

# ROBUST KERNEL-BASED REGRESSION USING ORTHOGONAL MATCHING PURSUIT

*George Papageorgiou, Pantelis Bouboulis, Sergios Theodoridis\**

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

## ABSTRACT

Kernel methods are widely used for approximation of non-linear functions in classic regression problems, using standard techniques, e.g., Least Squares, for denoising data samples in the presence of white Gaussian noise. However, the approximation deviates greatly, when impulse noise outlying the data enters the scene. We present a robust kernel-based method, which exploits greedy selection techniques, particularly *Orthogonal Matching Pursuit* (OMP), in order to recover the sparse support of the outlying vector; at the same time, it approximates the non-linear function via the mapping to a Reproducing Kernel Hilbert Space (RKHS).

**Index Terms**— Robust Least Squares, Greedy Algorithms, Outliers, Orthogonal Matching Pursuit (OMP), Kernel-Based Regression, Reproducing Kernel Hilbert Space (RKHS)

## 1. INTRODUCTION

A task of major interest in Machine Learning has always been that of parameter estimation. *Regression analysis* is the statistical technique for establishing the relation among a set of input-output variables. The performance of regression analysis methods, in practice, depends on the data generating process, where the existence of noise plays a key role.

Usually, the assumption of additive white (Gaussian) noise is the way to model, attack and finally solve many practical problems, including classic regression ones. The main drawback in this modelling is robustness. A number of issues are posed, concerning the performance of the method, in the presence of non-Gaussian noise, e.g., with extreme noise values. In the last few years, the outliers' modelling has gained in importance in the context of what is known as Big Data applications. Within our study, we will follow a path that explicitly models outliers and makes use of appropriate regularization techniques.

\*This work was carried out under the 621 ARISTEIA program, cofinanced by the Greek Secretariat for Research and Development and the EU.

In a set of numerical data, any value that is markedly smaller or larger than other surrounding values (locally extreme value) is called an *outlier* [1]. In general, outliers are quite difficult to be defined. There is as much controversy over what constitutes an outlier, as whether to remove them or not. Most often, is up to the analyst to decide. In an efficient mathematical approach, an outlier is “information” that is not an inlier, i.e., neither originated from the original data source nor from a common (usually expected) noise source. However, an important key characteristic that defines outliers is *sparsity*, i.e., extreme values are expected to be of insufficient amount.

Lately, there has been an increased interest in the development of robust methods for denoising data samples containing outliers, since it is known that classic ones, e.g., Least Squares, fail. Our focus in this paper, will be upon the implementation of a robust method, which detects the outlier support as well as the “extreme” values themselves, and at the same time it obtains estimates of the original data using kernel functions. At the heart of our method lies a greedy selection algorithm. In particular, the performance of the *Orthogonal Matching Pursuit* (OMP) towards error reduction, provided the necessary spark in order to turn our focus towards this specific direction. Furthermore, it is known that OMP performs best when trying to recover very sparse vectors, which is usually the case when dealing with outlying observations.

Although OMP lacks the stability and consistency towards recovering the sparsest vector in the general case of a redundant dictionary [2, 3, 4], it turns out the special structure of the matrix employed by the proposed algorithm ensures that, in most cases, the exact support of the outlier vector is recovered (see Section 4).

## 2. PROBLEM MODELLING AND PRIOR WORK

Consider a finite set of training points  $(y_k, \mathbf{x}_k)$ ,  $k = 1, 2, \dots, n$ , with  $y_k \in \mathbb{R}$  and  $\mathbf{x}_k \in \mathbb{R}^m$ . The goal of a typical regression

task is to estimate the input-output relation via a model of the form

$$y_k = f(\mathbf{x}_k) + \eta_k, \quad k = 1, 2, \dots, n, \quad (1)$$

where  $\eta_k$  is an unobservable noise sequence, usually assumed to be white (Gaussian) noise. In the case where  $f$  is a linear function, the problem is only limited in computing the corresponding coefficients that define a hyperplane, on which the solution lies. In the more general case, where  $f$  is a non-linear function, we will assume that  $f$  belongs to a space of “smooth” functions  $\mathcal{H}$ , which will be assumed to have a structure of a reproducing kernel Hilbert space (RKHS). The kernel function and the norm induced by the inner product are  $\kappa : \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}$  and  $\|\cdot\|_{\mathcal{H}} = \sqrt{\langle \cdot, \cdot \rangle_{\mathcal{H}}}$ , respectively.

The Representer Theorem guarantees that, over the training set, the minimizer of the regularized minimization problem

$$\min_f \sum_{k=1}^n (y_k - f(\mathbf{x}_k))^2 + \mu \|f\|_{\mathcal{H}}^2, \quad \mu \geq 0 \quad (2)$$

admits a representation  $\hat{f}(\mathbf{x}) = \sum_{i=1}^n \alpha_i \kappa(\mathbf{x}, \mathbf{x}_i)$ , where  $\alpha_i$  are the unknown (real) coefficients. The regularized term in (2), is used in order to guard our method against overfitting, a standard technique in Machine Learning tasks. For more details see [5], [6].

Problem (2) is actually a Least Squares task in the RKHS  $\mathcal{H}$ . Although similar tasks can be successfully applied to remove Gaussian noise [7, 8, 9], it has been established that the presence of outliers causes their solution to overfit [10]. Hence, a data sequence containing outliers should not be modelled via (1). To this end, a sequence,  $u_k$ , associated with the outliers is explicitly modelled and the input-output relation takes the form:

$$y_k = f(\mathbf{x}_k) + u_k + \eta_k, \quad k = 1, 2, \dots, n. \quad (3)$$

As outliers are expected to often comprise a small fraction of the training sample, most of the values of  $u_k$  are zeros. In general, a percentage of less than 20% of non zero values is expected, thus  $\mathbf{u} := (u_1, u_2, \dots, u_n)^T$  is modelled as a *sparse* vector. Now that we have paved the way, it seems appropriate to reveal the gains while working under the sparsity approximation umbrella.

Prior knowledge of sparsity over vector  $\mathbf{u}$ , provides the tools to form the nonconvex minimization problem

$$\min_{\mathbf{u}, f \in \mathcal{H}} \sum_{k=1}^n (y_k - f(\mathbf{x}_k) - u_k)^2 + \mu \|f\|_{\mathcal{H}}^2 + \lambda \|\mathbf{u}\|_0, \quad (4)$$

where  $\mu > 0$ , is a user defined parameter controlling the trade-off between the two main goals of this task, i.e., minimizing the error, while keeping the complexity of the model, i.e.,  $\|f\|_{\mathcal{H}}$ , low. Values of  $\lambda \geq 0$  are set in order to control the sparsity levels of vector  $\mathbf{u}$ . This formulation was introduced in [11]. In this paper, we cast the task in the following

formulation,

$$\begin{aligned} & \min_{\mathbf{u}, f \in \mathcal{H}} \|\mathbf{u}\|_0 \\ \text{s.t.} & \sum_{k=1}^n (y_k - f(\mathbf{x}_k) - u_k)^2 + \lambda \|f\|_{\mathcal{H}}^2 \leq \varepsilon, \end{aligned} \quad (5)$$

for fixed threshold parameters  $\varepsilon \geq 0$  and  $\lambda > 0$ .

Although problem (4) (or (5)) is of an NP-hard combinatorial nature, greedy selection algorithms succeed in recovering the solution for certain data and model parameters. Another notable fact is the relation between (4) and the variational Least Trimmed Squares (VLTS) problem, for certain levels of sparsity of vector  $\mathbf{u}$ , in [11].

## 2.1. Convex relaxation

It is evident that problem (4) is a nonconvex optimization task. To achieve stable solutions and robust properties, many authors prefer to consider an alternative convex task, which (in some sense) is close to the original problem. This is the popular convex relaxation technique. In our case, substituting the  $\ell_0$  with the  $\ell_1$  norm, problem (4) can be cast as:

$$\min_{\mathbf{u}, f \in \mathcal{H}} \sum_{k=1}^n (y_k - f(\mathbf{x}_k) - u_k)^2 + \mu \|f\|_{\mathcal{H}}^2 + \lambda \|\mathbf{u}\|_1.$$

The precedent problem, presented in [11], was solved using an alternating direction algorithm (ADM). Despite the fact that in this case we deal with a convex problem, relaxation seems to lose some of its immediacy. As a result, our gains towards estimation or error reduction may not be ultimately achieved.

## 3. KERNEL REGULARIZED OMP (KROMP)

The standard sparse approximation denoising problem has been studied in an extended list of papers, [3, 4, 12, 13, 14, 15, 16]. Our focus in this work is to provide a robust kernel-based denoising method that can efficiently remove not only the typical Gaussian noise, but also impulses and types of noise with heavy tailed distributions. To this end, we aim at efficiently solving problem (5).

Let  $\phi(\cdot) : \mathbb{R}^m \rightarrow \mathcal{H}$ ,  $\phi(\mathbf{x}) = \kappa(\cdot, \mathbf{x})$ , denote the feature map of  $\mathcal{H}$ , that transforms the data from the input to the feature space  $\mathcal{H}$ , where  $\mathcal{H}$  is the RKHS induced by the kernel  $\kappa$ . Under this framework, we map the data to a high dimensional space,  $\mathcal{H}$ , which gives us the luxury of adopting linear tools to attack the specific problem. Furthermore, the reproducing property of the RKHS, i.e.,  $\kappa(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathcal{H}}$ , ensures that the actual structure of the space may be ignored, as the computation of any inner product can be given by the kernel function. Recall that for every set of points,  $\mathbf{x}_i, \mathbf{x}_j$ ,  $i, j = 1, 2, \dots, n$ , the *Gram* matrix  $K_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$  is a positive (semi-)definite matrix. Although a variety of kernel functions

are available, the most standard, which is also used in our experiments, is the Gaussian radial basis function kernel with parameter  $\sigma$ , see [5].

In the following, we make the a priori assumption that the estimated function,  $f$ , can be expressed as a finite linear combination of kernel functions centered at the training data, i.e.,

$$f = \sum_{k=1}^n \alpha_k \kappa(\cdot, \mathbf{x}_k) + c,$$

Hence, instead of solving problem (5), we target our efforts at estimating the solution of

$$\text{s.t.} \quad \min_{\mathbf{u}, \boldsymbol{\alpha}, c} \|\mathbf{u}\|_0 \quad (6)$$

$$\|K\boldsymbol{\alpha} + c\mathbf{1} + \mathbf{u} - \mathbf{y}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_2^2 + \lambda c^2 \leq \varepsilon,$$

where  $\boldsymbol{\alpha} \in \mathbb{R}^n$ ,  $c \in \mathbb{R}$  are the kernel expansion coefficients and the bias,  $\mathbf{1} \in \mathbb{R}^n$  is the vector of ones and  $\mathbf{y}$ ,  $\mathbf{u} \in \mathbb{R}^n$  are the measurement and outlier vectors, respectively.

At this point, it is important to make the following remarks. The quadratic inequality constraint in (6) could also be written as  $J(\mathbf{z}) = \|A\mathbf{z} - \mathbf{y}\|_2^2 + \lambda\mathbf{z}^T B\mathbf{z} \leq \varepsilon$ , where

$$A = \begin{bmatrix} K & \mathbf{1} & I_n \end{bmatrix}, \quad \mathbf{z} = \begin{bmatrix} \boldsymbol{\alpha} \\ c \\ \mathbf{u} \end{bmatrix}, \quad B = \begin{bmatrix} I_n & \mathbf{0} & O_n \\ \mathbf{0}^T & 1 & \mathbf{0}^T \\ O_n & \mathbf{0} & O_n \end{bmatrix},$$

while  $I_n$  denotes the unitary matrix,  $\mathbf{0}$  the zero vector and  $O_n$  the all zero square matrix.

At each step, according to the OMP rationale, our algorithm selects the most correlated column and attempts to solve  $\min_{\mathbf{z}} J(\mathbf{z})$ . First of all, notice that the square symmetric matrix  $B$  is a projection matrix, i.e.,  $B = B^2$ , of vectors in  $\mathbb{R}^{2n+1}$  to the lower dimension subspace  $\mathbb{R}^{n+1}$ . Substituting in  $J(\mathbf{z})$  and reformulating, our minimization problem becomes equivalent to

$$\min_{\mathbf{z}} \left\| \begin{pmatrix} A \\ \sqrt{\lambda}B \end{pmatrix} \mathbf{z} - \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} \right\|_2^2, \quad (7)$$

which could be viewed as a classic Least Squares problem.

Next, we would like to see if  $J(\mathbf{z})$  in (7) attains a minimum and whether it is unique. Note, that for any data set  $(y_k, \mathbf{x}_k)$ ,  $k = 1, 2, \dots, n$ , and for all  $\varepsilon \geq 0$ , we can find  $\mathbf{z}$  such that  $J(\mathbf{z}) \leq \varepsilon$ . This means that the feasible set of (6) is always nonempty<sup>1</sup>. Finally, recall that (7) acquires a unique solution, if and only if the nullspaces of  $A$  and  $B$  intersect only trivially, i.e.,  $\mathcal{N}(A) \cap \mathcal{N}(B) = \{0\}$  ([17], [18]). For simplicity, let  $D = \begin{pmatrix} A \\ \sqrt{\lambda}B \end{pmatrix}$ . It is straightforward to prove that the set of *normal equations* obtained from (7) is

$$(A^T A + \lambda B)\mathbf{z} = A^T \mathbf{y},$$

where  $(A^T A + \lambda B)$  is invertible, as the following proposition establishes:

**Proposition 1** *Matrix  $A^T A + \lambda B$  is (strictly) positive definite, hence invertible.*

To prove this, decompose a nonzero vector  $\mathbf{x}$  into three parts (according to dimensions of  $A$ ) and  $\mathbf{x}^T M \mathbf{x}$  is a strictly positive quantity for arbitrary  $\mathbf{x} \neq \mathbf{0}$ . Moreover, we make use of the following well known theorem.

**Theorem 1** *Matrix  $A^T A + \lambda B$  is (strictly) positive definite if and only if the columns of matrix  $D$  are linearly independent, i.e.,  $\text{rank}(D) = 2n + 1$ .*

Consequently, the minimizer  $\mathbf{z}^* \in \mathbb{R}^{2n+1}$  of (7) is unique. See, also [19].

Another formulation of (7) for  $\delta > 0$  is

$$\min_{\mathbf{z}} \|A\mathbf{z} - \mathbf{y}\|_2^2, \quad \text{s.t.} \quad \|B\mathbf{z}\|_2 \leq \delta. \quad (8)$$

Equivalence between (8) and (7) is well established in [20]. Now suppose  $\mathbf{z}_{\lambda_*}$  is the unique minimizer of (7) for certain  $\lambda_* > 0$ . Then  $\mathbf{z}_{\lambda_*}$  solves problem (8) with  $\delta = \|B\mathbf{z}_{\lambda_*}\|_2$ . Problem (8) reveals the physical properties of the problem and opens the way to interpret the performance of the greedy algorithm.

### 3.1. Algorithm implementation

The basic concept of the algorithm, lies in the restriction of the column selection set over the last part of matrix  $A$ , i.e., matrix  $I_n$ , whose columns form an orthonormal basis in  $\mathbb{R}^n$ . This is due to our modelling; therefore, outliers are expected to appear in the third part of vector  $\mathbf{z}$ , i.e., vector  $\mathbf{u}$ , which is known to be sparse. This comprises the major task of our new algorithm, KROMP, which is to perform a selection over the active set of columns. An overview of the algorithm and its converging properties (error reduction per step) are described below.

Prior to the implementation, computation of the kernel matrix  $K$  is required, given the input vectors  $\mathbf{x}_k$ ,  $k = 1, 2, \dots, n$  and using the Gaussian radial basis kernel function with kernel parameter  $\sigma$ , i.e.,  $\kappa_\sigma(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right)$ . At this point, we should emphasize that a careful tuning of the kernel parameter should be made, since correct selection determines whether the algorithm identifies the actual outlier support or not. This is also the case for  $\lambda$  and  $\varepsilon$  values. The method used in the present work is cross validation, for  $\lambda$  as well as for  $\sigma$  values. The same level of sensitivity also holds for the convex relaxation problem, see [11].

At each step, we define two separate sets, one for the *active* and another for the *inactive* columns of matrix  $A$ , denoted  $S_{ac}$  and  $S_{inac}$ , respectively. Initially, we fix  $S_{inac}$  to include indices from the first  $n + 1$  columns of matrix  $A$  and define a)  $A_{inac}^{(0)}$  as the matrix that contains the columns of  $A$ , whose indices belong to  $S_{inac}$  and b)  $B_{inac}^{(0)}$  as the matrix containing the rows of  $B$ , whose indices belong to  $S_{inac}$ . During the

<sup>1</sup>For example, if we select  $\mathbf{z} = (0, 0, \mathbf{y})^T$ , then  $J(\mathbf{z}) = 0$ .

algorithmic process, matrix  $A$  is augmented by an optimal selected column and  $B$  is augmented by zeros, in order to match the column dimension of matrix  $A$ .

Let  $\mathbf{z}^{(0)}$  denote the solution of the regularized problem (8), restricted over the initial inactive columns of matrix  $A$  and let  $\mathbf{r}^{(0)} = A_{inac}^{(0)} \mathbf{z}^{(0)} - \mathbf{y}$  denote the initial residual. Since this is a noise removal method, it is expected that  $\mathbf{y} \notin \mathcal{R}(A_{inac}^{(0)})^2$ , irrespective of whether the added noise sequence is Gaussian, impulse (or both), or heavy tailed. Suppose that the  $\ell_2$  norm of the residual  $\mathbf{r}^{(0)}$  is below our threshold parameter  $\epsilon$ . This assumption underlies the fact that no impulse outlying the noise exists and that the problem is significantly simplified to solving a regularized Least Squares problem. In this case, the algorithm stops. However, if outliers are present, the algorithm will proceed in order to approximate the sparse outlier vector support.

At each iteration step,  $k$ , KROMP selects the index

$$\begin{aligned} j_k &:= \arg \min_{j \in S_{ac}} \|\mathbf{r}^{(k-1)} - \langle \mathbf{e}_j, \mathbf{r}^{(k-1)} \rangle \mathbf{e}_j\|_2^2 \\ &= \arg \max_{j \in S_{ac}} |r_j^{(k-1)}|, \end{aligned} \quad (9)$$

where  $r_j^{(k-1)}$  is the  $j$ -th coordinate of the residual vector  $\mathbf{r}^{(k-1)}$  and  $\mathbf{e}_j$  is the unit norm vector from the standard orthonormal basis of  $\mathbb{R}^n$ . Then,  $S_{inac}$  is enlarged by  $j_k$  and matrix  $A_{inac}^{(k-1)}$  is augmented by  $\mathbf{e}_{j_k}$ . Next, the solution of (8) is computed, i.e.,  $\mathbf{z}_*^{(k)} = (\boldsymbol{\alpha}^{(k)}, c^{(k)}, u_{j_1}, \dots, u_{j_k})^T \in \mathbb{R}^{n+k+1}$ , taking into account the replacement of matrix  $A$  with matrix  $A_{inac}^{(k)}$ . Finally, the residual is calculated as  $\mathbf{r}^{(k)} = A_{inac}^{(k)} \mathbf{z}_*^{(k)} - \mathbf{y}$ .

At  $k+1$  step, the process is repeated and another column  $\mathbf{e}_{j_{k+1}}$  is added to matrix  $A_{inac}^{(k)}$ . At this stage we have,

$$A_{inac}^{(k+1)} = [K \mathbf{1} \ \mathbf{e}_{j_1} \ \dots \ \mathbf{e}_{j_k} \ \mathbf{e}_{j_{k+1}}] = [A_{inac}^{(k)} \ \mathbf{e}_{j_{k+1}}].$$

Now, let  $\mathbf{z}_*^{(k+1)} \in \mathbb{R}^{n+k+2}$  be the unique minimizer of  $L_{k+1}(\mathbf{z}) = \|A_{inac}^{(k+1)} \mathbf{z} - \mathbf{y}\|_2^2$  subject to the constraint  $\|\mathbf{B}_{inac}^{(k)} \mathbf{z}\|_2^2 \leq \epsilon$  (i.e., the minimization function of (8) at the current step). It can be shown that the residual obtained at each iteration cycle is *strictly decreasing*. To this end, consider the vector  $\mathbf{z}_*^{(k)}$ , which denotes  $\mathbf{z}_*^{(k)}$  augmented by the opposite value of the  $j_{k+1}$ -th coordinate of the residual vector  $\mathbf{r}^{(k)}$ , i.e.,  $\mathbf{z}_*^{(k)} = (\mathbf{z}_*^{(k)}, -r_{j_{k+1}}^{(k)})^T$ . Observe that  $\mathbf{z}_*^{(k)}$  belongs to the feasible set defined by the inequality constraint of (8) at the current step<sup>3</sup>. Hence,  $L_{k+1}(\mathbf{z}_*^{(k+1)}) \leq L_{k+1}(\mathbf{z}_*^{(k)})$ .

<sup>2</sup>Denotes the range of a matrix.

<sup>3</sup>Geometrically the feasible set remains the same, while matrix  $B$  is augmented by zero elements at each step.

Moreover, we have that

$$\begin{aligned} L_{k+1}(\mathbf{z}_*^{(k)}) &= \|A_{inac}^{(k+1)} \mathbf{z}_*^{(k)} - \mathbf{y}\|_2^2 \\ &= \left\| \begin{bmatrix} A_{inac}^{(k)} & \mathbf{e}_{j_{k+1}} \end{bmatrix} \cdot \begin{pmatrix} \mathbf{z}_*^{(k)} \\ -r_{j_{k+1}}^{(k)} \end{pmatrix} - \mathbf{y} \right\|_2^2 \\ &= \|A_{inac}^{(k)} \mathbf{z}_*^{(k)} - r_{j_{k+1}}^{(k)} \mathbf{e}_{j_{k+1}} - \mathbf{y}\|_2^2 \\ &= \|\mathbf{r}^{(k)} - r_{j_{k+1}}^{(k)} \mathbf{e}_{j_{k+1}}\|_2^2 \\ &< \|\mathbf{r}^{(k)}\|_2^2, \end{aligned} \quad (10)$$

where the last strict inequality is due to the fact that  $|r_{j_{k+1}}^{(k)}| > 0$ , as  $j_{k+1}$  is selected according to (9)<sup>4</sup>. Thus, we conclude that

$$\|\mathbf{r}^{(k+1)}\|_2^2 = L_{k+1}(\mathbf{z}_*^{(k+1)}) \leq L_{k+1}(\mathbf{z}_*^{(k)}) < \|\mathbf{r}^{(k)}\|_2^2,$$

which proves our claim. Moreover, we can see that the residual eventually will drop below the predefined threshold,  $\epsilon$ , no matter how small this is. However, if the user selects a very small  $\epsilon$ , then the proposed procedure will continue and model all noise samples (even those originating from a Gaussian source) as impulses, filling up the vector  $\mathbf{u}$ , which will no longer be sparse. Hence, sensible tuning of  $\epsilon$  is of importance. Algorithm 1 describes the procedure in detail.

---

#### Algorithm 1 : Kernel Regularized OMP (KROMP)

---

**Input:**  $K, \mathbf{y}, \lambda, \epsilon$

**Initialization:**  $k := 0$

$S_{inac} = \{1, 2, \dots, n+1\}, S_{ac} = \{n+2, \dots, 2n+1\}$

$A_{inac} = [K \ \mathbf{1}], A_{ac} = I_n = [\mathbf{e}_1 \ \dots \ \mathbf{e}_n]$

**Solve:**  $\mathbf{z}^{(0)} := \arg \min_{\mathbf{z}} \|A_{inac} \mathbf{z} - \mathbf{y}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_2^2 + \lambda c^2$

**Initial Residual:**  $\mathbf{r}^{(0)} = A_{inac} \mathbf{z}^{(0)} - \mathbf{y}$

**while**  $\|\mathbf{r}^{(k)}\|_2 > \epsilon$  **do**

$k := k + 1$

**Find:**  $j_k := \arg \max_{j \in S_{ac}} |r_j^{(k-1)}|$

**Update Support:**

$S_{inac} = S_{inac} \cup \{j_k\}, S_{ac} = S_{ac} - \{j_k\}$

$A_{inac} = [A_{inac} \ \mathbf{e}_{j_k}]$

**Update Current solution:**

$\mathbf{z}^{(k)} := \arg \min_{\mathbf{z}} \|A_{inac} \mathbf{z} - \mathbf{y}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_2^2 + \lambda c^2$

**Update Residual:**  $\mathbf{r}^{(k)} = A_{inac} \mathbf{z}^{(k)} - \mathbf{y}$

**end while**

**Output:**  $\mathbf{z} = (\boldsymbol{\alpha}, c, \mathbf{u})^T$  after  $k$  iterations

---

Each step of KROMP involves solving a linear system using Cholesky decomposition, which has complexity of  $\mathcal{O}((n+k)^3)$ , where  $k \ll n$  and due to the fact that the decomposition is recomputed at each step. It should be noted, that the complexity of ADM, according to [11], is  $\mathcal{O}(n^2)$  per

<sup>4</sup>If  $r_{j_{k+1}}^{(k)} = 0$ , then  $\mathbf{r}^{(k)} = \mathbf{0}$  and the algorithm should have been terminated at iteration  $k$ .

iteration, plus a single step of  $\mathcal{O}(n^3)$  operations in order to initially decompose the matrix. However, the complexity of KROMP could be further reduced, by employing the matrix inversion Lemma (details are omitted, due to space limitations). This is possible since large blocks of the matrices  $A_{inac}$  and  $B_{inac}$  remain unchanged. Indeed, it is easy to check that

$$C^{(k)} = \begin{pmatrix} C^{(k-1)} & (A^{(k-1)})^T e_{j_k} \\ e_{j_k} A^{(k-1)} & 1 \end{pmatrix},$$

where  $C^{(k)}$  is the matrix that is inverted at iteration  $k$ , i.e.,  $C^{(k)} = (A^{(k)})^T A^{(k)} + \lambda B^{(k)}$ . In this case, the computational cost has been shown to reduce to  $\mathcal{O}((n+k)^2)$  per iteration, with an initial step of  $\mathcal{O}(n^3)$  complexity due to the inversion of

$$C^{(0)} = \begin{pmatrix} K^T K + \lambda I_n & K^T \mathbf{1} \\ \mathbf{1}^T K & n + \lambda \end{pmatrix}.$$

#### 4. EXPERIMENTAL RESULTS

To demonstrate the performance of the presented algorithm, we present two simple toy examples, as those used in [11], and compare it with the ADM algorithm.

In the first experiment (Figure 1 (a)), the goal is to estimate the true input-output relationship given by the two dimensional *sinc* function, i.e.,  $y = 20\text{sinc}(2\pi x)$  over the presence of 20 dB Gaussian noise and 10% outliers (with values taken randomly in the interval  $[-20, 20]$ ). The proposed algorithm (KROMP) attains a MSE equal to 0.051, while the ADM presented in [11] attains a MSE equal to 0.091.

In the second experiment (Figure 1 (b)), our data are generated by the three dimensional function

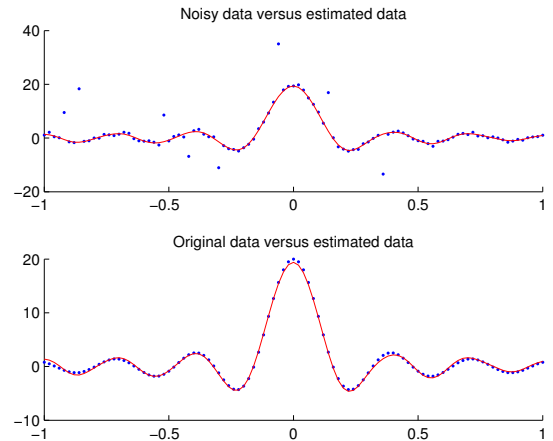
$$f(x, y) = 50\text{sinc}(\pi\sqrt{x^2 + y^2}),$$

corrupted by 15 dB Gaussian noise and 10% outliers (with values chosen randomly in the interval  $[-20, 20]$ ). Once again our algorithm excels, attaining a MSE equal to 1.022, while the ADM attains a MSE equal to 2.205.

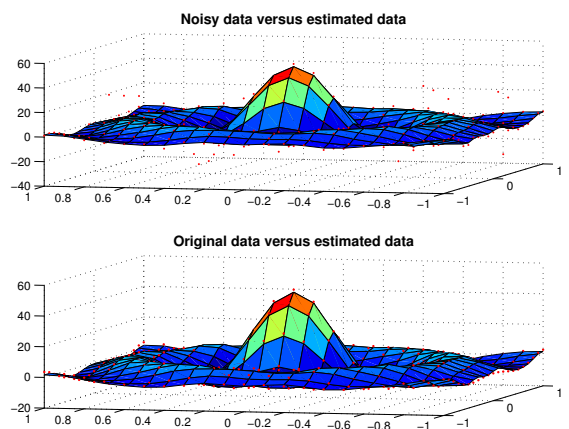
For both cases (a) and (b), the MSE is computed over 10000 computer generated data sets of 201 and  $21 \times 21 = 421$  points respectively. In both algorithms, the user-defined parameters were carefully tuned, so that to provide the smallest MSE possible. In particular, in the 2-D case, we set  $\epsilon = 0.5$  and  $\lambda = 1$  for KROMP, while we set  $\lambda = 1$  and  $\mu = 1$  for ADM. In the 3-D case, the parameters were set  $\epsilon = 1.2$  and  $\lambda = 0.7$  for KROMP, while  $\lambda = 0.2$  and  $\mu = 0.1$  for ADM.

In our last experiment, the input-output relationship is modelled by the function

$$y := g(x) = 50\sin(11(x - 1/3)\pi) + 40\cos(6(x + 2/3)\pi) - 20\sin(7(x - 1/2)\pi).$$



(a) The 2-D case over the presence of 20 dB Gaussian noise



(b) The 3-D case over the presence of 15 dB Gaussian noise

**Fig. 1.** KROMP approximation of *sinc* over the presence of Gaussian noise and 10% outliers.

In this case, 10000 data sets of 201 points each were also generated (which are also corrupted by 20 dB Gaussian noise and 10% outliers in the range  $[-200, 200]$ ). The respective MSE is 5.95 for KROMP and 8.68 for ADM. The parameters were set  $\epsilon = 5$  and  $\lambda = 0.125$  for KROMP, while  $\lambda = 10$  and  $\mu = 0.05$  for ADM. It is important to point out that in all tests performed, the support of the outlier vector was exactly recovered. The Matlab code can be found at <http://bouboulis.mysch.gr/kernels.html>.

The aforementioned experiments highlight the superior performance of the KROMP over the ADM, at approximately the same complexity level.

#### 5. CONCLUSIONS

In this paper, a robust method for nonlinear regression in the presence of outliers was presented. It has been proved, that

the algorithm monotonically reduces the error at each step, and it thus converges to a solution. Even though we cannot guarantee that our solution is optimal (due to the non convex nature of the problem), experimental results have shown that KROMP outperforms previous proposed methods, i.e., ADM.

Future research will include other, similar in philosophy greedy approaches, like *Orthogonal Least Squares* [21], [22], or an algorithm based on *Cyclic Matching Pursuit* [23], as well as a study of the algorithm's stability properties.

## 6. REFERENCES

- [1] F. E. Grubbs, "Procedures for detecting outlying observations in samples," *Technometrics*, vol. 11, no. 1, pp. 1–21, 1969.
- [2] Stephane Mallat, *A wavelet tour of signal processing: the sparse way*, Academic press, 2008.
- [3] 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.
- [4] Joel A Tropp, "Greed is good: Algorithmic results for sparse approximation," *Information Theory, IEEE Transactions on*, vol. 50, no. 10, pp. 2231–2242, 2004.
- [5] Alex J. Smola and Bernhard Schölkopf, *Learning with Kernels*, The MIT Press, 2002.
- [6] Sergios Theodoridis and Konstantinos Koutroumbas, *Pattern Recognition, 4th Edition*, Academic press, 2008.
- [7] Hiroyuki Takeda, Sina Farsiu, and Peyman Milanfar, "Kernel regression for image processing and reconstruction," *Image Processing, IEEE Transactions on*, vol. 16, no. 2, pp. 349–366, 2007.
- [8] Sylvain Durand and Jacques Froment, "Reconstruction of wavelet coefficients using total variation minimization," *SIAM Journal on Scientific computing*, vol. 24, no. 5, pp. 1754–1767, 2003.
- [9] Patrick L Combettes and J-C Pesquet, "Image restoration subject to a total variation constraint," *Image Processing, IEEE Transactions on*, vol. 13, no. 9, pp. 1213–1222, 2004.
- [10] Pantelis Bouboulis, Konstantinos Slavakis, and Sergios Theodoridis, "Adaptive kernel-based image denoising employing semi-parametric regularization," *Image Processing, IEEE Transactions on*, vol. 19, no. 6, pp. 1465–1479, 2010.
- [11] Gonzalo Mateos and Georgios B. Giannakis, "Robust nonparametric regression via sparsity control with application to load curve data cleansing," *Signal Processing, IEEE Transactions on*, vol. 60, no. 4, pp. 1571–1584, 2012.
- [12] Joel A Tropp and Anna C Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *Information Theory, IEEE Transactions on*, vol. 53, no. 12, pp. 4655–4666, 2007.
- [13] 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.
- [14] Jean-Jacques Fuchs, "Recovery of exact sparse representations in the presence of bounded noise," *Information Theory, IEEE Transactions on*, vol. 51, no. 10, pp. 3601–3608, 2005.
- [15] Brendt Wohlberg, "Noise sensitivity of sparse signal representations: Reconstruction error bounds for the inverse problem," *Signal Processing, IEEE Transactions on*, vol. 51, no. 12, pp. 3053–3060, 2003.
- [16] Hanxi Li, Yongsheng Gao, and Jun Sun, "Fast kernel sparse representation," in *Digital Image Computing Techniques and Applications (DICTA), 2011 International Conference on*. IEEE, 2011, pp. 72–77.
- [17] Walter Gander, *On the linear least squares problem with a quadratic constraint*, Computer Science Department, Stanford University, 1978.
- [18] Walter Gander, "Least squares with a quadratic constraint," *Numerische Mathematik*, vol. 36, no. 3, pp. 291–307, 1980.
- [19] Åke Björck, *Numerical methods for least squares problems*, Number 51. Society for Industrial and Applied Mathematics, 1996.
- [20] Marielba Rojas and Danny C Sorensen, "A trust-region approach to the regularization of large-scale discrete forms of ill-posed problems," *SIAM Journal on Scientific Computing*, vol. 23, no. 6, pp. 1842–1860, 2002.
- [21] Thomas Blumensath and Mike E Davies, "On the difference between orthogonal matching pursuit and orthogonal least squares," 2007.
- [22] Charles Soussen, Rémi Gribonval, Jérôme Idier, and Cédric Herzet, "Joint k-step analysis of orthogonal matching pursuit and orthogonal least squares," 2013.
- [23] Bob L Sturm, Mads G Christensen, and Rémi Gribonval, "Cyclic pure greedy algorithms for recovering compressively sampled sparse signals," in *Signals, Systems and Computers (ASILOMAR), 2011 Conference Record of the Forty Fifth Asilomar Conference on*. IEEE, 2011, pp. 1143–1147.