

RYSZARD ZIELIŃSKI (Warszawa)

## THE MOST STABLE ESTIMATOR OF LOCATION UNDER INTEGRABLE CONTAMINANTS

*Abstract.* If a symmetric distribution is  $\varepsilon$ -contaminated and the contaminants have finite first moments, the median may cease to be the most robust estimator of location.

**1. Statement of the problem.** The problem is to estimate the location  $\theta \in \mathbb{R}^1$  of the distribution  $F_\theta(x) = F(x - \theta)$ , where  $F$  is assumed to be a symmetric (around zero), unimodal (mode = 0), continuous and strictly increasing distribution function; here  $F = F_0$ . We denote by  $f$  the probability distribution function of  $F$ .

Suppose that the observations are  $\varepsilon$ -contaminated and their distribution is  $G_\theta(x) = G(x - \theta)$  such that  $G = (1 - \varepsilon)F + \varepsilon H$ ,  $H \in \mathcal{H}$ , where  $\mathcal{H}$  is a class of distributions and  $\varepsilon \in (0, 1/2)$  is a constant.

We consider as estimators the statistics  $T_n = T(G_n)$  derived from a functional  $T \in \mathcal{T}$ , where  $\mathcal{T}$  is the class of all translation invariant functionals on the space of all distribution functions; here  $G_n$  is the empirical distribution function.

We are interested in finding such a  $T$  which minimizes the maximum asymptotic oscillation of the bias  $B_\varepsilon(T) = \sup |T(G_1) - T(G_2)|$ , where the supremum is taken over all  $G_i = (1 - \varepsilon)F + \varepsilon H_i$ ,  $H_i \in \mathcal{H}$ ,  $i = 1, 2$  (“the most stable translation invariant estimator of location under  $\varepsilon$ -contamination”).

The median, trimmed means, and suitable  $L$ -,  $M$ -, and  $R$ -estimators as robust alternatives to the mean for estimating location in that model have a long history. If  $\mathcal{H}$  is the class of all distributions, the well known optimal

---

2000 *Mathematics Subject Classification*: 62F35, 62F10, 62F12.

*Key words and phrases*: robust estimation, location,  $\varepsilon$ -contamination, integrable contaminants.

solution (Huber 1981) is the sample median  $T_{0.5}$  with  $B_\varepsilon(T_{0.5}) = 2C_0$  where

$$C_0 = F^{-1}\left(\frac{1}{2(1-\varepsilon)}\right).$$

The distributions  $H_1$  and  $H_2$  for which  $\sup |T_{0.5}(G_1) - T_{0.5}(G_2)|$  is attained are those with supports in  $(-\infty, -C_0)$  and  $(C_0, +\infty)$ , respectively; here  $T_q$  is a translation invariant estimator of the  $q$ th quantile (the quantile of order  $q$ ) such that  $T_q(G) = G^{-1}(q)$  for all distribution functions  $G$ . The commonly accepted conclusion is: *the sample median is the most robust estimator of location if contaminants may spoil the sample* (see e.g. Borovkov 1998, Brown 1985, Shervish 1995). An optimal solution without the assumption of symmetry is given in Rychlik and Zieliński (1987).

It appears that if  $\mathcal{H}$  is a smaller class of distributions, then the optimal solution may be quite different (an example is given in Zieliński 1987). Below we consider the case of a class of distributions with finite first moments.

**2. The estimator.** We assume that  $F$  has a finite moment. For a given  $H$ , if  $0 < H(0) < 1$ , define

$$H^+(x) = \begin{cases} \frac{H(x) - H(0)}{1 - H(0)} & \text{if } x \geq 0, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$H^-(x) = \begin{cases} \frac{H(x)}{H(0)} & \text{if } x \leq 0, \\ 1 & \text{otherwise.} \end{cases}$$

If  $H(0) = 0$  then define  $H^-(x) = 0$  for  $x \leq 0$  and if  $H(0) = 1$  then define  $H^+(x) = 1$  for  $x \geq 0$ .

From the well known inequality for a positive random variable  $\xi$  with finite expectation  $E\xi$ ,

$$P\{\xi \geq t\} \leq \frac{E\xi}{t}, \quad t > 0,$$

we deduce that for  $H$  with a finite expectation,

$$(H) \quad \begin{aligned} H^+(x) &\geq 1 - C/x, & x > 0, \\ H^-(x) &\leq -C/x, & x < 0, \end{aligned}$$

with a finite  $C > 0$ . In what follows we assume that  $C > C_0$ . Note that if a contaminant  $\xi$  satisfies  $E\xi^+ < C$  and  $E\xi^- < C$  then the distribution of  $\xi$  satisfies (H).

Let

$$(1) \quad \begin{aligned} L(x) &= (1 - \varepsilon)F(x) + \varepsilon \begin{cases} 0 & \text{if } x \leq C, \\ 1 - C/x & \text{if } x > C, \end{cases} \\ U(x) &= (1 - \varepsilon)F(x) + \varepsilon \begin{cases} -C/x & \text{if } x \leq -C, \\ 1 & \text{if } x > -C, \end{cases} \end{aligned}$$

and define

$$\mathcal{N}(\varepsilon, C) = \{G = (1 - \varepsilon)F + \varepsilon H : L < G < U\}$$

For  $T \in \mathcal{T}$ , let

$$B_{\varepsilon, C}(T) = \sup_{G_1, G_2 \in \mathcal{N}(\varepsilon, C)} |T(G_1) - T(G_2)|.$$

For  $q \in (0, 1)$  define

$$\delta(q) = L^{-1}(q) - U^{-1}(q).$$

Suppose that there exists  $q^* \in (0, 1)$  such that

$$\delta(q^*) \leq \delta(q), \quad q \in (0, 1).$$

Let  $\Delta(x) = \delta(L(x))$ ,  $-\infty < x < \infty$ , and define  $\Delta^* = \frac{1}{2} \Delta(L^{-1}(q^*))$ . For  $q = 0.5$  we have  $\delta(q) = 2C_0$  so that  $\Delta^* \leq C_0$ .

As an estimator of location  $\theta$  we consider  $\hat{\theta}_{q^*} = T_{q^*} - F^{-1}(q^*)$ . Due to the fact that  $|\hat{\theta}_{q^*}(G_1) - \hat{\theta}_{q^*}(G_2)| = |T_{q^*}(G_1) - T_{q^*}(G_2)|$ , to demonstrate the optimality of  $\hat{\theta}_{q^*}$  it is enough to prove the following theorem.

**THEOREM.**  $B_{\varepsilon, C}(T_{q^*}) \leq B_{\varepsilon, C}(T)$  for all  $T \in \mathcal{T}$ .

*Proof.* Define

$$G^U(x) = \begin{cases} L(x + 2\Delta^*) & \text{if } x \leq -\Delta^*, \\ U(x) & \text{if } x > -\Delta^*. \end{cases}$$

By (1),

$$G^U(x) = \begin{cases} (1 - \varepsilon)F(x) + \varepsilon \cdot \frac{1 - \varepsilon}{\varepsilon} [F(x + 2\Delta^*) - F(x)] & \text{if } x \leq -\Delta^*, \\ (1 - \varepsilon)F(x) + \varepsilon & \text{if } x > -\Delta^*. \end{cases}$$

The function

$$H_U^0(x) = \frac{1 - \varepsilon}{\varepsilon} [F(x + 2\Delta^*) - F(x)], \quad x \leq -\Delta^*,$$

has the following properties:

- 1)  $H_U^0(x) \geq 0$ ;
- 2) by symmetry and unimodality,  $f(x + 2\Delta^*) - f(x) > 0$  for  $x \leq -\Delta^*$ , so that  $H_U^0(x)$  is increasing;

$$3) H_U^0(-\Delta^*) = \frac{1 - \varepsilon}{\varepsilon} [2F(\Delta^*) - 1] \leq \frac{1 - \varepsilon}{\varepsilon} [2F(C_0) - 1] = 1.$$

It follows that

$$H_U(x) = \begin{cases} H_U^0(x) & \text{if } x \leq -\Delta^*, \\ 1 & \text{if } x > -\Delta^*, \end{cases}$$

is a distribution function and in consequence  $G^U(x)$  is a distribution function of the form  $(1 - \varepsilon)F(x) + \varepsilon H_U(x)$  and belongs to  $\mathcal{N}(\varepsilon, C)$ .

Define

$$G^L(x) = \begin{cases} L(x) & \text{if } x \leq \Delta^*, \\ U(x - 2\Delta^*) & \text{if } x > \Delta^*. \end{cases}$$

By similar arguments to those concerning  $G^U(x)$  we conclude that  $G^L \in \mathcal{N}(\varepsilon, C)$ . It is easy to check that  $G^U(x) = G^L(x + 2\Delta^*)$  so that for  $T \in \mathcal{T}$  we have  $T(G^U) = T(G^L) + 2\Delta^*$  and in consequence  $B_{\varepsilon, C}(T) \geq 2\Delta^*$  for all  $T \in \mathcal{T}$ .

For  $G \in \mathcal{N}(\varepsilon, C)$  we have

$$T_{q^*}(U) \leq T_{q^*}(G) \leq T_{q^*}(L).$$

By the very definition of  $q^*$  we have  $T_{q^*}(L) - T_{q^*}(U) = 2\Delta^*$  so that  $B_{\varepsilon, C}(T_{q^*}) \leq 2\Delta^*$ . ■

If  $x \in (-C, C)$  then  $L(x) = (1 - \varepsilon)F(x)$  and  $U(x) = (1 - \varepsilon)F(x) + \varepsilon$ , so that

$$\begin{aligned} \min_{U(-C) \leq q \leq L(C)} \delta(q) &= \min_{U(-C) \leq q \leq L(C)} \left[ F^{-1}\left(\frac{q}{1 - \varepsilon}\right) - F^{-1}\left(\frac{q - \varepsilon}{1 - \varepsilon}\right) \right] \\ &= 2F^{-1}\left(\frac{1}{2(1 - \varepsilon)}\right) = 2C_0 \end{aligned}$$

for  $q = 1/2$ . It follows that without the moment condition, i.e. for  $C = +\infty$ , we have  $q^* = 1/2$ ; then the best estimator is the median  $T_{0.5}$ .

If  $C < +\infty$  then, given  $F$  and  $\varepsilon$ , it may happen that  $\Delta(x)$  has some other minima in  $\{x : x - \Delta(x) < -C\}$  or  $\{x : x > C\}$ , and the minima are smaller than  $\Delta(C_0)$  for  $q^* = 1/2$ . These minima give us more stable estimators. No general results for any class of  $F$  are known; a numerical study for the Gaussian case is presented in the next section.

**2. The Gaussian case: a numerical study.** The  $\varepsilon$ -contamination vicinity with “ $C$ -restriction on the first moment” for  $F = N(0, 1)$ ,  $\varepsilon = 0.2$ , and  $C = 0.7$  is exhibited in Fig. 1. Now  $C_0 = 0.3186$ , so that  $B_{0.2, +\infty}(T_{0.5}) = 0.6372$ . That is the maximal oscillation of the bias of the median (the optimal estimator with no moment restrictions). We shall construct the best estimator under the above restriction on the first moment.

Due to symmetry we may confine ourselves to considering the function  $\Delta(x)$  on the interval  $(C_0, +\infty)$  and to study its minimum on the interval  $(C, +\infty)$ . For  $\varepsilon = 0.2$  and  $C = 0.7$ , the function is presented in Fig. 2.

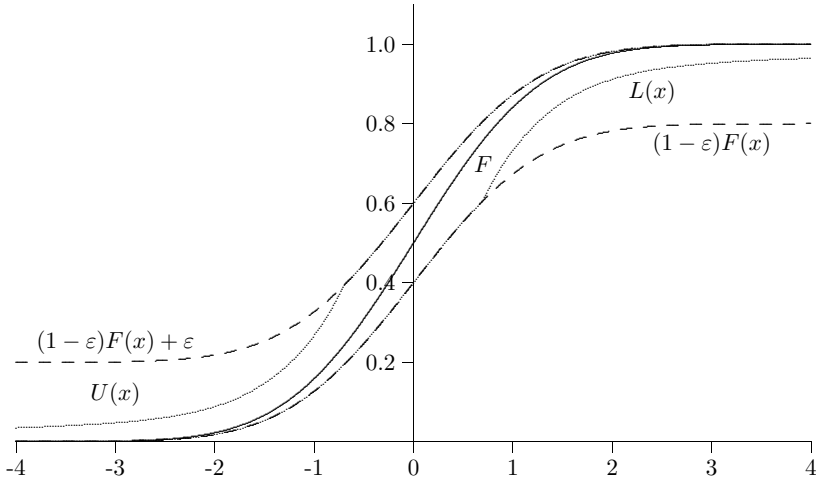


Fig. 1

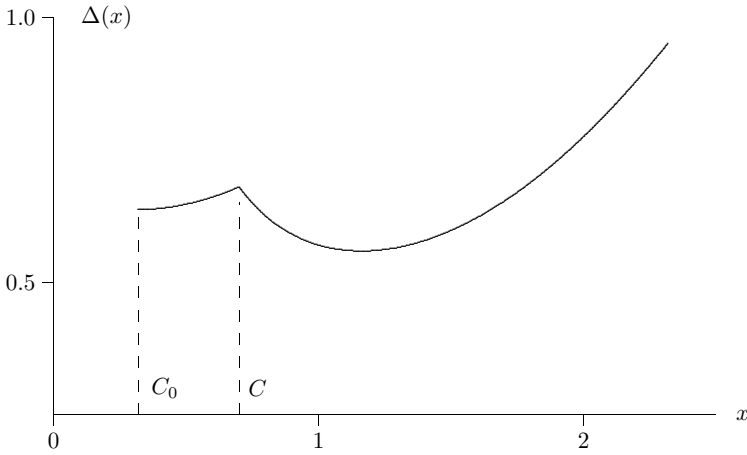


Fig. 2

Numerical calculations give us  $q^* = 0.7824$  with  $B_{0.2,0.7}(T_{0.7824}) = 0.5589$  which improves the estimator significantly.

The functions  $\Delta(x)$  for some other values of  $C$  are exhibited in Fig. 3. Numerical calculations lead to the conclusion: if  $C_0 \leq C < 0.8245$  then the optimal estimator is  $T_{q^*}$  with some  $q^* \neq 1/2$  and the median is not the best choice. If the expected value of the contaminant is large enough ( $C > 0.8245$ ), then the median is the most stable estimator.

*A comment.* T. Rychlik (2001) observed that also  $(\hat{\theta}_{q^*} + \hat{\theta}_{1-q^*})/2$  is an optimal estimator; the estimator does not depend on the constant  $F^{-1}(q^*)$  and in consequence it may be applied in our model with an unknown scale parameter.

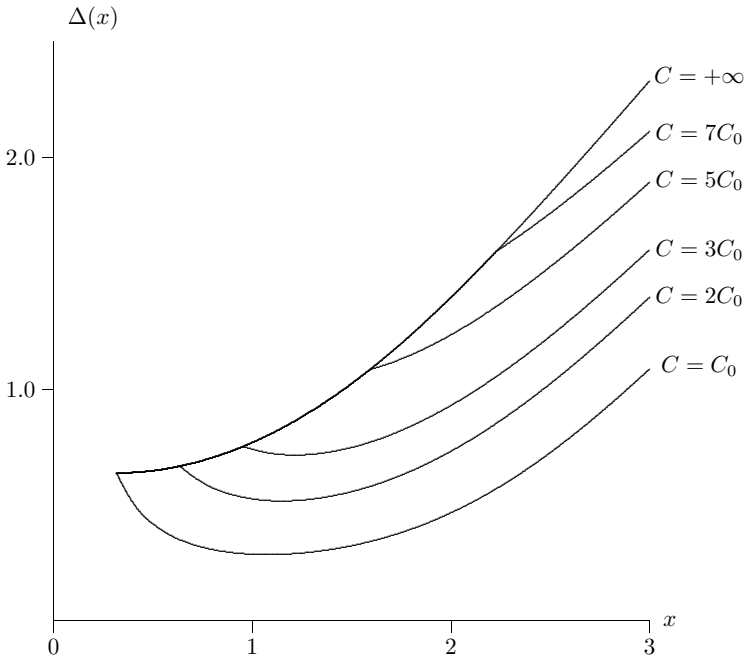


Fig. 3

**Acknowledgements.** The author thanks Tomasz Rychlik for useful discussions. The research was supported by Grant KBN 2 P03A 033 17.

### References

- A. A. Borovkov (1998), *Mathematical Statistics*, Gordon and Breach.  
 B. M. Brown (1985), *Median estimates and sign tests*, in: Encyclopedia of Statistical Sciences, Vol. 5, Wiley.  
 P. J. Huber (1981), *Robust Statistics*, Wiley.  
 T. Rychlik (2001), private communication.  
 T. Rychlik and R. Zieliński (1987), *An asymptotically most bias-robust invariant estimator of location*, in: Stability Problems for Stochastic Models, Lecture Notes in Math. 1233, V. V. Kalashnikov *et al.* (eds.), Springer, 156–171.  
 M. J. Shervish (1995), *Theory of Statistics*, Springer.  
 R. Zieliński (1987), *Robustness of sample mean and sample median under restrictions on outliers*, Appl. Math. (Warsaw) 19, 239–240.

Institute of Mathematics  
 Polish Academy of Sciences  
 P.O. Box 137  
 00-950 Warszawa, Poland  
 E-mail: rziel@impan.gov.pl