ON APPROXIMATION OF THE Q_n-ESTIMATE OF SCALE BY HIGHLY ROBUST AND EFFICIENT M-ESTIMATES

P. O. Smirnov[1], I. S. Shirokov[2], G. L. Shevlyakov[3]
[1,2,3] *Peter the Great St. Petersburg Polytechnic University*
[3] *Institute for Problems of Mechanical Engineering, Russian Academy of Sciences*
Saint Petersburg, RUSSIA
e-mail: [3] `Georgy.Shevlyakov@phmf.spbstu.ru`

Abstract

Low-complexity and computationally fast Huber M-estimates of scale are proposed to approximate the highly robust and efficient Q_n-estimate of scale of Rousseeuw and Croux (1993). The parameters of approximation are chosen to provide high robustness and efficiency of the proposed M-estimates of scale at an arbitrary underlying data distribution. A special attention is payed to the particular cases of the Gaussian and Cauchy distributions.

1 Introduction and Problem Set Up

The problem of estimation of a scale parameter is one of most important in statistical analysis. In present, the commonly used highly robust and efficient estimate of scale is given by the Q_n-estimate [4]. This estimate is defined as the first quartile of the distance between observations: $Q_n = c\{|x_i - x_j|\}_{(k)}$, where c is a constant that provides the consistency of estimation, $k = C_h^2$, $h = [n/2] + 1$.

The Q_n-estimate is highly robust with the highest breakdown point $\varepsilon^* = 0.5$ possible and high efficiency 82% at the Gaussian. Its drawback is the high asymptotic computational complexity: generally, it requires $O(n^2)$ of computational time and memory.

On the contrary, Hubers' robust M-estimates of scale are of low-complexity having a potential for enhancing their efficiency. Thus, the main goals of our work are:

1. to construct a low-complexity, computationally fast and highly robust approximation to the Q_n-estimate,

2. to adapt this approximation to data distributions of a general shape.

In what follows, we consider the class of Hubers' M-estimates $\widehat{S}$ of scale given by the implicit estimating equation [3]

$$\sum \chi(x_i/\widehat{S}) = 0, \tag{1}$$

where $\chi(x)$ is a score function commonly even and nondecreasing for $x > 0$.

2 Approximation of the Q_n-estimate: General Case

An important tool for the statistical analysis of estimation in robustness is given by the influence function $IF(x; S, F)$, which defines a measure of the resistance of an estimate functional $S = S(F)$ at a distribution F to gross errors at a point x [2]. Further, the asymptotic variance of the estimate $\widehat{S}$ is given by

$$AV(\widehat{S}, F) = \int IF(x; S, F)^2 \, dF(x) \,.$$

The class of Huber M-estimates of scale (1) has a convenient feature: the influence function $IF(x; S, F)$ is proportional to the score function $\chi(x)$: $IF(x; S, F) \propto \chi(x)$. Thus, it is possible to construct an M-estimate with any admissible influence function, and accordingly, efficiency.

It is known that the influence function of the Q_n-estimate is given by [4]

$$IF(x; Q, F) = c \cdot \left(\frac{1}{4} - F(x + c^{-1}) + F(x - c^{-1}) \right) \Big/ \left(\int f(y + c^{-1}) f(y) dy \right). \quad (2)$$

Since the score χ in Equation (1) is defined up to an arbitrary factor, the normalization integral in the denominator of (2) can be omitted. Then, the Q_n-estimate corresponds to the M-estimate generated by the score function

$$\chi_Q(x) = \frac{c}{4} - c \cdot (F(x + c^{-1}) - F(x - c^{-1})), \quad (3)$$

hence the influence function $IF(x; \chi_Q, F)$ is identical with $IF(x; Q, F)$, ensuring the match of the derivatives of its characteristics.

Now we transform Equation (3). At first, let us make the substitution $\alpha = c^{-1}$, generally not fixing α and considering it as an estimate parameter. Then we expand the distribution function F in a Taylor series, leaving only the first three terms:

$$F(x \pm \alpha) = F(x) \pm \alpha f(x) + \frac{1}{2}\alpha^2 f'(x) \pm \frac{1}{6}\alpha^3 f''(x) + o(\alpha^3). \quad (4)$$

Combination of (3) and (4) leads to the following.

Definition 1. Let f be an analytic probability density function on $\mathbf{R}$. One-parametric family of M-estimates with score functions

$$\chi_\alpha(x) = c_\alpha - 2f(x) - \frac{1}{3}\alpha^2 f''(x), \quad (5)$$

is called f-based MQ_n-family (of f-based MQ_n-estimates). The scalar constant c_α in Equation (5) provides consistency of defined MQ_n-estimates.

3 Approximation of the Q_n-estimate: Gaussian Case

In this section we use the recent results of [5].

Consider the proposed M-estimate in the case of the Gaussian distribution density: $f(x) = \varphi(x) = 2\pi^{-1/2}\exp(-x^2/2)$. Then $\varphi''(x) = (x^2-1)\varphi(x)$, and the score function takes the form

$$\chi_\alpha(x) = c_\alpha - \frac{1}{3}(6+\alpha^2(x^2-1))\varphi(x), \qquad c_\alpha = \frac{12-\alpha^2}{12\sqrt{\pi}}. \tag{6}$$

In the important special case when $\alpha = 0$, the expression takes the following form

$$\chi_0(x) = \frac{1}{\sqrt{\pi}} - 2\varphi(x). \tag{7}$$

This score is similar to a Welsh generalized error score [1] given by

$$\chi(x) = \sqrt{\frac{d}{d+2}} - \exp\left(-\frac{x^2}{d}\right), \qquad d > 0.$$

For $d = 2$ this score yields the same M-estimate of scale as the score given by (7). The highest possible asymptotic efficiency of estimates defined by (7) is 95.9%.

In the Gaussian case, the following result holds.

Theorem 1. *The Gaussian-based MQ_n-estimates for $\alpha \in [0;\sqrt{2}]$ at the Gaussian distribution are B-robust with the bounded influence function of the form*

$$IF(x; MQ, \Phi) = \frac{2(12-\alpha^2) - 8\sqrt{\pi}(6+\alpha^2(x^2-1))\varphi(x)}{3(4-\alpha^2)}.$$

The asymptotic efficiency of the fast low-complexity MQ_n-estimate with the score function (7) is 81%, just 1% less than that of the Q_n-estimate at the Gaussian.

4 Approximation of the Q_n-estimate: Cauchy Case

Now we consider the Cauchy-based MQ_n-estimates with the score (5) derived from the heavy-tailed Cauchy distribution density

$$f(x) = \frac{1}{\pi(1+x^2)}.$$

For the sake of low-complexity, take the MQ_n-estimate with $\alpha = 0$, as other parameter values are of no interest because of worse performance. In this case, the following result holds.

Theorem 2. *The Cauchy-based MQ_n-estimate with $\alpha = 0$ and score function*

$$\chi_0(x) = \frac{1}{\pi} \cdot \frac{x^2 - 1}{x^2 + 1}$$

coincides with the MLE estimate of scale for the Cauchy distribution.

The highest possible asymptotic efficiency of this estimate of scale is 100% at the Cauchy distribution but the asymptotic relative efficiency at the Gaussian is 50%.

5 Conclusions

1. A class MQ_n of low-complexity, computationally fast and highly robust M-estimates of scale close in efficiency to the highly efficient and robust Q_n-estimate is proposed.

2. The important Gaussian and Cauchy distribution particular cases are thoroughly studied both theoretically in asymptotics and experimentally on small samples—the obtained results confirm effectiveness of the proposed approach.

3. In our talk, we plan to exhibit the theoretical and Monte Carlo results of application of the proposed approach in the parametric families of t- and exponential-power distributions.

References

[1] Genton M. G. (1998). Asymptotic Variance of M-estimators for Dependent Gaussian Random Variables. *Statistics and Probability Letters*. Vol. **38**, pp. 255-261.

[2] Hampel F. R., Ronchetti E. M., Rousseeuw P. J., Stahel W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. John Wiley.

[3] Huber P. J. (1981). *Robust Statistics*. John Wiley.

[4] Rousseeuw P. J., Croux C. (1993). Alternatives to the Median Absolute Deviation. *Journal of the American Statistical Association*. Vol. **88**, pp. 1273-1283.

[5] Smirnov P.O., Shevlyakov G.L. (2014). Fast Highly Efficient and Robust One-Step M-Estimators of Scale Based on Qn. *Computational Statistics and Data Analysis*. Vol. **78**, pp. 153-158.