ELSEVIER Computational Statistics & Data Analysis 20 (1995) 467–479

COMPUTATIONAL STATISTICS & DATA ANALYSIS

The determination of a "least quantile of squares regression line" for all quantiles

Emilio Carrizosa[a], Frank Plastria[b,*]

[a] *Facultad de Matematicas, Universidad de Sevilla, Tarfia s/n. 41012 Sevilla, Spain*
[b] *Center for Industrial Location, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium*

Received 1 July 1992; revised 1 September 1994

Abstract Least median of squares regression has shown to be an extremely useful tool in robust regression analysis.

In this note, we extend this concept to least quantile of squares regression, and propose a polynomial algorithm that finds simultaneously an estimator for each quantile.

This leads to a proposal of a robust minimum scale regression line and a polynomial algorithm for its determination.

Keywords: Least median of squares regression; Robust regression; Sweep-line technique; Minquantile optimization

1. Introduction

Given a set $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ of points in $\mathbb{R}^2$, define the error function $r: \mathbb{R}^2 \to \mathbb{R}^n$, $r(a, b) = (y_1 - ax_1 - b, \ldots, y_n - ax_n - b)$.

As the classical least sum of squares line, the line $l_{\bar{a},\bar{b}} = \{(x, y): y = \bar{a}x + \bar{b}\}$ such that

$$\sum_{i=1}^{n} r_i(\bar{a}, \bar{b})^2 = \min_{(a,b)\in\mathbb{R}^2} \sum_{i=1}^{n} r_i(a, b)^2$$

does not perform well in presence of outliers in S, some more robust fitting estimators have been proposed.

SSDI 0167-9473(94)00059-X

Rousseeuw (1984) introduced the least median of squares (LMS) regression: an LMS estimator is a line $l_{\bar{a},\bar{b}} = \{(x, y): y = \bar{a}x + \bar{b}\}$ such that $\text{med}_i\,(\bar{a}, \bar{b})^2 = \min_{(a,b)\in\mathbb{R}^2} \text{med}_i\, r_i\,(a, b)^2$.

LMS has been shown to be an extremely useful tool for data analysis when there exists a very high (even close to 50%) degree of contamination in the sample S (see Rousseeuw and Leroy (1987) for a discussion of robust regression in general and LMS in particular.)

A deeper insight could be obtained if the methodology allowed for:

(a) The use of weights (frequencies) for different observations, as suggested by Souvaine and Steele (1987).
(b) Obtaining estimators for different quantiles, as suggested by Cook and Hawkins (1990), without an important increase in complexity.
(c) Enabling a choice among the different regression lines obtained for different quantiles, e.g., minimizing a scale parameter suggested by Rousseeuw (1984).

The aim of this paper is the design of an algorithm including the two first aspects, and showing how this leads to a solution of the third.

2. Minquantile lines

Let $\bar{S} = \{(\bar{x}_1, \bar{y}_1), \ldots, (\bar{x}_N, \bar{y}_N)\}$ be a sample in $\mathbb{R}^2$. Consider the set $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ of different points in $\bar{S}$, and associate with each (x_i, y_i) the frequency w_i of (x_i, y_i) in the original sample $\bar{S}$.

Throughout this paper, we assume w.l.o.g. that $x_1 \le x_2 \le \cdots \le x_n$; we also impose that $x_1 < x_n$ and $n \ge 3$: otherwise, the regression problem has no interest.

For each $I \subset \{1, \ldots, n\}$, denote by $W(I)$ the weight associated with the subset $\{(x_i, y_i): i \in I\}$ of S, i.e., $W(I) = \sum_{i\in I} w_i$.

For $m = 1, \ldots, N$, define the function $Q_m: \mathbb{R}^2 \to \mathbb{R}$,

$$Q_m(a, b) = \min_{\substack{I \subset \{1,\ldots,n\} \\ W(I) \ge m}} \max_{i\in I} r_i(a, b)^2$$

and

$$Q_m^* = \min_{(a,b)\in\mathbb{R}^2} Q_m(a, b).$$

Definition 1. The line $l_{\bar{a},\bar{b}} = \{(x, y): y = \bar{a}x + \bar{b}\}$ is said to be an m/N-minquantile line if $Q_m(\bar{a}, \bar{b}) = Q_m^*$.

This definition was already suggested by Rousseeuw and Leroy (1987) under the name of *least quantile of squares regression line*. Observe that the LMS lines are the $\lceil (N + 1)/2 \rceil / N$-minquantile lines.

For the unweighted problem (i.e. $S = \bar{S}$, $w_i = 1$, for all $i = 1, 2, \ldots, n = N$), it has been observed that any m/N-minquantile line (with $m \ge 3$) must be a Chebyshev line (i.e. a line minimizing the maximal absolute error) for a set of three observations

(see Stromberg, 1991). Hence, the search of minquantile lines is reduced to a finite search of cardinality $O(N^3)$. This result can be extended to the weighted case, in the sense that there exists a set of cardinality $O(n^3)$ containing, for each m, an m/N-minquantile line. In order to give an explicit representation of such a set, some notation is needed.

Definition 2. Given three points P_i, P_j, P_k of S, the line $l_{a,b}$ is said to be the *equidistance cutting line* (e.c.l.) of the triplet (P_i, P_j, P_k) iff

$$r_i(a, b) = -r_j(a, b) = r_k(a, b)$$

Remark 1. Any triplet $((x_i, y_i), (x_j, y_j), (x_k, y_k))$ of different points has a unique e.c.l. unless $x_i = x_k$ (in that case, no (nonvertical) e.c.l. exists).

It is well known (see e.g. Appa and Smith, 1973) that, when points are in general position, for any Chebyshev line $l_{a,b}$, there exists some triplet (P_i, P_j, P_k) which has $l_{a,b}$ as e.c.l. This result will be extended to the problem addressed here.

For $i, j \in \{1, 2, \ldots, n\}$, $x_i \neq x_j$, let

- $l_{a_{ij}, b_{ij}}$ be the line containing P_i and P_j,
- $z_{ij} = 0$,
- $\alpha_{ij} = W(\{k: l_{a_{ij}, b_{ij}} \text{ contains } P_k\})/N$,
- $v_{ij} = ((a_{ij}, b_{ij}), z_{ij}, \alpha_{ij})$.

Furthermore, for each triplet (P_i, P_j, P_k), $x_i \neq x_k$, of distinct points in S, let $l_{a_{ijk}, b_{ijk}}$ be its e.c.l., and define

- $z_{ijk} = r_i(a_{ijk}, b_{ijk})^2$,
- $\alpha_{ijk} = W(\{s: r_s(a_{ijk}, b_{ijk})^2 \leq z_{ijk}\})/N$,
- $v_{ijk} = \{(a_{ijk}, b_{ijk}), z_{ijk}, \alpha_{ijk})$.

With this notation, one has the following theorem.

Theorem 1. *Let B be the set*

$$B = \{v_{ij}: 1 \leq i < j \leq n, x_i < x_j\} \cup \{v_{ijk}: i \neq j \neq k, x_i \neq x_k\}.$$

For any m, $1 \leq m \leq N$, there exists $v = ((a, b), z, \alpha) \in B$ such that

(i) *$l_{a,b}$ is an m/N-minquantile line,*

(ii) *$Q_m^* = z$,*

(iii) *$m/N \leq \alpha$.*

Proof. Let m, $1 \leq m \leq N$. By definition,

$$Q_m^* = \min_{(a,b)} \min_{\substack{I \subset \{1,\ldots,n\} \\ W(I) \geq m}} \max_{i \in I} r_i(a, b)^2 = \min_{\substack{I \subset \{1,\ldots,n\} \\ W(I) \geq m}} \min_{(a,b)} \max_{i \in I} r_i(a, b)^2$$

Hence, there exists $I_1 \subset \{1, \ldots, n\}$, $W(I_1) \geq m$ such that

$$Q_m^* = \min_{(a,b)} \max_{i \in I_1} r_i(a, b)^2,$$

and any $(\bar{a}, \bar{b})$ solving min $\max_{i \in I_1} r_i^2$ gives an m/N-minquantile line.

The function r_i^2 is convex for all i; hence, (see, e.g, Drezner, 1982), there exists $I_2 \subset I_1$, $1 \leq \text{Card}(I_2) \leq 3$, such that

- $\min_{(a,b)} \max_{i \in I_2} r_i(a, b)^2 = \min_{(a,b)} \max_{i \in I_1} r_i(a, b)^2$.
- There exists (a^*, b^*), optimal solution to $\min \max_{i \in I_2} r_i^2$ which is also an optimal solution to $\min \max_{i \in I_1} r_i^2$.

Let

$$J = \left\{ j \in I_2 : r_j(a^*, b^*)^2 = \max_{i \in I_2} r_i(a^*, b^*)^2 \right\}.$$

It is obvious that J has cardinality $\text{Card}(J) \in \{1, 2, 3\}$, and, by the convexity of each r_i^2, (a^*, b^*) is an optimal solution to $\min \max_{i \in J} r_i^2$.

Let

$$z_J = \min_{(a,b)} \max_{i \in J} r_i(a, b)^2 = Q_m^*,$$

and consider the optimization subproblem:

$$(\text{SP}_J) \qquad \begin{aligned} &\max_{(a,b)} && \sum_{i=1}^{n} w_i t_i(a, b) \\ &\text{s.t.} && r_j(a, b)^2 = z_J \text{ for all } j \in J, \\ & && t_i(a, b) = \begin{cases} 1, & \text{if } r_i(a, b)^2 \leq z_J, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

By construction, (SP_J) is feasible, and any optimal solution $(\bar{a}, \bar{b})$ to (SP_J) yields an m/N- minquantile line satisfying $\sum_{i=1}^{n} w_i t_i(\bar{a}, \bar{b}) \geq m$. Hence, we only have to show that B contains an element of the form $((\bar{a}, \bar{b}), z_J, \sum_{i=1}^{n} w_i t_i(\bar{a}, \bar{b})/N)$ for some $(\bar{a}, \bar{b})$, optimal solution to (SP_J). We study separately the different values of $\text{Card}(J)$.

Case $\text{Card}(J) = 1$, $J = \{i\}$ (say)

First, observe that $z_J = 0$, thus the feasible points for (SP_J) correspond to the lines containing (x_i, y_i). Let (a, b) be a feasible solution to (SP_J), and let $K = \{j : t_j(a, b) = 1\}$. Clearly, if $\text{Card}(K) \geq 2$, then (a, b) is of the form (a_{ij}, b_{ij}). On the other hand, if $\text{Card}(K) = 1$, we can always take $j \neq i$ such that $x_j \neq x_i$. It is obvious that the line $l_{a_{ij}, b_{ij}}$ is feasible for (SP_J) and $\sum_{k=1}^{n} w_k t_k(a_{ij}, b_{ij}) \geq \sum_{k=1}^{n} w_k t_k(a, b)$. This implies that, whatever $\text{Card}(K)$, (SP_J) always admits an optimal solution of the form (a_{ij}, b_{ij}).

Case $\text{Card}(J) = 2$, $J = \{i, j\}$, *say, with* $i < j$

If $x_i < x_j$, the only Chebyshev solution for $\{P_j, P_j\}$ (thus the only feasible solution to (SP_J) is the line $l_{a_{ij}, b_{ij}}$, whose corresponding triplet v_{ij} is in B.

The case $x_i = x_j$ is similar to the case $\text{Card}(J) = 1$, except that the optimal lines are those containing the midpoint between P_i and P_j, and will not be repeated here.

Case $\text{Card}(J) = 3$

Straightforward: Any feasible solution to (SP_J) is, in particular, an e.c.l. for the points in J. □

Remark 2. Observe that, although only the m/N-minquantile lines with $[(m+1)/N] \geq \frac{1}{2}$ are of interest in robust analysis, the result above is valid for any m in $\{1, \ldots, N\}$.

3. Determination of minquantile lines

Several algorithms have been proposed in order to construct an exact LMS regression line in an efficient way; see, e.g., Steele and Steiger (1986), Souvaine and Steele (1987), Edelsbrunner and Souvaine (1990) and, Xu and Shiue (1993).

In this section we propose an algorithm that determines, for all m, $1 \leq m \leq N$, an m/N-minquantile line, for which we should expect a complexity higher than the $O(N^2)$-complexity obtained for the LMS estimation (Edelsbrunner and Souvaine, 1990) because of two reasons:

(1) Much more information must be obtained: in the worst case we will end up with N minquantile lines.

(2) The existing algorithms for the LMS estimator assume that points are in general position (no three points on a line, and no two points on the same vertical line), thus no frequencies (but the trivial) are allowed. When this assumption is highly violated (as might occur if $\bar{S}$ is a sample of a discrete random vector), such algorithms must be adapted, with a possible increase in complexity.

The output of the procedure we propose here consists of two lists of N elements, MINQ and Q, where, for any m, $1 \leq m \leq N$, MINQ$[m]$ stores the coefficients of an m/N-minquantile line, and $Q[m] = Q_m^*$. These two lists are obtained after performing three phases, which are described below.

1. Initialization of MINQ and Q,
2. Updating,
3. Garbage deletion.

Phase 1: (*Initialization of* MINQ and Q)

Set, for any m, MINQ$[m]$ equal to an arbitrary value, $(0, 0)$, say, and $Q[m] = +\infty$.

Phase 2: (*Updating*).

In order to update MINQ and Q, we modify the sweep-line technique used by Souvaine and Steele (1987) for the determination of the LMS line. For this purpose, define the mapping T, that takes $(a, b) \in \mathbb{R}^2$ to the line $l_{a,b}$, and the line $l_{a,b}$ to the point $(-a, b)$. (For a discussion of the properties of this mapping, refer to Souvaine and Steele, 1987). As T preserves vertical distances, the triplets v in B of Theorem 1 can be obtained in terms of T:

For any pair $P_i = (x_i, y_i)$, $P_j = (x_j, y_j)$, $x_i \neq x_j$, the associated triplet $v_{ij} = ((a_{ij}, b_{ij}), z_{ij}, \alpha_{ij})$ is obtained as follows:

- $l_{a_{ij}, b_{ij}} = T^{-1}(T(P_i) \cap T(P_j))$.
- $z_{ij} = 0$,
- $\alpha_{ij} = W(\{k: T(P_i) \cap T(P_j) \in T(P_k)\})/N$.

On the other hand, given three distinct points $P_i = (x_i, y_i)$, $P_j = (x_j, y_j)$, $P_k = (x_k, y_k)$, $x_i \neq x_k$, the triplet $v_{ijk} = ((a_{ijk}, b_{ijk}), z_{ijk}\ \alpha_{ijk})$ associated with the e.c.l.

$l_{a_{ijk},b_{ijk}}$ for (P_i, P_j, P_k) can be obtained as follows: Let $l_{a,b} = T^{-1}(T(P_i) \cap T(P_k))$ (i.e. $l_{a,b}$ is the line containing P_i and P_k). Then, the e.c.l. is obtained by considering the inverse by T of the middle point between $T(P_i) \cap T(P_k)$ and the point on $T(P_j)$ vertical from $T(P_i) \cap T(P_k)$. In other words, the e.c.l for (P_i, P_j, P_k) is $l_{a,(y_j - ax_j + b)/2}$, thus

- $(a_{ijk}, b_{ijk}) = (a, (b - x_j a + y_j)/2)$.
- $z_{ijk} = ((b + x_j a - y_j)/2)^2$,
- $\alpha_{ijk} = W(\{s: (-x_s a + y_s - (b - x_j a + y_j)/2)^2 \leq z_{ijk}\})/N$.

Following the notation of Souvaine and Steele (1987), we use two data structures of list type, LIST and HEAP. Initially, LIST is the list of lines $\{T(x_i, y_i), 1 \leq i \leq n\}$, arranged increasingly following the order $<$, given by

$$l_{a,b} < l_{c,d} \quad \text{iff } (a < c) \text{ or } (a = c, b > d).$$

Along with each $T(x_i, y_i)$, we define two pointers, UP and DOWN, pointing, respectively, at the immediate predecessor and successor of $T(x_i, y_i)$ following $<$ (if they exist). For each pair or nonparallel lines $T(x_i, y_i)$ and $T(x_j, y_j)$ which are adjacent in LIST, we add $T(x_i, y_i) \cap T(x_j, y_j)$ to HEAP in such a way that the top of HEAP contains the minimum point according to the lexicographical order. We also add the pointers from LIST to HEAP and from HEAP to LIST as described by Souvaine and Steele (1987).

Once the data structures have been initialized, the updating process is identical to the sweep-line method, taking into account that, if more than two lines intersect at $T(x_i, y_i) \cap T(x_k, y_k)$ (this is the degenerate case not considered there), the order of all such lines must be completely reversed in LIST. (Note that by incorporating this modification in the original Souvaine and Steele technique gets rid of the general position assumption in their LMS construction algorithm). Besides, at any step, with some $T(x_i, y_i) \cap T(x_k, y_k)$ at the top of HEAP, v_{ik} and *all* the triplets of the form v_{ijk} have to be evaluated (in the way mentioned above) and their corresponding z's compared with the values stored in Q. Observe that, after obtaining v_{ik}, the set $\{v_{ijk}\}_j$ can be obtained in $O(n)$ time, just by simultaneously sweeping LIST upwards from the highest among $T(P_i)$ and $T(P_k)$ and downwards from the lowest among $T(P_i)$ and $T(P_k)$.

During this operation, for any found triplet $v = ((a, b), z, \alpha)$, with $z < Q[\alpha N]$, we set

$$Q[\alpha N] \leftarrow z, \qquad \text{MINQ}[\alpha N] \leftarrow (a, b).$$

The process stops when HEAP is empty, that is, when all the triplets of B have been considered. Then, we go to Phase 3.

Phase 3. (*Garbage deletion*).

For any $i, j \in \{1, \ldots, N\}$, $i > j$, if $Q[i] \leq Q[j]$, then set

$$Q[j] \leftarrow Q[i], \qquad \text{MINQ}[j] \leftarrow \text{MINQ}[i].$$

Example. As a simple illustration, consider the sample $S = \{P_1, P_2, \ldots, P_6\}$ with coordinates and weights given in Table 1, and depicted in Fig. 1.

Table 1

i	P_i	w_i
1	(0, 0)	2
2	(2, 2)	2
3	(2, 3)	2
4	(3, 4)	2
5	(4, 3)	1
6	(8, 4)	1

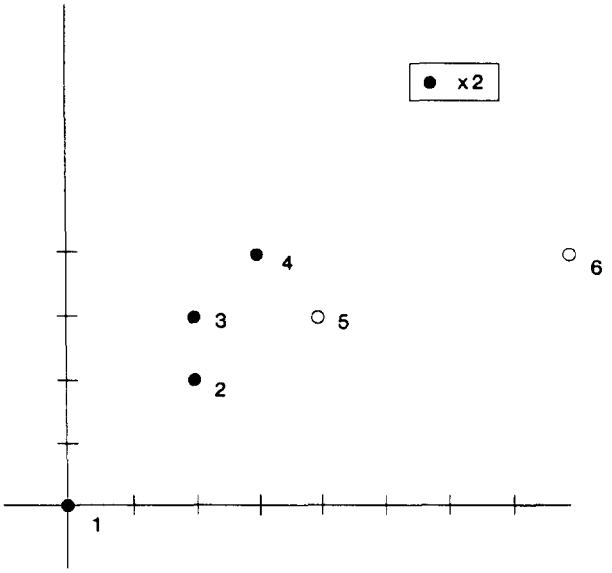

Fig. 1. The sample

Thus $N = 10$ and $n = 6$. In Phase 1, the lists Q and MINQ of 10 components each are built, and we set $Q[m] = +\infty$, MINQ$[m] = (0, 0)$ for all $m = 1, \ldots, 10$. Phase 2 starts with the construction of the structures LIST and HEAP, whose initialization values are shown in Fig. 2(a). Fig. 2(b) shows the initial situation under T-transform. The vertical sweepline is at the left, and the order in LIST may be read off downwards along this line. HEAP contains the intersection points at the right-hand vertex of each gray triangle (which represent the pointers between LIST and HEAP), ordered from left to right.

The lowest element of HEAP is $T(P_2) \cap T(P_4) = (-2, -2)$. Therefore, we compute the triplets v_{24} and $(v_{2j4})_j$, checking whether Q and MINQ must be updated:

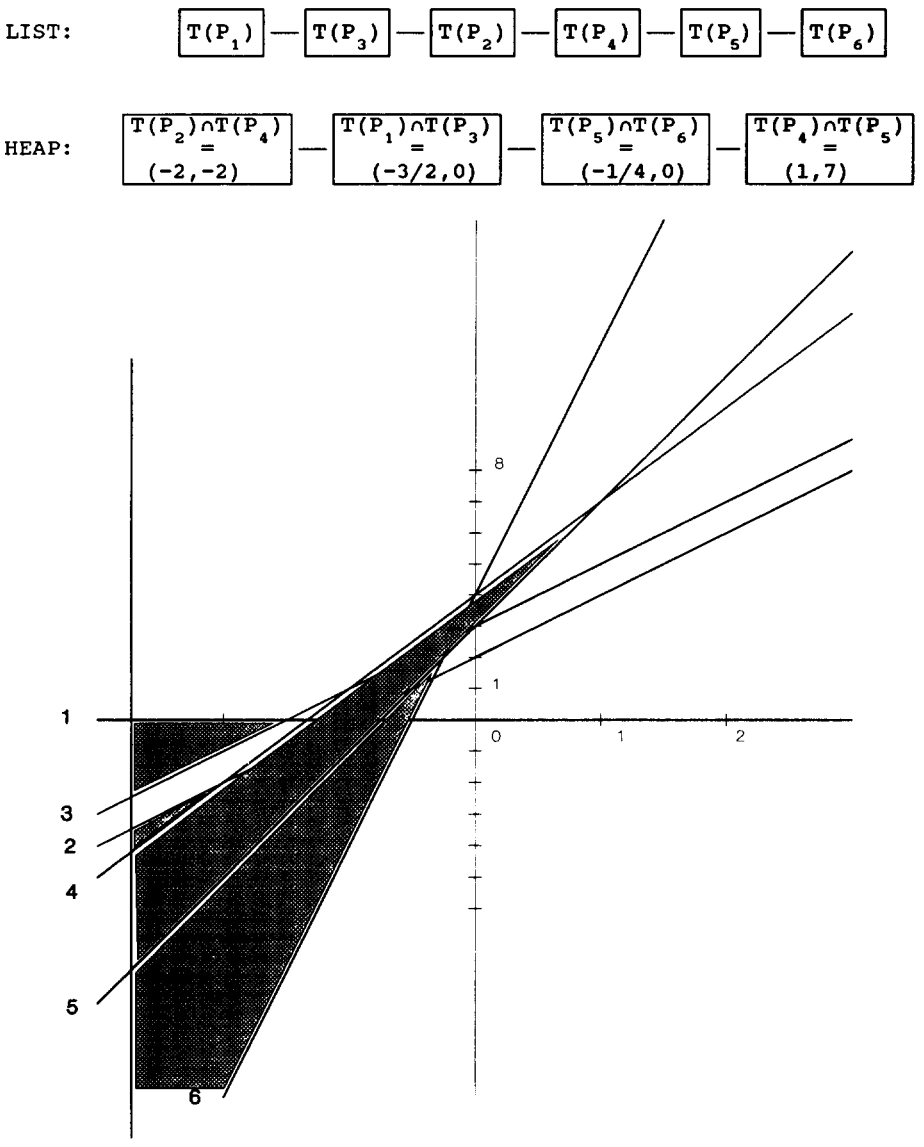

Fig. 2(a). Initial LIST and HEAP.(b).

- $v_{24} = ((2, -2), 0, 4/10)$. $Q[4] \leftarrow 0$, MINQ[4] $\leftarrow (2, -2)$,
- $v_{234} = ((2, -3/2), 1/4, 6/10)$ $Q[6] \leftarrow 1/4$ MINQ[6] $\leftarrow (3, -3/2)$,
- $v_{254} = ((2, 7/2), 9/4, 5/10)$ $Q[5] \leftarrow 9$ MINQ[5] $\leftarrow (2, 7/2)$,
- $v_{214} = ((2, -1), 1, 8/10)$ $Q[8] \leftarrow 1$ MINQ[8] $\leftarrow (2, -1)$,
- $v_{264} = ((2, -7), 25, 6/10)$ ($Q[6] = 1/4 < 25$, no updating).

After that, the order in LIST of $T(P_2)$ and $T(P_4)$ is interchanged, and HEAP is updated as shown in Fig. 3(a). Geometrically the sweepline has moved right across the point $(2, -2)$, as shown in Fig. 3(b).

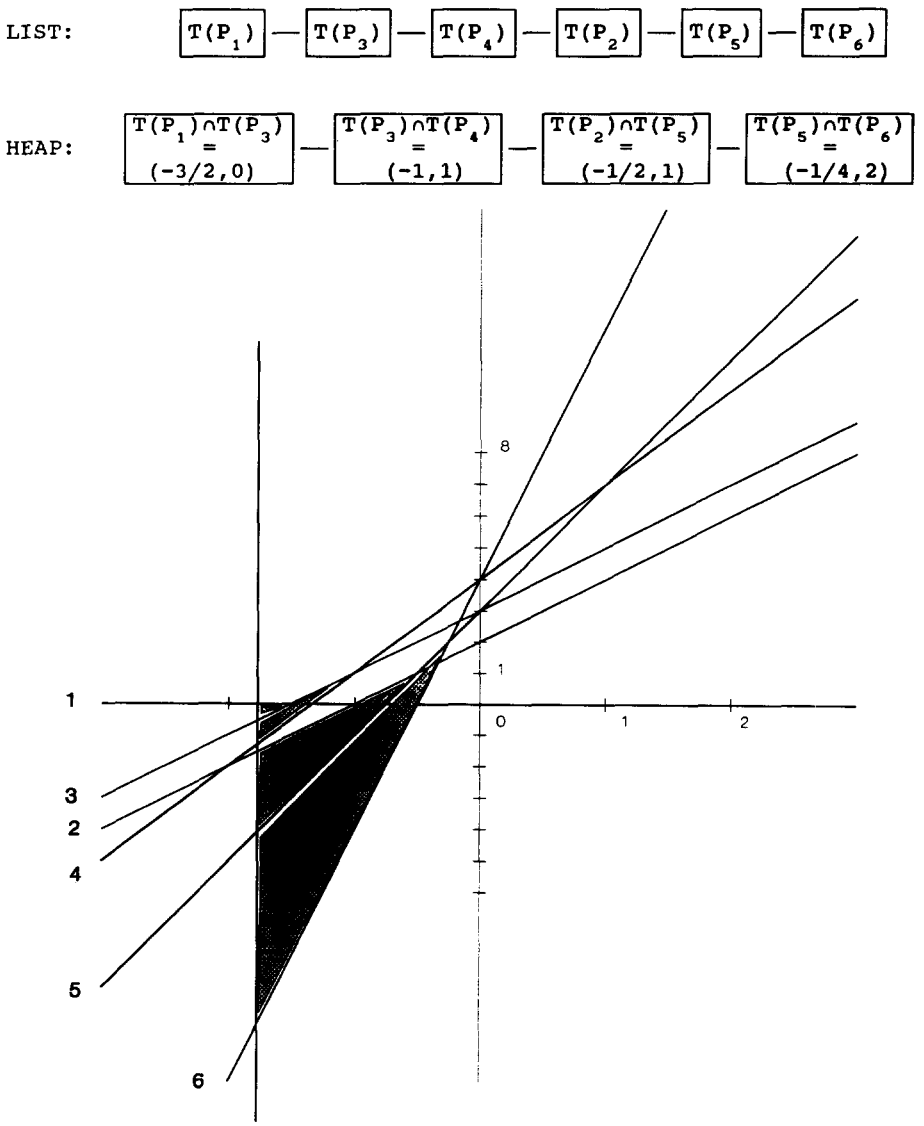

Fig. 3(a). First updating of LIST and HEAP.(b).

As the minimum of HEAP is $T(P_1) \cap T(P_3) = (-\frac{3}{2}, 0)$, one must evaluate v_{13} and $\{v_{1j3}\}_j$. For instance, $v_{143} = ((\frac{3}{2}, -\frac{1}{4}), 1/16, 6/10)$. As $Q[6] > 1/16$, $Q[6]$ and MINQ[6] must be updated, by setting $Q[6] \leftarrow 1/16$ and MINQ[6] $\leftarrow$ (3/2, $-1/2$).

The process goes on until HEAP becomes empty, the sweepline having moved across all intersection points, and Phase 3 starts, with Q and MINQ as shown in Table 2.

After the garbage deletion, one obtains the values given in Table 3.

Table 2

i	$Q[i]$	MINQ[i]
1	$+\infty$	(0, 0)
2	0	(1/4, 2)
3	0	(3/4, 0)
4	0	(2, −2)
5	1/16	(3/4, 1/4)
6	1/36	(4/3, 1/6)
7	1/4	(0, 7/2)
8	1/4	(1, 1/2)
9	49/64	(3/4, 7/8)
10	25/16	(1/2, 5/4)

Table 3

1	0	(2, −2)
2	0	(2, −2)
3	0	(2, −2)
4	0	(2, −2)
5	1/36	(4/3, 1/6)
6	1/36	(4/3, 1/6)
7	1/4	(1, 1/2)
8	1/4	(1, 1/2)
9	49/64	(3/4, 7/8)
10	25/16	(1/2, 5/4)

Theorem 2. *After having performed the three phases above, for any m,* MINQ[m], *and* $Q[m]$ *contain respectively the coefficients of an* m/N*-minquantile line and* Q_m^*.

Proof. Evident by Theorem 1. Observe that Phase 3 is necessary because in Theorem 1(iii), we only have $m/N \leq \alpha$. □

Theorem 3. MINQ *and* Q *can be obtained in* $O(n^3)$ *time with* $O(n)$ *space.*

Proof. Trivially Phase 1 is performed in $O(n)$ time and requires $O(n)$ space. In phase 2, the construction of LIST and HEAP requires $O(n \log(n))$ time and $O(n)$ space; furthermore, we have $O(n^2)$ steps in this phase (in fact, we have exactly $n(n+1)/2$ for points in general position), and at each of these steps, the computation of all corresponding v's, and updating of MINQ, Q, LIST and HEAP is done in $O(n)$ time. This gives to Phase 2 a complexity of $O(n^3)$ time and $O(n)$ space. As Phase 3 can be performed in $O(n)$ time with $O(n)$ space, the complexity of Phase 2 gives the total complexity of the algorithm. □

Remark 3. For simplicity, we have assumed that both LIST and HEAP are lists. Using, as in Souvaine and Steele (1987), more sophisticated data structures (e.g. heaps), one could alleviate the running time of some steps of the algorithm, but without reducing the overall $O(n^3)$ worst-case complexity.

Remark 4. After adaptation of the topological sweep of Souvaine and Steele to other quantiles, one could also obtain a $O(n^3)$ procedure just by repeating the method for each m. In our method, however, only one sweep is performed, thus only one data-structure construction is needed.

4. Robust minimum scale regression line

The choice of the robustness (in terms of breakdown point) is still left to the analyst. Of course, if this choice were made in advance, algorithms with lower complexity are available, by adapting those proposed for the LMS.

The availability of *all* minquantile lines allows however an "automatic" choice of the breakdown point. As shown by Carrizosa and Plastria (1992), the bi-criterion problem of finding a line $l_{a,b}$ which both maximizes m (i.e. the proportion of not-rejected data points) and minimizes $Q_m(a, b)$ (measuring within which range the nonrejected data points lie from the regression line), may be solved by inspecting the minquantile lines only. Indeed, consider any $m \in \{N/2, \ldots, N\}$ and $(a, b) \in \mathbb{R}^2$. If we set

$$m^* = W(i: r_i^2(a, b) \leq Q_m(a, b)\}),$$

we see that $m^* \geq m$ and $Q_{m^*}(a, b) = Q_m(a, b)$. By Theorem 1, there exists an m^*/N-minquantile line $l_{\bar{a},\bar{b}}$ such that

$$W(\{i: r_i^2(\bar{a}, \bar{b}) \leq Q_{m^*}(\bar{a}, \bar{b})\}) \geq m^*.$$

Hence, as $Q_{m^*}(\bar{a}, \bar{b}) \leq Q_{m^*}(a, b) = Q_m(a, b)$, the line $l_{\bar{a},\bar{b}}$ produces a smaller error than $l_{a,b}$ and, at the same time, rejects a lower number of observations: $W(\{i: r_i^2(\bar{a}, \bar{b}) > Q_{m^*}(\bar{a}, \bar{b})\})$ instead of $N - m^*$.

However, many different bi-criterion evaluation rules may still be applied. One proposal may be derived from the scale parameter suggested by Rouseeuw (1984) for the LMS regression line, by adapting it to other quantiles in the following way: For any line $l_{a,b}$ and m $(N/2 \leq m < N)$, define the scale parameter $S_m(a, b)$ as

$$S_m(a, b) = c(N, 2)\, Q_m(a, b)^{1/2} / \phi^{-1}((N + m)/2N)$$

and

$$S_N(a, b) = +\infty,$$

where $c(N, 2)$ is the finite-sample correction factor suggested by Rousseuw (1984). Since $S_m(a, b)$ is increasing in $Q_m(a, b)$ and decreasing in m, a regression line with minimal scale parameter will be found by calculating $S_m(a, b)$ for each m using Q_m^*, and selecting the m giving the lowest $S_m(a, b)$; the corresponding m/N-minquantile

Table 4

m	$S_m/c(N, 2)$
5	0.25
*6	0.20
7	0.48
8	0.39
9	0.53
10	$+\infty$

line will give the optimal solution. For instance, for the example of Section 3, one obtains the values (up to the constant $c(N, 2)$, which has no influence in the decision) shown in Table 4. As a result, the line $l_{(4/3,\ 1/6)}$ is chosen, rejecting the points 2, 5 and 6 (weight 4) as outliers.

5. Concluding remarks and extensions

In this paper we have addressed the problem of finding a least quantile of squares line for all quantiles. We have proposed a sweep-line algorithm that runs in polynomial time and allows for the use of frequencies associated with the different observations in the sample.

The knowledge of minquantile lines for all quantiles enables to accomodate the methodology of least median of squares to problems where the portion of outliers is lower than a half. In particular, an automatic choice of the portion of outliers may be obtained by minimizing a scale parameter, as shown in Section 4.

Part of the analysis carries over to the multiple linear regression model

$$Y_i = X_i\theta + \varepsilon_i \quad (i = 1, \ldots, n),$$

where $(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}$ are data points with weights (frequencies) w_i.

From the results of Carrizosa and Plastria (1992), it follows that an α-minquantile hyperplane may be determined for each α by considering all hyperplanes passing through at least p points, and for each set of $p + 1$ data points, all the hyperplanes yielding residuals which are equal in absolute residual $|r|$ but not in sign. For each such hyperplane, one determines the corresponding α-value, by evaluation of the proportion of the data with absolute residuals not exceeding $|r|$. Since there are at most

$$\binom{n}{p} + \binom{n}{p+1}(2^p - 2)$$

such hyperplanes, and each may be evaluated in $O(n)$ time, we obtain an $O(n^{p+2})$ algorithm. Thanks to the sweep-line technique we were able to reduce this complexity to $O(n^3)$ for $p = 2$.

It remains an open question whether similar techniques may be developed for higher dimension.

References

Appa, G. and C. Smith, On L_1 and Chebyshev estimation, *Math. Programming*, **5** (1973) 73–87.

Carrizosa, E. and F. Plastria, On minquantile and maxcovering optimisation, Working paper (Center for Management Informatics, Vrije Universiteit Brussel, Brussels, 1992). Submitted to: *Math. Programming*.

Cook, R.D. and D.M. Hawkins, Comment on unmasking multivariate outliers and leverage points, *J. Amer. Statist. Assoc.*, **85** (1990) 640–644.

Drezner, Z., On minimax optimization problems, *Math. Programming*, **22** (1982) 227–230.

Edelsbrunner, H. and D.L. Souvaine, Computing least median of squares regression lines and guided topological sweep, *J. Amer. Statist. Assoc.*, **85** (1990) 115–119.

Rousseeuw, P.J., Least median of squares regression, *J. Amer. Statist. Assoc.* **79** (1984) 871–880.

Rousseeuw, P.J. and A.M. Leroy, *Robust regression and outlier detection.* (Wiley, New York, 1987).

Souvaine, D.L. and J.M. Steele, Time- and space-efficient algorithms for least median of squares regresson, *J. Amer. Statist. Assoc.*, **82** (1987) 794–801.

Steele, J.M. and W.L. Steiger, Algorithms and complexity for least median of squares regression, *Discrete Appl. Math.* **14** (1986) 93–100.

Stromberg, A.J., Computing the exact value of the least median of squares estimate in multiple regression, Technical Report # 561 (Dept. of Statistics, University of Minnesota, 1991).

Xu, C.W. and W.K. Shiue, Parallel algorithms for least median of squares regression, *Comput. Statist. Data Anal.*, **16** (1993) 349–362.