한국통계학회논문집 제3권 2호
1996년 8월 pp. 299-307

L_∞-estimation based Algorithm for the Least Median of Squares Estimator[1)]

Bu-yong Kim[2)]

Abstract

This article is concerned with the algorithms for the least median of squares estimator. An algorithm based on the L_∞-estimation procedure is proposed in an attempt to improve the optimality of the estimate. And it is shown that the proposed algorithm yields more optimal estimate than the traditional resampling algorithms. The proposed algorithm employs a linear scaling transformation at each iteration of the L_∞-algorithm to deal with its computational inefficiency problem.

1. Introduction

Consider the problem of estimating the parameters of a multiple linear regression model

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $\boldsymbol{y}$ denotes an n-vector of response variable, X a full-rank $n \times p$ matrix of regressor variables with the first column of ones, $\boldsymbol{\beta}$ a p-vector of regression parameters, and $\boldsymbol{\varepsilon}$ an n-vector of random errors.

Rousseeuw (1984) has proposed a high breakdown point estimator which is called the least median of squares (LMS) estimator given by

$$\underset{\hat{\beta}}{\text{minimize}} \quad \underset{i}{\text{median}} \quad e_i^2, \tag{1.2}$$

where e_i denotes the i-th residual. It has been shown under an assumption that a solution to the problem (1.2) exists. Moreover, it turns out that this estimator has robustness property with respect to leverage points as well as vertical outliers. Rousseeuw (1984) compares the LMS estimate with four important robust competitors, and presents substantial advantage of

1) This paper was supported in part by Sookmyung Women's University, 1996.
2) Associate Professor, Department of Statistics, Sookmyung Women's University, Seoul 140-742, Korea.

the LMS estimator. In fact, it has been proved that if the observations are in general position, the finite sample breakdown point of the LMS estimator is $([\ n/2]\ -p+2)/n$, (where the notation $[\ \cdot\ ]$ stands for the largest integer function) which is asymptotically equal to 0.5, the best that can be expected. Another feature of interest is that the LMS estimator is regression equivariant, scale equivariant, and affine equivariant. These properties are extensively reviewed by Bassett (1991). Furthermore, it has the exact fit property, which means that when at least $n-[\ n/2]\ +1$ of the observations satisfy the relation $y=X\hat{\beta}$ exactly and are in general position, then the LMS estimate equals to $\hat{\beta}$ whatever the other observations are. Due to these desirable theoretical properties the LMS estimator is widely used for robust regression, regardless of the drawback that it is less efficient than other robust estimators. An approach to increase the efficiency of the LMS estimator has been suggested by Yohai (1987).

In this article we shall primarily concentrate on the optimality property of the LMS algorithm and on comparisons of the behavior and performance of the algorithms. Section 2 briefly describes the traditional resampling algorithms for LMS estimation suggested by Rousseeuw (1984) and Marazzi (1991). In Section 3 an algorithm based on the L_∞-estimation procedure is proposed in order to improve the optimality of the estimate. Section 4 discusses empirical comparisons of algorithms in terms of optimality, and it is shown that the proposed algorithm yields more optimal solution than the competitors.

2. Resampling Algorithms for LMS estimation

It is probably impossible to construct a closed form expression for the LMS estimator. For this reason, algorithms for the LMS estimation have been developed in the recent past. Steele and Steiger (1986) suggest an algorithm for the case of simple regression model, with the comment that since the LMS objective function can have high number $(O(n^2))$ of local minima, any gradient based minimization algorithms may fail to find the global minimum. Also, Souvaine and Steele (1987) provide two combinatorial algorithms for simple regression case by converting (1.2) into a discrete optimization problem. However, their computational complexity appears to be worse than $O(n^p)$. As a consequence, resampling approximation algorithms as suggested in Rousseeuw (1984) and Marazzi (1991) are of great interest.

The basic resampling algorithm starts with partitioning the n observations into two parts such that $y'=(\ y_E',\ y_{\bar{E}}')$, $X'=(X_E', X_{\bar{E}}')$, where $E=\{E_1, E_2, \ldots, E_p\}$ is a subset of distinct indices from the set $\{1,2,\ldots,n\}$, and $\bar{E}$ is the complementary set. For each subsample corresponding to the subset E, the solution of a system of p linear equations is

obtained, which is called the trial estimate $\widehat{\beta_E}$. And the predicted residuals $e = y - X\widehat{\beta_E}$ and the corresponding LMS objective function (median of the squared residuals) with respect to the whole points in the data set is computed. Actually, the h-th order statistic is usually taken as a median, where $h = [\ n/2] + [\ (p+1)/2]$ with which the maximum possible breakdown point is attained. Then the trial estimate with minimum value of objective function is finally treated as an approximate LMS estimate. In fact, the resampling algorithm does not, in general, provide the exact estimates, but the approximate ones.

Rousseeuw and Leroy (1987) provide a computer program PROGRESS which is designed to run on microcomputer. Actually, this program adapts the approximate method in which one may choose the faster version or the extensive search version since a lot of computation is required when all possible subsamples are considered. The program proceeds by repeatedly selecting m (predetermined) subsamples of p different observations. Rousseeuw and Leroy (1987) present the values of m which are actually employed in their program for different combinations of n and p.

On the other hand, Marazzi(1991) suggests a variant of resampling algorithm which adopts the least squares method based on the subsample of $q(>p)$ observations, rather than the exact fit method based on the subsample of p observations as in Rousseeuw's algorithm. This approximation approach is known to yield better estimate than the basic resampling scheme, and share the properties, such as the equivariance and high breakdown point, of the exact estimate. Marazzi(1993) provides the computer program HYLMSE for this algorithm.

3. Proposed Algorithm for LMS Estimation

From the computational studies, several notable facts have been found with respect to optimality of solutions from the traditional resampling algorithms. (Details of computational results are described in Section 4.) Thus, an algorithm is proposed in an attempt to obtain more optimal solutions for the LMS estimation. The LMS estimator, in fact, can be treated as a special case of a larger family of estimators, namely the least quantile of squares (LQS) estimator which is defined by

$$\min_{\hat{\beta}} \text{imize} \quad \tilde{e}_{([\ (1-\alpha)n] + [\ \alpha(p+1)])}, \quad 0 \le \alpha \le 0.5, \tag{3.1}$$

where $\tilde{e}_{(\cdot)}$ denotes the ordered absolute residual. Obviously, the LMS estimator is asymptotically equivalent to the LQS estimator for α tending to 0.5 since the squared residual in problem (1.2) can be replaced by the absolute residual. In this context, one possibility is to make use of the L_∞-estimation procedure in order to improve the algorithms

in terms of optimality. The reason is that the LMS estimation on the whole sample is the L_∞-estimation over subsample corresponding to the h smallest absolute residuals.

For completeness, this paper introduces the L_∞-estimation procedure for linear regression model (1.1). Suppose that a subsample of size h is taken. The L_∞-estimation problem is to find regression coefficients that minimize the L_∞-norm of the residual

$$\min_{\hat{\beta}} \text{imize} \ \| e \|_\infty . \tag{3.2}$$

This problem is a kind of minimax problem, and may be solved by linear programming technique. The linear programming formulation for the L_∞-estimation problem (3.2) is as follows

$$\begin{aligned} &\text{minimize} \quad [\ 0 \ \ 1] \begin{bmatrix} \hat{\beta} \\ \lambda \end{bmatrix} \\ &\text{subject to} \quad \begin{bmatrix} X & \ell \\ -X & \ell \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \lambda \end{bmatrix} \geq \begin{bmatrix} y \\ -y \end{bmatrix} , \end{aligned} \tag{3.3}$$

where λ denotes the value of the maximum absolute residual that is to be minimized, and $\ell = (1, \ldots, 1)' \in R^h$. For the computational efficiency, the dual problem corresponding to the formulation (3.3) is usually solved,

$$\text{maximize} \ \{ c'\gamma : \ A\gamma = b, \ \gamma \geq 0 \} \tag{3.4}$$

where

$$c = \begin{bmatrix} y \\ -y \end{bmatrix} , \ \gamma = \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} , \ A = \begin{bmatrix} X' & -X' \\ \ell' & \ell' \end{bmatrix} , \ b = \begin{bmatrix} 0 \\ 1 \end{bmatrix} ,$$

$\gamma \in R^{2h}$, and ξ_1 and ξ_2 represent the dual variables. The solution of this problem can be readily obtained by any variants of simplex method. However, since the matrix A is of large dimension in particular when the data set is large, it requires a great deal of computation to solve the problem by the simplex-type algorithms.

In an effort to deal with this computational inefficiency problem, in this article, a linear scaling transformation scheme is employed at each iteration of the L_∞-estimation procedure (3.4). This method commences with an initial feasible solution $\gamma_{\langle 0 \rangle} > 0$, which in practice can be obtained easily by adding one artificial variable γ_{2h+1} to the A matrix ; $\hat{A} = [\ A : b - A\hat{\ell}\ ]$, $\hat{\ell} = (1, \cdots, 1) \in R^{2h}$, and assigning big M to the artificial variable. For

notational simplicity, we shall use the structure same as the problem (3.4) in the following illustration. At every iteration, given a feasible solution $\gamma_{\langle k\rangle} > 0$ for (3.4), it employs the scaling linear transformation

$$\gamma = \Gamma_{\langle k\rangle} \tau, \quad \text{where } \Gamma_{\langle k\rangle} = \text{diag}\{\gamma_{\langle k\rangle 1}, \ldots, \gamma_{\langle k\rangle 2h}\}.$$

This transformation reformulates the problem (3.4) in terms of τ coordinates as follows,

$$\text{maximize } \{(\Gamma_{\langle k\rangle} c)'\tau : A\Gamma_{\langle k\rangle}\tau = b, \ \tau \geq 0\}. \tag{3.5}$$

Now, the projection $p_{\langle k\rangle}$ of the gradient of the objective function, in the transformed space, onto the null space of the equality constraints is determined as

$$p_{\langle k\rangle} = [I - (A\Gamma_{\langle k\rangle})'(A\Gamma^2_{\langle k\rangle}A')^{-1}A\Gamma_{\langle k\rangle}]\ \Gamma_{\langle k\rangle} c. \tag{3.6}$$

Because X is assumed to be of full column rank, A is of full row rank. Hence $A\Gamma^2_{\langle k\rangle}A'$ is nonsingular since $\gamma_{\langle k\rangle} > 0$. A step size $\eta_{\langle k\rangle}$ is taken from the current iterate $\tau_{\langle k\rangle} = \Gamma^{-1}_{\langle k\rangle}\gamma_{\langle k\rangle}$, an n vector of ones, along the projected gradient direction $p_{\langle k\rangle}$. This amounts to taking a step size $\eta_{\langle k\rangle}$ along $d_{\langle k\rangle} = \Gamma_{\langle k\rangle} p_{\langle k\rangle}$ in the γ space, yielding the new iterate such as

$$\gamma_{\langle k+1\rangle} = \gamma_{\langle k\rangle} + \eta_{\langle k\rangle} d_{\langle k\rangle}.$$

To ensure the feasibility of new point $\gamma_{\langle k+1\rangle}$, the step size $\eta_{\langle k\rangle}$ should be chosen as

$$\eta_{\langle k\rangle} = \delta\eta_{\max}, \text{ where } 1/\eta_{\max} = \max_{i=1,\ldots,2h}\{-d_{\langle k\rangle i}/\gamma_{\langle k\rangle i}\} > 0, \ 0 < \delta < 1.$$

This completes an iteration. It can be verified that if $p_{\langle k\rangle} = 0$ for some $\gamma > 0$, then any feasible solution is optimal, and if $p_{\langle k\rangle} \geq 0$ and $p_{\langle k\rangle} \neq 0$ for some γ, then the problem is unbounded. Note that the latter condition does not occur under the assumptions mentioned above, so the algorithm terminates. Therefore, the algorithm continues until the termination criterion

$$\| p_{\langle k\rangle} \|_\infty < \omega, \text{ for small enough } \omega > 0$$

is satisfied. Simulation studies indicate that $\delta = 0.97$ and $\omega = 10^{-8}$ appear to work quite well. The convergence of this type of algorithm has been proved by Sherali *et al.*(1988). Also it has been shown that this approach requires a reduced computational effort, in particular for large data sets.

When the termination criterion is met, the solution $\hat{\boldsymbol{\beta}}$ of the primal problem (3.3) can be obtained as follows. It follows from (3.6) that $\Gamma_{\langle k\rangle}\boldsymbol{c}$ is in the orthogonal complement of the null space of $A\Gamma_{\langle k\rangle}$. There exists a vector $\boldsymbol{w}$ such that

$$\boldsymbol{w}'A\Gamma_{\langle k\rangle}=(\Gamma_{\langle k\rangle}\boldsymbol{c})' \tag{3.7}$$

since the orthogonal complement of the null space of a matrix is the row space of that matrix. Furthermore, $\boldsymbol{w}$ is the vector of dual variables corresponding to the constraint $A\Gamma_{\langle k\rangle}\boldsymbol{r}=\boldsymbol{b}$ of the problem (3.5). And the scaling leaves the duals with respect to the problem (3.4) unchanged. Since $A\Gamma^2_{\langle k\rangle}A'$ is nonsingular from the assumption, (3.7) can be rewritten as

$$\boldsymbol{w}=(A\Gamma^2_{\langle k\rangle}A')^{-1}A\Gamma^2_{\langle k\rangle}\boldsymbol{c}.$$

Therefore, current estimate $\widehat{\boldsymbol{\beta}_E}$ consists of the first p entries of the vector $\boldsymbol{w}$.

In principle, L_∞-estimation should be applied to the subsample of size h. Unfortunately, in consequence of its combinatorial complexity, it requires so much computation that the proposed algorithm considers the subsample of size $p+2$ in order to avoid this problem. (Experience from the simulation study has shown that optimality of the algorithm with subsample of size $p+2$ is a little better than that with subsample of size $p+1$.) The proposed algorithm is differ from the traditional resampling algorithms in that the size of subsample is $p+2$ rather than p (in Rousseeuw's algorithm) or $p+1$ (one case in Marazzi's algorithm), and the trial estimate is computed by the L_∞-estimation procedure rather than by the exact fit of the system of linear equations (in Rousseeuw's algorithm) or the least squares method (in Marazzi's algorithm). In other words, this algorithm generates all possible subsamples of $p+2$ data points and, for each subsample, computes the L_∞-estimates that are the trial LMS estimates. The h-th largest absolute residual is picked as the median absolute residual, and then the trial estimate with the smallest median of absolute residual is determined as an final estimate. The proposed algorithm is described in detail as follows.

[Algorithm LINFLMS]

Initialization : Set the number of iteration $t=1$, and a very large number $Q^*=\infty$. Construct s all possible subsamples of size $p+2$ from the n observations.

Step 1 : For one of the subsamples, compute the L_∞-estimate $\widehat{\beta_E}$, and predicted residuals $e = y - X\widehat{\beta_E}$.

Step 2 : Find the objective function value, that is, the h-th absolute residual $Q = \tilde{e}_{(h)}$, where $h = [n/2] + [(p+1)/2]$.

Step 3 : If $Q < Q^*$, set $Q^* = Q$ and $\widehat{\beta}^* = \widehat{\beta_E}$.

Step 4 : If $t = s$, then return $\widehat{\beta}^*$ as the LMS estimate. Otherwise, set $t = t+1$ and go to step 1.

4. Performance Comparisons of Algorithms

The proposed algorithm LINFLMS is implemented on PC with the source program written in FORTRAN. In order to evaluate the optimality of LINFLMS, computational studies are conducted to measure to what extent the objective function values obtained by the LINFLMS differ from them computed by resampling algorithms PROGRESS and HYLMSE. Since any test problem generators for the LMS estimation such as L1GNR for the L_1-estimation have not been developed, it is not possible to generate data sets which have the same true regression coefficients. Therefore, computational studies are conducted on the basis of well-known real data sets which are introduced in the literatures on robust regression. (See Rousseeuw and Leroy (1987) and Hadi and Simonoff (1993) for further details on the data sets.)

The optimality of the proposed algorithm is compared empirically with other three algorithms (PROGRESS = Rousseeuw's algorithm; HYLMSE1 = Marazzi's algorithm with subsample of size $p+1$; HYLMSE2 = Marazzi's algorithm with subsample of size $p+2$) in terms of relative value of the least median of absolute residuals computed by the algorithms. To compare the performance of algorithms on the same basis, trial estimates are obtained for all possible subsamples in PROGRESS, HYLMSE1, and HYLMSE2. The computational results are summarized in Table 1. As expected, the results show that HYLMSE1 and HYLMSE2 yield more optimal estimates than PROGRESS, but there is no significant difference between HYLMSE1 and HYLMSE2. In addition, it is clear from the computational studies that the smallest value of the objective function is produced by the proposed algorithm, that is, LINFLMS leads to much improvement in optimality. It also implies that the estimates computed by PROGRESS and HYLMSE are, on the whole, very approximate ones.

Table 1. Relative comparisons on the least median of absolute residuals from 4 algorithms

Subject	(n p)	PROGRESS	HYLMSE1	HYLMSE2	LINFLMS
Inflation in China	(9 2)	1.000000	0.908380	0.885098	0.788043
Monthly payments	(12 2)	1.000000	0.841636	0.852258	0.812478
Pension funds	(18 2)	1.000000	0.957732	0.947260	0.938027
Phosphorus content	(18 3)	1.000000	0.909941	0.760703	0.745351
Cloud point	(19 2)	1.000000	0.984559	0.942859	0.910712
Pilot-plant	(20 2)	1.000000	0.948619	0.920867	0.899457
Wood specific gravity	(20 6)	1.000000	0.828716	0.992682	0.834814
Coleman	(20 6)	1.000000	0.781332	0.808537	0.618161
Stackloss	(21 4)	1.000000	0.959200	0.974322	0.911852
Aircraft	(23 5)	1.000000	0.830028	0.789298	0.692597
Number of phone calls	(24 2)	1.000000	0.991830	0.978707	0.963791
Hadi-Simonoff	(25 3)	1.000000	0.999979	0.970289	0.919915
Delivery time	(25 3)	1.000000	0.950108	0.943446	0.918436
Salinity	(28 4)	1.000000	0.889584	0.871475	0.840329
Air quality	(31 4)	1.000000	0.995728	0.982626	0.960438
Hertzsprung-Russell	(47 2)	1.000000	0.934140	0.930275	0.928572
Average		1.000000	0.919470	0.909419	0.855186

5. Concluding Remarks

An attempt has been made to modify the traditional resampling algorithms and to develop a better algorithm. It is worth noting that if the subsample size is set equal to p instead of $p+2$, then the proposed algorithm LINFLMS is nothing else but the algorithm PROGRESS. Although only the limited number of data sets are investigated in this short contribution, the computational results indicate that the proposed algorithm yields more optimal estimates in the LMS estimation. However, the proposed algorithm is so computationally intractable that much effort should be made to improve the algorithm with respect to computational efficiency. One of the possibilities may be to update the projection $\boldsymbol{p}_{\langle k \rangle}$ at each iteration of the L_∞ -algorithm. Of course, this algorithm can also make use of random selection when less computation is required. That is, any approximate versions can be employed in the proposed algorithm by assigning the appropriate replication number less than s.

Acknowledgements : The author is very grateful to Dr. Peter Rousseeuw for providing a copy of the PROGRESS program and data sets, and to a referee for useful suggestion on the entries of Table 1.

References

[1] Basset, Jr. G. W. (1991). Equivariant, Monotonic, 50% Breakdown Estimators, *The American Statistician*, Vol. 45, 135-137.

[2] Hadi, A. S. and Simonoff, J. S. (1993). Procedures for the Identification of Multiple Outliers in Linear Models, *Journal of the American Statistical Association*, Vol. 88, 1264-1272.

[3] Marazzi, A. (1991). Algorithms and Programs for Robust Linear Regression, in *Directions in Robust Statistics and Diagnostics : Part I*, eds. W. Stahel and S. Weisberg, New York: Springer-Verlag, 183-199.

[4] Marazzi, A. (1993). *Algorithms, Routines, and S Functions for Robust Statistics*, Wadsworth, Inc., California.

[5] Rousseeuw, P. J. (1984). Least Median of Squares Regression, *Journal of the American Statistical Association*, Vol. 79, 871-880.

[6] Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*, Wiley-Interscience, New York.

[7] Sherali, H. D., Skarpness, B. O., and Kim, B. Y. (1988). An Assumption-Free Convergence Analysis for a Perturbation of the Scaling Algorithm for Linear Programs, with Application to the L_1 Estimation Problem, *Naval Research Logistics*, Vol. 35, 473-492.

[8] Souvaine, D. L. and Steele, J. M. (1987). Time- and Space-Efficient Algorithms for Least Median of Squares Regression, *Journal of the American Statistical Association*, Vol. 82, 794-801.

[9] Steele, J. M. and Steiger, W. L. (1986). Algorithms and Complexity for Least Median of Squares Regression, *Discrete Applied Mathematics*, Vol. 14, 93-100.

[10] Yohai, V. J. (1987). High Breakdown Point and High Efficiency Robust Estimates for Regression, *The Annals of Statistics*, Vol. 15, 642-656.