The Annals of Statistics
1997, Vol. 25, No. 5, 2024–2026

A NOTE ON TYLER'S MODIFICATION OF THE MAD FOR THE STAHEL–DONOHO ESTIMATOR

By Ursula Gather and Torsten Hilker

University of Dortmund

In 1994 Tyler proposed a modification of the MAD in order to achieve the maximal possible finite sample breakdown point for the Stahel–Donoho estimator based upon the median and the modified MAD. A proof is not given, however, and as far as we know has not been published yet. Moreover there seems to be an error concerning the exact definition of the modified MAD which is needed to achieve this goal. For the sake of theoretical completeness and in order to correct this error we calculate the finite sample breakdown point of modifications of the MAD and point out which of these modifications lead to a Stahel–Donoho estimator with maximal finite sample breakdown point.

Projection based multivariate location and scatter statistics lift univariate location and scale estimators to higher dimensions by considering the value of those statistics for all univariate projections of the data. Consequently, the properties of the multivariate estimators and the univariate statistics are closely related. Therefore, we first consider breakdown properties of a modification of the univariate MAD.

Let $X_n = \{x_1, \dots, x_n\}$ be a sample of n points in $\mathbb{R}$ with ordered values $x_{(1)} \le \cdots \le x_{(n)}$. The multiplicity of x_i in X_n is denoted by $c_i(X_n) = \#\{k\colon x_k = x_i\}$, $i = 1, \dots, n$, and the maximal multiplicity by $c(X_n) = \max\{c_1, \dots, c_n\}$. We only consider samples X_n where $c(X_n) \le n-1$. Let $[\mu(\cdot), \sigma(\cdot)]$ be a pair of translation and scale equivariant univariate location and scale statistics. In this situation Tyler (1994) proves that the finite sample replacement breakdown point [for a definition see Donoho and Huber (1983) and Tyler (1994)] fulfills

$$\varepsilon^*(X_n; \mu, \sigma) \le \lfloor (n - c(X_n) + 1)/2 \rfloor / n, \tag{1}$$

where $\lfloor x \rfloor$ denotes the greatest integer smaller than or equal to x. In the same way $\lceil x \rceil$ denotes the smallest integer greater than or equal to x. Consider the median $\operatorname{med}(X_n) = (x_{(\lceil n/2 \rceil)} + x_{(\lfloor n/2 \rfloor + 1)})/2$ and the following modifications of the MAD:

$$\operatorname{mad}_k(X_n) = \operatorname{med}_k(\{|x_1 - \operatorname{med}(X_n)|, \dots, |x_n - \operatorname{med}(X_n)|\}), \tag{2}$$

where

$$\operatorname{med}_k(X_n) = (x_{(\lceil (n+k)/2 \rceil)} + x_{(\lfloor (n+k)/2 \rfloor + 1)})/2 \quad \text{for } k = 0, \dots, n-1.$$

Note that $\operatorname{mad}_k(X_n) = 0$ if $2c(X_n) - n > k$.

Received September 1996; revised March 1997.

AMS 1991 *subject classifications*. Primary 62F35; secondary 62H12.

Key words and phrases. Breakdown point, modification of the MAD, multivariate location and scatter, Stahel–Donoho estimator.

LEMMA 1. *The finite sample replacement breakdown point of* $[\mathrm{med}, \mathrm{mad}_k]$ *is given by*

$$\varepsilon^*(X_n; \mathrm{med}, \mathrm{mad}_k) = \lfloor (n - 2c(X_n) + k + 2)/2 \rfloor / n \tag{3}$$

for $k \in \{\max\{0, 2c(X_n) - n\}, \dots, c(X_n) - 1\}$ *and by*

$$\varepsilon^*(X_n; \mathrm{med}, \mathrm{mad}_k) = \lfloor (n - k + 1)/2 \rfloor / n \tag{4}$$

for $k \in \{c(X_n), \dots, n - 1\}$.

The proof is given in the Appendix. It shows that (3) corresponds to implosion and (4) to explosion breakdown. By Lemma 1 one gets a pair of estimators with maximal possible finite sample breakdown point if $k = c(X_n) - 1$ or $k = c(X_n)$.

Turning to a multivariate sample X_n in $\mathbb{R}^p$, $p \geq 2$, we let $\varepsilon^{**}(X_n; \mu, \sigma)$ denote the uniform finite sample replacement breakdown point of $[\mu(\cdot), \sigma(\cdot)]$ as defined in Tyler (1994) when all univariate projections of the data are considered. Further, let S_{p-1} denote the p-dimensional unit sphere. Then Tyler states that $\varepsilon^{**}(X_n; \mu, \sigma) \leq \inf_{a \in S_{p-1}} \varepsilon^*(a' X_n; \mu, \sigma)$, where $a' X_n$ is the projected sample $\{a' x_1, \dots, a' x_n\}$. Equality holds if for all samples $Z = \{z_1, \dots, z_n\}$ the statistics $\mu(a' Z)$ and $\sigma(a' Z)$ are continuous functions of a. This is the case if μ and σ are the median and the modified MAD, respectively.

If X_n is in general position with $n \geq p + 1$, then there exists a direction $a_0 \in S_{p-1}$ such that $c(a_0' X_n) = p$. Using $[\mathrm{med}, \mathrm{mad}_{p-1}]$ as location and scale statistics, Lemma 1 leads to $\varepsilon^*(a_0' X_n; \mathrm{med}, \mathrm{mad}_{p-1}) = \lfloor (n - p + 1)/2 \rfloor / n$. For a direction $a_1 \in S_{p-1}$ with $c(a_1' X_n) < p$ we find $\varepsilon^*(a_1' X_n; \mathrm{med}, \mathrm{mad}_{p-1}) = \lfloor (n - p + 2)/2 \rfloor / n$. As X_n is in general position, we have $c(a' X_n) \leq p$ for all $a \in S_{p-1}$. Therefore $\varepsilon^{**}(X_n; \mathrm{med}, \mathrm{mad}_{p-1}) = \lfloor (n - p + 1)/2 \rfloor / n$. Applying Theorem 3.1 of Tyler (1994), we get that the corresponding Stahel–Donoho location and scatter statistics achieve the maximal possible finite sample breakdown point for affine equivariant multivariate location and scatter statistics [see also Davies (1987)]. With the same arguments we get that the Stahel–Donoho statistics based on $[\mathrm{med}, \mathrm{mad}_p]$ achieve the maximal possible finite sample breakdown point, too.

The problem in Tyler [(1994), equation 3.5] is that he proposes to use the average of the $\lceil n/2 \rceil + k$ and the $\lfloor n/2 \rfloor + k + 1$ smallest absolute deviations about the median as a modification of the MAD and that he concludes that with the special choice $k = p - 1$ these statistics yield the maximal possible breakdown point. With the above definition this proposal means to take $[\mathrm{med}, \mathrm{mad}_{2(p-1)}]$ as location and scale statistics. As shown before, however, this implies that $\varepsilon^{**}(X_n; \mathrm{med}, \mathrm{mad}_{2(p-1)}) = \lfloor (n - 2p + 3)/2 \rfloor / n$, such that the corresponding Stahel–Donoho estimator does not achieve the maximal possible finite sample breakdown point. Maronna and Yohai (1995) use the wrong modification, too. Contrary to this, our modification of the MAD as given in (2) and with $k \in \{p - 1, p\}$ leads to the desired optimality.

A referee has pointed out the possibility of obtaining results for the modified MAD by using arguments from the fourth section of Tyler (1994).

APPENDIX

SKETCH OF THE PROOF OF LEMMA 1. It is well known that the finite sample breakdown point of the univariate median is given by $\varepsilon^*(X_n; \text{med}) = \lfloor (n+1)/2 \rfloor / n$. Without loss of generality we consider a permutation of the points in X_n such that $x_1 = \cdots = x_{c(X_n)}$. A corrupted sample $X_{n,m}$ is constructed from X_n by replacing m points of X_n by arbitrary values. In the following, only sequences of corrupted samples are considered for which the median is bounded, that is, it does not break down.

If one takes the special corrupted sample $X_{n,m}$, $m = \lfloor (n+k)/2 \rfloor - c(X_n) + 1$, consisting of the points $x_1, \dots, x_{n-m}$ of X_n and the m corrupted values which are all selected as x_1, we get $\text{mad}_k(X_{n,m}) = 0$, such that the scale estimator breaks down. On the other hand one easily sees that, for all corrupted samples $X_{n,m}$ with $m \le \lfloor (n+k)/2 \rfloor - c(X_n)$, one has $\text{mad}_k(X_{n,m}) \ge \Delta(X_n)/2 > 0$ with $\Delta(X_n) = \min_{1 \le i < j \le n : x_i \ne x_j} |x_i - x_j|$. Therefore $m \ge \lfloor (n+k)/2 \rfloor - c(X_n) + 1 = \lfloor (n - 2c(X_n) + k + 2)/2 \rfloor$ must be fulfilled in order to achieve an implosion of $\text{mad}_k(X_{n,m})$. Hence the finite sample implosion breakdown point of mad_k is $\varepsilon^*_-(X_n; \text{mad}_k) = \lfloor (n - 2c(X_n) + k + 2)/2 \rfloor / n$.

We now consider explosion of the scale statistic. We use corrupted samples $X_{n,m}$, $m = \lfloor (n-k+1)/2 \rfloor$. For $k \in \{n-2, n-1\}$ we get $m = 1$ and we use $X_{n,m} = \{x_1, \dots, x_{n-1}, x_{(n)} + t\}$, $t > 0$. For $k \le n-3$ we get $m \ge 2$ and we use $X_{n,m} = \{x_1, \dots, x_{n-m}, x_{(1)} - t, x_{(n)} + t, \dots, x_{(n)} + t\}$, $t > 0$. In both cases $\text{med}(X_{n,m}) \in [x_{(1)}, x_{(n)}]$ and $\text{mad}_k(X_{n,m}) \ge t/2$, such that the scale statistic breaks down by choosing an arbitrarily large value of t. On the other hand we cannot achieve explosion of the scale statistic with less contamination since then the absolute deviations about the median for $\lfloor (n+k)/2 \rfloor + 1$ observations remain bounded. Therefore $m \ge \lfloor (n-k+1)/2 \rfloor$ in order to achieve explosion of $\text{mad}_k(X_{n,m})$. Thus the finite sample explosion breakdown point of mad_k is $\varepsilon^*_+(X_n; \text{mad}_k) = \lfloor (n-k+1)/2 \rfloor / n$.

Combining these results for $k \le c(X_n) - 1$ and $k \ge c(X_n)$, respectively, completes the proof. □

REFERENCES

DAVIES, P. L. (1987). Asymptotic behavior of S-estimates of multivariate location parameters and dispersion matrices. *Ann. Statist.* **15** 1269–1292.

DONOHO, D. L. and HUBER, P. J. (1983). The notion of breakdown point. In *A Festschrift for Erich Lehmann* (P. J. Bickel, K. A. Doksum and J. L. Hodges, eds.) 157–184. Wadsworth, Belmont, CA.

MARONNA, R. A. and YOHAI, V. J. (1995). The behavior of the Stahel–Donoho robust multivariate estimator. *J. Amer. Statist. Assoc.* **90** 330–341.

TYLER, D. E. (1994). Finite sample breakdown points of projection based multivariate location and scatter statistics. *Ann. Statist.* **22** 1024–1044.

LEHRSTUHL MATHEMATISCHE STATISTIK
UND INDUSTRIELLE ANWENDUNGEN
UNIVERSITÄT DORTMUND
D44221 DORTMUND
GERMANY
E-MAIL: gather@omega.statistik.uni-dortmund.de