Pergamon

Appl. Math. Lett. Vol. 10, No. 4, pp. 65-70, 1997
Copyright©1997 Elsevier Science Ltd
Printed in Great Britain. All rights reserved
0893-9659/97 $17.00 + 0.00

PII: S0893-9659(97)00061-X

Duality in Robust Linear Regression Using Huber's M-Estimator

M. Ç. PINAR
Industrial Engineering Department, Bilkent University
06533 Bilkent, Ankara, Turkey
mustafap@bilkent.edu.tr

(Received October 1996; accepted December 1996)

Communicated by R. H. Byrd

Abstract—The robust linear regression problem using Huber's piecewise-quadratic M-estimator function is considered. Without exception, computational algorithms for this problem have been primal in nature. In this note, a dual formulation of this problem is derived using Lagrangean duality. It is shown that the dual problem is a strictly convex separable quadratic minimization problem with linear equality and box constraints. Furthermore, the primal solution (Huber's M-estimate) is obtained as the optimal values of the Lagrange multipliers associated with the dual problem. As a result, Huber's M-estimate can be computed using off-the-shelf optimization software.

Keywords—Lagrangean duality, Huber's M-estimator, Robust regression, Quadratic programming.

1. INTRODUCTION

There has been considerable interest in the theory and algorithms for robust estimation in the past two decades. In particular, Huber's M-estimator [1] has received a great deal of attention from both theoretical and computational points of view. Robust estimation is concerned with identifying "outliers" among data points and giving them less weight. Huber's M-estimator is essentially the least squares estimator, which uses the ℓ_1-norm for points that are considered outliers with respect to a certain threshold. Hence, the Huber criterion is less sensitive to the presence of outliers.

More precisely, Huber's M-estimate is a minimizer $x^* \in \Re^n$ of the function

$$F(x) = \sum_{i=1}^{m} \rho\left(\frac{r_i(x)}{\sigma}\right), \tag{1}$$

where

$$\rho(t) = \begin{cases} \dfrac{1}{2\gamma} t^2, & \text{if } |t| < \gamma, \\ |t| - \dfrac{1}{2}\gamma, & \text{if } |t| \geq \gamma, \end{cases} \tag{2}$$

with a tuning constant $\gamma > 0$, and a scaling factor σ that depends on the data to be estimated. The residual $r_i(x)$ is defined as

$$r_i(x) = b_i - a_i^{\mathsf{T}} x, \tag{3}$$

Special thanks are due to R. H. Byrd for his careful handling of the paper. One referee pointed out the QP form of the primal added to the revised version. Another referee clarified several issues concerning robust estimators and data analysis, and suggested directions for future research. I am grateful to both.

for all $i = 1, \dots, m$ with $r = b - A^{\mathsf{T}}x$. To view this minimization problem in a more familiar format, define a "sign vector"

$$\mathbf{s}_\gamma(x) = [s_{\gamma 1}(x), \dots, s_{\gamma m}(x)] \tag{4}$$

with

$$s_{\gamma i}(x) = \begin{cases} -1, & \text{if } r_i(x) \le -\gamma, \\ 0, & \text{if } |r_i(x)| < \gamma, \\ 1, & \text{if } r_i(x) \ge \gamma, \end{cases} \tag{5}$$

and

$$\mathbf{W_s} = \operatorname{diag}(w_1, \dots, w_m), \tag{6}$$

where

$$w_i = 1 - s_i^2. \tag{7}$$

Now, assuming a unit σ for the moment, Huber's M-estimation problem can be expressed as the following minimization problem.

PROBLEM [**P**].

$$\text{minimize } F(x) \equiv \frac{1}{2\gamma} r^{\mathsf{T}} \mathbf{W_s} r + \mathbf{s}_\gamma^{\mathsf{T}} \left[r - \frac{1}{2}\gamma \mathbf{s}_\gamma \right], \tag{8}$$

where the argument x of r is dropped for notational convenience. Clearly, F measures the "small" residuals ($|r_i(x)| < \gamma$) by their squares, while the "large" residuals are measured by the ℓ_1 function. Thus, F is a piecewise quadratic function, and it is once continuously differentiable in $\Re^n$.

The contribution of the present note is to introduce a dual approach to this computationally intensive problem. A simple derivation is used to give a dual problem which turns out to be a problem familiar to the numerical optimization community: linearly constrained separable convex quadratic programming. To the best of the author's knowledge, this duality relationship has not been noticed before. This simple result reduces the M-estimation problem to one that is easily solved using off-the-shelf optimization software.

The interest in Huber's M-estimator derives from its perceived utility as a robust estimation procedure. In this context, γ is an important quantity as it controls the spread of the residuals. This suggests that it should be related to the scaling factor σ. Some algorithms [2,3] estimate σ only once at the beginning using some rules of thumb, and keep it fixed throughout the computation. Another approach is to estimate σ iteratively by treating it as an independent variable. This is used by Huber in [4] who suggests σ to be computed from an auxiliary equation involving σ, γ, and x. Shanno and Rocke [5] and Ekblom [6] also use this approach in their respective papers. An alternative way to connect these two parameters and to iteratively estimate σ is to set $\sigma = 1$ for any values of γ and σ, and to replace γ with $\gamma\sigma$. Hence, there is no loss of information in using γ only, while γ and σ are both allowed to vary during the computation. Clark and Osborne [7] use this observation in their algorithm that computes both σ and γ. This is a continuation method, so that at each stage they have the M-estimate for the current γ. Clark and Osborne also describe a partitioning algorithm to resolve degenerate situations in the continuation algorithm. The continuation method is essentially based on tracing a curve of M-estimates as a function of γ. Their approach consists of two stages.

1. Construction of a solution for a particular γ. The easiest cases are $\gamma = 0$ (ℓ_1 estimate) and $\gamma = \infty$ (least squares). These are both used in [7].
2. Continuation with respect to γ to solve the data analysis problem.

A potential contribution of the present paper in the above context is that it suggests a new method for obtaining Huber's M-estimate for a particular value of γ. Then, continuation with respect to γ of a solution obtained by the dual method can be pursued using methods of parametric quadratic programming. In this connection, the rule of thumb choices that have been suggested for γ in the literature [2-4] can be used to see if they provide better starting points for the continuation method than least squares and ℓ_1 estimates. This is a topic for future research.

When σ and γ are fixed, most algorithms that have proved successful for the computation of Huber's M-estimate have been iterative in nature. Since F is only once continuously differentiable, research concentrated on developing successful applications of Newton's method to this problem, and on studying ideas such as the iteratively reweighted least squares (IRLS). Among these, Huber and Dutter's method [8] and Huber's [4] apply Newton's method to the nonlinear equation system:

$$A\left(\frac{\mathbf{W}_{\mathbf{s}} r(x)}{\gamma} + \mathbf{s}_\gamma(x)\right) = \mathbf{0}, \tag{9}$$

which represent the optimality conditions. The IRLS algorithm is attributed to Beaton and Tukey [9]. This algorithm has been discussed in several other papers as well [2,3,10-12]. A brief review of these algorithms is given in [7].

Madsen and Nielsen gave finite modified Newton algorithms for the minimization of the Huber function in [13]. These algorithms capitalize on the piecewise quadratic nature of the function. The idea is that if the sign vector associated with a minimizer x^*, $\mathbf{s}^*$ say, were known, then the minimizer could be computed using one step of Newton's method. Since sign vectors correspond to a subdivision of $\Re^n$ into a finite number of subregions, the methods of Madsen and Nielsen reduce to a search for the correct subregion.

2. A DUAL PROBLEM

In this section, a dual problem to [**P**] is derived using Lagrangean duality. The interested reader is directed to the book by Rockafellar [14] for a detailed exposition of Lagrangean duality. We use a single parameter γ to mean $\gamma\sigma$ as discussed in the previous section.

Consider the problem [**P**] in a slightly different form:

$$\min_x F(x) \equiv \Phi\left(b - A^\top x\right). \tag{10}$$

Let $u = b - A^\top x$, and rewrite the problem as:

$$\begin{array}{ll} \min & \Phi(u) \\ \text{s.t.} & b - A^\top x = u. \end{array}$$

Associating the multipliers $y \in \Re^m$ with the equality constraints $b - A^\top x = u$, we get the following Lagrangean problem:

$$\max_y \min_{x,u} \left\{\Phi(u) + y^\top \left(b - A^\top x - u\right)\right\}. \tag{11}$$

This is equivalent to:

$$\max_y \left[y^\top b + \min_u \left\{\Phi(u) - y^\top u\right\} + \min_x \left\{-y^\top A^\top x\right\}\right]. \tag{12}$$

Observing that x is a free variable, the term

$$\min_x \left\{-y^\top A^\top x\right\} \tag{13}$$

yields the constraint

$$Ay = \mathbf{0}. \tag{14}$$

It remains to simplify the term

$$\min_u \left\{\Phi(u) - y^\top u\right\}. \tag{15}$$

Simple calculus shows that

$$\min_u \{\Phi(u) - y^\top u)\} = \begin{cases} -\frac{1}{2}\gamma y^\top y, & \text{if } -\mathbf{1} \le y \le \mathbf{1}, \\ -\infty, & \text{otherwise.} \end{cases} \tag{16}$$

Hence, the dual problem is the following.
PROBLEM [**D**].

$$\begin{aligned} \max \quad & b^\top y - \frac{1}{2}\gamma y^\top y \\ \text{s.t.} \quad & Ay = \mathbf{0} \\ & -\mathbf{1} \le y \le \mathbf{1}. \end{aligned}$$

An alternative way to arrive at the above dual is to pose the primal problem [**P**] as a quadratic programming problem:

$$\begin{aligned} \max \quad & \frac{1}{2\gamma}\sum_{i=1}^m p_i^2 + \sum_{i=1}^m \left(q_i - \frac{\gamma}{2}\right) \\ \text{s.t.} \quad & -p - q \le b - A^\top x \le p + q \\ & p \le \gamma e \\ & q \ge \mathbf{0}. \end{aligned}$$

where e denotes a vector with all components unity. This is not surprising since the dual of a quadratic program is again a quadratic program. However, this alternative derivation is substantially longer since it requires some transformations on the resulting dual to be cast in the form [**D**]. So, it is not included into the present paper.

The remarkable fact about the duality result is that the dual we have derived is a quadratic programming problem with a strictly concave separable objective function, linear equality, and box constraints. This is important in two respects. First, this dual problem is an intensely researched, numerically well-solved problem. An excellent software system is available from Stanford University for the solution of quadratic programming problems [15,16]. Furthermore, numerical procedures for solving this problem are part of almost any subroutine library. Such procedures are also available through matrix manipulation packages such as Matlab™ and Octave [17]. Second, it has been shown to be polynomially solvable [18], and efforts to turn related algorithms into reliable software on this front are also under way. In particular, quadratic programming versions of the software systems CPLEX™ and LoQo are available [19]. Hence, any advances made in the fast and accurate solution of the quadratic programming problem will benefit Huber's M-estimation problem.

Note that any numerical procedure which yields a primal-dual solution to [**D**] gives a minimizer x^* of the primal problem, which is really of interest here. To see this, it is instructive to derive a dual problem to [**D**] using Lagrangean duality.

Associating multipliers $v \in \Re^n$ with the constraints $Ay = \mathbf{0}$, we get the following Lagrangean problem:

$$\min_v \max_{-\mathbf{1}\le y \le \mathbf{1}} \left\{ b^\top y - \frac{1}{2}\gamma y^\top y + v^\top(-Ay + \mathbf{0}) \right\}. \tag{17}$$

Now, rearranging terms and using $r \equiv b - A^\top v$, simple calculus shows that:

$$\max_{-1\le y_i \le 1} \left\{ y_i r_i - \frac{1}{2}\gamma y_i^2 \right\} = \begin{cases} \frac{1}{2\gamma} r_i^2, & \text{if } |r_i| < \gamma, \\ |r_i| - \frac{1}{2}\gamma, & \text{if } |r_i| \ge \gamma. \end{cases} \tag{18}$$

But, this is precisely Huber's M-estimator function. Therefore, the optimal values of the dual multipliers associated with the equality constraints in [**D**] give precisely Huber's M-estimate.

Further insight into the relationship between [P] and [D] is gained through Theorem 1 below which links the optimal solutions of [P] and [D]. This theorem shows that the optimal solution to [D] is obtained as the first derivative of the function F with respect to r at any primal optimal point. Before stating this result, we quote the following property that was proved in [20].

LEMMA 2.1. *Let x be a minimizer of F for some value of $\gamma > 0$, and let $\mathbf{s} = \mathbf{s}_\gamma(x)$ with $\mathbf{W_s}$ defined accordingly. Also, let $r = b - A^{\mathsf{T}}x$. Then, r_i is constant for all i such that $s_i = 0$. Furthermore, $\mathbf{s}_\gamma$ is constant for any x that minimizes F.*

By strict concavity, it is clear that the optimal solution to [D] is unique.

THEOREM 2.1. *Let x be a minimizer of F for some value of $\gamma > 0$, and let $\mathbf{s} = \mathbf{s}_\gamma(x)$ with $\mathbf{W_s}$ defined accordingly. Then, the unique optimal solution of [D] is given as:*

$$y \equiv \frac{\mathbf{W_s} r(x)}{\gamma} + \mathbf{s}. \tag{19}$$

PROOF. Associate multipliers x with the equality constraints and nonnegative multipliers $\alpha, \beta \in \Re^m$ with the box constraints. From [6, Theorem 28.3], it is well known that for y to be an optimal solution for [D], and for (x, α, β) to be a Lagrange multiplier vector, it is necessary and sufficient that (y, x, α, β) be a saddlepoint of the Lagrangean of [D] as defined in [6, p. 280]. This condition holds if and only if the components of (y, x, α, β) satisfy the following conditions:

$$\gamma y - b + A^{\mathsf{T}}x + \alpha - \beta = \mathbf{0}, \tag{20}$$

$$Ay = \mathbf{0}, \tag{21}$$

$$\alpha_i(y_i - 1) = 0, \qquad \text{for all } i = 1, \dots, m, \qquad \text{and} \tag{22}$$

$$\beta_i(-y_i - 1) = 0, \qquad \text{for all } i = 1, \dots, m. \tag{23}$$

From the first condition (20), we obtain

$$y = \frac{b - A^{\mathsf{T}}x}{\gamma} - \frac{\alpha - \beta}{\gamma}. \tag{24}$$

Let $r \equiv b - A^{\mathsf{T}}x$ and consider three cases.

CASE 1. If $-1 < y_i < 1$, clearly, $\alpha_i = \beta_i = 0$. This implies that $|r_i(x)| < \gamma$ with $\mathbf{s}_{\gamma i}(x) = 0$, i.e., $y_i = r_i(x)/\gamma$.

CASE 2. If $y_i = 1$, this implies that $\beta_i = 0$. Hence, $r_i(x) = \gamma + \alpha_i$. Therefore, $r_i(x) \ge \gamma$ with $\mathbf{s}_{\gamma i}(x) = 1$, i.e., $y_i = \mathbf{s}_{\gamma i}(x) = 1$.

CASE 3. If $y_i = -1$, one has $\alpha_i = 0$. Hence, similarly to Case 2 above, $r_i(x) \le -\gamma$ with $\mathbf{s}_{\gamma i}(x) = -1$, i.e., $y_i = \mathbf{s}_{\gamma i}(x) = -1$.

Therefore, an alternative expression for y is given as:

$$y \equiv \frac{\mathbf{W_s} r(x)}{\gamma} + \mathbf{s}_\gamma. \tag{25}$$

The result now follows from the previous lemma. ∎

REFERENCES

1. P. Huber, *Robust Statistics*, John Wiley, New York, (1981).
2. J.B. Birch, Some convergence properties of iterated reweighted least squares in the location model, *Comm. Statist.* **B9**, 359–369 (1980).
3. P.W. Holland and R.E. Welsch, Robust regression using iteratively reweighted least squares, *Comm. Statist.* **A6**, 813–827 (1977).
4. P. Huber, Robust regression: Asymptotics, conjectures and Monte Carlo, *Annals of Statistics* **1**, 799–821 (1973).
5. D.F. Shanno and D.M. Rocke, Numerical methods for robust regression: Linear models, *SIAM J. on Scientific and Statistical Computing* **7**, 86–97 (1986).
6. H. Ekblom, A new algorithm for the Huber estimator in linear models, *BIT* **28**, 123–132 (1988).
7. D.I. Clark and M.R. Osborne, Finite algorithms for Huber's M-Estimator, *SIAM J. on Scientific and Statistical Computing* **7**, 72–85 (1986).
8. P. Huber and R. Dutter, Numerical solution of robust regression problems, In *COMPSTAT 1974 Proc. Symposium on Computational Statistics*, (Edited by G. Brushmann), pp. 165–172, Physike Verlag, Berlin, (1974).
9. A.E. Beaton and J.W. Tukey, The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data, *Technometrics* **16**, 147–185 (1974).
10. J.B. Birch, Effects of the starting value and stopping rule on robust estimates obtained by iterated weighted least squares, *Comm. Statist.* **B9**, 133–140 (1980).
11. R.H. Byrd and D.A. Pyne, Some results on the convergence of the iteratively reweighted least squares algorithm for robust regression, In *Proc. Statistical Computing Section*, pp. 87–90, American Statistical Association, (1979).
12. S.A. Ruzinsky and E.T. Olsen, L_1 and L_∞ minimization via a variant of Karmarkar's algorithm, *IEEE Transactions on Acoustics, Speech and Signal Processing* **37**, 245–253 (1989).
13. K. Madsen and H.B. Nielsen, Finite algorithms for robust linear regression, *BIT* **30**, 682–699 (1990).
14. R.T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, (1970).
15. P.E. Gill, W. Murray, M.A. Saunders and M.H. Wright, User's guide for SOL/QSPOL (Version 3.2), Technical Report SOL 84-6, Systems Optimization Laboratory, Department of Operations Research, Stanford University, Stanford, CA, (1984).
16. P.E. Gill, S.J. Hammarling, W. Murray, M.A. Saunders and M.H. Wright, User's guide for LSSOL (Version 1.0): A Fortran package for constrained linear least squares and convex quadratic programming, Technical Report SOL 86-1, Systems Optimization Laboratory, Department of Operations Research, Stanford University, Stanford, CA, (1986).
17. J.W. Eaton, *OCTAVE: A High-Level Interactive Language for Numerical Computation*, (1994).
18. Y. Ye and E. Tse, An extension of Karmarkar's projective algorithm for convex quadratic programming, *Mathematical Programming* **44**, 157–179 (1989).
19. R.J. Vanderbei and T.J. Carpenter, Symmetric indefinite systems for interior point methods, *Mathematical Programming* **58**, 1–32 (1993).
20. K. Madsen, H.B. Nielsen and M.Ç. Pınar, New characterizations of ℓ_1 solutions to overdetermined systems of linear equations, *Operations Research Letters* **16**, 159–166 (1994).
21. D.I. Clark, The mathematical structure of Huber's M-estimator, *SIAM J. on Scientific and Statistical Computing* **6**, 209–219 (1985).