

Maxbias Curve and Robust Identification Property

Curve di Massimo Errore e Proprietà di Identificazione Robusta

Claudio Agostinelli

Dipartimento di Statistica

Università Ca' Foscari di Venezia

30125 Venezia, Italia

claudio@unive.it

Riassunto: Le curve di errore massimo (*maxbias curve*, Huber, 1964) sono uno degli strumenti impiegati per lo studio della robustezza di uno stimatore. Esse consentono di valutare la massima distorsione (asintotica) di uno stimatore in presenza di contaminazione. Recentemente Agostinelli (2004) ha studiato le proprietà di identificazione di una famiglia parametrica quando è ammessa una contaminazione definendo il concetto di Proprietà di Identificazione Robusta (*Robust Identification Property*). In questo lavoro presentiamo alcune relazioni tra i due concetti per la famiglia parametrica normale di posizione.

Keywords: Maxbias curve, Robust identification property

1. Introduction

Huber (1964) introduced the *contamination “neighborhood”* of a distribution function F , given by

$$N_\varepsilon = \{(1 - \varepsilon)F + \varepsilon H\}$$

where H is any distribution and $0 \leq \varepsilon < 0.5$. For a univariate location estimator T , Huber defined its *maxbias* (see for a summary account on the topics Rousseeuw, 1998) as

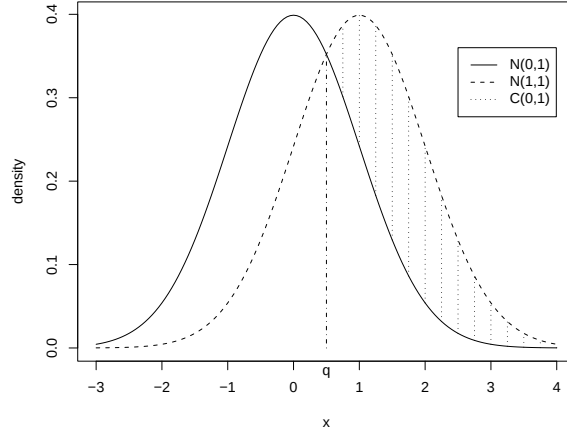
$$\text{maxbias}(\varepsilon; T, F) = \sup_{G \in N_\varepsilon} |T(G) - T(F)| \quad .$$

The maxbias curve is a plot of the $\text{maxbias}(\varepsilon; T, F)$ for a given T and F versus ε . The maxbias curve combines several well known robustness properties. If the curve is continuous at ε , then T is robust in the qualitative sense at F . When the maxbias is differentiable, often its derivative at $\varepsilon = 0$ equals the *gross error sensitivity* (Hampel *et al.*, 1986). Moreover, the ε at which the maxbias curve becomes infinite is the asymptotic *breakdown point*

$$\varepsilon^*(T, F) = \inf\{\varepsilon : \text{maxbias}(\varepsilon; T, F) = \infty\}$$

Agostinelli (2004) introduced the *Robust Identification Property*. Let $\mathcal{M} = \{m(x, \theta); \theta \in \Theta\}$ be a parametric family where $m(x; \theta)$ is the density with respect to some dominating measure ν and Θ is the parametric space. Two distinct elements ($\theta_a \neq \theta_b$) are said *identified* (using the classical *identification property*) if there exists a measurable set A such that $\int_A m(x; \theta_a) d\nu(x) \neq \int_A m(x; \theta_b) d\nu(x)$. A parametric family is said identified if every two distinct elements of the parametric family are identified.

Figure 1: The normal $\mu_a = 0$ and $\mu_b = 1$. In dashed $C(0, 1)$.



The Robust Identification Property characterizes the identification property of a parametric family when an outliers contamination is allowed. To fix the idea, let us consider the following simple example reported in figure 1. A normal location parametric family (with scale $\sigma = 1$) is considered and the observations are from a *gross error model* of the form $F(x) = (1 - \varepsilon)N(x; \mu_a) + \varepsilon H(x)$ (with $H(x)$ any distribution function).

In order to make the distribution $m(x; \mu_a = 0)$ not identified from $m(x; \mu_b = 1)$ we need to check if the area ($C(\mu_a = 0, \mu_b = 1)$) of the dashed part is less than $\frac{\varepsilon}{1-\varepsilon}$. In fact, when this condition holds, it is sufficient to choose the outliers contamination distribution H as a normalized version of the shaded area and we will not be able (even asymptotically) to distinguish if the majority of the observations are either from $m(x; \mu_a = 0)$ or $m(x; \mu_b = 1)$. In this example $C(0, 1) = 0.383$ then we need to have an outliers contamination level not less than 0.277 in order to find a H that make the two values not identified.

In the general case let the normalized version of the dashed area, namely the *worst (outliers) contamination density function* be:

$$w_o(x; \theta_a, \theta_b) = \frac{[m(x; \theta_b) - m(x; \theta_a)]^+}{C(\theta_a, \theta_b)}$$

$$C(\theta_a, \theta_b) = \int_{\Omega} [m(x; \theta_b) - m(x; \theta_a)]^+ d\nu(x)$$

and $W_o(x; \theta_a, \theta_b) = \int_{-\infty}^x w_o(t; \theta_a, \theta_b) d\nu(t)$ the corresponding *worst contamination distribution function*. Then our definition of the *Robust Identification Property* is

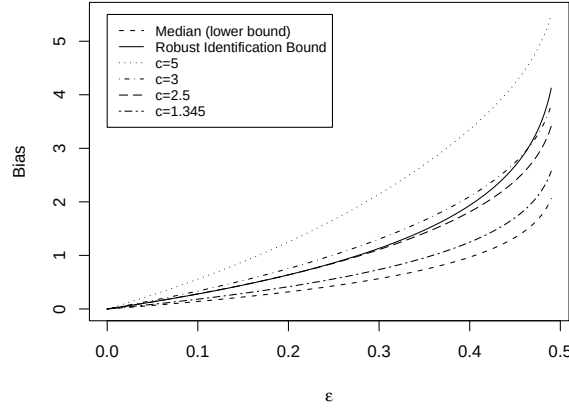
Definition: Robust Identification property (against outlier contamination)

A parametric family $\mathcal{M} = \{m(x; \theta) : \theta \in \Theta\}$ is *robust identified* for a level contamination $0 \leq \varepsilon^* < 0.5$ if for every value $\theta_a \in \Theta$ and for every $0 \leq \varepsilon \leq \varepsilon^*$ the set $\Theta_o(\theta_a) = \{\theta_b : C(\theta_a, \theta_b) \leq \varepsilon/(1 - \varepsilon)\}$ is bounded and such that $\Theta_o(\theta_a) \subsetneq \Theta$.

2. Relations between maxbias and robust identification property

In this section we study the relation among the maxbias curve of the median estimator T_M (which attains the lower bound between equivariant estimators, see He and Simp-

Figure 2: The maxbias curve of several Huber M-estimator with different c constant, the lower bound B and the bound of the robust identification property B_r .



son, 1993), Huber M-estimators T_ψ and the robust identification property in the normal location parametric family. Let ϕ and Φ be the density and the distribution function of the standard normal distribution. The maxbias of the M-estimators (Martin and Zamar, 1993) with known scale and ψ function continuous, monotone, odd and bounded is

$$\text{maxbias}(\varepsilon; T_\psi, F) = g^{-1} \left(\psi(\infty) \frac{\varepsilon}{1 - \varepsilon} \right)$$

where $g(t) = \int_0^\infty \psi(y) (\phi(y - t) - \phi(y + t)) d\nu(y)$ which specialized to (Huber, 1964)

$$B = \text{maxbias}(\varepsilon; T_M, F) = \Phi^{-1} \left(\frac{\varepsilon}{2(1 - \varepsilon)} \right)$$

It is easy to check that (for our case of the a normal location parametric family with scale one)

$$\Theta_o(\mu_a) = \{\mu_a - 2q, \mu_a + 2q\}$$

where $B_r = 2q = 2 \text{maxbias}(\varepsilon; T_M, F)$. Hence the length of the maxbias suggested by the robust identification property (B_r) is twice the lower bound (B). Figure 2 illustrated the maxbias curve for different estimators and B_r .

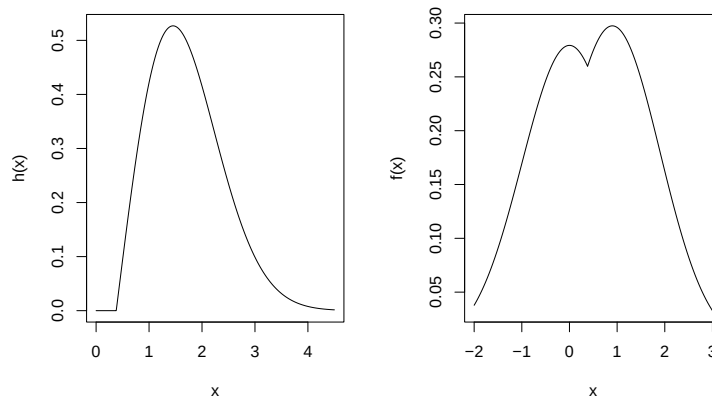
For Huber M-estimators with tuning constant c big enough (for example $c = 5$) we have a maxbias greater than B_r . Hence, for these estimators there exists a contamination distribution such that the value is outside B_r even if the majority of the observations are still from the central model. So, the robust identification property could be used to find a suitable upper bound for an estimator to be considered robust.

But, what happens to the estimators with maxbias curve between B and B_r ? Let us consider this example, where $F(x) = 0.7N(x; 0, 1) + 0.3 H(x)$ and choose $H(x)$ such that its density $h(x)$ is the plot in figure 3 (left) which leads to an $F(x)$ with density the plot in figure 3 (right).

It is easy to check that in this case $F(x)$ can be rewritten as

$$F(x) = 0.746N(x; 0.9, 1) + 0.254H'(x) \quad (1)$$

Figure 3: Distributions for the example, left: density $h(x)$ and right: $f(x)$.



for some $H'(x)$ which means that there exists inside the parametric family a model that better describes the majority of the observations (infact the level of contamination decreases from 0.3 to 0.254, which is a save of about 5%).

Since formula 1 a good robust location estimator of the majority of the observations (when the model is normal) would be very close to 0.9, but there no exists a contamination distribution that leads to $T = 0.9$ for every M-estimators with $c < 1.911$.

References

- Agostinelli C. (2004) Robust identification property, redazione provvisoria 2004.1, Dipartimento di Statistica, Università Ca' Foscari di Venezia.
- Hampel F., Ronchetti E., Rousseeuw P. and Stahel W. (1986) *Robust Statistics: The approach based on Influence Functions*, Wiley.
- He X. and Simpson D. (1993) Lower bounds for contamination bias: Globally minimax versus locally linear estimation, *Annals of Statistics*, 21, 1, 314–337.
- Huber P. (1964) Robust estimation of a location parameter, *Annals of Mathematical Statistics*, 35, 1, 73–101.
- Martin R. and Zamar R. (1993) Efficiency–constrained bias–robust estimation of location, *Annals of Statistics*, 21, 1, 338–354.
- Rousseeuw P. (1998) Maxbias curve, in: *Encyclopedia of Statistical Science*, Kotz S., Read C. and Bank D., eds., Wiley, volume Update 3, 441–443.