Analysis of Sparse Regularization Based Robust Regression Approaches

Kaushik Mitra, *Member, IEEE*, Ashok Veeraraghavan, *Member, IEEE*, and Rama Chellappa, *Fellow, IEEE*

***Abstract*—Regression in the presence of outliers is an inherently combinatorial problem. However, compressive sensing theory suggests that certain combinatorial optimization problems can be exactly solved using polynomial-time algorithms. Motivated by this connection, several research groups have proposed polynomial-time algorithms for robust regression. In this paper we specifically address the traditional robust regression problem, where the number of observations is more than the number of unknown regression parameters and the structure of the regressor matrix is defined by the training dataset (and hence it may not satisfy properties such as Restricted Isometry Property or incoherence). We derive the precise conditions under which the sparse regularization (l_0 and l_1-norm) approaches solve the robust regression problem. We show that the smallest principal angle between the regressor subspace and all k-dimensional outlier subspaces is the fundamental quantity that determines the performance of these algorithms. In terms of this angle we provide an estimate of the number of outliers the sparse regularization based approaches can handle. We then empirically evaluate the sparse (l_1-norm) regularization approach against other traditional robust regression algorithms to identify accurate and efficient algorithms for high-dimensional regression problems.**

***Index Terms*—Compressive sensing, robust regression, sparse representation.**

I. Introduction

THE goal of regression is to infer a functional relationship between two sets of variables from a given training data set. Many times the functional form is already known and the parameters of the function are estimated from the training data set. In most of the training data sets, there are some data points which differ markedly from the rest of the data; these are known as outliers. The goal of robust regression techniques is to properly account for the outliers while estimating the model parameters. Since any subset of the data could be outliers, robust regression is in general a combinatorial problem and robust algorithms such as *least median squares* (LMedS) [19] and *random sample consensus* (RANSAC) [10] inherit this combinatorial nature. However, compressive sensing theory [3], [7] has shown that certain combinatorial optimization problems (sparse solution of certain under-determined linear equations) can be exactly solved using polynomial-time algorithms. Motivated by this connection, several research groups [4], [13], [14], [23], [24], including our group [18], have suggested variations of this theme to design polynomial-time algorithms for robust regression. In this paper, we derive the precise conditions under which the sparse regularization (l_0 and l_1-norm) based approaches can solve the robust regression problem correctly.

We address the traditional robust regression problem where the number of observations N is larger than the number of unknown regression parameters D. As is now the standard practice for handling outliers, we express the regression error as a sum of two error terms: a sparse outlier error term and a dense inlier (small) error term [1], [13], [14], [18], [23], [24]. Under the reasonable assumption that there are fewer outliers than inliers in a training dataset, the robust regression problem can be formulated as a l_0 regularization problem. We state the conditions under which the l_0 regularization approach will correctly estimate the regression parameters. We show that a quantity θ_k, defined as the smallest principal angle between the regressor subspace and all k-dimensional outlier subspaces, is the fundamental quantity that determines the performance of this approach. More specifically we show that if the regressor matrix is full column rank and $\theta_{2k} > 0$, then the l_0 regularization approach can handle k outliers. Since, the l_0 regularization problem is a combinatorial problem, we relax it to a l_1-norm regularized problem. We then show that if the regressor matrix is full column rank and $\theta_{2k} > \cos^{-1}\left(\frac{2}{3}\right)$, then the l_1-norm regularization approach can handle k outliers. For a summary of our main results, see Fig. 1. We also study the theoretical computational complexity and empirically performance of various robust regression algorithms to identify algorithms that are efficient for solving high-dimensional regression problems.

A. Contributions

The technical contributions of this paper are as follows:

- We state the sufficient conditions for the sparse regularization (l_0 and l_1-norm) approaches to correctly solve the traditional robust regression problem. We show that a quantity θ_k, which measures the angular separation between the regressor subspace and all k-dimensional outlier subspaces, is the fundamental quantity that determines the performance of these algorithms.

Manuscript received December 03, 2011; revised July 20, 2012; accepted November 01, 2012. Date of publication November 27, 2012; date of current version February 12, 2013. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Konstantinos I. Diamantaras. This research was mainly supported by an Army Research Office MURI Grant W911NF-09-1-0383. It was also supported by funding from NSF via Grants NSF-CCF-1117939 and NSF-IIS-1116718, and a research grant from Samsung Telecommunications America.

K. Mitra and A. Veeraraghavan are with the Department of Electrical and Computer Engineering, Rice University, Houston, TX 77005 USA (e-mail: Kaushik.Mitra@rice.edu; vashok@rice.edu).

R. Chellappa is with the Department of Electrical and Computer Engineering, University of Maryland, College Park 20742 USA (e-mail: rama@umiacs.umd.edu).

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.2012.2229992

Robust regression model: $y=Xw+s+n$, where
y,X: constructed from training data;
w: regression parameter;
s: sparse outliers; n: dense inlier noise
Goal: Estimate w

Our contribution: Provide sufficient conditions for correctly estimating w from y
Summary of our results: 1)Assume X is full column rank
2) θ_k defined as the smallest principal angle between regressor subspace and all k-dimensional outlier subspaces

Different Cases	Assumptions	Optimization problem	Sufficient conditions
Exact recovery via l_0-norm regularization	$n=0$	$\min_{\{s,w\}} \|\|s\|\|_0$ such that $y=Xw+s$ (l_0 regression)	If $\theta_{2k}>0$ and $\|\|s\|\|_0 \le k$, then w can be estimated exactly (Prop II.1)
Exact recovery via l_1-norm regularization	$n=0$	$\min_{\{s,w\}} \|\|s\|\|_1$ such that $y=Xw+s$ (l_1 regression)	If $\theta_{2k}>\cos^{-1}(2/3)$ and $\|\|s\|\|_0 \le k$, then w can be estimated exactly (Fact II.1)
Bounded error recovery via l_1-norm regularization	$\|\|n\|\|_2 \le \varepsilon$	$\min_{\{s,w\}} \|\|s\|\|_1$ such that $y=Xw+s+n$, $\|\|n\|\|_2 \le \varepsilon$ (BPRR)	If $\theta_{2k}>\cos^{-1}(2/3)$ and $\|\|s\|\|_0 \le k$, then estimation error in w is bounded (Th II.1)

Fig. 1. The main contribution of this paper is to state the sufficient conditions under which the sparse regularization (l_0 and l_1-norm) approaches correctly solve the robust regression problem.

- Our Proposition II.1 and Theorem II.1 gives an estimate on the number of outliers the sparse regularization approaches can handle.
- We empirically compare the sparse (l_1-norm) regularization approach with the traditional robust algorithms to identify accurate and efficient algorithms for solving high-dimensional problems.

B. Prior Work

Various robust regression approaches have been proposed in the statistics and signal processing literature. We mention some of the major classes of approaches such as LMedS, RANSAC and M-estimators. In LMedS [19], the median of the squared residues is minimized using a random sampling algorithm. This sampling algorithm is combinatorial in the dimension (number of regression parameters) of the problem, which makes LMedS impractical for solving high-dimensional regression problems. RANSAC [10] and its improvements such as MSAC, MLESAC [26] are the most widely used robust approaches in computer vision [22]. RANSAC estimates the model parameters by minimizing the number of outliers, which are defined as data points that have residual greater than a pre-defined threshold. The same random sampling algorithm (as used in LMedS) is used for solving this problem, which makes RANSAC, MSAC and MLESAC impractical for high-dimension problems.

Another popular class of robust approaches is the M-estimates [12]. M-estimates are a generalization of the maximum likelihood estimates (MLEs), where the negative log likelihood function of the data is replaced by a robust cost function. Many of these robust cost functions are non-convex. Generally a polynomial time algorithm *iteratively reweighted least squares* (IRLS) is used for solving the optimization problem, which often converges to a local minimum. Many other robust approaches have been proposed such as S-estimates [21], L-estimates [20] and MM-estimates [28], but all of them are solved using a (combinatorial) random sampling algorithm, and hence are not attractive for solving high-dimensional problems [17].

A similar mathematical formulation (as robust regression) arises in the context of error-correcting codes over the reals [1], [4]. The decoding schemes for this formulation are very similar to robust regression algorithms. The decoding scheme used in [4] is the l_1-regression. It was shown that if a certain orthogonal matrix, related to the encoding matrix, satisfies the Restricted Isometry Property (RIP) and the gross error vector is sufficiently sparse, then the message can be successfully recovered. In [1], this error-correcting scheme was further extended to the case where the channel could introduce (dense) small errors along with sparse gross errors. However, the robust regression problem is different from the error-correction problem in the sense that in error-correction one is free to design the encoding matrix, whereas in robust regression the training dataset dictates the structure of the regressor matrix (which plays a similar role as the encoding matrix). Also, the conditions that we provide are more appropriate in the context of robust regression and are tighter than that provided in [1].

Recently, many algorithms have been proposed to handle outliers in the compressive sensing framework [5], [14]. Our framework is different from them since we consider the traditional regression problem, where there are more observations (data points) than the unknown model parameters and we do not have the freedom to design the regressor matrix. As an alternative to sparse regularization based robust regression

approaches, a Bayesian approach has been proposed in [13], [18]. In this approach, a sparse prior [25] is assumed on the outliers and the resulting problem is solved using the maximum a-posterior (MAP) criterion. Another strain of related results studies the recovery and separation of sparsely corrupted signal [23], [24]. These results, however, rely on the coherence parameters of the regressor and outlier matrices, rather than on the principle angle between them.

C. Outline of the Paper

The remainder of the paper is organized as follows: in Section II we formulate the robust regression problem as a l_0 regularization problem and its relaxed convex version, a l_1-norm regularization problem, and state conditions under which the proposed optimization problems solves the robust regression problem. In Section III we prove our main result and in Section IV we perform several empirical experiments to compare various robust approaches.

II. Robust Regression Based on Sparse Regularization

Regression is the problem of estimating the functional relation f between two sets of variables: independent variable (or regressor) $x \in \mathbb{R}^D$, and dependent variable (or regressand) $y \in \mathbb{R}$, given many training pairs (y_i, x_i), $i = 1, 2 \ldots, N$. In linear regression, the function f is a linear function of the vector of model parameters $w \in \mathbb{R}^D$:

$$y_i = x_i^T w + e, \tag{1}$$

where e is the observation noise. We wish to estimate w from the given training dataset of N observations. We can write all the observation equations collectively as:

$$y = Xw + e, \tag{2}$$

where $y = (y_1, \ldots, y_N)^T$, $X = [x_1, \ldots, x_N]^T \in \mathbb{R}^{N \times D}$ (the regressor matrix) and $e = (e_1, \ldots, e_N)^T$. In this paper we consider the traditional regression framework, where there are more observations than the unknown model parameters, i.e., $N > D$. The most popular estimator of w is the least squares (LS) estimate, which is statistically optimal (in the maximum likelihood sense) for the case when the noise is i.i.d Gaussian. However, in the presence of outliers or gross error, the noise distribution is far from Gaussian and, hence, LS gives poor estimates of w.

A. Robust Regression as a l_0 Regularization Problem

As is now a standard practice for handling outliers, we express the noise variable e as sum of two independent components, $e = s + n$, where s represents the sparse outlier noise and n represents the dense inlier noise [1], [13], [14], [18], [23]. With this the robust linear regression model is given by:

$$y = Xw + s + n. \tag{3}$$

This is an ill-posed problem as there are more unknowns, w and s, than equations and hence there are infinitely many solutions. Clearly, we need to restrict the solution space in order to find a unique solution. A reasonable assumption/restriction could be that outliers are sparse in a training dataset, i.e., there are fewer outliers than inliers in a dataset. Under this sparse outliers assumption, the appropriate optimization problem to solve would be:

$$\min_{s,w} \|s\|_0 \quad \text{such that} \quad \|y - Xw - s\|_2 \leq \epsilon, \tag{4}$$

where $\|s\|_0$ is the number of non-zero elements in s and ϵ is a measure of the magnitude of the small noise n. Before looking at the case where both outliers and small noise is present, we first treat the case where only outliers are present, i.e., $n = 0$.

When $n = 0$, we should solve:

$$\min_{s,w} \|s\|_0 \quad \text{such that} \quad y = Xw + s. \tag{5}$$

Note that the above problem can be rewritten as $\min_w \|y - Xw\|_0$, and hence can be termed the l_0 regression problem. We are interested in answering the following question: Under what conditions, by solving the above equation, can we recover the original w from the observation y? One obvious condition is that X should be a full column rank matrix (remember $N \geq D$), otherwise, even when there are no outliers, we will not be able to recover the original w. To discover the other conditions, we rewrite the constraint in (5) as

$$y = [X\ I]w_s, \tag{6}$$

where I is a $N \times N$ identity matrix and $w_s = [w; s]$[1] is the augmented vector of unknowns. Now, consider a particular dataset (y, X) where amongst the N data points, characterized by the index set $J = [1, 2, \ldots, N]$, k of them are affected by outliers. Let these k outlier affected data points be specified by the subset $T \subset J$. Then, (6) can be written as

$$y = [X\ I_T]w_{sT}, \tag{7}$$

where I_T is a matrix consisting of column vectors from I indexed by T, $s_T \in \mathbb{R}^k$ represents the k non-zero outlier noise and $w_{s_T} = [w; s_T]$. Given the information about the index subset T, i.e. given which data (indices) are affected by outliers, we can recover w and the non-zero outliers s_T from (7) if and only if $[X\ I_T]$ is full column rank. The condition for $[X\ I_T]$ to be full rank can also be expressed in terms of the smallest *principal angle* between the subspace spanned by the regressor, $\mathrm{span}(X)$, and the subspace spanned by outliers, $\mathrm{span}(I_T)$. The smallest principle angle θ between two subspaces $\mathcal{U}$ and $\mathcal{W}$ of $\mathbb{R}^N$ is defined as the smallest angle between a vector in $\mathcal{U}$ and a vector in $\mathcal{W}$ [11]:

$$\cos(\theta) = \max_{u \in \mathcal{U}} \max_{w \in \mathcal{W}} \frac{|u^T w|}{\|u\|_2 \|w\|_2}. \tag{8}$$

Let θ_T denote the smallest principal angle between the subspaces $\mathrm{span}(X)$ and $\mathrm{span}(I_T)$, then for any vectors $u \in \mathrm{span}(X)$ and $w \in \mathrm{span}(I_T)$:

$$|u^T w| \leq \delta_T \|u\|_2 \|w\|_2, \tag{9}$$

where $\delta_T = cos(\theta_T)$ is the smallest such number. We then generalize the definition of the smallest principal angle θ_T to a new quantity θ_k:

$$\theta_k = \min_{|T| \leq k} \theta_T, \quad k = 1, 2, \ldots, N, \tag{10}$$

[1]Throughout this paper, we will use the MATLAB notation $[w; s]$ to mean $[w^T\ s^T]^T$

i.e., θ_k is the smallest principal angle between the regression subspace and all the k-dimensional outlier subspaces. $\delta_k = cos(\theta_k)$ is then the smallest number such that for any vectors $u \in \text{span}(X)$ and $w \in \text{span}(I_T)$ with $|T| \leq k$:

$$|u^T w| \leq \delta_k \|u\|_2 \|w\|_2. \tag{11}$$

The quantity $\delta_k \in [0,1]$ (or equivalently $\theta_k \in [0^\circ, 90^\circ]$) is a measure of how well separated the regressor subspace is from the all the k-dimensional outlier subspaces. When $\delta_k = 1$ (or equivalently $\theta_k = 0^\circ$), the regressor subspace and one of the k dimensional outlier subspaces, share at least a common vector, whereas, when $\delta_k = 0$ (or equivalently $\theta_k = 90^\circ$), the regressor subspace is orthogonal to all the k-dimensional outlier subspaces. With the definition of δ_k, we are now in a position to state the sufficient conditions for recovering w by solving the l_0 regression problem (5).

Proposition II.1: Assume that $\delta_{2k} < 1$ (or equivalently $\theta_{2k} > 0$), X is a full column rank matrix and $y = Xw + s$. Then, by solving the l_0 regression problem (5), we can estimate w without any error if $\|s\|_0 \leq k$ (i.e., if there are at most k outliers in the y variable).

Proof: The conditions $\delta_{2k} < 1$ and X a full rank matrix together implies that all matrices of the form $[X\ I_T]$ with $|T| \leq 2k$ are full rank. This fact can be proved by the principle of contradiction.

Now, suppose w_0 and s_0 with $\|s_0\|_0 \leq k$ satisfy the equation

$$y = Xw + s. \tag{12}$$

Then to show that we can recover w_0 and s_0 by solving (5), it is sufficient to show that there exists no other w and s, with $\|s\|_0 \leq k$, which also satisfy (12). We show this by contradiction: Suppose there is another such pair, say w_1 and s_1 with $\|s_1\|_0 \leq k$, which also satisfies (12). Then $Xw_0 + s_0 = Xw_1 + s_1$. Re-arranging, we have:

$$[X\ I]\Delta w_s = 0 \tag{13}$$

where $\Delta w_s = [\Delta w; \Delta s]$, $\Delta w = (w_0 - w_1)$ and $\Delta s = (s_0 - s_1)$. Since $\|s_0\|_0 \leq k$ and $\|s_1\|_0 \leq k$, $\|\Delta s\|_0 \leq 2k$. If T_Δ denotes the corresponding non-zero index set, then T_Δ has a cardinality of at most $2k$ and, thus, $[X\ I_{T_\Delta}]$ is a full rank matrix. This in turn implies that $\Delta w_s = 0$, i.e. $w_0 = w_1$ and $s_0 = s_1$. ∎

From the above result, we can find a lower bound on the maximum number of outliers (in the y variable) that the l_0 regression (5) can handle in a dataset characterized by the regressor matrix X. This is given by the largest integer k such that $\delta_{2k} < 1$.

B. Robust Regression as a l_1-Norm Regularization Problem

The l_0 regularization problem (5) is a hard combinatorial problem to solve. So, we approximate it by the following convex problem:

$$\min_{s,w} \|s\|_1 \quad \text{such that} \quad y = Xw + s \tag{14}$$

where the $\|s\|_0$ term is replaced by the l_1 norm of s. Note that the above problem can be re-written as $\min_w \|y - Xw\|_1$, and hence this is the l_1 regression problem. Again, we are interested in the question: Under what conditions, by solving the l_1 regression problem (14), can we recover the original w? Not surprisingly, the answer is that we need a bigger angular separation between the regressor subspace and the outlier subspaces.

Fact II.1: Assume that $\delta_{2k} < \frac{2}{3}$ (or equivalently $\theta_{2k} > \cos^{-1}\left(\frac{2}{3}\right)$), X is a full column rank matrix and $y = Xw + s$. Then, by solving the l_1 regression problem (14), we can estimate w without any error if $\|s\|_0 \leq k$ (i.e., if there are at most k outliers in the y variable).

Proof: Proved as a special case of the main Theorem II.1. ∎

Similar to the l_0 regression case, we can also obtain a lower bound on the maximum number of outliers that the l_1 regression can handle in the y variable; this is given by the largest integer k for which $\delta_{2k} < \frac{2}{3}$.

Next, we consider the case where the observations y_i are corrupted by gross as well as small noise. In the presence of small bounded noise $\|n\|_2 \leq \epsilon$, we propose to solve the following convex relaxation of the combinatorial problem (4)

$$\min_{s,w} \|s\|_1 \quad \text{such that} \quad y = Xw + s + n, \|n\|_2 \leq \epsilon. \tag{15}$$

The above problem is a related to the basis pursuit denoising problem (BPDN) [6] and we will refer to it as basis pursuit robust regression (BPRR). Under the same conditions on the angular separation between the regressor subspace and the outliers subspaces, we have the following result.

Theorem II.1: Assume that $\delta_{2k} < \frac{2}{3}$ (or equivalently $\theta_{2k} > \cos^{-1}\left(\frac{2}{3}\right)$), X is a full column rank matrix and $y = Xw + s + n$ with $\|n\|_2 \leq \epsilon$. Then the error in the estimation of w, Δw, by BPRR (15) is related to s_k, the best k-sparse approximation of s, and ϵ as:

$$\|\Delta w\|_2 \leq \tau^{-1}\left(C_0 k^{-\frac{1}{2}} \|s - s_k\|_1 + C_1 \epsilon\right), \tag{16}$$

where τ is the smallest singular value of X, and C_0, C_1 are constants which depend only on δ_{2k} ($C_0 = \frac{2\delta_{2k}}{1 - \frac{3}{2}\delta_{2k}}$, $C_1 = \frac{2\sqrt{1+\delta_{2k}}}{1 - \frac{3}{2}\delta_{2k}}$).

Note that if there are at most k outliers, $s_k = s$ and the estimation error $\|\Delta w\|_2$ is bounded by a constant times ϵ. Furthermore, in the absence of small noise ($\epsilon = 0$), we can recover w without any error, which is the claim of Fact II.1. We prove our main Theorem II.1 in the next Section.

III. Proof of the Main Theorem II.1

The main assumption of the Theorem is in terms of the smallest principal angle between the regressor subspace, $\text{span}(X)$, and the outlier subspaces, $\text{span}(I_T)$. This angle is best expressed in terms of orthonormal bases of the subspaces. I_T is already an orthonormal basis, but the same can not be said about X. Hence we first ortho-normalize X by the reduced QR decomposition, i.e. $X = QR$ where Q is an $N \times D$ matrix which forms an orthonormal basis for X and R is an $D \times D$ upper triangular matrix. Since X is assumed to be a full column rank matrix, R is a full rank matrix. Using this decomposition of X, we can solve (15) in an alternative way. First, we can substitute $z = Rw$ and solve for z from:

$$\min_{s,z} \|s\|_1 \quad \text{such that} \quad \|y - Qz - s\|_2 \leq \epsilon. \tag{17}$$

We can then obtain w from $w = R^{-1}z$. This way of solving for w is exactly equivalent to that of (15), and hence for solving practical problems any of the two approaches can be used. However, the proof of the Theorem is based on this alternative approach. We first obtain an estimation error bound on z and then use $w = R^{-1}z$ to obtain a bound on w.

For the main proof we will need some more results. One of the results is on the relation between δ_k and a quantity μ_k, defined below, which is very similar to the concept of restricted isometry constant [4].

Definition III.1: For orthonormal matrix Q, we define a constant μ_k, $k = 1, 2, \ldots, N$ as the smallest number such that

$$(1 - \mu_k)\|x\|_2^2 \leq \|[Q\ I_T]x\|_2^2 \leq (1 + \mu_k)\|x\|_2^2 \tag{18}$$

for all T with cardinality at most k.

Lemma III.1: For orthonormal regressor matrix Q, $\delta_k = \mu_k$, $k = 1, 2, \ldots, N$.

Proof: From definition of δ_k, for any I_T (with $|T| \leq k$), z and $s \in \mathbb{R}^k$:

$$|\langle Qz, I_T s\rangle| \leq \delta_k \|z\|_2 \|s\|_2 \tag{19}$$

where we have used $\|Qz\|_2 = \|z\|_2$ and $\|I_T s\|_2 = \|s\|_2$ since Q and I_T are orthonormal matrices. Writing $x = [z; s]$, $\|[Q\ I_T]x\|_2^2$ is given by

$$\begin{aligned}\|[Q\ I_T]x\|_2^2 &= \|z\|_2^2 + \|s\|_2^2 + 2\langle Qz, I_T s\rangle \\ &\leq \|z\|_2^2 + \|s\|_2^2 + 2\delta_k \|z\|_2 \|s\|_2 \end{aligned} \tag{20}$$

Note that, from the definition of δ_k, the above inequality is tight, i.e., there exists z and s for which the inequality is satisfied with an equality. Using the fact $2\|z\|_2\|s\|_2 \leq \|z\|_2^2 + \|s\|_2^2$ we get

$$\|[Q\ I_T]x\|_2^2 \leq \|z\|_2^2 + \|s\|_2^2 + \delta_k \left(\|z\|_2^2 + \|s\|_2^2\right). \tag{21}$$

Further, using the fact $\|x\|_2^2 = \|z\|_2^2 + \|s\|_2^2$, we get $\|[Q\ \ I_T]x\|_2^2 \leq (1 + \delta_k)\|x\|_2^2$. Using the inequality $\langle Qz, I_T s\rangle \geq -\delta_k \|z\|_2\|s\|_2$, it is easy to show that $\|[QI_T]x\|_2^2 \geq (1 - \delta_k)\|x\|_2^2$. Thus, we have

$$(1 - \delta_k)\|x\|_2^2 \leq \|[Q\ I_T]x\|_2^2 \leq (1 + \delta_k)\|x\|_2^2, \tag{22}$$

which implies $\delta_k \geq \mu_k$. However, there exists $x = [z; s]$ which satisfies both the inequalities (20) and (21) with equality and hence $\delta_k = \mu_k$. ∎

The proof of the main Theorem parallels that in [2]. Suppose for a given (y, X), (z,s) satisfy $y = Qz + s + n$ with $\|n\|_2 \leq \epsilon$. And let z^* and s^* be the solution of (17) for this (y, X). Then

$$\|Q(z - z^*) + (s - s^*)\|_2 \leq \|Qz + s - y\|_2 + \|y - z^* - s^*\|_2 \leq 2\epsilon \tag{23}$$

This follows from the triangle inequality and the fact that both (z, s) and (z^*, s^*) are feasible for problem (17). Let $\Delta z = z^* - z$ and $h = s^* - s$. For the rest of the proof, we use the following notation: vector x_T is equal to x on the index set T and zero elsewhere.[2] Now, let's decompose h as $h = h_{T_0} + h_{T_1} + h_{T_2} + \ldots$, where each of the index set T_i, $i = 0, 1, 2, \ldots$, is of cardinality k except for the last index set which can be of lesser cardinality. The index T_0 corresponds to the locations of k largest coefficients of s, T_1 to the locations of k largest coefficients of $h_{T_0^c}$, T_2 to that of the next largest k coefficients of $h_{T_0^c}$ and so on. In the main proof, we will need a bound on the quantity $\sum_{j\geq 2}\|h_{T_j}\|_2$, which we obtain first. We use the following results from [2]:

$$\sum_{j\geq 2} \left\|h_{T_j}\right\|_2 \leq k^{-\frac{1}{2}} \left\|h_{T_0^c}\right\|_1 \tag{24}$$

and

$$\left\|h_{T_0^c}\right\|_1 \leq \|h_{T_0}\|_1 + 2\|s_{T_0^c}\|_1. \tag{25}$$

These results correspond to (10) and (12) in [2], with some changes in notations. The first result holds because of the way h has been decomposed into $h_{T_0}, h_{T_1}, h_{T_2}, \ldots$, and the second result is based on $\|s + h\|_1 \leq \|s\|_1$, which holds because $s + h = s^*$ is the minimum l_1-norm solution of (17). Based on the above two equations, we have

$$\sum_{j\geq 2} \left\|h_{T_j}\right\|_2 \leq k^{-\frac{1}{2}} \|h_{T_0}\|_1 + 2k^{-\frac{1}{2}} \left\|s_{T_0^c}\right\|_1 \leq \|h_{T_0}\|_2 + 2e_0, \tag{26}$$

where we have used the inequality $k^{-\frac{1}{2}}\|h_{T_0}\|_1 \leq \|h_{T_0}\|_2$ and e_0 is defined as $e_0 = k^{-\frac{1}{2}}\|s_{T_0^c}\|_1$. Since by definition $s_{T_0} = s_k$, the best k-sparse approximation of s, $s_{T_0^c} = s - s_k$ and hence $e_0 = k^{-\frac{1}{2}}\|s - s_k\|_1$. With these results, we are in a position to prove Theorem II.1.

Proof of Theorem II.1: Our goal is to find a bound on Δz, from which we can find a bound on Δw. We do this by first finding a bound for $[\Delta z; h_{T_0\cup T_1}]$ through bounds on the quantity $\|Q\Delta z + h_{T_0\cup T_1}\|^2$. Using $h_{T_0\cup T_1} = h - \sum_{j\geq 2} h_{T_j}$, we get

$$\begin{aligned}\|Q\Delta z + h_{T_0\cup T_1}\|^2 &= \langle Q\Delta z + h_{T_0\cup T_1}, Q\Delta z + h\rangle \\ &\quad - \left\langle Q\Delta z + h_{T_0\cup T_1}, \sum_{j\geq 2} h_{T_j}\right\rangle. \end{aligned} \tag{27}$$

Using triangular inequality, the first term in the right hand side can be bounded as

$$\langle Q\Delta z + h_{T_0\cup T_1}, Q\Delta z + h\rangle \leq \|Q\Delta z + h_{T_0\cup T_1}\|_2 \, \|Q\Delta z + h\|_2. \tag{28}$$

Since $h_{T_0\cup T_1}$ is $2k$ sparse, using (22), we get

$$\|Q\Delta z + h_{T_0\cup T_1}\|_2 \leq \sqrt{1 + \delta_{2k}}\, \|[\Delta z; h_{T_0\cup T_1}]\|_2.$$

Further, using the bound $\|Q\Delta z + h\|_2 \leq 2\epsilon$, see (23), we get

$$\langle Q\Delta z + h_{T_0\cup T_1}, Q\Delta z + h\rangle \leq 2\epsilon\sqrt{1 + \delta_{2k}}\|[\Delta z; h_{T_0\cup T_1}]\|_2. \tag{29}$$

Now, we look at the second term in the right hand side of (27). Since the support of $h_{T_0\cup T_1}$ and $h_{T_j}, j \geq 2$ are different, $\langle h_{T_0\cup T_1}, h_{T_j}\rangle = 0$ for all $j \geq 2$, and we get

$$\begin{aligned} -\left\langle Q\Delta z + h_{T_0\cup T_1}, \sum_{j\geq 2} h_{T_j}\right\rangle &= \sum_{j\geq 2} \langle Q\Delta z, -h_{T_j}\rangle \\ &\leq \delta_{2k}\|\Delta z\|_2 \sum_{j\geq 2} \left\|h_{T_j}\right\|_2, \end{aligned}$$

[2]Note that we have used the subscript notation in a slightly different sense earlier. However, it should be easy to distinguish between the two usages from the context.

where we used the definition of δ_{2k} and the fact that h_{T_j} is k-sparse, and hence also $2k$ sparse. Further, using (26), $\|h_{T_0}\|_2 \leq \|h_{T_0\cup T_1}\|_2$ and $\|\Delta z\|_2 \leq \|[\Delta z; h_{T_0\cup T_1}]\|_2$

$$\delta_{2k}\|\Delta z\|_2 \sum_{j\geq 2} \left\|h_{T_j}\right\|_2 \leq \delta_{2k}\|\Delta z\|_2 \|h_{T_0\cup T_1}\|_2 + 2e_0\delta_{2k}\left\|[\Delta z; h_{T_0\cup T_1}]\right\|_2 \quad (30)$$

The quantity $\|\Delta z\|_2\|h_{T_0\cup T_1}\|_2$ can be further bounded by $\frac{1}{2}\|[\Delta z; h_{T_0\cup T_1}]\|_2^2$ (by applying the inequality $2ab \leq a^2 + b^2$). Therefore,

$$\delta_{2k}\|\Delta z\|_2 \sum_{j\geq 2} \left\|h_{T_j}\right\|_2 \leq \frac{\delta_{2k}}{2}\left\|[\Delta z; h_{T_0\cup T_1}]\right\|_2^2 + 2e_0\delta_{2k}\left\|[\Delta z; h_{T_0\cup T_1}]\right\|_2. \quad (31)$$

Finally, we obtain the following bound for $\|Q\Delta z + h_{T_0\cup T_1}\|^2$

$$\|Q\Delta z + h_{T_0\cup T_1}\|^2 \leq (2\epsilon\sqrt{1+\delta_{2k}} + 2e_0\delta_{2k})\left\|[\Delta z; h_{T_0\cup T_1}]\right\|_2 + \frac{\delta_{2k}}{2}\left\|[\Delta z; h_{T_0\cup T_1}]\right\|_2^2.$$

Since $h_{T_0\cup T_1}$ is $2k$ sparse, from (22), we get

$$(1-\delta_{2k})\left\|[\Delta z; h_{T_0\cup T_1}]\right\|_2^2 \leq \|Q\Delta z + h_{T_0\cup T_1}\|_2^2. \quad (32)$$

From the above two equations, it follows that

$$\left(1 - \frac{3}{2}\delta_{2k}\right)\left\|[\Delta z; h_{T_0\cup T_1}]\right\|_2 \leq 2e_0\delta_{2k} + 2\epsilon\sqrt{1+\delta_{2k}}. \quad (33)$$

Since $\delta_{2k} < \frac{2}{3}$ is an assumption of the Theorem, $1 - \frac{3}{2}\delta_{2k} > 0$, and hence

$$\left\|[\Delta z; h_{T_0\cup T_1}]\right\|_2 \leq \frac{2e_0\delta_{2k}}{1-\frac{3}{2}\delta_{2k}} + \frac{2\epsilon\sqrt{1+\delta_{2k}}}{1-\frac{3}{2}\delta_{2k}}. \quad (34)$$

Since $\|\Delta z\|_2 \leq \|[\Delta z; h_{T_0\cup T_1}]\|_2$, we obtain

$$\|z\|_2 \leq C_0 k^{-\frac{1}{2}}\|s - s_k\|_1 + C_1\epsilon$$
$$\text{where } C_0 = \frac{2\delta_{2k}}{1-\frac{3}{2}\delta_{2k}}, C_1 = \frac{2\sqrt{1+\delta_{2k}}}{1-\frac{3}{2}\delta_{2k}}. \quad (35)$$

Using the definition $w = R^{-1}z$, we get $\Delta w \leq \|R^{-1}\|_2\|\Delta z\|_2$, where $\|R^{-1}\|_2$ is the spectral norm of R^{-1}. Note that the spectral norm of R^{-1} is given by its largest singular value, which is the reciprocal of the smallest singular value of R. Further, since $X = QR$ and R share the same singular values, $\|R^{-1}\|_2 = \tau^{-1}$, where τ is the smallest singular value of X. Hence, we have the final result

$$\Delta w \leq \tau^{-1}\left(C_0 k^{-\frac{1}{2}}\|s - s_k\|_1 + C_1\epsilon\right). \quad (36)$$

IV. Empirical Studies of the Robust Regression Algorithms

In the previous Section, we have shown that if X is full column rank and $\theta_{2k} > \cos^{-1}\left(\frac{2}{3}\right)$ (where θ_{2k} is the smallest principal angle between the regression and $2k$-dimensional outlier subspaces), then the l_1-norm regularization approach (BPRR) can handle k outliers. However, computing the quantity θ_{2k} is in itself a combinatorial problem. Hence, there is no easy way to characterize the performance of BPRR. In this Section we empirically characterize the performance of BPRR and compare it with other robust approaches. We classify the robust approaches into two major classes: 1) tradition approaches such as M-estimators, LMedS, RANSAC and 2) approaches based on compressive sensing theory such as the BPRR and a Bayesian alternative to the sparse regularization approach proposed in [13], [18]. Three important parameters of the robust regression problem are: fraction of outliers in the dataset f, dimension of the problem D and inlier noise variance σ^2. We study the performances of the algorithms with respect to these parameters. The performance criteria are estimation accuracy and computational complexity. In Section IV-A we briefly introduce the robust approaches and discuss the theoretical computational complexity of their associated algorithms and in Section IV-B we empirically study the performance of these algorithms.

A. Robust Regression Approaches and Computational Complexity of Their Associated Algorithms

1) Compressive Sensing Based Robust Approaches:

- *BPRR*: We formulate BPRR (15) as a second order cone programming problem. Since there are $N + D$ variables and one cone constraint of dimension N, the computational complexity of this algorithm is $O((N + D)^{2.5}N))$ [15].
- *Bayesian Robust Regression (BRR)*: As an alternative to the sparse regularization approach, a Bayesian approach was proposed in [13], [18] towards solving the robust regression problem (4). In this approach the outliers are modeled by sparse priors [25] and they are then estimated using the MAP criterion (see [18] for more details). The main computational step of this approach (BRR) is the MAP estimation step whose computational complexity is $O(N^3)$.

2) Traditional Robust Approaches:

- *M-estimates*: In M-estimates [12] a robust cost function $\rho(e_i)$ of the residual error $e_i = y_i - w^T x_i, i = 1, 2, \ldots, N$ is minimized:

$$\hat{w} = \arg\min_w \sum_{i=1}^{N} \rho(e_i) \quad (37)$$

  where the robust function $\rho(e)$ should satisfy certain properties (see [12]). In our experiments, we have used the popular Tukey's biweight function which is a robust but non-convex function. The IRLS algorithm, which is used for solving the optimization problem, is a polynomial time algorithm with a computational complexity of $O\left(\left(\frac{N+D}{3}\right)D^2\right)$ per iteration [17].
- *LMedS*: In LMedS the median of the residual error e_i is minimized, i.e.,

$$\hat{w} = \min_w \text{median}(e_i^2) \quad (38)$$

This problem is solved by a random sampling algorithm, which is combinatorial in the dimension of the problem D [10], [20]. Thus, LMedS becomes impractical for solving high-dimensional problems.

- *RANSAC*: In RANSAC the model parameter is estimated by minimizing the number of outliers (which are defined as data points that have residual greater than a pre-defined threshold):

$$\min_{s,w} \|s\|_0 \quad \text{such that} \quad \|y - Xw - s\|_\infty \leq c, \tag{39}$$

where c is the pre-defined threshold and is related to the standard deviation of the inlier noise. The same combinatorial random sampling algorithm as used in LMedS is used for solving this problem, which makes it impractical for solving high-dimensional problems.

B. Empirical Studies

We perform a series of empirical experiments to characterize the performance of the robust regression approaches. For each trial in the experiments, we generate the dataset (x_i, y_i), $i = 1, 2, \ldots, N$, $x_i \in \mathbb{R}^D$, $y \in \mathbb{R}$, and the model parameters $w \in \mathbb{R}^D$ in the following manner: x_i's are obtained by uniformly sampling a D-dimensional hypercube centered around the origin and w is a randomly sampled from a standard Gaussian random variable. Depending on the outlier fraction f, we randomly categorize the N indices into either inlier or outlier indices. The y_i's corresponding to the inlier indices are obtained from $y_i = x_i^T w + n$, where n is the inlier noise, which we choose to be a Gaussian random variable $\mathbb{N}(0, \sigma^2)$. The y_i corresponding to the outlier indices are obtained by uniformly sampling the interval $[-r, r]$, where $r = \max_i |y_i|$. Regression accuracy is measured by the estimation error $\|\Delta w\|_2$. BPRR, BRR and RANSAC need estimates of the inlier noise standard deviation, which we provide as the median absolute residual of the l_1 regression. In our experiments, we have used the MATLAB implementation of bisquare (Tukey's biweight) M-estimates, other M-estimates give similar results.

1) Studies by Varying the Fraction of Outliers: In the first experiment, we study the performance of the robust approaches as a function of outlier fraction and dimension. We generate $N = 500$ synthetic data with inlier noise standard deviation $\sigma = 0.001$. Fig. 2 shows the mean estimation error over 20 trials vs. outlier fraction for dimension 2 and 25. For dimension 25, we only show BPRR, BRR and M-estimates as the other approaches, LMedS and RANSAC, are combinatorial in nature and hence very slow. Fig. 2 suggests that, overall, compressive sensing based robust approaches perform better than the traditional approaches.

2) Phase Transition Curves: We further study the performance of the robust approaches with respect to outlier fraction and dimension using *phase transition curves* [8], [9]. In compressive sensing theory, where the goal is to find the sparsest solution for an under-determined system of equations, it has been observed that many algorithms exhibit a sharp transition from success to failure cases: For a given level of under-determinacy, the algorithms successfully recovers the correct solution (with high probability) if the sparsity is below a certain level and fails

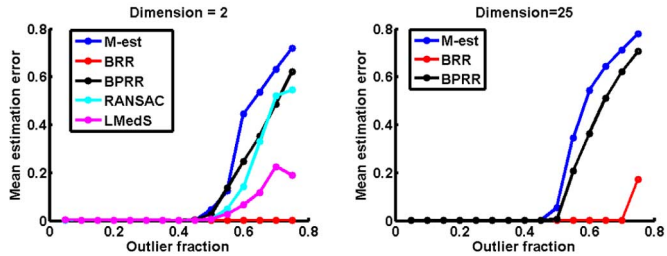


Fig. 2. Mean estimation error vs. outlier fraction for dimension 2 and 25 respectively. For dimension 25 we only show the plots for BPRR, BRR and M-estimator, as the other approaches (LMedS and RANSAC), being combinatorial in nature, are very slow. This plot suggests that, overall, compressive sensing based robust approaches perform better than the traditional approaches.

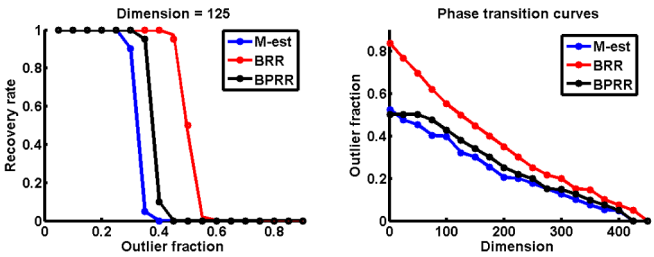
to do so (with high probability) if the sparsity is above that level [8], [9], [16]. This phenomenon is termed phase transition in the compressive sensing literature and it has been used to characterize and compare the performances of several compressive sensing algorithms [16]. We also use this measure to compare the various robust regression algorithms. In the context of robust regression, the notion of under-determinacy depends on N and D. Since, there are N observations and $N + D$ unknowns, by varying D for a fixed N we can vary the level of under-determinacy. The notion of sparsity is associated with the outlier fraction. Hence, to obtain the phase transition curves, we vary the dimension D of the problem for a fixed N and for each D find the outlier fraction where the transition from success to failure occurs.

As before, we choose $N = 500$ and $\sigma = 0.001$. We vary D over a range of values from 1 to 450. At each D, we vary the outlier fractions over a range of values and measure the fraction of trials in which the algorithms successfully found the correct solution.[3] Fig. 3(a) shows the fraction of successful recovery vs. outlier fraction for dimension 125 for approaches BPRR, BRR and M-estimators (we do not show LMedS and RANSAC as these approaches are very slow). From this Figure we conclude that each of the approaches exhibit a sharp transition from success to failure at a certain outlier fraction. This confirms that phase transition do occur in robust regression also. Next, for each regression approach and dimension, we find the outlier fraction where the probability of success is 0.5. Similar to [16], we use logistic regression to find this outlier fraction. We then plot this outlier fraction against the dimension to obtain the phase transition curves for each of the approaches. Fig. 3(b) shows the phase transition curves for BPRR, BRR and M-estimators. Again, compressive sensing based robust approaches (especially BRR) gives very good performance.

3) Studies by Varying the Amount of Inlier Noise: We also study the effect of inlier noise variance on the performance of the approaches. For this we fixed the dimension at 6, the outlier fraction at 0.4 and the number of data points at 500. Fig. 4 shows that all approaches, except for LMedS, perform well.

Finally based on the above experiments, we conclude that overall compressive sensing based robust approaches (especially BRR) perform better than the traditional robust approaches. It has been suggested in [27] that the sparse

[3]We consider a solution to be correct if $\frac{\|w - \hat{w}\|_2}{\|w\|_2} \leq 0.01$.



(a) *Recovery rate vs. outlier fraction* (b) *Phase transition curves*

Fig. 3. Subplot (a) shows recovery rate, i.e. the fraction of successful recovery, vs. outlier fraction for dimension 125 of BPRR, BRR and M-estimators. From this plot we conclude that each of the approaches exhibit a sharp transition from success to failure at a certain outlier fraction. Subplot (b) shows the phase transition curves for BPRR, BRR and M-estimator. The phase transition curve for any approach is obtained by computing for each dimension the outlier fraction where the recovery rate is 0.5. From the phase transition curves we conclude that the compressive sensing based robust approaches (especially BRR) gives very good performance. (a) Recovery rate vs. outlier fraction; (b) phase transition curves.

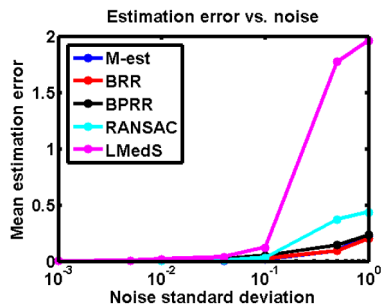


Fig. 4. Mean angle error vs. inlier noise standard deviation for dimension 6 and 0.4 outlier fraction. All approaches, except for LMedS, perform well.

Bayesian approach (BRR) is a better approximation of the l_0 regularization problem than the l_1-norm formulation, which might explain the better performance of BRR over the BPRR. However, analytical characterization of the Bayesian approach is very difficult and could be an interesting direction of future research.

V. Discussion

In this paper we addressed the traditional robust regression problem and stated the precise conditions under which sparse regularization (l_0 and l_1-norm) approaches can solve the robust regression problem. We showed that θ_k (the smallest principal angle between the regressor subspace and all k-dimensional outlier subspaces) is the fundamental quantity that determines the performance of these algorithms. Specifically, we showed if the regressor matrix X is full column rank and $\theta_{2k} > 0$, then the l_0 regularization can handle k outliers. Since, l_0 optimization is a combinatorial problem, we looked at its relaxed convex version BPRR. We then showed that if X is a full column rank matrix and $\theta_{2k} > \cos^{-1}\left(\frac{2}{3}\right)$, then BPRR can handle k outliers.

However, computing the quantity θ_k is in itself a combinatorial problem. Hence, we characterize the BPRR algorithm empirically and compare it with other robust algorithms such as M-estimates, LMedS, RANSAC and a Bayesian alternative to the sparse regularization approach (BRR). Our experiments show that BRR gives very good performance. It has been suggested in [27] that the sparse Bayesian approach is a better approximation of the l_0 regularization problem than the l_1-norm formulation, which might explain the better performance of BRR over the BPRR. However, analytical characterization of the Bayesian approach is very difficult and is an interesting direction of future research. Another interesting direction of future research would be to find greedy algorithms that can provide lower and upper bounds on the quantity θ_k.

References

[1] E. Candes and P. Randall, "Highly robust error correction byconvex programming," *IEEE Trans. Inf. Theory*, vol. 54, no. , pp. 2829–2840, 2008.
[2] E. J. Candes, "The restricted isometry property and its implications for compressed sensing," *Comptes Rendus Mathematique*, 2008.
[3] E. J. Candès, J. K. 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, 2006.
[4] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.
[5] R. E. Carrillo, K. E. Barner, and T. C. Aysal, "Robust sampling and reconstruction methods for sparse signals in the presence of impulsive noise," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 392–408, 2010.
[6] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Scientif. Comput.*, vol. 20, pp. 33–61, 1998.
[7] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.
[8] D. L. Donoho, "High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension," *Discrete Comput. Geometry*, 2006.
[9] D. L. Donoho and J. Tanner, "Counting faces of randomly projected polytopes when the projection radically lowers dimension," *J. Amer. Math. Soc.*, 2009.
[10] M. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Comm. Assoc. Mach.*, 1981.
[11] G. H. Golub and C. F. V. Loan, *Matrix Computations*. Baltimore, MD, USA: Johns Hopkins Univ. Press, 1996.
[12] P. Huber, *Robust Statistics*. : Wiley Series in Probability and Statistics, 1981.
[13] Y. Jin and B. D. Rao, "Algorithms for robust linear regression by exploiting the connection to sparse signal recovery," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2010, pp. 3830–3833.
[14] J. N. Laska, M. A. Davenport, and R. G. Baraniuk, "Exact signal recovery from sparsely corrupted measurements through the pursuit of justice," in *Conf. Rec. Asilomar Conf. Signals, Syst., Comput.*, 2009, pp. 1556–1560.
[15] M. S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, "Applications of second-order cone programming," *Linear Algebra Appl.*, 1998.
[16] A. Maleki and D. L. Donoho, "Optimally tuned iterative reconstruction algorithms for compressed sensing," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 330–341, 2010.
[17] R. A. Maronna, R. D. Martin, and V. J. Yohai, *Robust Statistics, Theory and Methods*, ser. Wiley Series in Probability and Statistics. New York, NY, USA: Wiley, 2006.
[18] K. Mitra, A. Veeraraghavan, and R. Chellappa, "Robust regression using sparse learning for high dimensional parameter estimation problems," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2010, pp. 3846–3849.
[19] P. J. Rousseeuw, "Least median of squares regression," *J. Amer. Stat. Assoc.*, 1984.
[20] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*, ser. Wiley Series in Probability and Mathematical Statistics. New York, NY, USA: Wiley, 1986.
[21] P. J. Rousseeuw and V. Yohai, "Robust regression by means of s-estimators," in *Robust and Nonlinear Time Series Analysis*, ser. Lecture Notes in Statistics. New York, NY, USA: Springer-Verlag, 1984.
[22] C. V. Stewart, "Robust parameter estimation in computer vision," *SIAM Rev.*, 1999.
[23] C. Studer and R. G. Baraniuk, "Stable restoration and separation of approximately sparse signals," *Appl. Comput. Harmon. Anal.*, 2011, submitted for publication.

[24] 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, 2012.
[25] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, 2001.
[26] P. H. S. Torr and A. Zisserman, "Mlesac: A new robust estimator with application to estimating image geometry," *Comput. Vis. Image Understand.*, 2000.
[27] D. P. Wipf and B. D. Rao, "Sparse Bayesian learning for basis selection," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2153–2164, 2004.
[28] V. J. Yohai, "High breakdown-point and high efficiency robust estimates for regression," *Ann. Statist.*, Jun. 1987.

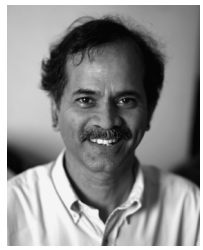
Kaushik Mitra (M'11) received the Ph.D. degree from the Electrical and Computer Engineering Department at the University of Maryland, College Park, MD, USA, in 2011 under the supervision of Prof. Ra. Chellappais.

He is currently a Postdoctoral Research Associate in the Electrical and Computer Engineering Department of Rice University, Houston, TX, USA. His areas of research interests are computational imaging, computer vision, and statistical learning theory.

Ashok Veeraraghavan (M'08) received the Bachelor's degree in electrical engineering from the Indian Institute of Technology, Madras, India, in 2002 and the M.S. and Ph.D. degrees from the Department of Electrical and Computer Engineering at the University of Maryland, College Park, MD, USA, in 2004 and 2008, respectively.

Before joining Rice University, he spent three years as a Research Scientist at Mitsubishi Electric Research Labs, Cambridge, MA, USA. He is currently Assistant Professor of Electrical and Computer Engineering at Rice University, Houston, TX, USA. His research interests are broadly in the areas of computational imaging, computer vision and robotics.

Dr. Veeraraghavan received the Doctoral Dissertation award for his thesis from the Department of Electrical and Computer Engineering at the University of Maryland.



Rama Chellappa (F'92) received the B.E. (Hons.) degree from the University of Madras, Madras, India, in 1975, the M.E. (Distinction) degree from the Indian Institute of Science, Bangalore, India, in 1977, and the M.S.E.E. and Ph.D. degrees in electrical engineering from Purdue University, West Lafayette, IN, USA, in 1978 and 1981 respectively.

Prior to joining the University of Maryland, he was an Assistant (1981–1986) and Associate Professor (1986–1991) and Director of the Signal and Image Processing Institute (1988–1990) at the University of Southern California (USC), Los Angeles, CA, USA. Since 1991, he has been a Professor of Electrical Engineering and an affiliate Professor of Computer Science at the University of Maryland, College Park, MD, USA. He is also affiliated with the Center for Automation Research and the Institute for Advanced Computer Studies (Permanent Member). In 2005, he was named a Minta Martin Professor of Engineering. Over the last 31 years, he has published numerous book chapters and peer-reviewed journal and conference papers in the areas of image processing, computer vision, and pattern recognition. He has coauthored and edited books on MRFs and face and gait recognition and collected works on image processing and analysis. His current research interests are face and gait analysis, markerless motion capture, 3D modeling from video, image and video-based recognition and exploitation, compressive sensing, sparse representations, and domain adaptation methods.

Prof. Chellappa served as an Associate Editor of four IEEE Transactions, as a co-Guest Editor of several special issues, as a Co-Editor-in-Chief of Graphical Models and Image Processing, and as the Editor-in-Chief of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. He served as a member of the IEEE Signal Processing Society Board of Governors and as its Vice-President of Awards and Membership. Recently, he completed a two-year term as the President of the IEEE Biometrics Council. He has received several awards, including an NSF Presidential Young Investigator Award; four IBM Faculty Development Awards; an Excellence in Teaching Award from the School of Engineering at USC; two paper awards from the International Association of Pattern Recognition (IAPR); the Society, Technical Achievement, and Meritorious Service Awards from the IEEE Signal Processing Society; and the Technical Achievement and Meritorious Service Awards from the IEEE Computer Society. He has been selected to receive the K. S. Fu Prize from IAPR. At the University of Maryland, he was elected as a Distinguished Faculty Research Fellow and as a Distinguished Scholar-Teacher, received the Outstanding Faculty Research Award from the College of Engineering, an Outstanding Innovator Award from the Office of Technology Commercialization, the Outstanding GEMSTONE Mentor Award, and the Poole and Kent Teaching Award for Senior Faculty. He is a Fellow of IAPR, OSA, and AAAS. In 2010, he received the Outstanding ECE Award from Purdue University. He has served as a General and Technical Program Chair for several IEEE international and national conferences and workshops. He is a Golden Core Member of the IEEE Computer Society and served a two-year term as a Distinguished Lecturer of the IEEE Signal Processing Society.