The influence of the design on the breakdown point of ℓ_1-type M-estimators

Ivan Mizera
Christine H. Müller

ABSTRACT: Mizera and Müller (1999) showed that the breakdown of M-estimators with bounded score function in the general linear model depends via a quantity $\mathcal{M}(X, r)$ only on the design X and the variation exponent r of the objective function. In the case $r = 0$, e.g. for the Cauchy M-estimator, the quantity $\mathcal{M}(X, r)$ is the maximum number of design points at which the unknown parameter vector is not identifiable. We study the case $r = 1$, the case of the so-called ℓ_1-type M-estimators—including the ℓ_1-estimator itself, Huber's M-estimator, and others—where the quantity $\mathcal{M}(X) = \mathcal{M}(X, 1)$ has the potential of detecting the presence of leverage points in the design. We give a numerical algorithm for computing $\mathcal{M}(X)$ and calculate it for several examples, showing how $\mathcal{M}(X)$ can be used as a diagnostic tool for detecting leverage points and how the design influences the breakdown point of ℓ_1-type M-estimators.

KEYWORDS: Breakdown point; Optimum design; Leverage point; ℓ_1-estimator

1 Introduction

Consider the general linear model with N-dimensional univariate response vector $y = (y_1, y_2, \dots, y_N)^\top$ and the $N \times p$ matrix of design points $X = (x_1, x_2, \dots, x_N)^\top$:

$$y = X\beta + \varepsilon$$

—where, as usual, β is the vector of unknown parameters, and ε the vector of random errors. Let

$$m(\beta, E, X) = \frac{\sum_{n \in E} |x_n^\top \beta|}{\sum_{n=1}^{N} |x_n^\top \beta|},$$

and let

$$m(E, X) = \max_{\beta} m(\beta, E, X). \tag{1.1}$$

Finally, let

$$\mathcal{M}(X) = \min\{\text{card } E : m(E, X) \geq 1/2\}. \tag{1.2}$$

The relationship of $\mathcal{M}(X)$ to the breakdown point of the ℓ_1 and ℓ_1-type estimators was pointed out by He et al. (1990), and also by Ellis and Morgenthaler (1992), who added more insights and facts about the exactfit degree; finally it was firmly established by Mizera and Müller (1999) in the framework of their general theory for M-estimators in the linear regression model. In this theory, (1.2) is a special case of a certain general quantity, depending on the regular variation exponent r of the objective function of the M-estimator. The value $r = 1$ corresponds to the ℓ_1-estimator (and other ℓ_1-type estimators—Huber's estimator, for instance); the breakdown point is equal to $\mathcal{M}(X)/N$. If $r = 0$, then the breakdown point depends on the design only via the *identification index*

$$\mathcal{N}(X) = \max_{\beta \neq 0} \text{card}\{n : x_n^\top \beta = 0\}$$

and is equal to $\lfloor (N - \mathcal{N}(X) + 1)/2 \rfloor / N$. This value is maximal possible for equivariant estimators: the inequality

$$\mathcal{M}(X) \leq \left\lfloor \frac{N - \mathcal{N}(X) + 1}{2} \right\rfloor \tag{1.3}$$

holds. For more details, we refer the reader to Mizera and Müller (1999). It is important to note that the version of the breakdown theory under discussion keeps the design points fixed and admits outliers only in response. Although fractional values of r are possible, only integer values of r seem of practical interest; the values $r = 0$ and $r = 1$ correspond to robust M-estimators, $r = 2$ to the classical least squares estimator.

In this note, we further examine the case $r = 1$. We study some consequences of the breakdown theory in optimum design considerations and diagnostics of leverage/influential points. We illustrate, on several examples, how $\mathcal{M}(X)$ may serve for evaluation of designs and how $\mathcal{M}(X)$ and related quantities may be used as a regression diagnostic tool. To this end, we need a numerical algorithm for their computation.

2 Computation of $\mathcal{M}(X)$

For computing $\mathcal{M}(X)$ numerically, the following observations are useful. First, if $\mathcal{N}(X) < N$, then $\mathcal{M}(X) - 1$ is equal to the maximal cardinality of a set $E \subset \{1, 2, \ldots, N\}$ such that $m(E, X) < 1/2$. For $c \neq 0$, $m(c\beta, E, X) = m(\beta, E, X)$; thus, since we may restrict the domain of maximization to a compact set

$$A = \left\{ \beta : \sum_{n=1}^{N} |x_n^\top \beta| = 1 \right\},$$

the maximum is attained for every E. Let $\mathcal{X}$ be *the set* of all design points x_n (that is, multiple design points enter $\mathcal{X}$ only once). Let $\mathcal{D}$ be the set of all subsets D of $\mathcal{X}$ with the property that the linear space spanned by the elements of D is $(p-1)$-dimensional. Let $D^\perp$ be the orthogonal complement of this space, the set of all β such that $x_n^\top \beta = 0$ for all $x_n \in D$. Let

$$\mathcal{B} = \bigcup_{D \in \mathcal{D}} (A \cap D^\perp).$$

Let $\mathcal{S} = \{-1, 1\}^N$. For $s \in \mathcal{S}$ we define

$$A_s = \{\beta \in A\colon |x_n^\top \beta| = s_n x_n^\top \beta \text{ for } n = 1, 2, \ldots, N\}$$

—if $\beta \in A_s$, then $\operatorname{sign}(x_n^\top \beta) = s_n$ or 0.

Proposition 1 *For every $s \in \mathcal{S}$, the set A_s is compact and convex (possibly empty). All extremal points of A_s lie in $\mathcal{B}$; $m(\cdot, E, X)$ is linear on A_s for every E.*

Proof. The set A_s is a subset of A, hence it is bounded. The properties that determine A_s are preserved under a passage to limit; hence A_s is closed. Convex combinations also remain in A_s; since

$$\sum_{n=1}^N |x_n^\top \beta| = \sum_{n=1}^N s_n x_n^\top \beta,$$

all convex combinations of elements of A_s belong to A. Also linearity of $m(\cdot, E, X)$ on A_s follows in the analogous way.
Fix $s \in \mathcal{S}$. Let β be an extremal point of A_s and let C be the set of all x_n such that $x_n^\top \beta = 0$. Suppose that the dimension of the linear space spanned by C does not exceed $p-2$. Then the linear space spanned by $C \cup \{\sum_{n=1}^N s_n x_n\}$ has dimension at most $p-1$; consequently, there is a $\tilde\beta \neq 0$ such that $x_n^\top \tilde\beta = 0$ for all $x_n \in C$ and $\sum_{n=1}^N s_n x_n^\top \tilde\beta = 0$. For all $x_n \in \mathcal{X} \setminus C$, $|x_n^\top \beta| \geq \eta$ for some $\eta > 0$; we may suppose that $|x_n^\top \tilde\beta| \leq \eta/2$ for all $n = 1, 2, \ldots, N$, taking a suitable nonzero multiple of $\tilde\beta$ otherwise. As a consequence,

$$\operatorname{sign}(x_n^\top \beta + x_n^\top \tilde\beta) = \operatorname{sign}(x_n^\top \beta) = \operatorname{sign}(x_n^\top \beta - x_n^\top \tilde\beta) \tag{1.4}$$

for all $x_n \in \mathcal{X} \setminus C$, and also for all $x_n \in C$. Therefore,

$$\sum_{n=1}^N \left| x_n^\top (\beta + \tilde\beta) \right| = \sum_{n=1}^N s_n x_n^\top (\beta + \tilde\beta) = \sum_{n=1}^N s_n x_n^\top \beta + \sum_{n=1}^N s_n x_n^\top \tilde\beta = 1,$$

and analogously $\sum_{n=1}^N |x_n^\top (\beta - \tilde\beta)| = 1$. We obtain, in view of (1.4), that β is a convex combination of $\beta + \tilde\beta$ and $\beta - \tilde\beta$, both elements of A_s. This contradicts the extremality of β; hence, the space spanned by C has dimension at least $p-1$. It cannot have dimension p since then $x_n^\top \beta = 0$ for all $x_n \in C$ would imply $\beta = 0$, a contradiction with the fact that $\beta \in A$. Thus, the dimension is $p-1$. □

Proposition 2 *For every $E \subseteq \{1, 2, \ldots, N\}$, the maximum $m(E, X)$ of $m(\beta, E, X)$ over β is attained on some point of $\mathcal{B}$ satisfying*

$$\frac{1}{2}\,\mathrm{card}\ \mathcal{B} \leq \binom{\mathrm{card}\ \mathcal{X}}{p-1}. \tag{1.5}$$

Proof. Since $D^\perp$ is a one-dimensional linear subspace, the set $A \cap D^\perp$ contains, for any $D \in \mathcal{D}$, at most two elements; this proves (1.5). The remaining part of the statement follows from Proposition 1 and the fact that $A = \bigcup_{s \in \mathcal{S}} A_s$—since the maximum of a linear function over a compact convex set is attained in one of its extremal points. □

Proposition 2 says that for computing $\mathcal{M}(X)$, we have only to inspect $\beta \in \mathcal{B}$. Note that, in fact, we have to consider only one β from each $A \cap D^\perp$, the second one is $-\beta$, with the same $|x_n^\top \beta|$'s for all $n = 1, 2, \ldots, N$. For simplicity, we used $\{x_1, \ldots, x_N\}$ instead of $\mathcal{X}$ in our computer program—despite the fact that the speed of the algorithm would improve if $\mathcal{X}$ would be used for designs with many repetitions. The computer program written in `S` (the `R` implementation, version `1.0.0`) consists of a main routine called `M.X` and a recursive function `M.X.rek` that determines all subsets of $\{x_1, \ldots, x_N\}$ with $p-1$ elements. The main function `M.X` is a function of X and has the form

```
N <- number of rows of X
p <- rank of X
cardE.opt <- N
E.opt <- 1:N
beta.opt <- vector of p zeros
for (n in 1:N) {
    subset <- n
    result <- result of M.X.rek at X, subset, E.opt,
              cardE.opt, beta.opt, N, p }
```

The recursive function `M.X.rek` is a function of X, N, p, the subset E and the candidates `cardE.opt`, `E.opt` and `beta.opt` for $\mathcal{M}(X)$, the minimal cardinality of E such that $m(E, X) \geq 1/2$, and $\arg\max_\beta m(\beta, E.opt, X)$, respectively. Its form is

```
sall <- 1:n
if (length of subset < p-1)
    for (snew in sall[-subset])
      if (snew > max(subset)) {
         subset.new <- cbind(subset,new)
         result <- result of M.X.rek at X, subset.new, E.opt,
                   cardE.opt, beta.opt, N, p }
```

```
else{
    beta <- an element of the orthogonal complement of X[subset,]
    Xbeta <- abs(X %*% beta) # vector of |x_n^\top \beta|
    Xorder <- order(Xbeta)
    Xbeta.order <- Xbeta[Xorder] # Xbeta ordered
    n <- p-1
    while(n<N & 2*sum(Xbeta.order[1:n])-sum(Xbeta)<0.000001)
         {cardE <- N-n
         n <- n+1}
    if (cardE <- cardE.opt) {
       cardE.opt <- cardE
       E.opt <- Xorder[n:N]
       beta.opt <- beta }}
```

3 $\mathcal{M}(X)$ as a design criterion

A usual way to choose an optimal design is to maximize the information matrix $\mathcal{I}(X) = X^\top X/N$ or a function of it. This ensures the smallest variance of the least squares estimator—but also of the ℓ_1-estimator, if no outliers are present. As soon as outliers can appear, an additional criterion on the design should be adopted, ensuring maximal possible robustness of the given estimator. For ℓ_1 estimators this calls for $\mathcal{I}(X)$ as small as possible and $\mathcal{M}(X)$ as large as possible. The following examples indicate that the D-optimal designs, the designs minimizing $\det \mathcal{I}(X)$, also maximize $\mathcal{M}(X)$ (however, we have no rigorous proof of this assertion so far).

Example 1. Simple regression
In Mizera and Müller (1999), an explicit formula for $\mathcal{M}(X)$ for simple regression was given, when $x_n = (1, t_n)^\top$. This implies that any equispaced design—and thus the D-optimal design—maximizes $\mathcal{M}(X)$.

Example 2. Quadratic regression
For the quadratic regression with $x_n = (1, t_n, t_n^2)^\top$, we calculated $\mathcal{M}(X)$ for four different designs. For a design with 40 repetitions at -1, 0, 1 and for that with 30 repetitions at -1, -0.3, 0.3, 1, we obtained $\mathcal{M}(X) = 20$. For equispaced design with the t_n's from -1 up to 1, the result was smaller: $\mathcal{M}(X) = 18$. Putting more repetitions at $t_n = 0$ reduces $\mathcal{M}(X)$ as well: for instance, we obtained $\mathcal{M}(X) = 15$ for a design with 30 repetitions at -1 and 1 and 60 repetitions at 0. Apparently, 20 is the maximal value for $\mathcal{M}(X)$—and it is reached by the D-optimal design.
Note that $20/120 = 1/6$ is the maximal possible breakdown point for any equivariant estimator at the D-optimal design; this follows from (1.3) since $\mathcal{N}(X) = 80$. Thus, at the D-optimal design, the highest breakdown point is attained by the ℓ_1-estimator—this is not true for other designs, however. For example, for the equispaced design (design points from -1 up to 1), the maximal breakdown point is $59/120$ (according to (1.3)); this is much larger than $18/120$ provided by the ℓ_1-estimator.

Example 3. Two-factor model
For a model with two factors each with three levels 1, 2, 3 and without interactions we compared the following two designs (each level combination with one observation):

Factor 1	1	1	1	2	2	2	3	3	3
Factor 2	1	1	2	2	2	3	1	3	3
Factor 1	1	1	1	2	2	2	3	3	3
Factor 2	1	2	3	1	2	3	1	2	3

The result was $\mathcal{M}(X) = 1$ for the first design and $\mathcal{M}(X) = 2$ for the second design. The D-optimal design again gives the maximal breakdown point. Since $\mathcal{N}(X) = 1$ for the first design and $\mathcal{N}(X) = 2$ for the second design, the ℓ_1-estimator has at both designs the maximal breakdown point—according to (1.3)

4 Applications for leverage/influence diagnostics

Ellis and Morgenthaler (1992) introduced $m(E, X)$ as an ℓ_1 analog of the quantity

$$\max_{\beta} \frac{\sum_{n\in E}(x_n^\top \beta)^2}{\sum_{n=1}^{N}(x_n^\top \beta)^2}, \tag{1.6}$$

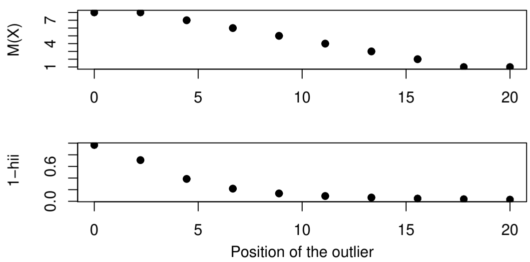
proposed as a diagnostic for the classical least squares method quantifying the influence/leverage of the group of observations represented by E. They interpreted (1.6) as the squared cosine of the minimal angle between two directions, one belonging to the linear space generated by the design points, the other to the linear space generated by the unit vectors e_i, $i \in E$. Particularly, if $E = \{i\}$, then (1.6) is equal to h_{ii}, the i-th diagonal element of the hat (projection) matrix $X(X^\top X)^{-1}X^\top$, whose diagnostic value is well-known; see Atkinson (1985), or Rousseeuw and Leroy (1987).
As already mentioned, in the theory of Mizera and Müller (1999) both (1.1) and (1.6) are special cases of the same general quantity, corresponding to values $r = 1$ and $r = 2$, respectively, of the regular variation exponent of the objective function of the M-estimator. In view of this breakdown point theory, $m(E, X)$ can be interpreted as a "breaking power", a capability to break the ℓ_1-estimator down, if the response is altered at the points with indices in E. This suggests the interpretation of $m(E, X)$ as the measure of leverage of the observations corresponding to the set E.
Many of regression diagnostics appearing in the literature are designed to evaluate the potential of observations to influence the least squares fit. Their rationale is often based on the influence function of the least squares estimator, or some of its finite-sample substitutes. As a result, they usually depend on h_{ii} as well as on the residual $y_i - x_i^\top \widehat{\beta}$. The influence function of the ℓ_1-estimator—unlike that of the least squares estimator—involves only the sign of the residual, suggesting that potential influence measures

designed for the ℓ_1 estimator—like (1.1)—may depend only on the design matrix.

To evaluate the leverage of individual points, it is natural to look at values of $m(\{i\}, X)$ (we use the shorthand $m(i, X)$). As with other measures of influence, it is important to know what values are "large". To this end, $\mathcal{M}(X)$ provides a valid comparison, evaluating how "stable" is the design: any E with $m(E, X) \geq 1/2$ is a "strategic set of points", capable of breaking down the estimator completely. Moreover, the set $E.opt$ on which the minimal cardinality appearing in $\mathcal{M}(X)$ is attained, is often the set of "principal outliers"; for instance, if there is one extremely outlying covariate point, then its index is contained in $E.opt$.

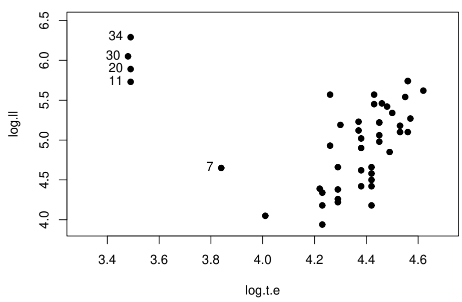
Example 4. Simple regression with one outlier

We considered again the simple regression, like in Example 1. The 30 points t_i were generated from the uniform distribution on $[-1, 1]$. Then we added an outlying design point $(1, T)$, where T took 10 equispaced values from 0 to 20. We plotted the dependence of $\mathcal{M}(X)$ on T; for comparison, we include also the analogous plot for the diagonal element of the hat matrix corresponding to the outlying design point.



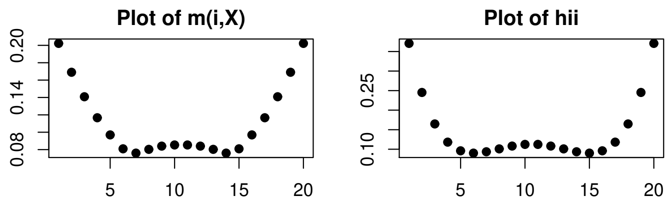
Example 5. The Hertzsprung-Rusell data

These data were introduced and analyzed by Rousseeuw and Leroy (1987). Using again simple regression for the data, $E.opt$ coincides with the 5 marked points, four of them corresponding to so-called "giants".



Example 6. Quadratic regression again

In certain situations, the influence may be distributed more uniformly: for instance, in polynomial regression with equispaced design, the points near the boundary are more influential than those in the middle and the influence is symmetric. However, *E.opt* returned by `M.X` may reflect just one side of the data points. To appreciate the full picture, we recommend to look at the individual influence of the points expressed by $m(i, X)$, and then to try various possible competitors of *E.opt*. We plotted $m(i, X)$ for the equispaced design in quadratic regression like in Example 2 (20 points from -1 to 1); for comparison, we plot again also the corresponding diagonal elements of the hat matrix. Apparently, points with indices 1, 2, 3 possess



the same leverage as those from *E.opt*: 18, 19, 20. Looking at $m(E, X)$ for both sets, we realize that they are indeed the same.

Acknowledgments: We gratefully acknowledge the Deutsche Forschungsgemeinschaft 436 SLK 17/1/00 support for the first author's visit of the second author in April, 2000.

References

Atkinson, A. C. (1985). *Plots, Transformations, and Regression.* Clarendon Press, Oxford.

Ellis, S. P. and Morgenthaler, S. (1992). Leverage and breakdown in L_1 regression. *Journal of the American Statistical Association*, **87**, 143-148.

He, X., Jurečková, J., Koenker, R. and Portnoy, S. (1990). Tail behavior of regression estimators and their breakdown points. *Econometrica* **58**, 1195–1214.

Mizera, I. and Müller, Ch. H. (1999). Breakdown properties and variation exponents of robust M-estimators in linear models. *Annals of Statistics* **27**, 1164–1177.

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