

# ROBUST REGRESSION USING SPARSE LEARNING FOR HIGH DIMENSIONAL PARAMETER ESTIMATION PROBLEMS

*Kaushik Mitra, Ashok Veeraraghavan\*, Rama Chellappa*

Department of Electrical and Computer Engineering  
University of Maryland, College Park

\*Mitsubishi Electric Research Laboratories(MERL)  
Cambridge MA

## ABSTRACT

Algorithms such as Least Median of Squares (LMedS) and Random Sample Consensus (RANSAC) have been very successful for low-dimensional robust regression problems. However, the combinatorial nature of these algorithms makes them practically unusable for high-dimensional applications. In this paper, we introduce algorithms that have cubic time complexity in the dimension of the problem, which make them computationally efficient for high-dimensional problems. We formulate the robust regression problem by projecting the dependent variable onto the null space of the independent variables which receives significant contributions only from the outliers. We then identify the outliers using sparse representation/learning based algorithms. Under certain conditions, that follow from the theory of sparse representation, these polynomial algorithms can accurately solve the robust regression problem which is, in general, a combinatorial problem. We present experimental results that demonstrate the efficacy of the proposed algorithms. We also analyze the intrinsic parameter space of robust regression and identify an efficient and accurate class of algorithms for different operating conditions. An application to facial age estimation is presented.

*Index Terms*— Robust Regression, Sparse Representation, Sparse Bayesian Learning

## 1. INTRODUCTION

The goal of regression is to estimate the parameters of a model relating two sets of variables, given a training dataset. However, the presence of outliers in the training dataset will make this estimate unreliable. Outliers are those data that differ markedly from other data present in the dataset. Real world data is almost always corrupted with outliers and hence robust parameter estimation is of paramount importance.

Solving the robust regression problem requires estimating the parameters of the model, and identifying the outliers. Since any subset of the data could be outliers, this is a combinatorial problem in general. There are two major approaches [1] for solving the robust regression problem: *i*) Estimate the parameters of the model using a robust cost function and then identify the outliers as data that deviate by a large amount from the model. M-estimators [2], LMedS [1] and RANSAC [3] follow this approach. The second approach is: *ii*) First identify the outliers, remove them and then use a (non-robust)

regression algorithm like Least Squares (LS) to estimate the model parameters.

M-estimators [2] are a generalization of maximum likelihood estimators (MLEs) where the (negative) log likelihood function of the data is replaced by a robust cost function. Amongst the many possible choices of cost functions, redescending cost functions [2] are the most robust ones. These cost functions are non-convex and the resulting non-convex optimization problem has many local minima. Generally, a polynomial algorithm called *iteratively reweighted least squares* (IRLS) [2] is used for solving the optimization problem which often converges to local minima. The quality of parameter estimation depends on a good initialization which itself is very difficult to obtain, especially for high-dimensional problems. The LMedS technique [1] is another widely used robust regression method. In LMedS, the median of the squared residuals is minimized. A random sampling algorithm [1] is used for solving this problem. This algorithm is combinatorial in the dimension (number of the parameters) of the problem which makes LMedS impractical for high-dimensional regression problems. The RANSAC algorithm [3] and its improvements such as MSAC, MLESAC [4] are the most widely used and successful robust methods in computer vision [5]. 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. A similar random sampling algorithm is used for solving this problem which makes RANSAC, MSAC and MLESAC impractical for high-dimension problems.

There are many heuristic methods that follow the second approach of first identifying the outliers, however, most of them fail in the presence of multiple outliers [1]. In this paper, we take a systematic approach towards outlier identification. We formulate the robust regression problem by projecting the dependent variable onto the null space of the independent variables. This projection receives significant contributions only from the outliers. The robust regression problem in this form is exactly of the same form as the sparse representation/learning problem [6, 7, 8, 9]. We then use polynomial algorithms such as Basis Pursuit [6, 7] and the sparse Bayesian learning algorithm [8, 9] to solve the robust regression problem. Under certain conditions [7, 10], these algorithms can accurately solve the sparse learning problem which in turn implies an accurate solution for the robust regression problem. We also undertake theoretical and empirical studies of the intrinsic parameter space of the robust regression problem to identify efficient and accurate class of algorithms for different operating conditions. By intrinsic parameters of robust regression, we mean: the outlier fraction  $f$ , the dimensionality of the problem  $D$ , the number of data points  $N$  and the inlier noise variance  $\sigma^2$ .

We would like to note that a similar mathematical formulation arises when considering channel coding problem over the reals, which is addressed in [11].

## 2. ROBUST REGRESSION AND SPARSE LEARNING

We consider a simple linear regression model, where the dependent variable  $y$  is related to the independent variable  $\mathbf{x}$  and a parameter vector  $\mathbf{w}$  through the relation:

$$\begin{aligned} y &= w_0 + x^1 w^1 + \dots + x^D w^D + e \\ &= \mathbf{x}^T \mathbf{w} + e \end{aligned} \quad (1)$$

where  $e$  is the observation noise. In regression, the objective is to estimate  $\mathbf{w}$  from a training dataset of  $N$  observations  $(y_i, \mathbf{x}_i), i = 1, 2, \dots, N$ . In the absence of outliers,  $e_i = n_i$  where  $n_i$ 's are generally considered as independent, zero mean Gaussian noise. In the presence of outliers, the observation error  $e_i$  can be modeled as  $e_i = n_i + s_i$  where  $n_i$  and  $s_i$  are independent of each other,  $n_i$  is the inlier Gaussian noise and  $s_i$  represents the outlier error. The regression model then becomes:

$$y_i = \mathbf{x}_i^T \mathbf{w} + n_i + s_i$$

The above model can also be written as

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{n} + \mathbf{s} \quad (2)$$

where  $\mathbf{y} = (y_1, \dots, y_N)^T$ ,  $\mathbf{n} = (n_1, \dots, n_N)^T$ ,  $\mathbf{s} = (s_1, \dots, s_N)^T$  and  $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N]^T$ .

Our goal is to estimate  $\mathbf{w}$  while being robust to outliers. Similar to RANSAC, a robust way to do this will be to find a  $\mathbf{w}$  which minimizes the number of outliers. Mathematically this can be stated as:

$$\min_{\mathbf{s}, \mathbf{w}} \|\mathbf{s}\|_0 \text{ such that } \|\mathbf{y} - \mathbf{X}\mathbf{w} - \mathbf{s}\|_2 \leq \mu \quad (3)$$

where  $\|\mathbf{s}\|_0$  is the  $L_0$  norm of  $\mathbf{s}$  which counts the number of non-zero elements in  $\mathbf{s}$ ,  $\mu$  is a threshold which depends on the variance of the inlier noise  $\mathbf{n}$ . The problem in (3) has two unknowns:  $\mathbf{w}$  and  $\mathbf{s}$ . It can be simplified by removing one of the unknowns, the parameter  $\mathbf{w}$ , by projecting  $\mathbf{y}$  onto the left null space of  $\mathbf{X}$ . The matrix  $\mathbf{X}$  has dimension  $N \times D$  where the rows correspond to the  $N$  data points and the columns correspond to the  $D$  dimensions, with  $N > D$ . The  $N$ -dimensional columns of  $\mathbf{X}$  span a  $D$ -dimensional subspace, known as the column space. This is assuming that the columns are linearly independent; if they are not, then one can reduce the dimensionality of the problem by eliminating the dependent columns. The orthogonal complement to the column space of  $\mathbf{X}$  is the left null space of  $\mathbf{X}$  which is a  $(N - D)$  dimensional subspace. Let  $\mathbf{C}$  be a matrix whose rows form an orthonormal basis for the left null space of  $\mathbf{X}$ , that is,  $\mathbf{C}$  is a  $(N - D) \times N$  matrix with  $\mathbf{C}\mathbf{X} = \mathbf{0}$ . Pre-multiplying (2) by  $\mathbf{C}$ , we get

$$\begin{aligned} \mathbf{C}\mathbf{y} &= \mathbf{C}\mathbf{X}\mathbf{w} + \mathbf{C}\mathbf{n} + \mathbf{C}\mathbf{s} \\ \mathbf{z} &= \mathbf{g} + \mathbf{C}\mathbf{s} \end{aligned} \quad (4)$$

where  $\mathbf{z} = \mathbf{C}\mathbf{y}$  and  $\mathbf{g} = \mathbf{C}\mathbf{n}$ , which is again Gaussian noise.

The robust regression problem stated in (3) now becomes:

$$\min_{\mathbf{s}} \|\mathbf{s}\|_0 \text{ such that } \|\mathbf{z} - \mathbf{C}\mathbf{s}\|_2 \leq \epsilon \quad (5)$$

where  $\epsilon = \mu\sqrt{(N - D)/N}$ . This formulation is an equivalent but a simpler version of (3). Once we find  $\mathbf{s}$ , by solving (5), we can

identify the outliers as those data that correspond to the non-zero entries of  $\mathbf{s}$ . We can then remove these outliers and find the least-squares (LS) estimate of  $\mathbf{w}$  using the remaining data. Note that the LS estimate is statistically optimal for Gaussian noise.

A naive way to solve (5) would be to do a combinatorial search. However, recently there has been a lot of work on sparse representation/learning [6, 7, 8, 9] which essentially tries to solve the above problem. Two of the major approaches for solving the sparse learning problem are: 1) the convex relaxation approach [6, 7] and 2) the Bayesian approach [8, 9]. The convex relaxation approach is also related to the emerging field of Compressive Sensing (CS) [10]. We use these two approaches to develop two robust regression algorithms.

### 2.1. Basis Pursuit Robust Regression

Instead of solving (5), we solve the following problem:

$$\min_{\mathbf{s}} \|\mathbf{s}\|_1 \text{ such that } \|\mathbf{z} - \mathbf{C}\mathbf{s}\|_2 \leq \epsilon \quad (6)$$

The above problem is a convex relaxation of the original problem (5), in which the  $L_0$  norm has been replaced by the  $L_1$  norm [7]. (6) is closely related to Basis Pursuit Denoising [7, 6] and we will refer to the robust regression algorithm that uses this algorithm as the Basis Pursuit Robust Regression (BPRR). It has been shown in [10, 7] that if  $\mathbf{s}$  was sparse to begin with then under certain condition ('Restricted Isometry Property' or 'incoherence') on the matrix  $\mathbf{C}$ , (5) and (6) will have the same solution up to a bounded uncertainty due to  $\epsilon$ . However, in our case,  $\mathbf{C}$  depends on the training dataset and may not satisfy those conditions.

### 2.2. Bayesian Sparse Robust Regression

The Bayesian approach for solving (5) is known as the sparse Bayesian learning approach [8, 9]. In this approach, a sparsity enforcing prior is imposed on  $\mathbf{s}$ . Each element of  $\mathbf{s}$  is assumed to be a zero-mean Gaussian random variable

$$p(\mathbf{s}|\alpha) = \prod_{i=1}^N \mathcal{N}(s_i|0, \alpha_i^{-1})$$

where  $\alpha$  is a vector of hyper-parameters. An individual hyper-parameter is associated with each element of  $\mathbf{s}$ . The likelihood term is given by

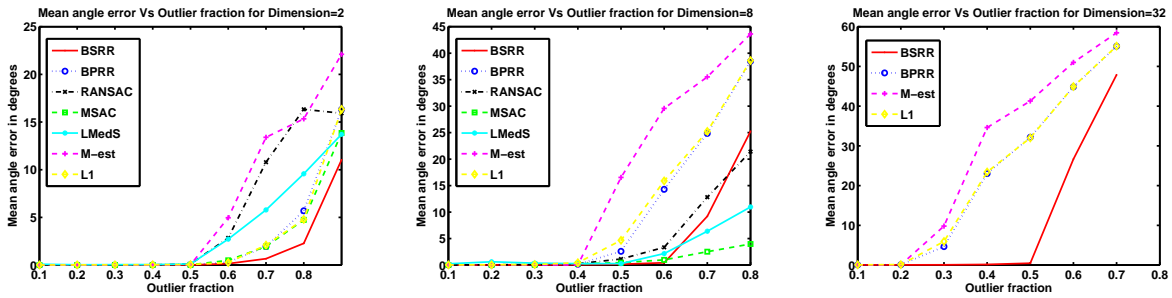
$$p(\mathbf{z}|\mathbf{s}, \sigma^2) = \mathcal{N}(\mathbf{z} - \mathbf{C}\mathbf{s}, \sigma^2)$$

where  $\sigma^2$  is the variance of the Gaussian noise  $\mathbf{C}\mathbf{n}$ . During inference, the hyper-parameters  $\alpha_i$  and  $\sigma$  are first estimated using the evidence maximization framework[8] which are then used for finding the MAP estimate of  $\mathbf{s}$ . We will, henceforth, refer to the robust regression algorithm that uses this algorithm as Bayesian Sparse Robust Regression (BSRR).

The regression model in (1) is linear in both the unknown parameter  $\mathbf{w}$  and the independent variable  $\mathbf{x}$ . However, all the above analysis also applies to models linear only in  $\mathbf{w}$ , that is, models of the form

$$\begin{aligned} y &= \sum_{j=0}^{M-1} w^j \phi^j(\mathbf{x}) + e \\ &= \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}) + e \end{aligned} \quad (7)$$

where  $\phi^j(\mathbf{x})$  could be nonlinear functions of  $\mathbf{x}$ .



**Fig. 1.** Mean angle error in degrees vs. outlier fraction for dimension 2, 8 and 32 respectively. Only BSRR, BPRR, M-estimator and  $L_1$  regression are shown for dimension 32 as the other algorithms are too slow to be practical. BSRR performs very well for all the dimensions, more so for dimension 32 where the performance of the other feasible algorithms such as BPRR, M-estimators and  $L_1$  regression degrades considerably.

### 3. THEORETICAL AND EMPIRICAL STUDIES OF THE INTRINSIC PARAMETER SPACE OF ROBUST REGRESSION

The four important intrinsic parameters of the robust regression problem are the outlier fraction  $f$ , the dimensionality of the problem  $D$ , the number of data points  $N$  and the inlier noise variance  $\sigma^2$ . Here, we study the performance of the proposed algorithms, BPRR and BSRR, and compare it to that of M-estimators, LMedS, RANSAC, MSAC and  $L_1$  regression (least absolute errors regression). The performance criteria are regression accuracy and computational complexity. We first discuss the theoretical computational complexity of the algorithms and then empirically study them for regression accuracy.

BPRR and BSRR have a computational complexity of  $O(N^3 + D^3)$ . The IRLS algorithm, used for solving the M-estimators, has a complexity of  $O(D^3/3 + D^2N)$  and the complexity of  $L_1$  regression is  $O(N^3)$ . None of them have any direct dependence on  $f$  or  $\sigma^2$ . The number of selections  $k$  that the random sampling algorithm in LMedS, RANSAC and MSAC have to make to successfully find a good solution with probability  $p$  is given by [3]

$$k = \min \left( \frac{\log(1-p)}{\log(1-(1-f)^D)}, \binom{N}{D} \right) \quad (8)$$

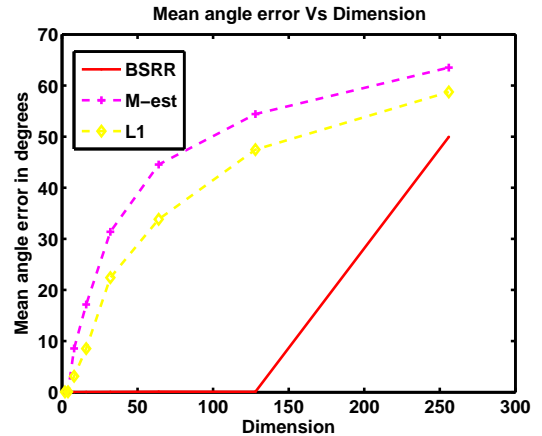
So, these algorithms are combinatorial in  $D$ . We can easily conclude from this discussion that BSRR, BPRR, M-estimators and  $L_1$  regression are the feasible algorithms for high-dimensional robust regression problems whereas LMedS, RANSAC and MSAC are not.

Next, we perform a series of experiments using synthetically generated data. The inlier data is obtained by sampling a Gaussian distribution  $\eta(\mathbf{0}, \sigma^2 \mathbf{I})$  around a randomly generated  $(D-1)$ -dimensional hyperplane in a  $R^D$  space. The outlier data is obtained by uniformly sampling a bounded space containing the hyperplane. Regression accuracy is measured by the angle error between the estimated normal to the hyperplane and the ground truth normal. BSRR, BPRR, RANSAC and MSAC need estimates of the inlier noise standard deviation which we provide as the median absolute residual of the least squares estimate. We have used the MATLAB implementation of bisquare (Tukey's biweight) M-estimator. Other M-estimators give similar results.

In the first experiment, we study the performance of the algorithms as a function of outlier fraction and dimension. The number of data points was fixed at 1000 and noise variance  $\sigma^2$  at 4. Fig. 1 shows the variation of mean angle error with outlier fraction for dimension 2, 8 and 32. For dimension 32, we only show BSRR,

BPRR, M-estimator and  $L_1$  regression as the other algorithms are too slow to be practical. BSRR performs very well for all the dimensions, more so for dimension 32 where the performance of the other feasible algorithms such as BPRR, M-estimators and  $L_1$  regression degrades considerably. At low-dimensions, MSAC also gives good performance.

Next, we vary the dimension while keeping the number of data points fixed at 5000, the outlier fraction at 0.4 and noise variance  $\sigma^2$  at 4. Fig. 2 shows that BSRR performs very well up to dimension 128. BSRR performs well at even higher dimensions if the number of data points is increased proportionally with the dimension. The time taken by BSRR was about 6 minutes for all the dimensions. We do not show LMedS, RANSAC and MSAC results as they are impractical for dimension 16 onwards. BPRR is also slow and hence was not studied here.



**Fig. 2.** Mean angle error vs. dimension for outlier fraction 0.4. BSRR performs very well upto dimension 128. It can perform well at even higher dimensions if the number of data points is increased proportionally with the dimension.

We also study the effect of inlier noise variance on the performance of the algorithms. For this we fixed the dimension at 8, the outlier fraction at 0.4 and the number of data points at 1000. Fig. 3 shows that BSRR performs robustly for a wide range of inlier noise variances. From the above experiments, we can easily conclude that BSRR should be the preferred algorithm for high-dimensional robust regression problems. At low-dimensions both BSRR and MSAC show good performance.

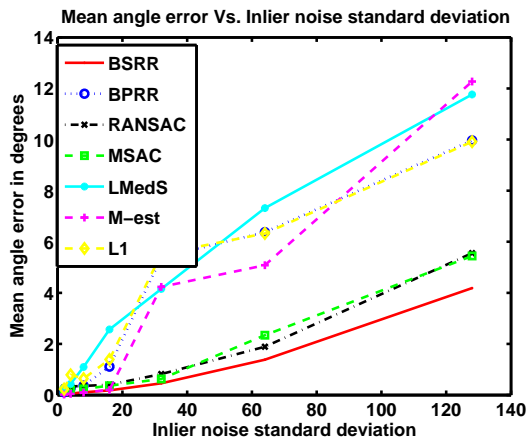


Fig. 3. Mean angle error vs. inlier noise standard deviation. BSRR performs well for a wide range of inlier noise variance.

#### 4. ROBUST REGRESSION FOR AGE ESTIMATION

In this section, we use the BSRR algorithm for robust age estimation from face images. We use the publicly available FG-Net dataset<sup>1</sup>, which contains 1002 facial images of 82 subjects along with their ages. The dependent variable for this problem is the age and the independent variable is a geometric feature obtained by computing the flow field at 68 fiducial features on each image with respect to a reference face image.

We use the BSRR algorithm to categorize the whole dataset into inliers and outliers. The algorithm found 177 outliers out of the total database of 1002 images. Some of the inliers and outliers are shown in figure 4. Most of the outliers were images of older subjects. The reason for this could be due to the fact that the geometric features do not vary much in older subjects and that there are less number of samples of older subjects in the FG-Net database. Next, we perform a leave-one-out testing in which the regression algorithm is trained on the entire dataset except for one sample on which testing is done. We measure the mean absolute error (MAE) of age estimation for inliers and outliers separately. The results are shown in Table 1. The low inlier MAE and the high outlier MAE indicates that the inlier vs outlier categorization was good.

	Inlier MAE	Outlier MAE	All MAE
BSRR	3.73	19.14	6.45

Table 1. Mean absolute error (MAE) of age estimation for inliers and outliers using BSRR. The low inlier MAE and the high outlier MAE indicates that the inlier vs outlier categorization was good.

#### 5. CONCLUSIONS

We presented a systematic approach towards outlier identification by noting that the projection of the dependent variable onto the null space of the independent variables has significant contributions only from the outliers. We then proposed two algorithms, BSRR and BPRR, to solve the resulting problem. These are polynomial algorithms and hence can be used for solving high-dimensional robust regression problems. We also performed an empirical study on the intrinsic parameter space of robust regression which highlighted the excellent performance of BSRR under varying operating conditions.

<sup>1</sup>The fg-net aging database, <http://www.fgnet.rsunit.com>

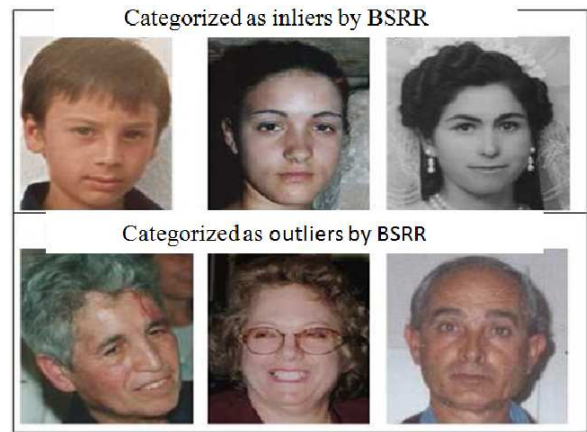


Fig. 4. Some outlier and inliers found by BSRR. Most of the outliers were images of older subjects. The reason for this could be due to the fact that the geometric features do not vary much in older subjects and that there are less number of samples of older subjects in the FG-Net database.

#### 6. REFERENCES

- [1] P. J. Rousseeuw and A. M. Leroy, "Robust regression and outlier detection," *Wiley Series in Prob. and Math. Stat.*, 1986.
- [2] P. J. Huber and E. M. Ronchetti, "Robust statistics," *Wiley Series in Probability and Statistics*, 2009.
- [3] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Comm. Assoc. Mach.*, 1981.
- [4] P. H. S. Torr and A. Zisserman, "Mlesac: A new robust estimator with application to estimating image geometry," *Computer Vision and Image Understanding*, 2000.
- [5] C. V. Stewart, "Robust parameter estimation in computer vision," *SIAM Reviews*, 1999.
- [6] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, 1998.
- [7] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Info. Theory*, vol. 52, 2006.
- [8] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, 2001.
- [9] D. P. Wipf and B. D. Rao, "Sparse bayesian learning for basis selection," *IEEE Trans. Signal Process*, vol. 52, 2004.
- [10] E. J. Candes and M. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, 2008.
- [11] E. J. Candès and P. A. Randall, "Highly robust error correction byconvex programming," *IEEE Transactions on Information Theory*, 2008.