stichting
mathematisch
centrum

AFDELING MATHEMATISCHE STATISTIEK SW 66/79 SEPTEMBER
(DEPARTMENT OF MATHEMATICAL STATISTICS)

J. JUREČKOVÁ

ESTIMATION OF LOCATION AND CRITERION OF TAILS

Preprint

2e boerhaavestraat 49 amsterdam

Printed at the Mathematical Centre, 49, 2e Boerhaavestraat, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O).

1980 Mathematical subject classification: 62G05, 62G35, 62F10

Estimation of location and criterion of tails*)

by

Jana Jurecková

ABSTRACT

Let $X_1,\ldots,X_n$ be a sample from a population with the density $f(x-\theta)$ such that f is symmetric and positive. It is proved that the tails of distribution of a translation-equivariant estimator of θ tend to 0 at most n times faster than the tails of basic distribution. The sample mean is shown being good in this sense for exponentially-tailed distributions while it becomes poor if there is a contaminacy by a heavy-tailed distribution. The rates of convergence of the tails of robust estimators are shown to be bounded away from the lower as well as from the upper bound.

KEY WORDS & PHRASES: *tails of a distributions, sample mean, L-estimator, trimmed L-estimator, M-estimator, median, Hodges-Lehmann's estimator.*

---

*) This report will be submitted for publication elsewhere.

1. INTRODUCTION

Let $X_1,\dots,X_n$ be a sequence of independent random variables identically distributed according to an absolutely continuous distribution function $F(x-\theta)$ with the density $f(x-\theta)$ such that $f(-x) = f(x) > 0$, $x \in R^1$; otherwise f is unspecified. The problem is that of estimating θ as a center of symmetry of an unknown symmetric absolutely continuous distribution. For each fixed n let $T_n = T_n(X_1,\dots,X_n)$ be an estimator of θ based on $X_1,\dots,X_n$.

Different measures of performance of T_n have been suggested and investigated. Besides the classical mean-square-error approach, the probability

$$(1.1) \qquad P_\theta(|T_n-\theta| > a)$$

of the absolute error not exceeding a fixed number $a > 0$ has been considered by several authors. If the sequence $\{T_n\}$ is consistent for θ, then the inaccuracy (1.1) tends to 0 as $n \to \infty$. BAHADUR [1], [2] proposed the limit

$$(1.2) \qquad \lim_{n\to\infty}\{ -\frac{1}{n}\,\ell n\; P_\theta(|T_n-\theta| > a)\} = e$$

for a fixed $\underline{a} > 0$ as a measure of performance of T_n, if the limit exists. BAHADUR [2] and FU [4] gave an upper bound for e for consistent sequences of estimators. SIEVERS [6] evaluated the limits e and their upper bounds for several estimators and several distribution shapes. From this point of view he found the sample median less efficient than the sample mean not only for normal but also for logistic distribution. We observed a similar feature even in the case of double-exponential distribution unless the values of a were small.

We shall consider estimators based on a finite sample $X_1,\dots,X_n$. One intuitively expects from a good estimator T_n that the inaccuracy (1.1) will tend to 0 as fast as possible i.e. that the distribution of T_n will have the least possible tails. The tails of an estimator cannot be made arbitrarily small; for instance, if the sample comes from the Cauchy distribution one cannot find an estimator with exponentially decreasing tails.

We shall prove that the tails of a translation-equivariant estimator could decrease to 0 at most n times faster than the tails of the basic distribution and that, on the other hand, there are estimators which behave from this point of view in the same way as one single observation (Theorem 2.1). Moreover, we shall show that both extreme cases may happen for the sample mean $\bar{X}_n$; $\bar{X}_n$ attains the upper bound if the basic distribution has exponentially decreasing tails of the type $\exp[-ba^r]$, $b > 0$, $r \geq 1$ and $\bar{X}_n$ attains only the lower bound if the basic distribution is heavy-tailed with the tails of the type $b.a^{-m}$, $b > 0$, $m > 0$.

Estimating the centre of symmetry of an unknown symmetric distribution, we want to find an estimator which has small tails for as large family of distributions as possible. An exponentially-tailed distribution, being contaminated by a heavy-tailed distribution, becomes heavy-tailed; it implies that the sample mean $\bar{X}_n$ is not too good for such families of distributions. On the other hand, $\bar{X}_n$ remains good for such cases as a mixture of two normal distributions, for the normal distributions contaminated by the double-exponential distribution, etc.

If we trim-off some extreme observations, then the rate of convergence of the tails of any resulting L-estimator attains neither the upper nor the lower bound (Theorem 3.1). The situation is similar for the estimators based on the ranks, e.g. for Hodges-Lehmann's estimator (Theorem 3.3). The tails of the sample median decrease exactly $\frac{n+1}{2}$ times faster (for n odd) than the tails of the basic distribution, for both exponentially-tailed as well as for heavy-tailed distributions. The same holds for the Huber M-estimator generated by a bounded monotone odd function ψ (Theorem 3.2).

2. BEHAVIOUR OF THE SAMPLE MEAN

Let us consider the model satisfying the following assumption:

ASSUMPTION A. $X_1,\ldots,X_n$ are random variables identically distributed according to the distribution function $F(x-\theta)$ with the density $f(x-\theta)$ such that $f(-x) = f(x) > 0$, $x \in R^1$; $\theta \in R^1$ is the parameter to be estimated.

All estimators we consider are translation-equivariant, i.e. they satisfy the condition

$$T_n(X_1+c,\ldots,X_n+c) = T_n(X_1,\ldots,X_n) + c \tag{2.1}$$

for any $c \in R^1$. If T_n is translation-equivariant then $P_\theta(|T_n-\theta| > a) = P_0(|T_n| > a)$ for any $\theta \in R^1$ and the inaccuracy (1.1) does not depend on θ.

The following theorem gives upper and lower bounds for the rate of convergence of the tails of a translation-equivariant estimator.

THEOREM 2.1. *Let $T_n = T_n(X_1,\dots,X_n)$ be a translation-equivariant estimator of θ such that*

$$\text{(2.2)} \qquad \begin{aligned} X^{(1)} > 0 &\Rightarrow T_n(X_1,\dots,X_n) > 0 \\ X^{(n)} < 0 &\Rightarrow T_n(X_1,\dots,X_n) < 0 \end{aligned}$$

where $X^{(1)} \le X^{(2)} \le \dots \le X^{(n)}$ are the order statistics corresponding to $X_1,\dots,X_n$. Then, under the Assumption A, it holds

$$\text{(2.3)} \qquad 1 \le \varliminf_{a\to\infty} B(a,T_n) \le \varlimsup_{a\to\infty} B(a,T_n) \le n$$

where

$$\text{(2.4)} \qquad B(a;T_n) = \frac{-\ln P_0(|T_n|>a)}{-\ln P_0(|X_1|>a)}$$

and P_0 is the probability distribution corresponding to F.

PROOF. We have

$$\begin{aligned} P_0(|T_n| > a) &= P_0(T_n > a) + P_0(T_n < -a) \\ &= P_0(T_n(X_1-a,\dots,X_n-a) > 0) + P_0(T_n(X_1+a,\dots,X_n+a) < 0) \\ &\ge P_0(X^{(1)} > a) + P_0(X^{(n)} < -a) = 2^{-n+1}[P_0(|X_1| > a)]^n \end{aligned}$$

which implies the second inequality in (2.3). Similarly,

$$\begin{aligned} P_0(|T_n| > a) &\le P_0(X^{(n)} \ge a) + P_0(X^{(1)} \le -a) \\ &= 2\{1 - [1 - \tfrac{1}{2} P_0(|X_1| > a)]^n\} \end{aligned}$$

and this implies the first inequality in (2.3). □

In the subsequent text, we shall investigate which estimators attains the upper bound in (2.3), which estimators are so poor that they attain the lower bound only and generally, what is the position of some well-known estimators from this point of view. We shall first consider the sample mean $\bar{X}_n$. The next theorem shows that the $\bar{X}_n$ attains the upper bound if the basic distribution has exponentially decreasing tails while it is a poor estimator for a heavy-tailed basic distribution.

THEOREM 2.2. *Let* $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ *be the sample mean, let* $X_1,\dots,X_n$ *satisfy the Assumption A.*

(i) *If*

$$(2.5)\qquad \lim_{a\to\infty} \frac{-\ln(1-F(x))}{b.a^r} = 1 \qquad \textit{for some } b > 0,\ r \geq 1$$

then

$$(2.6)\qquad \lim_{a\to\infty} B(a;\bar{X}_n) = n.$$

(ii) *If*

$$(2.7)\qquad \lim_{a\to\infty} \frac{-\ln(1-F(a))}{m.\ln a} = 1, \qquad m > 0$$

then

$$(2.8)\qquad \lim_{a\to\infty} B(a;\bar{X}_n) = 1.$$

PROOF. The part (i) was proved by the author in [5]. Considering the part (ii), we have

$$\begin{aligned} P_0(|\bar{X}_n| > a) &= P_0(\bar{X}_n > a) + P(\bar{X}_n < -a) \\ &\geq P_0(X_1>-a,\dots,X_{n-1}>-a,\ X_n > (2n-1)a) + \\ &+ P_0(X_1<a,\dots,X_n<a,\ X_n < -(2n-1)a) = \\ &= 2(F(a))^{n-1}[1-F((2n-1)a)] \end{aligned}$$

so that

$$\overline{\lim_{a\to\infty}}\, B(a;\bar{X}_n) \le \overline{\lim_{a\to\infty}}\, \frac{-\ln[1-F((2n-1)a)]}{m.\ln[(2n-1)a]} = 1. \qquad \square$$

The part (i) concerns not only the normal ($r = 2$) but also the logistic and double-exponential distributions ($r = 1$); the part (ii) covers Cauchy distributions ($m = 1$) and t-distribution with m degrees of freedom (m integer $m > 1$). Theorem 2.2 says that $\bar{X}_n$ is a good estimator for the case (i) while it is a poor estimator for the case (ii). Now, what is the situation of $\bar{X}_n$ if F is a mixture of two distributions, one from each group?

The following lemma shows that if a distribution is contaminated by a heavy-tailed distribution then the resulting distribution is heavy-tailed. It means that the sample mean $\bar{X}_n$ is a poor estimator in such case.

<u>LEMMA 2.1</u>. *Let* $F(x) = (1-\varepsilon)G(x) + \varepsilon H(x)$ where G and H are absolutely continuous distribution functions with the respective densities g and h such that $g(-x) = g(x) > 0$, $h(-x) = h(x) > 0$, $x \in R^1$; $0 \le \varepsilon < 1$. If it holds

$$\lim_{x\to\infty} \frac{1-G(x)}{1-H(x)} = 0 \tag{2.9}$$

and

$$\lim_{x\to\infty} \frac{g(x)}{h(x)} = 0 \tag{2.10}$$

then

$$\lim_{x\to\infty} \frac{\ln(1-F(x))}{\ln(1-H(x))} = 1. \tag{2.11}$$

<u>PROOF</u>.

$$\lim_{x\to\infty} \frac{\ln(1-F(x))}{\ln(1-H(x))} = \lim_{x\to\infty} \frac{(1-H(x))f(x)}{(1-F(x))h(x)} =$$

$$= \lim_{x\to\infty} \frac{(1-\varepsilon)\frac{g(x)}{h(x)} + \varepsilon}{(1-\varepsilon)\frac{1-G(x)}{1-H(x)} + \varepsilon} = 1. \qquad \square$$

3. BEHAVIOUR OF SOME ROBUST ESTIMATORS

If it is possible that the distribution of $X_1,\dots,X_n$ is contaminated by a heavy-tailed distribution we must look for some more robust estimators of location. Let us consider what is the position of three basic types of robust estimators; L-estimators, M-estimators and R-estimators.

We shall show that the rate of convergence of the tails of such estimators is more or less bounded away from the lower as well as from the upper bound in (2.3). It means that the estimators are not optimal but, on the other hand, they may not be very poor.

3.1. L-estimators

THEOREM 3.1. *Let* T_n *be an* L-*estimator of the form*

$$T_n = \sum_{i=1}^{n} c_i X^{(i)} \tag{3.1}$$

where $X^{(1)} \le \dots \le X^{(n)}$ *are the order statistics corresponding to* $X_1,\dots,X_n$ *and* $c_i \ge 0$, $i = 1,\dots,n$ *and* $\Sigma_{i=1}^{n} c_i = 1$. *Put* $c_0 = c_{n+1} = 0$ *and assume that* $c_i = c_{n-i+1} = 0$ *for* $i = 0,1,\dots,k$ *where* $0 \le k < \frac{n}{2}$. *Then, under the assumption A, it holds*

$$k+1 \le \varliminf_{a\to\infty} B(a;T_n) \le \varlimsup_{a\to\infty} B(a;T_n) \le n-k. \tag{3.2}$$

PROOF. The theorem was proved by the author in [5].

COROLLARY. *Let* T_n *be the sample median corresponding to* $X_1,\dots,X_n$. *Then, under the Assumption A, it holds*

$$\frac{n}{2} \le \varliminf_{a\to\infty} B(a;T_n) \le \varlimsup_{a\to\infty} B(a;T_n) \le \frac{n}{2} + 1 \tag{3.3}$$

for n *even, and*

$$\lim_{a\to\infty} B(a;T_n) = \frac{n+1}{2} \tag{3.4}$$

for n *odd.*

3.2 M-estimators

M-estimator T_n is defined as any solution of the equation

$$\sum_{i=1}^{n} \psi(X_i - t) = 0 \tag{3.5}$$

with respect to t; ψ is an appropriate non-decreasing odd function. We shall show that T_n behaves similarly as the sample median, at least for the distributions with exponentially decreasing and slowly decreasing tails.

THEOREM 3.2. *Let T_n be an M-estimator corresponding to the nondecreasing odd function ψ such that $\psi(x) = \psi(k)$ for $x \geq k$, $k > 0$. Suppose that the common distribution of $X_1,\dots,X_n$ satisfies the Assumption A and either of the following conditions:*

$$\lim_{a\to\infty} \frac{-\ln P_0(|X_1|>a)}{b.a^r} = 1; \qquad b > 0,\ r > 0 \tag{3.6}$$

$$\lim_{a\to\infty} \frac{-\ln P_0(|X_1|>a)}{m.\ln a} = 1, \qquad m > 0. \tag{3.7}$$

Then T_n satisfies (3.3) *and* (3.4).

PROOF.
(a) Suppose that n is even and denote $s = \frac{n}{2}$. Then

$$\begin{aligned} P_0(|T_n| > a) &= P_0(T_n > a) + P_0(T_n < -a) \\ &\geq P_0\left(\sum_{i=1}^{n} \psi(X_i - a) > 0\right) + P_0\left(\sum_{i=1}^{n} \psi(X_i + a) < 0\right) \\ &\geq P_0(X^{(s)} - a > k) + P_0(X^{(s+1)} + a < -k) \\ &\geq 2\binom{n}{s-1}(F(a+k))^{s-1}(1-F(a+k))^{s+1}; \end{aligned}$$

thus

$$\overline{\lim_{a\to\infty}}\, B(a;T_n) \leq (s+1)\ \overline{\lim_{a\to\infty}} \frac{\ln(1-F(a+k))}{\ln(1-F(a))} = s+1.$$

Analogously,

$$P_0(|T_n| > a) \le P_0(X^{(s+1)} \ge a-k) + P_0(X^{(s)} \le -a+k)$$

$$= 2n\binom{n-1}{s} \int_{F(a-k)}^{1} t^s(1-t)^{s-1}dt \le 4\binom{n-1}{s}(1-F(a-k))^s$$

so that

$$\underline{\lim}_{a\to\infty} B(a;T_n) \ge s \ \underline{\lim}_{a\to\infty} \frac{(1-F(a-k))}{(1-F(a))} = s.$$

(b) The proof for n odd is analogous. □

3.3. R-estimators

We shall consider in details only the Hodges-Lehmann's estimator which has the form

$$T_n = \underset{1\le i\le j\le n}{\text{med}} \frac{X_i+X_j}{2}. \tag{3.8}$$

Other R-estimators could be investigated by the same method but it provides only the numerical values of the lower and upper bounds for the rate of convergence of the tails; we do not yet have an analytical formula expressing the bounds through the score-generating function of the underlying signed-rank test.

THEOREM 3.3. *Let* T_n *be the Hodges-Lehmann's estimator* (3.8). *Then, under the Assumption A, it holds*

$$k_n + 1 \le \underline{\lim}_{a\to\infty} B(a;T_n) \le \overline{\lim}_{a\to\infty} B(a;T_n) \le n - k_n \tag{3.9}$$

where k_n *is the largest integer not exceeding* $0.2n$.

PROOF. We shall first prove a simple lemma.

LEMMA 3.1. *Let* $y_1,\dots,y_n$ *be integers satisfying* $|y_i| = i$, $i = 1,\dots,n$. *If* *at least* 0.8n *of those numbers are negative, then* $\Sigma_{i=1}^{n} y_i < 0$.

<u>PROOF OF LEMMA 3.1</u>. If 0.8n is an integer, then

$$\sum_{i=1}^{n} y_i \le -\sum_{i=1}^{0.8n} i + \sum_{i=0.8n+1}^{n} i = -0.14n^2 - 0.3n < 0;$$

$\Sigma_{i=1}^{n} y_i$ is still less in the case that 0.8n is not an integer. □

<u>PROOF OF THEOREM 3.3</u>. For any $t \in R^1$, let $R^+(|X_i-t|)$ be the rank of $|X_i-t|$ among $|X_1-t|,\ldots,|X_n-t|$. T_n is an inversion of Wilcoxon signed rank test, i.e. $T_n = \frac{1}{2}(T_n^* + T_n^{**})$ where

$$\begin{aligned} T_n^* &= \sup\{t: \sum_{i=1}^{n} \operatorname{sign}(X_i-t)R^+(|X_i-t|) > 0\} \\ T_n^{**} &= \inf\{t: \sum_{i=1}^{n} \operatorname{sign}(X_i-t)R^+(X_i-t) < 0\}. \end{aligned} \tag{3.10}$$

Then

$$\begin{aligned} P_0(|T_n| > a) &\le 2P_0\{\sum_{i=1}^{n} \operatorname{sign}(X_i-a)R^+(|X_i-a|) \ge 0\} \\ &\le 2P_0(X^{(n-k_n)} \ge a) \le 2\binom{n-1}{k_n}\frac{(1-F(a))^{k_n+1}}{k_n+1}, \end{aligned}$$

thus

$$\underline{\lim}_{a\to\infty} B(a;T_n) \ge k_n + 1.$$

Similarly,

$$\begin{aligned} P_0(|T_n| > a) &\ge 2P_0(\sum_{i=1}^{n} \operatorname{sign}(X_i-a)R^+(|X_i-a|) > 0) \\ &\ge 2P_0(X^{(k_n+1)} > a) \ge 2\binom{n}{k_n}(F(a))^{k_n}(1-F(a))^{n-k_n} \end{aligned}$$

so that

$$\overline{\lim}_{a\to\infty} B(a;T_n) \le n - k_n. \qquad \square$$

REFERENCES

[1] BAHADUR, R.R. (1960), *Asymptotic efficiency of tests and estimates,* Sankhyā 22, 229-252.

[2] BAHADUR, R.R. (1967), *Rates of convergence of estimates and test statistics,* Ann. Math. Statist. 38, 303-324.

[3] BAHADUR, R.R. (1971), *Some Limit Theorems in Statistics,* SIAM, Philadelphia.

[4] FU, J.C. (1975), *The rate of convergence of consistent point estimators,* Ann. Statist. 3, 234-240.

[5] JUREČKOVÁ, J. (1979), *Finite-sample comparison of L-estimators of location,* Commentationes mathematicae Universitatis Carolinae 20, 509-518.

[6] SIEVERS, G.L. (1978), *Estimation of location; A large deviation comparison,* Ann. Statist. 6, 610-618.