A Note on Finite Sample Breakdown Points of Projection Based Multivariate Location and Scatter Statistics

Yijun Zuo [1]

Department of Mathematics
Arizona State University
Tempe, AZ 85287-1804
USA

E-mail: yijun.zuo@asu.edu

May 16, 1999 (revised January, 2000)

[1]Research partially supported by a grant (FGIA) from Arizona State University.

Abstract

In this note, general results of finite sample breakdown point are obtained for two classes of projection based location and scatter statistics: the Stahel-Donoho statistics and the Maronna-Yohai statistics. It is shown that these projection based location and scatter statistics can achieve the maximum breakdown point of affine equivariant multivariate location and scatter statistics. General relationships between the finite sample breakdown point of these statistics and the uniform finite sample breakdown point of the sample median and a modified sample median absolute deviation are formally established.

AMS 1991 Classification: Primary 62F35; Secondary 62G05, 62H12.

Key words and phrase: Breakdown point; location and scatter statistics; median absolute deviation; affine equivariance; projection pursuit.

A note on finite sample breakdown points of projection based multivariate location and scatter statistics

Yijun Zuo

Department of Mathematics, Arizona State University, Tempe, AZ 85287-1804, USA

(*E-mail: yijun.zuo@asu.edu*)

Abstract. In this note, general results of finite sample breakdown point are obtained for two classes of projection based location and scatter statistics: the Stahel-Donoho statistics and the Maronna-Yohai statistics. It is shown that these projection based location and scatter statistics can achieve the maximum breakdown point of affine equivariant multivariate location and scatter statistics. General relationships between the finite sample breakdown point of these statistics and the uniform finite sample breakdown point of the sample median and a modified sample median absolute deviation are formally established.

Key Words: Breakdown point, location and scatter statistics, median absolute deviation, affine equivariance, projection pursuit.

1 Introduction

Donoho (1982) and Donoho and Huber (1983) introduced a notion of finite sample replacement breakdown point as a measure of robustness of an estimator. Based on their idea, Davies (1987) and Tyler (1994) defined the finite sample replacement breakdown point of multivariate location and scatter statistics. Let X^n be a dataset of size n in $\mathbf{R}^d$ with $d \geq 1$, and $T(\cdot)$ and $V(\cdot)$ be multivariate location and scatter statistics respectively. The *finite sample replacement breakdown point* of multivariate location and scatter statistics $[T(\cdot), V(\cdot)]$ at X^n is defined to be

$$\epsilon^*(T, V; X^n) = \min\{\frac{m}{n} : b_m(T, V; X^n) = \infty\},$$

where

$$b_m(T, V; X^n) \equiv \sup_{X_m^n} (\|V(X^n)^{-1/2}(T(X_m^n) - T(X^n))\| + trace(V(X^n)V(X_m^n)^{-1} + V(X^n)^{-1}V(X_m^n)),$$

and X_m^n denotes the contaminated dataset resulting from replacing m points of X^n with arbitrary vectors. $b_m(T, V; X^n)$ is defined to be infinity if T or V (V^{-1}) does not exist (see Tyler (1994)). Roughly speaking, the breakdown point is the minimum fraction of contaminated points which can either drive the location estimator beyond any bound or lead the scatter matrix to have eigenvalues close to zero or infinity.

One desirable property of location and scatter statistics is affine equivariance. $T(\cdot)$ and $V(\cdot)$ are said to be *affine equivariant*, if $T(AX^n + b) = AT(X^n) + b$ and $V(AX^n + b) = AV(X^n)A'$ respectively, for any nonsingular matrix A and vector b. Clearly, high breakdown point $\epsilon^*(T, V; X^n)$ is another desirable property of location and scatter statistics. How high could $\epsilon^*(T, V; X^n)$ be? For affine equivariant multivariate location and scatter statistics, Davies (1987) showed that $\lfloor (n - d + 1)/2 \rfloor / n$ is an upper bound of $\epsilon^*(T, V; X^n)$ for a random sample X^n *in general position* (i.e., no more than d points lie in any $(d-1)$-dimensional hyperplane in $\mathbf{R}^d$), where "$\lfloor x \rfloor$" denotes the greatest integer no greater than x. For an *arbitrary* random sample, this upper bound becomes $\lfloor (n - c(X^n) + 1)/2 \rfloor / n$, as pointed out in Tyler (1994), where $c(X^n)$ is the maximum number of points of X^n contained in any $(d-1)$-dimensional hyperplane.

Projection based multivariate location and scatter statistics, such as the Stahel-Donoho (S-D) statistics (Stahel (1981) and Donoho (1982)) and the Maronna-Yohai (M-Y) statistics (Maronna and Yohai (1989) and Tyler (1994)) possess high breakdown point. The S-D statistics $T(\cdot)$ and $V(\cdot)$ are defined as:

$$T(X^n) = \frac{\sum_i w_i X_i}{\sum_i w_i}, \quad V(X^n) = \frac{\sum_i w_i (X_i - T(X^n))(X_i - T(X^n))'}{\sum_i w_i},$$

where $w_i = w(r(X_i; X^n))$ for $i = 1, \ldots, n$ with $w: \mathbf{R} \to \mathbf{R}^+$ being bounded and continuous, $w(\infty) = 0$ and $r^2 w(r)$ being bounded, $r(v; X^n) = \sup_{\|a\|=1} \{|a'v - \mu(a'X^n)|/\sigma(a'X^n)\}$ for any $v \in \mathbf{R}^d$, and $\mu(\cdot)$ and $\sigma(\cdot)$ are location and scale statistics in $\mathbf{R}$, and $\mu(\cdot)$ and $\sigma^2(\cdot)$ are affine equivariant. Here $\mathbf{R}^+$ is the set $\{x : x > 0\}$. Thus, $r(v; X^n)$ is a measure of "outlyingness" of point v with respect to X^n, $w(r)$ is a weight function which downweights outlying observations, and S-D statistics are outlyingness weighted mean and covariance matrix. Tyler (1994) modified a scatter statistic of Maronna and Yohai (1989) and extended their approach to multivariate location. The resulting location and scatter statistics $T(\cdot)$ and $V(\cdot)$, called M-Y statistics here, are defined as:

$$T(X^n) = \arg \inf_{t \in \mathbf{R}^d} \{r(t; X)\}, \quad V(X^n) = \arg \inf_V (\sup_{\|a\|=1} g(\frac{\sigma^2(a'X^n)}{a'Va})),$$

where $g: \mathbf{R}^+ \to \mathbf{R}$ is nonnegative, continuous with $g(0^+) = \infty$ and $g(\infty) = \infty$, and V is positive definite. T here is the point with minimum outlyingness. The $a'Va$ above represents an approximation of $\sigma^2(a'X^n)$, the sup part above may be interpreted as a measure of worst case discrepancy between $a'Va$ and $\sigma^2(a'X^n)$, V then is the matrix which minimizes this worst case discrepancy. It is not difficult to verify that both T and V of the S-D and M-Y statistics are affine equivariant.

Clearly, the behavior of the above projection based multivariate location and scatter statistics depends on the nature of the corresponding univariate location and scale statistics $\mu(\cdot)$ and $\sigma(\cdot)$. The finite sample breakdown points of the multivariate statistics, however, are not directly related to those of the corresponding univariate statistics. Consequently, Tyler (1994) introduced the uniform finite sample replacement breakdown point concept. The *uniform finite sample replacement breakdown point* of $[\mu(\cdot), \sigma^2(\cdot)]$ at X^n in $\mathbf{R}^d$ $(d \geq 1)$ is

$$\epsilon^{**}(\mu, \sigma^2; X^n) = \min\{\frac{m}{n} : \sup_{\|a\|=1} b_m(\mu, \sigma^2; a'X^n) = \infty\},$$

where $a'X^n = a'\{X_1, \ldots X_n\} = \{a'X_1, \ldots, a'X_n\}$.

Take μ and σ to be the sample median (Med) and the sample median absolute deviation (MAD) respectively, Donoho (1982) and Donoho and Gasko (1992) obtained the first pair of affine equivariant multivariate location and scatter statistics with high breakdown point. The replacement breakdown point $\epsilon^*(T, V; X^n)$ is as high as $\lfloor (n - 2d + 2)/2 \rfloor / n$, see Davies (1987). For $d > 1$, this replacement breakdown point, however, is still less than the upper bound $\lfloor (n - d + 1)/2 \rfloor / n$ of affine equivariant multivariate location and scatter statistics.

To achieve the upper bound of the finite sample breakdown point of affine equivariant location and scatter statistics, Tyler (1994) used the Med and a modified median absolute deviation MAD_k in the S-D and M-Y statistics. The idea of modifying the MAD to achieve higher breakdown point for related estimators was proposed in a personal communication of A. Siegel with Rousseeuw (see Rousseeuw, 1984) and in Lopuhaä (1991). For X^n *in general position*, Tyler (1994) presented, with no proof, the $\epsilon^*(T, V; X^n)$ of the S-D and M-Y statistics based on the Med and the modified MAD, and its relationship with the $\epsilon^{**}(\mu, \sigma^2; X^n)$ of the Med and the modified MAD.

Gather and Hilker (1997) believe that there seems to be an error concerning the exact definition of Tyler's modified MAD which is needed to achieve the maximum breakdown point. Consequently, they proposed a different modification of the MAD (which turns out to be connected with and enlarging Tyler's modification). For X^n *in general position*, they considered the relationship between the $\epsilon^*(T, V; X^n)$ of the S-D statistics and the $\epsilon^{**}(\mu, \sigma^2; X^n)$ of the Med and a specific case of their modified MAD.

In this note: (i) new results of $\epsilon^*(T, V; X^n)$ for *arbitrary* X^n are obtained for both the S-D and the M-Y statistics, generalizing the corresponding results in Tyler (1994) and Gather and Hilker (1997), and correcting some errors in the breakdown points given in Tyler (1994); (ii) it is shown that for some special values of k the S-D and M-Y statistics based on the MAD_k (defined below) attain the maximum breakdown point $\lfloor (n-d+1)/2 \rfloor / n$ of affine equivariant multivariate location and scatter statistics; (iii) for *arbitrary* X^n and *general* k, relationships between the $\epsilon^*(T, V; X^n)$ of the S-D and M-Y statistics and the $\epsilon^{**}(\mu, \sigma^2; X^n)$ of the Med and MAD_k are *formally* established, generalizing and complementing the corresponding work of Tyler (1994) and Gather and Hilker (1997).

2 Main results

For a random sample $X^n = \{X_1, \ldots, X_n\}$ in $\mathbf{R}$, the sample Med and the sample MAD are defined as follows.

$$\mathrm{Med}(X^n) = (X_{(\lfloor (n+1)/2 \rfloor)} + X_{(\lfloor (n+2)/2 \rfloor)})/2, \quad \mathrm{MAD}(X^n) = \mathrm{Med}_{1\le i\le n}\{|X_i - \mathrm{Med}(X^n)|\},$$

where $X_{(1)} < \ldots < X_{(n)}$ is the ordered values of $X_1, \ldots, X_n$. To achieve the maximum breakdown point of affine equivariant location and scatter statistics, Tyler (1994) used the Med and the following modified MAD in the S-D and M-Y statistics.

$$\mathrm{MAD}_k(X^n) = \mathrm{Med}_k\{|X_1 - \mathrm{Med}(X^n)|, \ldots, |X_n - \mathrm{Med}(X^n)|\},$$

where $0 \le k < \lfloor (n+1)/2 \rfloor$ and

$$\mathrm{Med}_k(X^n) = \left(X_{(k+\lfloor (n+1)/2 \rfloor)} + X_{(k+\lfloor (n+2)/2 \rfloor)}\right)/2.$$

Gather and Hilker (1997) believe that there seems to be an error concerning the exact definition of the above modification which is needed for the resulting S-D statistics to achieve maximum replacement breakdown point of affine equivariant multivariate location and scatter statistics. Consequently, they proposed a modification of the MAD, which is the same as the one above except

$$\mathrm{Med}_k(X^n) = \left(X_{(\lceil (n+k)/2 \rceil)} + X_{(\lfloor (n+k)/2 \rfloor + 1)}\right)/2, \quad \text{for} \quad k = 0, 1, \ldots, n-1,$$

where "$\lceil x \rceil$" denotes the smallest integer no smaller than x. It is easy to see that this second modification of the MAD is intimately connected with Tyler's modification in the sense that the MAD_{2k} of Gather and Hilker (1997) equals the MAD_k of Tyler (1994), and it is also enlarging Tyler's modification. In our following discussion, the first modification of the MAD is adopted, and we will show that for carefully chosen k this modification can lead the corresponding S-D and M-Y statistics to achieve the maximum replacement breakdown point. We first present the uniform finite sample replacement breakdown point of the Med and the MAD_k.

Theorem 2.1 *Let X^n be an arbitrary dataset in $\mathbf{R}^d$ with $d \geq 1$. Then the uniform finite sample replacement breakdown point of the* Med *and the* MAD_k *is*

$$\epsilon^{**}(\mathrm{Med}, \mathrm{MAD}_k; X^n) = \begin{cases} \frac{\lfloor (n+2+2k-2c(X^n))/2 \rfloor}{n} & \max\{0, 2c(X^n)-n\} \leq 2k < c(X^n), \\ \frac{\lfloor (n+1-2k))/2 \rfloor}{n} & c(X^n) \leq 2k < n. \end{cases}$$

PROOF: Clearly, for any unit vector a, $\lfloor (n+1)/2 \rfloor$ contaminated points are necessary and sufficient to breakdown $\mathrm{Med}(a'X^n)$.

To implode the MAD_k at some projection of X^n, move m points onto a hyperplane containing $c(X^n)$ original data points such that $m + c(X^n) \geq k + (n+1)/2$ or $m + c(X^n) \geq k + (n+2)/2$ when n is odd or even respectively. It is then not difficult to see that $\mathrm{MAD}_k(a'Z^n) = 0$ for a unit vector a perpendicular to the hyperplane, where Z^n is the contaminated dataset resulting from the moving of m original data points in X^n onto the hyperplane. Thus $m = \lfloor (n + 2k + 2 - 2c(X^n))/2 \rfloor$ contaminated points are sufficient to implode the MAD_k. It is not difficult to show that $m = \lfloor (n + 2k + 2 - 2c(X^n))/2 \rfloor$ contaminated points are also necessary to implode the MAD_k.

To explode the MAD_k at some projection of X^n, move m ($< \lfloor n+1 \rfloor/2$) points to the same site y far away from the original data cloud with $m \geq n-(k+(n+1)/2-1)$ or $m \geq n-(k+(n+2)/2-1)$ when n is odd or even respectively. Let $\|y\| \to \infty$, and let a to be a unit vector parallel to y, then $\mathrm{MAD}_k(a'Z^n) \to \infty$, where Z^n is the contaminated dataset after moving m points from X^n to the same site y. Thus $m = \lfloor (n+1-2k)/2 \rfloor$ contaminated points are sufficient to explode the MAD_k. It is not difficult to show that $m = \lfloor (n+1-2k)/2 \rfloor$ contaminated points are also necessary to explode the MAD_k.

Hence $m = \min\{\lfloor (n + 2k + 2 - 2c(X^n))/2 \rfloor, \lfloor (n+1-2k)/2 \rfloor\}$ contaminated points are sufficient and necessary to breakdown the MAD_k for some projected data. The result follows. □

Remark 2.1 (i) For X^n *in general position*, Tyler (1994) claimed, with no proof, that

$$\epsilon^{**}(\mathrm{Med}, \mathrm{MAD}_k; X^n) = \lfloor (n - 2d + k + 2)/2 \rfloor / n, \quad k = 0, 1, \ldots, d-1.$$

The theorem above holds for *arbitrary* X^n in $\mathbf{R}^d$. When X^n is in general position, $c(X^n)$ will equal d, the theorem then yields a result corresponding to the claim, in the meantime an error in the claim is corrected. (ii) When $\mathbf{R}^d = \mathbf{R}$, the theorem immediately yields the Lemma 1 of Gather and Hilker (1997). On the other hand, employing the argument of Gather and Hilker (1997) (page 2025), one can show that the Lemma 1 of Gather and Hilker (1997) also holds for the uniform breakdown point. $\square$

In the following, it is shown that the finite sample replacement breakdown points of the S-D and M-Y statistics based on the Med and MAD_k are the same as the uniform finite sample replacement breakdown point of the Med and MAD_k.

Theorem 2.2 *Let X^n be an arbitrary dataset in $\mathbf{R}^d$ with $d \geq 1$. Then the finite sample replacement breakdown points of the S-D and M-Y statistics based on the* Med *and* MAD_k *equal* $\epsilon^{**}(\mathrm{Med}, \mathrm{MAD}_k; X^n)$ *given in Theorem 2.1.*

PROOF: We prove the assertion only for the S-D statistics, the assertion for the M-Y statistics can be proved accordingly. Let T and V be the S-D location and scatter statistics respectively.

Case (A): $\max\{0, 2c(X^n) - n\} \leq 2k < c(X^n)$. Move $m = \lfloor (n + 2 + 2k - 2c(X^n))/2 \rfloor$ points from X^n onto a hyperplane H containing $c(X^n)$ original data points. Assume, w.l.o.g., that $X_1, \ldots, X_{m^*}$ are data points which are not on H, where $m^* = n - c(X^n) - m$. Choose a unit vector a perpendicular to H, then it is not difficult to see that $\mathrm{MAD}_k(a'Z^n) = 0$, where Z^n is the contaminated dataset resulting from moving of m points of X^n onto H. Hence $r(X_i, Z^n) = \infty$, and thus $w_i = 0$ for $1 \leq i \leq m^*$. It follows that either $\sum_{i=1}^n w_i = 0$ or

$$a'V(Z^n)a \;=\; \frac{\sum_{i>m^*}^n w_i a'(Z_i - T(Z^n))(Z_i - T(Z^n))'a}{\sum_{i>m^*}^n w_i} \;=\; 0.$$

The latter holds since $Z_i - T(Z^n)$ is a vector contained in the hyperplane H for all $i > m^*$. The two cases lead to either undefined T and V or undefined V^{-1} (since $V(Z^n)$ has a zero eigenvalue). It follows that $b_m = \infty$. Hence, $\epsilon^*(T, V; X^n) \leq m/n$.

On the other hand, $\epsilon^*(T, V; X^n) \geq m/n$ follows immediately from Theorem 2.1 and Tyler's (1994) Theorem 3.1.

Case (B): $c(X^n) \leq 2k < n$. We first show that $m = \lfloor (n+1-2k))/2 \rfloor$ points are sufficient to breakdown T and (or) V. It is not difficult to see in this case that $\text{Med}(a'Z^n) \leq \mu_\infty < \infty$ for some μ_∞ and any unit vector a and contaminated dataset Z^n. Move m points in X^n, say $X_1, \ldots, X_m$, to the same site y far away from the original convex hull of X^n such that any hyperplane containing y does not contain $c(X^n)$ points from $\{X_{m+1}, \ldots, X_n\}$. Let $\|y\|$ approach infinity. It is not difficult to show that $r(Z_i, Z^n) \to \frac{1}{2}$ (or 1) for n even (or odd) or $r(Z_i, Z^n) = |\ \text{Med}(a'Z^n)|/\ \text{MAD}_k(a'Z^n) \equiv C$ with a perpendicular to y and a constant C for $1 \leq i \leq m$. Hence $\lim \sum_{i=1}^m w_i > 0$ as $\|y\| \to \infty$. Thus we have

$$\begin{aligned} \|T(Z^n)\| &\geq \frac{m \cdot w_1 \cdot \|y\| - \sum_{i=m+1}^n w_i \|X_i\|}{\sum_{i=1}^n w_i} \\ &\to \infty \qquad \text{as } \|y\| \to \infty. \end{aligned}$$

Thus $m = \lfloor (n+1-2k))/2 \rfloor$ points are sufficient to breakdown T.

On the other hand, $\epsilon^*(T, V; X^n) \geq m/n$ again follows immediately from Theorem 2.1 and Tyler's (1994) Theorem 3.1. □

Remark 2.2 In the above theorem, we have established, for *arbitrary* X^n and *general* k, relationships of the finite sample breakdown points of the S-D and the M-Y statistics with the uniform finite sample breakdown point of the Med and MAD_k. For X^n *in general position*, these relationships were mentioned, with no proof, in Tyler (1994). Gather and Hilker (1997) considered, for X^n *in general position* and a *special choice* of k, the relationship of the finite sample breakdown point of the S-D statistics with the uniform finite sample breakdown point of the Med and their MAD_k. The results given here generalize their findings. □

Remark 2.3 Take $2k$ to be $c(X^n) - 1$ (or $c(X^n)$) when $c(X^n)$ is odd (or even), then the S-D or M-Y statistics based on the Med and MAD_k achieve the maximum finite sample replacement breakdown point of affine equivariant multivariate location and scatter statistics. Particularly, for X^n in general position, taking $2k$ to be $d-1$ (or d) when d is odd (or even), these statistics achieve the maximum breakdown point. Note that the MAD_{2k} of Gather and Hilker (1997) equals the

MAD_k of Tyler (1994). So, taking $2k$ to be $d-1$ (or d, respectively) above exactly means applying the MAD_k of Gather and Hilker with k being $d-1$ (or d, respectively). $\square$

3 Acknowledgments

The author greatly appreciates two referees' helpful comments and suggestions, which led to several important corrections and improvements. He thanks Professor Robert Serfling for his insightful comments and suggestions. The work is partially supported by a grant (FGIA) from Arizona State University.

References

Davies PL (1987) Asymptotic behavior of S-estimates of multivariate location parameters and dispersion matrices. Ann. Statist. 15:1269-92

Donoho DL (1982) Breakdown properties of multivariate location estimators. Ph.D. qualifying paper, Department of Statistics, Harvard University

Donoho DL, Huber PJ (1983) The notion of breakdown point. In: Bickel PJ, Doksum KA, Hodges JL Jr (eds.) A Festschrift for Erich L. Lehmann (Wadsworth, Belmont,CA), pp. 157-184

Donoho DL, Gasko M (1992) Breakdown properties of location estimates based on halfspace depth and projected outlyingness. Ann. Statist. 20: 1803–1827

Gather U, Hilker T (1997) A note on Tyler's modification of the MAD for the Stahel-Donoho estimator. Ann. Statist. 25: 2024-2026

Lopuhaä HP (1991) Breakdown point and asymptotic properties of multivariate S-estimators and τ-estimators: a summary. In: Stahel W, Weisberg S (eds.) Directions in Robust Statistics and Diagnostics. Springer-Verlag, New York, pp. 167-182

Maronna RA, Yohai VJ (1989) Bias-robust estimates of multivariate scatter based on projections. Unpublished manuscript

Rousseeuw PJ (1984) Least Median of Squares Regression. J. Amer. Statist. Assoc. 79: 871-880

Stahel WA (1981) Breakdown of covariance estimators. Research Report 31, Fachgruppe für Statistik, ETH, Zürich

Tyler DE (1994) Finite sample breakdown points of projection based multivariate location and scatter statistics. Ann. Statist. 22: 1024-1044