Robust Algorithms for Linear and Nonlinear Regression via Spare Modeling Methods: Theory, Algorithms and Applications to Image Denoising

George K. Papageorgiou*

National and Kapodistrian University of Athens
Department of Informatics and Telecommunications
geo_papag@hotmail.com

Abstract. In this dissertation, the problem of robust regression is studied, for both the linear and the nonlinear case. For the former case, a novel algorithm, Greedy Algorithm for Robust Denoising (GARD), which is based on sparse optimization techniques, is derived. Moreover, theoretical conditions, which guarantee the identification of the outliers and a bound on the estimation error, are provided. Next, we focus on the nonlinear case, where it is assumed that the unknown nonlinear function belongs to a Reproducing Kernel Hilbert Space (RKHS). A robust scheme, Kernel Greedy Algorithm for Robust Denoising (KGARD), which shares the same concept with GARD, is proposed. The algorithm is compared against other cutting edge methods via extensive simulations, where its enhanced performance is demonstrated. In addition, theoretical results regarding the identification of the outliers are provided. Finally, the proposed robust estimation framework is applied to the task of image denoising, where the advantages of the proposed method are unveiled. The experiments verify that KGARD improves the denoising process significantly, when outliers are present.

Keywords: robust linear regression, robust nonlinear regression in RKHS, greedy algorithm for robust denoising, kernel greedy algorithm for robust denoising, image denoising, outliers

1 Introduction

At the heart of Machine Learning is the task of *regression* or *regression analysis*. In a classic regression task, given a set of training data, the goal is to learn a set of unknown parameters in order to make predictions. In simple words, the task could be seen as a curve fitting problem. Consider a set of training points $(y_i, \boldsymbol{x}_i)$, $y_i \in \mathbb{R}$ and $\boldsymbol{x}_i \in \mathbb{R}^M$ for $i = 1, \ldots, N$. The task is to estimate a function, f, whose graph fits the data. The target function, f, of the independent variables, $\boldsymbol{x}$, is called the *regression function* and can be either linear or nonlinear. The difference

* Dissertation Advisor: Sergios Theodoridis, Professor.

between regression and classification is that in regression the dependent variable belongs to an interval in the real axis (or region in the complex plane), while in classification it is a discrete variable.

Regression analysis is widely used for prediction and forecasting. It is also used as a means to extract information concerning the degree of dependence among the dependent (output) and the independent (input) variables. Thus, useful information and related implications of such dependencies can be revealed.

The earliest form of regression was the method of Least Squares (LS), which was published by Legendre in 1805 and by Gauss in 1809. Legendre and Gauss both applied the method to the problem of determining the orbits of comets, based on astronomical observations. Many techniques that perform regression analysis have been developed, since then. Familiar methods such as linear regression and ordinary Least Squares regression belong to the parametric class of learning techniques; that is, the model function is defined in terms of a finite number of unknown parameters that are estimated from the data. In contrast, nonparametric regression refers to techniques that bypass the need for explicit parameterization of the unknown functional dependence. For example, the regression function can be assumed to lie in a specific set of functions, which may also be infinite-dimensional. A popular example, that will be adopted in the current thesis for the estimation of a nonlinear function, is to assume that the regression function lies in a Reproducing Kernel Hilbert Space (RKHS).

The performance of regression methods, in practice, depends on the form of the data-generating mechanism and how this relates to the regression model being used. Since the true form of the data-generating process is generally unknown, regression analysis often depends, to a large extent, on making assumptions concerning this process. Regression models, that are designed for prediction, are often useful even when the assumptions are moderately violated, although they may not perform optimally. However, if our goal is to make accurate predictions, we should look for a model/method that is *robust* enough, i.e., it can tolerate abnormalities on the data so that the estimation is not significantly affected.

The notion of robustness, i.e., the efficiency of a method to solve a learning task from data under noise uncertainties of various types, has been a major issue in the scientific community for over half a century. The goal is to minimize the effect of the observations that have been corrupted by unexpected high values of noise, known as *outliers*. Outliers are often regarded as erroneous measurements that deviate greatly from the rest of the observations. This is due to the fact: either their values are heavily influenced by another source or they are generated by a different mechanism/distribution.

In such cases, classic estimators, e.g., the Least Squares, are known to fail to perform well. This problem was originally addressed since the 1950s and it was actually solved more than a decade later, by Huber. Eventually, it led to the development of a new field in Statistics, known as *Robust Statistics*. However, the need for development of robust estimators was not only limited within the Statistics scientific community. Similar tasks (involving robust estimators) emerged in

the context of many fields such as Physics, Medicine, Biology, Engineering and Computer Science, to name a few.

The robust tools that have been developed over the years for handling outliers can be classified into two major categories. The first one includes tools that rely on the use of *diagnostics*, whereas the second direction is based on *robust regression* methods. Diagnostics and robust regression have the same goals, only obtained in the opposite order; both approaches have a long history in the field of Robust Statistics. Lately, a different approach has emerged. The recent development of methods in the spirit of robust analysis owes a lot to the emergence of *sparse modeling* methods, during the past decade.

Sparsity-aware learning and related optimization techniques have been at the forefront of the research in signal processing, encompassing a wide range of topics, such as compressed sensing, signal denoising and approximation techniques. Sparsity is closely related to sufficiency or economy of a representation, a mechanism that harmonizes with nature, which tends to be parsimonious. At the heart of this problem lies an underdetermined set of linear equations, which, in general, accepts an infinite number of solutions. Imposing sparsity, is interpreted as seeking for a solution where only a few of the unknown coordinates, which we attempt to estimate, are nonzero. There are two major paths, towards modeling sparse vectors/signals. The first one focuses on minimizing the ℓ_0(pseudo)-norm of a vector, which equals the number of its nonzero coordinates. However, since this is a non-convex optimization task, approximate methods have been established. The family of algorithms that have been developed to address problems involving the ℓ_0 (pseudo)-norm, comprises *greedy* methods, which have been shown to provide the solution of the related minimization task, under certain reasonable assumptions. Even though, in general, this is an NP-Hard task, it has been shown that such methods can efficiently recover a solution in polynomial time. On the other hand, the family of algorithms developed around the methods that employ the ℓ_1-norm, embraces convex optimization, providing a broader set of tools and stronger guarantees for convergence. Both methods have been shown to generate sparse solutions.

A more recent application of sparse modeling and optimization methods, which is also the focus of this work, is that of signal denoising. There, one is interested in recovering the original signal, which apart from the standard inlier noise, e.g., Gaussian, has also been corrupted by outliers. The key to this modeling is to assume that the outliers comprise only a small fraction of the entire data set, thus the outlier vector is modeled as a sparse one.

The goal of this dissertation, is to address the task of robust linear and nonlinear regression via sparse modeling methods, within the context of machine learning. The proposed methods are built on the popular Orthogonal Matching Pursuit algorithm (OMP), by imposing sparsity constraints on the outliers. In particular, two novel robust algorithms are developed. One for the task of linear regression and a second one for the task of nonlinear regression, where it has been also assumed that the function to be estimated lies in a Reproducing Kernel Hilbert Space (RKHS). Various experiments are performed, where both of the

algorithms are compared against state-of-the-art methods. The obtained results demonstrate their performance and highlight their advantages. Moreover, the study of the algorithms has led to the establishment of sound theoretical results. Finally, the focus is turned on the applications of the nonlinear regression scheme to the task of image denoising. As a result, two methods are introduced for the removal of impulsive noise. The most significant results of this novel robust approach are outlined next.

2 Robust Linear Regression

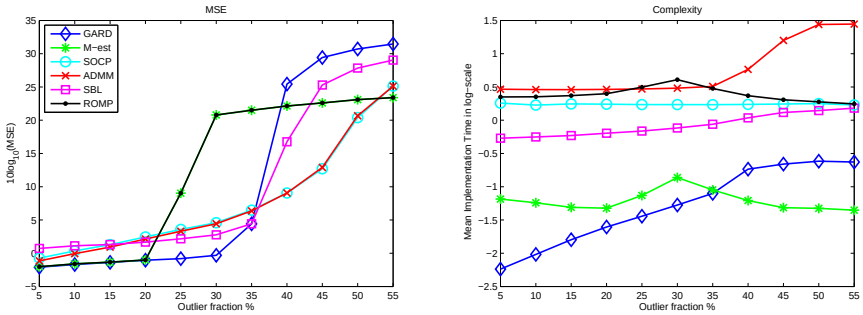
For the linear regression task we have assumed that the output data are corrupted by inlier and outlier noise. Moreover, we have assumed that the outliers are only few compared to the number of the data (thus the outlier vector can be modeled as a sparse one) and that the number, N, of the available data is sufficiently greater than the number, M, of the unknown coefficients. The proposed algorithm is called Greedy Algorithm for Robust Denoising (GARD), and it is based on the classic Orthogonal Matching Pursuit (OMP). The method alternates between a Least Squares (LS) optimization criterion and an OMP-like selection step, that identifies the outliers. The theoretical results that have been established for GARD are:

- The convergence of the scheme in a finite number of steps.
- A bound on the Restricted Isometry Property (RIP) constant, for the case where only outliers are present, which guarantees that GARD successfully identifies the outliers. Moreover, the method recovers both the regression solution and the sparse outlier vector, exactly (with no error), under the existence and uniqueness conditions.
- A second bound on the Restricted Isometry Property (RIP) constant, for the case where the data is corrupted by both inlier and outlier noise, which guarantees that GARD successfully identifies the outliers, assuming that the inlier noise is bounded.
- Performance bounds on the approximation, which guarantee the stability of the algorithm.

It should be noted that, the result concerning the identification of the outliers in the presence of both inlier and outlier noise has been derived for the first time in the robust regression framework.

Next, follows an extended set of experiments that are performed and demonstrate the performance of GARD against other comparative cutting edge methods. For each method, we have computed the Mean-square-error (MSE) and the Mean Implementation Time (MIT), while varying the fraction of the outliers. The most significant results for GARD are:

- It attains the lowest MSE.
- It demonstrates enhanced robustness, compared to all other methods.
- It has very low computational requirements.



(a): $M = 100$ (b): $M = 100$

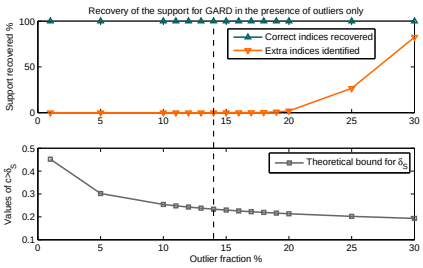
Fig. 1: (a): The attained Mean-square-error (MSE) (in logarithmic scale-dB) versus the fraction of outliers in the output data. (b): Logarithmic scale of the Mean Implementation Time (MIT) versus the outlier fraction. The number of the data is $N = 600$.

In Figure 1 (a), the MSE (in dBs) attained by each method versus the fraction of outliers is depicted, for a fix dimension of the unknown vector at $M = 100$. The Mean Implementation Time (MIT) is also plotted in logarithmic scale in Figure 1 (b). Observe that GARD attains the lowest MSE among its competitors, while in parallel it seems to be the most efficient, operating at the lowest computational cost (the interesting "zone" is for fractions of less than 30%, that is $10\% - 20\%$).

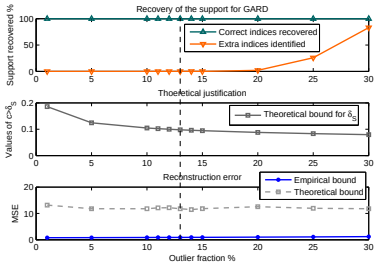
Moreover, in Figure 2 the capability of KGARD to identify the outliers is demonstrated. The green line pointing upwards corresponds to successful outlier identifications, while the orange one pointing downwards corresponds to extra indices that GARD has classified as outliers. In parallel, the relation of the percentage of outliers to the bound of the RIP constant is shown (grey line). Figure 2 (a) corresponds to the noiseless case, while in (b), the data is corrupted by outlier and bounded inlier noise, as the resulting theorem suggests. It is clear, that for small fractions of outliers the support is recovered (one-to-one index), thus we conclude that the condition is valid (the RIP constant cannot be computed).

Figure 3 (a) demonstrates the probability of recovery for each method tested, while varying the fraction of the outliers. In Figure 3 (b), the phase transition curves for each method are given. For each dimension of the unknown vector, we have computed the fraction of outliers for which the method transits from success to failure with probability $p = 0.5$. For example, for $M = 100$ (Figure 3 (a)), the horizontal line at 0.5 corresponds to fractions of outliers (for each method) that are located in the y-axis of Figure 3 (b) for the dimension of $M = 100$. Here, it is clear that up to $M = 200$, GARD succeeds to recover the solution with a higher probability than the rest of the methods.

Finally, in Table 1 we have measured the attained MSE for the case where the noise follows a more general distribution. In columns A, B and C the noise originates from the Lévy alpha-stable distribution, while in column D the noise

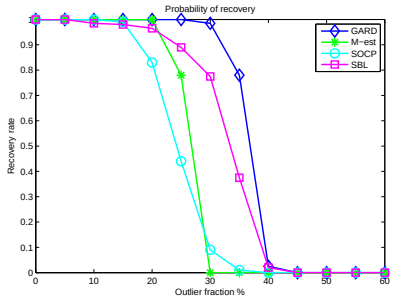


(a): Noise: outliers only.

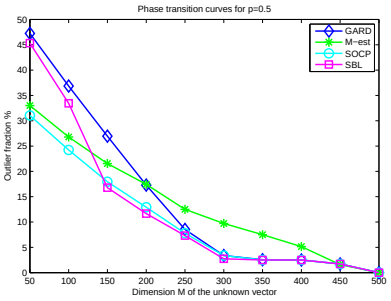


(b): Noise: outliers and bounded inlier.

Fig. 2: The identification of the outliers and the relation to the theoretical bound of the Restricted Isometry Property (RIP), δ_S. (a): The data is corrupted by outliers only. (b): The data is corrupted by outliers and bounded inlier noise. Moreover, the empirical error is computed and the relation to its theoretical upper bound is depicted.



(a): Probability of recovery for $M = 100$



(b): Transition from success to failure.

Fig. 3: (a): The probability of recovery while varying the fraction of outliers, for the the dimension $M = 100$ of the unknown vector, $\boldsymbol{\theta}$, and $N = 600$ observations. As the fraction of the outliers increases, the probability for an accurate estimation drops. (b): Transition from success to failure with probability $p = 0.5$. A vertical line at $M = 100$ indicates the percentage of outliers (for each method respectively) that correspond to the values of the x-axis for probability $p = 0.5$, in (a).

Table 1: Computed MSE, for various experiments. In tests A, B and C, the noise is drawn from the heavy-tailed distribution alpha-stable of Lévy distribution. In test D, noise consists of a sum of two vectors, drawn from 2 independent Gaussian distributions with different variance, plus an outlier noise vector of impulsive noise.

Algorithm	Test A	Test B	Test C	Test D
GARD	0.1772	0.0180	0.0586	0.690
M-est	0.2248	0.2859	1.844e+06	0.704
SOCP	0.4990	0.3502	5.852e+05	1.011
SBL	0.9859	58.3489	2.165e+06	1.292
ROMP	0.2248	0.2859	1.844e+06	0.704

consists of outliers plus inlier noise, with values drawn from two independent Gaussian distributions with different variance.

3 Robust Nonlinear Regression

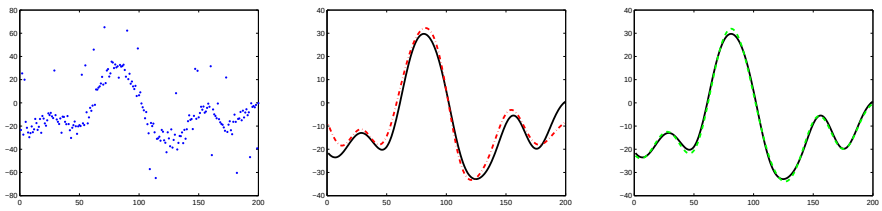
For the study of the nonlinear regression task we have assumed that the original function to be estimated lies in a Reproducing Kernel Hilbert Space (RKHS). Thus, we resort to simple manipulations by replacing the regression matrix with a kernel one. However, since this is a nonparametric estimation task, the proposed robust algorithm had to be modified again (with respect to GARD). The novel scheme, Kernel Greedy Algorithm for Robust Denoising (KGARD), alternates between a Kernel Ridge Regression (KRR) task and an OMP-like selection step. The addition of a regularization term at the estimation steps cannot be avoided and leads to a more complex theoretical analysis for the method. Thus, a different path, than the previously reported one (linear case) is followed. The study of this greedy-based selection scheme led to some interesting results:

- The solution to the regularized Least Squares task, which is performed at each step, is unique.
- The establishment of a bound on the maximum singular value of the kernel matrix, which guarantees that the method identifies the correct locations of all the outliers, first.

However, the method still manages to recover the correct support of the sparse outlier vector in many cases where the theoretical result does not hold. This leads to the conclusion that the provided conditions can be loosen up significantly in the future. The reason that the analysis is carried out for the case where inlier noise is not present is due to the fact that the analysis gets highly involved. The absence of the inlier noise makes the analysis easier and it highlights some theoretical aspects on why the method works. It must be emphasized that, such a

theoretical analysis appears for the first time in the related bibliography. Moreover, in practice, where inlier noise also exists, the method succeeds to correctly identify the majority of the outliers. The significance of the robust nonlinear regression task, is demonstrated in Figure 4, where the estimation with KGARD is compared against the non-robust Kernel Ridge Regression (KRR) method.

On the experimental section, various simulations are performed designating the overall advantages of KGARD against its competitors. In the tests performed, we have measured the MSE, the Mean Implementation Time (MIT) and the number of correct and wrong indices that each method has classified as outliers. In Table 2, the results of the estimation over the nonlinear function $\mathrm{f} = 20sinc(2\pi x)$ are depicted, for various levels of noise (inlier-outlier). It is observed that, KGARD attains the lowest MSE for most of the cases, except for the fraction of outliers at 20%. It should also be noted that, for small fractions of outliers the computational cost of the method is very low, and additionally, it successfully manages to identify the outliers.



(a) (b) (c)

Fig. 4: The significance of robust estimation: (a) Data corrupted by both inlier and 10% of outlier noise. (b) The black and the red dashed lines correspond to the uncorrupted data and the non-robust estimation performed, respectively. The MSE over the training set is 10.79. (c) The black and the green dashed lines correspond to the uncorrupted data and the robust estimation performed with KGARD, respectively. The MSE over the training set is 1.21.

4 Applications to Image Denoising

Finally, we present the applications of the proposed method, i.e., KGARD, in the context of image denoising. In particular, the goal is to approximate the original image that is corrupted by Gaussian (inlier) plus salt and pepper noise (outliers). To this end, the method has been slightly modified and adapted to the task, so that no tuning parameters are involved; instead, the parameters are automatically tuned by the method. As a result, two novel methods are proposed for the task of robust denoising: a) a direct KGARD implementation

Table 2: Computed MSE for f(x) = $20sinc(2\pi x)$ over the training and validation set. Additionally, the percentage of correct and wrong indices that each method has classified as outliers and the Mean Implementation Time (MIT), for various levels of inlier and outlier noise, are evaluated.

Algorithm	MSE_{tr}	MSE_{val}	Cor. ind.	Wr. ind.	MIT (sec)	Inlier - Outlier
RB-RVM	0.0850	0.0851	-	-	0.298	20 dB - 5%
RAM ($\lambda = 0.07, \mu = 2.5$)	0.0344	0.0345	100 %	0.2 %	0.005	20 dB - 5%
KGARD ($\lambda = 0.2, \varepsilon = 10$)	**0.0285**	**0.0285**	100 %	0 %	0.004	20 dB - 5%
RB-RVM	0.0911	0.0912	-	-	0.298	20 dB - 10%
RAM ($\lambda = 0.07, \mu = 2.5$)	0.0371	0.0372	100 %	0.1 %	0.007	20 dB - 10%
KGARD ($\lambda = 0.2, \varepsilon = 10$)	**0.0305**	**0.0305**	100 %	0 %	0.008	20 dB - 10%
RB-RVM	0.0992	0.0994	-	-	0.299	20 dB - 15%
RAM ($\lambda = 0.07, \mu = 2$)	0.0393	0.0393	100 %	0.6 %	0.008	20 dB - 15%
KGARD ($\lambda = 0.3, \varepsilon = 10$)	**0.0330**	**0.0330**	100 %	0 %	0.012	20 dB - 15%
RB-RVM	0.1189	0.1184	-	-	0.305	20 dB - 20%
RAM ($\lambda = 0.07, \mu = 2$)	**0.0421**	**0.0422**	100 %	0.4 %	0.010	20 dB - 20%
KGARD ($\lambda = 1, \varepsilon = 10$)	0.0626	0.0626	100 %	0 %	0.017	20 dB - 20%
RB-RVM	0.3630	0.3631	-	-	0.327	15 dB - 5%
RAM ($\lambda = 0.15, \mu = 5$)	0.1035	0.1036	100%	0.7 %	0.005	15 dB - 5%
KGARD ($\lambda = 0.3, \varepsilon = 15$)	**0.0862**	**0.0862**	100 %	0.1 %	0.005	15 dB - 5%
RB-RVM	0.3828	0.3830	-	-	0.319	15 dB - 10%
RAM ($\lambda = 0.15, \mu = 5$)	0.1117	0.1118	100%	0.4 %	0.006	15 dB - 10%
KGARD ($\lambda = 0.3, \varepsilon = 15$)	**0.0925**	**0.0925**	100 %	0 %	0.008	15 dB - 10%
RB-RVM	0.4165	0.4166	-	-	0.317	15 dB - 15%
RAM ($\lambda = 0.15, \mu = 5$)	0.1186	0.1186	100%	0.3 %	0.007	15 dB - 15%
KGARD ($\lambda = 0.3, \varepsilon = 15$)	**0.1001**	**0.1003**	100 %	0 %	0.012	15 dB - 15%
RB-RVM	0.4793	0.4798	-	-	0.312	15 dB - 20%
RAM ($\lambda = 0.15, \mu = 4$)	**0.1281**	**0.1282**	100%	1.4 %	0.008	15 dB - 20%
KGARD ($\lambda = 0.7, \varepsilon = 15$)	0.1340	0.1349	100 %	0 %	0.016	15 dB - 20%

that can perform the estimation and b) a KGARD scheme combined with a popular wavelet-based method, i.e., Block Matching and 3-D filtering (BM3D). The latter scheme, which first performs the identification and estimation of the outliers via the proposed algorithm (KGARD) and then it removes the remaining of the noise via the BM3D, demonstrated enhanced performance in terms of approximation. The results have been averaged based on the measured Peak signal-to-noise ratio (PSNR).

In Table 3, various results are given for the denoising of the Lena image. In Figure 5, the result of the process is clearly demonstrated. Finally, in Table 4 various results on the denoising of the boat image are depicted, while in Figure 6 the improvement achieved by the combined KGARD-BM3D method is observed.

Table 3: Denoising performed on the *Lena* image corrupted by various types and intensities of noise using the proposed methods, the robust RVM (RB-RVM) approach and the state-of-the-art wavelet method BM3D.

Method	Parameters	Gaussian Noise	Impulses (±100)	PSNR
BM3D	$s = 30$	25 dB	10%	30.84 dB
RB-RVM	$\sigma = 0.3$	25 dB	10%	31.25 dB
KGARD	$\sigma = 0.3$, $\lambda = 1$	25 dB	10%	33.49 dB
KGARD-BM3D	$\sigma = 0.3$, $\lambda = 1$, $s = 10$	25 dB	10%	**35.67 dB**
BM3D	$s = 35$	20 dB	10%	30.66 dB
RB-RVM	$\sigma = 0.4$	20 dB	10%	29.09 dB
KGARD	$\sigma = 0.3$, $\lambda = 1$	20 dB	10%	31.94 dB
KGARD-BM3D	$\sigma = 0.3$, $\lambda = 1$, $s = 15$	20 dB	10%	**33.81 dB**
BM3D	$s = 40$	15 dB	10%	29.94 dB
RB-RVM	$\sigma = 0.4$	15 dB	10%	25.85 dB
KGARD	$\sigma = 0.3$, $\lambda = 2$	15 dB	10%	28.47 dB
KGARD-BM3D	$\sigma = 0.3$, $\lambda = 1$, $s = 25$	15 dB	10%	**30.77 dB**

Table 4: Denoising performed on the *boat* image corrupted by various types and intensities of noise using the state-of-the-art wavelet method BM3D with and without outlier detection.

Method	Parameters	Gaussian Noise	Impulses (±100)	PSNR
BM3D	$s = 25$	25 dB	5%	30.57 dB
KGARD-BM3D	$\sigma = 0.3$, $\lambda = 1$, $s = 10$	25 dB	5%	**34.61 dB**
BM3D	$s = 35$	20 dB	10%	28.97 dB
KGARD-BM3D	$\sigma = 0.3$, $\lambda = 1$, $s = 15$	20 dB	10%	**31.52 dB**
BM3D	$s = 50$	20 dB	20%	27.49 dB
KGARD-BM3D	$\sigma = 0.4$, $\lambda = 1$, $s = 15$	20 dB	20%	**29.7 dB**

(a) (b)

(c) (d)

Fig. 5: (a) The *Lena* image corrupted by 20 dB of Gaussian noise and 10% outliers. (b) Denoising with BM3D (30.66 dB). (c) Denoising with KGARD (31.94 dB). (d) Denoising with joint KGARD-BM3D (33.81 dB).

(a) (b) (c)

Fig. 6: (a) The *boat* image corrupted by 20 dB of Gaussian noise and 10% outliers. (b) Denoising with BM3D (28.97 dB). (c) Denoising with joint KGARD-BM3D (31.52 dB).

5 Conclusions

In this dissertation we studied the problem of robust regression, for both the linear and nonlinear case, under the framework of sparse optimization techniques. Two novel algorithms are derived, for each case, and they are compared against state-of-the-art methods through extensive simulations. The results demonstrated enhanced performance, in terms of estimation and computational cost. Moreover, theoretical results, which guarantee the identification of the outliers, are provided. Finally, the proposed framework is applied to the task of image denoising, where it is shown that the process is significantly improved.

References

1. Papageorgiou, G., Bouboulis, P., Theodoridis, S.: Robust non-linear Regression: A Greedy Approach Employing Kernels. IEEE Transactions on Signal Processing, accepted (2017)
2. Papageorgiou, G., Bouboulis, P., Theodoridis, S.: Robust Linear Regression Analysis - A Greedy Approach. IEEE Transactions on Signal Processing 63(15), 3872–3887 (2015)
3. Papageorgiou, G., Bouboulis, P., Theodoridis, S.: Robust Regression in RKHS - An Overview. In: Proceedings of the European Signal Processing Conference, pp. 2874–2878. IEEE Press, New York (2015)
4. Papageorgiou, G., Bouboulis, P., Theodoridis, S.: Robust Linear Regression Analysis - The Greedy Way. In: Proceedings of the European Signal Processing Conference, pp. 16–20. IEEE Press, New York (2014)
5. Papageorgiou, G., Bouboulis, P., Theodoridis, S.: Robust Image Denoising in RKHS via Orthogonal Matching Pursuit. In: International Workshop on Cognitive Information Processing, pp. 1–6. IEEE Press, New York (2014)
6. Papageorgiou, G., Bouboulis, P., Theodoridis, S.: Robust Kernel-Based Regression Using Orthogonal Matching Pursuit. In: International Workshop on Machine Learning for Signal Processing, pp. 1–6. IEEE Press, New York (2013)