OWA Operators in Regression Problems

Ronald R. Yager, *Fellow, IEEE*, and Gleb Beliakov, *Senior Member, IEEE*

***Abstract*—We consider an application of fuzzy logic connectives to statistical regression. We replace the standard least squares, least absolute deviation, and maximum likelihood criteria with an ordered weighted averaging (OWA) function of the residuals. Depending on the choice of the weights, we obtain the standard regression problems, high-breakdown robust methods (least median, least trimmed squares, and trimmed likelihood methods), as well as new formulations. We present various approaches to numerical solution of such regression problems. OWA-based regression is particularly useful in the presence of outliers, and we illustrate the performance of the new methods on several instances of linear regression problems with multiple outliers.**

***Index Terms*—Aggregation operators, least trimmed squares (LTS), outliers, ordered weighted averaging (OWA), robust regression.**

I. Introduction

In THIS paper, we will look at an application of fuzzy logic connectives, in particular, the popular ordered weighted averaging (OWA) functions, to statistical regression. Outliers—atypical data that do not follow the regression model—can be very problematic in regression analysis. Even a single outlier can affect the regression model so much that it does not stand out. As the consequences, the computed model could be grossly erroneous, and the outlier becomes undetectable. The notion of an outlier is somewhat fuzzy: All data can be qualified as outliers to some degree, based on how well the regression model fits each datum. It makes sense to weight the contribution of each datum in the regression analysis based on the distance of this datum from the model, called the residual. By down-weighting poorly fitted data, which are considered outliers, we can limit their effect on the model. We investigate aggregation of residuals with various forms of OWA functions [1], and show how robust alternatives to traditional regression can be obtained as a result.

We consider the classical regression problem: Given a set of pairs $\{(x_k, y_k)\}, k = 1, \ldots, K$: $x_k \in \Re^n, y_k \in \Re$ (data), and a set of models $f_\theta : \Re^n \to \Re$ parameterized by a vector of parameters $\theta \in \Omega \subseteq \Re^p$, determine the parameter vector θ^*, such that f_{θ^*} fits the data best. The goodness of fit can be measured in different ways. Three classical instances are the least squares (LS) regression, the least absolute deviation (LAD) regression, and Chebychev (minimax) approximation. The maximum likelihood (ML) estimators give another set of instances. When functions f_θ depend on θ linearly, the problem is called linear regression, otherwise, it becomes a nonlinear regression problem.

We concentrate on linear regression, where the model is as follows:

$$y_i = x_{i1}\theta_1 + \cdots + x_{ip}\theta_p + \varepsilon_i, \quad i = 1, \ldots, K$$

with $x_{ip} = 1$ for regression with an intercept term. $\{x_{ij}\} = X \in \Re^{K \times p}$ is the matrix of explanatory variables and ε is a K-vector of independent identically distributed random errors with zero mean and (unknown) variance σ^2. The goal is to determine the vector of unknown parameters θ.

Simple regression ($p = n + 1 = 2$ with the intercept term) and multiple regression ($p > 2$) are classical instances of such a problem. Polynomial and spline regression (in one or multiple variables) can also be viewed as instances of such a problem: If $\{B_1, \ldots, B_p\}$ is a set of basis functions of a single variable t, then the explanatory variables are $x_{ik} = B_k(t_i)$. Another instance of linear regression problem we will be dealing with, is when there are additional linear constraints on the parameters θ. For example, when f_θ is chosen from the class of weighted arithmetic means $f_\theta(x) = x_1\theta_1 + \cdots + x_n\theta_n$, then, we have the constraints $\theta_i \geq 0$ and $\theta_1 + \cdots + \theta_n = 1$.

The goodness of fit is typically expressed in terms of either squared or absolute deviations (residuals) $r_k = f_\theta(x_k) - y_k$, namely, the weighted averages $\sum_{k=1}^K w_k r_k^2$ or $\sum_{k=1}^K w_k |r_k|$. More generally, in the ML estimators, one maximizes the log-likelihood function $\sum_{k=1}^K w_k l(\theta; r_k)$, where $l(\theta; r_k) = \log(\rho(r_k; \theta))$ is the logarithm of the probability density of the random variable R. The weights w_k reflect the relative importance of the kth datum: The larger the weight, the better $f_\theta(x_k)$ must approximate y_k. With no *a priori* information, typically equal weights are chosen.

Both the LS and LAD regression are sensitive to outliers—atypical points that do not follow the regression model. There are two types of outliers: The vertical outlier (only the value of y_i is atypical), and leverage points (the values x_{ik} are atypical). Leverage points often happen when some data are missing, and are replaced with some default values (like the notorious 9999). LAD regression handles well vertical outliers, but lacks robustness with respect to leverage points in the same way as LS regression [2], [3]. ML estimators are not robust against leverage points either [4].

The concept of the breakdown point ε^* was introduced in [5]. ε^* is the smallest proportion of contaminated data than can cause the regression estimator to take arbitrary large aberrant values. In the cases of LS and LAD, as well as ML estimators, $\varepsilon^* = 0$. Hence, even a single outlier can cause a wrong estimator. There are many studies in robust regression in which higher breakdown values were achieved, up to the maximum $\varepsilon^* = 0.5$ (see Section II).

Manuscript received April 30, 2009; revised August 10, 2009; accepted August 22, 2009. First published November 17, 2009; current version published February 5, 2010.

R. R. Yager is with Machine Intelligence Institute, Iona College, New Rochelle, NY 10801 USA (e-mail: yager@panix.com).

G. Beliakov is with the School of Information Technology, Deakin University, Burwood 3125, Australia (e-mail: gleb@deakin.edu.au).

Digital Object Identifier 10.1109/TFUZZ.2009.2036908

We look at an alternative way to aggregate the goodness of fit of individual data, by using OWA operators introduced in [1]. OWA operators allow us to associate nonnegative weights not with individual data, but with the magnitude of the residual r_k (or an appropriate function of r_k). Thus, we can either penalize large or small residuals, or, alternatively, not penalize the largest residuals, treating these data as outliers. The weighting vector of the OWA operator will control the penalties associated with each residual based on its ranking. It turns out that the standard LS, LAD, ML methods, and Chebyshev approximation, as well as several methods of robust regression, like the least median of squares (LMS), the least trimmed squares (LTS), trimmed absolute deviations, and trimmed likelihood methods, arise as special instances of OWA-based regression. We also present several numerical techniques for solving OWA-based regression problems.

The paper is structured as follows. In Section II, we formulate the regression problem and discuss previous work done in the field of high-breakdown robust regression. In Section III, we discuss various methods of numerical solution of the OWA-based regression problems. These methods will depend on the OWA weighting vector. In some cases, the optimization problem is convex, while in other cases, it is concave, or neither convex nor concave, and in all cases, it is nonsmooth. In Section IV, we present our numerical results when fitting datasets with multiple outliers, and show that OWA-based regression can effectively identify and filter out the outliers in the data. This section is followed with conclusions.

II. Problem Formulation and Some Prior Work

In the ordinary weighted linear LS regression, the optimal vector of parameters θ is found by minimizing

$$\text{Minimize } F(\theta) = \sum_{k=1}^{K} w_k (r_k(\theta))^2 \tag{1}$$

where the residuals are $r_k(\theta) = f_\theta(x_k) - y_k$. In the LAD regression, the parameters are found by minimizing

$$\text{Minimize } F(\theta) = \sum_{k=1}^{K} w_k |r_k(\theta)|.$$

Huber [6] suggested the use of criteria less sensitive to outliers, namely,

$$\text{Minimize } F(\theta) = \sum_{k=1}^{K} w_k \rho(|r_k(\theta)|)$$

with specially chosen functions ρ (continuous, strictly increasing, with $\rho(0) = 0$), which produce M-estimators (ML type estimators).

All mentioned methods are sensitive to the leverage points, and their breakdown point $\varepsilon^* = 0$. In order to make the estimators robust to outliers, the method of LMS was proposed in [2]. In this method, the following expression is minimized:

$$\text{Minimize } F(\theta) = \text{median}(r_k(\theta))^2.$$

In order to achieve the maximal breakdown point $\varepsilon^* = 0.5$, the median is replaced by the $[(K+p+1)/2]$th quantile (the $[x]$ denotes the nearest integer larger than or equal to x). Nowadays, Rousseeuw and Driessen [7] consider the method of the LTS, also proposed in [2], superior to the LMS, because the objective function is more smooth, its statistical efficiency is better and the convergence rate is higher, while it has the same breakdown point [2], [7], [8]. Here, the expression to be minimized is as follows:

$$\text{Minimize } F(\theta) = \sum_{k=1}^{h} (r_{(k)}(\theta))^2$$

where the residuals are ordered in the increasing order $|r_{(1)}| \le |r_{(2)}| \le \cdots \le |r_{(K)}|$, and $h = [(K+p+1)/2]$. The method of least trimmed absolute deviation (LTA) was advocated in [9] and [10]. The minimization problem is as follows:

$$\text{Minimize } F(\theta) = \sum_{k=1}^{h} |r_{(k)}(\theta)|.$$

In the maximum trimmed likelihood (MTL) methods, the expression

$$-\sum_{k=1}^{h} w_k l(\theta; r_{(k)})$$

is minimized, where l are log-likelihood functions [4].

In essence, in the LTS, LTA, and MTL methods, half of the sample is discarded as potential outliers, and the model is fitted to the remaining half. This makes the estimator not sensitive to up to the half of contaminated data. The problem of course is to decide, which half of the data should be discarded. The problem becomes NP-hard (like all robust estimation problems with high breakdown point [11], including LTS and the least quantile regression, another method in which instead of the median, the hth quantile of the absolute deviations is minimized [4]). The choice of $h = [(K+p+1)/2]$, while giving the highest breakdown point, limits the convergence and coverage of the method. Instead, often the values $h = 0.75K$ or $h = 0.9K$ are used (e.g., in statistical package SPlus), and in [12] an adaptive choice of h was proposed, based on the data.

Note that in all mentioned methods, instances of an OWA function (with different weighting vectors) are used. Let us formally define OWA operators and OWA-based regression.

Definition 1: An OWA function with the weighting vector $w \in [0,1]^K, \sum w_i = 1$ is the function

$$\text{OWA}_w(x) = \sum_{i=1}^{K} w_i x_{(i)}$$

where $x_{(i)}$ is the ith largest component of x.

We remind the basic facts about OWA function, for details the reader can consult [1], [13]–[17].

Definition 2: The orness of an OWA function OWA_w

$$\text{orness}(\text{OWA}_w) = \sum_{i=1}^{K} w_i \frac{K-i}{K-1}. \tag{2}$$

The entropy (dispersion) of an OWA function OWA_w is as follows:

$$\text{Disp}(\text{OWA}_w) = -\sum_{i=1}^{K} w_i \log w_i.$$

Special cases of OWA include the maximum [$w = (1, 0, \ldots, 0)$], the minimum [$w = (0, \ldots, 0, 1)$], the arithmetic mean [$w = ((1/K), \ldots, (1/K))$], and the median (for odd K, $w_{(K+1)/2} = 1$, for even K, $w_{K/2} = w_{K/2+1} = 1/2$). Depending on the properties of the weighting vector, one can obtain many other interesting cases, like the "olympic" OWA, which discards the largest and the smallest components of x ($w_1 = w_K = 0$), the average of the h largest (or smallest) components, etc.

Let us formulate an OWA-based regression problem.

$$\text{Minimize } F(\theta) = \sum_{k=1}^{K} w_k \rho(|r_{(k)}(\theta)|) \tag{3}$$

subject to $\theta \in \Omega \subseteq \Re^p$, where $r_{(k)}$ stands for the kth largest residual, and ρ is one of the functions appearing in M-estimators. The right hand side of (3) is the OWA function with the weights w_k. We now concentrate on the the functions $\rho(r) = r^q$ with $q \geq 1$, and particularly, on the cases $q = 1$ and $q = 2$. Müller [18] considered other choices for q (trimed weighted L_q estimators).

As special cases of (3), when $\Omega = \Re^p$, we obtain the LS and LAD, the LMS method, the least quantile regression, the LTS and LTA methods (with $w = (0, 0, \ldots, 0, u, u, \ldots, u)$, $u = 1/h$, and $q = 2$ and $q = 1$, respectively), and the MTL (when choosing ρ as negative log-likelihood function). Note that since a constant factor can be factored out from the objective function F in (3), we can use instead the weighting vectors defined by $w_k = 0$, $k = 1, \ldots, K - h$, and $w_k = 1$ for $k = K - h + 1, \ldots, K$, with $h = [(K + p + 1)/2]$, or $h = [0.9K]$, etc., i.e., we will not require w_i to sum to one.

In such cases, when the dimension of OWA operator changes from one problem instance to another (in our case it is the number of data K), it is convenient to use stress function to determine the weights [19].

Definition 3 Let $h : [0, 1] \to \Re_+$ be a nonnegative function on the unit interval. OWA weights are defined as follows:

$$w_i = \frac{1}{H} h\left(\frac{i}{K}\right), \qquad i = 1, \ldots, K \tag{4}$$

with $H = \sum_{i=1}^{K} h\,(i/K)$, the normalization constant.

Stress function is related to fuzzy linguistic quantifiers [15], [20], [21], by $Q'(t) = Hh(t)$, where H is the normalization constant, $Q : [0, 1] \to [0, 1], Q(0) = 0, Q(1) = 1$, a continuous monotone increasing function, called the regular increasing monotone (RIM) quantifier, and the value $Q(t)$ represents the degree to which t satisfies the fuzzy concept represented by the quantifier. Examples of such quantifiers for fuzzy sets are as "for all", "there exists", "identity", "most", "at least half", "as many as possible", etc.

The use of OWA operator in problem (3) allows us to model the following verbally expressed informal requirements.

1) We need to fit *all* data (standard LS or LAD problem). Here, we take $h(t) = 1$, for all t.
2) We need to fit, even *the worst* datum (Chebyshev approximation problem). Here, we take $h(0) = 1$ and $h(t) = 0, t > 0$.
3) We need to fit *most* data. An example of the corresponding stress function is $h(t) = 0$ for $t < 1/10$ and $h(t) = 1$, otherwise.
4) We need to fit *the majority of* the data. An example is

$$h(t) = \begin{cases} 0, & t < a \\ \frac{t-a}{b-a}, & a \leq t < b \\ 1, & t \geq b \end{cases} \tag{5}$$

with say, $a = 0$ and $b = 1/4$.
5) We need to fit *at least half* the data. We can take a piecewise linear h in (5), with $a = 1/3$, $b = 2/3$, etc.

For each of the mentioned requirements, we choose an appropriate stress function, generate the corresponding weighting vector w, and solve problem (3) for θ. Depending on the vector w, the methods of solution will be different. We outline them in the next section.

III. Methods of Solution

A. Decreasing Weighting Vector w

Let the vector w have the following property: $w_i \geq w_j$, for all $i < j$. We note that this happens when the RIM quantifier is a concave function (the stress function h is decreasing). One special case is $w = (1, 0, \ldots, 0)$, which results in Chebyshev approximation. The alternatives could be as follows.

1) $w = (\alpha, 1 - \alpha, 0, \ldots, 0)$, $\alpha > 1/2$, i.e., minimize the weighted mean of the two largest (squared, absolute) residuals.
2) $w = ((1/m), \ldots, (1/m), 0, \ldots, 0)$, i.e., minimize the mean of the m largest (squared, absolute) residuals, etc.

The problem (3) is reformulated as follows:

$$\text{Minimize} \max_{\pi} \sum_{k=1}^{K} w_k |r_{\pi(k)}(\theta)|^q \tag{6}$$

subject to $\theta \in \Omega$, where π denotes a permutation of the indexes $1, 2, \ldots, K$. This formulation follows from the observation that $\sum w_k |r_{(k)}(\theta)|^q \geq \sum w_k |r_{\pi(k)}(\theta)|^q$ for any π.

The implication of this result is that the objective function in (6) is convex. Hence, there exists a unique minimum of F, as long as Ω is convex. Note that for many other choices of ρ in M-estimators, when ρ is not convex, the objective is neither convex nor quasi-convex.

There are three main approaches to numerical solution of (3) in the case of decreasing weighting vectors.

1) Direct Method: The first method is to solve (3) directly by using methods of nonsmooth optimization, e.g., the discrete gradient, or bundle methods discussed in [22]–[25], and implemented in [26]. We note that the objective in (3) is not differentiable, and this creates difficulties for most off-the-shelf optimization packages. However, recent developments in nonsmooth optimization make this issue less relevant.

2) Linear Programming Formulation: The second approach, applicable to the case of $q = 1$ and Ω being a polytope, is to formulate an equivalent linear programming (LP) problem

$$\begin{aligned}
\text{minimize} \quad & \varepsilon \\
\text{s.t.} \quad & \varepsilon \geq \sum_{k=1}^{K} w_k \left(d^+_{\pi_1(k)} + d^-_{\pi_1(k)} \right) \\
& \varepsilon \geq \sum_{k=1}^{K} w_k \left(d^+_{\pi_2(k)} + d^-_{\pi_2(k)} \right) \\
& \vdots \\
& \varepsilon \geq \sum_{k=1}^{K} w_k \left(d^+_{\pi_M(k)} + d^-_{\pi_M(k)} \right) \\
& d^+_{\pi_m(k)} - d^-_{\pi_m(k)} = \sum_{j=1}^{n} \theta_j x_{kj} - y_k, \\
& d^+_{\pi_m(k)}, d^-_{\pi_m(k)} \geq 0, \ m = 1, \ldots, M. \qquad (7)
\end{aligned}$$

Here, $M = K!$, the total number of possible permutations π, π_m is the mth permutation, and the auxiliary nonnegative variables $d^+_{\pi_m(k)}, d^-_{\pi_m(k)}$ are simply the positive and negative parts of the residuals $f_\theta(x_k) - y_k$, so that $d^+_{\pi_m(k)} + d^-_{\pi_m(k)} = |f_\theta(x_k) - y_k| = |r_{\pi_m(k)}|$.

Of course the size of such an LP problem is $K!$, which means it will be numerically expensive, even for relatively small datasets. However, the size of this problem can be reduced in several important special cases, when some of the weights w_k coincide. For instance, consider the vector $w = ((1/2), (1/2), 0, \ldots, 0)$. Then, all permutations in which the first two or the last $K - 2$ indexes differ, are equivalent. In this case, instead of $M = K!$, we can take $M = K!/(2!(K-2)!) = K(K-1)/2$.

Similarly, when we have a weighting vector with three groups of identical values, like $w = (a, a, \ldots, a, b, \ldots, b, c, \ldots, c)$, we have $M = K!/(k!m!(K-k-m)!)$, where k is the number of as and m is the number of bs.

3) Mixed Integer Programming Formulation: The third method is based on mixed integer programming formulation, and was presented in [27]. Here, we use auxiliary integer variables $Z \in \{0,1\}^{K \times (K-2)}$ to represent and enforce ordering in the vector of deviations. Consider the following optimization problem:

$$\begin{aligned}
\text{minimize} \quad & \sum_{k=1}^{K} w_k c_k \\
\text{s.t.} \quad & d_k = |\theta_j x_{kj} - y_k|^q \quad (= |r_k|^q) \\
& c_1 - d_k \geq 0, k = 1, \ldots, K \\
& c_{i+1} - d_k + M \sum_{m=1}^{i} z_{mk} \geq 0 \\
& \text{for all } i = 1, \ldots, K-2, k = 1, \ldots, K \\
& \sum_{i=1}^{K} z_{ij} = 1, \text{ for all } j = 1, \ldots, K-1 \\
& \sum_{j=1}^{K-1} z_{ij} = 1, \text{ for all } j = 1, \ldots, K-1 \\
& z_{ij} \in \{0,1\}, i, j = 1, \ldots, K. \qquad (8)
\end{aligned}$$

The first set of inequality constraints $c_1 - d_k \geq 0$ ensures that $c_1 = d_{(1)}$, the largest absolute deviation. The second set of constraints ensures that $c_{i+1} = d_{(i)}$, the ith largest absolute deviation. Here, M is some large constant, larger than any possible value of c_i or d_k. Let us see how it works. c_2 must be larger than all, but smaller than one d_k. If $z_{1k} = 0$, then $c_2 \geq d_k$, the constraint is enforced, and if $z_{1k} = 1$, then, the constraint is relieved. There must be only one such inactive constraint, hence, the condition $\sum_{k=1}^{K} z_{1k} = 1$. Consider next c_3, which must be the third largest $d_{(3)}$, i.e., it must be larger than all, but smaller than two d_k. Therefore, we must relieve two constraints, but one of these is exactly the constraint with the same index k we relieved for c_2. Therefore, we use the term Mz_{1k} to relieve that constraint, and Mz_{2k} to relieve a new constraint. To ensure that we do not relieve the same constraint, we have the condition

$$\sum_{j=1}^{K-1} z_{ij} = 1, \text{ for all } i = 1, 2.$$

Proceeding in the same way for $c_j, j = 4, \ldots, K-1$, we obtain the other constraints, each time relieving one additional constraint. For c_K all the constraints are relieved. The binary variables z_{ij} can be conveniently represented as a matrix Z of size $K \times (K-2)$. Each row and each column of Z contains exactly one nonzero entry, and the kth element of column i is the index of the additional inequality constraint, which is relieved for c_{i+1}, so that $c_{i+1} = d_{(i+1)} = d_k$ when $z_{ik} = 1$.

If $q > 1$, the problem (8) is a convex mixed integer programming problem in which the variables are Z, θ, c_i. When $q = 1$, by splitting the components of d_k into positive and negative parts $d_k = d^+_k - d^-_k$, as it was done in (7), we obtain an equivalent mixed LP problem in variables $Z, \theta, c_i, d^+_k, d^-_k$, which can be solved by standard methods, like the branch-and-bound method [28].

B. Increasing Weighting Vector w

This is a very important case suitable for robust regression. The weights $w_i \leq w_j$, for all $i < j$. We note that this happens when the stress function is nondecreasing and the RIM quantifier is a convex function. Various methods of robust regression, notably, the LTS, LTA, and MTL, arise as special cases of (3) with the increasing weighting vector. For instance, when $\Omega = \Re^p$

1) $w = (0, \ldots, 0, 1, \ldots, 1)$, with $h = [(K + p + 1)/2]$ last components $w_k = 1$, $q = 2$ corresponds to LTS.
2) The same weighing vector, but $q = 1$ corresponds to LTA.
3) Other choices of ρ result in MTL methods.
4) w defined by (5) corresponds to a fuzzified, or weighted versions of LMS or LTA. Here, the outliers are not

eliminated, but down-weighted. We can call them *fuzzy* outliers, with the weights giving their membership grades in the set *outliers*.

In the case of increasing w, the problem (3) can be reformulated as follows:

$$\text{Minimize } \min_{\pi} \sum_{k=1}^{K} w_k |r_{\pi(k)}(\theta)|^q \tag{9}$$

subject to $\theta \in \Omega$, and subsequently, as

$$\min_{\pi} \min_{\theta \in \Omega} \sum_{k=1}^{K} w_k |r_{\pi(k)}(\theta)|^q. \tag{10}$$

The inner problem is a convex optimization problem as long as Ω is convex (formulated as an LP problem for $q = 1$ or as a quadratic programming (QP) problem for $q = 2$), with a unique solution. However, $M = K!$ such problems have to be solved. The implication of this result is that problem (9) will, in general, have $M = K!$ different locally optimal solutions. We recognize that the objective function in (9) as concave, hence, we have an instance of concave programming problem, which is NP-hard [28]. Finding the globally optimal solution numerically is feasible only for small K.

Heuristics and branch-and-bound methods can be used to solve (9) (or, alternatively, a combination of heuristics with nonsmooth optimization methods when solving (3) directly).

Since various methods of robust regression arise as special cases of (3) with the increasing weighting vector, we will have a look at various algorithms for LTS and LTA problems, which can be applied for OWA-based regression with these weighting vectors, or perhaps modified for more general OWA.

1) Subset Selection Methods and Heuristics: The methods of robust regression are generally based on the subset selection problem. The direct approach is to try out all possible subsets of "good" data of size h, solve the LS or LAD regression problem for these data, and compare the values of the objective function. The MVELMS code is presented in [29]. Of course, the number of LS or LAD problems that need to be solved is $\binom{h}{K}$. An exact branch-and-bound type method was proposed in [30], but of course, it is applicable only to small datasets (e.g., $n = 3$, $K < 200$, or $n = 5$, $K < 50$). The FAST-LTS method proposed in [7], is a heuristic, based on random start (a subset of "good" data H_1), and a "C-step" in which a better subset H_2 is constructed from the residuals of all data with respect to the best LS fit to H_1. The C-step is relatively cheap $O(K)$, and is repeated until the convergence. The FAST-LTS method allows one to handle tens of thousands of data, but it may not produce the global minimum of (3) in a fixed number of random starts. Rousseeuw and Driessen [7] also proposed an alternative selection of H_1 based on PROGRESS algorithm in [3].

2) Methods Based on Elemental Sets: When $p = 1$ (the LTA method), there is another alternative. Since LAD regression corresponds to an exact fit to some subset of data of size p, instead of trying all subsets of h out of K "good" data, one can fit hyperplanes to all "elemental" subsets of p out of K, and simply calculate and sort the corresponding residuals, and then, calculate the value of the objective F. Of course the complexity (the number of elemental datasets) is reduced to $\binom{p}{K}$. This is the idea explored in [10].

A similar approach can be taken to the LMS regression problem, another instance of OWA-based regression, where OWA=median, see Section III-C. The LMS fit is a Chebyshev fit to a suitably chosen subset of size $p + 1$, and the number of such subsets is $\binom{p+1}{K}$.

In this context, let us outline an algorithm for a general increasing vector w in (3), based on the same idea (the case $q = 1$ and $\Omega = \Re^p$). Whether or not w has any zero entries, the LAD fit corresponds to an exact fit to a subset of p data (elemental set). Then, we follow the steps.

1) Generate every elemental set.
2) For each elemental set, compute the exactly fitting regression function, and get the residuals on all data.
3) Calculate the objective $F(\theta)$ in (3).
4) Choose θ, which minimizes $F(\theta)$.

C. Unimodal Weighting Vector w

The concept of centered OWA operators was proposed by Yager in [31], and later, also investigated in [32]. Here, the weights are symmetric ($w_j = w_{K+1-j}$), strongly decaying ($w_i < w_j$, if either $i < j \leq (K+1)/2$ or $i > j \geq (K+1)/2$), and inclusive ($w_j > 0$). We can relax the second and the third conditions to get noninclusive, soft-decaying centered OWA. A prototypical centered OWA operator is the median, $w = (0, \ldots, 0, 1, 0 \ldots, 0)$, with $w_h = 1$ and $h = [(K+1)/2]$.

Here, we recall the method of LMS [2], one of the high breakdown methods mentioned earlier. To get the highest breakdown point of $\varepsilon^* = 0.5$, in LMS one chooses, in fact $h = [(K + p + 1)/2]$. To apply the concept of centered OWA, we therefore, relax the symmetry condition, which effectively leaves us with the following class of unimodal weighting vectors, satisfying the conditions.

1) The maximal weight w_h is achieved at some $1 < h < K$.
2) The weights are soft-decaying $w_i \leq w_j$, if $i < j \leq h$ or $i > j \geq h$.

Thus, the middle-sized absolute or squared residuals are penalized most in (3), with the largest an the smallest residuals having limited, if any, contribution. The LMS method illustrates this well.

Here, we shall make an interesting observation. Because the LTA method (when the weighting vector is increasing) corresponds to an exact fit of an elemental subset of size p, effectively the smallest p residuals make no contribution to the objective in (3). Then, we can replace the increasing weighting vector with the unimodal vector by zeroing the last p components (or replacing them with arbitrary numbers), and obtain the same optimal solution.

Another way of defining the unimodal weighting vector is the following. Let $w_1 = \cdots = w_h = 0$, $w_{h+1} = 1$, and weights $w_j, j > h + 1$ soft-decaying to 0. The objective will discard the h largest residuals as outliers, and at the same time will not penalize small residuals.

The solution methods for this class of OWA-based objective are the same as those in the case of increasing weights.

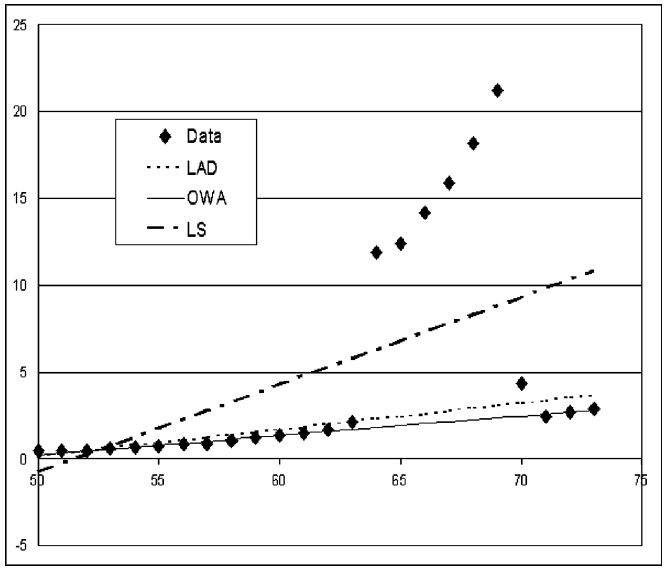

Fig. 1. Telephone data from [3], and the regression lines obtained by LS, LAD, and OWA-based regression. All outliers are vertical outliers. We see that the LS regression line is affected significantly, while the outliers have little effect on LAD. However, OWA regression is not affected at all.

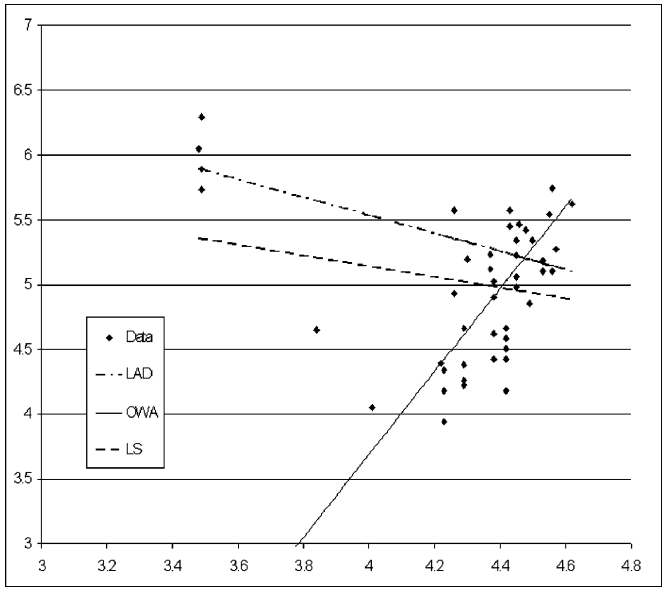

Fig. 2. Stars data, and the regression lines obtained by LS, LAD, and OWA-based regression. Both LS and LAD regression lines are severely affected by the four leverage points (red giants). OWA-based regression is not affected, and the outliers are identified by large residuals.

IV. Numerical Experiments

The purpose of this section is to illustrate the usefulness of OWA-based regression in the problem of outlier detection, and show feasibility of the direct approach for solving (3).

A. Description of the Datasets

In our numerical experiments, we concentrated on fitting difficult datasets containing outliers, such as real data: Modified wood gravity data, Hertzsprung–Russell stars data, and telephone data all discussed in [3], and data artificially generated to test robust regression algorithms.

1) Telephone data relate the number of telephone calls in Belgium to the variable *year*, for 24 years ($n = 1, K = 24$). Cases 15–20 are unusually high with cases 14 and 21 marginal (see Fig. 1).
2) Hertzsprung–Russell stars data (Stars) contain 47 measurements of the logarithm of effective temperature of the star and the logarithm of the light intensity $n = 1, K = 47$. The four red giants (cases 11, 20, 30, and 34) are clear outliers and leverage points.
3) Modified wood gravity data (Wood) $n = 5, K = 20$ is based on real data, but modified in [2] to contain outliers at cases 4, 6, 8, and 19.
4) Hawkins, Bradu, and Kass (HBK) artificial dataset [33] $n = 3, K = 75$, outliers are cases 1–10.
5) Hadi and Simonoff (H–S) artificial dataset [34] $n = 2, K = 25$, with three outliers (1, 2, 3). The data were generated randomly, with the dependent variable consistent with the model $y = x_1 + x_2 + \varepsilon$ with $\varepsilon_i \sim N(0, 1)$ ($N(a, b)$ stands for the normal distribution with the mean a and standard deviation b).
6) Artificial data generated following [7] (R–D), namely, $n = 1, K = 1000$, the data generated using $y = x_1 + 1 + \varepsilon$ for the first 800 observations, $x_i \sim N(0, 100)$ and $\varepsilon_i \sim N(0, 1)$, and for the remaining 200 observations (x_i, y_i) were drawn from the bivariate normal distribution with $\mu = (50, 0)$ and $\Sigma = 25I$ (see Fig. 2).

B. Results and Discussion

We used increasing OWA weighting vectors given by (5) with parameters $a = 0.2$ and $b = 0.5$. As the method of solution to (3), we minimized $F(\theta)$ directly, with $q = 1$, using the derivative free bundle method (DFBM) from [22], [25], implemented in GANSO library [26] available from http://www.ganso.com.au.

We make three observations here. First, the objective F in (3) has many local minima, we already mentioned that the optimization problem with such an objective is NP-hard. Therefore, we need to use a global search strategy. In this study, we used random start heuristic (i.e., starting DFBM from 1000 randomly chosen starting points), as well as used the solution to LAD problem as the starting point for DFBM. We used the Sobol quasi-random number generator [35] with the purpose of obtaining starting points.

The second observation is that the objective F in (3) is not smooth, but is Lipschitz-continuous, and hence, DFBM is applicable. DFBM [22], [25] uses a descent strategy similar to quasi-Newton type methods, but calculates the direction of descent differently, by using an approximation to Clarke-subdifferential. This method is guaranteed to converge to a local minimum of the objective, and has been shown to escape shallow local minima in multiextremal problems.

The third observation is about the impact of the continuous reordering of the residuals in OWA-based regression. We note

TABLE I
RESULTS OF NUMERICAL EXPERIMENTS

Data	K	p	outliers	computing time (s)	starting point
Telephone	24	2	14-21	0.1	LAD
Stars	47	2	11,20,30,34	3	random start (1000)
Wood	20	6	4,6,8,19	15	random start (1000)
HBK	75	4	1-10	8	random start (1000)
H-S	25	3	1-3	5	random start (1000)
R-D	1000	2	801-1000	0.2	LAD

Right column indicates whether the solution to LAD was used as a starting point for DFBM, or DFBM was combined with the the random start heuristic, in the latter case, the number of starting points is in the brackets.

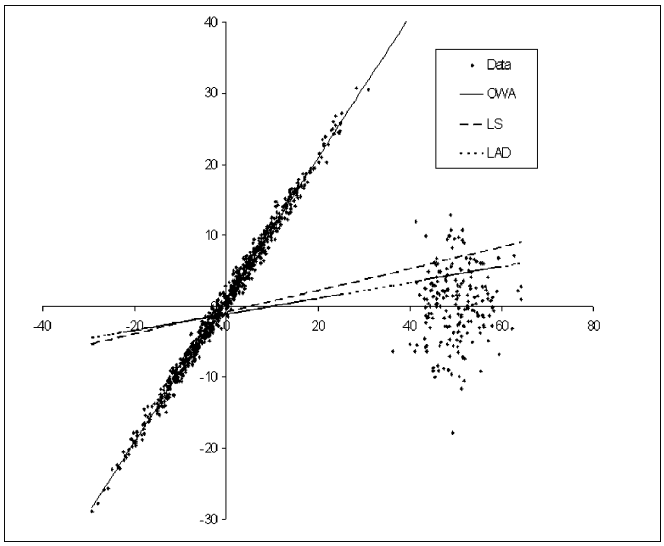

Fig. 3. Artificial data R–D from [7], and the regression lines obtained by LS, LAD, and OWA-based regression. LS and LAD are severely affected by outliers, but OWA-based regression correctly discards the outliers.

that the computational complexity of calculating a single value of F is the same of the sorting methods $O(K \log K)$. This bound on the computational cost is very attractive, and the sort operation has a limited impact on the total cost of OWA-based regression, which is mainly due to multiple local optima of F. Furthermore, it is possible to parallelize both the calculation of the residuals and the sort operation on modern general purpose graphic processing units (GPUs) [36], i.e., to use the workstation's graphics card to off-load computation of F, which makes direct minimization of F an attractive strategy, even for very large datasets ($K > 10^6$).

Table I presents the results of our experiments. Our solution method has correctly identified the outliers in all cases. Small running times of the algorithm illustrate its efficiency, even for large datasets. Computations were performed on a workstation with Pentium 2.3 GHz processor and 2 GB RAM.

Figs. 1–3 illustrate OWA-based regression, and compares it to LS and LAD regression. We can see that identification and elimination of outliers is quite effective with OWA-based regression. Note that several existing methods of robust regression, namely, the LTS and LTA (which minimize special instances of OWA functions) are also effective in eliminating the outliers in these cases, as reported in [2], [7], [34]. We shall make two points here. First, when we use gradually increasing weights on OWA-based regression, as opposed to 0 and 1 weights in LTS and LTA, we still obtain correct regression lines. It seems that the location of the global minimum of F in (3) is not drastically affected by the choice of weights. Second, we applied a very different approach to solving (3), applicable to OWA with any weighting vectors, and obtained the solution as efficiently, if not more efficiently, as the alternative methods used in LTS and LTA. This indicates the potential of the proposed method.

V. CONCLUSION

We introduced OWA-based regression as an alternative to the ordinary LS, LAD, and M-estimators. Depending on the OWA weighting vector, we obtain the classical instances of the regression problem, as well as various high-breakdown robust methods, such as the LMS, LTS, quantile regression, and least trimmed likelihood methods. We discussed various alternative methods of numerical solution of the regression problem. These methods vary depending on the OWA weighting vector. In particular, for the most interesting cases of increasing and unimodal weighting vectors, the regression problem becomes nonconvex and NP-hard (which is true for all high breakdown methods). We have used a method of nonsmooth global optimization to minimize the total fitness function in our numerical experiments. In all cases, we were able to identify correctly the outliers in the data, consistently with the latest methods of robust regression.

What makes OWA-based regression advantageous is that 1) it provides a generic problem formulation in which the existing classical and robust methods are special cases; and 2) it allows one to use an alternative method of numerical solution, with less rugged objective. We see the potential of this method in identifying outliers in large datasets, as the complexity of the method is not exponential in K.

REFERENCES

[1] R. Yager, "On ordered weighted averaging aggregation operators in multicriteria decision making," *IEEE Trans. Syst., Man Cybern.*, vol. 18, no. 1, pp. 183–190, Jan./Feb. 1988.
[2] P. Rousseeuw, "Least median of squares regression," *J. Amer. Statist. Assoc*, vol. 79, pp. 871–880, 1984.
[3] P. Rousseeuw and A. Leroy, *Robust Regression and Outlier Detection*. New York: Wiley, 2003.
[4] A. Hadi and A. Luceño, "Maximum trimmed likelihood estimators: A unified approach, examples, and algorithms," *Comput. Statist. Data Anal.*, vol. 25, pp. 251–272, 1997.
[5] F. Hampel, "A general qualitative definition of robustness," *Ann. Math. Statist.*, vol. 42, pp. 1887–1896, 1971.
[6] P. Huber, *Robust Statistics*. New York: Wiley, 2003.
[7] P. Rousseeuw and K. Van Driessen, "Computing lts regression for large data sets," *Data Mining Knowl. Discov.*, vol. 12, pp. 29–45, 2006.
[8] P. Rousseeuw and C. Croux, "Alternatives to the median absolute deviation," *J. Amer. Statist. Assoc*, vol. 88, pp. 1273–1283, 1993.
[9] A. Stromberg, O. Hossjer, and D. M. Hawkins, "The least trimmed differences regression estimator and alternatives," *J. Amer. Statist. Assoc*, vol. 95, pp. 853–864, 2000.
[10] D. M. Hawkins and D. J. Olive, "Applications and algorithms for least trimmed sum of absolute deviations regression," *Comput. Statist. Data Anal.*, vol. 32, pp. 119–134, 1999.
[11] T. Bernholt. (2005). "Robust estimators are hard to compute," Univ. Dortmund, Dortmund, Germany, Tech. Rep., [Online]. Available: http://www.statistik.tu-dortmund.de/fileadmin/user_upload/Lehrstuehle/MSind /SFB_475/2005/tr52-05.pdf
[12] D. Olive and D. M. Hawkins, "Robust regression with high coverage," *Statist. Probability Lett.*, vol. 63, pp. 259–266, 2003.

[13] G. Beliakov, A. Pradera, and T. Calvo, *Aggregation Functions: A Guide for Practitioners*. Heidelberg, Berlin, New York: Springer-Verlag, 2007.
[14] R. Yager and J. Kacprzyk, Eds., *The Ordered Weighted Averaging Operators. Theory and Applications*. Boston, MA: Kluwer, 1997.
[15] R. Yager, "Quantifier guided aggregation using OWA operators," *Int. J. Intell. Syst.*, vol. 11, pp. 49–73, 1996.
[16] M. Grabisch, J.-L. Marichal, R. Mesiar, and E. Pap, *Aggregation Functions*. Cambridge, U.K.: Cambridge Univ. Press, 2009.
[17] Y. Torra and V. Narukawa, *Modeling Decisions. Information Fusion and Aggregation Operators*. Berlin, Heidelberg, Germany: Springer-Verlag, 2007.
[18] C. H. Müller, "Breakdown points for designed experiments," *J. Statist. Plann. Inference*, vol. 45, no. 3, pp. 413–427, 1995, 0378-3758, DOI:10.1016/0378-3758(94)00086-B.
[19] R. Yager, "Using stress functions to obtain OWA operators," *IEEE Trans. Fuzzy Syst.*, vol. 15, no. 6, pp. 1122–1129, Dec. 2007.
[20] R. Yager, "Connectives and quantifiers in fuzzy sets," *Fuzzy Sets Syst.*, vol. 40, pp. 39–76, 1991.
[21] X. W. Liu and H. W. Lou, "On the equivalence of some approaches to the OWA operator and RIM quantifier determination," *Fuzzy Sets Syst.*, vol. 159, no. 13, pp. 1673–1688, 2008.
[22] A. Bagirov, "A method for minimization of quasidifferentiable functions," *Optim. Methods Softw.*, vol. 17, pp. 31–60, 2002.
[23] C. Lemarechal, "Bundle methods in non-smooth optimization," in *Nonsmooth Optimization*, C. Lemarechal and R. Mifflin, Eds. Oxford, U.K.: Pergamon, 1978, pp. 78–102.
[24] A. M. Bagirov, "Numerical methods for minimizing quasi-differentiable functions: A survey and comparison," in *Quasidifferentiability and Related Topics*, V. Demyanov and A. Rubinov, Eds., Dordrecht, The Netherlands/Boston, MA: Kluwer, 2000, pp. 33–71.
[25] A. Bagirov, "Derivative-free methods for non-smooth optimization," in *Encyclopedia of Optimization*, Heidelberg, Berlin, Germany, New York: Springer-Verlag, 2009.
[26] G. Beliakov and J. Ugon, "Implementation of novel methods of global and non-smooth optimization: GANSO programming library," *Optimization*, vol. 56, pp. 543–546, 2007.
[27] R. Yager, "Constrained OWA aggregation," *Fuzzy Sets Syst.*, vol. 81, pp. 89–101, 1996.
[28] R. Horst and P. M. Pardalos, *Handbook of Global Optimization*. Dordrecht, The Netherlands/Boston, MA: Kluwer, 1995.
[29] D. M. Hawkins and J. Simonoff, "High breakdown regression and multivariate estimation," *Appl. Statist.*, vol. 42, pp. 423–432, 1993.
[30] J. Agulló, "New algorithms for computing the least trimmed squares regression estimator," *Comput. Statist. Data Anal.*, vol. 36, no. 4, pp. 425–439, 2001.
[31] R. Yager, "Centered OWA operators," *Soft Comput.*, vol. 11, pp. 631–639, 2007.
[32] M. Zarghami, F. Szidarovszky, and R. Ardakanian, "Sensitivity analysis of the OWA operator," *IEEE Trans. Syst. Man Cybern. B-Cybern.*, vol. 38, no. 2, pp. 547–552, Apr. 2008.
[33] D. M. Hawkins, D. Bradu, and V. Kass, "Location of several outliers in multiple-regression data using elemental sets," *Technometrics*, vol. 26, pp. 197–208, 1984.
[34] A. Hadi and J. Simonoff, "Procedures for the identification of multiple outliers in linear models," *J. Amer. Statist. Assoc*, vol. 88, pp. 1264–1272, 1993.
[35] I. Sobol, "The production of points uniformly distributed in a multidimensional cube," *USSR Comput. Math. Math. Phys.*, vol. 16, pp. 236–242, 1977.
[36] E. Sintorn and U. Assarson, "Fast parallel GPU-sorting using a hybrid algorithm," *J. Parallel Distrib. Comput.*, vol. 68, pp. 1381–1388, 2008.

Ronald R. Yager (S'66–M'68–SM'93–F'97) received the B.E.E. degree from the City College of New York, New York, and the Ph.D. degree from the Polytechnic University of New York, New York.

He was a National Aeronautics and Space Administration (NASA)/Stanford Visiting Fellow and a Research Associate at the University of California, Berkeley. He has been a Lecturer at North Atlantic Treaty Organization (NATO) Advanced Study Institutes. He is currently the Director of the Machine Intelligence Institute, Iona College, New Rochelle, NY, where he is also a Professor of information and decision technologies. He is the Editor-in-Chief of the *International Journal of Intelligent Systems*. He serves on the Editorial Boards of a number of journals including *Neural Networks*, *Data Mining and Knowledge Discovery*, *Fuzzy Sets and Systems*, *Journal of Approximate Reasoning*, and *International Journal of General Systems*. He has been engaged in the area of fuzzy sets and related disciplines of computational intelligence for over 25 years. In addition to his pioneering work in the area of fuzzy logic, he has made fundamental contributions in decision making under uncertainty and the fusion of information. He has authored or coauthored over 500 papers published and 15 books.

Prof. Yager was the recipient of the IEEE Computational Intelligence Society Pioneer Award in Fuzzy Systems. He is a Fellow of the New York Academy of Sciences and the Fuzzy Systems Association. He was given an award by the Polish Academy of Sciences for his contributions. He was the Program Director in the Information Sciences program at the National Science Foundation. He is a member of the Editorial Boards of a number of journals, including the IEEE TRANSACTIONS ON FUZZY SYSTEMS and IEEE INTELLIGENT SYSTEMS.

Gleb Beliakov (SM'08) received the Ph.D. degree in physics and mathematics from the Russian Peoples Friendship University, Moscow, Russia, in 1992.

He was a Lecturer and Research Fellow with Los Andes University, the Universities of Melbourne and South Australia, and Deakin University, Melbourne, Australia. He is currently an Associate Professor with the School of IT, Deakin University. His research interests are in the areas of aggregation operators, multivariate constrained approximation, global optimization, decision support systems, and applications of fuzzy systems in healthcare. He is the author of 90 research papers in the mentioned areas and a number of software packages.