# High breakdown point designs

**Christine H. Müller**
Freie Universität Berlin, Fachbereich Mathematik und Informatik, WE 1,
Arnimallee 2-6, D-14195 Berlin, Germany, e-mail: mueller@math.fu-berlin.de

**Abstract.** A linear model is assumed where the experimental conditions are given by the experimenter so that they are without gross errors. For this situation we regard $h$-trimmed $L_p$-estimators which generalize the least median of squares estimator and the least trimmed squares estimators. The breakdown point of these estimators depends strongly on the underlying design and breakdown point maximizing estimators can be derived within these estimators. Also breakdown point maximizing designs can be derived. But these designs are often very different from the classically optimal designs which ensure efficiency also for robust procedures. Therefore we regard compromises providing high breakdown point and high efficient designs.

## 1 Breakdown points for designed experiments

We regard a general linear model

$$y = X\beta + z$$

where $y = (y_1, ..., y_N)' \in I\!R^N$ is the observation vector, $X = (x_1, ..., x_N)' \in \mathcal{X}^N \subset I\!R^{N\times r}$ the known design matrix, briefly also called 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 $x_n,\ n = 1, ..., N$, from the experimental region $\mathcal{X}$.

If we have an estimator $\widehat{\beta} : I\!R^N \times \mathcal{X}^N \to I\!R^r$ for the unknown parameter $\beta$ 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 $(y, X)$ was defined by Donoho and Huber (1983) by allowing also outliers in the experimental conditions as

$$(1.1)\quad \epsilon(\widehat{\beta}, y, X) := \min\{\frac{M}{N};\ M \in I\!N \text{ and } \sup_{(\overline{y},\overline{X})\in\mathcal{Y}_M(y,X)} |\widehat{\beta}(\overline{y},\overline{X}) - \widehat{\beta}(y,X)| = \infty\ \},$$

where $\mathcal{Y}_M(y, X)$ is the set of all contaminated samples $(\overline{y}, \overline{X}) \in I\!R^N \times \mathcal{X}^N$ for which $m(1) < ... < m(N - M)$ exist so that $(\overline{y}_{m(i)}, \overline{x}_{m(i)}) = (y_{m(i)}, x_{m(i)})$ for all $i \in \{1, ..., N - M\}$.

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 $(y, X)$ is defined as in Ellis and Morgenthaler (1992) and Müller (1993) by

$$(1.2)\quad \epsilon^*(\widehat{\beta}, y, X) := \min\{\frac{M}{N};\ M \in I\!N \text{ and } \sup_{\overline{y}\in\mathcal{Y}_M(y)} |\widehat{\beta}(\overline{y}, X) - \widehat{\beta}(y, X)| = \infty\ \},$$

where $\mathcal{Y}_M(y)$ is the set of all contaminated observation vectors $\overline{y} \in I\!R^N$ for which $m(1) < ... < m(N-M)$ exist 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 $X$, i.e.

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

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

# 2 Breakdown point maximizing estimators

The breakdown point of the classically optimal estimator for $\beta$, 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 (1.2) equal to $\frac{1}{N}$. To improve the breakdown point alternative estimators were proposed. One is the least median of squares estimator (LMS estimator) $\widehat{\beta}_{LMS}$ which minimizes the median of squared residuals instead of the sum of squared residuals, i.e. which is given by

$$\widehat{\beta}_{LMS}(y, X) = \arg \min_{\beta\in I\!R^r} \text{med}_{n=1,\ldots,N}(y_n - x_n'\beta)^2$$

(see Rousseeuw (1984), Rousseeuw and Leroy (1987)).

But if we have a linear regression model $y_n = \beta_0 + \beta_1 t_n + z_n$ with $t_n \in [-1, 1]$ and $x_n = (1, t_n)'$ and use the classically optimal design (A-, D-, E-optimal) design $X^0$ with $\frac{N}{2}$ observations at $x = (1, -1)'$ and $\frac{N}{2}$ observations at $x = (1, 1)'$ (see also Section 3) then the breakdown point of the LMS estimator is $\epsilon^*(\widehat{\beta}_{LMS}, X^0) = \frac{1}{N}$ because the exact fit point satisfies $\epsilon^*(\widehat{\beta}_{LMS}, X^0\beta, X^0) = \frac{1}{N}$. I.e. at $X^0$ the breakdown point of the LMS estimator coincides with the breakdown point of the LS estimator.

To improve the breakdown point at the classically optimal design $X^0$ we can regard a generalization of the LMS estimator, namely the $h$-trimmed L$_p$-estimators which were regarded in Müller (1993) and which also generalize the least trimmed squares estimators (see also Rousseeuw and Leroy (1987), Coakley (1991), Coakley and Mili (1992), Mili and Coakley (1993)). A $h$-trimmed L$_p$-estimator is given by

$$\widehat{\beta}_h(y, X) = \arg \min_{\beta\in I\!R^r} \sum\nolimits_{n=1}^{h} w_n\ |y_{(n)} - x_{(n)}'\beta|^p$$

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

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 $X$ 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 $X$ only via $N(X)$ and increases when $N(X)$ decreases. Thereby $N(X)$ is the maximal number of experimental conditions $x_n$ which are lying in a $(r-1)$-dimensional subspace of $\mathbb{R}^r$, i.e.

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

where $N_V := \sum_{n=1}^N 1_V(x_n)$ (compare also with the quantity $\Delta(Q)$ used in Davies (1993), (1994)). In particular for $N(X) < h \leq \frac{N+N(X)+1}{2}$ the breakdown point of the $h$-trimmed $\mathrm{L}_p$-estimator is

$$\epsilon^*(\widehat{\beta}_h, X) = \frac{1}{N}(h - N(X)) \tag{2.3}$$

and for $[\frac{N+N(X)+1}{2}] \leq h \leq [\frac{N+N(X)+2}{2}]$ we have

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

where $[z] := \max\{n \in \mathbb{N};\ n \leq z\}$ (see Müller (1993)).

Every regession equivariant estimator, i.e. every estimator $\widehat{\beta}$ satisfying $\widehat{\beta}(y + X\theta, X) = \widehat{\beta}(y, X) + \theta$ for all $y \in \mathbb{R}^N$ and $\theta \in \mathbb{R}^r$ as the $h$-trimmed $\mathrm{L}_p$-estimators, has a breakdown point with

$$\epsilon^*(\widehat{\beta}, y, X) \leq \frac{1}{N}\left[\frac{N-N(X)+1}{2}\right]$$

for all $y \in \mathbb{R}^N$ (see Müller (1993)). Hence the maximum breakdown point at the design $X$ is

$$\epsilon^*(X) := \max\{\epsilon^*(\widehat{\beta}, X);\ \widehat{\beta} \text{ regression equivariant}\} = \frac{1}{N}\left[\frac{N-N(X)+1}{2}\right]$$

and is attained by $h$-trimmed $\mathrm{L}_p$-estimators with $[\frac{N+N(X)+1}{2}] \leq h \leq [\frac{N+N(X)+2}{2}]$ so that these $h$-trimmed $\mathrm{L}_p$-estimators are breakdown point maximizing estimators.

# 3 Breakdown point maximizing designs

A breakdown maximizing design can be defined as a design $X^*$ which maximize the maximum breakdown point $\epsilon^*(X)$ within a given class $\Delta$ of designs $X$, i.e.

$$X^* = \arg\max\{\epsilon^*(X);\ X \in \Delta\}.$$

Because the maximum breakdown point is a decreasing function of $N(X)$ we also can define a breakdown point maximizing design as a design $X^*$ which minimizes $N(X)$, i.e.

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

(see Müller (1993)). Because of (2.3) the breakdown point maximizing design maximizes not only the breakdown point of the breakdown point maximizing estimator but also the breakdown point of every $h$-trimmed $L_p$-estimator with $N(X^*) < h \leq \frac{N+N(X^*)+1}{2}$.

Because we always have $N(X) \geq r-1$ a design $X^*$ with $N(X^*) = r-1$, i.e. a design where the experimental conditions $x_1, ..., x_N$ are in general position (see Rousseeuw and Leroy (1987)), is breakdown point maximizing.

For example for linear regression $N(X)$ is the maximal number of repetitions of a experimental condition $x_n$ so that a design $X^*$ with no repeated experimental conditions $x_n$ satisfies $N(X^*) = 1$ and therefore is breakdown point maximizing. But these breakdown point maximizing designs are very different from the classically optimal design which minimizes some functional $\phi$ of the covariance matrix $(X'X)^{-1}$ of the least squares estimator as the trace (A-optimality), the determinant (D-optimality) or the maximum eigenvalue (E-optimality) of $(X'X)^{-1}$. The equivalence theorems for A-, D- and E-optimality (see Pukelsheim (1993) p. 221, 212 and 181-182, respectively) provide that in $\Delta = \mathcal{X}^N$ the A-, D- and E-optimal design has $\frac{N}{2}$ repetitions of $x = (1,-1)'$ and $\frac{N}{2}$ repetitions of $x = (1,1)'$ if $\mathcal{X} = \{1\} \times [-1,1]$ is the experimental region. Hence for linear regression we have a contradiction between maximum breakdown point and classic optimality.

This contradiction is in particular grave because classically optimal designs provide also high efficiency of some robust procedures. In particular rank-based estimators as regarded in Hössjer (1994) and Huber M-estimators have an asymptotic covariance matrix which coincides with the covariance matrix of the least squares estimator except for some constant which does not depend on the design (see Hössjer (1994), Bickel (1975), Yohai and Maronna (1979)). Also at A-optimal designs the trace of the asymptotic covariance matrix of the optimal robust Hampel-Krasker estimator is minimized (see Müller (1994)). Moreover the classically optimal design for linear regression with $\mathcal{X} = \{1\} \times [-1,1]$ maximizes the breakdown point of the $L_1$-estimator (see Ellis and Morgenthaler (1992)).

# 4 Combining high breakdown point and high efficiency

Estimators which combine high breakdown point and high efficiency can be easily constructed by one-step M-estimators

$$\widehat{\beta}(y,X) = \widehat{\beta}^0(y,X) + \frac{1}{N}\sum\nolimits_{n=1}^{N} \psi(y_n - x_n'\widehat{\beta}^0(y,X), x_n),$$

where the initial estimator $\widehat{\beta}^0$ is a high breakdown point estimator and the score function $\psi$ provides high asymptotic efficiency. If the initial estimator is asymptotically normally distributed with convergence rate of $N^{-1/2}$ as many $h$-trimmed $L_p$-estimators (see Hössjer (1994)) then the finite sample breakdown point of $\widehat{\beta}$ coincides with that of the initial estimator $\widehat{\beta}^0$ and the asymptotic covariance matrix is only determined by the score function $\psi$ (see Jurečková and Portnoy (1987)).

To find a design which combines high breakdown point and high efficiency is more difficult. There are situations where even designs exist which are classically optimal as well as breakdown point maximizing (see Müller (1994)) but these situations are rare. More typical is the situation of linear regression where classically optimal designs are very different from the breakdown point maximizing designs. In all these situations constrained problems will be a good compromise to combine high breakdown point and high asymptotic efficiency.

One constrained problem is to minimize a functional $\phi$ of the covariance matrix $(X'X)^{-1}$ under the side condition that the breakdown point is not too bad, i.e.

$$X^{**} = \arg\min\{\phi((X'X)^{-1});\ \epsilon^*(X) \geq B,\ X \in \Delta\} = \arg\min\{\phi((X'X)^{-1});\ N(X) \leq \overline{B},\ X \in \Delta\}.$$

For the particular case that $N(X)$ is the maximal number of repetitions as for linear regression this problem was already treated by Wynn (1982), Wierich (1985) and Fedorov (1989) by constructing optimal designs with bounded number of repetitions.

Another constrained problem is to maximize the breakdown point under the side condition that the functional $\phi$ of the covariance matrix $(X'X)^{-1}$ is not too large, i.e.

$$X^{**} = \arg\max\{\epsilon^*(X);\ \phi((X'X)^{-1}) \leq B,\ X \in \Delta\} = \arg\min\{N(X);\ \phi((X'X)^{-1}) \leq B,\ X \in \Delta\}.$$

In particular we can set

$$B = \min\{\phi((X'X)^{-1});\ X \in \Delta\}$$

so that we restrict ourselves to $\phi$-optimal designs. This of course makes only sense if we have several $\phi$-optimal designs. For example for linear regression this makes no sense because the A-, D- and E-optimal design is unique. But there are other problems for which we have several classically optimal designs as the following examples show.

# 5 Examples

**A) 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 whether it will participate and if it will participate on which side of the balance the object should be put. Hence for the $n$'th weighing realization we have to choose

$$x_n \in \mathcal{X} = \{-1, 0, 1\}^4$$

where $x_n^i = 0$ means that in the $n$'th realization the $i$'th object does not participate while $x_n^i = -1$ and $x_n^i = 1$ means that the $i$'th object is put on the lefthand side and righthand side of the balance, respectively. Then the result of the $n$'th weighing realization will be

$$y_n = x_n'\beta + z_n = \beta^1 x_n^1 + ... + \beta^4 x_n^4 + z_n$$

where $\beta = (\beta^1, \beta^2, \beta^3, \beta^4)' \in I\!R^4$, $x_n = (x_n^1, x_n^2, x_n^3, x_n^4)' \in \mathcal{X}$ and $z_n$ is an error which may be also an outlier if some extraordinary shake happens. Then the equivalence theorems (see as in Section 3 for example Pukelsheim (1993)) provide that every design $X \in \mathcal{X}^N$ satisfying $X'X = N\,E_4$, where $E_4$ denotes the $4 \times 4$-identity matrix, is A-, D- and E-optimal in $\Delta = \mathcal{X}^N$, i.e. classically optimal. In particular if the sample size is $N = 8L$, where $L$ is an integer, then classically optimal designs are $X = (X_1|...|X_1)'$ and $X^* = (X_2|...|X_2)'$, where the matrix $X_1 = (1_4|X_0)' \in I\!R^{4\times 4}$ is repeated $2L$-times and the matrix $X_2 = ((1_4|X_0)'|(1_4|-X_0)') \in I\!R^{4\times 8}$ is repeated $L$ times. Thereby we have $1_4 = (1,1,1,1)'$ and

$$X_0 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \end{pmatrix}.$$

The design $X$ has minimum support but the maximum breakdown point at $X$ is approximately the half of that of $X^*$ which is the breakdown point maximizing design within the classically optimal designs in $\mathcal{X}^N$ (see Müller (1994)). Note that $X^*$ is not breakdown point maximizing within all designs of $\mathcal{X}^N$. But if we reduce $\mathcal{X}$ to $\mathcal{X}_0 = \{-1,1\}^4$ then $X^*$ also is breakdown point maximizing in $\mathcal{X}_0^N$ as Müller (1994) showed.

**B) Multiple regression on balls.** Another example where several classically optimal designs exist is the problem of multiple regression on balls which also was regarded in Müller (1994). For this problem Müller (1994) showed that classically optimal designs exist which are also breakdown point maximizing within all designs so that in particular they are breakdown point maximizing within the classically optimal designs.

# References


[1] Bickel, P.J. (1975). One-step Huber estimates in the linear model. *J. Amer. Statist. Assoc.* **70**, 428-434.

[2] 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.

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

[4] Davies, P.L. (1993). Aspects of robust linear regression. *Ann. Statist.* **21**, 1843-1899.

[5] Davies, P.L. (1994). Desirable properties, breakdown and efficiency in the linear regression model. *Statist. Probab. Lett.* **19**, 361-370.

[6] 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.

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

[8] Fedorov, V.V. (1989). Optimal design with bounded density: Optimization algorithms of the exchange type. *J. Statist. Plann. Inference* **22**, 1-13.

[9] Hössjer, O. (1991). Rank-based estimates in the linear model with high breakdown point. *J. Amer. Statist. Assoc.* **89**, 149-158.

[10] Jurečková, J. and Portnoy, S. (1987). Asymptotics for one-step M-estimators in regression with application to combining efficiency and high breakdown point. *Commun. Statist.-Theory Meth.* **16**, 2187-2199.

[11] 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.

[12] 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.*

[13] Müller, Ch.H. (1994). Optimal designs for robust estimation in conditionally contaminated linear models. *J. Statist. Plann. Inference.* **38**, 125-140.

[14] Müller, Ch.H. (1994). Optimal breakdown point maximizing designs. *Submitted to the Proceedings of PROBASTAT'94.* Tatra Mountains Mathematical Publications, Bratislava.

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

[16] Rousseuw, P.J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.* **79**, 871-880.

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

[18] Wierich, W. (1985). Optimum designs under experimental constraints for a covariate model and an intra-class regression model. *J. Statist. Plann. Inference* **12**, 27-40.

[19] Wynn, H. (1982). Optimum submeasures with applications to finite population sampling. In: *Statistical Decision Theory and Related Topics III*, Vol. 2. Academic Press, New York, 485-495.

[20] Yohai, V.J. and Maronna, R.A. (1979). Asymptotic behavior of M-estimators for the linear model. *Ann. Statist.* **7**, 258-268.