ROBUST LINEAR REGRESSION ANALYSIS - THE GREEDY WAY

George Papageorgiou, Pantelis Bouboulis, Sergios Theodoridis, Konstantinos Themelis

Department of Informatics
and Telecommunications
University of Athens
Athens, Greece, 157 84
Emails: geopapag, stheodor@di.uoa.gr,
panbouboulis@gmail.com, themelis@noa.gr

ABSTRACT

In this paper, the task of robust estimation in the presence of outliers is presented. Outliers are explicitly modeled by employing sparsity arguments. A novel efficient algorithm, based on the greedy Orthogonal Matching Pursuit (OMP) scheme, is derived. Theoretical results concerning the recovery of the solution as well as simulation experiments, which verify the comparative advantages of the new technique, are discussed.

Index Terms— Greedy Algorithm for Robust Denoising (GARD), Robust based Regression, Robust Least Squares, Greedy algorithms, Outlier detection

1. INTRODUCTION

Sparsity-aware learning and related optimization techniques have been at the forefront of the research happening in signal processing, encompassing a wide range of topics, such as compressed sensing, denoising and signal approximation techniques [1–3].

The idea of sparsity (also related to sufficiency or economy of a representation), recently, seems to fit in more applications than initially expected, despite the fact that similar techniques (such as the minimization of cost functions involving ℓ_1 norms) have been used since the 1970s. There are two basic paths for obtaining a solution to the respective problem. The first one, comprises the family of greedy algorithms which provide the solution of tasks that minimize the ℓ_0 (pseudo) norm, under certain assumptions. Even though, in general, this is an NP-hard problem, methods belonging to greedy class of schemes can efficiently recover the solution in polynomial time. The second path is to employ the ℓ_1 norm minimization, the closest convex relaxation of the ℓ_0 norm.

This research has been co-financed by the European Union (European Social Fund - ESF) and Greek National funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Aristeia I - 621.

A recent application of sparsity-aware optimization methods is that of signal denoising, where one is interested to recover the original signal, which (in this case) has been corrupted not only by standard (inlier) noise (e.g., Gaussian noise, etc.), but also by outliers. The main idea, which is followed in this work, is to model the noise via two separate components; one corresponding to the inlier and one to the outlier. Since the outlier part is not present to all samples, sparse modeling arguments are mobilized to model the outlier component. This concept has been previously employed in [4, 5]. The novelty of our approach lies in the different modeling compared to the previous works, by treating the task in terms of the ℓ_0 minimization via greedy algorithmic concepts. A novel scheme has been derived which exhibits notable performance gains, compared to previously published works, both computationally as well as in the quality of the recovered results.

Notations: We use capital bold to denote a matrix, $\mathbf{X}$, bold lower case for a vector, $\boldsymbol{\theta}$, while u_i denotes the $i-$th element of vector $\boldsymbol{u}$ and $\boldsymbol{x}_i$ the $i-$th column of matrix $\mathbf{X}$. Finally, the unitary matrix of dimension n, will be denoted $\mathbf{I}_n$ and the vector of zero elements as $\mathbf{0}$.

2. PROBLEM MODELLING

In a linear regression task, we are interested in estimating the relation between two sets of variables, $(y_i, \boldsymbol{x}_i),\ i = 1, ..., n$, where $y_i \in \mathbb{R}$ are the output values produced by the independent variable (input) $\boldsymbol{x}_i \in \mathbb{R}^m$, through the model:

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\theta} + e_i,\ i = 1, ..., n, \tag{1}$$

where e_i is the observation noise.

Our goal is to estimate $\boldsymbol{\theta} \in \mathbb{R}^m$ from the given training dataset of n measurements. Collecting over n, we can rewrite (1):

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{e}, \tag{2}$$

where $\boldsymbol{y} = (y_1, y_2, ..., y_n)^T$, $\boldsymbol{e} = (e_1, e_2, ..., e_n)^T$ and matrix $\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n]^T \in \mathbb{R}^{n \times m}$.

As it is common in regression analysis, we consider that the number of observations exceeds the number of unknowns, i.e., $n > m$. In order to seek for a solution, we should assume that $\mathbf{X}$ is a full rank matrix, i.e. $rank(\mathbf{X}) = m$. If the noise is i.i.d Gaussian, the most common estimator, which is statistically optimal is the Least Squares (LS) estimator. However, this estimator fails in the presence of outliers or when the noise distribution has long tails.

The key aspect exploited in this work, that provides an efficient approach to overcome the aforementioned problem, is *sparsity*. In this context, we assume that the outlier noise values are significantly fewer (i.e., sparse) than the size of the input data. To this end, we express the noise vector as a sum of two independent components, $\boldsymbol{e} = \boldsymbol{u}+\boldsymbol{\eta}$, where $\boldsymbol{\eta}$ is assumed to be a dense vector of bounded inlier noise of energy ϵ_0 and $\boldsymbol{u} \in \mathbb{R}^n$ the sparse outlier noise vector over the support set, T, with cardinality $|T| \leq s$. The support set is defined as the index set for which $\boldsymbol{u}$ is nonzero. Hence, equation (2) can be recast as:

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{u} + \boldsymbol{\eta}. \tag{3}$$

There are two basic paths towards an estimation of the solution of the problem (3). The first one, which reflects on the combinatorial nature of the problem is to employ the ℓ_0 (pseudo)-norm on vector $\boldsymbol{u}$. Although this task is impractical, greedy selection algorithms seem to work under specific assumptions and leads to good estimates of the solution, as we will later demonstrate. In this case, the associated optimization problem becomes:

$$\min_{\boldsymbol{\theta},\boldsymbol{u}} ||\boldsymbol{u}||_0, \text{ s.t. } ||\boldsymbol{y} - \mathbf{X}\boldsymbol{\theta} - \boldsymbol{u}||_2 \leq \epsilon_0. \tag{4}$$

The second classic approach, is to relax the ℓ_0 with the convex ℓ_1 norm, using a similar formulation:

$$\min_{\boldsymbol{\theta},\boldsymbol{u}} ||\boldsymbol{u}||_1, \text{ s.t. } ||\boldsymbol{y} - \mathbf{X}\boldsymbol{\theta} - \boldsymbol{u}||_2 \leq \epsilon_0. \tag{5}$$

Problem (5) is also known as *Robust Regression Basis Pursuit* (BPRR) and it can be solved using Second Order Cone Programming (SOCP) techniques [6, 7]. An equivalent approach is to define the Lagrangian form of the problem for appropriate multiplier values $\lambda > 0$:

$$\min_{\boldsymbol{\theta},\boldsymbol{u}} \frac{1}{2}||\boldsymbol{y} - \mathbf{X}\boldsymbol{\theta} - \boldsymbol{u}||_2^2 + \lambda||\boldsymbol{u}||_1, \tag{6}$$

and solving using a popular method by the name of *Alternating Direction Method of Multipliers*-(ADMM)[1].

3. GREEDY ALGORITHM FOR ROBUST DENOISING - (GARD)

3.1. Algorithm

Our algorithm attempts to solve problem (4) and it will be developed along the classic Orthogonal Matching Pursuit version [8–10]. Because of space limitations, the classic (OMP) will not be discussed in details here.

We reformulate (4) into the equivalent form:

$$\min_{\boldsymbol{\theta},\boldsymbol{u}} ||\boldsymbol{u}||_0, \text{ s.t. } \left\| \boldsymbol{y} - \mathbf{A}\begin{pmatrix}\boldsymbol{\theta}\\ \boldsymbol{u}\end{pmatrix} \right\|_2 \leq \epsilon_0, \tag{7}$$

where $\mathbf{A} = [\mathbf{X}\ \mathbf{I}_n]$. The basic idea is the restriction of the selection over atoms of the second half of matrix $\mathbf{A}$, i.e., matrix $\mathbf{I}_n = [\boldsymbol{e}_1\ \boldsymbol{e}_2\ ...\ \boldsymbol{e}_n]$. The notably improved performance of the proposed scheme is due to the orthogonality between the columns of $\mathbf{I}_n$ (standard Euclidean basis). Following the basic steps of the (OMP), the following algorithm is derived:

Input: $\mathbf{X}$, $\boldsymbol{y}$, ϵ_0
Output: $\hat{\boldsymbol{z}} = (\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{u}})^T$
Initialization: $k := 0$, $S_{ac} = \{1, 2, ..., m\}$, $S_{inac} = \{m+1, ..., m+n\}$, $\mathbf{A}_{ac}^{(0)} = \mathbf{X}$
Solve: $\hat{\boldsymbol{z}}^{(0)} := \arg\min_{\boldsymbol{z}} ||\boldsymbol{y} - \mathbf{A}_{ac}^{(0)}\boldsymbol{z}||_2^2$
Initial Residual: $\boldsymbol{r}^{(0)} = \boldsymbol{y} - \mathbf{A}_{ac}^{(0)}\hat{\boldsymbol{z}}^{(0)}$
while $||\boldsymbol{r}^{(k)}||_2 > \epsilon_0$ **do**
  $k := k + 1$
  Find: $j_k := \arg\max_{j \in S_{inac}} |r_{j-m}^{(k-1)}|$
  Update Support:
  $S_{ac} := S_{ac} \cup \{j_k\}$, $\mathbf{A}_{ac}^{(k)} = [\mathbf{A}_{ac}^{(k-1)}\ \boldsymbol{e}_{j_k}]$
  Update Current solution:
  $\hat{\boldsymbol{z}}^{(k)} := \arg\min_{\boldsymbol{z}} ||\boldsymbol{y} - \mathbf{A}_{ac}^{(k)}\boldsymbol{z}||_2^2$
  Update Residual: $\boldsymbol{r}^{(k)} = \boldsymbol{y} - \mathbf{A}_{ac}^{(k)}\hat{\boldsymbol{z}}^{(k)}$
end

Algorithm 1: (GARD) - Greedy Algorithm for Robust Denoising

Remarks:

- The algorithm starts with a Least Squares solution to obtain $\hat{\boldsymbol{z}}^{(0)}$. Thus if no outliers exist, (GARD) solves a Least Squares problem.
- The residual is orthogonal to each column that participates in the representation, i.e., $\langle \mathbf{r}^{(k)}, \boldsymbol{e}_{j_k} \rangle = r_{j_k}^{(k)} = 0,\ \forall\ k = 1, 2, \ldots$. Thus, column $\boldsymbol{e}_{j_k}$ of matrix $\mathbf{I}_n$, is not selected once again.
- Although the complexity of the algorithm is $O(m+k)^3$ at each step k, $(k << n)$, it could be further reduced using *Cholesky* decomposition. At the initial step $(k := 0)$, let $\mathbf{L}^{(0)}$ denote the lower triangular matrix of the Cholesky decomposition of $\mathbf{X}^T\mathbf{X}$. The total cost after solving the normal equations using forward and backward substitution is $O(m^3/3 + nm^2)$. At each following step, we compute $\boldsymbol{v}$, such that $\mathbf{L}^{(k-1)}\boldsymbol{v} = \mathbf{A}_{ac}^{(k-1)^T}\boldsymbol{e}_{j_k}$ and $b = \sqrt{1 - ||\boldsymbol{v}||_2^2}$, so that

$$\mathbf{L}^{(k)} = \begin{pmatrix}\mathbf{L}^{(k-1)} & \mathbf{0}\\ \boldsymbol{v}^T & b\end{pmatrix}.$$

[1] The two forms ((5) and (6)) are equivalent for positive values of λ depending on values of the inlier error bound ϵ_0.

The interested reader is referred to [11]. Hence, by updating the Cholesky matrix $\mathbf{L}$ at each step, we reduce the computational cost down to $O\big((3/2)(m+k)^2+3n(m+k)\big)$, $k << n$, since we only need to solve the normal equations for the augmented matrix $\mathbf{A}_{ac}^{(k)}$ using forward and backward substitution with matrix $\mathbf{L}^{(k)}$.

3.2. Properties

We derive the basic properties of (GARD). Due to space limitations the proofs will not be presented here.

Lemma 1 *At every $k \leq n-m$ step, (GARD) selects a column vector $\boldsymbol{e}_{j_k}$ from matrix $\mathbf{I}_n$, that is linearly independent of all the column vectors in matrix $\mathbf{A}_{ac}^{(k-1)}$. Hence, $\mathbf{A}_{ac}^{(k)}$ is full rank and the solution to the Least Squares problem at each step is unique.*

The proof relies on mathematical induction.

Theorem 1 *The norm of the residual vector $\boldsymbol{r}^{(k)} = \boldsymbol{y} - \mathbf{A}_{ac}^{(k)}\hat{\boldsymbol{z}}^{(k)}$ in algorithm 1 is strictly decreasing. Moreover, the algorithm will always converge.*

Theorem 2 *Assume that $\boldsymbol{y} \in \mathbb{R}^n$ admits a representation of the form $\boldsymbol{y} = \mathbf{X}\boldsymbol{\theta}_* + \boldsymbol{u}_* + \boldsymbol{\eta}$, where $T = supp(\boldsymbol{u}_*)$, $|T| \leq s << n$ and $||\boldsymbol{\eta}||_2 \leq \epsilon_0$. Then, (GARD) guarantees that the support of the sparse vector is recovered, if there exists $\delta \in (0, \min(|u_i|)/2)$, $i \in T$, with $||\mathbf{X}(\boldsymbol{\theta}_* - \hat{\boldsymbol{\theta}}^{(k)})||_\infty \leq \delta, \forall k$, such that*

$$2(\epsilon_0 + \delta) < \min(|u_i|),\ i \in T, \tag{8}$$

where $\hat{\boldsymbol{\theta}}^{(k)}$ is the Least Squares estimate of $\boldsymbol{\theta}_*$, at each step.

The proof borrows the main concepts behind Lemmas 5.2 and 5.3 in [12], to ensure that at every step the correct support is established. It is obvious that (8), could be violated, hence the stability result holds only locally. Another point of view of Lemma 2 is the separation of the inlier and outlier noise. Only the noise values outside the zone $[-\epsilon_0 - \delta, \epsilon_0 + \delta]$ are considered to be outliers, which could also be used as the definition of outlier noise.

4. COMPARISON OF METHODS

Methods tested within this work and compared against (GARD) are listed bellow:

- M-estimates: In M-est a robust cost function ρ (satisfying certain properties) of the residual error $r_i = y_i - \boldsymbol{x}_i^T\boldsymbol{\theta}$, $i = 1, 2, ..., n$ is minimized, scaled by a statistical factor σ, so that $\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \rho(r_i/\sigma)$ [13], [14]. Another way to interpret M-est is by solving a Weighted Least Squares problem iteratively,

$$\min_{\boldsymbol{\theta}} ||\mathbf{W_r}^{1/2}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\theta})||_2^2, \tag{9}$$

  using (IRLS) algorithm, where the diagonal weight matrix $\mathbf{W_r}$ assigns the weights, with values depending on the Robust function (for classic Least Squares each weight on the diagonal equals 1). In our experiments we used Tukey's biweight (or bisquare) robust (but nonconvex) function, which is the routine *robustfit* in MATLAB. For σ, we have used the default parameter settings for the mean absolute deviation (MAD) to compute it [14, 15].
- (SOCP): Reformulate problem (5) as (SOCP) problem [16], [6]. The function we used in order solve the (SOCP) problem is *SeDuMi* of the optimization package *CVX* for MATLAB of Stanford University, CVX RESEARCH: http://cvxr.com/ (31/01/2014).
- (ADMM): Solve problem (6) using the Alternating Direction method of multipliers described and used in many papers [17, 18].
- (SBL): Solve problem using Sparse Bayesian Learning algorithm described in [4, 5].
- (ROMP): Use OMP algorithm, but solve a robust Least Squares step (M-est) at each iteration instead via

$$\min_{\boldsymbol{\theta}} ||\mathbf{W_r}^{1/2}(\boldsymbol{y} - \mathbf{A}_{\Omega^t}\boldsymbol{\theta})||_2^2, \tag{10}$$

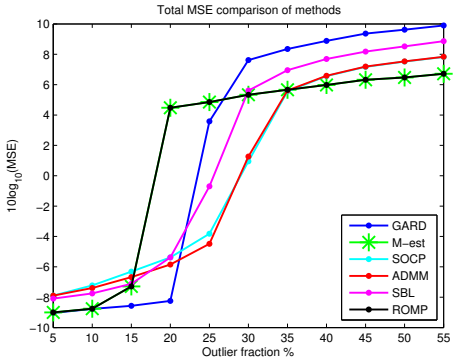
  where Ω^t is the set of active columns of matrix $\mathbf{A}$. Once again, we have used Tukey's biweight function (*robustfit*) with the default parameter settings as in M-est [19]. Complexity of the method depends on various parameters and is not given in closed form.

Algorithm	Solves	Complexity
(GARD)	(4)	$O\big(m^3/3 + k^3/2 + (n+3k)m^2 + 3kmn\big)$
(M-est)	(9)	$O\big(m^3/3 + nm^2\big)$/step
(SOCP)	(5)	$O\big((n+m)^{2.5}n\big)$
(ADMM)	(6)	$O\big((n+m)^3/3 + n(n+m)^2\big)$/step
(SBL)	(6)	$O\big(m^3/3 + nm^2\big)$/step

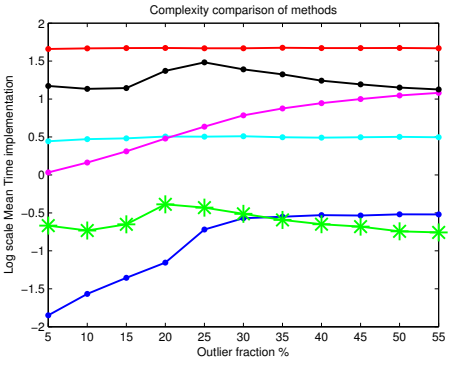
Table 1. For (GARD) and (SOCP) total complexity is given. For the rest total complexity depends on the number of iterations for convergence.

5. EXPERIMENTAL RESULTS

In the experiments below, we have tested how each algorithm performs and we highlighted their pros and cons. The experimental set up parallels [16]. In all our experiments, we have used equation (3) to generate our data $(y_i, \boldsymbol{x}_i)$, $i = 1, 2, ..., n$, $\boldsymbol{x}_i \in \mathbb{R}^m$, with $n = 600$ measurements. The columns of matrix $\mathbf{X}$, are obtained by uniformly sampling an m-dimensional hypercube centered around the origin and $\boldsymbol{\theta} \in \mathbb{R}^m$ are random vectors chosen from the multivariate normal distribution



(a) Log scale of MSE



(b) Time graph

Fig. 1. Mean Square Error and complexity for Robust Regression methods.

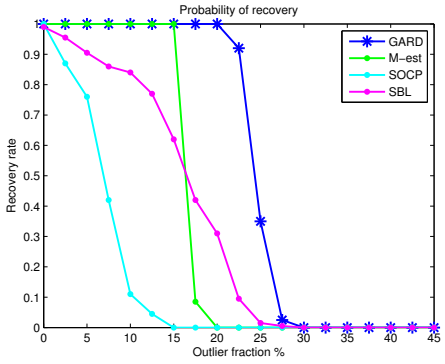


Fig. 2. Recovery rate for dimension of the unknown vector $m = 200$.

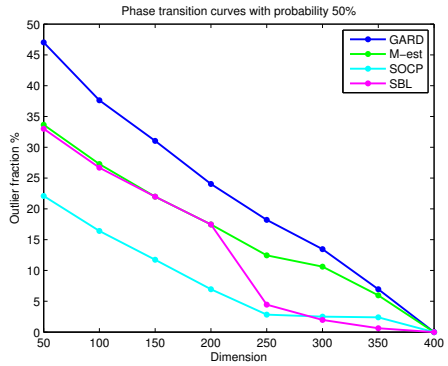


Fig. 3. Phase transition curves. Under each curve, the probability of recovery is greater than 50%.

with mean value 0 and standard deviation 0.03. The inlier noise is drawn from the normal distribution $\mathcal{N}(0, 0.1^2)$, while the outlier noise signal is sparse impulsive. Outlier values, equal ± 5 for the first experiment and ± 100 for the final two, in s indices, uniformly sampled over n coordinates ($s < n$).

The input parameter for (GARD), (SOCP) and (ROMP) is the inlier noise bound ϵ_0 (which could be assumed to be known or otherwise estimated using the cross validation method). The parameters required for (ADMM) and (SBL), are the regularization parameter λ and the pruning threshold plus the initial vector of hyperparameters, respectively. For all methods parameters have been selected, so that to optimize performance. Concerning (SBL) a major drawback of the latter method is that it is very sensitive to the initialization (recall that this is a non-convex method, which cannot guarantee that the global minimum is attained for each dimension m), while the time needed for each implementation cannot be assured, since the number of iterations until convergence strongly depends on those parameters. For (SBL) random initialization was performed a number of times and the best solution was selected. Finally, the (M-est) does not require any parameters to be predefined.

a) In the first experiment, the dimension of the unknown vector $\boldsymbol{\theta}$ is $m = 200$. The percentage of outliers varies from 5% to 55% and the Mean Square Error of the estimated solution is computed over a set of 40 experiments for each outlier density. In Figure 1, it is shown that for the case of outlier fraction under 20%, (GARD) excels; the mean implementation time is the lowest, while the support of the unknown outlier vector is fully recovered. (M-est) seems to perform well too, but its performance breaks down for a fraction of outliers greater than 15%; moreover, it operates at a higher computational cost, too. Another algorithm that is also worth to be noticed is (SBL); it is clear, that the greedy approach provides better performance when the outlier vector is very sparse (compared to the ℓ_1 minimization).

b) In our second experiment (Figure 2), we have tested the most efficient algorithms overall (in terms of performance and complexity). The dimension of the unknown vector, is also $m = 200$. Input noise is considered as in experiment a), with the outlier fraction varying over 0% to 45%. Finally, the probability of recovery [2] is computed over 200

[2] Solution recovered if $||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||_2 / ||\boldsymbol{\theta}||_2 \leq 0.07$.

experiments at each outlier fraction. Note that (GARD) reaches the phase transition (from success to failure) for the highest fraction of outliers.

c) In our last experiment (Figure 3), we have produced the phase transition curves for (GARD), (M-est), (SOCP) and (SBL). For each dimension, we have computed the fraction of outliers of the sparse impulse noise vector that corresponds to probability of recovery at 50% $\big($solution recovered as in experiment b)$\big)$. Once again, probability of recovery is computed over 200 experiments at each outlier fraction. Note that (GARD) reaches the phase transition for the highest fraction of outliers, for the whole range of dimensions of the unknown vector examined.

6. CONCLUSIONS AND FUTURE WORK

In this work, a new approach to the Linear Robust Regression problem has been proposed. Although other previous methods have a balanced performance, all tests against standard methods, reveal that due to the nature of the problem (sparse outlier vector), (GARD) manages to perform distinctly better than the other methods, while it is also computationally more efficient. Other issues to be discussed in the near future, include an a priori estimation of the solution error, as well as the development of an automatic termination criterion of the algorithm, depending on the statistics of the noise.

REFERENCES

[1] Scott Shaobing Chen, David L Donoho, and Michael A Saunders, "Atomic decomposition by basis pursuit," *SIAM journal on scientific computing*, vol. 20, no. 1, pp. 33–61, 1998.

[2] Emmanuel J Candes, Justin K Romberg, and Terence Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on pure and applied mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[3] Robert Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[4] Yuzhe Jin and Bhaskar D Rao, "Algorithms for robust linear regression by exploiting the connection to sparse signal recovery," in *Acoustics Speech and Signal Processing (ICASSP), International Conference on*. IEEE, 2010, pp. 3830–3833.

[5] Kaushik Mitra, Ashok Veeraraghavan, and Rama Chellappa, "Robust rvm regression using sparse outlier model," in *Computer Vision and Pattern Recognition (CVPR), Conference on*. IEEE, 2010, pp. 1887–1894.

[6] Miguel Sousa Lobo, Lieven Vandenberghe, Stephen Boyd, and Hervé Lebret, "Applications of second-order cone programming," *Linear algebra and its applications*, vol. 284, 1998.

[7] Stephen P Boyd and Lieven Vandenberghe, *Convex optimization*, Cambridge university press, 2004.

[8] Alfred M. Bruckstein, David L. Donoho, and Michael Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM review*, vol. 51, no. 1, pp. 34–81, 2009.

[9] Yagyensh Chandra Pati, Ramin Rezaiifar, and PS Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Signals, Systems and Computers, Conference Record of The Twenty-Seventh Asilomar Conference on*. IEEE, 1993, pp. 40–44.

[10] Stephane Mallat, *A wavelet tour of signal processing: the sparse way*, Academic press, 2008.

[11] Bob L Sturm and Mads Graesboll Christensen, "Comparison of orthogonal matching pursuit implementations," in *Signal Processing Conference (EUSIPCO), Proceedings of the 20th European*. IEEE, 2012, pp. 220–224.

[12] David Leigh Donoho, Michael Elad, and Vladimir N Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *Information Theory, IEEE Transactions on*, vol. 52, no. 1, pp. 6–18, 2006.

[13] Peter J Huber, *Wiley Series in Probability and Mathematics Statistics*, Wiley Online Library, 1981.

[14] Ricardo A Maronna, R Douglas Martin, and Victor J Yohai, *Robust statistics*, J. Wiley, 2006.

[15] Peter J Rousseeuw and Annick M Leroy, *Robust regression and outlier detection*, vol. 589, John Wiley & Sons, 2005.

[16] Kaushik Mitra, Ashok Veeraraghavan, and Rama Chellappa, "Analysis of sparse regularization based robust regression approaches," *Signal Processing, IEEE Transactions on*, vol. 61, no. 5, pp. 1249–1257, 2013.

[17] Stephen Boyd, "Alternating direction method of multipliers," in *Talk at NIPS Workshop on Optimization and Machine Learning*, 2011.

[18] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[19] S Alireza Razavi, Esa Ollila, and Visa Koivunen, "Robust greedy algorithms for compressed sensing," in *Signal Processing Conference (EUSIPCO), Proceedings of the 20th European*. IEEE, 2012, pp. 969–973.