_{\substack{g \in \mathbb{R}^n \\ g \neq 0}} \frac{\sum_{i \in \boldsymbol{M} \setminus \boldsymbol{E}} |{a_i}^T g|}{\sum_{i \in \boldsymbol{M}} |{a_i}^T g|} = \min_{\substack{E \subset \boldsymbol{M} \\ |\boldsymbol{E}| = q}} \min_{g_2 = 1} \frac{\sum_{i \in \boldsymbol{M} \setminus \boldsymbol{E}} |{a_i}^T g|}{\sum_{i \in \boldsymbol{M}} |{a_i}^T g|}$ and $\boldsymbol{m}(A) = \max\{q \in \boldsymbol{M} | c_q(A) > \frac{1}{2}\}$.

Note that [39] provides an algorithm to compute $\boldsymbol{m}(A)$ for any given A. The complexity is $O(\binom{m}{n}(n^3 + m^2))$, which is prohibitive for large m and n, making the computation of $\boldsymbol{m}(A)$ limited to a small size matrix A.

Proposition 1 (from [37]): $c_0(A) = 1,\ c_m(A) = 0$, and for every $q \in \{1, \ldots, m\}$, $c_q(A) \leq c_{q-1}(A)$.

Proposition 2: If $\boldsymbol{m}(A) \geq q$, then we must have $c_q(A) > \frac{1}{2}$, and $m > 2q$.

The proof is omitted due to the space limit and can be found in the supplemental material.

In [37], it is shown that the regression breakdown point of the ℓ_1 estimator is $\boldsymbol{m}(A) + 1$. Since in the iterations of AROSI, it detects and removes 'outliers' and uses the remaining entries to do ℓ_1 estimation, two fundamental questions arise: When deleting some entries, 1) will the regression matrix become singular? 2) how does $\boldsymbol{m}(A)$ change (will it suddenly become 0)? The following Lemma 1 and Theorem 2 address these concerns.

Lemma 1: Let matrix A be full column rank and $\boldsymbol{m}(A) \geq q$. Then for any index set $\boldsymbol{T} \subset \boldsymbol{M},\ |\boldsymbol{T}| = t \leq q$, we have that $A_{\boldsymbol{T}^c}$ must be full column rank, $\boldsymbol{m}(A_{\boldsymbol{T}^c}) \geq q - 0.5t \geq q - t$, and $c_{q-t}(A_{\boldsymbol{T}^c}) \geq c_{q-0.5t}(A_{\boldsymbol{T}^c}) \geq c_q(A) > \frac{1}{2}$.

The proof of the lemma is in Appendix B.

Theorem 2: Let matrix A be full column rank and $\boldsymbol{m}(A) \geq q > 0$. Then for any index set $\boldsymbol{T} \subset \boldsymbol{M},\ |\boldsymbol{T}| = t \leq 2q - 1$, we have that $A_{\boldsymbol{T}^c}$ must be full column rank, $\boldsymbol{m}(A_{\boldsymbol{T}^c}) \geq q - 0.5t$, and $c_{q-0.5t}(A_{\boldsymbol{T}^c}) \geq c_q(A) > \frac{1}{2}$.

The proof utilizes the above Lemma and is omitted due to the space limit and can be found in the supplemental material.

The above theorem is significant because it characterizes the slowly decreasing property of $\boldsymbol{m}(A)$ w.r.t. m (the number of rows of A), which enables AROSI to go beyond ℓ_1 estimation and deal with more outliers, as we will show later.

Now we first introduce our main theorem of exact recovery when $\|e\|_0 \leq \boldsymbol{m}(A)$.

Theorem 3: AROSI running with *any* $\alpha > 0$ will find x exactly if $\|e\|_0 \leq \boldsymbol{m}(A)$. If additionally $\alpha < \min\{|e_i| :\ e_i \neq 0\}$, AROSI will find both x and e exactly.

Proof: Proved as a special case of Theorem 6 with $\eta = 0$. ■

Actually when $\|e\|_0 \leq \boldsymbol{m}(A)$, AROSI running with *any* $\alpha > 0$ recovers x exactly in every iteration, so it will converge in 2 iterations.

The above theorem shows the robustness of AROSI in two contexts: First, it succeeds in a wide range of parameter settings; Second, it is robust to the undetected outliers (even if α is set too large such that only a few outliers are detected). This robustness is a result of the slowly decreasing property of $\boldsymbol{m}(A)$ w.r.t. m. When only sparse outliers are present, we want the first term in the objective function (4) to be 0, as there is no dense inlier noise. We need to put infinitely large weight on the first term, or equivalently, set $\alpha \to 0^+$ in the second term. So $\alpha < \min\{|e_i| : e_i \neq 0\}$ will be satisfied. Then we can recover both x and e exactly under the given condition. When $\alpha \to 0^+$, minimizing the objective function (4) is equivalent to the following problem:

$$\min_{e,x} \|e\|_0 \ s.t.\ y = Ax + e, \tag{8}$$

which is the problem of interest when there is no dense noise, under the principle of fitting most of the data, and which would give exact recovery under mild conditions [20]. To minimize our objective function (4) with $\alpha \to 0^+$, AROSI starts with $\min_x \|y - Ax\|_1$, which is proven to give exactly the same solution as (8) under certain conditions [17], [20]. The above analysis gives a justification for our objective function (4) and AROSI.

Next, we deal with the case where $\|e\|_0 > \boldsymbol{m}(A)$.

Suppose $\|e\|_0 \leq \boldsymbol{m}(A)$ is not satisfied for the ℓ_1 estimator, which is also Step 1 in our first iteration. In the following steps we remove some entries that may contain both inliers and outliers. If the number of remaining outliers $\|e_{\boldsymbol{S}_k}\|_0 \leq \boldsymbol{m}(A_{\boldsymbol{S}_k})$, we can recover x exactly (see quoted Theorem in Appendix C).

The key question is whether it is possible that $\|e_{\boldsymbol{S}_k}\|_0 \leq \boldsymbol{m}(A_{\boldsymbol{S}_k})$, given that $\|e\|_0 > \boldsymbol{m}(A)$. Theorem 2 shows the slowly decreasing property of $\boldsymbol{m}(A)$, which makes it possible.

Remark 1: Suppose that $\|e\|_0 > \boldsymbol{m}(A) \geq q$, and that when AROSI converges at the $(k+1)$st iteration, $|\boldsymbol{S}_k^{\boldsymbol{c}}| = t$, i.e., we have removed t entries. Among these t entries, $\boldsymbol{p} \times t$ of them are outliers, so $\|e_{\boldsymbol{S}_k}\|_0 = \|e\|_0 - \boldsymbol{p} \times t$. When $t \leq 2q-1$, from Theorem 2, we know that $A_{\boldsymbol{S}_k}$ is full column rank and $\boldsymbol{m}(A_{\boldsymbol{S}_k}) \geq q - \lceil 0.5t \rceil$. So if $\|e\|_0 - \boldsymbol{p} \times t \leq q - \lceil 0.5t \rceil$, i.e., $\boldsymbol{p} \geq \frac{\|e\|_0 + \lceil 0.5t \rceil - q}{t}$, we can guarantee the exact recovery of x. When $t > 2q-1$, then a sufficient condition for exact recovery of x is that $A_{\boldsymbol{S}_k}$ has full column rank and $\boldsymbol{p} = \frac{\|e\|_0}{t}$, i.e., all the outliers are within the t removed entries.

The exact recovery test in Section IV-A demonstrates that there are cases where the ℓ_1 estimator fails (this must be the case $\|e\|_0 > \boldsymbol{m}(A)$ according to the quoted theorem in Appendix C) while AROSI gives exact recovery.

Remark 2: In case both large outliers and moderate outliers exist, as a special case of Theorem 7 with $\eta = 0$, we show that under certain conditions AROSI can recover x exactly even if there are up to $\lfloor 1.5 \times m(A) \rfloor$ outliers. More specifically, when $0 < \boldsymbol{m}(A) \leq \|e\|_0 \leq \boldsymbol{m}(A) + \lfloor \frac{t}{2} \rfloor$, where $1 \leq t \leq \boldsymbol{m}(A)$, define $\boldsymbol{G} := \{\text{indices of } \boldsymbol{m}(A) \text{ largest entries of } |e|\}, \boldsymbol{P} := \{\text{indices of } t \text{ largest entries of } |e|\}$. If $\min\{|e_i| : i \in \boldsymbol{P}\} > \frac{2\sum_{i \in \boldsymbol{E} \backslash \boldsymbol{G}} |e_i|}{c_{\boldsymbol{m}(A)}(A) - 0.5}$, then any α satisfying $\frac{\sum_{i \in \boldsymbol{E} \backslash \boldsymbol{G}} |e_i|}{c_{\boldsymbol{m}(A)}(A) - 0.5} < \alpha < \min\{|e_i| : i \in \boldsymbol{P}\} - \frac{\sum_{i \in \boldsymbol{E} \backslash \boldsymbol{G}} |e_i|}{c_{\boldsymbol{m}(A)}(A) - 0.5}$ guarantees the exact recovery of x from the second iteration, and it will converge in no more than three iterations. It is natural to think about this guarantee in comparison with the so called "masking effect" [40], where some extreme outliers (e.g., those indexed by $\boldsymbol{P}$), help hide another group of mild but perhaps more structured outliers (e.g., indexed by $\boldsymbol{E} \backslash \boldsymbol{G}$), which are usually more difficult to detect. AROSI effectively identifies and removes those extreme outliers, and more importantly, is resistant to the remaining unidentified outliers and recovers x exactly.

C. Both Dense Noise and Sparse Outliers Present

Now we deal with the more general case where both dense inlier noise and sparse outliers exist. In the first subsection, we establish the error bound for AROSI. Then we characterize the behaviors of AROSI in the second subsection.

1) Recovery Error Bound: We first quote a definition and theorem from [39] regarding the ℓ_1 estimation error bound, and present our Corollary 1, which establishes the bound for AROSI.

Definition 2 (from [39]): Given an arbitrary $q \in \{0, 1, \ldots, m\}$, we call a set $\boldsymbol{B}$ a *possibly extreme set* if there exists a set $\boldsymbol{L}$, $\boldsymbol{L} \supseteq \boldsymbol{B}$, $|\boldsymbol{L}| = m - q$, such that the following holds:

$$\sum_{i \in \boldsymbol{B} \cup \boldsymbol{L}^{\boldsymbol{c}}} \left| {a_i}^T v \right| \geq \sum_{i \in (\boldsymbol{L} \backslash \boldsymbol{B})} \left| {a_i}^T v \right| \tag{9}$$

where v is any of the singular vectors corresponding to the smallest singular value of the $|\boldsymbol{B}| \times n$ submatrix $A_{\boldsymbol{B}}$ of A: $\|A_{\boldsymbol{B}} v\|_2 = \sigma_{\min}(A_{\boldsymbol{B}}) \|v\|_2$. We define $\boldsymbol{Q}_q$ to be the set of all possibly extreme sets for a given q.

Theorem 4 (from [39]): Let $y = Ax + e + \eta$, $\boldsymbol{E} = supp(e)$, the ℓ_1 estimation error is bounded as follows:

$$\|x_{\ell_1} - x\|_2 \leq \left(\max_{\boldsymbol{B} \in \boldsymbol{Q}_{|\boldsymbol{E}|}} \frac{1}{\sigma_{\min}(A_{\boldsymbol{B}})} \right) \|\eta\|_2 \tag{10}$$

It can be proved that if $|\boldsymbol{E}| \leq \boldsymbol{m}(A)$, then $\forall \boldsymbol{B} \in \boldsymbol{Q}_{|\boldsymbol{E}|}$, $\sigma_{\min}(A_{\boldsymbol{B}}) > 0$.

Now we are ready to establish the error bound for AROSI.

Corollary 1: In the $(k+1)$st iteration of AROSI, define the index set $\boldsymbol{R} := \boldsymbol{E} \cap \boldsymbol{S}_k$. If $|\boldsymbol{R}| \leq \boldsymbol{m}(A_{\boldsymbol{S}_k})$, and $A_{\boldsymbol{S}_k}$ has full column rank, then the following holds for $x^{(k+1)}$:

$$\left\| x^{(k+1)} - x \right\|_2 \leq \left(\max_{\boldsymbol{B}' \in \boldsymbol{Q}'_{|\boldsymbol{R}|}} \frac{1}{\sigma_{\min}\left((A_{\boldsymbol{S}_k})_{\boldsymbol{B}'} \right)} \right) \|\eta_{\boldsymbol{S}_k}\|_2 \tag{11}$$

where $\sigma_{\min}((A_{\boldsymbol{S}_k})_{\boldsymbol{B}'}) > 0, \forall \boldsymbol{B}' \in \boldsymbol{Q}'_{|\boldsymbol{R}|}$. Here $\boldsymbol{Q}'_q$ follows the same definition in Definition 2, except that A is replaced by $A_{\boldsymbol{S}_k}$, and m is replaced by the number of rows of $A_{\boldsymbol{S}_k}$.

Proof: This is apparent from Theorem 4, as $x^{(k+1)}$ is the ℓ_1 estimate on the model $y_{\boldsymbol{S}_k} = A_{\boldsymbol{S}_k} x + e_{\boldsymbol{S}_k} + \eta_{\boldsymbol{S}_k}$, and $\boldsymbol{R} = \boldsymbol{E} \cap \boldsymbol{S}_k$ corresponds to $supp(e_{\boldsymbol{S}_k})$. ■

Remark 3: $\boldsymbol{R} := \boldsymbol{E} \cap \boldsymbol{S}_k$ is the index set of outliers that remained in $\boldsymbol{S}_k$. Note that Corollary 1 does not need the initial condition $|\boldsymbol{E}| \leq \boldsymbol{m}(A)$. It only needs the number of remaining outliers $|\boldsymbol{R}| \leq \boldsymbol{m}(A_{\boldsymbol{S}_k})$, which can be guaranteed by $|\boldsymbol{E}| \leq \boldsymbol{m}(A)$ and proper α (see Remark 4) for any $k \in \mathbb{Z}_{\geq 0}$. Even if $|\boldsymbol{E}| > \boldsymbol{m}(A)$, it is still possible that $|\boldsymbol{R}| \leq \boldsymbol{m}(A_{\boldsymbol{S}_k})$ for any $k \in \mathbb{Z}_{\geq 1}$, e.g., under the condition of Theorem 7 (details can be found in the proof).

Then a natural question of interest is whether the bound for AROSI is better than that of the ℓ_1 estimator. The following theorem provides a positive answer.

Theorem 5: Let $y = Ax + e + \eta$, $\boldsymbol{E} = supp(e)$, $|\boldsymbol{E}| = q \leq \boldsymbol{m}(A)$. In the $(k+1)$st iteration of AROSI, if $\boldsymbol{E}^{\boldsymbol{c}} \subseteq \boldsymbol{S}_k$, then $\|x^{(k+1)} - x\|_2$ is bounded as in (11), and the bound is smaller than or equal to the bound in (10).

The proof of the theorem is in Appendix E.

Theorem 5 is applicable for *any* iteration. The condition $\boldsymbol{E}^{\boldsymbol{c}} \subseteq \boldsymbol{S}_k$ required by Theorem 5 can be guaranteed with proper α, given $|\boldsymbol{E}| \leq m(A)$, as we will see in Theorem 6a), and it follows immediately that the bound for Theorem 6c) is smaller than or equal to the bound for ℓ_1 estimation error provided in Theorem 4.

2) Characterization of AROSI in Noisy Case: In this subsection, we first present Lemma 2, which describes the behavior of AROSI in any iteration and is an important step in deriving our main results in Theorems 6 and 7.

Lemma 2: Let $y = Ax + e + \eta$ and $\boldsymbol{E} = supp(e)$ satisfying $|\boldsymbol{E}| = q \leq \boldsymbol{m}(A)$. Denote $r_{\boldsymbol{S}_k}^{(k+1)} = y_{\boldsymbol{S}_k} - A_{\boldsymbol{S}_k} x^{(k+1)}$. If $\boldsymbol{S}_k \supseteq \boldsymbol{E}^{\boldsymbol{c}}$ for a particular k, then we must have $A_{\boldsymbol{S}_k}$ full column rank, $\boldsymbol{m}(A_{\boldsymbol{S}_k}) \geq q - |\boldsymbol{S}_k^{\boldsymbol{c}}|$, and $\|(e+\eta)_{\boldsymbol{S}_k} - r_{\boldsymbol{S}_k}^{(k+1)}\|_1 \leq \frac{\sqrt{m-q}\,\|\eta\|_2}{c_q(A) - 0.5}$. Also $\forall i \in \boldsymbol{E}^{\boldsymbol{c}}$, $|r_i^{(k+1)}| \leq \|\eta\|_\infty + \frac{\sqrt{m-q}\,\|\eta\|_2}{c_q(A) - 0.5}$.

The proof of the theorem is in Appendix F.

Now we are in position to present our main results in the noisy case. Theorem 6 shows that when $\|e\|_0 \leq \boldsymbol{m}(A)$, the estimation error of AROSI (with proper α) is bounded, and from Theorem 5

we know its bound is smaller than or equal to the ℓ_1 estimation error bound.

Theorem 6: Let $y = Ax + e + \eta$, $\boldsymbol{E} = supp(e)$ and $|\boldsymbol{E}| = q \leq \boldsymbol{m}(A)$. Define $C_1 = \frac{\sqrt{m-q}\,\|\eta\|_2}{c_q(A)-0.5}$, $C_2 = \max(C_1, \frac{2\sqrt{m-q}\,\|\eta\|_2 \sigma_{\max}(A_{\boldsymbol{E}})}{\sigma_{\min}(A_{\boldsymbol{E}^c})})$, $C_3 = \frac{\sigma_{\max}(A_{\boldsymbol{E}})}{\sigma_{\min}(A_{\boldsymbol{E}^c})} C_1$. For any $\alpha > \|\eta\|_\infty + C_1$, AROSI guarantees that:

a) All the inlier entries (indexed by $\boldsymbol{E}^c$) are kept in *every* iteration (i.e., $\boldsymbol{E}^c \subseteq \boldsymbol{S}_k$ for any $k \in \mathbb{Z}_{\geq 0}$);
b) Significant outlier entries indexed by $\boldsymbol{P} := \{i : |e_i| > \alpha + \|\eta\|_\infty + C_3\}$ are identified and removed in *every* iteration (i.e., $\boldsymbol{P} \subseteq \boldsymbol{S}_{k+1}^{\boldsymbol{c}}$ for any $k \in \mathbb{Z}_{\geq 0}$);
c) $\|x^{(k+1)} - x\|_2$ is bounded for any $k \in \mathbb{Z}_{\geq 0}$.

Moreover, if $\min\{|e_i| : \ e_i \neq 0\} > 2\|\eta\|_\infty + C_1 + C_2$, then any α satisfying $\|\eta\|_\infty + C_1 < \alpha < \min\{|e_i| : \ e_i \neq 0\} - \|\eta\|_\infty - C_2$ for AROSI guarantees that:

d) AROSI converges in 3 iterations, and the support of e is recovered exactly;
e) After the reprojection step (whose threshold is within the range $(\|\eta\|_\infty + C_1, \min\{|e_i| : \ e_i \neq 0\} - \|\eta\|_\infty - C_2)$), we have $\|\hat{x} - \mathrm{x}\|_2 \leq \|\eta_{\boldsymbol{E}^c}\|_2 / \sigma_{\min}(A_{\boldsymbol{E}^c})$.

The proof of the theorem is in Appendix G.

Remark 4: In Theorem 6e), $\hat{x}$ is equivalent to the least squares solution on all the inlier entries. The bound is tight and is better than the bound in (10) (details in the proof). Theorem 6 is an exciting result for the noisy case: If the large magnitude corruptions are sparse ($\|e\|_0 \leq m(A)$), with proper value of α (which depends on the inlier noise level, matrix A, and the sparsity of outliers, and does *not* depend on the magnitude of outliers), we can guarantee that all the inliers are kept in *every* iteration. At the same time, *all* the removed entries are guaranteed to be outliers. This shows another aspect of AROSI robustness: under certain conditions there are no false alarms when identifying and removing outliers during iterations. Purely removing some outliers often leads to better signal estimation in our Step 1 ($x^{(k+1)} = \arg\min_x \|y_{\boldsymbol{S}_k} - A_{\boldsymbol{S}_k} x\|_1$) than ℓ_1 estimation, especially as we can also guarantee (by Lemma 2) that $A_{\boldsymbol{S}_k}$ is full column rank and the number of remaining outliers $(|\boldsymbol{E}| - |\boldsymbol{S}_k^{\boldsymbol{c}}|) \leq \boldsymbol{m}(A_{\boldsymbol{S}_k})$ for any $k \in \mathbb{Z}_{\geq 1}$. In addition, we can also guarantee that the *significant* outliers, which are usually the most troublesome ones, are identified and removed in *every* iteration. Further, if the magnitudes of the corruptions are all large enough, we can even guarantee all the outliers are removed in *every* iteration. Finally, note that we showed that the estimation error is bounded in *every* iteration.

The following Remark 5 and Theorem 7 demonstrate that even if $\|e\|_0 > \boldsymbol{m}(A)$ (recall that $\boldsymbol{m}(A) + 1$ is the regression breakdown point of the ℓ_1 estimator [37]), AROSI can still provide a bounded estimation error.

Remark 5: When $\|e\|_0 > \boldsymbol{m}(A)$, we have provided a sufficient requirement in Remark 1 to satisfy the condition of Corollary 1, thus guaranteeing that the estimation error of x by AROSI is bounded in the noisy case.

In the following theorem, we establish conditions under which AROSI is guaranteed to handle more than $\boldsymbol{m}(A)$ outliers.

Theorem 7: Suppose $y = Ax + e + \eta$, $\boldsymbol{E} = supp(e)$, $0 < \boldsymbol{m}(A) \leq |\boldsymbol{E}| = q \leq \boldsymbol{m}(A) + \lfloor \frac{t}{2} \rfloor$, where $1 \leq t \leq \boldsymbol{m}(A)$. Define $\boldsymbol{G} := \{\text{indices of } \boldsymbol{m}(A) \text{ largest entries of } |e|\}$, $\boldsymbol{P} =: \{\text{indices of } t \text{ largest entries of } |e|\}$, $q_1 = \boldsymbol{m}(A)$, $q_2 = \boldsymbol{m}(A_{\boldsymbol{P}^c})$, $w_1 = \max(\frac{\sqrt{m-q_1}\,\|\eta\|_2 + \sum_{i \in \boldsymbol{E}\setminus\boldsymbol{G}} |e_i|}{c_{q_1}(A)-0.5}, \frac{\sqrt{m-q}\,\|\eta\|_2}{c_{q_2}(A_{\boldsymbol{P}^c})-0.5})$, $w_2 = \max(\frac{\sqrt{m-q_1}\,\|\eta\|_2 + \sum_{i \in \boldsymbol{E}\setminus\boldsymbol{G}} |e_i|}{c_{q_1}(A)-0.5}, \frac{\sigma_{\max}(A_{\boldsymbol{P}})\sqrt{m-q}\,\|\eta\|_2}{\sigma_{\min}(A_{\boldsymbol{E}^c}) \times (c_{q_2}(A_{\boldsymbol{P}^c})-0.5)})$. If $\min\{|e_i| : \ i \in \boldsymbol{P}\} > 2\|\eta\|_\infty + w_1 + w_2$, then any α satisfying $\|\eta\|_\infty + w_1 < \alpha < \min\{|e_i| : i \in \boldsymbol{P}\} - \|\eta\|_\infty - w_2$ for AROSI guarantees that:

a) All the inlier entries (indexed by $\boldsymbol{E}^c$) are kept in *every* iteration (i.e., $\boldsymbol{E}^c \subseteq \boldsymbol{S}_k$ for any $k \in \mathbb{Z}_{\geq 0}$);
b) Significant outlier entries indexed by $\boldsymbol{P}$ are identified and removed in *every* iteration (i.e., $\boldsymbol{P} \subseteq \boldsymbol{S}_{k+1}^{\boldsymbol{c}}$ for any $k \in \mathbb{Z}_{\geq 0}$);
c) $\|x^{(k+1)} - x\|_2$ is bounded for any $k \in \mathbb{Z}_{\geq 1}$.

The proof of the theorem is in Appendix H.

As our first iteration is equivalent to ℓ_1 estimation, we can not guarantee the estimation error is bounded when there are more than $\boldsymbol{m}(A)$ outliers. However, we can guarantee it is bounded in the following iterations.

The basic idea underlying behind Theorem 7 is based on the following intuition: when there are $\|e\|_0 > \boldsymbol{m}(A)$ outliers, if the smallest $\|e\|_0 - \boldsymbol{m}(A)$ of them are moderate, we can treat them as very noisy inliers, so the number of outliers reduces to $\boldsymbol{m}(A)$. Then according to Theorem 6, we can use a large α to safely remove the very large outliers.

IV. Empirical Studies

Although we provided some theoretical guarantees/bounds for AROSI, they often involve $c_q(A)$, which itself is hard to compute. In this section, we empirically study the performance of AROSI (including the reprojection step unless noted) as well as the following state-of-the-art methods, where the complexity analysis is presented for $m > n$.

1. ℓ_1 estimator [17]: $x_{\ell_1} = \arg\min_x \|y - Ax\|_1$. We also add a reprojection step for comparison.
   The complexity in practice is $\mathrm{O}(m^3)$ [41].
2. Second-Order Cone Programming (SOCP) [8], which is a direct application (via the Projection Approach) of ℓ_1 minimization sparse recovery [23]–[25] to model (3). There is a reprojection step in the end. The complexity of this method is $\mathrm{O}(m^3)$ [42].
3. Ideal solution where we know e exactly:
$$x_{Ideal} = \arg\min_x \|y - e - Ax\|_2$$
4. Oracle solution [8] where we know the support of e exactly: $x_{Oracle} = \arg\min_x \|y_{\boldsymbol{S}} - A_{\boldsymbol{S}} x\|_2$.
   where $\boldsymbol{S} := \{i : e_i = 0\}$ is the index set of all the inliers.
5. Bayesian Sparse Robust Regression (BSRR) [19], which is a direct application (via the Projection Approach) of Sparse Bayesian Learning to model (3). The complexity of each iteration is $\mathrm{O}(m^3)$. We add a reprojection step.
6. Generalized M-estimators with Bisquare weighting function [5], [43]–[46]. It is solved via Iteratively Reweighted Least Squares (IRLS), and the complexity of each iteration is $\mathrm{O}(mn^2)$. We set its tuning constant $c = 3$ to generate better results than the default value.
7. ℓ_1 regularization algorithm [18], [26], which solves $\min_{x,e} \|y - Ax - e\|_2^2 + \lambda \|e\|_1$, where the parameter λ

is set as $\frac{\sigma\sqrt{2\log(m)}}{3}$ according to [26]. It can be solved using the approach described in [1], where the complexity is $\mathrm{O}(m^3)$ per iteration [1]. We add a reprojection step in the end.

8. Greedy Algorithm for Robust Denoising (GARD) [1], which aims to minimize the number of outliers via OMP by restricting the selection over columns of $I_{m\times m}$:

$$\min_{x,\,e}\|e\|_0 \; s.t. \; \left\| y - [\mathrm{A}\; I_{m\times m}] \begin{bmatrix} x \\ e \end{bmatrix} \right\|_2^2 \leq \varepsilon^2$$

The total complexity is $\mathrm{O}(\frac{K^3}{2} + (m\ +\ 3K)n^2 + 3Kmn)$, where K is the total number of iterations. We add a reprojection step in the end.

9. Thresholding-based Iterative Procedure for Outlier Detection (Θ-IPOD) [9], which iterates between least squares regression and hard thresholding. We initialize it by ℓ_1 estimation, and set the threshold to 5σ. The algorithm's pre-computation costs $\mathrm{O}(mn^2)$, and each iteration costs $\mathrm{O}(mn)$. We add a reprojection step in the end.

For AROSI, we fix α as 5σ throughout the experiments unless otherwise noted. In the reprojection step of BSRR, SOCP, AROSI, Θ-IPOD, GARD and the ℓ_1 regularization method, the threshold is tuned individually from $\{p\sigma : p = 1, 2, 3, 4, 5\}$ for each method.

For our experimental setup, below are the general steps:

1. Choose a fraction ρ of grossly corrupted entries and define the number of corrupted entries as $k = \mathrm{round}(\rho \cdot m)$;
2. Generate an m by n standard Gaussian matrix A.
3. Generate $x \in \mathbb{R}^n$ with i.i.d. $\mathcal{N}(0, {\sigma_x}^2)$ entries. Compute Ax.
4. Select k locations uniformly at random and add corruptions to these locations.
5. Generate the vector $\eta = (\eta_1, \ldots, \eta_m)$ of smaller errors with η_i i.i.d. $\mathcal{N}(0, \sigma^2)$, and add η to the outcome of the previous step. Obtain y.
6. Estimate x using different methods.

We first set $m = 512$, $\sigma_x = 1$, and $\sigma = median(|Ax|)/16$ as in [8]. The corruption values are drawn from $0.5 \times \mathcal{N}(12\sigma, (4\sigma)^2) + 0.5 \times \mathcal{N}(-12\sigma, (4\sigma)^2)$. For each $n \in \{256, 128,\ 64\}$, we repeat Step 2 - Step 6 fifty times for each corruption rate. We denote this setting as experimental setup A.

Next, we use the experimental setup in [1] (denoted as experimental setup B), where $m = 600$, $\sigma_x = 5$, $\sigma = 1$, and the rows of matrix A are obtained by uniformly sampling an n-dimensional hypercube centered around the origin; i.e., $A_{ij} \sim \mathrm{U}(-1, 1)$. The corruption values are drawn from $\{-25, 25\}$ with equal chance. For each $n \in \{170, 100, 50\}$, we repeat Step 2 - Step 6 fifty times for each corruption rate.

For evaluation, each estimate is compared with ground truth x. We measure its Relative ℓ_2-Error [47]: $\|\hat{x} - x\|_2/\|x\|_2$. We also compute the distance between the supports of e and $\hat{e}$. Denoting the two supports as $\boldsymbol{E}$ and $\hat{\boldsymbol{E}}$, $\hat{\boldsymbol{E}}$ is estimated by thresholding $|\hat{e}|$ or $|y - A\hat{x}|$ with $p\sigma$, where p is tuned individually for each method. The distance is defined as in [47]: $dist(\hat{\boldsymbol{E}}, \boldsymbol{E}) = \frac{\max\{|\hat{\boldsymbol{E}}|,|\boldsymbol{E}|\}-|\hat{\boldsymbol{E}}\cap\boldsymbol{E}|}{\max\{|\hat{\boldsymbol{E}}|,|\boldsymbol{E}|\}}$. We denote the average of $dist(\hat{\boldsymbol{E}}, \boldsymbol{E})$ over Monte Carlo runs as the Probability of Error in Support (PES) [47].

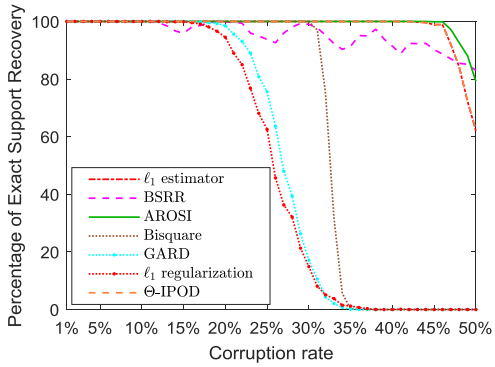

Fig. 1. Percentage of exact support recovery vs. corruption rate.

A. Exact Recovery Test

In this subsection, we empirically verify the exact recovery performance of AROSI when only sparse outliers are present, i.e., $y = Ax + e$. Recall that in the reprojection step, exact recovery of the support of e will suffice for the exact recovery of both x and e, as long as $[A\ (I_{\boldsymbol{E}})^T]$ is full column rank.

We use the same experimental setup as the Support Recovery Test in [1]. This is under experimental setup B with $n = 100$, except that there is no dense inlier noise. Fig. 1 shows the percentage of exact support recovery for each corruption rate (over 1000 trials) for each method. The support of BSRR, ℓ_1 estimator, Bisquare, AROSI, Θ-IPOD and the ℓ_1 regularization method (all without reprojection) is estimated by thresholding $|\hat{e}|$ or $|y - A\hat{x}|$ with a small numerical constant 1×10^{-4}.

Over 1000 trials, Bisquare keeps fully exact support recovery up to 29% corruption rate. For BSRR, ℓ_1 regularization method, GARD, ℓ_1 estimator, Θ-IPOD, and AROSI, it is up to 11%, 12%, 16%, 42%, 42%, and 44%, respectively. Θ-IPOD performs similarly to its initialization (ℓ_1 estimation), while AROSI demonstrates an improvement over ℓ_1 estimation.

When the corruption rates are 43% and 44%, there are cases where AROSI has exact support recovery while the ℓ_1 estimator does not. From the quoted theorem in Appendix C, we know it must be the case that $\|e\|_0 > \boldsymbol{m}(A)$. Since we also use the same ℓ_1 estimation in our first iteration, we do not have a perfect initialization. However, at the end of the iterations, we are able to identify and remove some outliers through the index set $\boldsymbol{S}_k$. The number of remaining outliers is very likely less than $\boldsymbol{m}(A_{\boldsymbol{S}_k})$, thus we get the exact solution. This shows the advantage of AROSI over the ℓ_1 estimator.

B. Both Dense Noise and Sparse Outliers Present

In this subsection, we test and compare the performance of each method in the noisy case under experimental setup A. Fig. 2 shows the average Relative ℓ_2-Error and the PES from 50 samples vs. corruption rate.

In general, AROSI has similar performance to BSRR and outperforms other methods. We can see that the reprojection step alone does help improve the performance of the ℓ_1 estimator. However, AROSI performs even better, which verifies that the advantage of AROSI over the ℓ_1 estimator is non-trivial. We can also see that, under the same corruption rate, increasing the signal dimension n makes the recovery harder for all methods, as the number of unknowns gets larger.

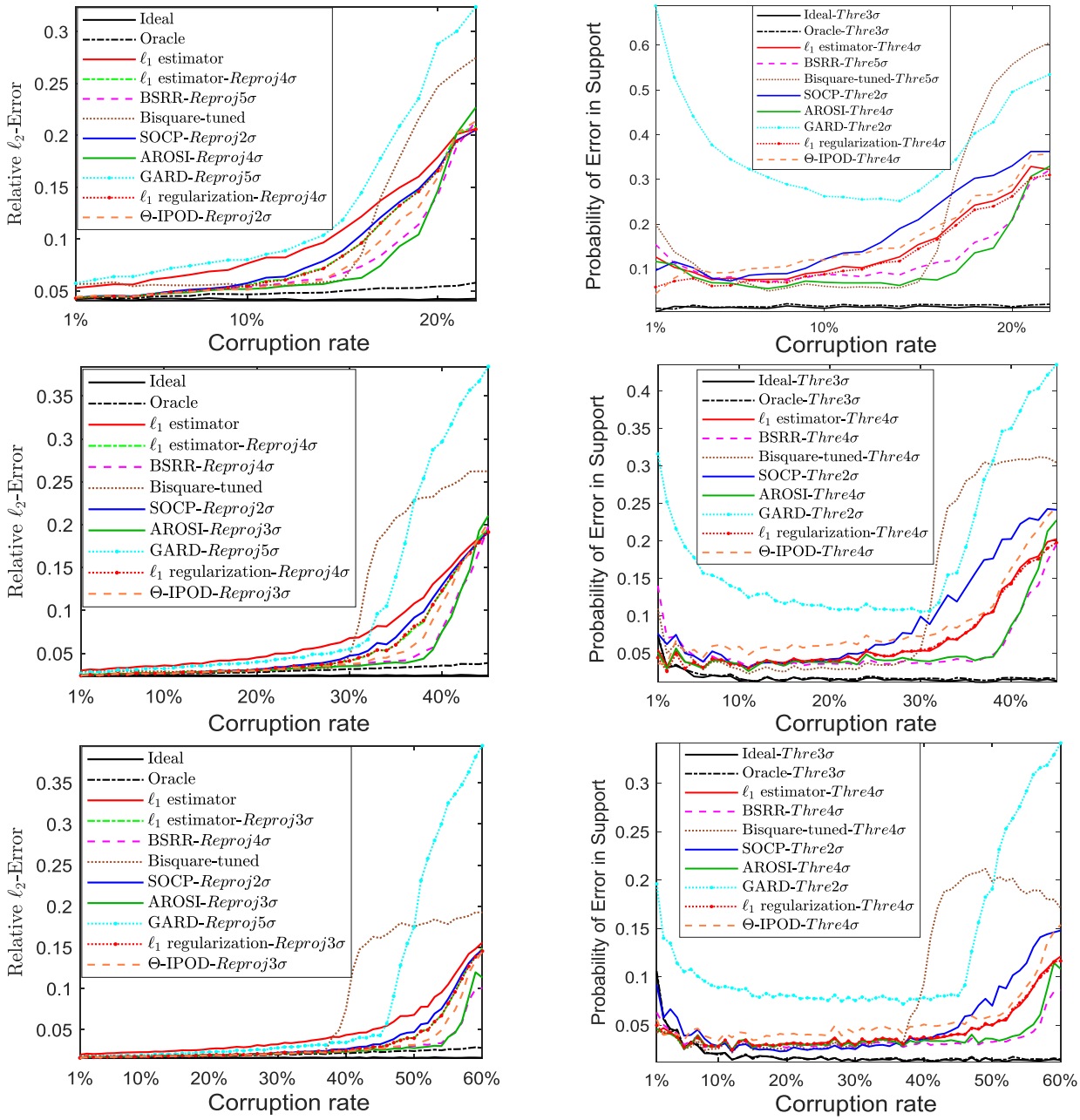

Fig. 2. Average relative ℓ_2-error (left) and PES (right) vs. corruption rate with different n (upper: 256; middle: 128; bottom: 64).

We have also tested on several non-Gaussian regression matrices, which can be found in the supplemental material. The relative performance of each method is almost unchanged, except some degradation of the relative performance of the ℓ_1 regularization method under some regression matrices.

C. Phase Transition Curves

We measure the Phase Transition Curves of each method under experimental setup A. For each dimension of x and each method, we test each outlier fraction and find the maximum fraction where the probability of successful recovery (Relative ℓ_2-Error less than 1.3 $\times$ that of Oracle) remains greater than 0.5. Fig. 3 shows the Phase Transition Curves of each method. AROSI outperforms all the other methods.

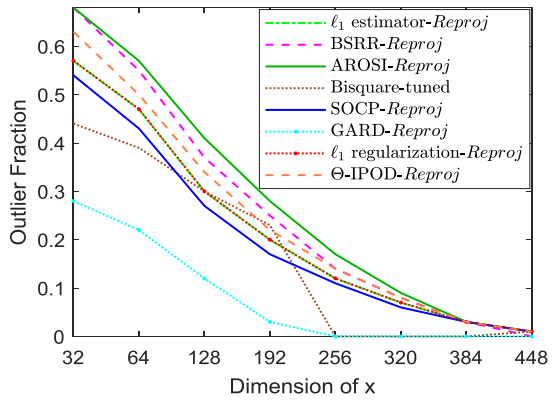

Fig. 3. Phase transition curves.

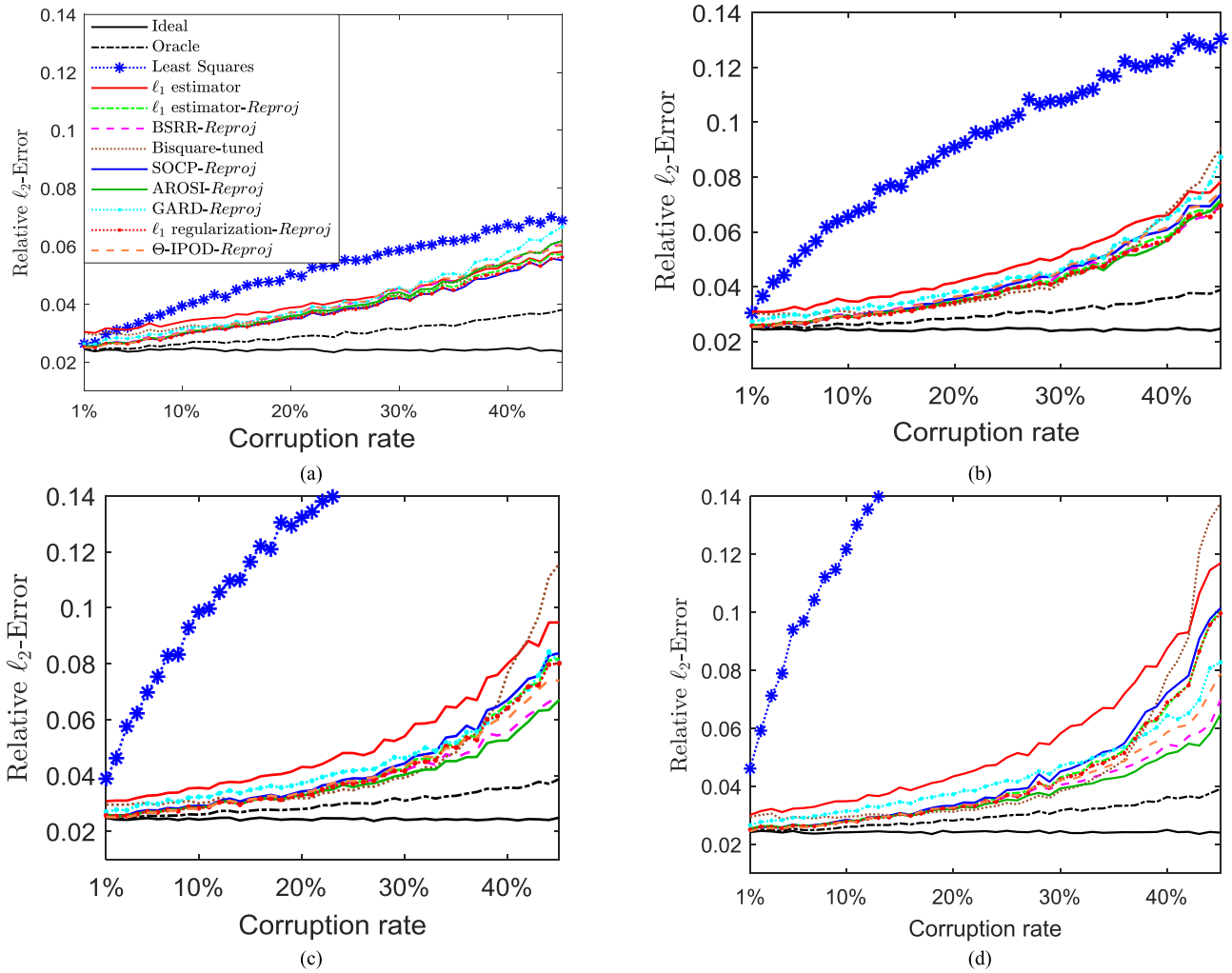

Fig. 4. Average relative ℓ_2-error vs. corruption rate for different scales ($\kappa\sigma$) of Gaussian corruptions: a) $\kappa = 4$; b) $\kappa = 8$; c) $\kappa = 12$; d) $\kappa = 16$.

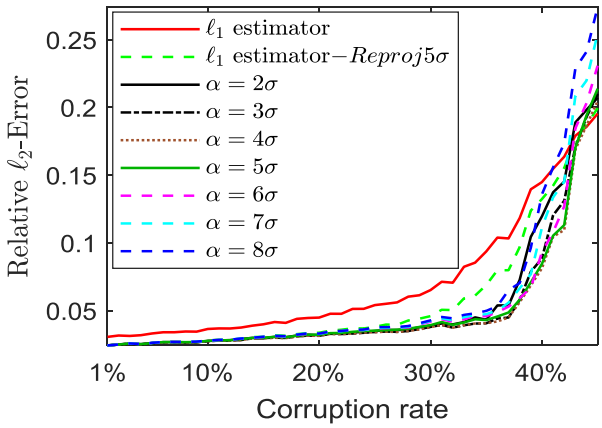

Fig. 5. Average relative ℓ_2-error vs. corruption rate for ℓ_1 estimator and AROSI with different α. In the reprojection step of AROSI, $p = 5$.

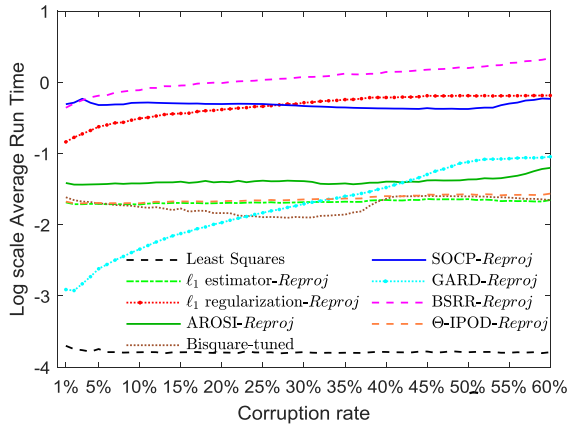

Fig. 6. Log scale average run time vs. corruption rate.

D. Different Magnitude of Corruptions

In this subsection, we use experimental setup A but with corruption values drawn from $\mathcal{N}(0, (\kappa\sigma)^2)$ instead (recall that $\sigma = median(|Ax|)/16$). We gradually increase the magnitude of corruptions (by increasing κ) to see how each method behaves. Fig. 4 shows the average Relative ℓ_2-Error on 50 samples vs. corruption rate for different scales ($\kappa\sigma$) of corruptions. We can see that, when the magnitude of corruptions is small (e.g., $\kappa = 4$), even the least squares works well, and all the robust lin-

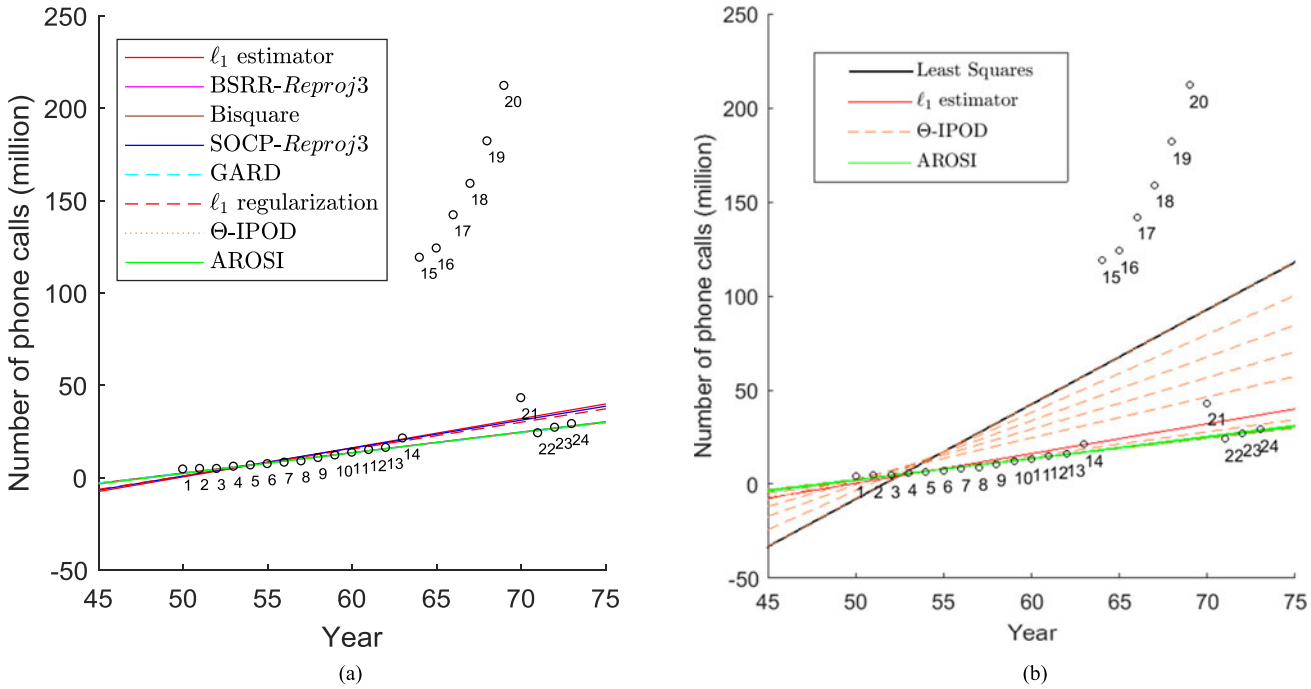

Fig. 7. Number of phone calls (million) in the years 1950-1973 fitted by: (a) all methods (with tuned parameter). (b) Least Squares, ℓ_1 estimator, AROSI, and Θ-IPOD (the parameters of AROSI and Θ-IPOD both vary from 3 to 180).

ear regression methods have very nominal differences and are slightly better than the least squares (we note that the performance of AROSI can be slightly improved if we set α larger). As κ is increased further, the robust linear regression methods begin to show their benefits. We note that when κ increases from 4 to 16, the performances of the ℓ_1 estimator (with or without the reprojection step), Bisquare, SOCP, and the ℓ_1 regularization method degrade. In contrast, BSRR and AROSI are quite resistant to the larger magnitudes of corruptions.

E. Sensitivity to Parameter α of AROSI

SOCP, GARD, ℓ_1 regularization method, Θ-IPOD, AROSI, and the initialization of BSRR all need the knowledge of inlier noise level. In the previous experiments, we assume we know the standard deviation σ of the inlier noise, and set $\alpha = 5\sigma$ for AROSI. However, in practice, the estimated $\hat{\sigma}$ may be slightly greater or less than the true σ, which is equivalent to setting α slightly greater or less than 5σ. We test AROSI with α varying from 2σ to 8σ. In the reprojection step of AROSI and the ℓ_1 estimator, we fix $p = 5$.

Fig. 5 shows the average Relative ℓ_2-Error on 50 samples vs. corruption rate for ℓ_1 estimation (with or without the reprojection step) and AROSI with different α, under experimental setup A with $n = 128$.

When the corruption rate is moderate (e.g., $\leq 35\%$ when $n = 128$), we have two observations:

- AROSI often performs better than the ℓ_1 estimator even with different α (from 2σ to 8σ).
- With α ranging from 2σ to 8σ, AROSI has similar performance, which indicates the method is not very sensitive to small variations of α.

F. Run Time

In this subsection, we compare run times under experimental setup A. Fig. 6 shows the Average Run Time (seconds) on 100 samples vs. corruption rate with $n = 64$. We can see that AROSI is an order of magnitude faster than BSRR.

G. Real Data

Finally, we compare the performance of each method on a real dataset, the Belgian Phone data, from the Belgian Statistical Survey (published by the Ministry of Economy). It contains large outliers as well as moderate outliers, and the swamping/masking effects could arise. There are 24 measurements. The response is the number of international phone calls (in millions), and the predictor is the year. It is known afterwards that observations 15-20 are large outliers and observations 14 and 21 are moderate outliers. For such a small size regression matrix A, using the algorithm provided in [39], we easily get $m(A) = 5$, which is unfortunately smaller than the number of the outliers.

To see the difference between each method more clearly, we do not perform the reprojection step, except for the Projection Approach methods, i.e., for BSRR and SOCP, the threshold is tuned to obtain the best result. The results are plotted in Fig. 7(a). Most methods have very similar results on this data, and fit the inliers very well, except for the ℓ_1 estimator, SOCP, and the ℓ_1 regularization method. We can see that these three methods are biased by outliers, and the residual of the outlier observation 14 is very small (it is perfectly masked by large outliers), even smaller than many inlier observations, e.g., observations 1, 2, 22-24. So, even if we add a reprojection step for the ℓ_1 estimator and the ℓ_1 regularization method, the outlier observation 14 is hard to get rid of.

TABLE II
BEHAVIOR OF AROSI UNDER DIFFERENT α

	1st iter.		2nd iter.		3rd iter.	Estimated $\hat{x}$ (no reprojection)	
α	$\boldsymbol{S}_1^{\boldsymbol{c}}$	$m(A_{S_1})$	$\boldsymbol{S}_2^{\boldsymbol{c}}$	$m(A_{S_2})$	$\boldsymbol{S}_3^{\boldsymbol{c}}$	$\hat{x}(1)$	$\hat{x}(2)$
3	1,13, 15-24	2	14-21	2	14-21	1.125	-54.000
4	15-24	3	14-21	2	14-21	1.125	-54.000
5-7	15-24	3	15-21	2	15-21	1.115	-53.280
8	15-23	3	15-21	2	15-21	1.115	-53.280
9	15-22	2	15-21	2	15-21	1.115	-53.280
10	15-21	2	15-21	2	CNVG	1.115	-53.280
11-18	15-20	3	15-21	2	15-21	1.115	-53.280
19-96	15-20	3	15-20	3	CNVG	1.115	-53.280
97-99	16-20	3	15-20	3	15-20	1.115	-53.280
100	17-20	4	15-20	3	15-20	1.115	-53.280
101-104	17-20	4	16-20	3	16-20	1.133	-54.233
105-116	17-20	4	17-20	4	CNVG	1.151	-55.309
117-121	18-20	4	17-20	4	17-20	1.152	-55.349
122-131	18-20	4	18-20	4	CNVG	1.173	-56.609
132-137	19-20	5	18-20	4	18-20	1.173	-56.609
138-153	19-20	5	19-20	5	CNVG	1.173	-56.609
154-158	20	5	19-20	5	19-20	1.173	-56.609
159-180	20	5	20	5	CNVG	1.173	-56.609

Though AROSI is equivalent to the ℓ_1 estimator at the beginning, it successfully eliminates the effect of outliers with a wide range of parameter α. Fig. 7(b) shows the results of the ℓ_1 estimator and AROSI with integer α ranging from 3 to 180, as well as Θ-IPOD with the same threshold ranging from 3 to 180, all without the reprojection step. We can see that even with very different α, AROSI still fits the inliers very well, and is better than the ℓ_1 estimator. While Θ-IPOD (initialized by the ℓ_1 estimator) is sometimes severely biased by the outliers; it only works better than the ℓ_1 estimator when its threshold is set properly such that the outlier observations 15-20 are *all* identified at the beginning. When the threshold of Θ-IPOD is set larger than 137 (the outlier observations 19 and 20 can still be detected at the beginning), it will finally converge to the least squares solution. This demonstrates one important robustness property of AROSI over Θ-IPOD: the tolerance to unidentified outliers.

Table II documents the details of AROSI regarding its estimated outlier support set $\boldsymbol{S_k^c}$ and the corresponding $\boldsymbol{m}(A_{\boldsymbol{S}_k})$ at the end of each iteration k under different α, as well as the estimated $\hat{x}$ upon convergence (without the reprojection step). AROSI converges in either 2 or 3 iterations (note that $\boldsymbol{S}_k^{\boldsymbol{c}} = \boldsymbol{S}_{k-1}^{\boldsymbol{c}}$ implies convergence). The least squares gives the solution $\hat{x}_{LS} = (5.041,\ -260.059)$, which is severely biased by the outliers. The ℓ_1 estimator gives the solution $\hat{x}_{\ell 1} = (1.580,\ -78.522)$. As $\boldsymbol{m}(A) = 5$, which is smaller than the number of outliers (there are 6 large outliers and 2 moderate outliers), the performance of the ℓ_1 estimator is not guaranteed. However, we can see that with α ranging from 3 to 121, AROSI successfully identifies some outliers, and more importantly, the number of remaining outliers contained in $\boldsymbol{S}_{k-1}$ is less than the corresponding $\boldsymbol{m}(A_{\boldsymbol{S}_{k-1}})$, which guarantees the performance of AROSI in the last iteration k.

V. CONCLUSION

We proposed a novel robust linear regression method AROSI based on ℓ_0 regularization. It assumes that outliers are sparse and result in large observation errors. Several properties of AROSI such as convergence, exact recovery or recovery error are derived.

Through extensive simulation studies and comparisons with state-of-the-art methods, we have shown that AROSI achieves the overall best quality of recovery (in terms of exact recovery, recovery error, outlier support recovery), and it runs much faster than the competing methods like BSRR. Comparisons on a real dataset further demonstrate the robustness of AROSI and its advantage over the ℓ_1 estimator, Θ-IPOD, and the ℓ_1 regularization method.

APPENDICES

A. Proof of Theorem 1

The proof is divided into the following three parts: a) monotonic decrease in the objective function prior to convergence, b) convergence in a finite number of steps, and c) local optimality of the cluster point.

a) Strictly decreasing behavior of $J(x^{(k)}, e^{(k)})$ before convergence

As defined earlier, $\boldsymbol{S}_k := \{i : e_i^{(k)} = 0\}$. We now denote its complementary set $\boldsymbol{S}_k^{\boldsymbol{c}} := \{i : e_i^{(k)} \neq 0\}$. Define $J_{\boldsymbol{S}_k}(x, e) \triangleq \sum_{i \in \boldsymbol{S}_k} (|(y - Ax - e)_i| + \alpha I(e_i \neq 0))$ and $J_{\boldsymbol{S}_k^{\boldsymbol{c}}}(x, e) \triangleq \sum_{i \in \boldsymbol{S}_k^{\boldsymbol{c}}} (|(y - Ax - e)_i| + \alpha I(e_i \neq 0))$. So we have $J(x, e) = J_{\boldsymbol{S}_k}(x, e) + J_{\boldsymbol{S}_k^{\boldsymbol{c}}}(x, e)$.

For any $i \in \boldsymbol{S}_k$, $e_i^{(k)} = 0$. Hence

$$J_{\boldsymbol{S}_k}(x, e^{(k)}) = \sum_{i \in \boldsymbol{S}_k} \left(\left| \left(y - Ax - e^{(k)} \right)_i \right| + \alpha I \left(e_i^{(k)} \neq 0 \right) \right)$$
$$= \sum_{i \in \boldsymbol{S}_k} |(y - Ax)_i| = \| y_{\boldsymbol{S}_k} - A_{\boldsymbol{S}_k} x \|_1 \qquad (12)$$

In Step 1, since $x^{(k+1)} \in \arg\min_x \| y_{\boldsymbol{S}_k} - A_{\boldsymbol{S}_k} x \|_1$, we have

$$J_{\boldsymbol{S}_k} \left(x^{(k+1)}, e^{(k)} \right) \leq J_{\boldsymbol{S}_k} \left(x^{(k)}, e^{(k)} \right), \qquad (13)$$

where the equality holds if and only if

$$\left\| y_{\boldsymbol{S}_k} - A_{\boldsymbol{S}_k} x^{(k+1)} \right\|_1 = \left\| y_{\boldsymbol{S}_k} - A_{\boldsymbol{S}_k} x^{(k)} \right\|_1 \qquad (14)$$

In Step 2, $J_{\boldsymbol{S}_k}(x^{(k+1)}, e) = \sum_{i \in \boldsymbol{S}_k} (|(y - Ax^{(k+1)})_i - e_i| + \alpha I(e_i \neq 0))$, and from (5) we know that $e_i^{(k+1)} \in \arg\min_{e_i} (|(y - Ax^{(k+1)})_i - e_i| + \alpha I(e_i \neq 0))$. Thus

$$J_{\boldsymbol{S}_k} \left(x^{(k+1)}, e^{(k+1)} \right) \leq J_{\boldsymbol{S}_k} \left(x^{(k+1)}, e^{(k)} \right).$$

Utilizing (13) we have $J_{\boldsymbol{S}_k}(x^{(k+1)}, e^{(k+1)}) \leq J_{\boldsymbol{S}_k}(x^{(k)}, e^{(k)})$.

For any $i \in \boldsymbol{S}_k^{\boldsymbol{c}}$, $e_i^{(k)} \neq 0$. From (5)-(6), we know that the upper bound for $J_{\boldsymbol{S}_k^{\boldsymbol{c}}}(x^{(j)}, e^{(j)})$, $j = 1, 2, \ldots$ is $\alpha \times |\boldsymbol{S}_k^{\boldsymbol{c}}|$, and $J_{\boldsymbol{S}_k^{\boldsymbol{c}}}(x^{(k)}, e^{(k)})$ equals this upper bound. Hence $J_{\boldsymbol{S}_k^{\boldsymbol{c}}}(x^{(k+1)}, e^{(k+1)}) \leq J_{\boldsymbol{S}_k^{\boldsymbol{c}}}(x^{(k)}, e^{(k)})$.

In sum, we have $J(x^{(k+1)}, e^{(k+1)}) \leq J(x^{(k)}, e^{(k)})$. So the value of the objective function is non-increasing in each itera-

tion. As the objective function is non-negative, it will always converge.

If $J(x^{(k+1)}, e^{(k+1)}) = J(x^{(k)}, e^{(k)})$, we must have equality to hold in (13), which implies $x^{(k+1)} = x^{(k)}$ according to (14) and Step 1. $x^{(k+1)} = x^{(k)}$ ensures $e^{(k+1)} = e^{(k)}$ and $\boldsymbol{S}_{k+1} = \boldsymbol{S}_k$. Similarly $\boldsymbol{S}_{k+1} = \boldsymbol{S}_k$ implies $x^{(k+2)} = x^{(k+1)}$, and further $e^{(k+2)} = e^{(k+1)}$ and $\boldsymbol{S}_{k+2} = \boldsymbol{S}_{k+1}$ and so on. So $(x^{(k)}, e^{(k)}) = (x^{(k+1)}, e^{(k+1)}) = (x^{(k+2)}, e^{(k+2)}) = \ldots$, which is a fixed point of AROSI.

Thus it follows that the objective function is strictly decreasing before convergence.

b) Convergence in a finite number of iterations

Now, we show that the objective function must converge in a finite number of iterations. As the number of different index sets $\boldsymbol{S}_k$ is finite (less than 2^m), it suffices to show that the same index set will not appear again before the objective function converges.

Note that the value of the objective function $J(x^{(k+1)}, e^{(k+1)})$ is determined by $x^{(k+1)}$ (as $e^{(k+1)}$ is also determined by $x^{(k+1)}$ according to Step 2).

We first show that the same index set can not reappear in nearby iterations before convergence. Suppose $\boldsymbol{S}_p = \boldsymbol{S}_{p-1}$, as $x^{(p)} = \arg \min_x \|y_{\boldsymbol{S}_{p-1}} - A_{\boldsymbol{S}_{p-1}} x\|_1 = \arg \min_x \|y_{\boldsymbol{S}_p} - A_{\boldsymbol{S}_p} x\|_1$, and $x^{(p+1)} = \arg \min_x \|y_{\boldsymbol{S}_p} - A_{\boldsymbol{S}_p} x\|_1$, we must have $\|y_{\boldsymbol{S}_p} - A_{\boldsymbol{S}_p} x^{(p+1)}\|_1 = \|y_{\boldsymbol{S}_p} - A_{\boldsymbol{S}_p} x^{(p)}\|_1$, so the algorithm sets $x^{(p+1)} = x^{(p)}$ in Step 1. Then we must have convergence of the objective function.

Then it remains to show that the same index set can not reappear in non-consecutive iterations before convergence.

Before convergence, we have $J(x^{(1)}, e^{(1)}) > \ldots > J(x^{(p+1)}, e^{(p+1)}) > \ldots > J(x^{(r)}, e^{(r)}) > J(x^{(r+1)}, e^{(r+1)}) > \ldots$. The corresponding index sets in Step 1 of each iteration are $\boldsymbol{S}_0, \ldots, \boldsymbol{S}_p, \ldots, \boldsymbol{S}_{r-1}, \boldsymbol{S}_r, \ldots$. We only need to show that $\boldsymbol{S}_r \neq \boldsymbol{S}_p$ for any $r > p+1$. As proved earlier, any $x^{(r+1)} \in \arg \min_x \|y_{\boldsymbol{S}_r} - A_{\boldsymbol{S}_r} x\|_1$ ensures $J(x^{(r+1)}, e^{(r+1)}) \leq J(x^{(r)}, e^{(r)})$, see (13). Suppose $\boldsymbol{S}_r = \boldsymbol{S}_p$, then for any $x^{(p+1)} \in \arg \min_x \|y_{\boldsymbol{S}_p} - A_{\boldsymbol{S}_p} x\|_1$, $x^{(p+1)} \in \arg \min_x \|y_{\boldsymbol{S}_r} - A_{\boldsymbol{S}_r} x\|_1$, thus $J(x^{(p+1)}, e^{(p+1)}) \leq J(x^{(r)}, e^{(r)})$, which is contradictory to $J(x^{(p+1)}, e^{(p+1)}) > J(x^{(r)}, e^{(r)})$.

c) Convergence to a local optimum

We now prove that when $J(x, e)$ converges ($J(x^{(k+1)}, e^{(k+1)}) = J(x^{(k)}, e^{(k)})$), $(x^{(k)}, e^{(k)})$ is a local optimum. From (4), we have

$$J\left(x^{(k)}, e^{(k)}\right) = \left\|y - Ax^{(k)} - e^{(k)}\right\|_1 + \alpha \left\|e^{(k)}\right\|_0$$

Let $(\Delta x, \Delta e)$ be a small deformation vector around $(x^{(k)}, e^{(k)})$. Then

$$\begin{aligned} &J\left(x^{(k)} + \Delta x, e^{(k)} + \Delta e\right) \\ &= \left\|y - A\left(x^{(k)} + \Delta x\right) - \left(e^{(k)} + \Delta e\right)\right\|_1 + \alpha \left\|e^{(k)} + \Delta e\right\|_0 \end{aligned} \tag{15}$$

Next we will show that $J(x^{(k)} + \Delta x, e^{(k)} + \Delta e) \geq J(x^{(k)}, e^{(k)})$ as long as $\|\Delta e\|_1$ is small enough.

Notice that when $\|\Delta e\|_1$ is small enough,

$$\alpha I(e_i^{(k)} + \Delta e_i \neq 0) = \begin{cases} \alpha I(\Delta e_i \neq 0), & e_i^{(k)} = 0 \\ \alpha I(e_i^{(k)} \neq 0), & \text{otherwise} \end{cases}$$

$$\text{So } \alpha \left\|e^{(k)} + \Delta e\right\|_0 = \alpha \left\|e^{(k)}\right\|_0 + \alpha \sum_{i \in \boldsymbol{S}_k} I\left(\Delta e_i \neq 0\right)$$

$$= \alpha \left\|e^{(k)}\right\|_0 + \alpha \|\Delta e_{\boldsymbol{S}_k}\|_0 \tag{16}$$

$$\begin{aligned} &\text{As } \left\|y - A\left(x^{(k)} + \Delta x\right) - \left(e^{(k)} + \Delta e\right)\right\|_1 \\ &\geq \left\|y_{\boldsymbol{S}_k} - A_{\boldsymbol{S}_k}\left(x^{(k)} + \Delta x\right) - \left(e^{(k)}_{\boldsymbol{S}_k} + \Delta e_{\boldsymbol{S}_k}\right)\right\|_1 \\ &\overset{(a)}{=} \left\|y_{\boldsymbol{S}_k} - A_{\boldsymbol{S}_k}\left(x^{(k)} + \Delta x\right) - \Delta e_{\boldsymbol{S}_k}\right\|_1 \\ &\geq \left\|y_{\boldsymbol{S}_k} - A_{\boldsymbol{S}_k}\left(x^{(k)} + \Delta x\right)\right\|_1 - \|\Delta e_{\boldsymbol{S}_k}\|_1 \\ &\overset{(b)}{\geq} \left\|y_{\boldsymbol{S}_k} - A_{\boldsymbol{S}_k} x^{(k+1)}\right\|_1 - \|\Delta e_{\boldsymbol{S}_k}\|_1 \\ &\overset{(c)}{=} \left\|y_{\boldsymbol{S}_k} - A_{\boldsymbol{S}_k} x^{(k)}\right\|_1 - \|\Delta e_{\boldsymbol{S}_k}\|_1 \\ &\overset{(d)}{=} \left\|y_{\boldsymbol{S}_k} - A_{\boldsymbol{S}_k} x^{(k)} - e^{(k)}_{\boldsymbol{S}_k}\right\|_1 - \|\Delta e_{\boldsymbol{S}_k}\|_1 \\ &\overset{(e)}{=} \left\|y - Ax^{(k)} - e^{(k)}\right\|_1 - \|\Delta e_{\boldsymbol{S}_k}\|_1 \end{aligned} \tag{17}$$

where step (a) and (d) follow from the fact that $e^{(k)}_{\boldsymbol{S}_k} = 0$, step (b) is from our Step 1, step (c) is from the convergence, see (14), and step (e) is from (5).

Substituting (16) and (17) in (15), we have

$$\begin{aligned} &J(x^{(k)} + \Delta x, e^{(k)} + \Delta e) \\ &\geq \left\|y - Ax^{(k)} - e^{(k)}\right\|_1 + \alpha \left\|e^{(k)}\right\|_0 + \alpha \|\Delta e_{\boldsymbol{S}_k}\|_0 - \|\Delta e_{\boldsymbol{S}_k}\|_1 \\ &= J\left(x^{(k)}, e^{(k)}\right) + \alpha \|\Delta e_{\boldsymbol{S}_k}\|_0 - \|\Delta e_{\boldsymbol{S}_k}\|_1. \end{aligned}$$

As long as $\|\Delta e\|_1$ is small enough (as $\|\Delta e_{\boldsymbol{S}_k}\|_1 \leq \|\Delta e\|_1$, then $\|\Delta e_{\boldsymbol{S}_k}\|_1$ is also small enough), we will have $\alpha \|\Delta e_{\boldsymbol{S}_k}\|_0 - \|\Delta e_{\boldsymbol{S}_k}\|_1 \geq 0$, and thus $J(x^{(k)} + \Delta x, e^{(k)} + \Delta e) \geq J(x^{(k)}, e^{(k)})$. So $(x^{(k)}, e^{(k)})$ is a local optimum of $J(x, e)$.

In the extreme case where $\boldsymbol{S}_k = \emptyset$, AROSI also sets $x^{(k+1)} = x^{(k)}$. Theorem 1 still holds. ■

B. Proof of Lemma 1

Let us first show that $A_{\boldsymbol{T}^c}$ must be full column rank for any $\boldsymbol{T} \subset \boldsymbol{M}$ with $|\boldsymbol{T}| = t \leq q$. As $\boldsymbol{m}(A) \geq q \geq t$, from Proposition 2, $c_t(A) > \frac{1}{2}$. If $t = 0$, $A_{\boldsymbol{T}^c} = A$ is full column rank. If $t > 0$, suppose $A_{\boldsymbol{T}^c}$ is not full column rank. Then there exists $g \in \mathbb{R}^n$ and $g \neq 0$, such that $A_{\boldsymbol{T}^c} g = 0$. Thus $\min_{\substack{g \in \mathbb{R}^n \\ g \neq 0}} \frac{\sum_{i \in \boldsymbol{T}^c} |a_i{}^T g|}{\sum_{i \in \boldsymbol{M}} |a_i{}^T g|} = 0$. This contradicts $c_t(A) = \min_{\substack{g \in T \subset M \\ |\boldsymbol{T}| = t}} \frac{\min_{\substack{g \in \mathbb{R}^n \\ g \neq 0}} \sum_{i \in \boldsymbol{T}^c} |a_i{}^T g|}{\sum_{i \in \boldsymbol{M}} |a_i{}^T g|} > \frac{1}{2}$, so $A_{\boldsymbol{T}^c}$ must be full column rank.

The following proof is motivated by the proof of Theorem 3.4 in [48].

From Proposition 2, we must have $m > 2q$. Let $q' = q - \lceil 0.5t \rceil$. As $0 \le t \le q$, we have $0 \le q' \le q$, so $m > 2q \ge t + q'$. For any given index set $\boldsymbol{T} \subset \boldsymbol{M}$ with $|\boldsymbol{T}| = t$, and any index set $\boldsymbol{R} \subset \boldsymbol{T^c}$ with $|\boldsymbol{R}| = q'$, and any $g \in \mathbb{R}^n, g \neq 0$, define index set $\boldsymbol{L} :=$ {indices of the largest $\lceil 0.5t \rceil$ entries of $|A_{\boldsymbol{T}} g|$}, and index set $\boldsymbol{E} = \boldsymbol{R} \cup \boldsymbol{L}$. As $\boldsymbol{R} \subset \boldsymbol{T^c}$ and $\boldsymbol{L} \subset \boldsymbol{T}$, so $\boldsymbol{R} \cap \boldsymbol{L} = \emptyset$, $|\boldsymbol{E}| = |\boldsymbol{R}| + |\boldsymbol{L}| = q' + \lceil 0.5t \rceil = q$. We have $\boldsymbol{T} = (\boldsymbol{T} \backslash \boldsymbol{L}) \cup \boldsymbol{L}, \boldsymbol{T^c} = \boldsymbol{R} \cup (\boldsymbol{T^c} \backslash \boldsymbol{R}), \boldsymbol{M} = \boldsymbol{T} \cup \boldsymbol{T^c} = (\boldsymbol{T} \backslash \boldsymbol{L}) \cup \boldsymbol{L} \cup \boldsymbol{R} \cup (\boldsymbol{T^c} \backslash \boldsymbol{R}) = (\boldsymbol{T} \backslash \boldsymbol{L}) \cup \boldsymbol{E} \cup (\boldsymbol{T^c} \backslash \boldsymbol{R}), \boldsymbol{E^c} = (\boldsymbol{T} \backslash \boldsymbol{L}) \cup (\boldsymbol{T^c} \backslash \boldsymbol{R})$. As $(\boldsymbol{T} \backslash \boldsymbol{L}) \cap (\boldsymbol{T^c} \backslash \boldsymbol{R}) = \emptyset$, we have

$$\sum_{i \in \boldsymbol{T^c} \backslash \boldsymbol{R}} \left| {a_i}^T g \right| = \sum_{i \in \boldsymbol{E^c}} \left| {a_i}^T g \right| - \sum_{i \in \boldsymbol{T} \backslash \boldsymbol{L}} \left| {a_i}^T g \right|.$$

Let us first consider the case $q > 0$.

As $\boldsymbol{m}(A) \ge q = |\boldsymbol{E}|$, from Definition 1 we know $\frac{\sum_{i \in \boldsymbol{M} \backslash \boldsymbol{E}} |{a_i}^T g|}{\sum_{i \in \boldsymbol{M}} |{a_i}^T g|} \ge c_q(A)$ with $\frac{1}{2} < c_q(A) < 1$, this leads to $\sum_{i \in \boldsymbol{E^c}} |{a_i}^T g| \ge \frac{c_q(A)}{1 - c_q(A)} \sum_{i \in \boldsymbol{E}} |{a_i}^T g|$, where $\frac{c_q(A)}{1 - c_q(A)} > 1$.

So we have

$$\begin{aligned}
&\sum_{i \in \boldsymbol{T^c} \backslash \boldsymbol{R}} \left| {a_i}^T g \right| \\
&\ge \frac{c_q(A)}{1 - c_q(A)} \sum_{i \in \boldsymbol{E}} \left| {a_i}^T g \right| - \sum_{i \in \boldsymbol{T} \backslash \boldsymbol{L}} \left| {a_i}^T g \right| \\
&= \frac{c_q(A)}{1 - c_q(A)} \left(\sum_{i \in \boldsymbol{R}} \left| {a_i}^T g \right| + \sum_{i \in \boldsymbol{L}} \left| {a_i}^T g \right| \right) - \sum_{i \in \boldsymbol{T} \backslash \boldsymbol{L}} \left| {a_i}^T g \right| \\
&\ge \frac{c_q(A)}{1 - c_q(A)} \sum_{i \in \boldsymbol{R}} \left| {a_i}^T g \right| + \sum_{i \in \boldsymbol{L}} \left| {a_i}^T g \right| - \sum_{i \in \boldsymbol{T} \backslash \boldsymbol{L}} \left| {a_i}^T g \right| \quad (18)
\end{aligned}$$

As $|\boldsymbol{T}| = t$, $|\boldsymbol{L}| = \lceil 0.5t \rceil$, by the definition of index set $\boldsymbol{L}$, we must have

$$\sum_{i \in \boldsymbol{L}} \left| {a_i}^T g \right| - \sum_{i \in \boldsymbol{T} \backslash \boldsymbol{L}} \left| {a_i}^T g \right| \ge 0. \quad (19)$$

So from (18) and (19), we have

$$\sum_{i \in \boldsymbol{T^c} \backslash \boldsymbol{R}} \left| {a_i}^T g \right| \ge \frac{c_q(A)}{1 - c_q(A)} \sum_{i \in \boldsymbol{R}} \left| {a_i}^T g \right|. \quad (20)$$

As $\frac{1}{2} < c_q(A) < 1$, (20) implies $\frac{\sum_{i \in \boldsymbol{T^c} \backslash \boldsymbol{R}} |{a_i}^T g|}{\sum_{i \in \boldsymbol{T^c}} |{a_i}^T g|} \ge c_q(A)$. So $c_{q-0.5t}(A_{\boldsymbol{T^c}}) = \min_{\substack{R \subset \boldsymbol{T^c} \\ |\boldsymbol{R}| = q - \lceil 0.5t \rceil}} \min_{\substack{g \in \mathbb{R}^n \\ g \neq 0}} \frac{\sum_{i \in \boldsymbol{T^c} \backslash \boldsymbol{R}} |{a_i}^T g|}{\sum_{i \in \boldsymbol{T^c}} |{a_i}^T g|} \ge c_q(A)$.

For the case $q = 0$, t must be zero. So $c_{q-0.5t}(A_{\boldsymbol{T^c}}) = c_0(A_{\boldsymbol{T^c}}) = 1 = c_q(A)$.

In sum, we have $c_{q-0.5t}(A_{\boldsymbol{T^c}}) \ge c_q(A)$. As $q - t \le q - \lceil 0.5t \rceil$, from Proposition 1, we further have $c_{q-t}(A_{\boldsymbol{T^c}}) \ge c_{q - \lceil 0.5t \rceil}(A_{\boldsymbol{T^c}}) \ge c_q(A) > \frac{1}{2}$. From Definition 1, we must have $\boldsymbol{m}(A_{\boldsymbol{T^c}}) \ge q - \lceil 0.5t \rceil \ge q - t$. ■

C. Theorem 2 of [37]

Let $x \in \mathbb{R}^n$, $e \in \mathbb{R}^m$, and set $y = Ax + e$, where $A \in \mathbb{R}^{m \times n}$ is full column rank. Then, x is the unique solution of the problem $\min_{g \in \mathbb{R}^n} \|y - Ag\|_1$ for any $\|e\|_0 \le q$ if and only if $q \le \boldsymbol{m}(A)$.

D. Lemma 3

The following Lemma facilitates the proof of Lemma 2 and Theorem 7, and is not introduced in the main text.

Let $y = Ax + e + \eta$ and $\boldsymbol{E} = supp(e)$ satisfying $|\boldsymbol{E}| = q \le \boldsymbol{m}(A)$. Denote $r_{\ell_1} = y - Ax_{\ell_1}$, where $x_{\ell_1} = \arg\min_x \|y - Ax\|_1$. Then $\|(e + \eta) - r_{\ell_1}\|_1 \le \frac{\sum_{i \in \boldsymbol{E^c}} |\eta_i|}{c_q(A) - 0.5} \le \frac{\sqrt{m - q}\, \|\eta\|_2}{c_q(A) - 0.5}$.

Proof: Let us first quote an important Lemma, from Lemma 1 of [37]: Let $\boldsymbol{E} \subset \boldsymbol{M}$, and $y, b^* \in \mathbb{R}^m$, as well as $g^*, g \in \mathbb{R}^n$ be arbitrary. Define $\boldsymbol{E^c} = \boldsymbol{M} \backslash \boldsymbol{E}$. If $|\boldsymbol{E}| = q \le \boldsymbol{m}(A)$, then $\|y - Ag - b^*\|_1 - \|y - Ag^* - b^*\|_1 \ge (2c_q(A) - 1) \times \|A(g - g^*)\|_1 - 2\sum_{i \in \boldsymbol{E^c}} |y_i - {a_i}^T g^* - b^*|$.

Setting $b^* = 0$, $g = x_{\ell_1}$, and $g^* = x$ in this Lemma, we have $0 \ge \|y - Ax_{\ell_1}\|_1 - \|y - Ax\|_1 \ge (2c_q(A) - 1) \times \|A(x_{\ell_1} - x)\|_1 - 2\sum_{i \in \boldsymbol{E^c}} |y_i - {a_i}^T x| = (2c_q(A) - 1) \times \|A(x_{\ell_1} - x)\|_1 - 2\sum_{i \in \boldsymbol{E^c}} |\eta_i|$, where the first inequality is from the optimality of x_{ℓ_1}, and the last equality is from the fact that $y_i = {a_i}^T x + \eta_i$, $\forall\, i \in \boldsymbol{E^c}$.

As $q \le m(A)$, from Proposition 2, we have $c_q(A) > \frac{1}{2}$. So we have $\frac{\sum_{i \in \boldsymbol{E^c}} |\eta_i|}{c_q(A) - 0.5} \ge \|A(x_{\ell_1} - x)\|_1 = \|(y - Ax) - (y - Ax_{\ell_1})\|_1 = \|(e + \eta) - r_{\ell_1}\|_1$.

Using the inequality of the norm, we have

$$\sum_{i \in \boldsymbol{E^c}} |\eta_i| \le \sqrt{|\boldsymbol{E^c}|} \sqrt{\sum_{i \in \boldsymbol{E^c}} |\eta_i|^2} \le \sqrt{m - q}\, \|\eta\|_2$$

So $\|(e + \eta) - r_{\ell_1}\|_1 \le \frac{\sum_{i \in \boldsymbol{E^c}} |\eta_i|}{c_q(A) - 0.5} \le \frac{\sqrt{m - q}\, \|\eta\|_2}{c_q(A) - 0.5}$. ■

E. Proof of Theorem 5

As $|\boldsymbol{E}| = q \le m(A)$ and $\boldsymbol{E^c} \subseteq \boldsymbol{S}_k$, we have $\boldsymbol{S}_k^{\boldsymbol{c}} \subseteq \boldsymbol{E}$ and $|\boldsymbol{R}| = |\boldsymbol{E} \cap \boldsymbol{S}_k| = |\boldsymbol{E}| - |\boldsymbol{S}_k^{\boldsymbol{c}}| \le \boldsymbol{m}(A_{\boldsymbol{S}_k})$, and $A_{\boldsymbol{S}_k}$ is full column rank from Lemma 2, thus the condition of Corollary 1 is satisfied, $\|x^{(k+1)} - x\|_2$ is bounded as in (11).

As $\boldsymbol{S}_k \subseteq \boldsymbol{M}$, for $\forall \boldsymbol{B'} \in \boldsymbol{Q'}_{|\boldsymbol{R}|}$ defined on $A_{\boldsymbol{S}_k}$, it has corresponding index set $\boldsymbol{B}$ defined on A, and $(A_{\boldsymbol{S}_k})_{\boldsymbol{B'}} = A_{\boldsymbol{B}}$. From the definition of $\boldsymbol{Q'}_{|\boldsymbol{R}|}$, there exists a set $\boldsymbol{L} \subseteq \boldsymbol{S}_k$ (both defined on A), $\boldsymbol{L} \supseteq \boldsymbol{B}$, $|\boldsymbol{L}| = |\boldsymbol{S}_k| - |\boldsymbol{R}|$, such that the following holds: $\sum_{i \in \boldsymbol{B} \cup (\boldsymbol{S}_k \backslash \boldsymbol{L})} |{a_i}^T v| \ge \sum_{i \in (\boldsymbol{L} \backslash \boldsymbol{B})} |{a_i}^T v|$, where v is any of the singular vectors corresponding to the smallest singular value of the $|\boldsymbol{B}| \times n$ submatrix $A_{\boldsymbol{B}}$ (of $A_{\boldsymbol{S}_k}$): $\|A_{\boldsymbol{B}} v\|_2 = \sigma_{\min}(A_{\boldsymbol{B}}) \|v\|_2$. Then we have $\sum_{i \in \boldsymbol{B} \cup (\boldsymbol{M} \backslash \boldsymbol{L})} |{a_i}^T v| \ge \sum_{i \in \boldsymbol{B} \cup (\boldsymbol{S_k} \backslash \boldsymbol{L})} |{a_i}^T v| \ge \sum_{i \in (\boldsymbol{L} \backslash \boldsymbol{B})} |{a_i}^T v|$. As $|\boldsymbol{L}| = |\boldsymbol{S}_k| - |\boldsymbol{R}| = |\boldsymbol{S}_k| - (|\boldsymbol{E}| - |\boldsymbol{S}_k^{\boldsymbol{c}}|) = |\boldsymbol{S}_k| + |\boldsymbol{S}_k^{\boldsymbol{c}}| - |\boldsymbol{E}| = m - |\boldsymbol{E}|$, from Definition 2 we know that $\boldsymbol{B} \in \boldsymbol{Q}_{|\boldsymbol{E}|}$. So $\boldsymbol{Q'}_{|\boldsymbol{R}|}$ (defined in terms of $A_{\boldsymbol{S}_k}$) corresponds to a subset of $\boldsymbol{Q}_{|\boldsymbol{E}|}$ (defined in terms of A). Thus $\{(A_{\boldsymbol{S}_k})_{\boldsymbol{B'}} : \boldsymbol{B'} \in \boldsymbol{Q'}_{|\boldsymbol{R}|}\} \subseteq \{A_{\boldsymbol{B}} : \boldsymbol{B} \in \boldsymbol{Q}_{|\boldsymbol{E}|}\}$. So $\max_{\boldsymbol{B'} \in \boldsymbol{Q'}_{|\boldsymbol{R}|}} \frac{1}{\sigma_{\min}((A_{\boldsymbol{S}_k})_{\boldsymbol{B'}})} \le \max_{\boldsymbol{B} \in \boldsymbol{Q}_{|\boldsymbol{E}|}} \frac{1}{\sigma_{\min}(A_{\boldsymbol{B}})}$. Together with $\|\eta_{\boldsymbol{S}_k}\|_2 \le \|\eta\|_2$, this shows that the bound in (11) is smaller than or equal to the bound in (10). ■

F. Proof of Lemma 2

As $\boldsymbol{S}_k \supseteq \boldsymbol{E}^{\boldsymbol{c}}$, so $\boldsymbol{S}_k^{\boldsymbol{c}} \subseteq \boldsymbol{E}$ and $|\boldsymbol{S}_k^{\boldsymbol{c}}| \leq |\boldsymbol{E}| = q$. From Lemma 1 we know that $A_{\boldsymbol{S}_k}$ is full column rank, $\boldsymbol{m}(A_{\boldsymbol{S}_k}) \geq q - |\boldsymbol{S}_k^{\boldsymbol{c}}|$ and $c_{(q-|\boldsymbol{S}_k^{\boldsymbol{c}}|)}(A_{\boldsymbol{S}_k}) \geq c_q(A) > \frac{1}{2}$. So

$$c_{\left(q-|\boldsymbol{S}_k^{\boldsymbol{c}}|\right)}\left(A_{\boldsymbol{S}_k}\right) - 0.5 \geq c_q\left(A\right) - 0.5 > 0. \tag{21}$$

As $\boldsymbol{S}_k^{\boldsymbol{c}} \subseteq \boldsymbol{E}$, so $|supp(e_{\boldsymbol{S}_k})| = \|e_{\boldsymbol{S}_k}\|_0 = \|e\|_0 - |\boldsymbol{S}_k^{\boldsymbol{c}}| = q - |\boldsymbol{S}_k^{\boldsymbol{c}}| \leq \boldsymbol{m}(A_{\boldsymbol{S}_k})$. From Lemma 3 and (21), we have

$$\begin{aligned}\left\|(e+\eta)_{\boldsymbol{S}_k} - r_{\boldsymbol{S}_k}^{(k+1)}\right\|_1 &\leq \frac{\sqrt{m-q}\,\|\eta\|_2}{c_{\left(q-|\boldsymbol{S}_k^{\boldsymbol{c}}|\right)}\left(A_{\boldsymbol{S}_k}\right) - 0.5} \\ &\leq \frac{\sqrt{m-q}\,\|\eta\|_2}{c_q\left(A\right) - 0.5}.\end{aligned}$$

For $\forall i \in \boldsymbol{E}^{\boldsymbol{c}} \subseteq \boldsymbol{S}_k$, $e_i = 0$, $|r_i^{(k+1)}| - |\eta_i| \leq |\eta_i - r_i^{(k+1)}| = |(e+\eta)_i - r_i^{(k+1)}| \leq \|(e+\eta)_{\boldsymbol{S}_k} - r_{\boldsymbol{S}_k}^{(k+1)}\|_1$.
So $|r_i^{(k+1)}| \leq |\eta_i| + \|(e+\eta)_{\boldsymbol{S}_k} - r_{\boldsymbol{S}_k}^{(k+1)}\|_1 \leq \|\eta\|_\infty + \|(e+\eta)_{\boldsymbol{S}_k} - r_{\boldsymbol{S}_k}^{(k+1)}\|_1 \leq \|\eta\|_\infty + \frac{\sqrt{m-q}\,\|\eta\|_2}{c_q(A)-0.5}$. ■

G. Proof of Theorem 6

a) In Step 1 of the $(k+1)$st (e.g., $k = 0, 1, \ldots$) iteration, if $\boldsymbol{S}_k \supseteq \boldsymbol{E}^{\boldsymbol{c}}$, from Lemma 2 we have $\forall i \in \boldsymbol{E}^{\boldsymbol{c}}$, $|r_i^{(k+1)}| \leq \|\eta\|_\infty + C_1 < \alpha$, then $e_i^{(k+1)} = 0$ according to (5). Then $\boldsymbol{S}_{k+1} := \{i : e_i^{(k+1)} = 0\} \supseteq \boldsymbol{E}^{\boldsymbol{c}}$.

As $\boldsymbol{S}_0 = \boldsymbol{M} \supseteq \boldsymbol{E}^{\boldsymbol{c}}$, we will have $\boldsymbol{S}_k \supseteq \boldsymbol{E}^{\boldsymbol{c}}$ for any $k \in \mathbb{Z}_{\geq 0}$.

b) As $\boldsymbol{S}_k \supseteq \boldsymbol{E}^{\boldsymbol{c}}$ for any $k \in \mathbb{Z}_{\geq 0}$, from Lemma 2 we have $\|(e+\eta)_{\boldsymbol{S}_k} - r_{\boldsymbol{S}_k}^{(k+1)}\|_1 \leq \frac{\sqrt{m-q}\,\|\eta\|_2}{c_q(A)-0.5}$ for any $k \in \mathbb{Z}_{\geq 0}$. From Lemma 1, we know $A_{\boldsymbol{E}^{\boldsymbol{c}}}$ is full column rank and thus $\sigma_{\min}(A_{\boldsymbol{E}^{\boldsymbol{c}}}) > 0$. As $\frac{\sqrt{m-q}\,\|\eta\|_2}{c_q(A)-0.5} \geq \|(e+\eta)_{\boldsymbol{S}_k} - r_{\boldsymbol{S}_k}^{(k+1)}\|_1 = \|(y_{\boldsymbol{S}_k} - A_{\boldsymbol{S}_k}x) - (y_{\boldsymbol{S}_k} - A_{\boldsymbol{S}_k}x^{(k+1)})\|_1 = \|A_{\boldsymbol{S}_k}(x - x^{(k+1)})\|_1 \geq \|A_{\boldsymbol{E}^{\boldsymbol{c}}}(x - x^{(k+1)})\|_1 \geq \|A_{\boldsymbol{E}^{\boldsymbol{c}}}(x - x^{(k+1)})\|_2 \geq \sigma_{\min}(A_{\boldsymbol{E}^{\boldsymbol{c}}})\|x - x^{(k+1)}\|_2$, we have $\|x - x^{(k+1)}\|_2 \leq \frac{\sqrt{m-q}\,\|\eta\|_2}{\sigma_{\min}(A_{\boldsymbol{E}^{\boldsymbol{c}}}) \times (c_q(A)-0.5)}$ for any $k \in \mathbb{Z}_{\geq 0}$.

For any $k \in \mathbb{Z}_{\geq 0}, \forall i \in \boldsymbol{P} \subseteq \boldsymbol{E}$, we have $|e_i| - |\eta_i| - |r_i^{(k+1)}| \leq |(e+\eta)_i| - |r_i^{(k+1)}| \leq |(e+\eta)_i - r_i^{(k+1)}| \leq \|(e + \eta - r^{(k+1)})_{\boldsymbol{E}}\|_2 = \|(y - Ax)_{\boldsymbol{E}} - (y - Ax^{(k+1)})_{\boldsymbol{E}}\|_2 = \|A_{\boldsymbol{E}}(x - x^{(k+1)})\|_2 \leq \sigma_{\max}(A_{\boldsymbol{E}})\|x - x^{(k+1)}\|_2 \leq \frac{\sigma_{\max}(A_{\boldsymbol{E}})\sqrt{m-q}\,\|\eta\|_2}{\sigma_{\min}(A_{\boldsymbol{E}^{\boldsymbol{c}}}) \times (c_q(A)-0.5)} = C_3$, so $|r_i^{(k+1)}| \geq |e_i| - |\eta_i| - C_3 \geq \min\{|e_i| : i \in \boldsymbol{P}\} - \|\eta\|_\infty - C_3 > \alpha$, then $e_i^{(k+1)} \neq 0$ according to (5). Then $\boldsymbol{P} \subseteq \boldsymbol{S}_{k+1}^{\boldsymbol{c}} := \{i : e_i^{(k+1)} \neq 0\}$ for any $k \in \mathbb{Z}_{\geq 0}$.

c) For any $k \in \mathbb{Z}_{\geq 0}$, as $\boldsymbol{S}_k \supseteq \boldsymbol{E}^{\boldsymbol{c}}$ and $|\boldsymbol{E}| = q \leq m(A)$, from Theorem 5 we know $\|x^{(k+1)} - x\|_2$ is bounded.

d) In Step 1 of the first iteration, as the condition of Lemma 2 is satisfied, we have $\|e + \eta - r^{(1)}\|_1 \leq C_1$. So $\forall i \in \boldsymbol{E}$, we have $|e_i| - |\eta_i| - |r_i^{(1)}| \leq |(e+\eta)_i| - |r_i^{(1)}| \leq |(e+\eta)_i - r_i^{(1)}| \leq \|e + \eta - r^{(1)}\|_1 \leq C_1$, thus $|r_i^{(1)}| \geq |e_i| - |\eta_i| - C_1 \geq \min\{|e_i|: e_i \neq 0\} - \|\eta\|_\infty - C_1 \geq \min\{|e_i| : e_i \neq 0\} - \|\eta\|_\infty - C_2 > \alpha$. Then $e_i^{(1)} \neq 0$ according to (5). Then $\boldsymbol{E} \subseteq \boldsymbol{S}_1^{\boldsymbol{c}} := \{i : e_i^{(1)} \neq 0\}$. As $\alpha > \|\eta\|_\infty + C_1$ guarantees $\boldsymbol{S}_1 \supseteq \boldsymbol{E}^{\boldsymbol{c}}$, we have $\boldsymbol{S}_1 = \boldsymbol{E}^{\boldsymbol{c}}$.

In Step 1 of the second iteration, as $\boldsymbol{S}_1 = \boldsymbol{E}^{\boldsymbol{c}}$, from Lemma 2, we have $\|(e+\eta)_{\boldsymbol{E}^{\boldsymbol{c}}} - r_{\boldsymbol{E}^{\boldsymbol{c}}}^{(2)}\|_1 \leq \frac{\sqrt{m-q}\,\|\eta\|_2}{c_0(A)-0.5} = 2\sqrt{m-q}\,\|\eta\|_2$

As $\|(e+\eta)_{\boldsymbol{E}^{\boldsymbol{c}}} - r_{\boldsymbol{E}^{\boldsymbol{c}}}^{(2)}\|_1 = \|(y_{\boldsymbol{E}^{\boldsymbol{c}}} - A_{\boldsymbol{E}^{\boldsymbol{c}}}x) - (y_{\boldsymbol{E}^{\boldsymbol{c}}} - A_{\boldsymbol{E}^{\boldsymbol{c}}}x^{(2)})\|_1 = \|A_{\boldsymbol{E}^{\boldsymbol{c}}}(x - x^{(2)})\|_1 \geq \|A_{\boldsymbol{E}^{\boldsymbol{c}}}(x - x^{(2)})\|_2 \geq \sigma_{\min}(A_{\boldsymbol{E}^{\boldsymbol{c}}})\|x - x^{(2)}\|_2$, we have $\|x - x^{(2)}\|_2 \leq 2\sqrt{m-q}\,\|\eta\|_2/\sigma_{\min}(A_{\boldsymbol{E}^{\boldsymbol{c}}})$.

For $\forall i \in \boldsymbol{E}$, we have $|e_i| - |\eta_i| - |r_i^{(2)}| \leq |(e+\eta)_i| - |r_i^{(2)}| \leq |(e+\eta)_i - r_i^{(2)}| \leq \|(e + \eta - r^{(2)})_{\boldsymbol{E}}\|_2 = \|(y - Ax)_{\boldsymbol{E}} - (y - Ax^{(2)})_{\boldsymbol{E}}\|_2 = \|A_{\boldsymbol{E}}(x - x^{(2)})\|_2 \leq \sigma_{\max}(A_{\boldsymbol{E}})\|x - x^{(2)}\|_2 \leq \sigma_{\max}(A_{\boldsymbol{E}}) \times \frac{2\sqrt{m-q}\,\|\eta\|_2}{\sigma_{\min}(A_{\boldsymbol{E}^{\boldsymbol{c}}})} \leq C_2$, thus $|r_i^{(2)}| \geq |e_i| - |\eta_i| - C_2 \geq \min\{|e_i| : e_i \neq 0\} - \|\eta\|_\infty - C_2 > \alpha$. Then $e_i^{(2)} \neq 0$ according to (5). Then $\boldsymbol{E} \subseteq \boldsymbol{S}_2^{\boldsymbol{c}} := \{i : e_i^{(2)} \neq 0\}$. As $\alpha > \|\eta\|_\infty + C_1$ guarantees $\boldsymbol{S}_2 \supseteq \boldsymbol{E}^{\boldsymbol{c}}$, we must have $\boldsymbol{S}_2 = \boldsymbol{E}^{\boldsymbol{c}}$.

Finally, $\boldsymbol{S}_2 = \boldsymbol{E}^{\boldsymbol{c}} = \boldsymbol{S}_1$ implies $x^{(3)} = x^{(2)}$, and further $\boldsymbol{S}_3 = \boldsymbol{S}_2 = \boldsymbol{E}^{\boldsymbol{c}}$. So AROSI converges in 3 iterations and recovers the support of outliers exactly.

e) In the reprojection step, with a threshold in the range of $(\|\eta\|_\infty + C_1, \min\{|e_i| : e_i \neq 0\} - \|\eta\|_\infty - C_2)$, we have $\hat{\boldsymbol{E}} = \boldsymbol{E}$ and $\hat{z} = \arg\ \min_z \|y - [A\ (I_{\boldsymbol{E}})^T]z\|_2$, $\hat{x} = \hat{z}_{\{1,\ldots,n\}}$. As $A_{\boldsymbol{E}^{\boldsymbol{c}}}$ is full column rank, $[A\ (I_{\boldsymbol{E}})^T]$ must also be full column rank (by inspecting the matrix structure). Actually, one can verify that the above $\hat{x}$ is also the unique solution of $\min_x \|y_{\boldsymbol{E}^{\boldsymbol{c}}} - A_{\boldsymbol{E}^{\boldsymbol{c}}}x\|_2$, so $\hat{x} = A_{\boldsymbol{E}^{\boldsymbol{c}}}{}^\dagger y_{\boldsymbol{E}^{\boldsymbol{c}}} = A_{\boldsymbol{E}^{\boldsymbol{c}}}{}^\dagger(A_{\boldsymbol{E}^{\boldsymbol{c}}}x + \eta_{\boldsymbol{E}^{\boldsymbol{c}}}) = x + A_{\boldsymbol{E}^{\boldsymbol{c}}}{}^\dagger\eta_{\boldsymbol{E}^{\boldsymbol{c}}}$. We have $\|\hat{x} - x\|_2 = \|A_{\boldsymbol{E}^{\boldsymbol{c}}}{}^\dagger\eta_{\boldsymbol{E}^{\boldsymbol{c}}}\|_2 \leq \|A_{\boldsymbol{E}^{\boldsymbol{c}}}{}^\dagger\|_2\,\|\eta_{\boldsymbol{E}^{\boldsymbol{c}}}\|_2 = \frac{\|\eta_{\boldsymbol{E}^{\boldsymbol{c}}}\|_2}{\sigma_{\min}(A_{\boldsymbol{E}^{\boldsymbol{c}}})}$.

Next, we want to show that the bound here is better than the bound in (10). Let v_1 be any of the singular vectors corresponding to the smallest singular value of the $A_{\boldsymbol{E}^{\boldsymbol{c}}}$. Since $A_{\boldsymbol{E}^{\boldsymbol{c}}}$ is full column rank, we have $\|A_{\boldsymbol{E}^{\boldsymbol{c}}}v_1\|_2 = \sigma_{\min}(A_{\boldsymbol{E}^{\boldsymbol{c}}})\|v_1\|_2 > 0$.

In Definition 2, we let the set $\boldsymbol{L} = \boldsymbol{E}^{\boldsymbol{c}}$, and let the set $\boldsymbol{B}$ be a subset of $\boldsymbol{E}^{\boldsymbol{c}}$ that corresponds to the $m - q - \boldsymbol{m}(A)$ (according to Proposition 2 it must be positive) smallest entries of $|A_{\boldsymbol{E}^{\boldsymbol{c}}}v_1|$. Then $|\boldsymbol{B} \cup \boldsymbol{L}^{\boldsymbol{c}}| = m - \boldsymbol{m}(A)$, and $|\boldsymbol{L}\backslash\boldsymbol{B}| = \boldsymbol{m}(A)$. Since $c_{\boldsymbol{m}(A)}(A) > 0.5$, we must have (9) holds. So the above set $\boldsymbol{B}$ is a possibly extreme set with q, i.e., $\boldsymbol{B} \in \boldsymbol{Q}_{|\boldsymbol{E}|}$. Since $\sigma_{\min}(A_{\boldsymbol{B}})\|v_1\|_2 \leq \|A_{\boldsymbol{B}}v_1\|_2 \leq \|A_{\boldsymbol{E}^{\boldsymbol{c}}}v_1\|_2 = \sigma_{\min}(A_{\boldsymbol{E}^{\boldsymbol{c}}})\|v_1\|_2$ (where the second inequality becomes a strict inequality as long as $\boldsymbol{m}(A) > 0$), we have $\sigma_{\min}(A_{\boldsymbol{B}}) \leq \sigma_{\min}(A_{\boldsymbol{E}^{\boldsymbol{c}}})$. Then it follows that the bound in (10) is larger or equal to $\|\eta_{\boldsymbol{E}^{\boldsymbol{c}}}\|_2/\sigma_{\min}(A_{\boldsymbol{E}^{\boldsymbol{c}}})$, and is strictly larger when $\boldsymbol{m}(A) > 0$. ■

H. Proof of Theorem 7

a-b) By definition, we have $|\boldsymbol{P}| = t \leq \boldsymbol{m}(A)$, and $\boldsymbol{P} \subseteq \boldsymbol{G} \subseteq \boldsymbol{E}$. We can view e_i indexed by $\boldsymbol{E}\backslash\boldsymbol{G}$ (can be an empty set) as part of the noise, i.e., we define the new noise and corruptions as $\eta_i' = \begin{cases} e_i + \eta_i, & i \in E\backslash G \\ \eta_i, & \text{otherwise} \end{cases}$, $e_i' = \begin{cases} e_i, & i \in G \\ 0, & \text{otherwise} \end{cases}$, then $y = Ax + e' + \eta'$ with $\|e'\|_0 = |\boldsymbol{G}| = \boldsymbol{m}(A) = q_1$.

In Step 1 of the first iteration, we have $\|(e+\eta) - r^{(1)}\|_1 = \|(e' + \eta') - r^{(1)}\|_1 \leq \frac{\sum_{i \in \boldsymbol{G}^{\boldsymbol{c}}} |\eta_i'|}{c_{q_1}(A) - 0.5}$ from Lemma 3. So $\|(e+\eta) - r^{(1)}\|_1 \leq \frac{\sum_{i \in \boldsymbol{G}^{\boldsymbol{c}}} |\eta_i'|}{c_{q_1}(A) - 0.5} \leq \frac{\sum_{i \in \boldsymbol{G}^{\boldsymbol{c}}} |\eta_i + e_i|}{c_{q_1}(A) - 0.5} \leq \frac{\sum_{i \in \boldsymbol{G}^{\boldsymbol{c}}} |\eta_i| + \sum_{i \in \boldsymbol{G}^{\boldsymbol{c}}} |e_i|}{c_{q_1}(A) - 0.5} = \frac{\sum_{i \in \boldsymbol{G}^{\boldsymbol{c}}} |\eta_i| + \sum_{i \in \boldsymbol{E}\backslash\boldsymbol{G}} |e_i|}{c_{q_1}(A) - 0.5} \leq \frac{\sqrt{m-q_1}\,\eta_2 + \sum_{i \in \boldsymbol{E}\backslash\boldsymbol{G}} |e_i|}{c_{q_1}(A) - 0.5}$.

For $\forall i \in \boldsymbol{P}$, we have $|e_i| - |\eta_i| - |r_i^{(1)}| \le |(e+\eta)_i| - |r_i^{(1)}| \le |(e+\eta)_i - r_i^{(1)}| \le \|e+\eta-r^{(1)}\|_1 \le \frac{\sqrt{m-q_1}\,\|\eta\|_2 + \sum_{i\in\boldsymbol{E}\backslash\boldsymbol{G}}|e_i|}{c_{q_1}(A)-0.5} \le w_2$, thus $|r_i^{(1)}| \ge |e_i| - |\eta_i| - w_2 \ge \min\{|e_i| : i \in \boldsymbol{P}\} - \|\eta\|_\infty - w_2 > \alpha$. Then $e_i^{(1)} \ne 0$ according to (5). Then $\boldsymbol{P} \subseteq \boldsymbol{S}_1^{\boldsymbol{c}} := \{i : e_i^{(1)} \ne 0\}$.

For $\forall i \in \boldsymbol{E^c}$, $e_i = 0$, $|r_i^{(1)}| - |\eta_i| \le |\eta_i - r_i^{(1)}| = |(e+\eta)_i - r_i^{(1)}| \le \|(e+\eta) - r^{(1)}\|_1 \le \frac{\sqrt{m-q_1}\,\|\eta\|_2 + \sum_{i\in\boldsymbol{E}\backslash\boldsymbol{G}}|e_i|}{c_{q_1}(A)-0.5} \le w_1$. So $|r_i^{(1)}| \le |\eta_i| + w_1 \le \|\eta\|_\infty + w_1 < \alpha$. Then $e_i^{(1)} = 0$ according to (5). Then $\boldsymbol{E^c} \subseteq \boldsymbol{S}_1 := \{i : e_i^{(1)} = 0\}$.

Next we will show for the $(k+1)$st (e.g., $k = 1, 2, \ldots$) iteration, if $\boldsymbol{E^c} \subseteq \boldsymbol{S}_k$ and $\boldsymbol{P} \subseteq \boldsymbol{S}_k^{\boldsymbol{c}}$, then we will have $\boldsymbol{E^c} \subseteq \boldsymbol{S}_{k+1}$ and $\boldsymbol{P} \subseteq \boldsymbol{S}_{k+1}^{\boldsymbol{c}}$. Thus $\boldsymbol{E^c} \subseteq \boldsymbol{S}_k$ and $\boldsymbol{P} \subseteq \boldsymbol{S}_{k+1}^{\boldsymbol{c}}$ for any $k \in \mathbb{Z}_{\ge 0}$.

As $\boldsymbol{m}(A) \ge |\boldsymbol{P}|$, from Lemma 1, we know that $A_{\boldsymbol{P^c}}$ is full column rank, and $\boldsymbol{m}(A_{\boldsymbol{P^c}}) \ge \boldsymbol{m}(A) - \lceil 0.5 \times |\boldsymbol{P}| \rceil$. So $\boldsymbol{m}(A) \le \boldsymbol{m}(A_{\boldsymbol{P^c}}) + \lceil 0.5 \times |\boldsymbol{P}| \rceil$. Combined with $|\boldsymbol{E}| \le \boldsymbol{m}(A) + \lfloor \frac{t}{2} \rfloor = \boldsymbol{m}(A) + \lfloor 0.5 \times |\boldsymbol{P}| \rfloor$, we have $|\boldsymbol{E}| \le \boldsymbol{m}(A_{\boldsymbol{P^c}}) + \lceil 0.5 \times |\boldsymbol{P}| \rceil + \lfloor 0.5 \times |\boldsymbol{P}| \rfloor = \boldsymbol{m}(A_{\boldsymbol{P^c}}) + |\boldsymbol{P}|$. So

$$|\boldsymbol{E}| - |\boldsymbol{P}| \le \boldsymbol{m}\left(A_{\boldsymbol{P^c}}\right). \tag{22}$$

As $\boldsymbol{P} \subseteq \boldsymbol{S}_k^{\boldsymbol{c}}$, so $\boldsymbol{S}_k \subseteq \boldsymbol{P^c}$. We have $\boldsymbol{E^c} \subseteq \boldsymbol{S_k} \subseteq \boldsymbol{P^c}$ and thus $|\boldsymbol{P^c}\backslash\boldsymbol{S_k}| \le |\boldsymbol{P^c}\backslash\boldsymbol{E^c}| = |\boldsymbol{E}\backslash\boldsymbol{P}| = |\boldsymbol{E}| - |\boldsymbol{P}| \le \boldsymbol{m}(A_{\boldsymbol{P^c}}) = q_2$. From Lemma 1, we have that $A_{\boldsymbol{S}_k}$ is full column rank, $\boldsymbol{m}(A_{\boldsymbol{S}_k}) \ge \boldsymbol{m}(A_{\boldsymbol{P^c}}) - |\boldsymbol{P^c}\backslash\boldsymbol{S}_k| = q_2 - |\boldsymbol{P^c}\backslash\boldsymbol{S}_k|$, and

$$c_{q_2-|\boldsymbol{P^c}\backslash\boldsymbol{S}_k|}\left(A_{\boldsymbol{S}_k}\right) \ge c_{q_2}\left(A_{\boldsymbol{P^c}}\right) > \frac{1}{2}. \tag{23}$$

Combined with (22), we have

$$\begin{aligned}\boldsymbol{m}\left(A_{\boldsymbol{S}_k}\right) &\ge |\boldsymbol{E}| - |\boldsymbol{P}| - |\boldsymbol{P^c}\backslash\boldsymbol{S}_k| = |\boldsymbol{E}\backslash\boldsymbol{P}| - |\boldsymbol{P^c}\backslash\boldsymbol{S}_k| \\ &= |\boldsymbol{P^c}\backslash\boldsymbol{E^c}| - |\boldsymbol{P^c}\backslash\boldsymbol{S}_k| = (|\boldsymbol{P^c}| - |\boldsymbol{E^c}|) - (|\boldsymbol{P^c}| - |\boldsymbol{S}_k|) \\ &= |\boldsymbol{S}_k| - |\boldsymbol{E^c}| = |\boldsymbol{S}_k\backslash\boldsymbol{E^c}| \end{aligned} \tag{24}$$

From Lemma 3, we know that

$$\begin{aligned}&\left\|(e+\eta)_{\boldsymbol{S}_k} - r_{\boldsymbol{S}_k}^{(k+1)}\right\|_1 \\ &\le \frac{\sqrt{m-q}\,\|\eta\|_2}{c_{|\boldsymbol{S}_k\backslash\boldsymbol{E^c}|}\left(A_{\boldsymbol{S}_k}\right)-0.5} = \frac{\sqrt{m-q}\,\|\eta\|_2}{c_{|\boldsymbol{P^c}\backslash\boldsymbol{E^c}|-|\boldsymbol{P^c}\backslash\boldsymbol{S}_k|}\left(A_{\boldsymbol{S}_k}\right)-0.5}\end{aligned} \tag{25}$$

As $|\boldsymbol{P^c}\backslash\boldsymbol{E^c}| \le q_2$, so $|\boldsymbol{P^c}\backslash\boldsymbol{E^c}| - |\boldsymbol{P^c}\backslash\boldsymbol{S}_k| \le q_2 - |\boldsymbol{P^c}\backslash\boldsymbol{S}_k|$, and from Proposition 1 we have $c_{|\boldsymbol{P^c}\backslash\boldsymbol{E^c}|-|\boldsymbol{P^c}\backslash\boldsymbol{S}_k|}(A_{\boldsymbol{S}_k}) \ge c_{q_2-|\boldsymbol{P^c}\backslash\boldsymbol{S}_k|}(A_{\boldsymbol{S}_k})$. Together with (23), we have $c_{|\boldsymbol{P^c}\backslash\boldsymbol{E^c}|-|\boldsymbol{P^c}\backslash\boldsymbol{S}_k|}(A_{\boldsymbol{S}_k}) - 0.5 \ge c_{q_2}(A_{\boldsymbol{P^c}}) - 0.5 > 0$. Combined with (25) we have $\|(e+\eta)_{\boldsymbol{S}_k} - r_{\boldsymbol{S}_k}^{(k+1)}\|_1 \le \frac{\sqrt{m-q}\,\|\eta\|_2}{c_{q_2}(A_{\boldsymbol{P^c}})-0.5}$.

For $\forall i \in \boldsymbol{E^c} \subseteq \boldsymbol{S}_k$, $e_i = 0$, $|r_i^{(k+1)}| - |\eta_i| \le |\eta_i - r_i^{(k+1)}| = |(e+\eta)_i - r_i^{(k+1)}| \le \|(e+\eta)_{\boldsymbol{S}_k} - r_{\boldsymbol{S}_k}^{(k+1)}\|_1 \le \frac{\sqrt{m-q}\,\|\eta\|_2}{c_{q_2}(A_{\boldsymbol{P^c}})-0.5} \le w_1$. So $|r_i^{(k+1)}| \le |\eta_i| + w_1 \le \|\eta\|_\infty + w_1 < \alpha$. Then $e_i^{(k+1)} = 0$ according to (5). Then $\boldsymbol{E^c} \subseteq \boldsymbol{S}_{k+1} := \{i : e_i^{(k+1)} = 0\}$.

As $|\boldsymbol{P^c}\backslash\boldsymbol{E^c}| \le \boldsymbol{m}(A_{\boldsymbol{P^c}})$ and $A_{\boldsymbol{P^c}}$ is full column rank, from Lemma 1, we know $A_{\boldsymbol{E^c}}$ is also full column rank, so $\sigma_{\min}(A_{\boldsymbol{E^c}}) > 0$. As $\frac{\sqrt{m-q}\,\|\eta\|_2}{c_{q_2}(A_{\boldsymbol{P^c}})-0.5} \ge \|(e+\eta)_{\boldsymbol{S}_k} - r_{\boldsymbol{S}_k}^{(k+1)}\|_1 = \|(y_{\boldsymbol{S}_k} - A_{\boldsymbol{S}_k}x) - (y_{\boldsymbol{S}_k} - A_{\boldsymbol{S}_k}x^{(k+1)})\|_1 = \|A_{\boldsymbol{S}_k}(x - x^{(k+1)})\|_1 \ge \|A_{\boldsymbol{E^c}}(x - x^{(k+1)})\|_1 \ge \|A_{\boldsymbol{E^c}}(x - x^{(k+1)})\|_2 \ge \sigma_{\min}(A_{\boldsymbol{E^c}})\|x - x^{(k+1)}\|_2$, so we have

$$\left\|x - x^{(k+1)}\right\|_2 \le \frac{\sqrt{m-q}\,\|\eta\|_2}{\sigma_{\min}\left(A_{\boldsymbol{E^c}}\right) \times \left(c_{q_2}\left(A_{\boldsymbol{P^c}}\right) - 0.5\right)}$$

For $\forall i \in \boldsymbol{P}$, we have $.|e_i| - |\eta_i| - |r_i^{(k+1)}| \le |(e+\eta)_i| - |r_i^{(k+1)}| \le |(e+\eta)_i - r_i^{(k+1)}| \le \|(e+\eta-r^{(k+1)})_{\boldsymbol{P}}\|_2 = \|(y - Ax)_{\boldsymbol{P}} - (y - Ax^{(k+1)})_{\boldsymbol{P}}\|_2 = \|A_{\boldsymbol{P}}(x - x^{(k+1)})\|_2 \le \sigma_{\max}(A_{\boldsymbol{P}})\|x - x^{(k+1)}\|_2 \le \frac{\sigma_{\max}(A_{\boldsymbol{P}})\sqrt{m-q}\,\|\eta\|_2}{\sigma_{\min}(A_{\boldsymbol{E^c}}) \times (c_{q_2}(A_{\boldsymbol{P^c}})-0.5)} \le w_2$, so $|r_i^{(k+1)}| \ge |e_i| - |\eta_i| - w_2 \ge \min\{|e_i| : i \in \boldsymbol{P}\} - \|\eta\|_\infty - w_2 > \alpha$, then $e_i^{(k+1)} \ne 0$ according to (5). Then $\boldsymbol{P} \subseteq \boldsymbol{S}_{k+1}^{\boldsymbol{c}} := \{i : e_i^{(k+1)} \ne 0\}$.

c) As for any $k \in \mathbb{Z}_{\ge 1}$, we have $\boldsymbol{m}(A_{\boldsymbol{S}_k}) \ge |\boldsymbol{S}_k\backslash\boldsymbol{E^c}| = |\boldsymbol{E} \cap \boldsymbol{S}_k|$ from (24). Since $A_{\boldsymbol{S}_k}$ is full column rank, the condition of Corollary 1 is satisfied. Thus $\|x^{(k+1)} - x\|_2$ is bounded. ■

ACKNOWLEDGMENT

The authors would like to thank the associate editor Dr. G. Mateos and the anonymous reviewers for their insightful comments which helped improve the paper. One comment of a reviewer, Reviewer 1, is directly incorporated into Remark 2. J. Liu also thanks Y. Ding for the helpful discussions on SSR.

REFERENCES

[1] G. Papageorgiou, P. Bouboulis, and S. Theodoridis, "Robust linear regression analysis—A greedy approach," *IEEE Trans. Signal Process.*, vol. 63, no. 15, pp. 3872–3887, Aug. 2015.
[2] P. J. Huber, "The 1972 wald lecture robust statistics: A review," *Ann. Math. Statist.*, vol. 43, no. 4, pp. 1041–1067, 1972.
[3] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*. New York, NY, USA: Wiley, 1987.
[4] R. A. Maronna, R. D. Martin, and V. J. Yohai, *Robust Statistics: Theory and Methods*. New York, NY, USA: Wiley, 2006.
[5] P. J. Huber, *Robust Statistics*. New York, NY, USA: Wiley, 1981.
[6] P. J. Rousseeuw, "Least median of squares regression," *J. Amer. Stat. Assoc.*, vol. 79, no. 388, pp. 871–880, 1984.
[7] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.
[8] E. J. Candès and P. A. Randall, "Highly robust error correction byconvex programming," *IEEE Trans. Inf. Theory*, vol. 54, no. 7, pp. 2829–2840, Jul. 2008.
[9] Y. She and A. B. Owen, "Outlier detection using nonconvex penalized regression," *J. Amer. Stat. Assoc.*, vol. 106, no. 494, pp. 626–639, 2011.
[10] K. Bhatia, P. Jain, and P. Kar, "Robust regression via hard thresholding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 721–729.
[11] J. N. Laska, M. A. Davenport, and R. G. Baraniuk, "Exact signal recovery from sparsely corrupted measurements through the pursuit of justice," in *Proc. 43rd Asilomar Conf. Signals, Syst. Comput.*, 2009, pp. 1556–1560.
[12] J. Wright and Y. Ma, "Dense error correction via ℓ1-minimization," *IEEE Trans. Inf. Theory*, vol. 56, no. 7, pp. 3540–3560, Jul. 2010.
[13] N. H. Nguyen and T. D. Tran, "Exact recoverability from dense corrupted observations via ℓ1-Minimization," *IEEE Trans. Inf. Theory*, vol. 59, no. 4, pp. 2017–2035, Apr. 2013.
[14] N. H. Nguyen and T. D. Tran, "Robust lasso with missing and grossly corrupted observations," *IEEE Trans. Inf. Theory*, vol. 59, no. 4, pp. 2036–2058, Apr. 2013.
[15] C. Studer and R. G. Baraniuk, "Stable restoration and separation of approximately sparse signals," *Appl. Comput. Harmon. Anal.*, vol. 37, no. 1, pp. 12–35, 2014.

[16] C. Studer, P. Kuppinger, G. Pope, and H. Bölcskei, "Recovery of sparsely corrupted signals," *IEEE Trans. Inf. Theory*, vol. 58, no. 5, pp. 3115–3130, May 2012.
[17] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.
[18] J. J. Fuchs, "An inverse problem approach to robust regression," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.,* 1999, pp. 1809–1812.
[19] K. Mitra, A. Veeraraghavan, and R. Chellappa, "Robust regression using sparse learning for high dimensional parameter estimation problems," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.,* 2010, pp. 3846–3849.
[20] K. Mitra, A. Veeraraghavan, and R. Chellappa, "Analysis of sparse regularization based robust regression approaches," *IEEE Trans. Signal Process.*, vol. 61, no. 5, pp. 1249–1257, Mar. 2013.
[21] M. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, 2001.
[22] D. P. Wipf and B. D. Rao, "Sparse Bayesian learning for basis selection," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2153–2164, Aug. 2004.
[23] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, 2006.
[24] J. A. Tropp, "Just relax: Convex programming methods for identifying sparse signals in noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 1030–1051, Mar. 2006.
[25] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 6–18, Jan. 2006.
[26] Y. Jin and B. D. Rao, "Algorithms for robust linear regression by exploiting the connection to sparse signal recovery," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.,* 2010, pp. 3830–3833.
[27] V. Kekatos and G. B. Giannakis, "From sparse signals to sparse residuals for robust sensing," *IEEE Trans. Signal Process.*, vol. 59, no. 7, pp. 3355–3368, Jul. 2011.
[28] E. J. Candès, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted $\ell 1$ minimization," *J. Fourier Anal. Appl.*, vol. 14, no. 5, pp. 877–905, Dec. 2008.
[29] M. Fazel, "Matrix rank minimization with applications," Ph.D. dissertation, Stanford University, Stanford, CA, USA, 2002.
[30] M. S. Lobo, M. Fazel, and S. Boyd, "Portfolio optimization with linear and fixed transaction costs," *Ann. Oper. Res.*, vol. 152, no. 1, pp. 341–365, Jul. 2007.
[31] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.,* 2008, pp. 3869–3872.
[32] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. 27th Asilomar Conf. Signals, Syst. Comput.,* 1993, pp. 40–44.
[33] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.
[34] 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.
[35] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.
[36] A. Giloni and M. Padberg, "Alternative methods of linear regression," *Math. Comput. Model.*, vol. 35, no. 3, pp. 361–374, Feb. 2002.
[37] S. Flores, "Sharp non-asymptotic performance bounds for $\ell 1$ and Huber robust regression estimators," *TEST*, vol. 24, pp. 796–812, 2015.
[38] E. Candès and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," *Ann. Statist.*, vol. 35, no. 6, pp. 2313–2351, 2007.
[39] Y. Sharon, J. Wright, and Y. Ma, "Minimum sum of distances estimator: Robustness and stability," in *Proc. 2009 Amer. Control Conf.*, 2009, pp. 524–530.
[40] P. J. Rousseeuw and B. C. van Zomeren, "Unmasking multivariate outliers and leverage points.," *J. Amer. Stat. Assoc.*, vol. 85, no. 411, pp. 633–639, 1990.
[41] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K: Cambridge Univ. Press, 2004.
[42] M. S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, "Applications of second-order cone programming," *Linear Algebra Appl.*, vol. 284, no. 1–3, pp. 193–228, 1998.
[43] P. W. Holland and R. E. Welsch, "Robust regression using iteratively reweighted least-squares," *Commun. Statist.—Theory Methods*, vol. 6, no. 9, pp. 813–827, 1977.
[44] W. Dumouchel and F. O'Brien, "Integrating a robust option into a multiple regression computing environment," in *Computing and Graphics in Statistics*, A. Buja and P. A. Tukey, Eds. New York, NY, USA: Springer-Verlag, 1991, pp. 41–48.
[45] A. E. Beaton and J. W. Tukey, "The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data," *Technometrics*, vol. 16, no. 2, pp. 147–185, 1974.
[46] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust Statistics: The Approach Based on Influence Functions*. New York, NY, USA: Wiley, 2005.
[47] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. New York, NY, USA: Springer-Verlag, 2010.
[48] Y. Wang and W. Yin, "Sparse signal reconstruction via iterative support detection," *SIAM J. Imag. Sci.*, vol. 3, no. 3, pp. 462–491, Jan. 2010.

Jing Liu (S'15) received the B.E. degree in electronic engineering from BIT, Beijing, China, in 2010, and the M.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2013. He is currently working toward the Ph.D. degree in electrical and computer engineering at the University of California, San Diego, CA, USA. His research interests include signal processing, machine learning, and computer vision. He received the first prize in Science and Technology Award, Beijing, in 2013. He was the recipient of the National Fellowship of China, Guanghua Fellowship of Tsinghua University, Frontiers of Innovation Fellowship of UCSD, and the Silver Medal of BIT.

Pamela C. Cosman (S'88–M'93–SM'00–F'08) received the B.S. (Hons.) degree in electrical engineering from CalTech, Pasadena, CA, USA, in 1987, and the Ph.D. degree in EE from Stanford University, Stanford, CA, USA, in 1993. She is currently a Professor in the Department of Electrical and Computer Engineering, UC San Diego, San Diego, CA, USA. She is a past Director at the Center for Wireless Communications, an ECE Department Vice-Chair, and an Associate Dean for Students. Her research interests include image and video compression and processing, and wireless communications. She was an Associate Editor for the IEEE COMMUNICATIONS LETTERS, and the IEEE SIGNAL PROCESSING LETTERS, as well as the Editor-in-Chief (2006–2009) for the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS.

Bhaskar D. Rao (S'80–M'83–SM'91–F'00) is currently a Distinguished Professor in the Department of Electrical and Computer Engineeringand the holder of the Ericsson Endowed Chair in wireless access networks at the University of California, San Diego, CA, USA. He received the 2016 IEEE Signal Processing Society Technical Achievement Award. His research interests include digital signal processing, estimation theory, and optimization theory, with applications to digital communications, speech signal processing, and biomedical signal processing.