ROBUST ANALOGS OF THE Q_n-ESTIMATE OF SCALE FOR THE STUDENT DISTRIBUTIONS

G.L. Shevlyakov, I.S. Shirokov, V.G. Litvinova
Peter the Great St. Petersburg Polytechnic University
St. Petersburg, RUSSIA
e-mail: gshevlyakov@yahoo.com

Abstract

Robust computationally fast Huber's MQ_n-estimates of scale are designed to approximate the highly robust and efficient Q_n-estimate of scale proposed by Rousseeuw and Croux (1993). The parameters of this approximation are tuned to provide high robustness and efficiency of these M-estimates of scale for the Student distributions—the dependencies between the values of the estimate parameter and distribution shape parameter are written out and tabulated. The comparative study of robust estimates is performed by computation of their asymptotic efficiencies and breakdown points. A special attention is payed to the particular cases of the Gaussian and Cauchy distributions.
Keywords: data science, robust analog, Student distribution

1 Introduction

Estimation of scale is one of the most important problems in statistics (Hampel *et al.*, 1986; Huber, 1981). First of all, there are two natural goals in statistics: constructing measures of distribution location and spread, although the role of scale is secondary as compared to location: generally, the problem of estimation of scale is subordinated to the problem of estimation of location. However, we may enlist a number of important reasons for the direct use of scale estimates: (i) data standardizing, (ii) detection of outliers in the data, (iii) estimation of correlation, and (iv) estimation of regression.

Here, we restrict ourselves to robust estimation of scale. In present, one of the best robust estimates of scale is given by the Q_n-estimate (Rousseeuw and Croux, 1993). This robust estimate is defined as the first quartile of the pair-wise distances between observations:

$$Q_n = c\{|x_i - x_j|\}_{(k)},$$

where the factor c provides the consistency of estimation, $k = C_h^2$, $h = [n/2] + 1$. The Q_n-estimate is robust with the breakdown point $\varepsilon^* = 0.5$ highest possible and high efficiency 82% at the Gaussian. Its drawback is the high asymptotic computational complexity: generally, it takes $O(n \log n)$ of computational time.

Much more common, Huber's robust M-estimates $\widehat{S}$ of scale are given by the implicit estimating equation (Huber, 1981)

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

where $\chi(x)$ is an estimating (score) function commonly even and nondecreasing for $x > 0$. The classical particular cases of M-estimates of scale are: the standard deviation

$s = \sqrt{n^{-1}\sum x_i^2}$ with $\chi(x) = x^2 - 1$, the mean absolute deviation $d = n^{-1}\sum |x_i|$ with $\chi(x) = |x| - 1$ and the median absolute deviation $MAD = med_i|x_i|$ with $\chi(x) = sgn(|x| - 1)$ (the parameter of location is set to zero here).

In this work, we use the approximations of the Q_n-estimate of scale by low-complexity and computationally fast robust MQ_n-estimates of scale of high efficiency (Smirnov and Shevlyakov, 2014) with the parameters tuned for the Student distributions. This family of distributions comprises distributions with relatively heavy tails with the important particular cases, such as the Cauchy and Gaussian (the limiting case) distributions.

An outline of the remainder of the paper is as follows. In Section 2, general results on the approximation of the Q_n-estimate of scale by MQ_n-estimates of scale are given. In Section 3, the particular case of the Student distributions is considered. In Section 4, some conclusions are drawn.

2 Approximation of the Q_n-estimate by MQ_n-estimates

The notion of the influence function $IF(x; S, F)$ that defines a measure of the sensitivity of an estimate functional $S = S(F)$ at a distribution F to the perturbation at a point x is one of the central in robust statistical analysis (Hampel et al., 1986). It is important that the asymptotic variance $V(\widehat{S}, F)$ of the estimate $\widehat{S}$ is expressed through the influence function

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

Moreover, in the class of Huber's M-estimates of scale (1), the influence function $IF(x; S, F)$ is proportional to the estimating function $\chi(x)$:

$$IF(x; S, F) \propto \chi(x).$$

Basing on this result, it is possible to construct an M-estimate with any admissible influence function, in particular, with the influence function of the Q_n-estimate of scale. This idea is used for constructing the approximation of the Q_n-estimate of scale by an M-estimate of scale.

The sought approximation, namely the estimating function $\chi(x)$ for MQ_n-estimates of scale, naturally depends on the underlying distribution density $f(x)$ shape: the explicit result gives the following form of this connection (Smirnov and Shevlyakov, 2014)

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

where the constant c_α is chosen from the condition of consistency and α is a tuning parameter. So, we call M-estimates with estimating function χ_α as MQ_n-estimates.

In what follows, we apply Equation (2) to the Student distribution densities in order to design computationally fast highly robust and efficient MQ_n-estimates of scale .

3 MQ_n-estimates for the Student distributions

3.1 Estimating functions for robust MQ_n-estimates

In order to get the estimating function of MQ_n-estimates of scale, we substitute the expression for the Student distribution density

$$f(x) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(k/2)(1+x^2/k)^{\frac{k+1}{2}}}, \quad k = 1, 2, \ldots,$$

into Equation (2) and compute the constant c_α determined from the condition of consistency

$$\int \chi_\alpha(x) f(x)\, dx = 0.$$

The formula for this consistency constant $c_\alpha(x)$ is given by

$$c_\alpha(x) = \frac{\Gamma^4(\frac{k+1}{2})\Gamma(k+1/2)}{2^{1-2k}\pi^{3/2}k^{3/2}\Gamma^3(k)} \left(1 - \frac{\alpha^2(k+1)(2k+1)}{24k(k+2)}\right).$$

It can be shown that the tuning parameter α lies in the interval $[0, 1/\sqrt{2}]$. The results of computation are presented in Table 1, so, a potential user may choose a consistency constant and thus with Equation (2) an MQ_n-estimate.

We skip the general formula for the estimating function $\chi_\alpha(x)$—it is rather cumbersome; in the particular case $\alpha = 0$, it has the form

$$\chi_0(x) = \frac{\Gamma^4(\frac{k+1}{2})\Gamma(k+1/2)}{2^{1-2k}\pi^{3/2}k^{3/2}\Gamma^3(k)} - \frac{\Gamma^2(\frac{k+1}{2})}{2^{-k}\pi\sqrt{k}\Gamma(k)(1+x^2/k)^{\frac{k+1}{2}}}.$$

The breakdown point of MQ_n-estimates of scale for the Student distributions with the tuning parameter α lying in $[0, 1/\sqrt{2}]$ is given by

$$\varepsilon^* = 1 - \frac{\Gamma^2(\frac{k+1}{2})\Gamma(k+1/2)}{2^{-k}\pi^{1/2}k^{3/2}\Gamma^3(k)} - \frac{\Gamma^2(\frac{k+1}{2})}{2^{-k}\pi^{1/2}k\Gamma^2(k)\left(2 - \frac{\alpha^2(k+1)}{3k}\right)} \left(1 - \frac{\alpha^2(k+1)(2k+1)}{24k(k+2)}\right).$$

The maximum possible breakdown point equal to 50% is attained at $k = 1$ and $\alpha = 0$ (see Fig. 1). With increasing α, the breakdown point is decreasing for any k; with given α and increasing k, the breakdown point also is decreasing.

3.2 Asymptotic efficiency of robust MQ_n-estimates

The asymptotic efficiency $eff(\widehat{S}_\alpha)$ of MQ_n-estimates of scale with the estimating function $\chi_\alpha(x)$ is computed by the following formula

$$eff(\widehat{S}_\alpha) = \frac{1}{V(\widehat{S}_\alpha, F)J(F)},$$

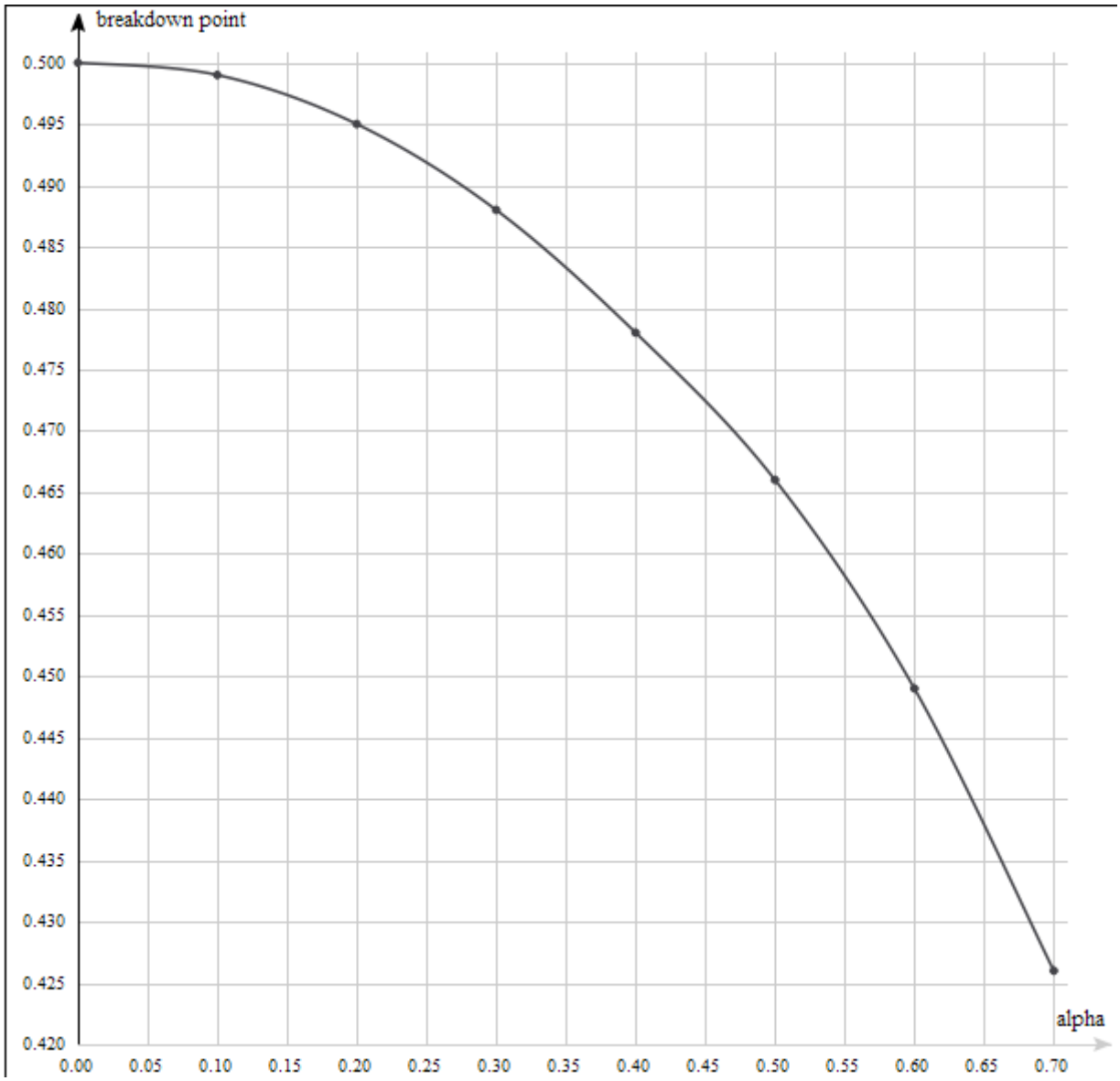

Figure 1: Breakdown Points of MQ_n-Estimates of Scale, $k = 1$

Table 1: Consistency Constant $c_\alpha(x)$ for MQ_n-Estimates of Scale

k, α	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
1	0.318	0.318	0.317	0.316	0.314	0.312	0.309	0.305
2	0.417	0.416	0.415	0.414	0.411	0.408	0.405	0.401
3	0.459	0.459	0.458	0.456	0.454	0.451	0.447	0.442
4	0.483	0.483	0.482	0.480	0.477	0.474	0.470	0.465
5	0.498	0.498	0.497	0.495	0.492	0.488	0.484	0.479
6	0.509	0.508	0.507	0.505	0.502	0.499	0.494	0.489
7	0.516	0.516	0.515	0.512	0.510	0.506	0.501	0.496
8	0.522	0.522	0.520	0.518	0.515	0.512	0.507	0.502
9	0.526	0.526	0.525	0.523	0.520	0.516	0.511	0.506
10	0.530	0.530	0.528	0.526	0.523	0.519	0.515	0.509
20	0.547	0.546	0.545	0.543	0.540	0.536	0.531	0.525
30	0.553	0.552	0.551	0.549	0.545	0.541	0.536	0.530
40	0.555	0.555	0.554	0.551	0.548	0.544	0.539	0.533
50	0.557	0.557	0.555	0.553	0.550	0.546	0.541	0.535
∞	0.564	0.564	0.562	0.560	0.557	0.552	0.547	0.541

where $V(\widehat{S}_\alpha, F)$ is the asymptotic variance of MQ_n-estimates of scale given by Equation (1), which takes the following form

$$V(\widehat{S}_\alpha, F) = \int IF(x; S_\alpha, F)^2 \, dF(x) = \frac{\int \chi_\alpha^2(x) \, dF(x)}{\left[\int x \chi'_\alpha(x) dF(x)\right]^2},$$

$J(F)$ is the Fisher information for scale

$$J(F) = \int \left[x \frac{f'(x)}{f(x)} + 1 \right]^2 dF(x) = \frac{2k}{k+3}.$$

The explicit expression for the asymptotic efficiency has been derived, but it is cumbersome and thus not written out; its numerical values are presented in Table 2.

4 Conclusions

1. The class of MQ_n-estimates of computationally fast and highly robust M-estimates of scale close in efficiency to the highly efficient and robust Q_n-estimate of scale is thoroughly studied for the Student distributions: explicit formulas are derived for the consistency constants, asymptotic efficiencies and breakdown points of those estimates.

2. The efficiency of the considered MQ_n-estimates are in the range $80\% - 100\%$, their breakdown points lie in the range $25\% - 50\%$—this means that MQ_n-estimates are highly efficient and robust.

Table 2: Asymptotic Efficiency of MQ_n-Estimates of Scale

k, α	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
1	1.000	1.000	1.000	0.999	0.998	0.995	0.989	0.979
2	0.984	0.985	0.986	0.988	0.991	0.993	0.995	0.996
3	0.962	0.962	0.964	0.967	0.971	0.976	0.981	0.986
4	0.942	0.943	0.945	0.948	0.953	0.958	0.965	0.972
5	0.926	0.927	0.929	0.932	0.937	0.944	0.939	0.948
6	0.913	0.914	0.916	0.920	0.925	0.931	0.939	0.948
7	0.903	0.903	0.906	0.909	0.915	0.921	0.929	0.939
8	0.894	0.895	0.897	0.901	0.906	0.913	0.921	0.931
9	0.887	0.888	0.888	0.890	0.894	0.906	0.914	0.924
10	0.881	0.881	0.884	0.887	0.893	0.900	0.908	0.918
20	0.849	0.849	0.852	0.855	0.861	0.867	0.876	0.886
30	0.836	0.837	0.839	0.843	0.848	0.855	0.863	0.874
40	0.830	0.830	0.832	0.836	0.841	0.848	0.856	0.867
50	0.826	0.826	0.828	0.832	0.837	0.844	0.852	0.862
∞	0.808	0.809	0.811	0.814	0.819	0.825	0.834	0.844

3. The asymptotic complexity of MQ_n-estimates is of order $O(n)$, much smaller than $O(n \log n)$ of the Q_n-estimate.

4. Note that in the case of the Cauchy distribution, the MQ_n-estimate is just the maximum likelihood estimate of scale with efficiency 100% and the breakdown point 50%; in the other limit case, for the Gaussian distribution, the efficiency and breakdown point are equal to 80.8% and 29.3%, respectively.

References

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

[2] Huber P.J. (1981). *Robust Statistics*. Wiley, New York.

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

[4] 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.