Optimal breakdown point maximizing designs

Christine H. Müller
Free University of Berlin, Department of Mathematics, Arnimallee 2-6,
D-14195 Berlin, Germany, e-mail: mueller@math.fu-berlin.de

Abstract. In linear models the breakdown point of an estimator depends strongly on the underlying design. This holds in particular for high breakdown point estimators as the least median of squares estimator or least trimmed squares estimators. It could be shown that the breakdown point is maximized if the number of regressors which lie in a subspace is minimized. This means in particular that the number of repetitions of experimental conditions should be minimized. Usually this leads to designs which are very different from the classically optimal designs. But in some situations breakdown point maximizing designs can be found which are also optimal in the classical sense. In this paper two examples are given where breakdown point maximizing designs are also classically optimal.

AMS Mathematical Subject Classification (1991): 62K05, 62G35.

Key words: Linear model, breakdown point, trimmed L_p-estimator, optimal design.

1 Introduction

We regard a general linear model

$$y = X_d\beta + z$$

where $y = (y_1, ..., y_N)' \in I\!R^N$ is the observation vector, $X_d = (x(t_1), ..., x(t_N))' \in I\!R^{N\times r}$ the known design matrix, $x : \mathcal{T} \to I\!R^r$ the known regression function, $d = (t_1, ..., t_N)' \in \mathcal{T}^N$ the design, $\beta \in I\!R^r$ the unknown parameter vector and $z = (z_1, ..., z_N)' \in I\!R^N$ the error vector. We assume that the experimenter can choose freely the experimental conditions t_n, $n = 1, ..., N$, from the experimental region $\mathcal{T}$.

If we have an estimator $\widehat{\beta} : I\!R^N \times \mathcal{T}^N \to I\!R^r$ for the unknown parameter β then we can regard as a robustness measure its breakdown point which is the minimal fraction of outliers which can provide an arbitrarily high bias of the estimator. Originally the breakdown point of an estimator $\widehat{\beta}$ at a finite sample (y, d) was defined by Donoho and Huber (1983) by allowing also outliers in the experimental conditions. In the literature mainly this definition is used by assuming that the experimental conditions are given per random (see for example Rousseeuw and Leroy (1987), Coakley (1991), Mili and Coakley (1993)). But for designed experiments the experimental conditions are given by the experimenter so that outlying experimental conditions can not appear. Therefore here the breakdown point at a given sample (y, d) is defined as in Ellis and Morgenthaler (1992) and Müller (1993) by

$$(1.1)\ \epsilon^*(\widehat{\beta}, y, d) := \min\{\frac{M}{N};\ M \in \mathbb{N} \text{ and } \sup_{\overline{y} \in \mathcal{Y}_M(y)} |\widehat{\beta}(\overline{y}, d) - \widehat{\beta}(y, d)| = \infty\ \},$$

where $\mathcal{Y}_M(y)$ is the set of all contaminated observation vectors $\overline{y} \in \mathbb{R}^N$ for which $m(1) < ... < m(N-M)$ exists so that $\overline{y}_{m(i)} = y_{m(i)}$ for all $i \in \{1, ..., N-M\}$. By regarding the minimum possible breakdown point for an estimator $\widehat{\beta}$ at a given design d, i.e.

$$\epsilon^*(\widehat{\beta}, d) := \min\{\epsilon^*(\widehat{\beta}, y, d);\ y \in \mathbb{R}^N\},$$

we get a breakdown point which is independent of the observation vector y.

The breakdown point of the classically optimal estimator for β, the least squares estimator (LS estimator) $\widehat{\beta}_{LS}$, is in the original definition of Donoho and Huber (1983) as well as in the definition of (1.1) equal to $\frac{1}{N}$. To improve the breakdown point we can regard the h-trimmed L_p-estimators which were regarded in Müller (1993) and which generalize the least median of squares estimator and the least trimmed squares estimators (see Rousseeuw and Leroy (1987), Coakley (1991), Coakley and Mili (1993), Mili and Coakley (1993)). A h-trimmed L_p-estimator is given by

$$\widehat{\beta}_h(y, d) = \arg \min_{\beta \in \mathbb{R}^r} \sum_{n=1}^{h} w_n\ |y_{(n)} - x(t_{(n)})'\beta|^p$$

where $w_1, ..., w_N > 0$, $p > 0$ and $|y_{(1)} - x(t_{(1)})'\beta|, ..., |y_{(N)} - x(t_{(N)})'\beta|$ are the ordered residuals for a given β, i.e. we have $|y_{(1)} - x(t_{(1)})'\beta| \leq ... \leq |y_{(N)} - x(t_{(N)})'\beta|$.

As Müller (1993) has shown for h not too large the breakdown point of a h-trimmed L_p-estimator depends only on the trimming value h, on the design d and the sample size N, and it does not depend on p and the weights w_n. For fixed h and N the breakdown point depends on the design d only via $N(d)$ and increases if $N(d)$ decreases. Thereby $N(d)$ is the maximal number of regressors $x(t_n)$ which are lying in a $(r-1)$-dimensional subspace of $\mathbb{R}^r$, i.e.

$$N(d) := \max\{N_V;\ V \text{ is a } (r-1)-\text{dimensional subspace of } \mathbb{R}^r\},$$

where $N_V := \sum_{n=1}^N 1_V(x(t_n))$. In particular for $[\frac{N+N(d)+1}{2}] \leq h \leq [\frac{N+N(d)+2}{2}]$ the h-trimmed L_p-estimator maximizes the breakdown point within all regression equivariant estimators with a maximum breakdown point of

$$\epsilon^*(d) := \epsilon^*(\widehat{\beta}_h, d) = \frac{1}{N}\left[\frac{N - N(d) + 1}{2}\right].$$

Hence in Müller (1993) a breakdown point maximizing design within a class Δ of designs is defined as

$$d^* = \arg\min\{N(d);\ d \in \Delta\}.$$

Because we always have $N(d) \geq r - 1$ a design d^* with $N(d^*) = r - 1$, i.e. a design where the regressors $x(t_1), ..., x(t_N)$ are in general position (see Rousseeuw and Leroy (1987)), is breakdown point maximizing.

For example for linear regression $N(d)$ is the maximal number of repetitions of a experimental condition t_n so that a design with no repeated experimental conditions t_n satisfies $N(d) = 1$ and therefore is breakdown point maximizing. But these breakdown point maximizing designs are very different from the classically optimal design which has $\frac{N}{2}$ repetitions of $t = -1$ and $\frac{N}{2}$ repetitions of $t = 1$ if $\mathcal{T} = [-1, 1]$ is the experimental region. Hence for linear regression we have a contradiction between maximum breakdown point and classic optimality.

But there are other problems for which it is possible to derive designs which are breakdown point maximizing as well as classically optimal. In this paper this is demonstrated for two problems, namely for the multiple regression on balls in Section 2 and for a weighing problem in Section 3.

2 Multiple regression on balls

In a r-dimensional multiple regression problem without constant term the observations are given by

$$y_n = x(t_n)'\beta + z_n = \sum_{i=1}^r \beta^i t_n^i + z_n$$

for $n = 1, ..., N$, i.e. $\beta = (\beta^1, ..., \beta^r)' \in I\!R^r$ and $x(t) = t = (t^1, ..., t^r)'$. Here we assume that the experimental region is a ball, i.e.

$$t_n \in \mathcal{T} = \{t \in I\!R^r;\ \textstyle\sum_{i=1}^{r} (t^i)^2 \leq R^2\}.$$

Without loss of generality we can assume that the radius of the ball is $R = 1$. Then the equivalence theorems (see for example Pukelsheim (1993)) provide that a design $d \in \mathcal{T}^N$ satisfying $X_d' X_d = \frac{N}{r} E_r$, where E_r denotes the $r \times r$-identity matrix, is A-, D- and E-optimal in $\Delta = \mathcal{T}^N$, i.e. classically optimal. In particular if the sample size N satisfies $N = K\,r$, where K is an integer, then the design given by $d = (E_r|...|E_r)'$, where the identity matrix E_r is repeated K times, is classically optimal because $X_d' X_d = \frac{N}{r} E_r$. This design has minimum support but for $r > 1$ it provides a low maximum breakdown point of

$$\epsilon^*(d) = \frac{1}{N} \left[\frac{N - (r-1)K + 1}{2} \right] = \frac{1}{N} \left[\frac{N - (r-1)\frac{N}{r} + 1}{2} \right] \stackrel{N \to \infty}{\longrightarrow} \frac{1}{2r}$$

because $N(d) = (r-1)K = N\frac{r-1}{r}$. But instead of repeating the identity matrix K times we can use K different orthonormal matrices $O_1, ..., O_K \in I\!R^{r \times r}$ with $O_k\,O_k' = E_r$ for $k = 1, ..., K$ so that the r unit vectors are rotated to different positions by keeping the angles between them. It is always possible to find K positions so that the design d^* with $d^* = (O_1|...|O_K)' = X_d$ satisfies $N(d^*) = r - 1$ so that d^* is breakdown point maximizing in $\Delta = \mathcal{T}^N$. In particular for $r = 2$ it is always possible to find $N = 2K$ different experimental conditions t_n with $|t_n| = 1$ so that every t_n is othogonal to one another experimental condition t_m and no t_n, t_m with $n \neq m$ are lying on a line through the origin. Hence we have the following lemma.

Lemma 2.1 *If $d^* = (O_1|...|O_K)'$ with $O_k\,O_k' = E_r$ for $k = 1, ..., K$ and $N(d^*) = r - 1$ then d^* is classically optimal in $\mathcal{T}^N$ as well as breakdown point maximizing in $\mathcal{T}^N$ with maximum breakdown point of*

$$\epsilon^*(d^*) = \frac{1}{N} \left[\frac{N - r + 2}{2} \right] \stackrel{N \to \infty}{\longrightarrow} \frac{1}{2}.$$

A design $d^* = (O_1|...|O_K)'$ with $O_k\,O_k' = E_r$ for $k = 1, ..., K$ and $N(d^*) = r - 1$ always exists.

3 Weighing designs for four objects

Here we regard the problem of weighing four objects with a chemical balance. The weights $\beta^1, ..., \beta^4$ of the four objects are unknown and we can use N weighing realizations for determining the weights of the objects. For each weighing realization for each object we have to decide on which side of the balance the object should be put. Hence for the n'th weighing realization we have to choose

$$t_n \in \mathcal{T} = \{-1, 1\}^4$$

where $t_n^i = -1$ and $t_n^i = 1$ means that in the n'th realization the i'th object is put on the lefthand side or righthand side of the balance, respectively. Then the result of the n'th weighing realization will be

$$y_n = x(t_n)'\beta + z_n = \beta^1 t_n^1 + ... + \beta^4 t_n^4 + z_n$$

where $\beta = (\beta^1, \beta^2, \beta^3, \beta^4)' \in I\!R^4$, $x(t) = t = (t^1, t^2, t^3, t^4)' \in \mathcal{T}$ and z_n is an error which may be also an outlier if some extraordinary shake happens.

If a design d satisfies $X_d'X_d = N\,E_4$, where E_4 denotes the 4×4 identity matrix, then as for multiple regression on balls d is A-, D- and E-optimal in $\Delta = \mathcal{T}^N$, i.e. classically optimal (see Section 2). In particular if the sample size is $N = 4K$, where K is an integer, then a classically optimal design is $d = (X_0|...|X_0)'$, where the matrix $X_0 = (1_4|X)' \in I\!R^{4\times 4}$ is repeated K-times, $1_4 = (1, 1, 1, 1)'$ and

$$X = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \end{pmatrix}.$$

The design d has minimum support but it provides a maximum breakdown point of

$$\epsilon^*(d) = \frac{1}{N}\left[\frac{N - 3K + 1}{2}\right] \xrightarrow{N\to\infty} \frac{1}{8}$$

because $N(d) = 3K = \frac{3N}{4}$. In the case of $N = 8\,L$, where L is an integer, the breakdown is improved by a design $d^* = (X_1|...|X_1)'$, where the matrix $X_1 = ((1_4|X)'|(1_4|-X)') \in I\!R^{4\times 8}$ is repeated L times. In $\Delta = \mathcal{T}^N$ this design is breakdown point maximizing and classically optimal which shows the following lemma.

Lemma 3.1 *If* $N = 8L$ *and* $d^* = (X_1|...|X_1)'$, *where* L *is an integer and* $X_1 = ((1_4|X)'|(1_4|-X)')$ *is repeated* L *times, then* d^* *is classically optimal in* $\mathcal{T}^N$ *as well as breakdown point maximizing in* $\mathcal{T}^N$ *with maximum breakdown point of*

$$\epsilon^*(d^*) = \frac{1}{N}\left[\frac{N-4L+1}{2}\right] \stackrel{N\to\infty}{\longrightarrow} \frac{1}{4}.$$

Proof. Let be $\mathcal{A}$ the set of columns of $(1_4|X)'$ and $\mathcal{B}$ the set of columns of $(1_4|-X)'$. Then we have for every $a \in \mathcal{A}$ and $b \in \mathcal{B}$ the property $(a'b)^2 = 4$. Hence for every $a \in \mathcal{A}$ and $b_1, b_2, b_3 \in \mathcal{B}$ with $b_1 \neq b_2 \neq b_3 \neq b_1$ we have $\det((a, b_1, b_2, b_3)'(a, b_1, b_2, b_3)) \neq 0$ so that a, b_1, b_2, b_3 are linearly independent. Similarly every $b \in \mathcal{B}$ and $a_1, a_2, a_3 \in \mathcal{B}$ with $a_1 \neq a_2 \neq a_3 \neq a_1$ are linearly independent. Hence it is impossible that more than 4 elements of $\mathcal{A} \cup \mathcal{B}$ are lying in a subspace of $I\!R^4$, i.e. $N(d^*) \leq 4L$. On the other hand for every $a_1, a_2 \in \mathcal{A}$ there exists $b_1, b_2 \in \mathcal{B}$ with $a_1 + b_1 = (2,0,0,0)' = a_2 + b_2$ so that a_1, a_2, b_1, b_2 are linearly dependent. Hence we have $N(d^*) = 4L = \frac{N}{2}$. Moreover if additionally $N = 6J$ for some integer J then for every $d^6 = (X_2|...|X_2)'$, where $X_2 = (c_1, ..., c_6)$ is repeated J times and $\{c_1, ..., c_6\} \subset \mathcal{A} \cup \mathcal{B}$, we have $N(d^6) = 4J = \frac{2N}{3} > \frac{N}{2} = N(d^*)$. If additionally $N = 5J$ for some integer J then for every $d^5 = (X_3|...|X_3)'$, where $X_3 = (c_1, ..., c_5)$ is repeated J times and $\{c_1, ..., c_5\} \subset \mathcal{A} \cup \mathcal{B}$, we have $N(d^5) \geq 3J = \frac{3N}{5} > \frac{N}{2} = N(d^*)$. The same holds for designs d^4 and d^7 with 4 and 7 different experimental conditions. Also designs which are a mixture of designs d^4, d^5, d^6, d^7 and d^* always satisfy $N(d) \geq N(d^*)$ so that d^* is breakdown point maximizing. □

Note that in many weighing problems it is also allowed that in some weighing realizations some of the objects do not participate so that for four objects the experimental region would be $\mathcal{T}_0 = \{-1, 0, 1\}^4$. But in $\mathcal{T}_0^N$ there is no breakdown point maximizing design which is also classically optimal.

Also note that the design used in Lemma 3.1 can be constructed by product designs (see for example Schwabe (1992)) so that for more than four objects breakdown point maximizing designs will be given by product designs. Often the product designs are criticized that they have too much support points. But for the breakdown point this is an advantage so that the breakdown point provides a justification of product designs.

References

[1] Coakley, C.W. (1991). Breakdown points under simple regression with replication. *Technical Report No. 91-21*, Department of Statistics, Virginia Polytechnic Institute and State University.

[2] Coakley, C.W. and Mili, L. (1993). Exact fit points under simple regression with replication. *Statist. Probab. Lett.* **17**, 265-271.

[3] Donoho, D.L. and Huber, P.J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann* (P.J. Bickel, K.A. Doksum and J.L. Hodges, Jr.,eds.) 157-184. Wadsworth, Belmont, Calif.

[4] Ellis, S.P. and Morgenthaler, S. (1992). Leverage and breakdown in L_1 regression. *J. Amer. Statist. Assoc.* **87**, 143-148.

[5] Mili, L. and Coakley, C.W. (1993). Robust estimation in structured linear regression. *Technical Report No. 93-13*, Department of Statistics, Virginia Polytechnic Institute and State University.

[6] Müller, Ch.H. (1993). Breakdown points for designed experiments. *Preprint No. A-93-29*, Freie Universität Berlin, Fachbereich Mathematik und Informatik. *To appear in J. Statist. Plann. Inference.*

[7] Pukelsheim, F. (1993). *Optimal Design of Experiments.* John Wiley, New York.

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

[9] Schwabe, R. (1992). Optimum designs for multi-factor models. *Habilitation thesis.* Freie Universität Berlin.