Finite Sample Tail Behavior of the Multivariate Trimmed Mean Based on Tukey-Donoho Halfspace Depth

Yijun Zuo [1]

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

E-mail: yijun.zuo@asu.edu
URL: http://math.la.asu.edu/~zuoyijun

June, 1999 (Revised April, 2000)

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

Abstract

Finite sample tail behavior of the Tukey-Donoho halfspace depth based multivariate trimmed mean is investigated with respect to a tail performance measure. It turns out that the tails of the sampling distribution of the α-depth-trimmed mean approach zero at least $\lfloor \alpha n \rfloor$ times as fast as the tails of the underlying population distribution and could be $n - \lfloor \alpha n \rfloor + 1$ times as fast. In addition, there is an intimate relationship among the tail behavior, the halfspace depth, and the finite sample breakdown point of the estimator. It is shown that the lower tail performance bound of the depth based trimmed mean is essentially the same as its halfspace depth and the breakdown point. This finding offers a new insight into the notion of the halfspace depth and extends the important role of the tail behavior as a quantitative assessment of robustness in the regression (He, Jurečková, Koenker and Portnoy (1990)) and the univariate location settings (Jurečková (1981)) to the multivariate location setting.

AMS 1991 Classification: Primary 62F35; Secondary 62G35, 62H12.

Key words and phrases: Depth-trimmed mean, tail behavior, halfspace depth, robustness, breakdown point.

Short Running title: Tail behavior of depth-trimmed mean.

1 Introduction

Let $T_n = T_n(X_1, \ldots, X_n)$ be an estimator of an unknown location parameter θ based on a random sample $X_1, \ldots, X_n$ from a random vector X in $\mathbf{R}^d$ ($d \geq 1$) with $F(x - \theta)$ and P_θ being the corresponding distribution and probability measure. θ could be the point of symmetry of a centrally, angularly or even halfspace symmetric distribution. For definitions and some discussions of these notions of symmetry, see Zuo and Serfling (2000c). In our following discussion, however, symmetry is not required.

How well does T_n perform as an estimator of θ? Various performance measures have been proposed and studied. Besides the classical (*asymptotic*) *unbiasedness* and *mean squared error* approaches, the "inaccuracy" probability of the error of the estimator exceeding a fixed number $a > 0$

$$P_\theta(|T_n - \theta| > a)$$

has been suggested and investigated. *Consistency* has been a typical asymptotic performance measure of T_n based on the inaccuracy probability. Bahadur (1967), Fu (1975), and Sievers (1978) have also investigated the asymptotic behavior of T_n in $\mathbf{R}^1$ based on the behavior of

$$-\frac{1}{n} \ln P_\theta(|T_n - \theta|) > a), \quad \text{as } n \to \infty,$$

for *fixed* a. The limit above (if exists) is called "inaccuracy rate" of T_n, which measures the rate of the inaccuracy (or tail) probability of T_n approaching zero. The inaccuracy rate is expected to be high for good estimator T_n.

In practice, however, the sample size n is *fixed* and *finite*. In sharp contrast with asymptotic approach to the performance of T_n, Jurečková (1981) studied, for centrally symmetric and continuous F in $\mathbf{R}^1$, the performance of location estimator T_n based on the measure

$$-\ln P_\theta(|T_n - \theta| > a) / -\ln(1 - F(a)), \quad \text{as } a \to \infty,$$

for *fixed* n. The limit above (if exists) measures the rate of the tail probability of the sampling distribution of T_n approaching zero relative to the tail probability of the underlying population distribution. He, Jurečková, Koenker and Portnoy (HJKP) (1990) extended this measure to the p-dimensional regression context and explored the tail behavior of various regression estimators.

Generalizing the idea of Jurečková (1981) in $\mathbf{R}^1$ to higher dimensions, we now introduce a tail performance measure for multivariate location estimators in $\mathbf{R}^d$ $(d \geq 1)$. For *fixed* sample size n and unit vector u, clearly one would like the "tail probability" $P_\theta(u'(T_n - \theta) > a)$ to be as small as possible for a fixed number $a > 0$ and to approach zero as fast as possible when $a \to \infty$. On the other hand, the behavior of this tail probability depends inherently on the behavior of the tail probability $P_\theta(u'(X - \theta) > a)$ of an underlying distribution. Thus we introduce a measure of tail performance, the *finite sample tail behavior*, of T_n in $\mathbf{R}^d$ based on

$$B(T_n; u, a) = \frac{-\ln P_\theta(u'(T_n - \theta) > a)}{-\ln P_\theta(u'(X - \theta) > a)}, \qquad \text{as } a \to \infty,$$

for a unit vector u and a *fixed* n. The limit above (if exists) measures the relative rate of the tail probability of the sampling distribution of T_n approaching zero with respect to the tail probability of the underlying population distribution.

Clearly, for good estimator T_n one would expect the tail performance measure $B(T_n; u, a)$ to be reasonably high for any unit vector u as $a \to \infty$. How high can $B(T_n; u, a)$ be as a approaches infinity? We answer the question in this paper for the α-depth-trimmed mean (defined below), a promising type of multivariate location estimators. Specifically, we examine for the α-depth-trimmed mean

$$\underline{B}(T_n; a) = \inf_{\|u\|=1} B(T_n; u, a) \ \text{ and } \ \overline{B}(T_n; a) = \sup_{\|u\|=1} B(T_n; u, a), \quad \text{as } a \to \infty$$

for fixed n. Lower and upper bounds for $\underline{B}(T_n; a)$ and $\overline{B}(T_n; a)$ are obtained for the estimator as $a \to \infty$. It turns out that the tail behavior of the α-depth-trimmed mean is inherently linked with its halfspace depth and breakdown point (defined below). This finding suggests an important role of the tail behavior of *multivariate location estimators* on the robustness scene, and offers a new interpretation of the notion of halfspace depth.

A location estimator T_n is said to be *translation equivariant* if $T_n(X^n + b) = T_n(X^n) + b$, for any vector $b \in \mathbf{R}^d$, where $X^n = \{X_1, \ldots, X_n\}$ and $X^n + b = \{X_1 + b, \ldots, X_n + b\}$. For translation equivariant T_n, $P_\theta(u'(T_n - \theta) > a) = P_0(u'T_n > a)$ and $P_\theta(u'(X - \theta) > a) = P_0(u'X > a)$, where P_0 is the probability measure corresponding to $X - \theta$. We thus assume in the following discussion that $\theta = 0$ and write P for P_0 for short.

2 Tail behavior of the depth-trimmed mean

Tukey (1975) introduced the notion of the "halfspace depth" and suggested its role in defining multivariate analogues of univariate rank and order statistics via depth-induced "contours". Further development in the halfspace depth was stimulated by Donoho (1982) and Donoho and Gasko (1992). The *halfspace depth* (HD) of a point $x \in \mathbf{R}^d$ with respect to a sample X^n is defined to be the minimum number of sample points contained in a closed halfspace containing x, that is,

$$\begin{aligned} HD_n(x; X^n) &= \inf_H \{\#\{X_i : X_i \in H\} : \ H \text{ a closed halfspace, } x \in H\} \\ &= \inf_{\|u\|=1} \sum_{i=1}^n \mathbf{I}(u'X_i \geq u'x), \end{aligned}$$

where u is a unit vector in $\mathbf{R}^d$ and $\mathbf{I}(\cdot)$ is an indicator function. Let $\lfloor x \rfloor$ be the largest integer $\leq x$ for $x \in \mathbf{R}^1$. The *α-depth-trimmed mean* $T_{\lfloor \alpha n \rfloor}$ in $\mathbf{R}^d$ $(0 \leq \alpha < 1/3)$, an analog of the univariate α-trimmed mean, is defined to be the average of all sample points which have halfspace depth no less than $\lfloor \alpha n \rfloor$ with respect to a given sample X^n, that is,

$$T_{\lfloor \alpha n \rfloor} = \text{Ave}\{X_i : \ HD_n(X_i; X^n) \geq \lfloor \alpha n \rfloor\}.$$

Here and in the following, $T_{\lfloor \alpha n \rfloor}$ is assume to be *well defined* (i. e., the above set on the right hand side is nonempty). It is easy to see that $T_{\lfloor \alpha n \rfloor}$ is translation equivariant.

To examine how well $T_{\lfloor \alpha n \rfloor}$ performs as a location estimator, we look at its tail behavior. Specifically, we would like to know how fast its tail probability approaches zero relative to the tail probability of the underlying population distribution. We have

Theorem 2.1 *Let $T_{\lfloor \alpha n \rfloor}$ be the α-depth-trimmed mean in $\mathbf{R}^d$ and well defined. Then*

$$\lfloor \alpha n \rfloor \leq \liminf_{a\to\infty} \underline{B}(T_{\lfloor \alpha n \rfloor}; a) \leq \limsup_{a\to\infty} \overline{B}(T_{\lfloor \alpha n \rfloor}; a) \leq n - \lfloor \alpha n \rfloor + 1.$$

PROOF: (i) The inequality in the middle is trivial. Let D_k be the *k-depth-trimmed region*, defined to be the set of all $x \in \mathbf{R}^d$ such that $HD_n(x; X^n) \geq k$ (we assume that D_k is not empty). Then by the definition of the halfspace depth, D_k is the intersection of all the d-dimensional halfspaces containing $n-k+1$ points of the dataset X^n. Consequently, D_k is convex and nested. Therefore $T_{\lfloor \alpha n \rfloor} \in D_{\lfloor \alpha n \rfloor}$. Hence if $u'T_{\lfloor \alpha n \rfloor} > a$ for a given unit vector u and a large positive number a, then there must exist at least some $\lfloor \alpha n \rfloor$ points in X^n such that the projection of each of these points to u is greater than a. Denote this latter event by $A(a,u)$. We have for large a

$$\begin{aligned} P(u'T_{\lfloor \alpha n \rfloor} > a) &\leq P(A(a,u)) \\ &= \sum_{s=\lfloor \alpha n \rfloor}^{n} \binom{n}{s} (P(u'X > a))^s (P(u'X \leq a))^{n-s} \\ &= (P(u'X > a))^{\lfloor \alpha n \rfloor} \left(\sum_{s=\lfloor \alpha n \rfloor}^{n} \binom{n}{s} (P(u'X > a))^{s-\lfloor \alpha n \rfloor} (P(u'X \leq a))^{n-s} \right). \end{aligned}$$

Write

$$f(u;a) = \sum_{s=\lfloor \alpha n \rfloor}^{n} \binom{n}{s} (P(u'X > a))^{s-\lfloor \alpha n \rfloor} (P(u'X \leq a))^{n-s},$$

then

$$\frac{-\ln P(u'T_{\lfloor \alpha n \rfloor} > a)}{-\ln P(u'X > a)} \geq \lfloor \alpha n \rfloor + \frac{-\ln f(u;a)}{-\ln P(u'X > a)}.$$

If we can show that the second term in the right hand side of the above inequality approaches zero *uniformly* with respect to u as a approaches infinity, then the desirable leftmost inequality follows. It follows, by replacing $P(u'X > a)$ and $P(u'X \leq a)$ with 1 in $f(u;a)$, that

$$f(u;a) \leq \sum_{s=s_0}^{n} \binom{n}{s} \equiv U,$$

with $s_0 = \lfloor \alpha n \rfloor$. Dropping all terms except the first term in $f(u; a)$ and utilizing Cauchy-Schwartz inequality, we have

$$f(u; a) \geq \binom{n}{s_0} (P(u'X \leq a))^{n-s_0} \geq \binom{n}{s_0} (P(\|X\| \leq a))^{n-s_0}.$$

For sufficiently large a, $P(\|X\| \leq a) \geq 1/2$, thus

$$f(u; a) \geq \binom{n}{s_0} (1/2)^{n-s_0} \equiv L.$$

Since $P(u'X > a) \leq P(\|X\| > a) \to 0$ uniformly with respect to u as $a \to \infty$, thus $\ln P(u'X > a) \to -\infty$ uniformly with respect to u as $a \to \infty$. Combining this with squeeze theorem and the fact that $\ln f(u; a)$ is uniformly bounded with respect to u ($\ln L \leq \ln f(u; a) \leq \ln U$) gives

$$\frac{-\ln f(u; a)}{-\ln P(u'X > a)} \to 0, \quad \text{uniformly with respect to } u \text{ as } a \to \infty.$$

The desired leftmost inequality now follows.

(ii) To prove the rightmost inequality, denote by $B(a, u)$ the event that there are at least some $(n - \lfloor \alpha n \rfloor + 1)$ sample points in X^n such that the projection of each of these points to u is greater than a for unit vector u and large positive number a. Now if $B(a, u)$ occurs, then $u'T_{\lfloor \alpha n \rfloor} > a$ (since $T_{\lfloor \alpha n \rfloor} \in D_{\lfloor \alpha n \rfloor}$). Define $m = n - \lfloor \alpha n \rfloor + 1$, we have,

$$\begin{aligned} P(u'T_{\lfloor \alpha n \rfloor} > a) &\geq P(B(a, u)) \\ &= \sum_{s=m}^{n} \binom{n}{s} (P(u'X > a))^{s} (P(u'X \leq a))^{n-s} \\ &= (P(u'X > a))^{m} \left(\sum_{s=m}^{n} \binom{n}{s} (P(u'X > a))^{s-m} (P(u'X \leq a))^{n-s} \right). \end{aligned}$$

Set

$$g(u; a) = \sum_{s=m}^{n} \binom{n}{s} (P(u'X > a))^{s-m} (P(u'X \leq a))^{n-s},$$

then

$$\frac{-\ln P(u'T_{\lfloor \alpha n \rfloor} > a)}{-\ln P(u'X > a)} \leq m + \frac{-\ln g(u; a)}{-\ln P(u'X > a)}.$$

Now employing an argument similar to that utilized in part (i), we have

$$\frac{-\ln g(u;a)}{-\ln P(u'X>a)} \to 0, \quad \text{uniformly with respect to } u \text{ as } a \to \infty.$$

The desired rightmost inequality follows. $\square$

Remark 2.1 (i) Roughly speaking, the theorem says that the tails of the sampling distribution of the α-depth-trimmed mean $T_{\lfloor \alpha n \rfloor}$ approach zero at least $\lfloor \alpha n \rfloor$ times as fast as the tails of underlying population distributions and could be $n - \lfloor \alpha n \rfloor + 1$ times as fast. (ii) The theorem reveals a close connection between the finite sample tail behavior and the halfspace depth of the $T_{\lfloor \alpha n \rfloor}$. The lower tail performance bound of the $T_{\lfloor \alpha n \rfloor}$ is essentially the same as the lower bound of the halfspace depth of the estimator. Thus the theorem gives an interpretation of the halfspace depth of the $T_{\lfloor \alpha n \rfloor}$ based on its tail behavior. $\square$

Remark 2.2 The tail behavior of the $T_{\lfloor \alpha n \rfloor}$ is also intimately connected with its breakdown point. The finite sample *replacement* breakdown point of an estimator $T_n(\cdot)$ at X^n is defined to be the minimum number of contaminating points m such that after replacing original m points by m contaminating points, one can drive the estimator beyond any bounds (see Donoho and Huber (1983)). It is shown in the Appendix that the replacement breakdown point of the $T_{\lfloor \alpha n \rfloor}$ is the same as its lower tail performance bound. $\square$

Now consider univariate α-trimmed mean $\overline{X}_\alpha$ $(0 \le \alpha < 1/2)$ defined as

$$\overline{X}_\alpha = \frac{1}{n - 2\lfloor \alpha n \rfloor} \sum_{i=\lfloor \alpha n \rfloor + 1}^{n - \lfloor \alpha n \rfloor} X_{(i)},$$

where $X_{(1)} \le \ldots \le X_{(n)}$ are ordered statistics of a sample $X_1, \ldots, X_n$ in $\mathbf{R}^1$. We have

Corollary 2.1 *Let $\overline{X}_\alpha$ be the α-trimmed mean in $\mathbf{R}^1$. Then*

$$\lfloor \alpha n \rfloor + 1 \le \liminf_{a\to\infty} \underline{B}(\overline{X}_\alpha; a) \le \limsup_{a\to\infty} \overline{B}(\overline{X}_\alpha; a) \le n - \lfloor \alpha n \rfloor.$$

PROOF: The α-trimmed mean $\overline{X}_\alpha$ is the average of the sample points with halfspace depth $\geq \lfloor \alpha n \rfloor + 1$. (Or the finite sample replacement breakdown point of the $\overline{X}_\alpha$ is $\lfloor \alpha n \rfloor + 1$.) The desirable result now follows immediately from Theorem 2.1. □

For the univariate sample median (Med_n) ($\lfloor \alpha n \rfloor + 1 = \lfloor (n+1)/2 \rfloor$), we have

Corollary 2.2 *Let* Med_n *be the univariate sample median. Then*

$$\lfloor (n+1)/2 \rfloor \leq \liminf_{a \to \infty} \underline{B}(\text{Med}_n; a) \leq \limsup_{a \to \infty} \overline{B}(\text{Med}_n; a) \leq \lfloor (n+2)/2 \rfloor.$$

Remark 2.3 The results in this section hold for arbitrary F in $\mathbf{R}^d$. The result in Corollary 2.2 was also given in Jurečková (1981) (there F was assumed to be continuous and symmetric for convenience). □

3 Conclusion

Lopuhaä and Rousseeuw (LR) (1991) made a first attempt to generalize a result of HJKP (1990) to the multivariate location setting. The tail performance measure for multivariate location estimators in LR (1991) is different from the one in this paper, and the estimators considered in LR (1991) satisfy a monotone condition and the population distribution are spherically symmetric. The result in this paper does not follow from LR (1991).

HJKP (1990) first revealed the connection between the tail behavior and the breakdown point of various regression and univariate monotone location estimators for continuous and symmetric F satisfying some tail conditions in $\mathbf{R}^1$. This connection is extended to the multivariate depth trimmed mean in this paper for *arbitrary* F in $\mathbf{R}^d$. Further, a connection of the halfspace depth with the tail behavior and the breakdown point of the estimator is revealed in the current paper, offering a new insight into the notion of halfspace depth.

Appendix

Proof of Remark 2.2. Assume that m points are sufficient to break down the $T_{\lfloor \alpha n \rfloor}$. Then there must exist at least one sample point X_i, which is far away from the original convex hull formed by the original sample points, with the halfspace depth at least $\lfloor \alpha n \rfloor$. By the separating hyperplane theorem, the number of contaminating points m must be at least the depth of the point X_i. Hence $m \geq \lfloor \alpha n \rfloor$.

Now we show that $m \leq \lfloor \alpha n \rfloor$. Move $\lfloor \alpha n \rfloor$ points from the original sample X^n to the same point y far away from the original convex hull. Clearly, the halfspace depth of y is at least $\lfloor \alpha n \rfloor$. Hence the $T_{\lfloor \alpha n \rfloor}$ breaks down as $\|y\|$ approaches infinity. Thus, $m \leq \lfloor \alpha n \rfloor$. $\square$

Acknowledgments. The author thanks the editor and two referees for their valuable comments and suggestions, which led to improvements in the paper. The research is partially supported by a grant (FGIA) from Arizona State University.

References

Bahadur RR (1967) Rates of convergence of estimators and test statistics. Ann. Math. Statist. 38: 303-324

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

Donoho DL, Huber PJ (1983) The notion of breakdown point. In: Bickel PJ, Doksum KA, Hodges JL Jr (eds.) A Festschrift foe 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

Fu JC (1975) The rate of convergence of consistent point estimators. Ann. Statist. 3: 234-240

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

Jurečková J (1981) Tail behavior of location estimators. Ann. Statist. 9: 578-585

Lopuhaä HP, Rousseeuw J (1991) Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. Ann. Statist. 19:, 229-248

Sievers GL (1978) Estimation of location: A large deviation comparison. Ann. Statist. 6: 610-618

Tukey JW (1975) Mathematics and picturing data. Proc. Intern. Congr. Math. Vancouver 1974 2: 523–531

Zuo Y, Serfling R (2000) On the performance of some robust nonparametric location measures relative to a general notion of multivariate symmetry. J. Statist. Plann. Inference 84: 55-79