An analysis of the Least Median of Squares regression problem

Nikolai Krivulin
Faculty of Mathematics and Mechanics, St.Petersburg State University

Abstract

The optimization problem that arises out of the least median of squared residuals method in linear regression is analyzed. To simplify the analysis, the problem is replaced by an equivalent one of minimizing the median of absolute residuals. A useful representation of the last problem is given to examine properties of the objective function and estimate the number of its local minima. It is shown that the exact number of local minima is equal to $\binom{p+\lfloor (n-1)/2\rfloor}{p}$, where p is the dimension of the regression model and n is the number of observations. As applications of the results, three algorithms are also outlined.

1. INTRODUCTION

The *least median of squares* (LMS) method has recently been proposed by Rousseeuw in [6] to provide a very robust estimate of parameters in linear regression problems. The LMS estimate can be obtained as the solution of the following optimization problem.

Let $x_i^\top = (x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$, and $y = (y_1, \ldots, y_n)^\top$ be given real vectors. We assume that $n/2 \geq p$ and the $(n \times p)$–matrix $X = [x_{ij}]$ is of full rank to avoid degenerate cases. Let $\theta = (\theta_1, \ldots, \theta_p)^\top$ be a vector of regression parameters. The optimization problem that arises out of the LMS method is to find θ^* providing

$$\min_\theta \ \mathrm{med}\,\{(y_i - x_i^\top \theta)^2\}. \tag{1}$$

It is known (see [2, 4, 6, 7]) that the objective function in (1) is hard to minimize. This function is multiextremal, it is considered as having $O(n^p)$ local minima. In fact, there are efficient and exact algorithms available only for the problems of the lowest dimensions. A simple algorithm for $p = 1$ can be found in [6]. Another two designed for the problems of the dimension $p = 2$ have been described in [2] and [7]. For other dimensions, there are probabilistic algorithms producing approximate solutions of the problem (see [1] and [4]).

The purpose of this paper is to present some new ideas concerning the LMS problem so as to provide some theoretical framework for efficient regression algorithms. In Section 2 we offer a useful representation of the problem. The representation is exploited in Section 3 to demonstrate properties of the objective function and estimate the number of its local minima. Section 4 includes our main result providing the exact number of local minima. Finally, in Section 5 we briefly outline three LMS regression algorithms based on the above results.

2. REPRESENTATION OF THE LMS PROBLEM

To produce our representations, we first replace (1) by an equivalent problem just examined below. Obviously, the solutions of (1) are exactly the same as those of the problem:

$$\min_{\theta} \ \mathrm{med}\,\{|y_i - x_i^\top \theta|\}. \tag{2}$$

A serious difficulty one meets in analyzing both problems (1) and (2) is that it is hard to understand how the median behaves as the objective function. The next result offers a useful representation for the median as well as for other operators defined by means of ordering.

Let $R = \{r_1, \ldots, r_n\}$ be a finite set of real numbers. Suppose that we arrange its elements in order of increase, and denote the kth smallest element by $r_{(k)}$. If there are elements of the equal value, we count them repeatedly in an arbitrary order.

Lemma 1 *For each $k = 1, \ldots, n$, the value of $r_{(k)}$ is given by*

$$r_{(k)} = \min_{I \in \Im_k} \ \max_{i \in I} r_i, \tag{3}$$

where $\Im_k$ is the set of all k-subsets of the set $N = \{1, \ldots, n\}$.

Proof. Denote the set of indices of the first k smallest elements by I^*. It is clear that $r_{(k)} = \max_{i \in I^*} r_i$. Consider an arbitrary subset $I \in \Im_k$. Obviously, if $I \neq I^*$, there is at least one index $j \in I$ such that $r_j \geq r_{(k)}$. Therefore, we have $r_{(k)} \leq \max_{i \in I} r_i$. It remains to take minimum over all $I \in \Im_k$ in the last inequality so as to get (3). □

Let $h = \lfloor n/2 \rfloor + 1$, where $\lfloor n/2 \rfloor$ is the largest integer less than or equal to $n/2$. For simplicity, we assume $\mathrm{med}_{i \in N}\, r_i = r_{(h)}$. (It is absolutely correct to define the median in this form if n is odd. However, for an even n, it is normally defined as $\frac{1}{2}(r_{(h-1)} + r_{(h)})$.) By using (3) with $k = h$ and $r_i = r_i(\theta) = |y_i - x_i^\top \theta|$, we may now rewrite (2) as follows:

$$\min_{\theta} \ \min_{I \in \Im_h} \ \max_{i \in I} \ |y_i - x_i^\top \theta|. \tag{4}$$

The obtained representation seems to be more useful than the original because it is based on the well-known functions *max* and *min*. Moreover, the representation allows of further reducing the problem. In particular, one may change the order of the operations of taking minimum in (4) and get

$$\min_{I \in \Im_h} \ \min_{\theta} \ \max_{i \in I} \ |y_i - x_i^\top \theta|. \tag{5}$$

Assume I to be a fixed subset of N. Consider the problem

$$P(I): \ \min_{\theta} \ \max_{i \in I} \ |y_i - x_i^\top \theta|. \tag{6}$$

This is the well-known problem of fitting a linear function according to the l_∞–criterion, first examined by Fourier in the early 19th century [3]. The method proposed by Fourier was actually a version of the simplex algorithm and therefore (6) may be regarded as one of the oldest problems in linear programming. For modern methods and ideas, one can be referred to [5]. Incidentally, by applying an additional variable ρ, we may shape (6) into a usual form of linear programming problems:

$$\begin{aligned} &\min\ \rho \\ &\text{subject to}\quad \rho - x_i^\top\theta \ge -y_i,\ \ \rho + x_i^\top\theta \ge y_i,\ i \in I. \end{aligned} \tag{7}$$

To conclude this section, note that (5 may be regarded as a "two–stage" problem of both combinatorial optimization and linear programming. It consists in minimizing a function defined on a discrete set by solving some linear programming problem.

3. AN ANALYSIS OF THE OBJECTIVE FUNCTION

In this section we examine properties of the objective function in (4), *i.e.*

$$F(\theta) = \min_{I\in\Im_h} \max_{i\in I} |y_i - x_i^\top\theta|. \tag{8}$$

The main question we will try to answer is how many local minima it can have. To start the discussion, consider the function $\varrho_I(\theta) = \max_{i\in I} |y_i - x_i^\top\theta|,\ I \subset N$. It is a piecewise linear and convex function bounded below. Clearly, the problem of minimizing $\varrho_I(\theta)$ always has the solution.

The function $\varrho_I(\theta)$ can be portrayed as the surface of a convex polyhedron in a $(p+1)$–dimensional space. It is not difficult to see that function (8), which one may now express as $F(\theta) = \min_{I\in\Im_h} \varrho_I(\theta)$, also allows of visualizing its graph as the surface of some polyhedron. It is that produced by taking union of the polyhedra associated with $\varrho_I(\theta)$, for all $I \in \Im_h$. Note that $F(\theta)$ is still piecewise linear, but fails to be convex. An illustration for $p = 1$ and $n = 5$ is given in Figure 1.

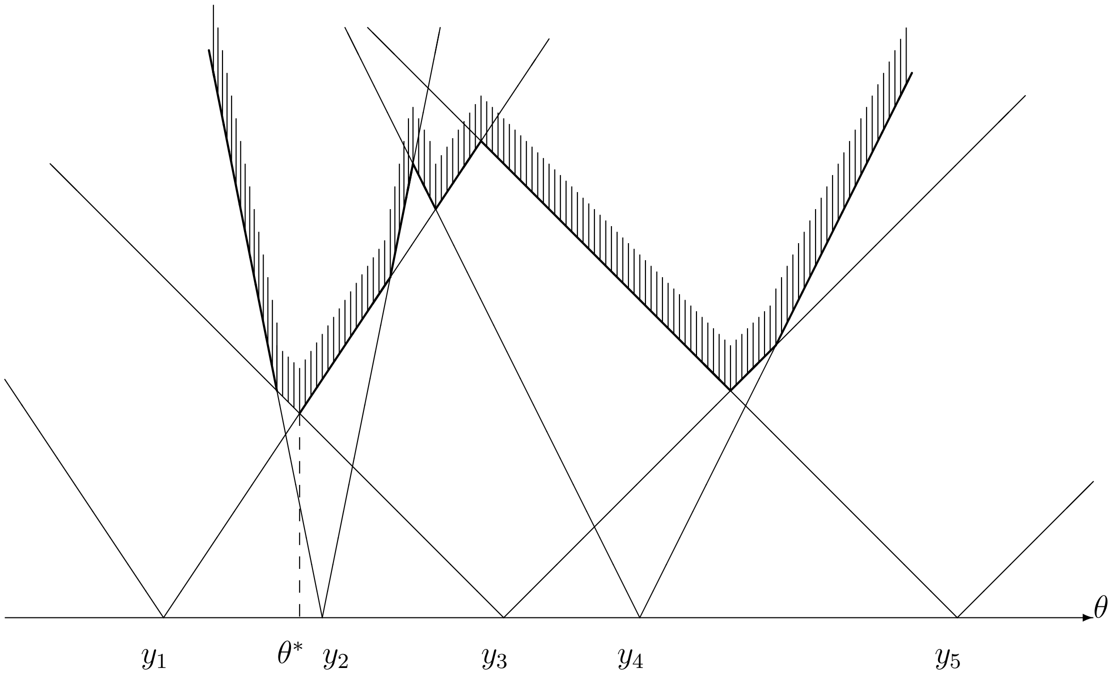

Figure 1: An objective function plot.

The objective function in Figure 1 is multiextremal, it has three local minima. It is clear that for practical problems, the number of the local minima of (8) can be enormous.

To take the first step to determining this number, we may conclude from representation (5) that it must not be greater than the number of problems $P(I)$ for all $I \in \Im_h$. This last is equal to $\binom{n}{h}$, *i.e.* the number of all h–subsets $I \in \Im_h$.

Suppose θ^* to be the solution of a problem $P(I)$, $|I| \geq p+1$. One can state the condition for the function $\varrho_I(\theta)$ to have the minimum at θ^* (see [5]): it is necessary that there exist a $(p+1)$–subset $I^* \subset I$ and real numbers λ_i to satisfy

$$\sum_{i \in I^*} \lambda_i \, \varepsilon_i \, x_i = 0, \quad \sum_{i \in I^*} \lambda_i = 1, \quad \lambda_i \geq 0, \; i \in I^*, \tag{9}$$

for some $\varepsilon_i \in \{-1, 1\}$. In other words, θ^* is defined by the point of intersection of $p+1$ "active" hyperplanes $\rho + \varepsilon_i \, x_i^\top \theta = \varepsilon_i \, y_i$ for some choice of $\varepsilon_i \in \{-1, 1\}$, $i \in I^*$, provided that the intersection point is an "acute" top of the corresponding polyhedron.

On the other hand, for any $(p+1)$–subset of indices, we are always able to choose both λ_i and ε_i suitable to satisfy (9). To illustrate this, let us examine an arbitrary $(p+1)$–subset. Without loss of generality, we assume it to be $\{1, \ldots, p+1\}$. Consider the equation $\sum_{i=1}^{p} t_i \, x_i = -t_{p+1} \, x_{p+1}$, and set $t_{p+1} = 1$ in it. Since $\operatorname{rank}(X) = p$, we may obtain values of $t_1, \ldots, t_p$ as the unique solution of the above equation. For every $i = 1, \ldots, p+1$, we define $\lambda_i = |t_i| / \sum_{j=1}^{p+1} |t_j|$, $\varepsilon_i = \operatorname{sign}(t_i)$. Obviously, $\lambda_i, i = 1, \ldots, p+1$, are just those required in (9).

As we have shown, the solution of any problem $P(I)$ is determined by $p+1$ vectors x_i. Conversely, any $p+1$ vectors x_i produce only one point which satisfies the necessary condition (9) and can therefore be treated as the solution of some problem. Clearly, the number of the local minima of $F(\theta)$ must not be greater than the number of such points, equaled $\binom{n}{p+1}$. Since we assume that $p \leq n/2$, our first estimate $\binom{n}{h}$ can be improved by replacing it by $\binom{n}{p+1}$. Although the last estimate is still rough, yet it is much lower than the quantity n^p considered in [2, 4, 7] as the order of the number of local minima.

4. THE EXACT NUMBER OF LOCAL MINIMA

We may now present our main result providing us with the exact number of local minima in (4). In fact, it allows of determining the number of local minima for any function of the absolute residuals $|y_i - x_i^\top \theta|$, $i \in N$, defined by using representation (3).

For each $k = 0, 1, \ldots, n-(p+1)$, let us introduce the function

$$f_k(\theta) = \min_{I \in \Im_{n-k}} \max_{i \in I} |y_i - x_i^\top \theta|, \tag{10}$$

and denote the number of its local minima by M_k. It should be noted that we have to set $k = n - h = \lfloor \frac{n-1}{2} \rfloor$ in (10) to produce the objective function of problem (4).

Theorem 2 *For each* $k = 0, 1, \ldots, n-(p+1)$*, it holds*

$$M_k = \binom{p+k}{p}. \tag{11}$$

Sketch of the proof. Let Π be the set of problems $P(I)$ for all $I \subset N$, $|I| \geq p+1$. To prove the theorem, we express $|\Pi|$, *i.e.* the number of all the problems in Π, in two ways. Firstly, it is easy to see that this number may be calculated as the sum

$$|\Pi| = \binom{n}{0} + \binom{n}{1} + \ldots + \binom{n}{n-(p+1)} = \sum_{j=0}^{n-(p+1)} \binom{n}{j}. \tag{12}$$

To produce the second representation, we examine a local minimum of the function $f_k(\theta)$ for an arbitrary k, $0 \leq k \leq n-(p+1)$. Assume θ^* to be the point of the local minimum. It is clear that $\theta^* = \theta^*(I)$ is the solution of some problem $P(I)$, where $|I| = n-k$. Since θ^* is actually determined by a subset $I^* \subset I$, which consists of $p+1$ "active" indices, it is also the solution of problems $P(I \setminus J)$ for all $J \subset I \setminus I^*$. The number of the problems having the solution at θ^* coincides with the number of all subsets of $I \setminus I^*$ including the empty set $\emptyset$, and equals $2^{n-(p+1)-k}$. In that case, the total number of the problems connected with the local minima of $f_k(\theta)$ is $2^{n-(p+1)-k}M_k$.

Now we may express $|\Pi|$ in the form:

$$|\Pi| = 2^{n-(p+1)}M_0 + 2^{n-(p+1)-1}M_1 + \ldots + M_{n-(p+1)} = \sum_{j=0}^{n-(p+1)} 2^{n-(p+1)-j}M_j. \tag{13}$$

From (12) and (13), we have

$$\sum_{j=0}^{n-(p+1)} 2^{n-(p+1)-j}M_j = \sum_{j=0}^{n-(p+1)} \binom{n}{j}. \tag{14}$$

It is not difficult to understand that for a fixed k, $0 \leq k \leq n-(p+1)$, the number M_k depends on p, but does not on n. One can consider M_0 as an illustration. Because the problem $P(N)$ has the unique solution (see [5]), M_0 is always equal to 1. Also, it holds $M_1 = p+1$ independently on n. To see this, note that every one of the local minima of $f_1(\theta)$ can be produced by relaxing only one of $p+1$ "active" constraints at the minimum point of $f_0(\theta)$.

Setting $n = p+1, p+2, p+3, \ldots$ in (14), we may successively get $M_0 = 1$, $M_1 = p+1$, $M_2 = \frac{(p+1)(p+2)}{2}, \ldots$. It is not difficult to verify that the general solution of (14) is represented as (11). □

Finally, substituting $k = \lfloor \frac{n-1}{2} \rfloor$ into (11), we conclude that the objective function of the LMS problem has $\binom{p+\lfloor (n-1)/2 \rfloor}{p}$ local minima.

5. APPLICATIONS

In this section we briefly outline LMS regression algorithms based on the above analysis of the problem. Only the main ideas that underlie the algorithms are presented.

"Greedy" algorithm. The algorithm produces an approximate solution and consists of solving the sequence of problems (6), $P(I_0), P(I_1), \ldots, P(I_{n-h})$, where $I_0 = N$ and the sets $I_1, I_2, \ldots, I_{n-h}$ are defined as follows. Let I_k^* be the set of $p+1$ "active" indices for the solution of a problem $P(I_k)$. Clearly, for each $i \in I_k^*$, the minimum of the objective function in the problem $P(I_k \setminus \{i\})$ is at least no greater than that in $P(I_k)$. Denote by i_k^* the index that yields the problem having the lowest solution. Finally, we define $I_{k+1} = I_k \setminus \{i_k^*\}$.

The "greedy" algorithm formally requires solving $(n-h)\times(p+1)+1$ optimization problems. In practice, however, an efficient procedure of transition between points, which yields the solutions of the problems, may be designed to avoid solving each of them.

Exhaustive search algorithm. This algorithm may be considered as the complete version of the previous one which actually uses a reduced search procedure. It exploits the classical depth-first search technique to provide all local minima of the objective function. From Theorem 2, one can conclude that it requires examining $\binom{n-h+p+1}{p+1}$ points to produce the exact solution. Because of its exponential time complexity, this search algorithm can hardly be applied to problems of high dimensions. Note, however, that it normally allows of solving problems with $p \leq 5$ within reasonable time.

Branch and probability bound algorithm. It is a random search algorithm based on the Branch and Probability Bound (BPB) technique which has been developed in [8] as an efficient tool for solving both continuous and discrete optimization problems. The BPB algorithm designed to solve the LMS problem is of combinatorial optimization. It produces an approximate solution by searching over $(p+1)$–subsets of N. As it follows from Section 3, each $(p+1)$–subset determines a point satisfying the condition (9), one of such points is the solution of the LMS problem.

In conclusion, I would like to thank Professor A.A. Zhigljavsky for drawing my attention to the problem and for valuable discussions, and Professor A.C. Atkinson for his kind interest in this work as well as for providing me with a reprint of paper [1].

6. REFERENCES

[1] Atkinson, A.C. and Weisberg, S. (1991). Simulated annealing for the detection of multiple outliers using least squares and least median of squares fitting. In Directions in Robust Statistics and Diagnostics, Part I (Eds. W. Stahel and S. Weisberg). Springer–Verlag, 7–20.

[2] Edelsbrunner, H. and Souvaine, D.L. (1990). Computing least median of squares regression lines and guided topological sweep. Journal of American Statistical Association 85, 115–119.

[3] Fourier, J.B.J. (1826). Analyse de travaux de 1^1. Academie Royale des Sciences pendant l'année 1823, Partie Mathématique, Histoire de l'Academie Royale des Sciences de l'Institute de France, 6, XXIV–xli.

[4] Joss, J. and Marazzi, A. (1990). Probabilistic algorithms for least median of squares regression. Computational Statistics and Data Analysis 9, 123–133.

[5] Polyak, B.T. (1989). Introduction to Optimization. Springer–Verlag.

[6] Rousseeuw, P.J. (1984). Least median of squares regression. Journal of American Statistical Association 79, 871–880.

[7] Souvaine, D.L. and Steel, J.M. (1987). Time– and space–efficient algorithms for least median of squares regression. Journal of American Statistical Association, 82, 794–801.

[8] Zhigljavsky, A.A. (1991). Theory of Global Random Search. Kluwer Academic Publishers.