Self Scaled Regularized Robust Regression*

Yin Wang Caglayan Dicle Mario Sznaier Octavia Camps
Electrical and Computer Engineering
Northeastern University, Boston, MA 02115
wang.yin@husky.neu.edu,cdicle@gmail.com,{msznaier,camps}@coe.neu.edu

Abstract

Linear Robust Regression (LRR) seeks to find the parameters of a linear mapping from noisy data corrupted from outliers, such that the number of inliers (i.e. pairs of points where the fitting error of the model is less than a given bound) is maximized. While this problem is known to be NP hard, several tractable relaxations have been recently proposed along with theoretical conditions guaranteeing exact recovery of the parameters of the model. However, these relaxations may perform poorly in cases where the fitting error for the outliers is large. In addition, these approaches cannot exploit available a-priori *information, such as co-occurrences. To circumvent these difficulties, in this paper we present an alternative approach to robust regression. Our main result shows that this approach is equivalent to a "self-scaled" ℓ_1 regularized robust regression problem, where the cost function is automatically scaled, with scalings that depend on the a-priori information. Thus, the proposed approach achieves substantially better performance than traditional regularized approaches in cases where the outliers are far from the linear manifold spanned by the inliers, while at the same time exhibits the same theoretical recovery properties. These results are illustrated with several application examples using both synthetic and real data.*

1. Introduction

Many computer vision problems involve finding a linear regression model relating a set of input and output variables. Examples, illustrated in Fig. 1, include line extraction from 2D images, planar surface fitting in range images, and classification using linear discriminant analysis (LDA), among others. When all the available data are inliers, least squares regression (LSR) provides good fitting regression parameters [9]. However, it is well known that in the presence of outlier data points, i.e. data points that do not fit the sought model, LSR can result in very poor fitting models [13].

The goal of robust regression is to find good fitting models in spite of the presence of outliers. Robust algorithms for linear regression include least median squares regression (LMedS) [22] and random sample consensus type methods (RANSAC) [8]. While these methods perform well, they are inherently combinatorial.

Alternative approaches exploit recent advances in compressive sensing [3, 6] by reformulating robust linear regression as an optimization problem with sparse regularizations [4, 14, 12, 19]. The advantage of these methods is that they admit convex relaxations which can be solved using polynomial-time algorithms. In [19], Mitra *et al.* derived conditions under which these relaxations solve the robust regression problem, which depend on the smallest principal angle between the regressor subspace and all outlier subspaces, in the case of noiseless inliers.

A drawback of the sparsity-based approaches is that the presence of a few gross outliers, outliers which are very far from the inlier data, can poison the optimization problem and lead to ill fitting models. Another limitation of the current sparsity-based methods is that they cannot accommodate a-priori semi-supervised knowledge such as co-occurrence information when it is known that a subset of points should have a single label – i.e. they are all inliers or all outliers. Thus, to address the above limitations, we propose a new formulation for robust linear regression which can handle gross outliers and a priori information.

The contributions of this paper are as follows:

- We provide a new sparsity-based formulation to maximize the number of inlier data points.
- We show that this new approach is equivalent to a *"self-scaled"* ℓ_1 regularized robust regression problem, where the cost function is automatically scaled and the scalings capture a-priori information. Hence, we have called the proposed method a "Self-Scaled Regularized Robust Regression" (S^2R^3) algorithm.
- We show that the self-scaling property of the proposed

*This work was supported in part by NSF grants IIS–1318145 and ECCS–1404163; AFOSR grant FA9550-12-1-0271, and the Alert DHS Center of Excellence under Award Number 2008-ST-061-ED0001.

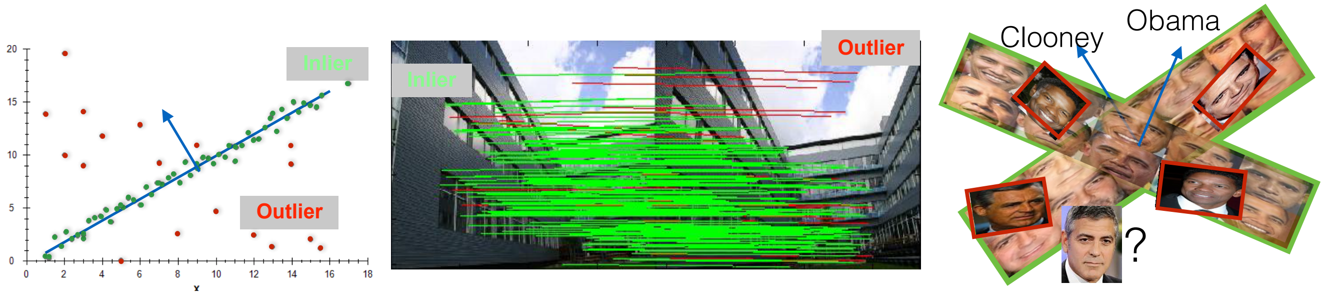

Figure 1. Sample Regression Problems in Computer Vision. Left: Line fitting; Center: Surface fitting from 3D cloud data points; Right: Linear discriminant analysis (LDA) for face recognition.

approach yields smaller fitting errors in the presence of gross outliers.

- We can incorporate a priori information by adding simple constraints to the problem.

The paper is organized as follows. Section 2 gives an overview of related work. In section 3 a new sparsity-based formulation to the robust linear regression problem is introduced. Section 4 presents the tightest, tractable convex relaxation of the proposed objective and investigates the relationship between the new and existing approaches. Section 5 describes how to incorporate prior information through the use of additional constraints in the optimization problem. Section 6 summarizes the proposed algorithm and section 7 illustrates its performance with several application examples using synthetic and real data. Finally, section 8 gives the conclusions.

2. Literature Review

Most current robust linear regression methods can be classified into one of four major classes: median-based, M-estimators, RANSAC, and sparsity-based convex methods.

Median-based approaches try to overcome the limitations of least squares regression by using, instead of the mean, the median of the fitting errors since it is more robust to outliers. For example, LMedS [22] seeks to minimize the median of the squared residues using a random sampling algorithm. However, this algorithm is combinatorial on the number of regression parameters and hence is not suitable for high dimensional regression problems.

An alternative for making least squares regression robust is to use M-estimators [13]. In this approach, the residual error in the maximum likelihood estimate is replaced by a non-linear function of the residuals that penalizes residuals from inliers almost linearly but saturates for residual errors due to outliers. A disadvantage of this approach is that the resulting optimization problem is non-convex. Solving this problem with iterative steepest descent methods is not guaranteed to converge to the true optimum. It has also been proposed to solve this problem using random sampling methods [18]. However, as mentioned above, this approach suffers from its combinatorial complexity.

Perhaps the most commonly used robust regression algorithms belong to the RANSAC family [8, 23, 5, 21, 7]. The main idea behind these approaches is to try to minimize the number of outliers. However, these techniques, like LMedS, rely on random sampling of the data to separate the inliers from the outliers, based on fitting error, and hence are also inherently combinatorial.

More recently, inspired by the success of compressive sensing, it has been proposed to use sparsity ideas to separate outliers from inliers [4, 14, 12, 19]. These methods reformulate the regression problem by minimizing the least square error for the inlier data while enforcing the natural assumption that there are more inliers than outliers data points . While these new formulations are non-convex, they can be relaxed to convex optimization problems which can be solved in polynomial time. The issue of whether the solutions obtained using these relaxations are also solutions to the original regression problem was addressed in [19]. There, the authors computed, for the case when there is no inlier noise, a lower bound of the maximum number of outliers that minimizing the number of outliers (such that the inliers fit the model) can handle. This bound is given by the smallest principal angle θ_k between the regressor subspace and all the k-dimensional outlier subspaces. However, the quality of the solutions obtained by these formulations suffers when the size of some of the outlier errors is very large.

In this paper, a *Self-Scaled Regularized Robust Regression* (S^2R^3) algorithm is proposed. The S^2R^3algorithm belongs to the last category, as it maximizes the number of inliers through a convex relaxation. However, the main advantage of the proposed formulation is that it is not sensitive to the scale of the outlier errors. Indeed, as shown in section 4 the proposed formulation is equivalent to a properly scaled ℓ_1 regularized regression. While the S^2R^3method has the desirable property that the data scaling is done automatically, it also suggests that previous methods would benefit from proper scaling of the data, as well. In addition, in contrast with previous approaches, the S^2R^3method can easily handle prior information such as co-occurrence labeling.

3. Preliminaries

3.1. Notation

$\mathbb{R}$	set of real numbers
$\mathbb{N}_n$	set of positive integers up to n: $\mathbb{N}_n \doteq \{1,\dots,n\}$
$\mathbf{x}$ ($\mathbf{X}$)	a vector (matrix) in $\mathbb{R}^N$ ($\mathbb{R}^{N\times d}$)
$\mathbf{x}^{(j)}$	j$^{\text{th}}$ component of the vector $\mathbf{x}$
$\mathbf{X}^{(j)}$	j$^{\text{th}}$ row of the matrix $\mathbf{X}$
$\Vert\mathbf{x}\Vert_p$	p-norm in $\mathbb{R}^N$
$\Vert\{\mathbf{x}\}\Vert_p$	ℓ_p norm of a vector valued sequence $\{\mathbf{x}_i\}_{t=1}^m$ where each $\mathbf{x}_i \in \mathbb{R}^N$ $\Vert\{\mathbf{x}\}\Vert_p \doteq \left(\sum_{i=1}^m \Vert\mathbf{x}_i\Vert_p^p\right)^{1/p}, 1 \le p < \infty$ $\Vert\{\mathbf{x}\}\Vert_\infty \doteq \max\limits_{1\le i\le m} \Vert\mathbf{x}_i\Vert_\infty$
$\Vert\{\mathbf{x}\}\Vert_o$	ℓ_o-quasinorm $\doteq$ number of non-zero vectors in the sequence (i.e. cardinality of the set $\{i \vert \mathbf{x}_i \neq \mathbf{0}, i \in [1,m]\}$)
$\vert S\vert$	cardinality of the set S
$\mathbf{e_i}$	i^{th} vector of the canonical basis: $\mathbf{e_i} \doteq [0,\dots,1,\dots 0]^T$
$\mathbf{I}_{\mathcal{I}}$	Matrix whose columns are the vectors $\mathbf{e}_i$ corresponding to indexes $i \in \mathcal{I} \subset \mathbb{N}_n$.

3.2. Statement of the problem

Given N data points $\mathbf{x}_i \in \mathbb{R}^d$, corresponding scalars $y_i,\ i = 1,\dots,N$, a vector $\mathbf{r} \in \mathbb{R}^d$, and a noise bound ϵ, define the set of inliers as:

$$\mathcal{S}_i(\mathbf{r}) = \left\{\mathbf{x}_i \colon |y_i - \mathbf{x}_i^T\mathbf{r}| \le \epsilon\right\} \tag{1}$$

The robust regression problem consists on determining a vector $\mathbf{r}$ such that the number of inliers is maximized, that is:

$$\mathbf{r}^* = \underset{\mathbf{r}}{argmax}\, |\mathcal{S}_i(\mathbf{r})| \tag{2}$$

By introducing additional variables $\mathbf{r}_i \in \mathbb{R}^d$ the problem above can be reformulated as:

$$\begin{array}{rcl} \mathbf{r}^* & = & argmin_{\mathbf{r},\mathbf{r}_i} \|\{\mathbf{r}-\mathbf{r}_i\}\|_o \text{ subject to:} \\ & & |y_i - \mathbf{x}_i^T\mathbf{r}_i| \le \epsilon,\ i = 1,\dots,N \end{array} \tag{3}$$

where $\|\{\mathbf{r}-\mathbf{r}_i\}\|_o$ denotes the number of non-zero elements of the vector sequence $\{\mathbf{r}-\mathbf{r}_i\}_{i=1}^N$.

Lemma 1. *Problems* (2) *and* (3) *are equivalent.*

Proof. Given $\mathbf{r}$, define

$$\begin{array}{rl} J(\mathbf{r}) \doteq & \min_{\mathbf{r}_i \in \mathbb{R}^d} \|\{\mathbf{r}-\mathbf{r}_i\}\|_o \text{ subject to} \\ & |y_i - \mathbf{x}_i^T\mathbf{r}_i| \le \epsilon,\ i = 1,\dots,N \end{array} \tag{4}$$

Since $\|\mathbf{r}-\mathbf{r}_i\|_o = 0 \iff \mathbf{r} = \mathbf{r}_i \iff \mathbf{x}_i \in \mathcal{S}_i(\mathbf{r})$, it follows that $\|\{\mathbf{r}-\mathbf{r}_i\}\|_o =$ number of outliers, or equivalently, $|S_i(\mathbf{r})| = N - J(\mathbf{r})$. Thus $\mathbf{r}^*$ maximizes $|\mathcal{S}_i(\mathbf{r})|$, and hence it is a solution to (2), if and only if it is a minimizer of $J(\mathbf{r})$. □

Note that the solution to problem (3) may not be unique. Conditions guaranteeing uniqueness and exact recovery are discussed below.

Theorem 1. *Let $\mathbf{X}^{(i)} \doteq \mathbf{x}_i^T$ and denote by $\mathcal{I}_k$ the set of subsets $S_k \subseteq S \doteq \{1,\dots,N\}$, with $|S_k| = k$. Then, in the noiseless case, if the matrix $[\mathbf{X}\ \mathbf{I}_{\mathcal{I}}]$ has full column rank for all $\mathcal{I} \in \mathcal{I}_k$ and Problem* (3) *admits a solution with $\|\{\mathbf{r}-\mathbf{r}_i\}\|_o < \frac{k}{2}$, the model $\mathbf{r}$ is unique.*

Proof. Define $s_i \doteq \mathbf{x}_i^T(\mathbf{r}_i - \mathbf{r})$ and consider the following related problem:

$$\begin{array}{l} \min_{\mathbf{r},\mathbf{s}} \|\mathbf{s}\|_o \text{ subject to} \\ y_i = \mathbf{x}_i^T\mathbf{r} + s_i,\ i = 1,\dots,N \end{array} \tag{5}$$

Note that in the noiseless case, (5) and (4) have the same constraint set. Since $\|\mathbf{s}\|_0 \le \|\{\mathbf{r}-\mathbf{r}_i\}\|_0$, it follows from the hypothesis and Proposition II.1 in [19], that if (4) admits a m-sparse solution, with $m < \frac{k}{2}$ then the solution to (5) is unique. To finish the proof, assume by contradiction that (4) admits multiple m-sparse solutions with different $\mathbf{r}$. Then, the corresponding vectors $\mathbf{r}$ and $\mathbf{s}$ solve (5), contradicting the fact that this problem has a unique solution. □

4. Main Results

While the results in section 3.2 guarantee exact recovery of the model under some conditions, they require solving problem (3), which can be easily shown to be generically NP-hard. In this section we present a tractable convex relaxation and investigate its relationship with existing approaches.

4.1. Self-Scaled Regularized Robust Regression

Recall that the convex envelope (that is the tightest convex approximation) of the cardinality of a vector sequence $\{\mathbf{v}_i\}$ [20] is given by:

$$\|\{\mathbf{v}\}\|_{0,env} = \sum_i \|\mathbf{v}_i\|_\infty \tag{6}$$

It follows that replacing $\|\{\mathbf{r}-\mathbf{r}_i\}\|_0$ by $\sum_{i=1}^N \|\mathbf{r}-\mathbf{r}_i\|_\infty$ provides the tightest convex approximation to the objective function, motivating the following convex relaxation of (3)

$$\begin{array}{l} \min_{\mathbf{r},\mathbf{r}_i} \sum_{i=1}^N \|\mathbf{r}-\mathbf{r}_i\|_\infty \text{ subject to:} \\ |y_i - \mathbf{x}_i^T\mathbf{r}_i| \le \epsilon,\ i = 1,\dots,N \end{array} \tag{7}$$

As we show next, in the absence of additional constraints, the problem above is equivalent to a suitable scaled traditional ℓ_1-regularized robust regression. Thus, in the sequel we will refer to problem (7) as the *self-scaled regularized robust regression* problem (S^{2}R^3).

Theorem 2. *Problem* (7) *is equivalent to the following optimization problem:*

$$\begin{array}{l} \min_{\mathbf{r},\boldsymbol{\eta}} \sum_{i=1}^{N} \frac{|y_i - \mathbf{x}_i^T \mathbf{r} + \eta_i|}{\|\mathbf{x}_i\|_1} \\ \text{subject to } |\eta_i| \leq \epsilon,\ i = 1, \ldots, N \end{array} \tag{8}$$

Proof. Rewriting the constraints in (7) as $y_i = \mathbf{x}_i^T \mathbf{r}_i + \eta_i$ for some $|\eta_i| \leq \epsilon$ leads to

$$y_i = \mathbf{x}_i^T (\mathbf{r}_i - \mathbf{r}) + \mathbf{x}_i^T \mathbf{r} + \eta_i$$

Thus

$$|\mathbf{x}_i^T (\mathbf{r} - \mathbf{r}_i)| = |\mathbf{x}_i^T \mathbf{r} + \eta_i - y_i| \tag{9}$$

Since the ℓ_1 and ℓ_∞ norms are dual [17], from the equation above it follows that

$$\begin{array}{rcl} \|\mathbf{x}_i\|_1 \|\mathbf{r} - \mathbf{r}_i\|_\infty & \geq & |\mathbf{x}_i^T \mathbf{r} + \eta_i - y_i| \Rightarrow \\ \|\mathbf{r} - \mathbf{r}_i\|_\infty & \geq & \frac{|\mathbf{x}_i^T \mathbf{r} + \eta_i - y_i|}{\|\mathbf{x}_i\|_1} \end{array} \tag{10}$$

(with equality holding when $\mathbf{x}$ and $\mathbf{r} - \mathbf{r}_i$ are aligned). For fixed $y_i, \mathbf{r}, \eta_i$, consider now the following minimization problem:

$$\min_{\mathbf{r}_i} \|\mathbf{r} - \mathbf{r}_i\|_\infty \text{ subject to: } \mathbf{x}_i^T \mathbf{r}_i + \eta_i - y_i = 0 \tag{11}$$

We claim that the solution to this problem is given (component-wise) by

$$\tilde{\mathbf{r}}_i^{(j)} = \mathbf{r}^{(j)} - \frac{\mathbf{x}_i^T \mathbf{r} - y_i + \eta_i}{\|\mathbf{x}_i\|_1} sign(\mathbf{x}_i^{(j)}) \tag{12}$$

To show this, note that $\tilde{\mathbf{r}}_i$ is a feasible solution of (11), and such that each component of the difference vector $\mathbf{r} - \tilde{\mathbf{r}}_i$ satisfies:

$$|\mathbf{r}^{(j)} - \tilde{\mathbf{r}}_i^{(j)}| = \frac{|\mathbf{x}_i^T \mathbf{r} - y_i + \eta_i|}{\|\mathbf{x}_i\|_1}, j = 1, \ldots, d$$

and hence

$$\|\mathbf{r} - \tilde{\mathbf{r}}_i\|_\infty = \frac{|\mathbf{x}_i^T \mathbf{r} - y_i + \eta_i|}{\|\mathbf{x}_i\|_1}, j = 1, \ldots, d \tag{13}$$

since, from(10), this is the lowest possible value of the objective, optimality of $\tilde{\mathbf{r}}_i$ follows. Replacing each term in the objective function in (7) by its optimal value leads to:

$$\begin{array}{cc} \begin{array}{c} \min \\ \mathbf{r}, \mathbf{r}_i, |\eta_i| \leq \epsilon \\ y_i = \mathbf{x}_i^T \mathbf{r}_i + \eta_i \end{array} & \sum_{i=1}^{N} \|\mathbf{r} - \mathbf{r}_i\|_\infty = \\ \min\limits_{\mathbf{r}, |\eta_i| \leq \epsilon} & \sum_{i=1}^{N} \frac{|\mathbf{x}_i^T \mathbf{r} + \eta_i - y_i|}{\|\mathbf{x}_i\|_1} \end{array} \tag{14}$$

□

4.2. Connections with regularized ℓ_1 robust regression

By introducing an outlier error vector $\mathbf{s}$, problem (2) can be reformulated as:

$$\begin{array}{l} \min_{\mathbf{r},\boldsymbol{\eta},\mathbf{s}} \|\mathbf{s}\|_o \text{ subject to:} \\ \mathbf{y} = \mathbf{X}\mathbf{r} + \boldsymbol{\eta} + \mathbf{s},\ \|\boldsymbol{\eta}\|_\infty \leq \epsilon \end{array} \tag{15}$$

where $\mathbf{X}^{(i)} = \mathbf{x}_i^T$. Since this problem is known to be NP hard, a convex relaxation can be obtained by using the ℓ_1 norm as surrogate for cardinality, leading to the ℓ_1 regularized robust regression problem introduced in [19].

$$\begin{array}{l} \min_{\mathbf{r},\boldsymbol{\eta}} \|\mathbf{s}\|_1 \text{ subject to:} \\ \mathbf{y} = \mathbf{X}\mathbf{r} + \boldsymbol{\eta} + \mathbf{s},\ \|\boldsymbol{\eta}\|_\infty \leq \epsilon \end{array} \tag{16}$$

From (8), (16) and Theorem 2, it follows that, in the unconstrained case, (7) can be considered as a scaled version of (16), where each data point is automatically scaled by its ℓ_1 norm. As we will illustrate with several examples, this scaling prevents small groups of outliers, far from the inlier manifold, from "poisoning" the optimization, hence leading to better fitting.

4.3. Exact Recovery Conditions and Bounds on the Estimation Error

From Theorem 2, it follows that the results in [19] can be directly applied to establish bounds on the norm of the difference between the solutions to (3) and its convex relaxation (7). To this effect, begin by defining the normalized data matrix $\mathbf{X}_n$, with rows given by $\mathbf{X}_n^{(i)} = \frac{\mathbf{x}_i^T}{\|\mathbf{x}_i\|_1}$. Next, perform a reduced QR decomposition $\mathbf{X}_n = \mathbf{Q}\mathbf{R}$, where $\mathbf{Q}$ is orthonormal and $\mathbf{R}$ is upper diagonal, and define $\mathbf{z} \doteq \mathbf{R}\mathbf{r}$. Proceeding as in [19], we will first find the estimation error $\Delta\mathbf{z}$ and use it to bound the estimation error $\Delta\mathbf{r}$. Define the isometry constant δ_k of $\mathbf{Q}$ as the smallest real such that

$$(1 - \delta_k)\|\mathbf{v}\|^2 \leq \|[\mathbf{Q}\ \mathbf{I}_{\mathcal{I}}]\mathbf{v}\|_2^2 \leq (1 + \delta_k)\|\mathbf{v}\|^2$$

for all $\mathcal{I}$ with $|\mathcal{I}| \leq k$, where $\mathbf{v}^T \doteq [\mathbf{z}^T \mathbf{s}^T]$ and $\mathbf{s} \in \mathbb{R}^k$.

Corollary 1. *Assume that* $\delta_{2k} < \frac{2}{3}$. *Then, the estimation error in* $\mathbf{z}$ *is bounded by*

$$\|\Delta\mathbf{z}\|_2^2 \leq C_o \|\hat{\mathbf{s}} - \hat{\mathbf{s}}^k\|_1 + C_1 \epsilon \tag{17}$$

where $C_o = \frac{2\delta_{2k}}{\sqrt{k}(1 - 1.5\delta_{2k})}$, $C_1 = \frac{\sqrt{1 + 2\delta_{2k}}}{1 - 1.5\delta_{2k}} (\sum\limits_{i=1}^{n} \|\mathbf{x}_i\|_1^{-1})^{\frac{1}{2}}$, $\hat{s}_i \doteq \frac{s_i}{\|\mathbf{x}_i\|_1}$, *denote the components of weighted true outlier approximation error vector and where* $\hat{\mathbf{s}}^k$ *denote the best (in the* ℓ_1 *sense), k-sparse approximation to* $\hat{\mathbf{s}}$. *In particular, if* $\|\hat{\mathbf{s}}\|_o \leq k$, *in the noiseless case,* (7) *recovers the exact* $\mathbf{z}$ *(and hence* $\mathbf{r}$*).*

Proof. Follows from Theorem 2 and Theorem II.1 in [19] by noting that (8) can be rewritten as:

$$\min_{\mathbf{r},\boldsymbol{\eta}} \|\hat{\mathbf{s}}\|_1 \text{ subject to: } \hat{\mathbf{y}} = \mathbf{X}_n\mathbf{r} + \hat{\boldsymbol{\eta}} + \hat{\mathbf{s}},$$

where $\hat{y}_i \doteq \frac{y_i}{\|\mathbf{x}_i\|_1}$, $\hat{s}_i \doteq \frac{s_i}{\|\mathbf{x}_i\|_1}$, and $\hat{\eta}_i \doteq \frac{\eta_i}{\|\mathbf{x}_i\|_1}$. □

Remark 1. *Note that, for inliers $s_i = s_i^k = 0$, and thus the scaling has no effect on the error bounds. A similar reasoning holds for the outliers corresponding to the largest k components of $\hat{\mathbf{s}}$, since here $\hat{s}_i = \hat{s}_i^k$. Indeed, it can be shown that as long as the only data points with $\|\mathbf{x}_i\|_1 < 1$ are amongst those corresponding to the k largest components of $\mathbf{s}$, then the solution to* (7) *yields a smaller upper bound on the estimation error than the solution to the classical ℓ_1 regularized regression.*

5. Incorporating priors

In many situations of practical interest, additional a-priori information is available that can be exploited to improve performance. In the sequel, we illustrate the ability of the proposed algorithm to exploit priors, using two commonly occurring scenarios.

5.1. Using co-occurrence information

Consider the case where it is known that certain sets of points should have the same label. An example of this situation arises for instance in motion-based segmentation problems, where often it is known that a group of points belongs to either the target or the background. As we show in the sequel, this information can be easily incorporated as additional constraints in the formulation (7). On the other hand, traditional ℓ_1 regularized regression cannot exploit this information, since the problem is formulated in terms of error indicator variables s_i, rather than candidate model parameters $\mathbf{r}_i$. Specifically, let $\mathcal{I}$ denote the set of indices of points $\mathbf{x}_i$ that should have the same label and denote by $\mathbf{X}_\mathcal{I}$ and $\mathbf{y}_\mathcal{I}$ the sub matrix of $\mathbf{X}$ formed by considering only the rows indexed by elements of $\mathcal{I}$, and the vector formed by the corresponding elements of $\mathbf{y}$, respectively. Consider first the noiseless case and assume that $\mathbf{y}_\mathcal{I} \in$ span-col$(\mathbf{X}_\mathcal{I})$ and rank$(\mathbf{X}_\mathcal{I}) \leq d$[1]. Under these conditions, there exist at least one $\mathbf{r}^*$ such that $\mathbf{y}_\mathcal{I} = \mathbf{X}_\mathcal{I}\mathbf{r}^*$. Thus adding the constraints $\mathbf{r}_i = \mathbf{r}_\mathcal{I}\ \forall\ i \in \mathcal{I}$ (enforced by simply using the same variable $\mathbf{r}_\mathcal{I}$ in all terms in (7) involving elements of $\mathcal{I}$, does not change the optimal solution. This follows from the fact that $\mathbf{r}_\mathcal{I}$ can be set to $\mathbf{r}$ if the points indexed by $\mathcal{I}$ are inliers, or to $\mathbf{r}^*$ if they are outliers, without changing the value of the objective. In the case of noisy data, the same reasoning can be applied as long as there exists some vector $\boldsymbol{\eta}_\mathcal{I}$, with $\|\boldsymbol{\eta}_\mathcal{I}\|_\infty \leq \epsilon$ such that $\mathbf{y}_\mathcal{I} - \boldsymbol{\eta}_\mathcal{I} \in$ span-col$(\mathbf{X}_\mathcal{I})$. As before, this condition holds trivially as long as $|\mathcal{I}| \leq d$.

5.2. Non-full rank $\mathbf{X}$

Conventional robust regression typically considers the case where $\mathbf{X}$ is full rank. However, this assumption does not always hold in practice. Indeed, several practical problems involve considering the case where $\mathbf{y}_i = 0$, and hence, if non-trivial solutions to (7) exist, they are not unique. An example of this situation is the problem of estimating the fundamental matrix [11], where the solution is unique up to a scaling factor. In these cases, in order to avoid ambiguities, it is of interest to impose additional constraints on $\mathbf{r}$. One such class of constraints is of the form $\mathbf{u}^T\mathbf{r} = 1$, for some suitable chosen $\mathbf{u}$. For instance, the choice $\mathbf{u} = \mathbf{1}$ leads to the constraint $\sum \mathbf{r}^{(j)} = 1$, while $\mathbf{u} = \mathbf{e}_i$ corresponds to $\mathbf{r}^{(i)} = 1$, both used in the context of fundamental matrix estimation.

In this case, it can be shown, by computing sub gradients, that the optimal solutions to (7) for the case $\sum \mathbf{r}_i^{(j)} = 1$ is:

$$\tilde{\mathbf{r}}_i^{(j)} = \mathbf{r}^{(j)} - (\mathbf{x}_i^T\mathbf{r} + \eta_i - y_i)\frac{sign(\mathbf{x}_i^{(j)} - \mathbf{x}_{median})}{\|\mathbf{x}_i - \mathbf{x}_{median}\|_1} \tag{18}$$

with associated cost

$$\|\tilde{\mathbf{r}}_i - \mathbf{r}\|_\infty = \frac{|\mathbf{x}_i^T\mathbf{r} + \eta_i - y_i|}{\|\mathbf{x}_i - \mathbf{x}_{median}\|_1} \tag{19}$$

leading again to a modified regularized ℓ_1 regression, where each term is now scaled by the factor $\|\mathbf{x}_i - \mathbf{x}_{median}\|_1$.

6. Self Scaled Regularized Regression Algorithm

Theorem 1 provides sufficient conditions guaranteeing that solving (7) will lead to the sparsest $\{\mathbf{r} - \mathbf{r}_i\}$ sequence and hence result in a model that maximizes the number of inliers. However, in many practical situations these conditions may not hold. In these cases, sparse solutions can be obtained by using a reweighted heuristic [16], leading to the following algorithm[2]

Algorithm 1 Reweighted Self Scaled Regression

```
$w^0 \leftarrow \{1, \ldots, 1\}$, $\tau = 10^{-2}$        ▷ Initialization
repeat
    Solve
        $\{\mathbf{r}^k, \mathbf{r}_i^k\} = \text{argmin}_{\mathbf{r},\mathbf{r}_i} \sum_{i=1}^N w^k(i)\|\mathbf{r} - \mathbf{r}_i\|_\infty$
                          s.t. $|y_i - \mathbf{x}_i^T\mathbf{r}_i| \leq \epsilon$,
                               $i = 1, \ldots, N$
    Update $w^{k+1}(i) = (\|\mathbf{r}^k - \mathbf{r}_i^k\|_\infty + \tau)^{-1}$.
until convergence
```

[1] This situation holds trivially when $|\mathcal{I}| \leq d$).

[2] A similar algorithm was proposed in [20] in the context of systems identification, but without an analysis of its recovery properties or relationship to traditional ℓ^1 regularized robust regression.

7. Experimental Results

In this section we describe two sets of experiments to evaluate the performance of the proposed algorithm. The first set of experiments uses synthetic data to fit a hyperplane while the second set uses real data to reconstruct corrupted face images from the Yale face dataset. In all cases, performance is compared against 8 existing regression methods that range from classic techniques using random approaches to state of the art convex formulations.

Randomized Algorithms. The methods using random approaches we compared against are: RANSAC, MSAC, MLEASAC and LMEDS. To ensure that comparisons are fair, i.e. all the methods are solving a robust regression problem, these techniques were used to solve the problem below:

$$\min_{r} \quad \|s\|_0 \quad s.t. \quad \|y - Xr + s\|_\infty \leq \epsilon$$

The most critical parameters for randomized algorithms are the inlier noise bound and the number of iterations. We set the inlier noise bound for all randomized algorithms equal to the inlier noise bound of convex formulations. In other words, all algorithms shared same inlier noise bound. The number of iterations was set to 500 for all noise levels and for all algorithms. We used the implementation from GML Toolbox from [1].

M-estimator. M-estimator is a standard robust regression method. We used the MATLAB implementation with "huber" weighting function, which is the common setup for it. We found the best parameter for M-estimator by line search in one dataset and use it for the corresponding experiment.

Constrained RPCA: Robust PCA is a recent robust regression method proposed by [2]. We modified the original formulation inline with [24]. In our formulation we find the smallest penalty parameter that gives a rank deficient data matrix. This step removes both inliers and outliers of the data and makes sure it has a null space of dimension 1. Then, we use the null space vector as the final model.

RR [12]: This formulation is similar to RPCA formulations with the inclusion of the model term in the optimization function. Their extended formulation is bilinear and solved with ALM. We implemented our own version following their supplementary material. This algorithm has 2 parameters which are difficult to tune. We used grid search to find the best setup given a dataset from each experiment.

BPRR and BSRR [19]: These are the most recent formulations for robust regression and they are the closest works to ours. We implemented them using CVX Toolbox [10]. The only parameter these formulations require is the inlier noise bound (same as ours).

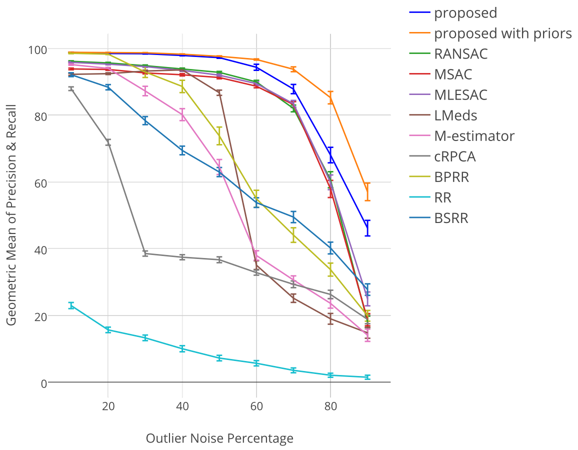

Figure 2. Synthetic Data Experiments: Fitting a 5-dimensional hyperplane. The plots show the Geometric Mean of Precision and Recall for all the evaluated algorithms.

7.1. Synthetic Data Experiments

This set of experiments attempts to recover hyperplanes from data corrupted by outliers. The data was generated as follows. First, a vector $\mathbf{r}$ was drawn using a Normal distribution $N(\mathbf{0}, \mathbf{I})$. Then, the input samples $\mathbf{x}_i$ were uniformly sampled from $[0,1]^{m-1}$, where $m = 5$ is the dimension of the data. Next, the outputs y_i were computed as $y_i = \mathbf{x}_i\mathbf{r} + e_i$, with e_i uniformly distributed from $[-\epsilon, \epsilon]$, where $\epsilon = 0.1$. Finally, the outliers were seeded by randomly sampling y_i and $\mathbf{x}_i$ from $N(0, 15)$ and $N(0, 1)$, respectively.

In all the experiments, the inlier noise bound was set to the value used to generate the data. The number of outliers was varied from 10% to 90%, in increments of 10%. The algorithm was run 100 times for each level of outliers. Performance was compared using two performance scores: geometric mean of precision and recall, and the regression recovery error.

Table 1. Running times for the experiments with synthetic data.

Method	Implementation	Times
proposed	Gurobi(LP)	0.0266
RANSAC	MATLAB	0.0491
MSAC	MATLAB	0.0495
MLESAC	MATLAB	0.1282
LMeds	MATLAB	0.0580
M-estimator	MATLAB	0.0080
cRPCA	MATLAB(ADMM)	2.5346
BPRR	CVX	0.9859
RR	MATLAB(ADMM)	1.9388
BSRR	MATLAB	0.7307

The results of this set of experiments can be seen in Figures 2 and 3 and the running times are given in Table 1.

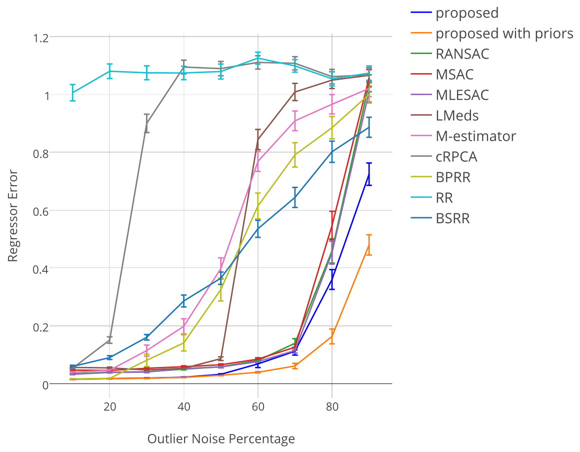

Figure 3. Synthetic Data Experiments: Fitting a 5-dimensional hyperplane. The plots show the Model Error for all the evaluated algorithms.

Note that our algorithm performs the best, both with low and high percentages of outliers. On the other hand, randomized algorithms show a significant performance drop when the percentage of outliers is 70% and above, showing the advantage of our formulation. Furthermore, it should be noted that not all the convex formulations have similar robustness under heavy outlier noise. In particular, the early failure of the BPRR algorithm illustrates the importance of the self scaling property of the proposed approach.

Using Priors. To evaluate the impact of using priors we proceeded as follows. After a run without priors was done, no more than half of the false positive points were paired randomly with a true negative point, and no more than half of the false negative points were paired randomly to a true point. As seen in Figure 2 and 3, the ability to incorporate the additional co-occurrence information can boost the performance of the proposed algorithm between 5 to 10 percent under heavy outlier noise.

7.2. Real Data Experiments

This set of experiments attempts to reconstruct face images that have been corrupted with heavy occlusion, where the occluding pixels constitute the outliers. The data used for these experiments is from the CroppedYale Dataset [15]. The dataset contains 38 subjects. We choose 8 face images per person, taken under mild illumination conditions and computed an eigenface set with 20 eigenfaces. Then, the goal of these experiments was: given a corrupted face image of a subject in the database (this (uncorrupted) image was not used to compute the eigenspace), get the best reconstruction/approximation of the true face image.

We reconstructed one image per person. Occlusion was simulated by randomly placing 10 blocks of size 30×30. To increase the difficulty of the problem and reduce the dimensionality, data was randomly sampled (400 pixels from the image and the basis). The performance of the algorithms was evaluated using the Root Mean Square metric (Table 2),

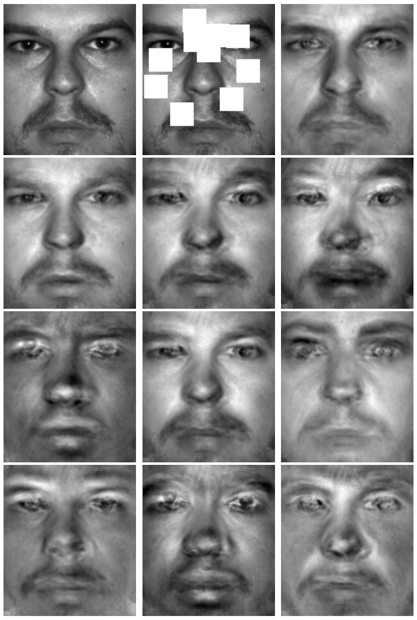

Figure 4. Face recovery results: In order from left to right, top to bottom: original image, occluded image, best possible recovery with given basis, proposed, BPRR, cRPCA, LMedS, Mestimator, MLESAC, MSAC, RANSAC, and RR.

$$RMS(I, \hat{I}) = \sqrt{||I - \hat{I}||_F^2 / N_{pixels}}$$

where I is the original image without occlusion and $\hat{I}$ is the reconstructed image. A visual comparison for one instance of recovery using all the evaluated methods is shown in Figure 4. We normalized all images to [0, 1] range to remove scaling effects of the pixel values on the RMS metric. We also computed a best possible reconstruction of the original face image by using the 20 eigenfaces. We used the model of this step as the ground truth model and computed the model recovery error as in the synthetic experiments (Table 3). The experiments show that the mean RMS and the model error are the best for our method and that the recovered images are visually closer to the un-occluded original image.

Finally, we ran another set of experiments where we gave all the *SAC algorithms (RANSAC, MSAC, MLESAC, LMeds) some extra time. For these experiments, we set the number of iterations so that these algorithms could use as much time or longer than the time used by the proposed algorithm. While the extra time improved the performance of the *SAC algorithms it was not enough to achieve the best performance, as summarized in Tables 4 and 5.

Table 2. Fitting to original image error.

	proposed	BPRR	BSRR	M-est.	RR	cRPCA	MLESAC	MSAC	RANSAC	LMedS
Mean RMS	**0.1320**	0.1397	0.1378	0.1345	0.1844	0.1854	0.1751	0.1773	0.1690	0.1835
stdev	0.0074	0.0052	0.0081	0.0074	0.0054	0.0071	0.0082	0.0067	0.0064	0.0085

Table 3. Model estimation error.

	proposed	BPRR	BSRR	M-est.	RR	cRPCA	MLESAC	MSAC	RANSAC	LMedS
Mean RMS	**0.7105**	0.9092	0.7428	0.7232	1.0663	1.0761	1.0682	1.0917	1.1013	1.0528
stdev	0.0382	0.0472	0.0533	0.0435	0.0441	0.0506	0.0329	0.0337	0.0338	0.0353

Table 4. Fitting to original image error (allowing extra time to the *SAC algorithms).

	proposed	BPRR	BSRR	M-est.	RR	cRPCA	MLESAC	MSAC	RANSAC	LMedS
Mean RMS	**0.1320**	0.1397	0.1378	0.1345	0.1844	0.1854	0.1704	0.1545	0.1588	0.1661
stdev	0.0074	0.0052	0.0081	0.0074	0.0054	0.0071	0.0069	0.0064	0.0074	0.0082
run time	1.5088	1.6553	51.1901	**0.0343**	19.5540	0.3533	3.3083	1.5997	1.5864	1.7923

Table 5. Model estimation error (allowing extra time to the *SAC algorithms).

	proposed	BPRR	BSRR	M-est.	RR	cRPCA	MLESAC	MSAC	RANSAC	LMedS
Mean RMS	**0.7105**	0.9092	0.7428	0.7232	1.0663	1.0761	0.8719	0.9098	0.9045	0.9183
stdev	0.0382	0.0472	0.0533	0.0435	0.0441	0.0506	0.0392	0.0366	0.0305	0.0369
run time	1.5088	1.6553	51.1901	**0.0343**	19.5540	0.3533	3.3083	1.5997	1.5864	1.7923

8. Conclusions

Robust regression is at the core of a large number of computer vision problems ranging from recovering 3D geometry, to classification and image reconstruction. While this problem has been the object of a very large research effort, it remains challenging in scenarios characterized by noisy correspondences and high percentage of gross outliers. The main result of this paper is a computationally tractable regression algorithm specifically tailored to this situation. Contrary to other sparsification based approaches, the proposed algorithm seeks to directly sparsify the set of models that explain the data, rather than the set of outlier errors. The intuition behind this approach is that this set of models can be normalized so that all its elements have comparable magnitude, a fact that prevents gross outliers from skewing the results. Surprisingly, as shown in the paper, the proposed approach is equivalent to a *self-scaled* robust regression, where the data points are automatically scaled by a problem dependent quantity, providing an alternative explanation of the reason behind its improved performance in the presence of gross outliers. In addition, working directly with models (rather than outlier errors) allows for exploiting existing a-priori information about co-occurrences to improve the resulting model, a feature hitherto beyond the ability of existing regression techniques. As shown in the paper, the combination of self-scaling and the ability to exploit priors allows the proposed algorithm to consistently outperform existing techniques, regardless of the percentage of outliers.

References

[1] GML RANSAC Matlab Toolbox.

[2] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[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*, 52(2):489–509, 2006.

[4] E. J. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. Inf. Theory*, 51(12):4203–4215, 2005.

[5] O. Chum, J. Matas, and J. Kittler. Locally optimized ransac. In *Pattern Recognition*, pages 236–243. Springer, 2003.

[6] D. L. Donoho. Compressed sensing. *IEEE Trans. Inf. Theory*, 52(4):1289–1306, 2006.

[7] O. Enqvist, E. Ask, F. Kahl, and K. Åström. Tractable algorithms for robust model estimation. *International Journal of Computer Vision*, pages 1–15, 2014.

[8] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[9] D. Freeman. *Statistical Models: Theory and Practice*. Cambridge University Press, 2009.

[10] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 2.1, Mar. 2014.

[11] R. Hartley and A. Zisserman. *Multiple View Geometry*. Cambridge University Press, 2003.

[12] D. Huang, R. Silveira, and F. D. la Torre. Robust regression. In *European Conf. on Computer Vision (ECCV)*. Elsevier, 2012.

[13] P. Huber. *Robust Statistics*. Wiley, 1981.

[14] Y. Jin and B. D. Rao. Algorithms for robust linear regression by exploiting the connection to sparse signal recovery. In *Int. Conf. Acoust, Speech, Signal Procsess (ICASSP)*, pages 3830–3833, 2010.

[15] K. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 27(5):684–698, 2005.

[16] M. Lobo, M. Fazel, and S. Boyd. Portfolio optimization with linear and fixed transaction costs. *Annals of Operations Research*, 152(1):376–394, July 2007.

[17] D. G. Luenberger. *Optimization by vector space methods*. Decision and control. Wiley, New York, NY, 1969.

[18] R. A. Maronna, R. D. Martin, and V. J. Yohai. *Robust Statistics: Theory and Methods*. Wiley, 2006.

[19] K. Mitra, A. Veeraraghavan, and R. Chellappa. Analysis of sparse regularization based robust regression algorithms. *IEEE Trans. Signal Processing*, 61(5):1249–1257, 2013.

[20] N. Ozay, M. Sznaier, C. M. Lagoa, and O. I. Camps. A sparsification approach to set membership identification of switched affine systems. *Automatic Control, IEEE Transactions on*, 57(3):634–648, 2012.

[21] R. Raguram, J.-M. Frahm, and M. Pollefeys. A comparative analysis of ransac techniques leading to adaptive real-time random sample consensus. In *Computer Vision–ECCV 2008*, pages 500–513. Springer, 2008.

[22] P. J. Rousseeuw. Least median of squares regression. *Journal of the American statistical association*, 79(388):871–880, 1984.

[23] P. H. Torr and A. Zisserman. Mlesac: A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding*, 78(1):138–156, 2000.

[24] H. Xu, C. Caramanis, and S. Sanghavi. Robust pca via outlier pursuit. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 2496–2504. Curran Associates, Inc., 2010.