EURASIP Journal on Applied Signal Processing 2004:12, 1762–1769

A Maximum Likelihood Approach to Least Absolute Deviation Regression

Yinbo Li

Department of Electrical and Computer Engineering, University of Delaware, Newark, DE 19716 3130, USA
Email: yli@eecis.udel.edu

Gonzalo R. Arce

Department of Electrical and Computer Engineering, University of Delaware, Newark, DE 19716 3130, USA
Email: arce@eecis.udel.edu

Received 7 October 2003; Revised 22 December 2003

Least absolute deviation (LAD) regression is an important tool used in numerous applications throughout science and engineering, mainly due to the intrinsic robust characteristics of LAD. In this paper, we show that the optimization needed to solve the LAD regression problem can be viewed as a sequence of maximum likelihood estimates (MLE) of location. The derived algorithm reduces to an iterative procedure where a simple coordinate transformation is applied during each iteration to direct the optimization procedure along edge lines of the cost surface, followed by an MLE of location which is executed by a weighted median operation. Requiring weighted medians only, the new algorithm can be easily modularized for hardware implementation, as opposed to most of the other existing LAD methods which require complicated operations such as matrix entry manipulations. One exception is Wesolowsky's direct descent algorithm, which among the top algorithms is also based on weighted median operations. Simulation shows that the new algorithm is superior in speed to Wesolowsky's algorithm, which is simple in structure as well. The new algorithm provides a better tradeoff solution between convergence speed and implementation complexity.

Keywords and phrases: least absolute deviation, linear regression, maximum likelihood estimation, weighted median filters.

1. INTRODUCTION

Linear regression has long been dominated by least squares (LS) techniques, mostly due to their elegant theoretical foundation and ease of implementation. The assumption in this method is that the model has normally distributed errors. In many applications, however, heavier-than-Gaussian tailed distributions may be encountered, where outliers in the measurements may easily ruin the estimates [1]. To address this problem, robust regression methods have been developed so as to mitigate the influence of outliers. Among all the approaches to robust regression, the least absolute deviations (LADs) method, or L_1-norm, is considered conceptually the simplest one since it does not require a "tuning" mechanism like most of other robust regression procedures. As a result, LAD regression has drawn significant attentions in statistics, finance, engineering, and other applied sciences as detailed in a series of studies on L_1-norm methods [2, 3, 4, 5]. LAD regression is based on the assumption that the model has Laplacian distributed errors. Unlike the LS approach though, LAD regression has no closed-form solution, hence numerical and iterative algorithms must be resorted to.

Surprisingly to many, the LAD regression method first suggested by Boscovich (1757) and studied by Laplace (1793) predated the LS technique originally developed by Legendre (1805) and Gauss (1823) [1, 2]. It was not until nearly a century later that Edgeworth [6] proposed a general numerical method to solve the unconstrained LAD problem, where the *weighted median* was introduced as the basic operation in each iteration. Edgeworth's method, however, suffers from cycling when data has degeneracies [7]. A breakthrough came in the 1950's when Harris [8] brought in the notion that linear programming techniques could be used to solve the LAD regression, and Charnes et al. [9] actually utilized the simplex method to minimize the LAD objective function. Many simplex-like methods blossomed thereafter, among which Barrodale and Roberts [10] and Armstrong et al. [11] are the most representative ones. Other efficient approaches include the active set method by Bloomfield and Steiger [12], the direct decent algorithm by Wesolowsky [13], and the interior point method proposed by Zhang [14]. More historical background about LAD estimate can be found in [2].

The simple LAD regression problem is formulated as follows. Consider N observation pairs (X_i, Y_i) modelled in a linear fashion

$$Y_i = aX_i + b + U_i, \quad i = 1, 2, \ldots, N, \tag{1}$$

where a is the unknown slope of the fitting line, b the intercept, and U_i are unobservable errors drawn from a random variable U obeying a zero-mean Laplacian distribution $f(U) = (1/2\lambda)e^{-|U|/\lambda}$ with variance $\sigma^2 = 2\lambda^2$. The LAD regression is found by choosing a pair of parameters a and b that minimizes the objective function

$$F(a, b) = \sum_{i=1}^{N} |\, Y_i - aX_i - b \,|, \tag{2}$$

which has long been known to be continuous and convex [1]. Moreover, the cost surface is of a polyhedron shape, and its edge lines are characterized by the sample pairs (X_i, Y_i).

Notably, the minimization of the LAD cost function (2) is closely related to the location estimation problem defined as follows. Let the random variable V be defined as $V = U + \mu$, where μ is an unknown constant location and U obeys the Laplacian distribution. The maximum likelihood estimate (MLE) of location on the sample set $\{V_i|_{i=1}^N\}$ is

$$\mu^* = \arg\min_{\mu} \sum_{i=1}^{N} |\, V_i - \mu \,|. \tag{3}$$

The solution to the above minimization problem is well known to be the sample *Median*

$$\mu^* = \text{MED}\left(V_i|_{i=1}^N\right). \tag{4}$$

The striking similarity between (2) and (3) infers that, for a fixed $a = a_0$, the minimizer of (2), say $b^*_{a_0}$, is essentially an MLE for location under the Laplacian assumption. For reasons that will be explained shortly in Section 2, the minimizer of (2) $a^*_{b_0}$, given $b = b_0$, is also an MLE for location under the Laplacian assumption with certain extensions. Thus, a very intuitive way of solving the LAD regression problem can be constructed as a "seesaw" procedure: first, hold one of the parameters a or b constant, optimize the other using the MLE concept, then alternate the role of the parameters, and repeat this process until both parameters converge. It will soon be shown in the paper that this method suffers from some intrinsic limitations that often leads to nonglobal optimal solutions despite its attractive simplicity. However, further inspection on this initial algorithm reveals that, with some specific guidance on how to do the MLE optimization and one simple coordinate transformation, a similar but more accurate algorithm can be formulated where the global optimum can be reached. In fact, in this paper, we derive a fast iterative solution where the concept of ML is applied jointly with coordinate transformations. It is also shown that the proposed method is comparable with the best algorithms used to date in terms of computational complexity, and has a greater potential to be implemented in hardware.

2. ALGORITHM DERIVATION

2.1. Basic understanding

Consider the linear regression model in (1). If the value of a is fixed at first, say $a = a_0$, the objective function (2) now becomes a one-parameter function of b:

$$F(b) = \sum_{i=1}^{N} |\, Y_i - a_0X_i - b \,|. \tag{5}$$

Assuming a Laplace distribution for the errors U_i, the above cost function reduces to an ML estimator of location for b. That is, we observe the sequence of random samples $\{Y_i - a_0X_i\}$, and the goal is to estimate the fixed but unknown location parameter b. Thus according to (4), the parameter b^* in this case can be obtained by

$$b^* = \text{MED}\left(Y_i - a_0X_i|_{i=1}^N\right). \tag{6}$$

If, on the other hand, we fix $b = b_0$, the objective function reduces to

$$\begin{aligned} F(a) &= \sum_{i=1}^{N} |\, Y_i - b_0 - aX_i \,| \\ &= \sum_{i=1}^{N} |\, X_i \,| \left| \frac{Y_i - b_0}{X_i} - a \right|. \end{aligned} \tag{7}$$

Again, if the error random variable U_i obeys a Laplacian distribution, the observed samples $\{(Y_i - b_0)/X_i\}$ are also Laplacian distributed, but with the difference that each sample in this set has different variance. The reason is obvious since for each known X_i and zero-mean U_i, U_i/X_i remains a zero-mean Laplacian with variance scaled by $1/X_i^2$. Thus the parameter a^* minimizing the cost function (7) can still be seen as the ML estimator of location for a, and can be calculated out as the *weighted median*

$$a^* = \text{MED}\left(|\, X_i \,| \diamond \frac{Y_i - b_0}{X_i} \Big|_{i=1}^N \right), \tag{8}$$

where $\diamond$ is the replication operator. For a positive integer $|X_i|$, $|X_i| \diamond Y_i$ means Y_i is replicated $|X_i|$ times. When the weights $|X_i|$ are not integers, the computation of the weighted median is outlined in the appendix.

A simple and intuitive approach to the LAD regression problem is through the following iterative algorithm.

(1) Set $k = 0$. Find an initial value a_0 for a, such as the LS solution.

(2) Set $k = k + 1$ and obtain a new estimate of b for a fixed a_{k-1} using

$$b_k = \text{MED}\left(Y_i - a_{k-1}X_i|_{i=1}^N\right). \tag{9}$$

(3) Obtain a new estimate of a for a fixed b_k using

$$a_k = \text{MED}\left(|\, X_i \,| \diamond \frac{Y_i - b_k}{X_i} \Big|_{i=1}^N \right). \tag{10}$$

(4) Once a_k and b_k do not deviate from a_{k-1} and b_{k-1} within a tolerance range, end the iteration. Otherwise, go back to step (2).

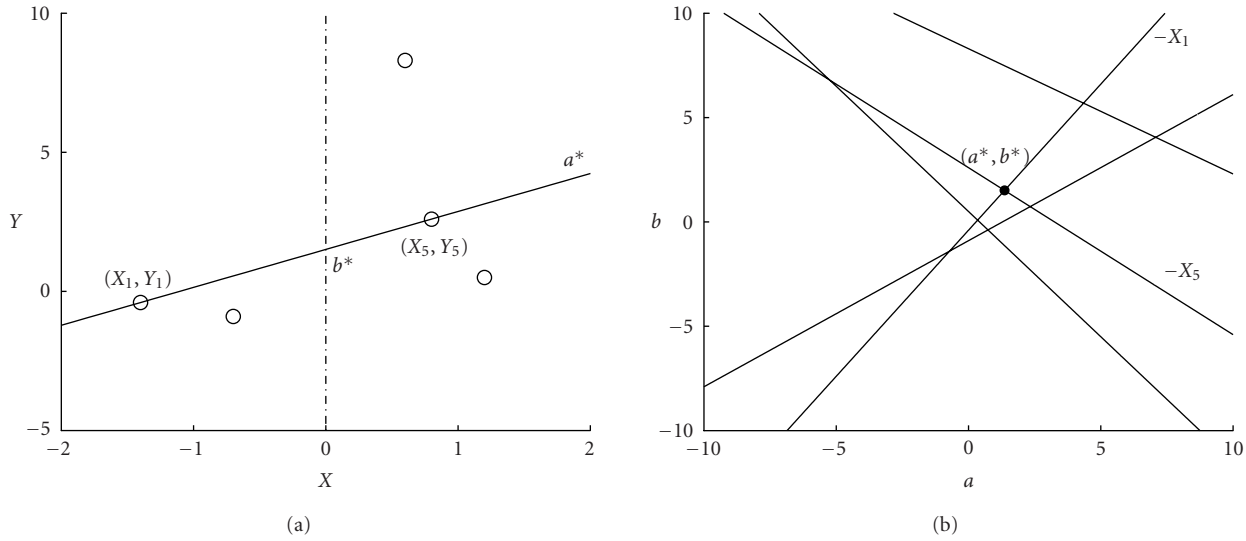

FIGURE 1: Illustration of (a) the sample space and (b) the parameter space in the simple linear regression problem. The circles in (a) represent the samples; the dot in (b) represents the global minimum.

Since the median and weighted median operations are both ML location estimators under the least absolute criterion, the cost functions will be nonincreasing throughout the iterative procedure, that is,

$$F(a_{k-1}, b_{k-1}) \geq F(a_{k-1}, b_k) \geq F(a_k, b_k). \tag{11}$$

The algorithm then converges iteratively. Since the objective function $F(a, b)$ is continuous and convex, one may readily conclude that the algorithm converges to the global minimum. However, careful inspection reveals that there are cases where the algorithm does not reach the global minimum. To see this, it is important to describe the relationship between the sample space and the parameter space.

As shown in Figure 1, the two spaces are dual to each other. In the sample space (Figure 1a), each sample pair (X_i, Y_i) represents a point on the plane. The solution to the problem (1), namely (a^*, b^*), is represented as a line with slope a^* and intercept b^*. If this line goes through some sample pair (X_i, Y_i), then the equation $Y_i = a^* X_i + b^*$ is satisfied. On the other hand, in the parameter space (Figure 1b), (a^*, b^*) is a point on the plane, and $(-X_i, Y_i)$ represents a line with slope $(-X_i)$ and intercept Y_i. When $b^* = (-X_i)a^* + Y_i$ holds, it can be inferred that the point (a^*, b^*) is on the line defined by $(-X_i, Y_i)$. As can be seen in Figure 1, the line going through (X_1, Y_1) and (X_5, Y_5) in the sample space has a slope a^* and an intercept b^*, but in the parameter space, it is represented as a point which is the intersection of two lines with slopes $(-X_1)$ and $(-X_5)$, respectively. The sample set used to generate Figure 1 is, in a (X_i, Y_i) manner, $[(-1.4, -0.4), (0.6, 8.3), (1.2, 0.5), (-0.7, -0.9), (0.8, 2.6)]$.

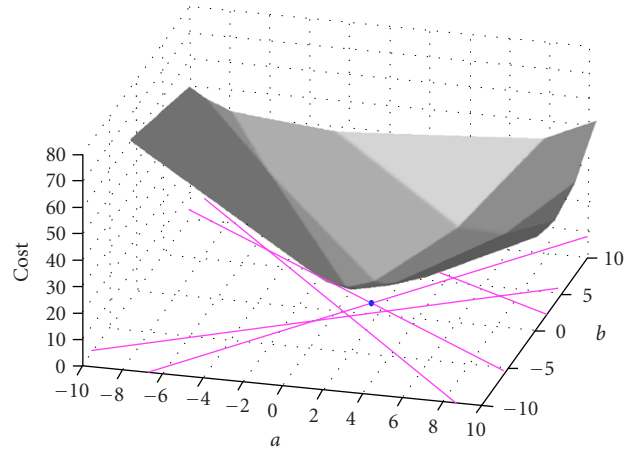

FIGURE 2: The cost surface of the LAD regression problem. The dot at an intersection on the a-b plane represents the global minimum. To better illustrate the inner topology of the function, the half surface that is towards the viewers is cut off.

The structure of the objective function $F(a, b)$ is well defined as a polyhedron sitting on top of the a-b plane, as seen in Figure 2. The projections of the polyhedron edges onto the plane are exactly the lines defined by sample pairs (X_i, Y_i), which is why the term "edge line" is used. In other words, every sample pair (X_i, Y_i) has a corresponding edge line in the parameter space. Moreover, the projections of the polyhedron corners are those locations on the a-b plane, where two or more of the edge lines intersect. Most importantly, the minimum of this convex, linearly-segmented error surface occurs at one of these corners.

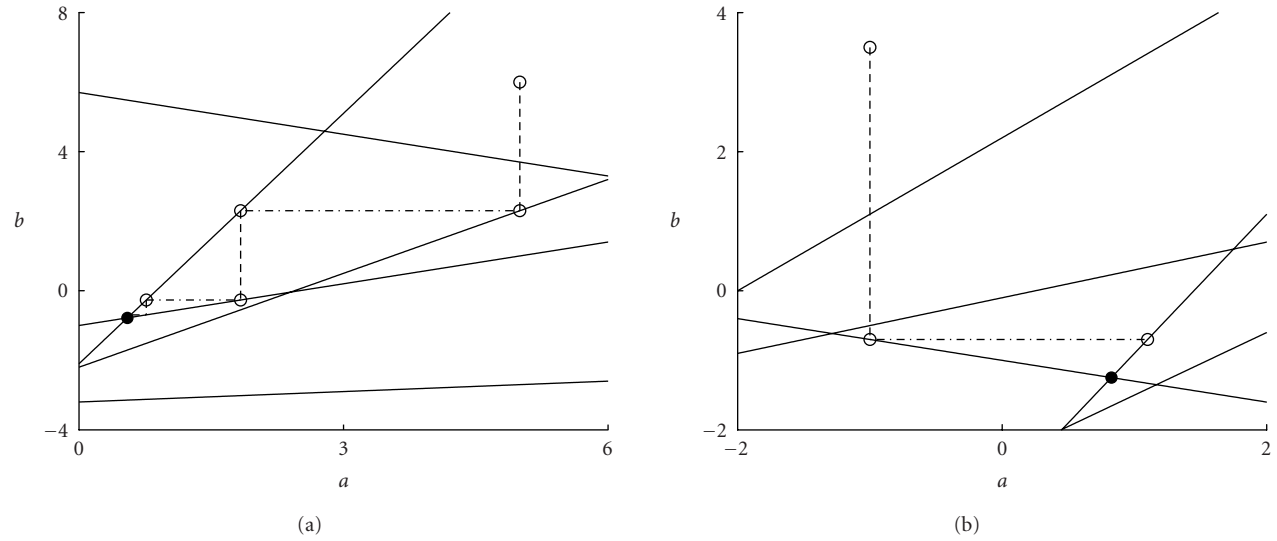

Figure 3: The parameters' trajectories during the iterations. Vertical dashed lines represent b updates, while horizontal dotted lines represent a updates; (a) zigzag case, (b) nonoptimal case. The marked dots represent the global minima. To better illustrate, the initial values for a and b are not set from the LS solution.

To describe the dynamics of this simple iterative method, consider Step (2) in the procedure, where a new estimate b_k is calculated based on a fixed, previously obtained a_{k-1} through a median operation. Since the median is of selection type, its output is always one of the inputs. Without loss of generality, assume $b_k = Y_j - a_{k-1}X_j$, which means that the newly estimated parameter pair (a_{k-1}, b_k) is on the edge line defined by $(-X_j)$ and Y_j. Thus, the geometrical interpretation of Step (2) can be derived as follows: draw a vertical line at $a = a_{k-1}$ in the parameter space and mark all the intersections of this line with N edge lines.[1] The intersection on the edge line defined by $(-X_j)$ and Y_j is vertically the median of all; thus its b-coordinate value is accepted as b_k, the new update for b. Similar interpretation can be made for Step (3), except that the chosen intersection is a weighted median output, and there may be some edge lines parallel to the a-axis.

The drawback of this algorithm is that the convergence dynamics depends on the geometry of the edge lines in the parameter space. As can be seen in Figure 3a, the iteration is carried on between edge lines in an inefficient zigzag manner, needing infinite steps to converge to the global minimum. Moreover, as illustrated in Figure 3b, it is possible that vertical optimization and horizontal optimization on the edge lines can both give the same results in each iteration. Thus the algorithm gets stuck in a nonoptimal solution. The sample set used for Figure 3a is $[(-0.1, -3.2), (-0.9, -2.2), (0.4, 5.7), (-2.4, -2.1), (-0.4, -1.0)]$, and the initial values for a and b are 5 and 6. The sample set used for Figure 3b is $[(0.3, -1.0), (-0.4, -0.1), (-2.0, -2.9), (-0.9, -2.4), (-1.1, 2.2)]$, and the initial values for a and b are -1 and 3.5.

2.2. New algorithm

To overcome these limitations, the iterative algorithm must be modified exploiting the fact that the optimal solution is at an intersection of edge lines. Thus, if the search is directed along the edge lines, then a more accurate and more efficient algorithm can be formulated. The approach proposed in this paper is through coordinates transformation. The basic idea is as follows. In the parameter space, if the coordinates are transformed so that the edge line containing the previous estimate (a_{k-1}, b_{k-1}) is parallel to the a'-axis at height b'_{k-1}, then the horizontal optimization based upon b'_{k-1} is essentially an optimization along this edge line. The resultant (a'_k, b'_k) will be one of the intersections that this line has with all other edge lines, thus avoiding possible zigzag dynamics during the iterations. Transforming the obtained parameter pair back to the original coordinates results in (a_k, b_k). This is illustrated in Figure 4. The only requirement for this method is that the shape of the cost surface must be preserved upon transformation; thus the same optimization result can be achieved. Notice that, if an edge line is horizontal, its slope $(-X_j)$ has to be 0. We will show shortly that a simple shifting in the sample space can satisfy the requirement.

The following is the proposed algorithm for LAD regression.

(1) Set $k = 0$. Initialize b to be b_0 using the LS solution

$$b_0 = \frac{\sum_{i=1}^{N} (X_i - \bar{X})(\bar{Y}X_i - \bar{X}Y_i)}{\sum_{i=1}^{N} (X_i - \bar{X})^2}. \tag{12}$$

[1] Since all meaningful samples are finite, no edge lines will be parallel to the b-axis; hence there must be N intersections.

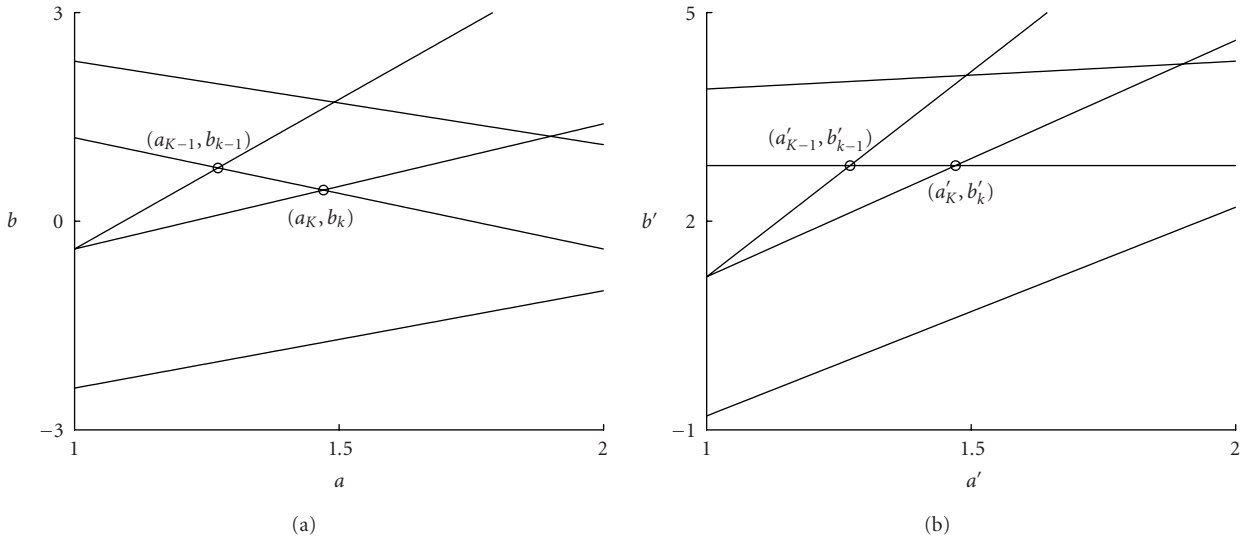

FIGURE 4: Illustration of one iteration. The previous estimate (a_{k-1}, b_{k-1}) is mapped into the transformed coordinates as (a'_{k-1}, b'_{k-1}); (a'_k, b'_k) is obtained through ML estimation in the transformed coordinates; the new estimate (a_k, b_k) is formed by mapping (a'_k, b'_k) back into the original coordinates. The sample set is $[(1.6, 2.8),\ (-1.4, -3.8),\ (1.2, 3.5),\ (-4.3, -4.7),\ (-1.8, -2.2)]$.

Calculate a_0 by a weighted median

$$a_0 = \text{MED}\left(|X_i| \diamond \frac{Y_i - b_0}{X_i} \Big|_{i=1}^{N} \right). \tag{13}$$

Keep the index j which satisfies $a_0 = (Y_j - b_0)/X_j$. In the parameter space, (a_0, b_0) is on the edge line with slope $(-X_j)$ and intercept Y_j.

(2) Set $k = k + 1$. In the sample space, right shift the coordinates by X_j so that the newly formed y'-axis goes through the original (X_j, Y_j). The transformations in the sample space are

$$X'_i = X_i - X_j, \qquad Y'_i = Y_i, \tag{14}$$

and the transformations in the parameter space are

$$a'_{k-1} = a_{k-1}, \qquad b'_k = b'_{k-1} = b_{k-1} + a_{k-1}X_j. \tag{15}$$

The shifted sample space (X', Y') corresponds to a new parameter space (a', b'), where $(-X'_j, Y'_j)$ represents a horizontal line.

(3) Perform a weighted median to get a new estimate of a':

$$a'_k = \text{MED}\left(|X'_i| \diamond \frac{Y'_i - b'_k}{X'_i} \Big|_{i=1}^{N} \right). \tag{16}$$

Keep the new index t which gives $a'_k = (Y'_t - b'_k)/X'_t$.

(4) Transform back to the original coordinates

$$a_k = a'_k, \qquad b_k = b'_k - a'_k X_j. \tag{17}$$

(5) Set $j = t$. If a_k is identical to a_{k-1} within the tolerance, end the program. Otherwise, go back to step (2).

It is simple to verify that the transformed cost function is the same as the original one using the relations in (14) and (15). For fixed b_k,

$$\begin{aligned} F'(a') &= \sum_{i=1}^{N} |\, Y'_i - a'X'_i - b'_k \,| \\ &= \sum_{i=1}^{N} |\, Y_i - a(X_i - X_j) - (aX_j + b_k) \,| \\ &= \sum_{i=1}^{N} |\, Y_i - aX_i - b_k \,| = F(a). \end{aligned} \tag{18}$$

This relationship guarantees that the new update in each iteration is correct.

3. SIMULATIONS

The major part of the computational power of the proposed algorithm is consumed in the weighted median operation at each iteration. Essentially, it is a sorting problem, which, for n samples, is in the order of $n \log n$. Fortunately, for this particular application, some speed-up can be achieved by not doing a full sorting every time. In [13], where the weighted median is also used as the kernel operation, a shortcut to circumvent this time-consuming full-sorting procedure is developed. The basic idea is the previous estimate can be considered close enough to the true value, thus "fine tuning" can be executed around this point by making use of the weighted median inequalities shown next in (21).

Consider a weighted median defined as follows:

$$a^* = \mathrm{MED}\left(W_i \diamond Z_i \,|_{i=1}^{n}\right) = \arg\min_{a} \sum_{i=1}^{N} W_i \,|\, Z_i - a \,|, \tag{19}$$

where the weights $W_i \geq 0$. If we order the samples Z_i as $Z_{(1)} \leq Z_{(2)} \leq \cdots \leq Z_{(N)}$, then the weight associated with the ith order statistic $Z_{(i)}$ is often referred to as the concomitant $W_{[i]}$ [15]. In this way, the weighted median a^* can always be identified as $Z_{(j)}$ whose index j satisfies the following inequalities:

$$\sum_{i=1}^{j-1} W_{[i]} < \sum_{i=j}^{N} W_{[i]}, \tag{20}$$

$$\sum_{i=1}^{j} W_{[i]} \geq \sum_{i=j+1}^{N} W_{[i]}. \tag{21}$$

Comparing to (16), we should notice that the weights W_i and samples Z_i in every LAD iteration are different. Suppose that the previous estimate a_{k-1}, which is also the output of a weighted median, corresponds to Z_j. We do not have to fully order all these samples, but classify them into two categories, the ones smaller than it and the ones larger. Check the inequalities to see if they still hold. If not, transfer the boundary sample and its weight into another group and recheck until the new weighted median output is found.

Two criteria are often used to compare LAD algorithms: speed of convergence and complexity. Most of the efficient algorithms, in terms of convergence speed (except for Wesolowsky's and its variations), are derived from linear programming (LP) perspectives, such as simplex and interior point. Take Barrodale and Roberts' algorithm[2] [10], for example; its basic idea is to apply row and column operations on a constructed $(N+K)\times(K+1)$ matrix $\mathbf{A}$. The initial value of $\mathbf{A}$ is

$$\mathbf{A} = \begin{bmatrix} \mathbf{X} & \mathbf{Y} \\ \mathbf{I} & \mathbf{0} \end{bmatrix}, \tag{22}$$

where $\mathbf{Y}$ is an $N \times 1$ vector of observations of the dependent variable and $\mathbf{X}$ is an $N \times K$ matrix of the independent variables. For the simple regression case, $K = 2$. BR-like algorithms usually consist of two phases: Phase I forms a set of independent edge direction vectors, Phase II updates the variable basis until it converges. In general, BR-like algorithms are slightly faster than other algorithms with simpler structures. Their computational complexity, however, is significantly higher. The complicated variable definition and logical branches used in BR-like algorithms cause tremendous efforts in their hardware implementations and are thus less attractive in such cases. Focusing on efficient algorithms that have a simple structure for ease of implementation, Wesolowsky's direct descent algorithm stands out. The algorithm is summarized below.

Step 1. Set $k = 0$. Choose the initial values a_0, b_0. Choose j so that $|Y_j - a_0 X_j - b_0|$ is a minimum.

Step 2. Set $k = k + 1$. Use the weighted median structure to get the update for b,

$$b_k = \mathrm{MED}\left(\left|1 - \frac{X_i}{X_j}\right| \diamond \frac{Y_i - Y_j X_i/X_j}{1 - X_i/X_j} \,\Bigg|_{i=1}^{N}\right). \tag{23}$$

Record the index i at which the term $(Y_i - Y_j X_i/X_j)/(1 - X_i/X_j)$ is the weighted median output.

Step 3. (a) If $b_k - b_{k-1} = 0$: if $k \geq 3$, go to Step 4; if not, set $j = i$ and go to Step 2.

(b) If $b_k - b_{k-1} \neq 0$: set $j = i$ and go to Step 2.

Step 4. Let $b^* = b_k$, $a^* = Y_j/X_j - b^*/X_j$.

The major difference between Wesolowsky's algorithm and ours is that the weighted median operations in their case are used for intercept b updates, while in our algorithm, they are used for slope a updates. Since the realization of the weighted median in both algorithms can benefit from the partial sorting scheme stated above, to compare them, we only need to count the iteration times. Also notice that in the initialization of Step 1, there is a minimum-finding procedure, which can be considered a sorting operation thus treated as having the same order of complexity as a weighted median, even though they may be implemented with totally different structures. For this reason, this step in Wesolowsky's algorithm will be counted as one iteration. Figure 5 depicts the comparison of the newly proposed algorithm and Wesolowsky's direct descent algorithm in terms of number of iterations. It can be observed from Figure 5 that, for large sample sets, the newly proposed LAD regression method needs 5% less iterations, and about 15% less for small sample sets.

4. CONCLUSIONS

A new iterative algorithm for LAD regression is developed based on MLEs of location. A simple coordinate transformation technique is used so that the optimization within each iteration is carried out by a weighted median operation, thus the proposed algorithm is well suited for hardware implementation. Simulation shows that the new algorithm is comparable in computational complexity with the best algorithms available to date.

[2] which can be considered as the basic form of the other two best simplex-type algorithms, namely, Bloomfield and Steiger's [1], and Armstrong, Frome, and Kung's [11], according to [2].

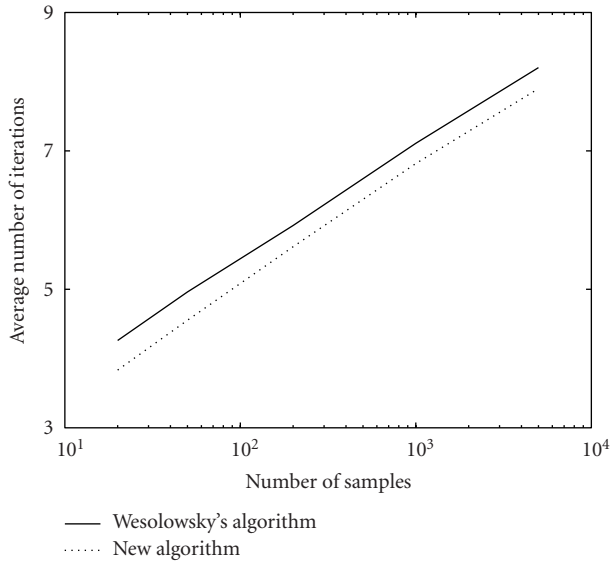

Figure 5: Comparison on the average number of iterations of Wesolowsky's and LA algorithms. The dimensions of the sample sets are chosen as [20, 50, 200, 1000, 5000], each having 1000 averaging runs.

APPENDIX

WEIGHTED MEDIAN COMPUTATION

The weighted median

$$Y = \text{MED}\left(W_i \diamond X_i \Big|_{i=1}^{N}\right), \tag{A.1}$$

having a set of positive real weights, can be computed out as follows.

(1) Calculate the threshold $W_0 = (1/2)\sum_{i=1}^{N} W_i$.
(2) Sort all the samples into $X_{(1)}, \ldots, X_{(N)}$ with the corresponding concomitant weights $W_{[1]}, \ldots, W_{[N]}$.
(3) Sum the concomitant weights beginning with $W_{[1]}$ and continuing up in order.
(4) The weighted median output is the sample $X_{(j)}$ whose weight causes the inequality $\sum_{i=1}^{j} W_{[i]} \geq W_0$ to hold first.

ACKNOWLEDGMENT

This work was supported in part by the Charles Black Evans Endowment and by collaborative participation in the Communications and Networks Consortium sponsored by the US Army Research Laboratory under the Collaborative Technology Alliance Program, Cooperative Agreement DAAD19-01-2-0011.

REFERENCES

[1] P. Bloomfield and W. L. Steiger, *Least Absolute Deviations: Theory, Applications, and Algorithms*, Progress in Probability and Statistics, Birkhäuser Boston, Boston, Mass, USA, 1983.

[2] Y. Dodge, Ed., *Statistical Data Analysis Based on the L_1-Norm and Related Methods*, Elsevier Science Publishers (North-Holland), Amsterdam, The Netherlands, 1987.

[3] Y. Dodge, Ed., *L_1-Statistical Analysis and Related Methods*, North-Holland Publishing, Amsterdam, The Netherlands, 1992.

[4] Y. Dodge, Ed., *L_1-Statistical Procedures and Related Topics*, Institute of Mathematical Statistics, Hayward, Calif, USA, 1997.

[5] Y. Dodge and W. Falconer, Eds., *Statistical Data Analysis Based on the L_1-Norm and Related Methods*, Barika Photography & Productions, New Bedford, Mass, USA, 2002.

[6] F. Y. Edgeworth, "A new method of reducing observations relating to several quantities," *Philosophical Magazine (Fifth Series)*, vol. 24, pp. 222–223, 1887.

[7] R. W. Hawley and N. C. Gallagher Jr., "On edgeworth's method for minimum absolute error linear regression," *IEEE Trans. Signal Processing*, vol. 42, no. 8, pp. 2045–2054, 1994.

[8] T. E. Harris, "Regression using minimum absolute deviations," *The American Statistician*, vol. 4, no. 1, pp. 14–15, 1950.

[9] A. Charnes, W. W. Cooper, and R. O. Ferguson, "Optimal estimation of executive compensation by linear programming," *Management Science*, vol. 1, no. 2, pp. 138–151, 1955.

[10] I. Barrodale and F. D. K. Roberts, "An improved algorithm for discrete l_1 linear approximation," *SIAM Journal on Numerical Analysis*, vol. 10, no. 5, pp. 839–848, 1973.

[11] R. D. Armstrong, E. L. Frome, and D. S. Kung, "A revised simplex algorithm for the absolute deviation curve fitting problem," *Communications in Statistics, Simulation and Computation*, vol. B8, no. 2, pp. 175–190, 1979.

[12] P. Bloomfield and W. Steiger, "Least absolute deviations curve-fitting," *SIAM Journal on Scientific and Statistical Computing*, vol. 1, no. 2, pp. 290–301, 1980.

[13] G. O. Wesolowsky, "A new descent algorithm for the least absolute value regression problem," *Communications in Statistics, Simulation and Computation*, vol. B10, no. 5, pp. 479–491, 1981.

[14] Y. Zhang, "Primal-dual interior point approach for computing l_1-solutions, and l_∞-solutions of overdetermined linear systems," *Journal of Optimization Theory and Applications*, vol. 77, no. 2, pp. 323–341, 1993.

[15] H. A. David, "Concomitants of order statistics," *Bulletin de l'Institut International de Statistique*, vol. 45, no. 1, pp. 295–300, 1973.

Yinbo Li was born in Mudanjiang, China, in 1973. He received the B.S. degree and M.S. degree in underwater acoustic and electrical engineering, both with the highest honors, from the Harbin Engineering University, Harbin, China, in 1994 and 1997, respectively. From 1997 to 1998, he was with the Institute of Acoustics, Chinese Academy of Sciences, Beijing, China, mainly focusing on signal processing and automatic system control. He was a Research and Development Engineer with the Beijing Division of Shenzhen Huawei Technology Co., Beijing, China, and a key member of the high-end router developing group from 1998 to 1999. He is currently a Research Assistant with the Department of Electrical and Computer Engineering, University of Delaware. He has been working with industry in the areas of signal processing and optical communications. His research interests include statistical signal processing, nonlinear signal processing and its applications, image processing, and optical and wireless communications.

Gonzalo R. Arce received the Ph.D. degree from Purdue University, West Lafayette, in 1982. Since 1982, he has been with the faculty of the Department of Electrical and Computer Engineering at the University of Delaware, where he is the Charles Black Evans Professor and Chairman of Electrical and Computer Engineering. His research interests include statistical and nonlinear signal processing, multimedia security, electronic imaging and display, and signal processing for communications. Dr. Arce received the Whittaker, Rehabilitation Engineering & Assistive Technology Society of North America (RESNA) and the Advanced Telecommunications/Information Distribution Research Program (ATIRP) Consortium best paper awards. He received the NSF Research Initiation Award. He is a Fellow of the IEEE. Dr. Arce was the Cochair of the 2001 EUSIPCO/IEEE Workshop on Nonlinear Signal and Image Processing (NSIP'01), Cochair of the 1991 SPIE's Symposium on Nonlinear Electronic Imaging, and the Cochair of the 2002 and 2003 SPIE ITCOM conferences. He has served as an Associate Editor for the IEEE Transactions on Signal Processing, and a Senior Editor of the Express.