

L₁ Norm Based Computational Algorithms

Bijan Bidabad¹

Keywords: L₁ norm, Regression, Algorithm, Computer

Abstract

This paper gives a rather general review of the L₁ norm algorithms. The chronology and historical development of the L₁ norm estimation theory for the period of 1632-1928 will be surveyed and the algorithms belonging to the after 1928 period will be categorized into three main classes of direct descent, simplex type and other algorithms.

Historical development (1632-1928)

The origin of L₁ norm estimation may be traced back to Galilei (1632). In determining the position of a newly discovered star, he proposed the least possible correction in order to obtain a reliable result (see, Ronchetti (1987) for some direct quotations). Boscovich (1757) for the first time formulated and applied the minimum sum of absolute errors for obtaining the best fitting line given three or more pairs of observations for a simple two variable regression model. He also restricts the line to pass through the means of the observation points. That is,

$$\begin{aligned} \min_{\beta_0, \beta_1} : & \sum_{i=1}^n |y_i - \beta_0 - \beta_1 x_{i1}| \\ \text{s.to: } & \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1}) = 0 \end{aligned} \quad (1)$$

Boscovich (1760) gives a simple geometrical solution to his previous suggestion. This paper has been discussed by Eisenhart (1961) and Sheynin (1973). In a manuscript Boscovich poses the problem to Simpson and Simpson gives an analytical solution to the problem (see, Stigler (1984)).

Laplace (1793) provides an algebraic formulation of an algorithm for the L₁ norm regression line which passes through the centroid of observations. In Laplace (1799), extension of L₁ norm regression to observations with different weights has also been discussed. Prony (1804) gives a geometric interpretation of Laplace's (1799) method and compares it with other methods through an example. Svanberg (1805) applies Laplace's

¹ Research professor of economics, Monetary and Banking Research Academy, Central Bank of Iran, Pasdaran, Zarabkhaneh, Tehran, 16619, Iran. bijan_bidabad@msn.com.

I should express my sincere thanks to Professor Yadolah Dodge from Neuchatel University of Switzerland who taught me many things about L₁ norm when he was my Ph.D.dissertation advisor.

method in determining a meridian arc and Von Lindenau (1806) uses this method in determination of the elliptic meridian.

Gauss (1809) suggests the minimization of sum of absolute errors without constraint. He concludes that this criterion necessarily sets m of the residuals equal to zero, where m is number of parameters and further, the solution obtained by this method is not changed if the value of dependent variable is increased or decreased without changing the sign of the residual. This conclusion, is recently discussed by Narula and Wellington (1985). He also noted that Boscovich or Laplace estimators which minimize the sum of absolute residuals with zero sum of residuals constraint, necessarily set $m-1$ of the residuals equal to zero (see, Stigler (1981), Farebrother (1987b)).

Mathieu (1816) used Laplace's method to compute the eccentricity of the earth. Van Beeck-Calkoen (1816) advocates the using of the least absolute values criterion in fitting curvilinear equation obtained by using powers of the independent variable.

Laplace (1818) adapted Boscovich's criterion again and gave an algebraic procedure (see, Farebrother (1987b)). Let \bar{x}_1 and \bar{y} be the means of x_{i1} and y_i then,

$$\beta_0 = \bar{y} - \beta_1 \bar{x}_1 \tag{2}$$

Value of β_1 is found by,

$$\min_{\beta_1} S = \sum_{i=1}^n |y_i - \beta_1 x_{i1}| \tag{3}$$

where x_{i1} and y_i are deviations of x_{i1} and y_i from their means respectively. By rearranging the observations in descending order of y_i/x_{i1} values, Laplace notes that S is infinite when β_1 is infinite and decreases as β_1 is reduced. β_1 reaches the critical value y_t/x_{t1} when it again begins to increase. This critical value of β_1 is determined when,

$$\sum_{i=1}^{t-1} |x_{i1}| < \frac{1}{2} \sum_{i=1}^n |x_{i1}| \leq \sum_{i=1}^t |x_{i1}| \tag{4}$$

This procedure to find β_1 is called weighted median, and has been used in many other algorithms such as Rhodes (1930), Singleton (1940), Karst (1958), Bloomfield and Steiger (1980), Bidabad (1987a,b,88a,b,89a,b) later. Bidabad (1987a,88a,89a,b) derives the condition (4) via discrete differentiation method.

Fourier (1824) formulates least absolute residuals regression as what we would now call linear programming; that is minimization of a linear objective function subject to linear inequality constraints.

Edgeworth (1883) presents a philosophical discussion on differences between minimizing mean square errors and mean absolute errors. Edgeworth (1887a,b) proposed a simple method for choosing the regression parameters. By fixing $m-1$ of the parameters, he used Laplace's procedure to determine the optimal value of the remaining parameter. Repeating this operation for a range of values for $m-1$ fixed parameters he obtained a set of results for each of m possible choices of the free parameters. Edgeworth drops the restriction of passing through the centroid of data. Turner (1887) discusses the problem of non unique solutions under the least absolute error criterion as a graphical variant of Edgeworth (1887a) as a possible drawback to the method. Edgeworth (1888) replies to Turner's criticism by proposing a second method for choosing the two parameters of least absolute error regression of a simple linear model which makes no use of the median loci of his first method. Edgeworth, in this paper, followed Turner's suggestion for graphical analysis of steps to reach the minimum solution.

Before referring to double median method of Edgeworth (1923), it should be noted that Bowley (1902) completes the Edgeworth's (1902) paper by a variant of double median

method which presented after him by Edgeworth (1923). This variant ignores the weights attached to errors.

Edgeworth (1923) discussed the more general problem of estimating the simple linear regression parameters by minimizing the weighted sum of the absolute residuals. He restates the rationale for the method and illustrates its usage through several examples. He also considers the non unique solution problem. His contribution is called double median method.

Estienne (1926-28) proposes replacing the classical theory of errors of data based on least squares with what he calls a rational theory based on the least absolute residual procedure. Bowley (1928) summarizes the Edgeworth's contributions to mathematical statistics which includes his work on L_1 norm regression. Dufton (1928) also gives a graphical method of fitting a regression line.

Farebrother (1987b) summarizes the important contributions to L_1 norm regression for the period of 1793-1930. For more references see also Crocker (1969), Harter (1974a,b,75a,b,c,76), Dielman (1984). Up to 1928, all algorithms had been proposed for simple linear regression. Though some of them use algebraic propositions, are not so organized to handle multiple L_1 norm regression problem. In the next section we will discuss the more elaborated computational methods for simple and multiple L_1 norm regressions not in a chronological sense; because many digressions have been occurred. We may denote the period of after 1928 the time of modern algorithms in the subject of L_1 norm regression.

Computational algorithms

Although, a closed form of the solution of L_1 norm regression has not been derived yet, many algorithms have been proposed to minimize its objective function (see, Cheney (1966), Chambers (1977), Dielman and Pfaffenberger (1982,84)). Generally, we can classify all L_1 norm algorithms in three major categories as,

- i) Direct descent algorithms
- ii) Simplex type algorithms
- iii) Other algorithms

which will be discussed in the following sections sequentially.

Direct descent algorithms

The essence of the algorithms which fall within this category is finding an steep path to descend down the polyhedron of the L_1 norm regression objective function. Although the Laplace's method (explained herein before) is a special type of direct descent algorithms; origin of this procedure in the area of L_1 norm can be traced back to the algorithms of Edgeworth which were explained in the previous section.

Rhodes (1930) found Edgeworth's graphical solution laborious, therefore, he suggested an alternative method for general linear model which may be summarized as follows (see, Farebrother (1987b)). Suppose, we have n equations with $m < n$ unknown parameters. To find L_1 norm solution of this overdetermined system of equations;

Step 1) Select $m-1$ equations arbitrarily.

Step 2) Solve these equations for $m-1$ parameters.

Step 3) Estimate the remaining m^{th} parameter by the Laplace's method.

Step 4) Recognize the resulting equation in step 3 and add it to the $m-1$ equations in step 1.

Step 5) If the set of m equations has recurred m times then stop; otherwise discard the oldest equation and go to step 2.

Rhodes (1930) explained his algorithm by an example and did not give any proof for convergence. Bruen (1938) reviews the L_1 norm regression methods presented by earlier authors. He also compares L_1 , L_2 and L_∞ norms regressions.

Singleton (1940) applied Cauchy's steepest descent method (see, Panik (1976)) for the general linear L_1 norm regression. In this paper a geometrical interpretation of gradient on L_1 norm polyhedron and some theorems about existence and uniqueness of solution and convexity property all were given. Although the paper has not been clearly written, the following steps summarize his algorithm.

Step 1) Select a point $\beta_j^{(0)}, j=1, \dots, m$ arbitrarily.

Step 2) Determine the gradient,

$$g_j^{(0)} = -\sum_{i=1}^n \text{sgn}(u_i \wedge | \beta_j^{(0)}, j=1, \dots, m) x_{ij}.$$

Step 3) Compute, $w_i^{(0)} = \sum_{j=1}^m x_{ij} g_j^{(0)}, z_i^{(0)} = y_i - \sum_{j=1}^m x_{ij} \beta_j^{(0)}$.

Step 4) Determine the value of $t^{(0)}$ as weighted median of $\sum_{i=1}^n |w_i t - z_i|$ by Laplace's method.

The t value is the length of movement along direction of the gradient.

Step 5) Compute $\beta_j^{(1)} = \beta_j^{(0)} + g_j^{(0)} t^{(0)}$.

Step 6) Test the optimality condition. Singleton gives a condition to stop, but it is not quite clear. In this step any other criterion relevant to L_1 norm function may be displaced.

Step 7) This step is not well defined by Singleton. In this phase he tries to choose the best gradient among usable gradients. Without this step algorithm is still operational, because, the steps are all standards of Cauchy steepest descent method and instead of choosing the best gradient we can proceed by going to step 2.

Bejar (1956,57) focuses on consideration of residuals rather than on the vector of parameters. He puts forth a procedure with the essence of Rhodes (1930). However, he is concerned with two and three parameter linear models.

Karst (1958) gives an expository paper for one and two parameter regression models. In his paper, Karst without referring to previous literature actually reaches to the Laplace proposition to solve the one parameter restricted linear model and for the two parameter model, he proposed an algorithm similar to that of Rhodes (1930). His viewpoint is both geometrical and algebraic and no proof of convergence for his iterative method is offered. Sadoski (1974) uses a simple "bubble sort" procedure and implements Karst algorithm in Fortran. Sposito (1976) pointed out that the Sadoski's program may not converge in general. Sposito and Smith (1976) offered another algorithm to remove this problem. Farebrother (1987c) recodes Sadoski's implementation in Pascal language with some improvement such as applying "straight insert sort".

Usow (1967b) presents an algorithm for L_1 norm approximation for discrete data and proves that it converges in a finite number of steps. A similar algorithm on L_1 norm approximation for continuous data is given by Usow (1967a). Given the function $f(\mathbf{x})$ defined on a finite subset $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ of an interval on the real line and also linearly independent continuous functions $\Phi_1(\mathbf{x}), \dots, \Phi_m(\mathbf{x})$ where $m < n$; consider the "polynomial"

$L(\mathbf{B}, \mathbf{x}) = \sum_{j=1}^m \beta_j \Phi_j(\mathbf{x})$. In Usow (1967b) the following function is to be minimized:

$$\min_{\mathbf{B}} S(\mathbf{B}) = \sum_{i=1}^n |L(\mathbf{B}, \mathbf{x}_i) - f(\mathbf{x}_i)| = \sum_{i=1}^n \left| \sum_{j=1}^m \beta_j \Phi_j(\mathbf{x}_i) - f(\mathbf{x}_i) \right| \quad (5)$$

where \mathbf{B} is vector of size m . The above form is general; if we let $f(\mathbf{x}_i) \equiv y_i$ and $\Phi_j(\mathbf{x}_i) \equiv x_{ij}$ the function S becomes the standard linear regression objective function. Now let the set K be:

$$K = \{(\mathbf{B}, d) \mid (\mathbf{B}, d) \in \mathbb{E}^{m+1}, S \leq d\} \quad (6)$$

Then K is a convex polytope, the vertices of which occur only when $L(\mathbf{B}, \mathbf{x}) - f(\mathbf{x})$ is zero at m or more points of \mathbf{X} . The Usow's algorithm is to descend on K from vertex to vertex along connecting edges of the polytope in such a way that certain intermediate vertices are by-passed. This descent continues until the lowest vertex (\mathbf{B}^*, d^*) is reached. To clarify the algorithm assume that we are at the vertex (\mathbf{B}^k, d^k) on K and the polynomial $L(\mathbf{B}^k, \mathbf{x})$ interpolates m points of \mathbf{X} denoted by $\mathbf{U}^k = \{\mathbf{u}_1^k, \dots, \mathbf{u}_m^k\}$. Thus,

$$L(\mathbf{F}^k, \mathbf{x}) = \sum_{i=1}^m f(\mathbf{u}_i^k) \pi_i(\mathbf{x}) \tag{7}$$

where,

$$\mathbf{F}^k = (f(\mathbf{u}_1^k), \dots, f(\mathbf{u}_m^k)) \tag{8}$$

$$\pi_i(\mathbf{x}) = \sum_{j=1}^m a_j^i \Phi_j(\mathbf{x}) \quad i=1, \dots, m \tag{9}$$

The m coefficients a_{ji} are calculated as follows. Form the matrix $(\Phi_j(\mathbf{u}_i^k))$ for $i, j=1, \dots, m$. Let $\boldsymbol{\pi}(\mathbf{x})$ and $\boldsymbol{\Phi}(\mathbf{x})$ be two $m \times 1$ vectors for any $\mathbf{x} \in \mathbf{X}$ whose elements are $\pi_i(\mathbf{x})$ and $\Phi_i(\mathbf{x})$ for $i=1, \dots, m$ respectively. Hence, we can derive,

$$\boldsymbol{\pi}(\mathbf{x}) = [(\Phi_j(\mathbf{u}_i^k))^T]^{-1} \boldsymbol{\Phi}(\mathbf{x}) \tag{10}$$

a_j^i 's are the elements of the i^{th} row of the matrix $[(\Phi_j(\mathbf{u}_i^k))^T]^{-1}$, where the superscript T denotes transposition. Let \mathbf{e}_i be a zero vector of size m where its i^{th} element is equal to one. Then if for some δ , $S(\mathbf{F}^k - \delta \mathbf{e}_i) < S(\mathbf{F}^k)$, there is a T_j such that $T_j \delta > 0$ and $S(\mathbf{F}^k - T_j \mathbf{e}_i) < S(\mathbf{F}^k)$. Also,

$$S(\mathbf{F}^k - T_j \mathbf{e}_i) = \min \{S(\mathbf{F}^k - t \mathbf{e}_i)\} \tag{11}$$

and $((\mathbf{F}^k - T_j \mathbf{e}_i), S(\mathbf{F}^k - T_j \mathbf{e}_i))$ is a vertex. On the other hand, a point $\mathbf{u}_i^k \in \mathbf{U}^k$ may be replaced by a point $\mathbf{u}_i^{k+1} \in \{\mathbf{X} - \mathbf{U}^k\}$ such that the polynomial $L(\mathbf{B}^k, \mathbf{x})$ interpolating $\mathbf{U}_i^k = \{\mathbf{u}_1^k, \dots, \mathbf{u}_i^{k+1}, \dots, \mathbf{u}_m^k\}$ and $S(\mathbf{B}^{ki}) < S(\mathbf{B}^k)$. $S(\mathbf{B}^{ki})$ is the minimum of all norms obtained if \mathbf{u}_i^k were replaced by the different points of the set $\{\mathbf{X} - \mathbf{U}^k\}$, as indicated by the above relation.

In going from vertex $(\mathbf{B}^k, S(\mathbf{B}^k))$ to $(\mathbf{B}^{ki}, S(\mathbf{B}^{ki}))$, one or more vertices on K might have been by-passed. The nearest vertex to $(\mathbf{F}^k, S(\mathbf{F}^k))$ and below it on the edge parallel to the i^{th} parameter space coordinate axis, say the vertex $((\mathbf{F}^k - t_r \mathbf{e}_i), S(\mathbf{F}^k - t_r \mathbf{e}_i))$, is obtained from:

$$|t_r| = \min_s \left\{ \frac{|L(\mathbf{F}^k, \mathbf{x}_s) - f(\mathbf{x}_s)|}{|\pi_i(\mathbf{x}_s)|} \right\} \quad \mathbf{x}_s \in \{\mathbf{X} - \mathbf{U}^k\} \tag{12}$$

The point \mathbf{x}_r is characterized by,

$$\text{sgn}[L(\mathbf{F}^k, \mathbf{x}_1) - f(\mathbf{x}_1)] = \text{sgn}[L(\mathbf{F}^{ki}, \mathbf{x}_1) - f(\mathbf{x}_1)], \quad \mathbf{x}_1 \in \{\mathbf{X} - \mathbf{U}^k - \mathbf{x}_r\} \tag{13}$$

Now, if there is not any δ such that $S(\mathbf{F}^k - \delta \mathbf{e}_i) < S(\mathbf{F}^k)$, then $S(\mathbf{F}^k)$ could not be reduced by moving on K along the edge parallel to the i^{th} parameter space coordinate axis and \mathbf{u}_i^k should not be replaced by another point from the set $\{\mathbf{X} - \mathbf{U}^k\}$. This iteration is repeated m times, once for each point in \mathbf{U}^k in succession. The whole cycle is then repeated a finite number of times until the solution (\mathbf{B}^*, d^*) is reached (see also, Abdelmalek (1974)).

Relation of this algorithm with simplex method has been discussed by Abdelmalek (1974). He shows that Usow's algorithm is completely equivalent to a dual simplex algorithm applied to a linear programming model with nonnegative bounded variables, and one iteration in the former is equivalent to one or more iterations in the latter. Bloomfield and Steiger (1980) devise an efficient algorithm based on the proposition of Usow explained above.

Sharpe (1971) by applying the L_1 norm regression to portfolio and its rate of return, gives an algorithm for the two parameters linear regression model. He argues that for the simple model with the objective function,

$$S = \sum_{i=1}^n |y_i - (\beta_0 + \beta_1 x_{i1})| \tag{14}$$

it must be possible to assign half of the points above and half below the regression line. With any given value of β_1 , we can derive β_0 as the median of $\beta_{0i} = y_i - \beta_1 x_{i1}$. Now segregate the points, such that:

$$S = \sum_{i_{above}}^{n_{above}} [y_i - (\beta_0 + \beta_1 x_{i1})] - \sum_{i_{below}}^{n_{below}} [y_i - (\beta_0 + \beta_1 x_{i1})] \tag{15}$$

Rearranging the terms and note that $n_{above} = n_{below}$ gives,

$$S = k_1 + k_2 \beta_1 \tag{16}$$

where,

$$k_1 = \sum_{i_{above}}^{n_{above}} y_i - \sum_{i_{below}}^{n_{below}} y_i \tag{17}$$

$$k_2 = -\sum_{i_{above}}^{n_{above}} x_{i1} + \sum_{i_{below}}^{n_{below}} x_{i1} \tag{18}$$

The overall solution strategy may now be formulated as follows. Any value of β_1 may be chosen at the outset. By computing β_{0i} , segregate the points above and below the line. By using equations (16) through (18) the corresponding segment of the S and β_1 relation is calculated. Sign of k_2 indicates the appropriate direction for the next iteration. If k_2 is positive, only smaller values of β_1 need be considered. If k_2 is negative, only larger values of β_1 need be considered. If k_2 is zero, the initial value of β_1 is a solution.

When a border line (median of β_{0i}) and the direction of search is determined, the nearest intersection of the present border line with another should be found. The calculations can be reduced by comparing slopes (x_{i1} values) to determine whether or not two lines intersect in the region of interest (thus avoiding needless division operations). The new values k_1 and k_2 must be computed. If this alteration causes k_2 to change sign, the solution has been obtained and algorithm stops.

Rao and Srinivasan (1972) interpret Sharpe's procedure as the solution of parametric dual linear programming formulation of the problem. They give an alternate and about equally efficient procedure for solving the same problem. Brown (1980) gives a distinct but similar approach to those of Edgeworth (1923) and Sharpe (1971). He emphasizes on the median properties of the estimator. The similarity comes from graphical approach of the three authors. Kawara (1979) also develops a graphical method for the simple regression model.

Bartels and Conn and Sinclair (1978) apply the method of Conn (1976) to the L_1 norm solution of overdetermined linear system. Their approach is minimization technique for piecewise differentiable functions. The algorithm may be reduced as follows.

Step 0) Select an arbitrary point \mathbf{b} .

Step 1) i) Identify the index set, $I = \{i_1, \dots, i_m\} = \{i \mid \mathbf{x}_i^T \mathbf{b} - y_i = 0\}$.

Let I^{th} row of \mathbf{X} be $\mathbf{X}_I = [x_{i_1}, \dots, x_{i_m}]$ where $i_j \in I$ and the nullspace $\mathbf{N} = \mathbf{N}(\mathbf{X}_I^T) = \{\delta \mid \mathbf{x}_i \delta = 0, i \in I\}$.

Let the orthogonal projector onto \mathbf{N} be denoted by \mathbf{P}_N .

ii) Let I^c be the complement of the set I . Compute the vector $\mathbf{h} = \sum_{i \in I^c} \text{sgn}(\mathbf{x}_i^T \mathbf{b} - y_i) \mathbf{x}_i$.

iii) Compute $\mathbf{p} = -\mathbf{P}_N \mathbf{h}$ that is the projection of $-\mathbf{h}$ onto the nullspace of \mathbf{X}_I , as long as this projection is non zero. If $\mathbf{p} \neq 0$, let $\mathbf{g} = \mathbf{h}$ and go to step 2.

- iv) compute \mathbf{w} according to $\mathbf{h}=\mathbf{X}_I\mathbf{w}=\sum_{j=1}^m w_j\mathbf{x}_i, i_j \in I$.
- v) if $|\mathbf{w}_j| \leq 1$ for all $j=1, \dots, m$ stop. In this case \mathbf{B} is optimal.
- vi) Find $i_{j0} \in I$ such that $|\mathbf{w}_{j0}| > 1$.
- vii) Change I to $I - \{i_{j0}\}$ and make corresponding changes to \mathbf{X}_I and N . Compute $\mathbf{p} = -\text{sgn}(w_{j0})\mathbf{P}_N\mathbf{x}_{i_{j0}}$ and let $\mathbf{g} = \mathbf{h} - \text{sgn}(w_{j0})\mathbf{x}_{i_{j0}}$.

Step 2) Determine $A = \{\alpha_i \mid i \in I^c, \alpha_i = (\mathbf{x}_i^T \mathbf{B} - y_i) / \mathbf{x}_i^T \mathbf{p}, \alpha_i > 0\}$ elements and order them such that $0 < \alpha_1 < \alpha_2 < \dots < \alpha_t$. Let $\tau = 1$.

Step 3) If $\mathbf{p}^T \mathbf{g} \geq 2 \text{sgn}(\mathbf{x}_\tau^T \mathbf{B} - y_\tau) \mathbf{p}^T \mathbf{x}_\tau$, then go to step 5.

Step 4) Change \mathbf{g} to $\mathbf{g} - 2 \text{sgn}(\mathbf{x}_\tau^T \mathbf{B} - y_\tau) \mathbf{x}_\tau$ and τ to $\tau + 1$ and go to step 3.

Step 5) Replace \mathbf{B} by $\mathbf{B} + \alpha_\tau \mathbf{p}$ and go to step 1.

This algorithm has also been modified for the case of degeneracy (see also, Bartels and Conn and Sinclair (1976)).

Bartels and Conn (1977) showed that how L_1 norm, restricted L_1 norm, L_∞ norm regressions and general linear programming can all be easily expressed as a piecewise linear minimization problem. Let \mathbf{U} and \mathbf{v} be of sizes $p \times m$ and $p \times 1$ respectively. Consider the function:

$$\Phi(\mathbf{B}) = \mathbf{h}^T \mathbf{B} + \sum_i |y_i - \mathbf{x}_i^T \mathbf{B}| + \sum_j \max(0, v_j - \mathbf{u}_j^T \mathbf{B}) \tag{19}$$

Where $y_i - \mathbf{x}_i^T \mathbf{B}$ represents the i^{th} element of the residual vector $\mathbf{y} - \mathbf{X}\mathbf{B}$ and $v_j - \mathbf{u}_j^T \mathbf{B}$ represents the j^{th} element of the residual vector $\mathbf{v} - \mathbf{U}\mathbf{B}$. To minimize Φ with respect to \mathbf{B} the following steps should be taken.

Step 0) Start with arbitrary $\mathbf{B}^{(k)}$.

Step 1) Find $\delta^{(k)}$ so that $\Phi(\mathbf{B}^{(k)} + \Theta \delta^{(k)}) \leq \Phi(\mathbf{B}^{(k)})$ for all $\Theta > 0$ small enough.

Step 2) Choose $\Theta^{(k)} \geq 0$ to obtain the largest possible decrease in Φ .

Step 3) Let $\mathbf{B}^{(k+1)} = \mathbf{B}^{(k)} + \Theta^{(k)} \delta^{(k)}$.

When $h=0$ and the sum on j is vacuous in the $\Phi(\mathbf{B})$ function (19), the resulting simplification of the above steps corresponds precisely to the algorithm proposed by Bartels and Conn and Sinclair (1978). Other related problems can be solved by modification of $\Phi(\mathbf{B})$ by a quantity μ to obtain a parameterized family of piecewise linear functions $\Phi_\mu(\mathbf{B})$, and take following steps to find the minimum solution.

Step 0) Set $\mu > 0$; select any $\mathbf{B} = \mathbf{B}^{(0)}$.

Step 1) Minimize $\Phi_\mu(\mathbf{B})$ with respect to \mathbf{B} according to the above procedure.

Step 2) Stop if a prescribed terminating condition on β is met; otherwise set $\mu = \mu/10$ and go to step 1.

The contribution of this paper is putting a wide class of problems in the mould of two algorithms mentioned above. The techniques are easily extended to the models with norm restrictions.

Bloomfield and Steiger (1980) proposed a descent method for the L_1 norm multiple regression. Their algorithm is also explained in Bloomfield and Steiger (1983). In some steps this algorithm is related to that of Singleton (1940) and Usow (1967b). The basis of this method is to search for a set of m observations which locate on the optimal L_1 norm regression. This set is found iteratively by successive improvement. In each iteration, one point from the current set is identified as a good prospect for deletion. This point is then replaced by the best alternative. The novel features of this method are in an efficient

procedure for finding the optimal replacement and a heuristic method for identifying the point to be deleted from the pivot.

Denote the $\mathbf{x}_1^T, \dots, \mathbf{x}_m^T$ as rows of the independent variables design matrix which correspond to the current set of points 1, ..., m; and \mathbf{x}_m^T for replacement. Set,

$$y_i = \mathbf{x}_i^T \mathbf{B} \quad i=1, \dots, m-1 \quad (20)$$

Redefine \mathbf{B} as,

$$\mathbf{B} = \mathbf{B}^0 + t\boldsymbol{\delta} \quad (21)$$

Where \mathbf{B}^0 is arbitrary member of the set and $\boldsymbol{\delta}$ vector obeys the following system,

$$\mathbf{x}_i^T \boldsymbol{\delta} = 0 \quad i=1, \dots, m-1 \quad (22)$$

Given this set of points, optimum value of S may be found by minimizing the following expression with respect to the scalar t.

$$\sum_{i=1}^n |y_i - \mathbf{x}_i^T(\mathbf{B}^0 + t\boldsymbol{\delta})| \quad (23)$$

Rearranging the terms, leads to:

$$\sum_{i=1}^n |w_i| |r_i - t| \quad (24)$$

where $w_i = \mathbf{x}_i^T \boldsymbol{\delta}$ and $r_i = (y_i - \mathbf{x}_i^T \mathbf{B}^0) / (\mathbf{x}_i^T \boldsymbol{\delta})$. Value of t may be found by Laplace's weighted median method. Bloomfield and Steiger propose a weighted modification of partial quicksort procedure of Chamber (1971) to find the t value efficiently. The data point corresponding to weighted median replaces \mathbf{x}_k^T . By using the y_i from (20) and the additional m^{th} equation, value of \mathbf{B}^0 is computed. Vector $\boldsymbol{\delta}$ is determined up to scalar multiples of (22). The new set of parameters values are computed by (21). Now a point should be deleted. Bloomfield and Steiger do not give an assured way to identify this point. They propose a heuristic method based on gradients and use the following quantity:

$$\rho = \frac{\left| \begin{array}{c} \sum w_i - \sum_{i:r_i < 0} w_i \\ \sum_{i:r_i > 0} w_i \\ - \sum_{i:r_i = 0} w_i \end{array} \right|}{\sum_{i=1}^n w_i} \quad (25)$$

Once ρ is calculated for each candidate point for deletion, and delete the point for which ρ is largest.

To start the algorithm, any set of m rows of \mathbf{X} may be chosen, with the appropriate \mathbf{B}^0 . Add variables stepwise until a fit for \mathbf{B}^0 and corresponding set of m points are derived. At each intermediate step, the fit involves k variables where $0 \leq k < m$ and corresponding set of k data points with zero residuals. Improving the fit by entering a new variable, thus increasing k to k+1. At each stage, it is the measure ρ that determines whether we set up to a larger model or improve the current one without setting up. Suppose at the current stage we are dealing with k variables. If the value of ρ is largest at the variable p, which p is not in the set of k variables above, k is increased to k+1 by entering the variable p and improving the fit with k+1 variables. If p is in the current set of k variables and ρ is in largest value, corresponding point to p is deleted and replaced in a manner described before. In this paper relationship of this algorithm to linear programming is also discussed.

Seneta and Steiger (1984) proposed an algorithm for L_1 norm solution of slightly overdetermined system of equations. Their proposition is based on the above algorithm of Bloomfield and Steiger. It is more efficient than the former if m is near n. Given $(\mathbf{x}_i, y_i) \in \mathbf{R}^{m+1}$, $i=1, \dots, n$ and $k=n-m$, $\mathbf{X}=(\mathbf{x}_1, \dots, \mathbf{x}_n)^T$, their algorithm may be described as follows:

Step 1) Renumber the rows of $(\mathbf{X}|\mathbf{y})$ such that \mathbf{X}_N the bottom m rows of \mathbf{X} , is invertible. Solve the k linear equations system $\mathbf{X}_N^T \mathbf{N} = -\mathbf{X}_T^T$ for \mathbf{N} , where \mathbf{X}_T denotes the top k rows of \mathbf{X} .

Step 2) Let $\mathbf{D}=(\mathbf{A}|\mathbf{c})$ where $\mathbf{A}=(\mathbf{I}|\mathbf{N})$ is of size $k \times n$ and $\mathbf{c}=\mathbf{A}\mathbf{y}$.

Step 3) Let $\sigma=(1, \dots, k)$, $\sigma^c=(k+1, \dots, n)$.

Step 4) Set $r_{\sigma}(i)=b_i$ for $i=1, \dots, k$; $r_{\sigma^c}(i)=0$ for $i=k+1, \dots, n$.

Step 5) Let $I=\{i \mid 1 \leq i \leq k, c_i=0\}$.

Step 6) Do loop for $j=1$ to m :

Let $v_i=D_{i\sigma(j)}$ for $j=1, \dots, k$. Let $M=\{i \mid \text{sgn}(c_i) \neq \text{sgn}(v_i)\}$ and $J=\{1, \dots, k\} \setminus M$.

$$\left| \sum_M |v_i| - \sum_J |v_i| \right| \quad \left| -1 - \sum_I |v_i| \right|$$

Let $\beta_j = \frac{\quad}{\quad}$

$$\frac{k}{1 + \sum_{i=1}^k |v_i|}$$

End loop.

Step 7) Set $S=\{1, \dots, m\}$.

Step 8) Choose β_q as $\max_s \{\beta_j\}$.

Step 9) If $\beta_q > 0$ go to step 11.

k

Step 10) If $\prod_{i=1}^k r_{\sigma}(i)=0$, the problem is degenerate; and stop.

$i=1$

m

Otherwise, solve, $y_{\sigma^c(i)} = \sum_{j=1}^m x_{\sigma^c(i)j} \Theta_j$ for Θ , stop.

$j=1$

Step 11) Let $v_i=D_{i,\sigma(q)}$ for $i=1, \dots, k$.

Step 12) Compute t^{\wedge} =weighted median of $c_1/v_1, \dots, c_k/v_k, 0$ with weights $|v_1|, \dots, |v_k|, 1$.

Step 13) If $t^{\wedge}=c_p/v_p \neq 0$ go to step 16.

Step 14) Let $S=S \setminus \{q\}$; if $S=\emptyset$ go to step 10.

Step 15) Go to step 8.

Step 16) Divide row p of \mathbf{D} by $D_{p\sigma(q)}$.

Step 17) For $i \neq p$ in \mathbf{D} , let $(\text{row } i) = (\text{row } i) - (\text{row } p) * D_{i\sigma(q)}$.

Step 18) Commute for pair $\sigma(p)$ and $\sigma^c(q)$.

Step 19) Let $r_{\sigma}(i)=b_i$ for $i=1, \dots, k$; and set $r_{\sigma^c(q)}=0$.

Step 20) Go to step 5.

Seneta (1983) reviews the iterative use of weighted median to estimate the parameters vector in the classical linear model when the fitting criterion is L_1 norm and also Cauchy criterion.

Wesolowsky (1981) presents an algorithm for multiple L_1 norm regression based on the notion of edge descent along the polyhedron of the objective function. This algorithm is closely related to those of Rhodes (1930) and Bartels and Conn and Sinclair (1978) which explained before. Consider the multiple linear regression as before. Select a set of m points $(x_{j1}^I, \dots, x_{jm}^I, y_j^I)$. The following system of equations can be solved for a unique set of coefficients.

$$y_j^I - \sum_{h=1}^m \beta_h x_{jh}^I = 0, \quad j=1, \dots, m \tag{26}$$

An edge is formed of any subset J consisting of $m-1$ equations. To minimize along an edge, set:

$$\mathbf{Y}_I = \begin{bmatrix} Y_1^I \\ \vdots \\ Y_m^I \end{bmatrix}, \mathbf{X}_I = \begin{pmatrix} x_{11}^I & \cdots & x_{1m}^I \\ \vdots & \ddots & \vdots \\ x_{m1}^I & \cdots & x_{mm}^I \end{pmatrix}, \mathbf{Y}_J = \begin{bmatrix} Y_1^J \\ \vdots \\ Y_{m-1}^J \end{bmatrix}, \mathbf{X}_J = \begin{pmatrix} x_{11}^J & \cdots & x_{1m}^J \\ \vdots & \ddots & \vdots \\ x_{m-1,1}^J & \cdots & x_{m-1,m}^J \end{pmatrix} \quad (27)$$

Let β_p be formed from β by deleting β_p and let x_p be the p^{th} column in \mathbf{X}_J and let \mathbf{X}_{pJ} be formed by removing x_p from \mathbf{X}_J . Then for a given β_p it can be shown that,

$$\beta_p = \mathbf{X}_{pJ}^{-1} \mathbf{Y}_J - \beta_p \mathbf{X}_{pJ}^{-1} x_p \quad (28)$$

Let the elements of β_p be

$$\beta_q = r_q - s_q \beta_p \quad \text{for } q \neq m \quad (29)$$

Substitution in the L_1 norm objective function gives,

$$\min_{\beta_1} : \sum_{i=1}^n \left| x_{ip} - \sum_{\substack{q=2 \\ q \neq p}}^m s_q x_{iq} \right| \left| \frac{y_i - \sum_{\substack{q=2 \\ q \neq p}}^m r_q x_{iq}}{\sum_{q=2}^m s_q x_{iq}} - \beta_1 \right| \quad (30)$$

Now the following steps should be taken.

Step 1) Set $k=1, l=0$, choose initial values for β_1, \dots, β_m . Least squares values are one possibility. Let $\mathbf{I}^{(1)} = \{j_1^{(1)}, \dots, j_m^{(1)}\}$ be a set of m data points chosen in sequence as follows. The point with the smallest squared residual is chosen each time subject to the condition that $\mathbf{X}_I^{(1)}$ is nonsingular. Find $\beta_q^{(1)}$ for $q=1, \dots, m$, by solving $\mathbf{Y}_I^{(1)} = \mathbf{X}_I^{(1)} \beta$. Set $\mathbf{I}^{(k)} = (j_1^{(k)}, \dots, j_m^{(k)})$; $\mathbf{J} = (j_2^{(k)}, \dots, j_m^{(k)})$.

Step 2) Set $k=k+1$. Obtain β_p for the smallest p from (30) by using the weighted median procedure. Set $\beta_p^{(k)} = \beta_p$ and let i be the index which defines the lower weighted median β_p in (30) for the lowest possible p .

Step 3) a) If $\beta_p^{(k)} - \beta_p^{(k-1)} = 0$ and if $l > m$, go to step 4. Otherwise, set $\mathbf{I} = (j_2^{(k-1)}, \dots, j_m^{(k-1)}, i)$ and $l=l+1, \beta_q^{(k)} = \beta_q^{(k-1)}$ for all q and go to step 2.

b) If $\beta_p^{(k)} - \beta_p^{(k-1)} \neq 0$, set $l=0$. Calculate $\beta_q^{(k)}$ for $q \neq p$ from (29). Set $\mathbf{I}^{(k)} = (j_2^{(k-1)}, \dots, j_m^{(k-1)}, i)$ and go to step 2.

Step 4) Calculate $\beta_q^{(k)}$ for $q \neq p$ from (29); set $\beta^* = \beta^{(k)}$ and stop.

In this paper Wesolowsky also discusses the problem of multicollinearity and gives an appropriate solution.

Josvanger and Sposito (1983) modify Wesolowsky's algorithm for the two parameter simple linear regression model. The modification is an alternative way to order observations instead of sorting all of them to find the necessary weighted median value. Suppose the problem has been reduced to a weighted median problem. They place smaller values of factors to be sorted with corresponding weights below $\beta_1^{(k-1)}$ and larger or equal values above it, then recheck the inequalities (4) of weighted median. If the inequalities do not satisfy then an appropriate adjustment is made. In particular, if the right hand side is overly weighted, then the weight corresponding to the smallest sorting factor is transferred to the left hand side, and the check is made again. A computer program for this algorithm is also given by the authors.

"Generalized gradient" method introduced by Clarke (see, Clarke (1983)) is a general procedure for nonsmooth optimization functions and problems (see, Osborne and Pruess and Womersley (1986)). A subclass of this method is called "reduced gradient" explained by Osborne (1985) is a general algorithm which contains linear programming, piecewise linear

optimization problems and polyhedral convex function optimization algorithms inside. The reduced gradient algorithm is a special case of descent method which possesses two important characteristics. Identify direction and taking a step in this direction to reduce the function value (see also, Anderson and Osborne (1976), Osborne and Watson (1985) Osborne (1985,87)). The algorithms of Bartels and Conn and Sinclair (1978), Armstrong and Frome and Kung (1979), Bloomfield and Steiger (1980) are all special cases of reduced gradient method.

Imai and Kato and Yamamoto (1987) present a linear time algorithm for computing the two parameter L_1 norm linear regression by applying the pruning technique. Since the optimal solution in the $\beta_0 \times \beta_1$ plane lies at the intersection of data lines, so, at each step a set of data lines which does not determine the optimum solution are discarded. In this paper algebraic explanation of the problem is also offered.

Pilibossian (1987) also gives an algorithm similar to Karst (1958) for the simple two parameter linear L_1 norm regression. Bidabad (1987a,b,88a,b) proposed descent methods for the simple and multiple L_1 norm regressions. These algorithms with many improvements will be discussed by Bidabad (1989a,b). Bidabad (1989a,b) introduces four descent algorithms which two of them are for simple and two others are for multiple regression models. His first algorithm is crude and tries to check many points to find optimal solution. Second algorithm is more efficient. Algorithm three is a partial descent procedure for general linear model. Algorithm four is a full descent method which has many proved properties. He also proves the convergence of all above four algorithms.

Simplex type algorithms

The essence of linear programming in solving L_1 norm problem may be found in the work of Edgeworth (1888). Harris (1950) suggested that the L_1 norm estimation problem is connected with linear programming. Charnes and Cooper and Ferguson (1955) formulated the problem as linear programming model. This article is the first known to use linear programming for this case. Adaptation of linear programming to L_1 norm estimation problem is shown below:

$$\begin{aligned} \min: & \mathbf{1}_n^T(\mathbf{w}+\mathbf{v}) \\ & \beta \\ \text{s.to: } & \mathbf{X}\beta + \mathbf{I}_n(\mathbf{w}-\mathbf{v}) = \mathbf{y} \\ & \mathbf{w}, \mathbf{v} \geq \mathbf{0} \\ & \beta \text{ unrestricted in sign} \end{aligned} \quad (31)$$

where $\mathbf{1}_n$ is a vector of size $n \times 1$ of 1's and \mathbf{I}_n is a n^{th} order identity matrix. The vectors \mathbf{v} and \mathbf{w} are of size $n \times 1$ and their elements may be interpreted as vertical deviations above and below the fitted regression hyperplane respectively. This problem has n equality constraints in $m+2n$ variables. When n is large, this formulation generally requires a large amount of storage and computation time.

Wagner (1959) shows that the formulation of the L_1 norm regression may be reduced to m equality constraints linear programming problem. Thus, this dual formulation reduces n equations of primal form to m equations of dual form and considerably reduces the storage and computation time.

Fisher (1961) reviews the formulation of the L_1 norm estimation in relation with primal form of linear programming. Barrodale and Young (1966) developed a modified simplex algorithm for determining the best fitting function to a set of discrete data under the L_1 norm criterion. The method is given as Algol codes (for critics see, McCormick and Sposito (1975)). Davies (1967) demonstrates the use of the L_1 norm regression estimates. Rabinowitz (1968) also discusses the application of linear programming in this field. Crocker (1969)

cautions against using the L_1 norm criterion merely to restrain unwanted negative coefficient estimates which occur in least squares regression. Multicollinearity is one of the cases which causes this result. Robers and Ben-Israel (1969) by using interval linear programming, proposed an algorithm to solve the L_1 norm estimation problem. Rabinowitz (1970), Shanno and Weil (1970) discuss some connections between linear programming and approximation problem. Barrodale (1970) summarizes the linear and nonlinear L_1 norm curve fitting on both continuous and discrete data. Spyropoulos and Kiountouzis and Young (1973) suggest two algorithms for fitting general functions and particularly fast algorithm with minimum storage requirements for fitting polynomials based on the algebraic properties of linear programming formulation. Robers and Robers (1973) have supplied a special version of the general method of Robers and Ben-Israel (1969) which is designed specifically for the L_1 norm problem. A related Fortran code is also provided.

Barrodale and Roberts (1973) present a modification of simplex method which needs smaller amount of storage and by skipping over simplex vertices is more efficient than usual simplex procedure. Define the vector \mathbf{b} as a difference of two nonnegative vectors \mathbf{c} and \mathbf{d} , their formulation can be stated as follows,

$$\begin{aligned} \min: & \mathbf{1}_n^T(\mathbf{w}+\mathbf{v}) \\ & \mathbf{c}, \mathbf{d} \\ \text{s.to: } & \mathbf{X}(\mathbf{c}-\mathbf{d})+\mathbf{I}_n(\mathbf{w}-\mathbf{v})=\mathbf{y} \\ & \mathbf{w}, \mathbf{v}, \mathbf{c}, \mathbf{d} \geq \mathbf{0} \end{aligned} \quad (32)$$

Because of the relationships among variables, computation can be performed by using only $(n+2) \times (m+2)$ amount of array storage, including labels for the basic and non-basic vectors. An initial basis is given by \mathbf{w} if all y_i are nonnegative. If a y_i is negative, sign of the corresponding row is changed and the unit column from the corresponding element of \mathbf{v} is taken as part of the basis. The algorithm is implemented in two stages. First stage restricts the choice of pivotal column during the first m iterations to the vectors elements c_j and d_j according to the associated maximum nonnegative marginal costs. The vector that leaves the basis causes the maximum decrease in the objective function. Thus the pivot element is not necessarily the same as in the usual simplex. Second stage involves interchanging non basic w_i or v_i with the basic w_i or v_i . The basic vectors corresponding to c_j and d_j are not allowed to leave the basis. The algorithm terminates when all marginal costs are nonpositive (see, Kennedy and Gentle (1980)). A Fortran code for this procedure is given by Barrodale and Roberts (1974). Peters and Willms (1983) give algorithms accompanying with computer codes for up-and-down dating the solution of the problem when a column or row inserted to or deleted from \mathbf{X} , or \mathbf{y} is changed. These algorithms are all based on Barrodale and Roberts (1973,74) procedure.

Abdelmalek (1974) describes a dual simplex algorithm for the L_1 norm problem with no use of artificial variables. For this algorithm, the Haar condition (see, Osborne (1985), Moroney (1961)) need not be satisfied anymore. This algorithm seemed to be very efficient at the time of publication. An improved dual simplex algorithm for L_1 norm approximation is proposed by Abdelmalek (1975a). In this algorithm, certain intermediate iterations are skipped and in the case of ill-conditioned problems, the basis matrix can lend itself to triangular factorization and thus ensure stable solution. Abdelmalek (1980a) improves his previous algorithm by using triangular decomposition. A Fortran translation of the algorithm is given by Abdelmalek (1980b). Sposito and McCormick and Kennedy (1975) summarizes much of the works on L_1 norm estimation including problem statement, linear programming formulation, efficient computational algorithms and properties of the estimators.

Armstrong and Kung (1978) propose an algorithm for simple two parameter L_1 norm regression. The method is a specification of linear programming of Barrodale and Roberts (1973) algorithm. A Fortran code is given too.

Armstrong and Frome and Kung (1979) use LU (Lower-Upper triangular) decomposition of Bartels and Golub (1969) in maintaining the current basis on revised simplex procedure. A Fortran translation is also enclosed. Armstrong and Godfrey (1979) show that the primal method of Barrodale and Roberts (1973) and dual method of Abdelmalek (1975a) are essentially equivalent. With a given initial basis for the two methods, they show that, both algorithms will generate corresponding bases at each iteration. The only difference is the choice of initial basis and heuristic rules for breaking ties. Armstrong and Kung (1982b) presents a dual linear programming formulation for the problem. Various basis entry and initialization procedures are considered. It has been shown that the dual approach is superior to primal one if a good dual feasible solution is readily available (see also, Steiger (1980)). Banks and Taylor (1980) suggest a modification of Barrodale and Roberts (1973) algorithm. The objective function is altered to include magnitudes of the elements of the both errors and solution vectors. For a general discussion on simplex for piecewise linear programming see Fourer (1985a,b) and for a survey of the corresponding problem on the L_1 norm see Fourer (1986). Narula and Wellington (1987) propose an efficient linear programming algorithm to solve the both L_1 and L_∞ norms linear multiple regressions. The algorithm exploits the special structure and similarities between the two problems.

Brennan and Seiford (1987) develop a geometrical interpretation of linear programming in L_1 norm regression. They give a geometric insight into the solving process in the space of observations. McConnell (1987) shows how the method of vanishing Jacobians which has been used to optimize quadratic programming problems can also be used to solve the special linear programming problem associated with computing linear discrete L_1 norm approximation. For the possibility of applying other types of linear programming solutions such as Karmarkar solution to L_1 norm problem see Meketon (1986).

Other algorithms

This category consists of algorithms which were not classified in the two last sections.

Rice (1964c) applies the bisection method to L_1 norm regression. In this method at each step the domain of S is broken to two segments and the appropriate segment is selected for the next iteration. Solution is reached when the last segment is less than a predetermined small value.

Abdelmalek (1971) develops an algorithm for fitting functions to discrete data points and solving overdetermined system of linear equations. The procedure is based on determining L_1 norm solution as the limiting case of L_p norm approximation when p tends to one from right in limit. This technique thus obtains a solution to a linear problem by solving a sequence of nonlinear problems.

Schlossmacher (1973) computed the L_1 norm estimates of regression parameters by an iterative weighted least squares procedure. Instead of minimizing sum of absolute deviations he minimized sum of weighted squared errors with $1/|u_i|$ as weights. Once least squares is applied to the problem and residuals are computed. The absolute value of the inverse of the residuals are again used as corresponding weights in the next iteration for minimizing the sum of weighted squared errors (see also, Holland and Welsh (1977)). Fair (1974) observed that the estimated values of β did not change after the second or third iterations. In cases where any residual is zero, continuation of procedure is impossible, because the corresponding weight to this residual is infinite. This problem is also discussed by Sposito

and Kennedy and Gentle (1977), Soliman and Christensen and Rouhi (1988). Absolute convergence of this algorithm has not been proved, but non-convergent experiment has not been reported.

Soliman and Christensen and Rouhi (1988) used left pseudoinverse (see, Dhrymes (1978) for description of this inverse) to solve the general linear L_1 norm regression. According to this procedure one should calculate the least squares solution using the left pseudoinverse or least squares approximation as $\mathbf{B}^{\wedge} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$. Calculate the residuals as $\mathbf{u} = |\mathbf{y} - \mathbf{X} \mathbf{B}^{\wedge}|$; where \mathbf{u} is $n \times 1$ column vector. Select the m observations with the smallest absolute values of the residuals and partition the matrices as the selected observations locate on the top,

$$\mathbf{u} = \begin{bmatrix} \hat{\mathbf{u}} \\ \cdots \\ \tilde{\mathbf{u}} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} \hat{\mathbf{X}} \\ \cdots \\ \tilde{\mathbf{X}} \end{bmatrix}, \mathbf{y} = \begin{bmatrix} \hat{\mathbf{y}} \\ \cdots \\ \tilde{\mathbf{y}} \end{bmatrix}. \quad (33)$$

Solve $\mathbf{y}^{\wedge} = \mathbf{X}^{\wedge} \mathbf{B}^{\wedge}$ for the top partitions as $\mathbf{B}^{\wedge} = \mathbf{X}^{\wedge^{-1}} \mathbf{y}$. Although this procedure is operationally simple, its solution is not the same as other exact methods and no proof is presented to show that the solution is in the neighborhood of the exact solution of the L_1 norm minimization problem.

Application of median polish (see, Tukey (1977)) and e-median polish to L_1 norm estimation are discussed and developed by Bloomfield and Steiger (1983), Kemperman (1984), Sposito (1987a), Bradu (1987a,b).

Application of Karmarkar's algorithm for linear programming and its relation to L_1 norm is given by Sherali and Skarpness and Kim (1987). For using homotopy method in L_1 norm see Garcia and Gould (1983), Schellhorn (1987). An algorithm for linear L_1 norm approximation for continuous function is given by Watson (1981), (see also, Baboolal and Watson (1981)).

BIBLIOGRAPHY

- N.N. Abdelmalek (1971) Linear approximation for a discrete point set and L_1 solutions of overdetermined linear equations. *J. ACM*, 18, 41-47.
- N.N. Abdelmalek (1974) On the discrete linear L_1 approximation and L_1 solutions of overdetermined linear equations. *J. of Approx. Theory*, 11, 38-53.
- N.N. Abdelmalek (1975a) An efficient method for the discrete L_1 approximation problem. *Math. Comput.*, 29, 844-850.
- N.N. Abdelmalek (1980a) L_1 solution of overdetermined systems of linear equations. *ACM Trans. Math. Soft.*, 6, 220-227.
- N.N. Abdelmalek (1980b) A Fortran subroutine for the L_1 solution of overdetermined systems of linear equations. *ACM Trans. Math. Soft.*, 6, 228-30.
- D.H. Anderson, M.R. Osborne (1976) Discrete linear approximation problems in polyhedral norms. *Numer. Math.* 26, 179-189.
- R.D. Armstrong, E.L. Frome, D.S. Kung (1979) A revised simplex algorithm for the absolute deviation curve fitting problem. *Commun. Stat. B8*, 175-190.
- R.D. Armstrong, J. Godfrey (1979) Two linear programming algorithms for the discrete L_1 problem. *Math. Comput.*, 33, 289-300.
- R.D. Armstrong, D.S. Kung (1978) AS132: Least absolute value estimates for a simple linear regression problem. *Appl. Stat.*, 27, 363-366.
- R.D. Armstrong, D.S. Kung (1982b) A dual algorithm to solve linear least absolute value problems. *J. Oper. Res. Soc.*, 33, 931-936.
- T.S. Arthanari, Y. Dodge (1981) *Mathematical programming in statistics*. John Wiley, Interscience division, New York.
- S. Baboolal, G.A. Watson (1981) Computational experience with an algorithm for discrete L_1 approximation. *Computing*, 27, 245-252.

- S.C. Banks, H.L. Taylor (1980) A modification to the discrete L_1 linear approximation algorithm of Barrodale and Roberts. *SIAM J. on Scientific and Stat. Comput.* 1, 187-190.
- Barrodale (1970) On computing best L_1 approximations. In A. Talbot, *Approximation theory* Academic Press, New York, 205-215.
- Barrodale, F.D.K. Roberts (1973) An improved algorithm for discrete L_1 linear approximation. *SIAM J. Numer. Anal.*, 10, 839-848.
- Barrodale, F.D.K. Roberts (1974) Algorithm 478: Solution of an overdetermined system of equations in the L_1 norm. *Commun. ACM*, 17, 319-320.
- Barrodale, A. Young (1966) Algorithms for best L_1 and L_∞ linear approximations on a discrete set. *Numer. Math.*, 8, 295-306.
- R.H. Bartels, A.R. Conn (1977) LAV regression: A special case of piecewise linear minimization. *Commun. Stat.*, B6, 329-340.
- R.H. Bartels, A.R. Conn, J. Sinclair (1976) The L_1 solution to an overdetermined linear system. *Proc. 9th Ann. Symp. Interface Statist.* In C.D.C. Hoaglin (ed.) Boston, Prindle, Weber and Schmidt Inc., 120-7.
- R.H. Bartels, A.R. Conn, J. Sinclair (1978) Minimization technique for piecewise differentiable functions: The L_1 solution to an overdetermined linear system. *SIAM J. Numer. Anal.* 15, 224-241.
- R.H. Bartels, G.H. Golub (1969) The simplex method of linear programming using LU decomposition. *Commun. ACM*, 12, 266-268.
- J. Bejar (1956) Regression en mediana y la programación lineal, *Trabajos de Estadística* 7, 141-58.
- J. Bejar (1957) Calculo practico de la regression en mediana *Trabajos de Estadística*, 8, 157-173.
- B. Bidabad (1987a) Least absolute error estimation. Submitted to the First International Conference on Statistical Data Analysis Based on the L_1 norm and Related Methods, Neuchatel, Switzerland.
- B. Bidabad (1987b) Least absolute error estimation, part II. Submitted to the First International Conference on Statistical Data Analysis Based on the L_1 norm and Related Methods, Neuchatel, Switzerland.
- B. Bidabad (1988a) A proposed algorithm for least absolute error estimation. *Proceedings of the Third Seminar of Mathematical Analysis.* Shiraz University, 24-34, Shiraz, Iran.
- B. Bidabad (1988b) A proposed algorithm for least absolute error estimation, part II. *Proceedings of the Third Seminar of Mathematical Analysis,* Shiraz University, 35-50, Shiraz, Iran.
- B. Bidabad (1989a) Discrete and continuous L_1 norm regressions, proposition of discrete approximation algorithms and continuous smoothing of concentration surface, Ph.D. thesis, Islamic Azad University, Tehran, Iran.
- B. Bidabad (1989b) Discrete and continuous L_1 norm regressions, proposition of discrete approximation algorithms and continuous smoothing of concentration surface, Ph.D. thesis, Islamic Azad University, Tehran, Iran. Farsi translation.
- P. Bloomfield, W. Steiger (1980) Least absolute deviations curve fitting. *SIAM J. Sci. Statist. Comput.* 1, 290-301.
- P. Bloomfield, W. Steiger (1983) *Least absolute deviations: theory, applications and algorithms.* Birkhauser, Boston.
- R.J. Boscovich (1757) *De litteraria expeditione per pontificiam ditionem, et synopsis amplioris operis...*, 'Bononiensis' scientiarum et artum instituto atque academia commetarii, vol.4, 353-396. Reprinted with a Serbo-Croatian translation by N. Cubranic, Institute of higher geodesy, University of Zagreb 1961.
- R.J. Boscovich (1760) *De recentissimis graduum dimensionibus et figura, ac magnitudine terrae inde derivanda.* *Philosophiae recentioris, a benedicto stay in Romano archigynasis publico eloquentare professore, vesibus traditae, libri X, cum adnotianibus et supplementas P. Rugerii Joseph Boscovich, S.J.,* 2, 406-426.
- A.L. Bowley (1902) Methods of representing the statistics of wages and other groups not fulfilling the normal law of error, II: applications to wage statistics and other groups. *J. of the Roy. Stat. Soc.*, 65, 331-54.
- A.L. Bowley (1928) *F.Y. Edgeworth's contributions to mathematical statistics.* London, Roy. Stat. Soc..
- D. Bradu (1987a) L_1 fit, median polish and conjugate gradients. *CSIR Tech. Rep. TWISK 509,* National Res. Inst. For Math Sci. CSIR, Pretoria.
- D. Bradu (1987b) An ε -median polish algorithm. *CSDA*, 5, 327-336.
- J.J. Brennan, L.M. Seiford (1987) Linear programming and L_1 approximation using the method of vanishing Jacobians. *CSDA*, 5, 263-276.
- B.M. Brown (1980) Median estimates in a simple linear regression. *Australian J. of Stat.*, 22, 154-165.
- C. Bruen (1938) Methods for the combination of observations modal points or most lesser-deviations, mean loci or least squares, and mid point of least range or least greatest-deviation. *Metron* 13, 61-140.
- J. Chamber (1971) Algorithm 410: partial sorting. *Comm. ACM*, 14, 357-358.
- J.M. Chambers (1977) *Computational methods for data analysis* Wiley, New York.
- Charnes, W.W. Cooper, R.O. Ferguson (1955) Optimal estimation of executive compensation by linear programming. *Manag. Sci.* 1, 138-151.
- E.W. Cheney (1966) *Introduction to approximation theory,* McGraw-Hill, New York.

- F.H. Clarke (1983) Optimization and nonsmooth analysis. Wiley, New York
- A.R. Conn (1976) linear programming via a nondifferentiable penalty function. *SIAM J. Numer. Anal.*, 13, 145-154.
- D.C. Crocker (1969) Linear programming technique in regression analysis, the hidden danger. *A.I.E.E. Trans.*, 1, 112-126.
- M. Davies (1976) Linear approximation using the criterion of least total deviations. *J. Roy. Stat. Soc. B29*, 101-109.
- T.E. Dielman (1984) Least absolute value estimation in regression models: An annotated bibliography. *Comm. Stat.* 13, 513-41.
- T. Dielman, R. Pfaffenberger (1982) LAV (Least Absolute Value) estimation in linear regression: A review, *TIMS studies in the Manag. Sci.*, 19, 31-52.
- T. Dielman, R. Pfaffenberger (1984) Computational algorithms for calculating least absolute value and Chebyshev estimates for multiple regression. *Amer. J. Math. Manag. Sci.*, 4, 169-197.
- P.J. Dhrymes (1978) Mathematics for econometrics. Springer-Verlag, New York.
- Y. Dodge (1987) An introduction to statistical data analysis L_1 -norm based. In Y. Dodge (ed.) *Statistical data analysis based on the L_1 norm and related methods*. North-Holland. Reprinted in *CSDA*, 5, 239-254.
- A.F. Dufton (1928) Correlation. *Nature*, 121, 866.
- F.Y. Edgeworth (1883) The method of least squares. *Philosophical Magazine*, 16, 360-375.
- F.Y. Edgeworth (1887a) On observations relating to several quantities. *Hermathena*, 6, 279-285.
- F.Y. Edgeworth (1887b) A new method of reducing observations relating to several quantities. *Philosophical Magazine*, 24, 222-223.
- F.Y. Edgeworth (1888) On a new method of reducing observation relating to several quantities. *Philosophical Magazine*, 25, 184-191.
- F.Y. Edgeworth (1902) Method of representing statistics of wage and other groups not fulfilling the normal law of error, I: mathematical considerations. *J. Roy. Stat. Soc.*, 65, 325-331.
- F.Y. Edgeworth (1923) On the use of medians for reducing observations relating to several quantities. *Philosophical Magazine*, 6th series, 46, 1074-1088.
- Eisenhart (1961) Boscovich and the combination of observations. Ch. 9 of Whyte (1961, 200-212) reprinted in Kendall and Plackett (1977) *studies in the history of statistics and probability*, vol.II, Charles Griffin and Co. Ltd., High Wycombe 88-100.
- J.E. Estienne (1926-28) Introduction a une theorie rationnelle des erreurs d'observation. *Revue d'artillerie* 97(1926), 421-441; 98(1928), 542-562; 100(1927), 471-487.
- R.C. Fair (1974) On the robust estimation of econometric models. *Ann. Econ. Soc. Measurement*, 3, 667-77.
- R.W. Farebrother (1987b) The historical development of the L_1 and L_∞ estimation procedures. In Y. Dodge (ed.) *Statistical data analysis based on the L_1 norm and related methods*. North-Holland. 37-64.
- R.W. Farebrother (1987c) A simple recursive procedure for the L_1 norm fitting of a straight line. *Work. Pap., Dept. of Econometrics and Social Stat. University of Manchester, Manchester, M13 9PL, UK.*
- W.D. Fisher (1961) A note on curve fitting with minimum deviations by linear programming. *JASA*, 11, 359-362.
- R. Fourer (1985a) A simplex algorithm for piecewise-linear programming I: derivation and proof. *Math. Prog.*, 33, 204-233.
- R. Fourer (1985b) A simplex algorithm for piecewise-linear programming II: Finiteness, feasibility and degeneracy. *Tech. Rep.*, 85-03 (revised), Dept. of Ind. Engin. and Manag. Sci., The Tech. Inst., Northwestern University, Evanston, Illinois.
- R. Fourer (1986) A simplex algorithm for piecewise-linear programming III: Computational analysis and applications. *Tech. Rep.*, 86-03, Dept. of Ind. Engin. and Manag. Sci., The Tech. Inst., Northwestern University, Evanston, Illinois.
- J.B.I. Fourier (1824) Solution d'une question particuliere au calcul des inegalites, second extrait. *Histoire de l'academie des sciences pour 1824*, 47-55. Reprinted in *oeuvres de Fourier*, 2. Paris, 1980, Gauthier-Villars, 325-328.
- G. Galilei (1632) *Dialogo dei massimi sistemi*.
- C.B. Garcia, F.G. Gould (1983) An application of homotopy to solving linear programs. *Math. Prog.* 27, 263-282.
- C.F. Gauss (1809) *Theoria motus corporum coelestium*. In F. Perthes, I.H. Besser, *Sectionbus conicis solem ambientium*, Hamburg. Reprinted in his *werke*, vol. 7, F. Pethes, Gotha 1871. English translation by C.H. Davis, Little, Brown and Co., Boston, 1857. Reprinted by Dover Pub. New York, 1963.
- T.E. Harris (1950) Regression using minimum absolute deviations. *Am. Statist.*, 4, 14-15.
- H.L. Harter (1974a) The method of least squares and some alternative, I. *Int. Stat. Rev.*, 42, 147-174.
- H.L. Harter (1974b) The method of least squares and some alternative, II. *Int. Stat. Rev.*, 42, 235-264.
- H.L. Harter (1975a) The method of least squares and some alternative, III. *Int. Stat. Rev.*, 43, 1-44.

- H.L. Harter (1975b) The method of least squares and some alternative, IV Int. Stat. Rev., 43, 125-190, 273-278.
- H.L. Harter (1975c) The method of least squares and some alternative, V. Int. Stat. Rev., 43, 269-272.
- H.L. Harter (1976) The method of least squares and some alternative, VI. Int. Stat. Rev., 44, 113-159.
- P.W. Holland, R.E. Welsch (1977) Robust regression using iteratively reweighted least-squares. Comm. Stat., A6, 813-827.
- L. Horvath (1987) Asymptotic normality of L_p -norms of density estimators. Tech. Rep. series of Lab Res. Stat. Prob., no.3, Carleton University, Ottawa, Canada.
- H. Imai, K. Kato, P. Yamamoto (1987) A linear-time algorithm for linear L_1 approximation of points. Tech. Rep. CSCE-87-C30. Dept. of Comp. Sci. and Commun. Engin., Kyushu University 36, Fukuoka 812, Japan.
- L.A. Josvanger, V.A. Sposito (1983) L_1 -norm estimates for the simple regression problem. Comm. Stat. B12, 215-21.
- O.J. Karst (1958) Linear curve fitting using least deviations. JASA, 53, 118-132.
- Y. Kawara (1979) Straight line fitting by minimizing the sum of absolute deviations. J. of the Japan Stat. Soc., 9, 47-64.
- J.H.B. Kemperman (1984) Least absolute value and median polish. In Y.L. Tong (ed.), Inequalities in statistics and probability (IMS Lecture notes monograph series, vol.5), Inst. of Math. Stat., Hayward, CA, 84-113.
- W.J. Kennedy, J.E. Gentle (1980) Statistical computing. New York, Marcel Dekker.
- P.S. Laplace (1793) Sur quelques points du system du monde. Memoires de l'Academie Royale des Science de Paris. Annee 1789, 1-87. Reprinted in Oeuvres completes de Laplace II. Paris, Gauthier-Villars, 1985, 477-558.
- P.S. Laplace (1799) Traite des mecanique celeste, 2. Paris; J.B.M. Depart. Reprinted as oeuvres completes de Laplace, 2. Paris; Gauthier-Villars 1878, 116-165.
- P.S. Laplace (1812) Theorie analytique des probabilités, Mme courcier Paris 1820 Reprinted in his oeuvres, vol.7, Imprimerie Royale, Paris, 1847, and Gauthier-Villars et fils, Paris 1886.
- P.S. Laplace (1818) Deuxieme supplement to Laplace (1812).
- C.L. Mathieu (1816) Sur les experiences du pendule, faites par les navigateurs espagnol, en differens points du globe. Connaissance des tems, 314-332.
- C.R. McConnell (1987) On computing a best discrete L_1 approximation using the method of vanishing Jacobians. CSDA, 5, 277-288.
- G.F. McCormick, V.A. Sposito (1975) A note on L_1 estimation based on the median positive quotient. Appl. Stat., 24, 347-350.
- M.S. Meketon (1986) Least absolute value regression. Work. Pap., AT&T Bell Laboratories, Holmdel, N.J.
- R.M. Moroney (1961) The Haar problem in L_1 . Proc. Amer. Math. Soc., 12, 793-795.
- S.C. Narula, J.F. Wellington (1985) Interior analysis for the minimum sum of absolute errors regression. Technometrics, 27, 181-188.
- S.C. Narula, J.F. Wellington (1987) An efficient algorithm for the MSAE and MMAE regression problems. Work. Pap., Virginia Commonwealth University, Richmond, VA 23284.
- M.R. Osborne (1987) The reduced gradient algorithm. In Y. Dodge (ed.) Statistical data analysis based on the L_1 norm and related methods. North-Holland. 95-108.
- M.R. Osborne, S.A. Pruess, R.S. Womersley (1986) Concise representation of generalized gradients. J. of Austra. Math. Soc., Ser. B, 28, 57-74.
- M.R. Osborne, G.A. Watson (1985) An analysis of the total approximation problem in separable norms, and an algorithm for the total L_1 problem. SIAM J. Sci. Stat. Comp., 6, 410-424.
- M.J. Panik (1976) Classical optimization: foundation and extensions. North-Holland, Amsterdam.
- U. Peters, C. Willms (1983) Up- and down-dating procedures for linear L_1 regression. OR Spektrum 5, 229-239.
- P. Pilibossian (1987) A direct solving algorithm for a linear regression according to L_1 -norm criteria. Work. Pap., L.S.T.A. Universite, Paris VI
- P. Rabinowitz (1968) Application of linear programming to numerical analysis. SIAM Rev., 10, 121-159.
- P. Rabinowitz (1970) Mathematical programming and approximation. In A. Talbot (ed.) Approximation Theory. Academic Press, 217-231.
- M.R. Rao, V. Srinivasan (1972) A note on Sharpe's algorithm for minimum sum of absolute deviations in a simple regression problem. Manag. Sci., 19, 222-225.
- E.C. Rhodes (1930) Reducing observations by the method of minimum deviations. Philo. Mag., 7th series, 9, 974-92.
- J.R. Rice (1964c) The approximation of functions, vol. I, linear theory. Reading Mass.; Addison-Wesley.
- P.D. Robers, A. Ben-Israel (1969) An interval programming algorithm for discrete linear L_1 approximation problem. J. Approx. Theory, 2, 323-336.

- P.D. Robers, S.S. Robers (1973) Algorithm 458: discrete linear L_1 approximation by interval linear programming. *Comm. ACM*, 16, 629-633.
- E. Ronchetti (1987) Bounded influence in regression: a review. In Y. Dodge (ed.) *Statistical data analysis based on the L_1 norm and related methods*. North-Holland, 65-80.
- A.N. Sadovski (1974) AS74: L_1 -norm fit of a straight line. *Appl. Stat.* 23, 244-248.
- J.P. Schellhorn (1987) Fitting data through homotopy methods In Y. Dodge (ed.) *Statistical data analysis based on the L_1 norm and related methods*. North-Holland. 131-138.
- E.J. Schlossmacher (1973) An iterative technique for absolute deviations curve fitting. *JASA* 68, 857-865.
- E. Seneta (1983) The weighted median and multiple regression. *Austral. J. Stat.*, 25(2), 370-377.
- E. Seneta, W.L. Steiger (1984) A new LAD curve-fitting algorithm: slightly overdetermined equation system in L_1 . *Discrete Applied Math.*, 7, 79-91.
- Shanno, R.L. Weil (1970) Linear programming with absolute value functionals. *Oper. Res.*, 19, 120-124.
- W.F. Sharpe (1971) Mean-absolute deviation characteristic lines for securities and portfolios. *Manag. Sci.*, 18, B1-B13.
- H.D. Sherali, B.O. Skarpness, B. Kim (1987) An assumption-free convergence analysis for a perturbation of the scaling algorithm for linear programs, with application to the L_1 estimation problem. Dept. of Ind. Engin. and OR, Virginia Polytechnic Inst. and State University, Blacksburg, Virginia.
- O.B. Sheynin (1973) R.J. Boscovich's work on probability. *Archive for history of exact sciences*, vol. 9, 306-324, and vol. 28, 173.
- R.R. Singleton (1940) A method for minimizing the sum of absolute values of deviations. *Annals of math. Stat.*, 11, 301-310.
- S.A. Soliman, G.S. Christensen, A. Rouhi (1988) A new technique for curve fitting based on minimum absolute deviations. *CSDA*, 6(4), 341-352.
- V.A. Sposito (1976) A remark on algorithm AS74, L_1 norm fit of a straight line. *Appl. Stat.*, 25, 96-97.
- V.A. Sposito (1987a) On median polish and L_1 estimators. *CSDA*, 5, 155-162.
- V.A. Sposito, W.J. Kennedy, J.E. Gentle (1977) AS110: L_p norm fit of a straight line. *Appl. Stat.*, 26, 114-118.
- V.A. Sposito, G.F. McCormick, W.J. Kennedy (1975) L_1 estimation strategies based on the simplex algorithm. In *Proc. of the eighth symposium on the interface*, J.W. France (ed.) Health science computing facility. UCLA, Los Angeles.
- V.A. Sposito, W.C. Smith (1976) On a sufficient and necessary condition for L_1 estimation. *Appl. Stat.*, 25, 154-157.
- K. Spyropoulos, E. Kiountouzis, A. Young (1973) Discrete approximation in the L_1 norm. *Comp. J.*, 16, 180-186.
- W.L. Steiger (1980) Linear programming via L_1 curve fitting beats simplex. *Abstracts, AMS*, 80T-C26, 385-386.
- S.M. Stigler (1981) Gauss and invention of least squares. *Annals of Stat.*, 9, 465-474.
- S.M. Stigler (1984) Studies in the history of probability and statistics XL, Boscovich, Simpson and a 1760 manuscript note on fitting a linear relation. *Biometrika*, 71, 3, 615-620.
- J. Svanberg (1805) *Exposition des operations faites en lappnie pour la determination d'un arc du meridien en 1801, 1802 et 1803,...* Stockholm.
- J.W. Tukey (1977) *Exploratory data analysis*. Reading, Mass. Addison-Wesley.
- H.H. Turner (1887) On Mr. Edgeworth's method of reducing observations relating to several quantities. *Phil. Mag.* (5th series), 24, 466-470.
- K.H. Usow (1967a) On L_1 approximation: computation for continuous functions and continuous dependence. *SIAM J. of Numer. Anal.*, 4, 70-88.
- K.H. Usow (1967b) On L_1 approximation: computation for discrete functions and discretization effect. *SIAM J. Numer. Anal.*, 4, 233-244.
- J.F. Van Beeck-Calkoen (1816) *Ver de theoric der Gemiddelde Waardij*. *Verhandlingen der K. Nederlandandsch Instituut Can Wetenschappen*, 2, 1-19.
- B.A. Von Lindenau (1806) *Uber den Gebrauch der Gradmessungen zur bestimmung der gestalt der erde*. *Monatliche correspondenz zur befar derung der Erd-und Himmels-kunde*, 14, 113-158.
- H.M. Wagner (1959) Linear programming technique for regression analysis. *JASA*, 54, 202-212.
- G.A. Watson (1981) An algorithm for linear L_1 approximation of continuous functions. *IMA J. Num. Anal.*, 1, 157-167.
- G.O Wesolowsky (1981) A new descent algorithm for least absolute value regression problem. *Comm. Stat.*, B10, 479-491.