Annals of Discrete Mathematics 16 (1982) 201–211

TWO LINES LEAST SQUARES

A.K. LENSTRA and J.K. LENSTRA
Mathematical Centre, Amsterdam, The Netherlands

A.H.G. RINNOOY KAN
Econometric Institute, Erasmus University, Rotterdam, The Netherlands

T.J. WANSBEEK
Netherlands Central Bureau of Statistics, Voorburg, The Netherlands

It is well known that the standard single line least squares problem for n points in the plane is solvable in linear time. We consider two generalizations of this problem, in which two lines have to be constructed in such a way that, after a certain assignment of each point to one of the lines, the sum of squared vertical distances is minimal. Polynomial time algorithms for the solution of these problems are presented.

1. Introduction

Given a set $P = \{(x_j, y_j) \mid x_j, y_j \in \mathbb{R}, j = 1, \ldots, n\}$ of n points in the plane, the *single line least squares* (1LLS) problem is to find a line

$$l(x) = ax + b$$

such that

$$\sum_{(x,y)\in P} (l(x) - y)^2$$

is minimized. As is well known, the solution is given by

$$a = \frac{n \sum_P xy - \sum_P x \sum_P y}{n \sum_P x^2 - \left(\sum_P x\right)^2}, \qquad b = \frac{1}{n}\left(\sum_P y - a \sum_P x\right),$$

and the line l can thus be determined in $O(n)$ time.

In this paper we shall study two variations on this problem, in which *two lines*

have to be constructed in such a way that, after a certain assignment of each point to one of the lines, the sum of the squared vertical distances is minimal.

The first and most obvious variation, the *two lines least squares* (2LLS) problem, is to find a set $Q \subseteq P$ and two lines

$$l_1(x) = a_1x + b_1\,, \qquad l_2(x) = a_2x + b_2$$

such that

$$\sum_{(x,y)\in Q} (l_1(x) - y)^2 + \sum_{(x,y)\in \bar{Q}} (l_2(x) - y)^2 \tag{1}$$

is minimized (cf. Fig. 1).

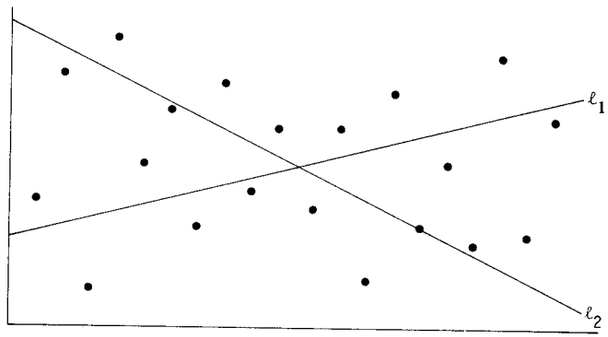

Fig. 1. 2LLS.

Another variation, the *bent line least squares* (BLLS) problem, is to find a *breakpoint* x^* and a *bent line*

$$l^*(x) = \begin{cases} a_1(x - x^*) + b & (x \leq x^*)\,, \\ a_2(x - x^*) + b & (x > x^*) \end{cases}$$

such that

$$\sum_{(x,y)\in P} (l^*(x) - y)^2 \tag{2}$$

is minimized (cf. Fig. 2). Note that the bent line l^* is continuous in the breakpoint x^*.

These types of problems may well arise in many situations in applied statistics, e.g., biometrics and econometrics. If the observations belong to either

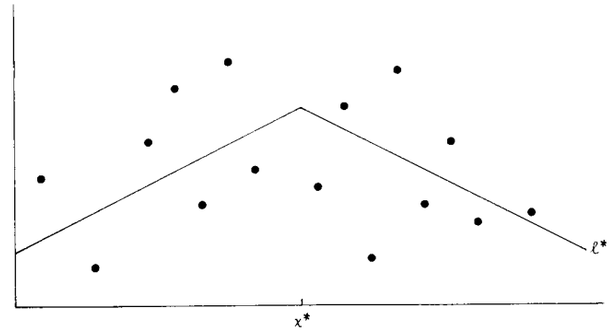

Fig. 2. BLLS.

of two regression regimes, with little or no classifying information being available, a two lines or bent line model may be more appropriate than the single line one.

The 2LLS problem occurs for example in the case of markets in disequilibrium, where data are available on prices and supply or demand, while it is not known whether each particular observation is generated by the supply curve or by the demand curve. This problem was considered by Fair and Jaffee [3], who suggested to obtain a solution by exhaustive search over all sets $Q \subseteq P$; this approach requires exponential time. In Section 2 we develop a polynomial algorithm to minimize (1) in $O(n^3)$ time.

The BLLS problem arises when, in the above situation, the observations correspond to the minimum of supply and demand. This is an example of a class of problems where the regression parameters may change as an independent variable increases. In this area an extensive literature has appeared; see [13] for a bibliography. The seminal paper on the BLLS problem is by Hudson [10], who also considered several generalizations involving multiple breakpoints. In Section 3 we use his results to minimize (2) in $O(n \log n)$ time.

Finally, in Section 4 we comment on the statistical properties of our estimators, on related previous work, and on possible extensions of our algorithms.

2. The two lines least squares problem

To solve the 2LLS problem, we start by observing two obvious properties of an optimal solution. First, the lines l_1 and l_2 are the ordinary 1LLS solutions for the sets Q and $\bar{Q}$ respectively. Secondly, the set $Q \subseteq P$ evidently satisfies

$$Q = \{(x, y) \,|\, |l_1(x) - y| \leq |l_2(x) - y|\}\,. \tag{3}$$

The partition of P into Q and $\bar{Q}$, given the lines l_1 and l_2, is therefore characterized by the set $Q_0 \subset \bar{\mathbb{R}}^2$ of points for which equality holds in (3):

$$\begin{aligned} Q_0 &= \{(x, y) \mid |l_1(x) - y| = |l_2(x) - y|\} \\ &= \{(x, y) \mid l_1(x) - y = l_2(x) - y\} \cup \{(x, y) \mid l_1(x) - y = -l_2(x) + y\} \\ &= \{(x_0, y)\} \cup \{(x, l_0(x))\} \end{aligned}$$

where

$$x_0 = \begin{cases} \dfrac{b_2 - b_1}{a_1 - a_2} & \text{if } a_1 \neq a_2, \\ +\infty & \text{if } a_1 = a_2, b_1 < b_2, \\ -\infty & \text{if } a_1 = a_2, b_1 \geq b_2, \end{cases}$$

$$l_0(x) = \tfrac{1}{2}(l_1(x) + l_2(x)).$$

Thus, the set Q_0 consists of two lines: the vertical line through the x-coordinate x_0 of the intersection of l_1 and l_2, and the median l_0 of l_1 and l_2. Under the assumption that $a_1 \geq a_2$, the set Q can now be rewritten as

$$Q = \{(x, y) \mid x \leq x_0, y \leq l_0(x)\} \cup \{(x, y) \mid x \geq x_0, y \geq l_0(x)\} \tag{4}$$

(cf. Fig. 3).

We conclude that we may restrict our attention to feasible solutions for which the partition of P into Q and $\bar{Q}$ is defined by a value x_0 and a line l_0 as in (4), and that a solution that is optimal with respect to such a partition is given by the 1LLS solutions for Q and $\bar{Q}$. (Note, however, that solutions satisfying (4) do not necessarily satisfy (3).) It follows that the 2LLS problem can be solved by generating all partitions of the above type, by solving two

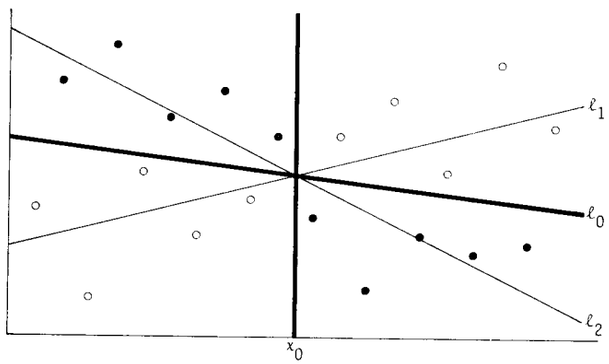

Fig. 3. The sets Q_0 (the heavy lines) and Q (the open points).

1LLS problems for each of them, and by selecting a solution for which the optimality criterion has minimal value.

What is the total number of different partitions of P of the type characterized by (4)? We may assume (if necessary, after a small perturbation) that no two points from P have the same x-coordinate and that no three points from P lie on the same line. First, it is clear that P can be separated into two subsets by a vertical line in n different ways, corresponding to the choices $x_0 = x_j$ $(j = 1, \ldots, n)$. Secondly, we claim that there is a one-to-one correspondence between separations of P by an arbitrary line and pairs of points from P, where the latter can be chosen in $\frac{1}{2}n(n-1)$ different ways. It follows that we have to consider no more than $\frac{1}{2}n^2(n-1)$ different partitions of P.

To see why the above claim is true, consider a separation of P into Q and $\bar{Q}$ by an arbitrary line l. Let $C(Q)$ and $C(\bar{Q})$ denote the convex hulls of Q and $\bar{Q}$ respectively. Since l also separates $C(Q)$ and $C(\bar{Q})$, there exists a unique line l_0 such that

(a) $y \geq l_0(x)$ for all points (x, y) in one of the convex hulls, say $C(Q)$,

(b) $y \leq l_0(x)$ for all points (x, y) in the other convex hull $C(\bar{Q})$,

(c) there are two points $p' = (x', y') \in C(Q)$ and $\bar{p}' = (\bar{x}', \bar{y}') \in C(\bar{Q})$ with $x' > \bar{x}'$ such that $y' = l_0(x')$ and $\bar{y}' = l_0(\bar{x}')$

(cf. Fig. 4); l_0 is obtained from l by turning l counterclockwise until l is tangent to both $C(Q)$ and $C(\bar{Q})$. Moreover, the assumption that no three points from P lie on the same line implies that $p' \in Q$ and $\bar{p}' \in \bar{Q}$.

This establishes one part of the correspondence. Conversely, consider two points $p'' = (x'', y'') \in P$ and $\bar{p}'' = (\bar{x}'', \bar{y}'') \in P$ with $x'' > \bar{x}''$. Let l_0 denote the line through p'' and $\bar{p}''$. A separation of P into Q and $\bar{Q}$ is now obtained by defining

$$\begin{aligned} Q &= \{(x, y) \mid y > l_0(x)\} \cup \{(x'', y'')\}, \\ \bar{Q} &= \{(x, y) \mid y < l_0(x)\} \cup \{(\bar{x}'', \bar{y}'')\}; \end{aligned} \tag{5}$$

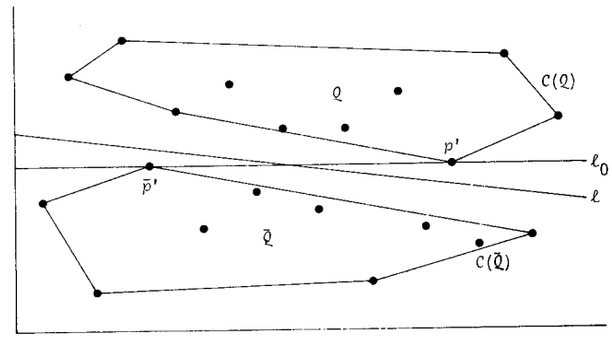

Fig. 4.

with respect to this separation, l_0 satisfies the conditions (a), (b) and (c) with $p' = p''$ and $\bar{p}' = \bar{p}''$.

If we start from a separation of P by an arbitrary line, apply the first transformation to find a pair of points and next apply the second transformation to this pair, we obtain the same separation once again. Thus, the correspondence is one-to-one, as claimed.

Our 2LLS algorithm can now be described as follows.

First, we renumber the points from P according to increasing x-coordinate in $O(n \log n)$ time.

Secondly, we consider all pairs of points $\{(x'', y''), (\bar{x}'', \bar{y}'')\}$ with $x'' > \bar{x}''$ in succession. For each such pair, we start by determining the sets Q and $\bar{Q}$ as in (5), calculating the partial sums

$$\sum_Q x, \quad \sum_Q y, \quad \sum_Q x^2, \quad \sum_Q xy, \tag{6}$$

and solving the 1LLS problem for Q and $\bar{Q}$; next, for $j = 1, \ldots, n-1$, we repeat this procedure for the partition induced by the vertical separation corresponding to $x_0 = x_j$. The sets Q and $\bar{Q}$ and the partial sums can be determined according to (5) and (6) in $O(n)$ time, and they can be adjusted for each successive value of x_0 in constant time. Given the partial sums (6), each 1LLS problem is solvable in constant time. It follows that this step requires $O(n)$ time for each pair of points, and $O(n^3)$ time overall.

Finally, we select the solution for which the optimality criterion achieves its minimal value. The entire procedure requires $O(n^3)$ time.

There are various ways in which the implementation of this algorithm can be improved that do not, however, reduce the running time by more than a constant factor. We note that, in general, it is impossible to generate all pairs of points from P in such a way that the partial sums (6) for a given separation can be derived from those for the preceding separation by interchanging a single point.

3. The bent line least squares problem

To solve the BLLS problem, we first renumber the points from P according to nondecreasing x-coordinate in $O(n \log n)$ time. We now reformulate the problem as follows: determine an index $k \in \{1, \ldots, n-1\}$ and two lines

$$l_1(x) = a_1 x + b_1, \qquad l_2(x) = a_2 x + b_2,$$

subject to a constraint on the x-coordinate of their intersection:

$$x_k \leq \frac{b_2 - b_1}{a_1 - a_2} \leq x_{k+1} \quad \text{if } a_1 \neq a_2, \tag{7}$$

$$b_1 = b_2 \quad \text{if } a_1 = a_2,$$

such that

$$\sum_{j=1}^{k} (l_1(x_j) - y_j)^2 + \sum_{j=k+1}^{n} (l_2(x_j) - y_j)^2$$

is minimized. Since the breakpoint x^* has to be located in one of the intervals $[x_k, x_{k+1}]$, both formulations are clearly equivalent.

We shall show how, after an $O(n)$ initialization, the determination of l_1 and l_2 for any given value of k can be carried out in constant time. To select the optimal value of k, this has to be done for $k = 1, \ldots, n-1$, and the entire procedure requires $O(n \log n)$ time, as announced. Note that, apart from sorting the set P, our BLLS algorithm requires linear time off-line, whereas the 1LLS problem is solvable in linear time on-line.

We start by calculating the partial sums

$$\sum_{j=1}^{k} x_j, \quad \sum_{j=1}^{k} y_j, \quad \sum_{j=1}^{k} x_j^2, \quad \sum_{j=1}^{k} x_j y_j \tag{8}$$

for $k = 1, \ldots, n$ in $O(n)$ time. Next, for a given value of k, we solve the ordinary 1LLS problem for $\{(x_j, y_j) | j = 1, \ldots, k\}$ and for $\{(x_j, y_j) | j = k+1, \ldots, n\}$ to find two lines l_1' and l_2' respectively; in view of the availability of the partial sums (8), this requires constant time.

In the x-coordinate of the intersection of l_1' and l_2' lies in $[x_k, x_{k+1}]$ or if $l_1' = l_2'$, the pair (l_1', l_2') defines a feasible and optimal solution with respect to the given value of k, and we are finished.

If this is not the case, we claim that the optimal pair of lines has its intersection on either x_k or x_{k+1}, i.e., one of the inequalities in (7) has to be satisfied as an equality (cf. [10]). To see why, compare the infeasible pair (l_1', l_2') to any feasible solution (l_1, l_2) for which both inequalities in (7) are strictly satisfied. Suppose that $l_2' \neq l_2$ and that these lines intersect in a point p (cf. Fig. 5). Then any line l_2'' through p whose direction is between the directions of l_2' and l_2 yields an improvement over l_2, since the 1LLS optimality criterion is quadratic and convex in the parameters a and b of a line $ax + b$ and since l_2'' is closer than l_2 to the optimal line l_2'. If moreover the x-coordinate of the intersection of l_1 and l_2'' lies in $[x_k, x_{k+1}]$, the pair (l_1, l_2'') defines a feasible solution that is better than (l_1, l_2), and the latter solution cannot be optimal. The argument is easily extended to the case that l_2' and l_2 are parallel.

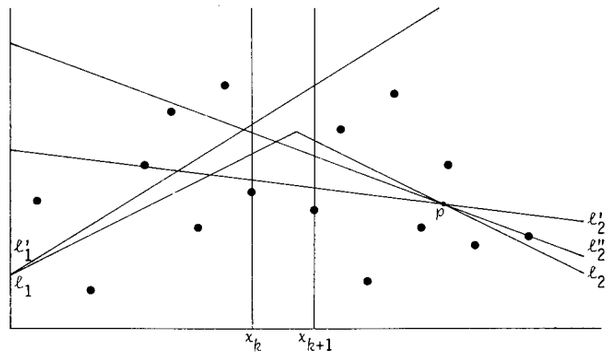

Fig. 5.

All we have to do in this case, therefore, is to determine two optimal bent lines l^* with fixed breakpoints $x^* = x_k$ and $x^* = x_{k+1}$ respectively and to select whichever of the two is best. The determination of l^* for a given value of x^* can be carried out in constant time, as follows. Starting from (2), we rewrite the optimality criterion as

$$\sum_{j=1}^{k} (a_1(x_j - x^*) + b - y_j)^2 + \sum_{j=k+1}^{n} (a_2(x_j - x^*) + b - y_j)^2 .$$

Taking first derivatives with respect to the parameters a_1, a_2 and b, we obtain the following linear equations:

$$\begin{cases} a_1 \sum_{j=1}^{k} (x_j - x^*)^2 + b \sum_{j=1}^{k} (x_j - x^*) = \sum_{j=1}^{k} (x_j - x^*) y_j , \\ a_2 \sum_{j=k+1}^{n} (x_j - x^*)^2 + b \sum_{j=k+1}^{n} (x_j - x^*) = \sum_{j=k+1}^{n} (x_j - x^*) y_j , \\ a_1 \sum_{j=1}^{k} (x_j - x^*) + a_2 \sum_{j=k+1}^{n} (x_j - x^*) + bn = \sum_{j=1}^{n} y_j . \end{cases}$$

Given the partial sums (8), this system is solvable in constant time to yield the required values of a_1, a_2 and b.

It is again an easy matter to conceive of various improvements in the implementation of this algorithm, when the above system has to be solved $O(n)$ times with only slight intermediate changes in the coefficients (cf. [10]). The worst case running time, however, will be affected by no more than a constant factor as a result.

4. Concluding remarks

The 2LLS and BLLS problems have so far been taken to be purely deterministic problems. Keeping in line with regression analysis tradition, we now assume that the parameters to be determined (Q, a_1, a_2, b_1, b_2 in 2LLS, x^*, a_1, a_2, b in BLLS) have true but unknown values that are to be estimated and raise the question what the statistical properties are of the outcomes of the minimization problems. With respect to the stochastic nature of the data generating process, we make the simple assumption that the x_j are non-stochastic and that the observations (x_j, y_j) satisfying the unknown linear relations are subject to vertical disturbances which are drawn from a normal distribution with zero expectation and constant variance. Under this interpretation, our *least squares* estimators are also *maximum likelihood* estimators.

The estimators in the 2LLS problem do probably not have any other desirable statistical properties. The following simple example shows that they are not in general consistent (which is not to say that consistent 2LLS estimators do not exist). Suppose that the true model is given by

$$y_1 = 1 + u , \qquad y_2 = -1 + u ,$$

where u has a standard normal distribution with expectation 0, variance 1 and density function ϕ, i.e., $a_1 = a_2 = 0$, $b_1 = 1$, $b_2 = -1$. Suppose further that the independent variable x assumes at least three equidistant values, and that, when n goes to infinity, either regime accounts for half of the data with an equal number of data for each value of x. Then it is easy to see that the estimators $\hat{a}_1$, $\hat{a}_2$, $\hat{b}_1$, $\hat{b}_2$ of a_1, a_2, b_1, b_2 satisfy

$$\operatorname{plim} \hat{a}_1 = a_1 , \qquad \operatorname{plim} \hat{b}_1 = b_1 + c ,$$

$$\operatorname{plim} \hat{a}_2 = a_2 , \qquad \operatorname{plim} \hat{b}_2 = b_2 - c ,$$

where

$$c = \int_1^{\infty} (2u - 2)\phi(u)\, du \approx 0.167 .$$

In this situation, the estimators of b_1 and b_2 are bounded away from the true values, due to persistent misclassification. (The assertion by Fair and Jaffee [3, p. 500] as to the consistency of the 2LLS estimators is incorrect. They neglect the implicit presence of dummy variables assigning observations to regimes. The fact that the number of such variables goes to infinity with n invalidates standard maximum likelihood theory.)

Statistical properties of the estimators in the BLLS problem follow from

results due to Feder [4], who considered the case of multiple breakpoints and a more general functional specification for the regression segments. As the finite sample distribution of the estimators is intractable, he concentrated on their asymptotic distribution. His results imply that, if $a_1 \neq a_2$, the estimators of x^*, a_1, a_2, b are consistent and have a certain multinormal asymptotic distribution [4, pp. 71, 77]. Feder and Sylwester [5] already established the asymptotic normality of the estimator of x^*. Hinkley [9] showed that the asymptotic distribution is not a good approximation of the small sample distribution and presented an alternative that performs better in small samples.

Due to its combinatorial nature and the statistical intractability of its estimators, the 2LLS problem has not spawned much research. Mustonen [12] and Hermann [8] considered the multidimensional case, in which two hyperplanes [8] or, more generally, two functions of a given form [12] have to be constructed. In view of the nonlinearity of the optimality criterion, they suggested to obtain an approximate solution by iterative numerical methods. We conjecture that our combinatorial approach can be extended to yield an optimal solution in polynomial time, as long as the dimension and the number of hyperplanes are constants. However, this generalization is likely to be extremely intricate.

Another reason for the relative neglect of the 2LLS problem is that, as far as disequilibrium econometrics is concerned, economic theory can be invoked to further model the regime choice mechanism. Fair and Jaffee [3] made several additional assumptions, the simplest one being that a price increase points to an excess demand regime and a price fall to an excess supply regime. A huge literature has developed in this direction (e.g. [7, 2]), a common trait being the use of nonlinear rather than combinatorial optimization methods. Estimators with favorable asymptotic properties have been derived (e.g. [11]).

It has been mentioned that the BLLS problem in which multiple breakpoints are allowed has been considered by Hudson [10]. In contrast to his analytical approach, Bellman and Roth [1] proposed a dynamic programming recursion to obtain a solution which is approximate in the sense that the breakpoints are to be chosen on a grid. The running time of their method depends heavily on the grid width; it is linear in the number of breakpoints, but only pseudo-polynomial [6] in the data (x_j, y_j). For the many variations on and extensions of the BLLS problem, the reader is again referred to the bibliography in [13].

The purpose of this paper has been to analyze the computational complexity of two combinatorial optimization problems arising in statistical analysis. Other results of a similar nature can be found in the work of Shamos [14]. These examples should serve to demonstrate the potential value of research in this interface area.

References

[1] R. Bellman and R. Roth, Curve fitting by segmented straight lines, J. Amer. Statist. Assoc. 64 (1969) 1079–1084.

[2] R.J. Bowden, The Econometrics of Disequilibrium (North-Holland, Amsterdam, 1978).

[3] R.C. Fair and D.M. Jaffee, Methods of estimation for markets in disequilibrium, Econometrica 40 (1972) 497–514.

[4] P.I. Feder, On asymptotic distribution theory in segmented regression problems—identified case, Ann. Statist. 3 (1975) 49–83.

[5] P.I. Feder and D.L. Sylwester, On the asymptotic theory of least squares estimation in segmented regression: identified case (abstract), Ann. Math. Statist. 39 (1968) 1362.

[6] M.R. Garey and D.S. Johnson, 'Strong' NP-completeness results: motivation, examples and implications, J. Assoc. Comput. Mach. 25 (1978) 499–508.

[7] S.M. Goldfeld and R.E. Quandt, The estimation of structural shifts by switching regressions, Ann. Econ. Soc. Measurement 2 (1973) 475–485.

[8] J. Hermann, Data-analytic concepts for isolating intersecting reaction surfaces, Institute for Econometrics and Operations Research, University of Bonn, 1980.

[9] D.V. Hinkley, Inference about the intersection in two-phase regression, Biometrika 56 (1969) 495–504.

[10] D.J. Hudson, Fitting segmented curves whose join points have to be estimated, J. Amer. Statist. Assoc. 61 (1966) 1097–1129.

[11] N.M. Kiefer, Discrete parameter variation: efficient estimation of a switching regression model, Econometrica 46 (1978) 427–434.

[12] S. Mustonen, Digression analysis: fitting alternative regression models to heterogeneous data, in: L.C.A. Corsten and J. Hermans, eds., COMPSTAT '78: Proc. Computational Statistics (Physica-Verlag, Vienna, 1978) pp. 95–101.

[13] S.A. Shaban, Change point problem and two-phase regression: an annotated bibliography, Internat. Statist. Rev. 48 (1980) 83–93.

[14] M.I. Shamos, Geometry and statistics: problems at the interface, in: J.F. Traub, ed., Algorithms and Complexity: New Directions and Recent Results (Academic Press, New York, 1976) pp. 251–280.