Conic Optimization for Robust Quadratic Regression: Deterministic Bounds and Statistical Analysis

Igor Molybog, Ramtin Madani, and Javad Lavaei

***Abstract*— This paper is concerned with the robust quadratic regression problem, where the goal is to find the unknown parameters (state) of a system modeled by nonconvex quadratic equations based on observational data. In this problem, a sparse subset of equations are subject to errors (noise values) of arbitrary magnitudes. We propose two techniques based on conic optimization to address this problem. The first one is a penalized conic relaxation, whereas the second one is a more complex iterative conic optimization equipped with a hard thresholding operator. We derive a deterministic bound for the penalized conic relaxation to quantify how many bad measurements the algorithm can tolerate without producing a nonzero estimation error. This bound is then analyzed for Gaussian systems, and it is proved that the proposed method allows up to a square root of the total number of measurements to be grossly erroneous. If the number of measurements is sufficiently large, we show that the iterative conic optimization method recovers the unknown state precisely even when up to a constant fraction of equations are arbitrarily wrong in the Gaussian case. The efficacy of the developed methods is demonstrated on synthetic data and a European power grid.**

I. Introduction

Nonlinear regression aims to find the parameters of a given model based on observational data. One may assume the existence of a parametrized function $f(\mathbf{x};\mathbf{a})$ defined over the set of all possible models $\mathbf{x} \in \mathcal{X}$ and all possible inputs $\mathbf{a} \in \mathcal{A}$, where the goal is to estimate the true model given a set of imperfect measurements y_i's:

$$y_i = f(\mathbf{x}, \mathbf{a}_i) + \eta_i, \qquad \forall i \in \{1, ..., m\}$$

In this formulation, unknown error vector η could be the measurement noise with modest values. However, a more drastic scenario corresponds to the case where the random vector η is sparse and its nonzero entries are allowed to be arbitrarily large. Under this circumstance, some *a priori* information about the probability distribution of the sparse vector η may be available, in addition to an upper bound on the cardinality of η. This important problem is referred to as *robust regression* and appears in real-world situations when some observations, named outliers, are completely wrong in an unpredictable way. This could occur during an image acquisition with several corrupted pixels, or result from communication issues during data transmission for sensor networks. Such problems arise in different domains of application and have been studied in the literature. In the context of electric power grid, the regression problem is known as state estimation, where the goal is to find the operating point of the system based on the voltages signals measured at buses and power signals measured over lines and at buses [1]–[3]. Outliers in this case are associated with faulty sensors, cyber attacks, or regional data manipulation to impact the electricity market [2], [4].

There are several classical works on robust regression and outliers detection. The book [5] offers an overview of many fundamental results in this area dating back to 1887 when Edgeworth proposed the least absolute values regression estimator. Modern techniques for handling sparse errors of arbitrary magnitudes vary with respect to different features: statistical properties of the error, class of the regression model $f(\mathbf{x};\mathbf{a})$, set of possible true models, type of theoretical guarantees, and characteristics of the adversary model generating errors [6]–[10]. There is a plethora of papers on this topic for the well-known linear regression problem [11]–[15]. In this case, the function $f(\mathbf{x};\mathbf{a})$ is linear in the model vector $\mathbf{x}$, and can be written as $\mathbf{a}^*\mathbf{x}$. Nevertheless, there are far less results for nonlinear regression. This is due to the fact that linear regression amounts to a system of linear equations with a cubic solution complexity if the measurements are error-free, whereas nonlinear regression is NP-hard and its complexity further increases with the inclusion of premeditated errors. However, very special cases of nonlinear regression have been extensively studied in the literature. In particular, the robust phase retrieval problem that can be formulated with $f(\mathbf{x};\mathbf{a}_i) = |\mathbf{a}_i^*\mathbf{x}|^2$ has received considerable attention [9], [16], [17].

To model a general nonlinear regression problem, notice that every smooth nonlinear function can be approximated arbitrarily precisely with a polynomial function, and that every polynomial function can be converted to a quadratic function subject to quadratic equality constraints (playing the role of error-free quadratic measurements) after introducing specific auxiliary variables [18]. This implies that every nonlinear regression could be approximated up to any arbitrary precision with a quadratic regression where the augmented model of the system is quadratic. As a far more general case of phase retrieval, a quadratic regression problem with the variable $\mathbf{x}$ can be modeled as $f(\mathbf{x};\mathbf{A}_i) = \mathbf{x}^*\mathbf{A}_i\mathbf{x}$. The state estimation problem for power systems belongs to the above model due to the quadratic laws of physics (i.e., the quadratic

Email: igormolybog@berkeley.edu, ramtin.madani@uta.edu, lavaei@berkeley.edu

Igor Molybog and Javad Lavaei are with the Department of Industrial Engineering and Operations Research, University of California, Berkeley. Ramtin Madani is with the Department of Electrical Engineering, University of Texas, Arlington. This work was supported by the ONR YIP Award, DARPA YFA Award, AFOSR YIP Award, NSF CAREER Award, and NSF EPCN Award.

relationship between voltage and power), where each matrix A_i is rank 1 or 2. Robust regression in power systems is referred to as *bad data detection*. This problem was first studied in 1971 [19], and there are many recent progresses on this topic [2], [20], [21]. It is worth to mention that there are similar problems that arise in power system analysis, like State Estimation which is robust to wrong topological information [**?**]. However, we do not consider them in this paper.

The existing approaches for robust regression include the analysis of the unconstrained case [8], [11], [13], [15], the constrained scenario with conditions on the sparsity of the solution vector $\mathbf{x}$ [7], [12], [22], [23], and more sophisticated scenarios in the context of matrix completion [6], [10], [24]. Motivated by applications in inverse covariance estimation [25], the papers [23], [26], [27] consider sparse noise in the input vector $\mathbf{a}_i$ as opposed to the additive error considered in this paper. The work [11] is based on l_1-minimization, whereas [7] solves an extended Lasso formulation defined as the minimization of $\|\mathbf{y}-\mathbf{A}\mathbf{x}+\nu\|_2^2+\mu_1\|\mathbf{x}\|_1+\mu_2\|\nu\|_1$. The work [28] proposes to solve a second-order cone programming (SOCP) for robust linear regression, which is related to the current paper with a focus on robust nonlinear regression. In contrast to the above-mentioned papers that aim to develop a single optimization problem to estimate the solution of a linear regression, there are iterative-based methods as well. For instance, [8], [14], [15] propose iterative algorithms via hard thresholding. This technique will be exploited in the current paper as well.

Due to the diversity in the problem formulation and approaches taken by different papers, it is difficult to compare the existing results since there is no single dominant method. However, the most common measures of performance for robust regression algorithms are the traditional algorithmic complexity and the permissible number of gross measurements $\|\eta\|_0$ compared to the total number of measurements m. In this paper, the objective is to design a polynomial-time algorithm, in contrast with potentially exponential-time approaches [29]. As far as the robustness of an algorithm is concerned, the existing works often provide probabilistic guarantees on the recoverability of the original parameter vector $\mathbf{x}$ for linear Gaussian stochastic systems under various assumptions on the relationship between $\|\eta\|_0$ and m. In this case, the ratio $\frac{\|\eta\|_0}{m}$, named breakdown point, is limited by a constant and could even approach 1 if the unknown solution $\mathbf{x}$ is sparse.

A. Contributions and Organization

The main objective of this paper is to analyze a robust regression problem for an arbitrary quadratic model that includes power system state estimation and phase retrieval as special cases. The focus is on the calculation of the maximum number of bad measurements that does not compromise the exact reconstruction of the model vector $\mathbf{x}$. In Section II, we formally state the problem. In Section III, we offer the main results of this paper. First, a penalized conic relaxation is proposed and its performance is analyzed via deterministic bounds. For Gaussian systems, the results are refined and it is shown that the proposed algorithm tolerates up to a square root of the total number of measurements to be arbitrarily wrong without creating any nonzero estimation error. Second, a more computationally-complex method based on iterative conic optimization and hard thresholding is proposed to solve the robust regression problem. In the Gaussian case with a high number of measurements, it is proved that this method allows up to a constant number of equations to be grossly wrong. Numerical results are presented in Section IV, which includes a case study on a European power grid. Concluding remarks are drawn in Section V, followed by the proofs in the appendix.

B. Notation

$\mathbb{R}^n$ and $\mathbb{C}^n$ denote the sets of real and complex n-dimensional vectors, respectively. $\mathbb{H}^n$ and $\mathbb{S}^n$ are the sets of $n \times n$ Hermitian and symmetric matrices. $\text{tr}(\mathbf{A})$ and $\langle \mathbf{A}, \mathbf{B} \rangle$ denote the trace of a matrix $\mathbf{A}$ and the Frobenius inner product of two matrices $\mathbf{A}$ and $\mathbf{B}$. The conjugate transpose of $\mathbf{A}$ is shown as $\mathbf{A}^*$. The notation $\mathbf{A} \circ \mathbf{B}$ refers to the Hadamard (entrywise) multiplication. The eigenvalues of a matrix $\mathbf{M} \in \mathbb{H}^n$ are denoted as $\lambda_1(\mathbf{M}), ..., \lambda_n(\mathbf{M})$ in descending order, from which three parameters are defined as $\lambda_{\max}(\mathbf{M}) = \lambda_1(\mathbf{M})$, $\lambda_{\min}(\mathbf{M}) = \lambda_n(\mathbf{M})$ and $\kappa(\mathbf{M}) = \lambda_{n-1}(\mathbf{M}) + \lambda_n(\mathbf{M})$. Given a matrix $\mathbf{A} \in \mathbb{C}^{n\times m}$ and a set $S \subset \{1, \dots, m\}$, the matrix $\mathbf{A}_S$ is defined to be a submatrix of $\mathbf{A}$ obtained by selecting those columns of $\mathbf{A}$ with indexes in S. The smallest and largest singular values of $\mathbf{A}$ are shown as $\sigma_{\min}$ and $\sigma_{\max}$, respectively. The symbol $\|v\|_0$ shows the cardinality of a vector v, i.e., the number of its nonzero elements. Given a matrix $\mathbf{A}$, the symbols $\|\mathbf{A}\|_1$, $\|\mathbf{A}\|_\infty$, $\|\mathbf{A}\|_2$, and $\|\mathbf{A}\|_F$ denote the maximum absolute column sum, maximum absolute row sum, maximum singular value, and the Frobenius norm of $\mathbf{A}$, respectively. The cardinality of a set $\mathcal{M}$ is indicated as $|\mathcal{M}|$. The operator $\text{vec}(\cdot)$ vectorizes its matrix argument.

II. Problem Formulation

The Robust Quadratic Regression aims to find a vector $\mathbf{x} \in \mathbb{D}^n$ such that

$$y_r = \mathbf{x}^*\mathbf{M}_r\mathbf{x} + \eta_r, \quad \forall r \in \{1, \dots, m\}, \tag{1}$$

where

- $\mathbb{D}$ is either $\mathbb{R}$ or $\mathbb{C}$
- $y_1, \dots, y_m$ are some known real-valued measurements.
- $\eta_1, \dots, \eta_m$ are unknown and sparsely occurring real-valued noise with arbitrary magnitudes.
- $\mathbf{M}_1, \dots \mathbf{M}_m$ are some known $n \times n$ Hermitian matrices.

The regression problem could have two solutions $\pm\mathbf{x}$ in the real-valued case, which increases to infinitely many in the form of $\mathbf{x} \times e^{\sqrt{-1}\theta}$ in the complex case. To avoid this ambiguity, the objective is to find the matrix $\mathbf{x}\mathbf{x}^*$ rather than $\mathbf{x}$ since this matrix is invariant if $\mathbf{x}$ rotates. At the same time, recovery of $\mathbf{x}$ from $\mathbf{x}\mathbf{x}^*$ is a simple problem that can be solved with spectral decomposition. If m is large enough, then $\mathbf{x}\mathbf{x}^*$ is unique. This paper aims to recover any

solution $\mathbf{x}\mathbf{x}^*$ in case there are multiple ones. To develop the theoretical results of this paper, it is essential to ensure that the matrices $\mathbf{M}_1, \dots \mathbf{M}_m$ are somehow comparable. To achieve this, one may appropriately rescale each individual measurement equation to make the norm of the resulting constant matrix equal to 1. Therefore, with no loss of generality, assume that $\|\mathbf{M}_r\|_2 = 1$ for $r = 1, ..., m$. In the robust regression problem, the vector η is assumed to be sparse. To distinguish between error-free and erroneous measurements, we introduce a partition of the set of measurements into two subsets of *good* and *bad* measurements:

$$\mathcal{G} = \{r|\eta_r = 0\}, \quad \mathcal{B} = \{1, \dots, m\}\backslash\mathcal{G}$$

To streamline the derivation of the analytical results of this paper, we assume that $\mathcal{G} = \{1, \dots, |\mathcal{G}|\}$ and $\mathcal{B} = \{|\mathcal{G}| + 1, \dots, m\}$. However, the algorithms to be designed are completely oblivious to the type of each measurement and its membership in either $\mathcal{G}$ or $\mathcal{B}$. Define the function $\mathbf{F} : \mathbb{C}^n \to \mathbb{C}^m$ as follows:

$$\mathbf{F}(\mathbf{z}) = [\mathbf{z}^*\mathbf{M}_1\mathbf{z} \ \dots \ \mathbf{z}^*\mathbf{M}_m\mathbf{z}]^T,$$

Define also the Jacobian of the above function at a point $\mathbf{x}$ with respect to the coordinates of good measurements

$$\nabla_{\mathcal{G}}\mathbf{F}(\mathbf{z})\big|_{\mathbf{z}=\mathbf{x}} = \mathbf{J}_{\mathcal{G}} = 2\left[\mathbf{M}_1\mathbf{x} \ \dots \ \mathbf{M}_{|\mathcal{G}|}\mathbf{x}\right] \tag{2}$$

and with respect to the coordinates of bad measurements

$$\nabla_{\mathcal{B}}\mathbf{F}(\mathbf{z})\big|_{\mathbf{z}=\mathbf{x}} = \mathbf{J}_{\mathcal{B}} = 2\left[\mathbf{M}_{|\mathcal{G}|+1}\mathbf{x} \ \dots \ \mathbf{M}_m\mathbf{x}\right] \tag{3}$$

Likewise, let $\mathbf{J}$ be the Jacobian of $\mathbf{F}(\mathbf{x})$. The objective of this paper is to develop efficient algorithms to find $\mathbf{x}$ precisely as long as η is sufficiently sparse. This statement will be formalized in the next section.

III. Main Results

Consider a variable matrix $\mathbf{W}$ playing the role of $\mathbf{x}\mathbf{x}^*$. This matrix is positive semidefinite and has rank 1. By dropping the rank constraint, we can cast the quadratic regression problem as a linear matrix regression problem. Motivated by this relaxation, consider the optimization problem

$$\begin{aligned} \underset{\substack{\mathbf{W}\in\mathbb{D}^{n\times n},\ \nu\in\mathbb{R}^n,\\ \omega\in\mathbb{R}^n}}{\text{minimize}} \quad & \langle\mathbf{W}, \mathbf{M}\rangle + \mu_1\|\nu\|_1 + \tfrac{\mu_2}{2}\|\omega\|_2 \\ \text{s.t.} \quad & \langle\mathbf{W}, \mathbf{M}_r\rangle + \nu_r = y_r + \mu_2\omega_r, \ \forall r \in \{1, \dots, m\} \\ & \mathbf{W} = \mathbf{W}^* \succeq_{\mathcal{C}} 0 \\ & \|\nu\|_0 \le k \end{aligned} \tag{4}$$

where $(\mathbf{M}, \mu_1, \mu_2, k)$ is the set of hyperparameters of the problem and the notation $\succeq_{\mathcal{C}}$ is the generalized inequality sign with respect to $\mathcal{C}$, which is either the cone of Hermitian positive semidefinite (PSD) matrices or the second-order (SO) cone defined as the set of all Hermitian matrices that satisfy:

$$\begin{bmatrix} \mathbf{W}_{ii} & \mathbf{W}_{ij} \\ \mathbf{W}_{ji} & \mathbf{W}_{jj} \end{bmatrix} \succeq 0, \qquad \forall (i,j) \in \{1, \dots, n\}. \tag{5}$$

In what follows, we will analyze this problem for special values of the hyperparameters and different numbers of measurements.

A.i. Penalized Conic Relaxation

Suppose that $\hat{\mathbf{x}} \in \mathbb{D}^n$ is an initial guess for the solution of the quadratic regression, serving as *a priori* information about the unknown vector $\mathbf{x}$. Consider a Hermitian positive-semidefinite matrix $\mathbf{M} \in \mathbb{D}^{n\times n}$ with the following properties:

$$\begin{aligned} &\mathbf{M}\hat{\mathbf{x}} = \mathbf{0}, \\ &\lambda_1(\mathbf{M}) = \dots = \lambda_{n-1}(\mathbf{M}) = 1 \end{aligned}$$

There are infinitely many choices for $\mathbf{M}$. It results from this definition that $\|\mathbf{M}\mathbf{x}\|_\infty$ is a measure of the alignment of the initial guess $\hat{\mathbf{x}}$ and the exact solution $\mathbf{x}$. Hence, define the function

$$\rho(\hat{\mathbf{x}}, \mathbf{x}) := \|\mathbf{M}\mathbf{x}\|_\infty$$

which serves as a distance between $\mathbf{x}$ and $\hat{\mathbf{x}}$ and represents a priori knowledge about the solution of the problem. Its value will be important in the following theoretical constructions.

In the special case $(\mu_1, \mu_2, k) = (\mu, 0, m)$, we obtain a penalized conic programming relaxation of the problem (1):

$$\begin{aligned} \underset{\mathbf{W}\in\mathbb{D}^{n\times n},\ \nu\in\mathbb{R}^m}{\text{minimize}} \quad & \langle\mathbf{W}, \mathbf{M}\rangle + \mu\|\nu\|_1 \\ \text{s.t.} \quad & \langle\mathbf{W}, \mathbf{M}_r\rangle + \nu_r = y_r, \quad \forall r \in \{1 \dots m\} \\ & \mathbf{W} = \mathbf{W}^* \succeq_{\mathcal{C}} 0 \end{aligned} \tag{6}$$

Note that since no rank constraint is imposed on $\mathbf{W}$ and that a regularization term is included in the objective function. We refer to this problem as *penalized conic relaxation*. This is a convex problem and can be solved in polynomial time up to any given accuracy.

A.ii. Upper Bound on Cardinality of Bad Measurement Set

In this subsection, we establish a uniform bound on the number of bad measurements that the penalized conic relaxation can tolerate. To do so, we make use of two matrix properties defined in [15].

Definition 1 (SSC property): A matrix $\mathbf{X} \in \mathbb{C}^{n\times m}$ is said to satisfy the Subset Strong Convexity Property at level p with constant λ_p if

$$\lambda_p \le \min_{|S|=p} \sqrt{\lambda_{min}(\mathbf{X}_S\mathbf{X}_S^T)}$$

Definition 2 (SSS property): A matrix $\mathbf{X} \in \mathbb{C}^{n\times m}$ is said to satisfy the Subset Strong Smoothness Property at level p with constant Λ_p if

$$\max_{|S|=p} \sqrt{\lambda_{max}(\mathbf{X}_S\mathbf{X}_S^T)} \le \Lambda_p$$

The relationship between the constants λ_p and Λ_{m-p} can be interpreted as a uniform condition number at level p. By leveraging the properties of these constants, the first main result of this paper aims to find a bound on the permissible number of bad measurements.

Theorem 1: Assume that there exists a number μ^* such that

- Condition 1.1: $1 - \sqrt{n|B|}\frac{\Lambda_{|\mathcal{B}|}}{\lambda_{|\mathcal{G}|}} > 0$

- Condition 1.2: $\left(\frac{\lambda_{|\mathcal{G}|}}{4\rho(\hat{\mathbf{x}},\mathbf{x})\sqrt{n}} - \sqrt{n|\mathcal{G}|}\right) \cdot \frac{1-\sqrt{n|B|}\frac{\Lambda_{|\mathcal{B}|}}{\lambda_{|\mathcal{G}|}}}{1+\sqrt{n|G|}\frac{\Lambda_{|\mathcal{B}|}}{\lambda_{|\mathcal{G}|}}} > |\mathcal{B}|$
- Condition 1.3: $\mu^* > \frac{2\rho(\hat{\mathbf{x}},\mathbf{x})\sqrt{n}}{\lambda_{|\mathcal{G}|}-\sqrt{n|\mathcal{B}|}\Lambda_{\mathcal{B}}}$
- Condition 1.4: $\mu^* < \frac{\lambda_{|\mathcal{G}|}-4n\rho(\hat{\mathbf{x}},\mathbf{x})\sqrt{|\mathcal{G}|}}{2|\mathcal{B}|\left(\sqrt{n|\mathcal{G}|}\Lambda_{\mathcal{B}}+\lambda_{|\mathcal{G}|}\right)}$

Then, $(\mathbf{W},\nu) = (\mathbf{x}\mathbf{x}^*,\eta)$ is the unique solution of the penalized conic relaxation (6) with $\mathcal{C}$ chosen as the PSD cone and $\mu = \mu^*$.

Proof: The proof is provided in the Appendix. ■

Theorem 1 states that the nonconvex robust regression problem can be solved precisely, leading to the recovery of the unknown solution and the detection of bad measurements, provided that certain conditions are satisfied. These conditions depend on the notions of SSC and SSS.

Lemma 1: As long as Conditions 1.1 and 1.2 in Theorem 1 are satisfied, there exists a number μ^* for which all conditions of this theorem are met.

Proof: μ^* must belong to the interval

$$\left[\frac{2\rho(\hat{\mathbf{x}},\mathbf{x})\sqrt{n}}{\lambda_{|\mathcal{G}|}-\sqrt{n|\mathcal{B}|}\Lambda_{\mathcal{B}}}, \frac{\lambda_{|\mathcal{G}|}-4n\rho(\hat{\mathbf{x}},\mathbf{x})\sqrt{|\mathcal{G}|}}{2|\mathcal{B}|\left(\sqrt{n|\mathcal{G}|}\Lambda_{\mathcal{B}}+\lambda_{|\mathcal{G}|}\right)}\right]$$

It is straightforward to verify that the interval is not empty if Conditions 1.1 and 1.2 in Theorem 1 are met. ■

Condition 1.1 ensures that a term in Condition 1.2 is nonnegative. On the other hand, Lemma 1 states that Condition 1.2 is the most important requirement for the success of the penalized conic relaxation in solving quadratic regression. We want to emphasize here that the precise knowledge of the regularization parameter μ^* seems to be not necessary in practice. It will be shown in the Experiments section (IV) that the heuristically chosen value of 10^{-2} may work quite well.

To refine the result of Theorem 1, we will study the special case of Gaussian systems next.

A.iii. Upper Bound for Gaussian Systems

Without loss of generality, assume throughout this subsection that $\mathbb{D} = \mathbb{R}$.

Definition 3: The matrix $\mathbf{J}$ is called standard Gaussian over $\mathbb{R}$ if its entries are independent and identically distributed random variables with a standard normal distribution.

Theorem 2: Assume that $\mathbf{J}$ is standard Gaussian over $\mathbb{R}$ and that there exist numbers $\Delta > 0$, μ^*, $c = 24e^2 \log\frac{3}{\varepsilon}$ and $c' = 24e^2$ such that

- Condition 2.1: $\frac{\sqrt{cn+c'\log\frac{2}{\delta}}}{(1-2\varepsilon)\sqrt{|\mathcal{B}|}} < \Delta$
- Condition 2.2: $\alpha = \frac{8\rho(\hat{\mathbf{x}},\mathbf{x})n\sqrt{1+\Delta}}{1-\Delta-4\rho(\hat{\mathbf{x}},\mathbf{x})n\sqrt{1-\Delta}} > 0$
- Condition 2.3: $\sqrt{|\mathcal{G}|} > \alpha|\mathcal{B}|^{\frac{3}{2}} + \sqrt{n\frac{1+\Delta}{1-\Delta}}|\mathcal{B}|$
- Condition 2.4: $\mu^* > \frac{2\rho(\hat{\mathbf{x}},\mathbf{x})\sqrt{n}}{\sqrt{(1-\Delta)|\mathcal{G}|}-\sqrt{n(1+\Delta)|\mathcal{B}|}}$
- Condition 2.5: $\mu^* < \frac{\sqrt{(1-\Delta)|\mathcal{G}|}-4n\rho(\hat{\mathbf{x}},\mathbf{x})\sqrt{|\mathcal{G}|}}{2|\mathcal{B}|\sqrt{|\mathcal{G}|}\left(\sqrt{n(1+\Delta)|\mathcal{B}|}+\sqrt{(1-\Delta)}\right)}$,

Then, with probability at least $(1-\delta)^2$, the point $(\mathbf{W},\nu) = (\mathbf{x}\mathbf{x}^*,\eta)$ is the unique solution of the penalized conic relaxation (6) with $\mathcal{C}$ equal to the SDP cone and $\mu = \mu^*$.

Proof: The proof is provided in the Appendix. ■

Every individual good measurement may be treated as a bad measurement where its corresponding η_i approaches zero. Using this subtle technique, the set of bad measurements could be expanded from its original (true) set. Condition 2.1 of Theorem 2 requires the (expanded) set $|\mathcal{B}|$ to be sufficiently large, which implies that the total number of measurements should be high. This theorem is most effective when $m \geq n^2$. As a consequence of the strict law of large numbers, the minimal and maximal singular values of a standard Gaussian matrix are concentrated around the square root of its width (number of columns). Using this observation, it can be inferred from Theorem 2 that if $\mathbf{J}$ is a standard Gaussian matrix, then the penalized conic relaxation (6) with a PSD cone recovers the exact solution of the quadratic regression with a high probability in the regime where the number of measurements is sufficiently large, provided that (i) $|\mathcal{B}| = O(|\mathcal{G}|^{\frac{1}{3}})$, or (ii) $|\mathcal{B}| = O(|\mathcal{G}|^{\frac{1}{2}})$ and $\rho(\hat{\mathbf{x}},\mathbf{x}) \sim 0$. These two asymptotic bounds are obtained from Condition 2.3 because the other conditions of the theorem do not matter if the number of measurements is sufficiently large.

B. Robust Least-Squares Regression

Consider the optimization problem (4) with the parameter set $(\mathbf{M},\mu_1,\mu_2,k) = (\mathbf{0},0,1,k)$, which reduces to

$$\begin{array}{ll} \underset{\mathbf{W}\in\mathbb{D}^{n\times n},\ \nu\in\mathbb{R}^n}{\text{minimize}} & \frac{1}{2}\sum_{r=1}^{m}\left(\langle\mathbf{W},\mathbf{M}_r\rangle + \nu_r - y_r\right)^2 \\ \text{subject to} & \mathbf{W}\succeq_{\mathcal{C}} 0 \\ & \|\nu\|_0 \leq k \end{array} \tag{7}$$

This problem is nonconvex due to a cardinality constraint. With no loss of generality, assume that $\mathbb{D} = \mathbb{R}$ and that $k \leq m$.

Definition 4: Define $HT_k(\mathbf{y}) : \mathbb{R}^m \to \mathbb{R}^m$ to be a hard thresholding operator such that

$$[HT_k(\mathbf{z})]_i = \begin{cases} z_i, & \text{if } |z_i| \text{ is among the } k \text{ largest} \\ & \text{entries of } \mathbf{z} \text{ in magnitude} \\ 0, & \text{otherwise} \end{cases}$$

for every $\mathbf{z}\in\mathbb{R}^n$, where $[HT_k(\mathbf{z})]_i$ denotes the i^{th} entry of $[HT_k(\mathbf{z})]$.

Consider the function

$$f(\nu) := \min_{\mathbf{W}\succeq_{\mathcal{C}}0} \frac{1}{2}\sum_{r=1}^{m}\left(\langle\mathbf{W},\mathbf{M}_r\rangle - (y_r-\nu_r)\right)^2$$

and let $\hat{\mathbf{W}}(\nu)$ denote a solution to this problem. The Hard Thresholding approach to be proposed for solving the quadratic regression problem consists of the iterative scheme

$$\nu^{t+1} = HT_k(\nu^t - \mathbf{d}(\nu^t)) \tag{8}$$

where

$$\mathbf{d}(\nu) = \frac{1}{2}\nabla_\nu\left(\sum_{r=1}^{m}\left(\langle\mathbf{W},\mathbf{M}_r\rangle - (y_r-\nu_r)\right)^2\right)\Big|_{\mathbf{W}=\hat{\mathbf{W}}(\nu)}$$

(the symbol ∇_ν denotes the gradient with respect to ν). By Lemma 3.3.1 in [30], if $\hat{\mathbf{W}}(\nu)$ is a continuously differentiable mapping, then $\nabla f(\nu) = \mathbf{d}(\nu)$. Inspired by this fact, one may informally regard $\mathbf{d}(\nu)$ as the gradient of the objective of the optimization problem (7) without its cardinality constraint and after fixing its variable ν. Define $\mathbf{w} = \text{vec}(\mathbf{W})$, $\hat{\mathbf{w}}(\nu) = \text{vec}(\hat{\mathbf{W}}(\nu))$, $\mathbf{a}_r = \text{vec}(\mathbf{M}_r)$ for $r = 1, ..., m$, and $\mathbf{A} = [\mathbf{a}_1 \ \dots \ \mathbf{a}_m]^T$. It can be verified that

$$\mathbf{d}(\nu) = \mathbf{A}\hat{\mathbf{w}}(\nu) - \mathbf{y} + \nu$$

which implies that

$$HT_k(\nu - \mathbf{d}(\nu)) = HT_k(\mathbf{y} - \mathbf{A} \cdot \text{vec}(\hat{\mathbf{W}}(\nu)))$$

Based on this formula, we propose a conic hard thresholding method in Algorithm 1.

Algorithm 1 Conic Hard Thresholding

Input: Covariates $\mathbf{A}$, responses $\mathbf{y}$, corruption index k, tolerance ε, and cone $\mathcal{C}$

Initialization :

$\nu^0 \leftarrow \mathbf{0}$, $t \leftarrow 0$;

LOOP Process

while $\|\nu^t - \nu^{t-1}\| > \varepsilon$ **do**

  $\hat{\mathbf{W}}^t = \arg\min_{\mathbf{W} \succeq_{\mathcal{C}} \mathbf{0}} \sum_{r=1}^{m} \left(\langle \mathbf{W}, \mathbf{M}_r \rangle - (y_r - \nu_r^t)\right)^2$;

  $\nu^{t+1} = HT_k(\mathbf{y} - \mathbf{A} \cdot \text{vec}(\hat{\mathbf{W}}^t))$;

  $t \leftarrow t + 1$;

end while

return $\hat{\mathbf{W}}^{t+1}$

Unlike the penalized conic relaxation, Algorithm 1 solves a sequence of conic programs to identify the set of bad measurements through a thresholding technique. In the regime where $m \geq n^2$, this algorithm with a high computational complexity can be further relaxed by letting the cone $\mathcal{C}$ to be the set of symmetric matrices. We refer to this as **Algorithm 2**, where the condition $\mathbf{W} \succeq_{\mathcal{C}} \mathbf{0}$ is reduced to $\mathbf{W} = \mathbf{W}^*$. Note that Algorithm 2 is not effective if $m < n(n+1)/2$ because the number of measurements becomes less than the number of scalar variables in $\mathbf{W}$. On the other hand, as m grows, the feasibility constraint $\mathbf{W} \succeq_{\mathcal{C}} \mathbf{0}$ becomes almost redundant and Algorithm 1 performs similarly to Algorithm 2. Inspired by this property, we analyze the asymptotic behavior of Algorithm 2 for Gaussian systems below.

Theorem 3: Suppose that $|\mathcal{B}| < \frac{m}{20000}$, $m \geq n^2$, $\mathbf{M}_r$ is a random normal Gaussian matrix for $r = 1, ..., m$, and there is Gaussian additive noise with variance σ^2. For every $\epsilon, \delta > 0$, Algorithm 2 recovers a matrix $\mathbf{W}$ such that $\|\mathbf{W} - \mathbf{x}\mathbf{x}^T\|_2 \leq \epsilon + \mathcal{O}(\sigma\sqrt{\frac{n}{m}\log\frac{nm}{\delta}})$ within $\mathcal{O}(\log(\frac{\|\eta\|_2}{\epsilon}) + \log(\frac{2m}{n^2+n}))$ operations with probability $1 - \delta$.

For every $\epsilon > 0$, Algorithm 2 recovers a matrix $\mathbf{W}$ such that $\|\mathbf{W} - \mathbf{x}\mathbf{x}^*\|_2 \leq \epsilon$ within $\mathcal{O}(\log(\frac{\|\eta\|_2}{\epsilon}) + \log(\frac{2m}{n^2+n}))$ operations.

Proof: The proof is provided in the Appendix. ∎

Let $\mathbf{W}$ be a solution found by Algorithm 2. Then, one can use its eigenvalue decomposition to find a vector $\mathbf{u}$ such that $\mathbf{u} = \arg\min_{\mathbf{v} \in \mathbb{C}^n} \|\mathbf{v}\mathbf{v}^* - \mathbf{W}\|_2$. Therefore,

$$\begin{aligned} \|\mathbf{u}\mathbf{u}^* - \mathbf{x}\mathbf{x}^*\|_2 &= \|(\mathbf{u}\mathbf{u}^* - \mathbf{W}) - (\mathbf{x}\mathbf{x}^* - \mathbf{W})\|_2 \\ &\leq \|\mathbf{u}\mathbf{u}^* - \mathbf{W}\|_2 + \|\mathbf{x}\mathbf{x}^* - \mathbf{W}\|_2 \leq 2\varepsilon \end{aligned} \tag{9}$$

This means that Algorithm 2 can be used to find an approximate solution $\mathbf{u}$ with any arbitrary precision for the robust regression problem for Gaussian systems with a large number of measurements and yet it allows up to a constant fraction of measurements to be completely wrong (i.e., $O(|\mathcal{B}|) = O(|\mathcal{G}|)$). Comparing this with the guarantee $O(|\mathcal{B}|) = O(|\mathcal{G}|^{\frac{1}{3}})$ for the penalized conic optimization, it can be concluded that Algorithm 1 (or 2) is more robust to outliers than the penalized conic program since it solves a sequence of optimization problems iteratively as opposed to a single one. This leads to a tradeoff between the complexity of an estimation method and its robustness level.

The theoretical analyses of this work were all on a regression model subject to a sparse error vector. However, the results can be slightly modified to account for modest noise values in addition to sparse errors. The bounds derived in this work remain the same, but the solutions found by the penalized conic relaxation and Algorithm 1 would no longer match the true regression solution being sought (as expected, due to a corruption in all equations). The mismatch error is a function of the modest noise values. The details are omitted for brevity; however, the result will later be demonstrated in numerical examples.

IV. Experiments

In this section, we study the numerical properties of the penalized conic relaxation (6) and the conic hard thresholding Algorithm 1.

A. Synthetic Data

Following [31], we define the sparsity pattern of an arbitrary matrix $\mathbf{X} \in \mathbb{H}^n$ as a binary matrix $\mathbf{N} \in \mathbb{S}^n$ whose (i,j)-entry is equal to 1 if and only if $X_{ij} \neq 0$. Define the set

$$\mathcal{S}(\mathbf{N}) \triangleq \{\mathbf{X} \in \mathbb{H}^n | \mathbf{X} \circ \mathbf{N} = \mathbf{X}\}$$

. We conduct experiments on synthetically generated quadratic regression datasets with corruptions. The true model vector $\mathbf{x}$ is chosen to be a random unit norm vector, while the input matrices $\mathbf{M}_r$'s are chosen from $\mathcal{S}(\mathbf{N})$ according to a common random sparsity pattern $\mathbf{N}$. The nonzero entries of $\mathbf{M}_r$'s are generated from normal standard distribution. The matrix $\mathbf{N}$ is formed by including all diagonal elements and $3n$ off-diagonal elements. The off-diagonal positions are picked uniformly. The responses to be corrupted are chosen uniformly at random and the value of each corruption is generated uniformly from the interval $[10, 20]$. Responses are then generated as $y_r = \mathbf{x}^*\mathbf{M}_r\mathbf{x} + \eta_r + \omega_r$, where in addition to the sparse error vector η we have a random dense noise vector ω whose entries are Gaussian with

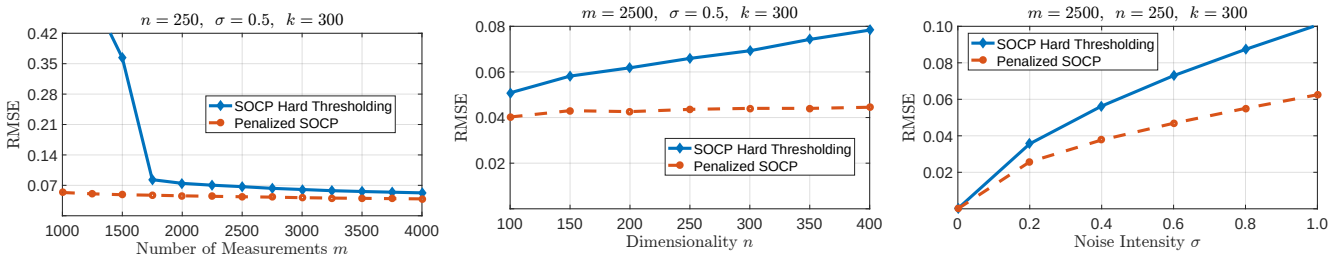

Fig. 1. Estimation error as a function of: (a) the number of data points m, (b) the dimensionality n, and (c) the additive white noise variance σ.

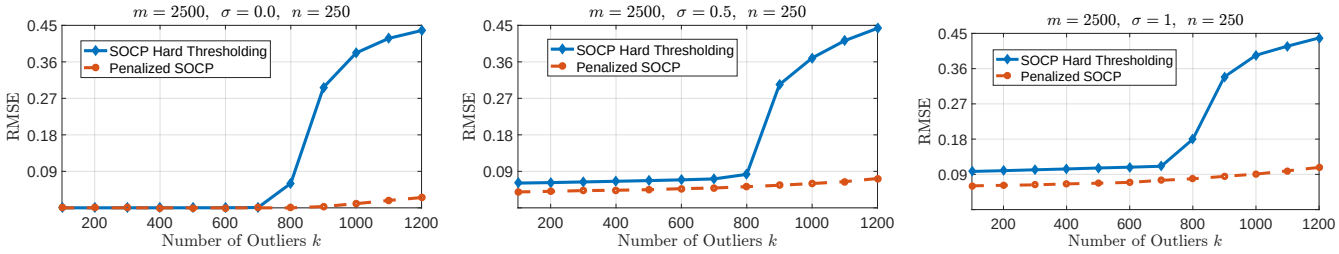

Fig. 2. Estimation error as a function of the number of bad measurements k for different magnitudes of additive dense Gaussian noise.

zero mean and variance σ. All reported results are averaged over 10 random trials.

By assuming that no prior information about the solution $\mathbf{x}$ is available, we set the matrix $\mathbf{M}$ to be equal to $\mathbf{I}_n$ in the penalized conic relaxation with the parameter μ chosen as 10^{-2}. Regarding Algorithm 1, the parameter k is selected as the true number of corrupted measurements, the tolerance ε is set to 10^{-3}, and the algorithm is terminated early if the number of conic iterations exceeds 50. In both of the methods, $\mathcal{C}$ is considered to be the SO cone. Hence, we refer to these methods as penalized SOCP and SOCP hard thresholding. Due to the sparsity in the data, the SOCP formulation can be simplified by only imposing those 2×2 constraints in (5) that correspond to the members of $\{(i,j) \mid N_{ij} = 1\}$.

We measure the performance of each algorithm using the root mean squared error (RMSE) defined as $\frac{\|\hat{\mathbf{x}}-\mathbf{x}\|_2}{\sqrt{n}}$. Figure 1 shows the RSME in three different plots as a function of the number of data points m, the dimensionality n, and the additive white noise variance σ. Figure 2 depicts the RSME as a function of the number of bad measurements k for different magnitudes of additive dense Gaussian noise. It can be observed that both the penalized conic relaxation and the conic hard thresholding algorithm exhibit an exact recovery property for systems with up to 700 randomly corrupted measurements out of 2500 measurements in the absence of dense Gaussian noise. The same behavior is observed in the presence of dense Gaussian noise of different magnitudes: the error of the penalized SOCP solution grows gradually, while the error of the the hard thresholding algorithm has a jump at around 800 bad measurements. These simulations support the statement that up to a constant fraction of measurements could be completely wrong, and yet the unknown regression solution is found precisely.

Although the theoretical analyses provided in this paper favor Algorithm 1 over the penalized conic relaxation, our empirical analysis shows that the penalized SOCP method has a better performance than the hard thresholding algorithm uniformly in the number of measurements, dimensionality, noise magnitude and the number of outliers. To explain this observation, note that the derived theoretical bounds correspond to the worst-case scenario and are more conservative for an average scenario. Moreover, the implementation of Algorithm 1 in this section has limited the number of iterations to 50, while Theorem 3 requires the number of iterations to grow with respect to the amount of corruption.

The results of this part are produced using the standard MOSEK v7. SOCP-solving procedure, run in MATLAB on a 12-core 2.2GHz machine with 256GB RAM. The CPU time for each round of solving SOCP ranges from 3s (for $n = 250$, $m = 2500$) to 30s (for $n = 400$, $m = 2500$).

B. State Estimation for Power Systems

In this subsection, we present empirical results for the penalized conic relaxation with a PSD cone $\mathcal{C}$ tested on the real data for the power flow state estimation with outliers. As discussed in [2], this problem can be formulated as robust quadratic regression. The experiment is run on the PEGASE 1354-bus European system borrowed from the MATPOWER package [32], [33]. This system has 1354 nodes and the objective is to estimate the nodal voltages based on voltage magnitude and power measurements of the form $y_r = \mathbf{x}^*\mathbf{M}_r\mathbf{x} + \eta_r + \omega_r$, where ω is a dense additive noise whose r^{th} entry is Gaussian with mean zero and the standard deviation equal to σ times the true value of the corresponding voltage/power parameter. The dimension of the complex vector $\mathbf{x}$ is 1354, which leads to 2708 real

variables in the problem. In this model, the measurements are voltage magnitude squares, active and reactive nodal power injections, and active and reactive power flows from both sides of every line of the power system. This amounts to $3n + 4l = 12026$ measurements, where $l = 1991$ denotes the number of lines in the system. Note that the quadratic regression problem is complex-valued in this case.

The penalty parameter μ of the penalized conic relaxation is set to 10^2 and the matrix $\mathbf{M}$ is chosen as $-\mathbf{Y} + \gamma\mathbf{I}$, where $\mathbf{Y}$ is the susceptance matrix of the system and γ is the smallest positive number that makes $\mathbf{M}$ positive semidefinite. Since the penalized SDP relaxation is large-scale, we employ a tree decomposition technique to leverage the sparsity of the problem to solve it more efficiently [34]. The width of the tree decomposition used to reduce the complexity is equal to 12. We do not report any results on Algorithm 1 because it requires solving large-scale SDPs successively and this could be time-consuming. Moreover, the number of measurements is not high enough to use Algorithm 2, so we can't use it too.

The numerical results are reported in Figure 3. Remarkably, if the dense Gaussian noise is non-existent, the conic relaxation recovers the solution precisely as long as the number of bad measurements is less than 150 (note that $\sqrt{m} \simeq 109$). Note that power systems are sparse and their models are far from Gaussian, but the results of Theorem 2 are still valid in this numerical example.

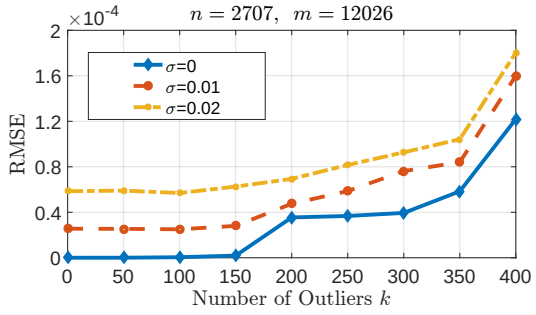

Fig. 3. This plot shows the RMSE with respect to the number of corrupted measurements k for the PEGASE 1354-bus system.

V. Conclusion

This paper is concerned with the robust quadratic regression problem, where the goal is to find the unknown parameters (state) of the system modeled by nonconvex quadratic equations based on observational data subject to sparse errors of arbitrary magnitudes. Two methods are developed in this paper, which rely on conic optimization. The first approach is a single optimization problem that includes a regularizer term in the objective function to cope with the sparse noise, whereas the second method is an iterative algorithm that requires solving a conic optimization at every iteration and performing a hard thresholding task. A deterministic bound is derived for the first method, named penalized conic relaxation, which quantifies how many bad measurement the algorithm can tolerate and yet recover the correct solution. This bound is further refined for Gaussian systems, and it is shown that up to a square root of the total number of measurements could be grossly erroneous without compromising the quality of the recovered solution. In the case where the number of measurements is sufficiently large, it is shown that the second algorithm allows up to a constant number of equations to be arbitrarily wrong for Gaussian systems. The results of this paper are demonstrated on synthetic data. In addition, a case study is provided on a European power grid to verify that the proposed technique can correctly identify the state of the system even if $O(\sqrt{m})$ measurements are completely wrong.

References

[1] A. Abur and A. G. Exposito, *Power system state estimation: theory and implementation*. CRC press, 2004.

[2] R. Madani, J. Lavaei, R. Baldick, and A. Atamtürk, "Power system state estimation and bad data detection by means of conic relaxation," in *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2017.

[3] Y. Zhang, R. Madani, and J. Lavaei, "Conic relaxations for power system state estimation with line measurements," *IEEE Transactions on Control of Network Systems*, 2017.

[4] M. Jin, J. Lavaei, and K. Johansson, "A semidefinite programming relaxation under false data injection attacks against power grid AC state estimation," in *55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2017, pp. 236–243.

[5] P. J. Rousseeuw and A. M. Leroy, *Robust regression and outlier detection*. John wiley & sons, 2005, vol. 589.

[6] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis," *Journal of the ACM (JACM)*, vol. 58, no. 3, p. 11, 2011.

[7] N. M. Nasrabadi, T. D. Tran, and N. Nguyen, "Robust lasso with missing and grossly corrupted observations," in *Advances in Neural Information Processing Systems*, 2011, pp. 1881–1889.

[8] K. Bhatia, P. Jain, and P. Kar, "Robust regression via hard thresholding," in *Advances in Neural Information Processing Systems*, 2015, pp. 721–729.

[9] H. Zhang, Y. Chi, and Y. Liang, "Provable non-convex phase retrieval with outliers: Median truncatedwirtinger flow," in *International conference on machine learning*, 2016, pp. 1022–1031.

[10] O. Klopp, K. Lounici, and A. B. Tsybakov, "Robust matrix completion," *Probability Theory and Related Fields*, vol. 169, no. 1-2, pp. 523–564, 2017.

[11] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE transactions on information theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[12] J. Wright and Y. Ma, "Dense error correction via l_1-minimization," *IEEE Transactions on Information Theory*, vol. 56, no. 7, pp. 3540–3560, 2010.

[13] C. Studer, P. Kuppinger, G. Pope, and H. Bolcskei, "Recovery of sparsely corrupted signals," *IEEE Transactions on Information Theory*, vol. 58, no. 5, pp. 3115–3130, 2012.

[14] Y. Chen, C. Caramanis, and S. Mannor, "Robust sparse regression under adversarial corruption," in *International Conference on Machine Learning*, 2013, pp. 774–782.

[15] K. Bhatia, P. Jain, P. Kamalaruban, and P. Kar, "Consistent robust regression," in *Advances in Neural Information Processing Systems*, 2017, pp. 2107–2116.

[16] P. Hand and V. Voroninski, "Corruption robust phase retrieval via linear programming," *arXiv preprint arXiv:1612.03547*, 2016.

[17] J. Chen, L. Wang, X. Zhang, and Q. Gu, "Robust wirtinger flow for phase retrieval with arbitrary corruption," *arXiv preprint arXiv:1704.06256*, 2017.

[18] S. Sojoudi, R. Madani, G. Fazelnia, and J. Lavaei, "Graph-theoretic algorithms for polynomial optimization problems," in *IEEE 53rd Conference on Decision and Control*. IEEE, 2014, pp. 2257–2271.

[19] H. M. Merrill and F. C. Schweppe, "Bad data suppression in power system static state estimation," *IEEE Transactions on Power Apparatus and Systems*, no. 6, pp. 2718–2725, 1971.

[20] D. Deka, R. Baldick, and S. Vishwanath, "Optimal data attacks on power grids: Leveraging detection & measurement jamming," in *Smart Grid Communications (SmartGridComm), 2015 IEEE International Conference on*. IEEE, 2015, pp. 392–397.
[21] Y. Weng, M. D. Ilić, Q. Li, and R. Negi, "Convexification of bad data and topology error detection and identification problems in ac electric power systems," *IET Generation, Transmission & Distribution*, vol. 9, no. 16, pp. 2760–2767, 2015.
[22] N. H. Nguyen and T. D. Tran, "Exact recoverability from dense corrupted observations via l_1-minimization," *IEEE transactions on information theory*, vol. 59, no. 4, pp. 2017–2035, 2013.
[23] B. McWilliams, G. Krummenacher, M. Lucic, and J. M. Buhmann, "Fast and robust least squares estimation in corrupted linear models," in *Advances in Neural Information Processing Systems*, 2014, pp. 415–423.
[24] Y. Chen, A. Jalali, S. Sanghavi, and C. Caramanis, "Low-rank matrix recovery from errors and erasures," *IEEE Transactions on Information Theory*, vol. 59, no. 7, pp. 4324–4337, 2013.
[25] J.-K. Wang *et al.*, "Robust inverse covariance estimation under noisy measurements," in *International Conference on Machine Learning*, 2014, pp. 928–936.
[26] H. Xu, C. Caramanis, and S. Mannor, "Robust regression and lasso," in *Advances in Neural Information Processing Systems*, 2009, pp. 1801–1808.
[27] W. Yang and H. Xu, "A unified robust regression model for lasso-like algorithms," in *International Conference on Machine Learning*, 2013, pp. 585–593.
[28] A. Dalalyan and Y. Chen, "Fused sparsity and robust estimation for linear models with unknown variance," in *Advances in Neural Information Processing Systems*, 2012, pp. 1259–1267.
[29] J. Á. Víšek, "The least trimmed squares. part i: Consistency," *Kybernetika*, vol. 42, no. 1, pp. 1–36, 2006.
[30] D. P. Bertsekas, D. P. Bertsekas, D. P. Bertsekas, and D. P. Bertsekas, *Dynamic programming and optimal control*. Athena scientific Belmont, MA, 1995, vol. 1, no. 2.
[31] R. Madani, A. Kalbat, and J. Lavaei, "A low-complexity parallelizable numerical algorithm for sparse semidefinite programming," *IEEE Transactions on Control of Network Systems*, 2017.
[32] C. Josz, S. Fliscounakis, J. Maeght, and P. Panciatici, "AC power flow data in MATPOWER and QCQP format: iTesla, RTE snapshots, and PEGASE," *arXiv preprint arXiv:1603.01533*, 2016.
[33] S. Fliscounakis, P. Panciatici, F. Capitanescu, and L. Wehenkel, "Contingency ranking with respect to overloads in very large power systems taking into account uncertainty, preventive, and corrective actions," *IEEE Transactions on Power Systems*, vol. 28, no. 4, pp. 4909–4917, 2013.
[34] R. Madani, M. Ashraphijuo, and J. Lavaei, "Promises of conic relaxation for contingency-constrained optimal power flow problem," *IEEE Transactions on Power Systems*, vol. 31, no. 2, pp. 1297–1307, 2016.
[35] I. Dokmanić and R. Gribonval, "Beyond moore-penrose Part I: Generalized inverses that minimize matrix norms," *arXiv preprint arXiv:1706.08349*, 2017.

APPENDIX

Lemma 2: Consider a full row-rank matrix $\mathbf{A} \in \mathbb{C}^{n\times m}$ and its Moore-Penrose pseudoinverse $\mathbf{A}^+ \in \mathbb{C}^{m\times n}$. The following inequalities hold:

$$\|\mathbf{A}\|_\infty \le \sqrt{m}\sigma_{max}(\mathbf{A}), \quad \|\mathbf{A}^+\|_\infty \le \sqrt{n}\frac{1}{\sigma_{min}(\mathbf{A})},$$
$$\|\mathbf{A}\|_1 \le \sqrt{n}\sigma_{max}(\mathbf{A}), \quad \|\mathbf{A}^+\|_1 \le \sqrt{m}\frac{1}{\sigma_{min}(\mathbf{A})}$$

Proof: Using the singular value decomposition, the matrix $\mathbf{A}$ can be written as

$$\mathbf{A} = \mathbf{U}\begin{bmatrix}\Sigma & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{V}_1^* \\ \mathbf{V}_2^*\end{bmatrix}$$

and therefore,

$$A^+ = \begin{bmatrix}\mathbf{V}_1 & \mathbf{V}_2\end{bmatrix}\begin{bmatrix}\Sigma^{-1} \\ \mathbf{0}\end{bmatrix}\mathbf{U}^*$$

Consider the following inequalities:

$$\begin{aligned}&\|\mathbf{A}\|_\infty \le \sqrt{m}\|\mathbf{A}\|_2, \qquad \|\mathbf{A}\|_1 \le \sqrt{n}\|\mathbf{A}\|_2\\ &\|\mathbf{A}^+\|_\infty \le \sqrt{n}\|\mathbf{A}^+\|_2, \quad \|\mathbf{A}^+\|_1 \le \sqrt{m}\|\mathbf{A}^+\|_2\end{aligned} \tag{10}$$

(please refer to Section 2.3 in [35] for more details). Now, one can use the fact that unitary transformations preserve the 2-norm:

$$\begin{aligned}&\|\mathbf{A}\|_2 = \|\mathbf{U}\Sigma\mathbf{V}_1^*\|_2 = \|\Sigma\|_2 = \sigma_{\max}(\mathbf{A})\\ &\|\mathbf{A}^+\|_2 = \|\mathbf{V}_1\Sigma^{-1}\mathbf{U}^*\|_2 = \|\Sigma^{-1}\|_2 = \frac{1}{\sigma_{\min}(\mathbf{A})}\end{aligned} \tag{11}$$

The proof follows from the above equations. ■

Lemma 3: If the conditions

i) $1 - \sqrt{n|B|}\frac{\sigma_{\max}(\mathbf{J}_\mathcal{B})}{\sigma_{\min}(\mathbf{J}_\mathcal{G})} > 0$

ii) $\mu^* > \frac{2\rho(\hat{\mathbf{x}},\mathbf{x})\sqrt{n}}{\sigma_{\min}(\mathbf{J}_\mathcal{G}) - \sqrt{n|\mathcal{B}|}\sigma_{\max}(\mathbf{J}_\mathcal{B})}$

iii) $\mu^* < \frac{\sigma_{\min}(\mathbf{J}_\mathcal{G}) - 4n\rho(\hat{\mathbf{x}},\mathbf{x})\sqrt{|\mathcal{G}|}}{2|\mathcal{B}|\left(\sqrt{n|\mathcal{G}|}\sigma_{\max}(\mathbf{J}_\mathcal{B}) + \sigma_{min}(\mathbf{J}_\mathcal{G})\right)}$

iv) $|\mathcal{B}| < \left(\frac{\sigma_{\min}(\mathbf{J}_\mathcal{G})}{4\rho(\hat{\mathbf{x}},\mathbf{x})\sqrt{n}} - \sqrt{n|\mathcal{G}|}\right)\cdot\frac{1-\sqrt{n|B|}\frac{\sigma_{\max}(\mathbf{J}_\mathcal{B})}{\sigma_{\min}(\mathbf{J}_\mathcal{G})}}{1+\sqrt{n|G|}\frac{\sigma_{\max}(\mathbf{J}_\mathcal{B})}{\sigma_{\min}(\mathbf{J}_\mathcal{G})}}$

are satisfied, then $(\mathbf{W},\nu) = (\mathbf{x}\mathbf{x}^*,\eta)$ is the unique solution of the penalized conic relaxation (6) with $\mathcal{C}$ equal to the PSD cone and $\mu = \mu^*$.

Proof: It follows from [2] that $(\mathbf{W},\nu) = (\mathbf{x}\mathbf{x}^*,\eta)$ is the unique solution of the penalized conic relaxation (6) based on a dual certificate if the followings conditions are all satisfied:

$$1 - \|\mathbf{J}_\mathcal{G}^+\mathbf{J}_\mathcal{B}\|_\infty > 0 \tag{12a}$$
$$\mu > \frac{2\|\mathbf{J}_\mathcal{G}^+\mathbf{M}\mathbf{x}\|_\infty}{1 - \|\mathbf{J}_\mathcal{G}^+\mathbf{J}_\mathcal{B}\|_\infty} \tag{12b}$$
$$\mu < \frac{1 - 4\|\mathbf{J}_\mathcal{G}^+\mathbf{M}\mathbf{x}\|_1}{2\left(\|\mathbf{J}_\mathcal{G}^+\mathbf{J}_\mathcal{B}\|_1 + 1\right)|\mathcal{B}|} \tag{12c}$$
$$|\mathcal{B}| < \frac{1 - 4\|\mathbf{J}_\mathcal{G}^+\mathbf{M}\mathbf{x}\|_1}{\|\mathbf{J}_\mathcal{G}^+\mathbf{J}_\mathcal{B}\|_1 + 1}\cdot\frac{1 - \|\mathbf{J}_\mathcal{G}^+\mathbf{J}_\mathcal{B}\|_\infty}{4\|\mathbf{J}_\mathcal{G}^+\mathbf{M}\mathbf{x}\|_\infty} \tag{12d}$$

Using the inequality $\|AB\| \le \|A\|\|B\|$, Lemma 2, and certain algebraic transformations, the above conditions could be relaxed to those stated in this lemma. ■

Proof of Theorem 1: The proof follows directly from Lemma 3 and the notions of SSC and SSS introduced in Definitions 1 and 2. ■

Proof of Theorem 2: In light of Lemma 14 in [8], any randomly sampled Gaussian matrix $\mathbf{X} \in \mathbb{R}^{n\times m}$ satisfies the inequalities

$$\lambda_{max}(\mathbf{X}\mathbf{X}^T) \le m + (1-2\varepsilon)^{-1}\sqrt{cmn + c'm\log\tfrac{2}{\delta}}$$
$$\lambda_{min}(\mathbf{X}\mathbf{X}^T) \ge m - (1-2\varepsilon)^{-1}\sqrt{cmn + c'm\log\tfrac{2}{\delta}}$$

with probability at least $1-\delta$ for every $\varepsilon > 0$, where $c = 24e^2\log\frac{3}{\varepsilon}$ and $c' = 24e^2$. This implies that the relations

$$\sigma_{min}(\mathbf{J}_\mathcal{G}) \in [\sqrt{|\mathcal{G}|(1-\Delta)}, \sqrt{|\mathcal{G}|(1+\Delta)}]$$

and

$$\sigma_{max}(\mathbf{J}_\mathcal{B}) \in [\sqrt{|\mathcal{B}|(1-\Delta)}, \sqrt{|\mathcal{B}|(1+\Delta)}]$$

are each satisfied with the probability $1-\delta$, and both are met simultaneously with probability at least $(1-\delta)^2$. As a result, Conditions (i), (iii) and (iv) in Lemma 3 hold with high probability. To analyze Condition (ii), one can write:

$$\left(\frac{\sigma_{\min}(\mathbf{J}_{\mathcal{G}})}{4\rho(\hat{\mathbf{x}},\mathbf{x})\sqrt{n}}-\sqrt{n|\mathcal{G}|}\right)\cdot\frac{1-\sqrt{n|B|}\frac{\sigma_{\max}(\mathbf{J}_{\mathcal{B}})}{\sigma_{\min}(\mathbf{J}_{\mathcal{G}})}}{1+\sqrt{n|G|}\frac{\sigma_{\max}(\mathbf{J}_{\mathcal{B}})}{\sigma_{\min}(\mathbf{J}_{\mathcal{G}})}}\geq$$

$$\left(\frac{\sqrt{|\mathcal{G}|(1-\Delta)}}{4\rho(\hat{\mathbf{x}},\mathbf{x})\sqrt{n}}-\sqrt{n|\mathcal{G}|}\right)\cdot\frac{1-\sqrt{n\frac{1+\Delta}{1-\Delta}}\frac{|\mathcal{B}|}{\sqrt{|\mathcal{G}|}}}{1+\sqrt{n|\mathcal{B}|\frac{1+\Delta}{1-\Delta}}}\geq$$

$$\sqrt{|\mathcal{G}|}\left(\frac{(1-\Delta)}{4\rho(\hat{\mathbf{x}},\mathbf{x})n\sqrt{1+\Delta}}-\sqrt{\frac{1-\Delta}{1+\Delta}}\right)\cdot\frac{1-\sqrt{n\frac{1+\Delta}{1-\Delta}}\frac{|\mathcal{B}|}{\sqrt{|\mathcal{G}|}}}{2\sqrt{|\mathcal{B}|}}$$

Since

$$\sqrt{|\mathcal{G}|}\left(1-\sqrt{n\frac{1+\Delta}{1-\Delta}}\frac{|\mathcal{B}|}{\sqrt{|\mathcal{G}|}}\right)>\alpha|\mathcal{B}|^{\frac{3}{2}},$$

Condition (ii) also holds true, and then the proof follows from Lemma 3. ■

Proof of Theorem 3: Define

$$\begin{aligned}&\tilde{\mathbf{w}}=\text{UPvec}(\mathbf{W})\\&\tilde{\mathbf{a}}=\text{UPvec}(\mathbf{M})\\&\tilde{\mathbf{a}}_{\mathbf{r}}=\text{UPvec}(\mathbf{M}_r),\qquad r=1,...,m\\&\tilde{\mathbf{A}}=[\tilde{\mathbf{a}}_1\ \ldots\ \tilde{\mathbf{a}}_m]^T\end{aligned}\tag{13}$$

Where UPvec$(\cdot)$ is an operator that returns a vector composed of the components of the upper triangle (including diagonal) of its matrix argument. The dimension of the resulting vector is $\frac{n^2+n}{2}$. In light of (1), the unknown vector $\mathbf{w}=\text{UPvec}(\mathbf{x}\mathbf{x}^*)$ satisfies the linear regression problem

$$\mathbf{y}=\mathbf{A}\mathbf{w}+\eta\tag{14}$$

Moreover, $\mathbf{a}_r$ is a random vector with a Gaussian probability distribution for every $r\in\{1,...,m\}$. It follows from [15] that Algorithm 2 recovers $\mathbf{w}$ correctly. This implies that the matrix $\mathbf{x}\mathbf{x}^*$ is found using this algorithm. ■