A FAST MAXIMUM LIKELIHOOD ESTIMATION APPROACH TO LAD REGRESSION

Yinbo Li and Gonzalo R. Arce

Department of Electrical and Computer Engineering
University of Delaware, Newark, DE 19716 USA
e-mail: yli@eecis.udel.edu, arce@eecis.udel.edu

ABSTRACT

In this paper, we show that the optimization needed to solve the Least Absolute Deviations (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 a MLE estimate 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. The new algorithm provides a better trade-off solution between convergence speed and implementation complexity compared to existing algorithms.

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 (LAD) 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 devoted studies on L_1-norm methods [1, 2]. 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.

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

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

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. The Least Absolute Deviation 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).

In this paper, we derive a fast iterative solution to the LAD regression problem where the concept of Maximum Likelihood 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. BASIC UNDERSTANDING

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

$$F(b) = \sum_{i=1}^{N} |Y_i - a_0 X_i - b|. \tag{3}$$

Assuming a Laplace distribution for the errors U_i, the above cost function greatly resembles a Maximum Likelihood location estimator for b. Thus, the parameter b^* in this case

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 U.S. Army Research Laboratory under the Collaborative Technology Alliance Program, Cooperative Agreement DAAD19-01-2-0011.

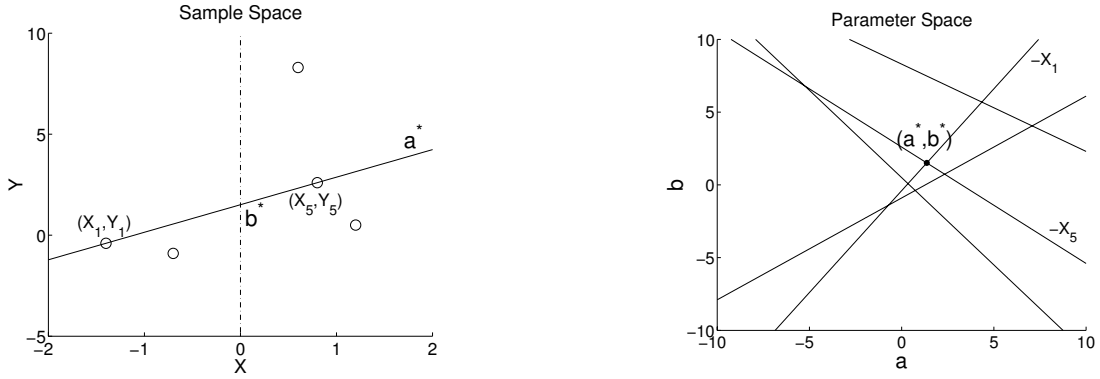

Fig. 1. Illustration of the sample space and the parameter space in the simple linear regression problem. The circles in the left plot represent the samples, the dot in the right plot represents the global minimum.

can be obtained by

$$b^* = \mathrm{MED}(Y_i - a_0 X_i \mid_{i=1}^{N}) \tag{4}$$

where $\mathrm{MED}(\cdot)$ stands for the well-known median operation. If, on the other hand, we fix $b = b_0$, the objective function reduces to

$$\begin{aligned} F(a) &= \sum_{i=1}^{N} |Y_i - b_0 - aX_i| \\ &= \sum_{i=1}^{N} |X_i| \left| \frac{Y_i - b_0}{X_i} - a \right|. \end{aligned} \tag{5}$$

Again, if the error random variable U_i obeys a Laplacian distribution, the observed samples $\{\frac{Y_i - b_0}{X_i}\}$ are also Laplacian distributed, but with the difference that each sample in this set has different variance. Thus the parameter a^* minimizing the cost function (5) can still be seen as the ML estimator of location for a, and can be calculated out as the *weighted median*,

$$a^* = \mathrm{MED}\left(|X_i| \diamond \frac{Y_i - b_0}{X_i} \,\middle|\, {}_{i=1}^{N} \right), \tag{6}$$

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 [3].

A simple and 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. 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 Fig. 1, the two spaces are dual to each other. In the sample space, 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, (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 $(-X_i, Y_i)$.

The structure of the objective function $F(a, b)$ is well defined as a polyhedron sitting on top of the a-b plane. 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. 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.

The geometrical interpretation of operation (4) can be derived as follows: draw a vertical line at $a = a_0$ in the parameter space, 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^*, the new update for b. Similar interpretation can be made for (6), 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 intuitive algorithm is that the convergence dynamics largely depends on the geometry of the cost surface. For example, if the very bottom of the cost function is a trench, then the optimization route will be bounced back and forth between two edges of this trench; correspondingly on the parameter plane, the convergence follows

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

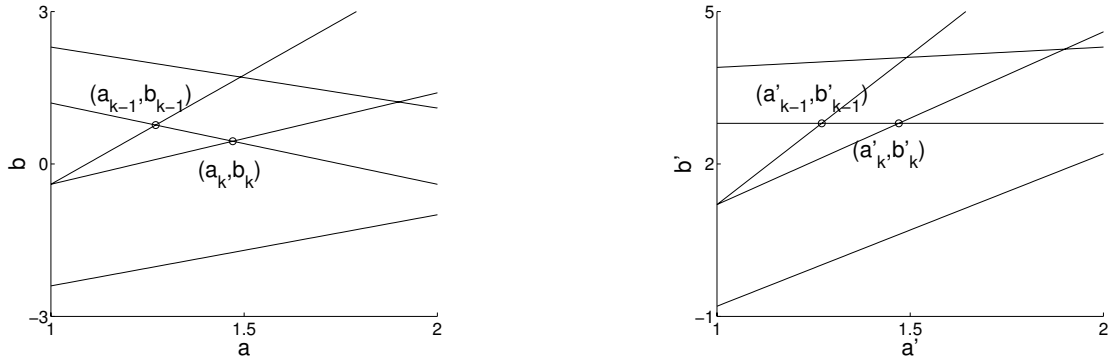

Fig. 2. Illustration of one iteration. The previous estimate (a_{k-1}, b_{k-1}) is mapped into the transformed coordinates as (a'_{k-1}, b'_{k-1}); (a'_k, b'_k) is obtained through ML estimation in the transformed coordinates; The new estimate (a_k, b_k) is formed by mapping (a'_k, b'_k) back into the original coordinates. The sample set is $[(1.6, 2.8), (-1.4, -3.8), (1.2, 3.5), (-4.3, -4.7), (-1.8, -2.2)]$.

a fairly inefficient zigzag manner. If the bottom of the cost function is a notch, then whenever the optimization gets on any part of that notch, it will stuck there, since the horizontal and vertical optimization will not bring it to the global minimum any further.

3. 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 Fig. 2.

The following is the proposed algorithm for LAD regression.

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

$$b_0 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(\bar{Y}X_i - \bar{X}Y_i)}{\sum_{i=1}^{N}(X_i - \bar{X})^2}. \qquad (7)$$

Calculate a_0 by a weighted median

$$a_0 = \text{MED}\left(|X_i| \diamond \frac{Y_i - b_0}{X_i} \,\middle|\, {}^{N}_{i=1}\right). \qquad (8)$$

Keep the index j which satisfies $a_0 = \frac{Y_j - b_0}{X_j}$. 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

$$X'_i = X_i - X_j \quad , \quad Y'_i = Y_i, \qquad (9)$$

and the transformations in the parameter space

$$a'_{k-1} = a_{k-1} \quad , \quad b'_k = b'_{k-1} = b_{k-1} + a_{k-1}X_j. \qquad (10)$$

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'

$$a'_k = \text{MED}\left(|X'_i| \diamond \frac{Y'_i - b'_k}{X'_i} \,\middle|\, {}^{N}_{i=1}\right). \qquad (11)$$

Keep the new index t which gives $a'_k = \frac{Y'_t - b'_k}{X'_t}$.

4) Transform back to the original coordinates

$$a_k = a'_k \quad , \quad b_k = b'_k - a'_k X_j \qquad (12)$$

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 (9) and (10). This relationship guarantees that the new update in each iteration is correct.

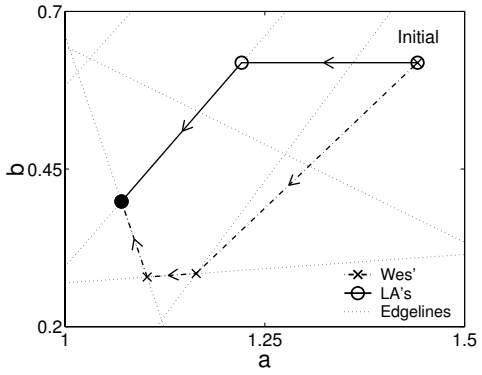

Fig. 3. Comparison on Wesolowsky's and Li and Arce's algorithms: the convergence of the algorithms on the parameter space. Two algorithms choose the same LS solution as the initial point. The marked dot represents the global minimum. Notice that not all the edgelines are plotted.

4. SIMULATION

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. In general, BR-like algorithms [2] 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' direct descent algorithm stands out [4].

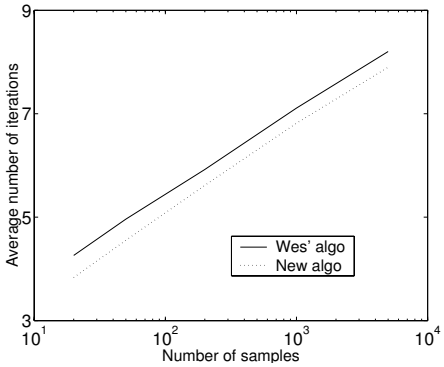

Fig. 4. Comparison on the average number of iterations of Wesolowsky's and LA algorithms. The dimensions of the sample sets are chosen as [20, 50, 200, 1000, 5000], each having 1000 averaging runs.

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. Also as depicted in Fig. 3, the first iterations of the two algorithms are different. LA algorithm picks the first a update horizontally, whereas Wesolowsky's algorithm chooses a nearby intersection based on a minimization operation. 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 Wesolowsky's algorithm, 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. Fig 4 depicts the comparison of the newly proposed algorithm and Wesolowsky's direct descent algorithm, in terms of number of iteration.

It can be observed from Fig. 4 that, for large sample sets, the newly proposed LAD regression method needs 5% less iterations, and about 15% less for small sample sets.

5. CONCLUSIONS

A new iterative algorithm for Least Absolute Deviation regression is developed based on Maximum Likelihood Estimates 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.

6. REFERENCES

[1] P. Bloomfield and W. L. Steiger, *Least Absolute Deviations: Theory, Applications, and Algorithms*. Boston: Birkhauser, 1983.

[2] Y. Dodge, Ed., *Statistical Data Analysis: Based on the L_1-Norm and Related Methods*. The Netherlands: Elsevier Science, 1987.

[3] G. R. Arce, "A general weighted median filter structure admitting negative weights," *IEEE Transactions on Signal Processing*, vol. 46, pp. 3195–3205, Dec. 1998.

[4] G. O. Wesolowsky, "A new descent algorithm for the least absolute value regression," *Commun. Statist.*, vol. B10, no. 5, pp. 479 – 491, 1981.