The Korean Communications
in Statistics Vol. 11 No. 1, 2004
pp. 153-160

Robust Estimator of Location Parameter[1]

Dongryeon Park[2]

Abstract

In recent years, the size of data set which we usually handle is enormous, so a lot of outliers could be included in data set. Therefore the robust procedures that automatically handle outliers become very importance issue. We consider the robust estimation problem of location parameter in the univariate case. In this paper, we propose a new method for defining robustness weights for the weighted mean based on the median distance of observations and compare its performance with several existing robust estimators by a simulation study. It turns out that the proposed method is very competitive.

Keywords : Location parameter, Median distance, Robust estimator, Robustness weight

1. Introduction

It is often assumed in the social sciences that data conform to a normal distribution. When estimating the location of a normal distribution, a sample mean is well known to be the best estimator according to many criteria. However, numerous studies (Hample et al., 1986; Hoaglin et al., 1976; Rousseeuw and Leroy, 1987) have strongly questioned normal assumption in real world data sets. In fact, a few large errors might infect the data set so that the tails of the underlying distribution are heavier than those of the normal distribution. In this situation, the sample mean is no longer a good estimate for the center of symmetry because all the observations equally contribute to the value of the sample mean, so the estimators which are insensitive to extreme values should have better performance.

The estimator is considered robust or resistant if small changes in many of the observations or large changes in only a few data points have small effect on its value. The median, the trimmed mean, and M-estimators are considered the examples of the robust measures of location parameter. These estimators define robustness weights in their own way to reduce the influence of outliers.

Suppose we have a random sample $X_1, \cdots, X_n$. The sample mean reduces to the

[1] This research was supported by the Hanshin University Special Research Grants in 2004.

[2] Associate Professor, Department of Statistics, Hanshin University, Osan 447-791, KOREA.
E-mail : drpark@hanshin.ac.kr

minimization of

$$\sum_{i=1}^{n} (X_i - \vartheta)^2. \tag{1}$$

Location M-estimators are defined by replacing the quadratic function in (1) with a objective function ρ :

$$\min_{\vartheta} \sum_{i=1}^{n} \rho(X_i - \vartheta). \tag{2}$$

The median corresponds to $\rho(x) = |x|$. Huber estimator and Tukey biweight estimator are defined by the objective function ρ_H and ρ_B, respectively :

$$\rho_H(x) = \begin{cases} x^2/2 & \text{for } |x| \le k \\ k|x| - k^2/2 & \text{for } |x| > k, \end{cases} \tag{3}$$

$$\rho_B(x) = \begin{cases} \frac{k^2}{6}\left\{1 - \left[1 - \left(\frac{x}{k}\right)^2\right]^3\right\} & \text{for } |x| \le k \\ k^2/6 & \text{for } |x| > k. \end{cases} \tag{4}$$

The value k for Huber and Tukey estimator is called a tuning parameter; smaller values of k produce more resistance to outliers, but at expense of lower efficiency when the errors are normally distributed. The trimmed mean is the mean of the central $n(1-2\alpha)$ observations in order.

In this paper, we propose a new method for defining robustness weights such that the weighted average using these weights is robust against outliers for estimating location parameter. In Section 2, the new estimator for the univariate case is defined and efficiency is compared with several existing methods by a simulation study in Section 3.

2. Defining New Robustness Weight

The idea of defining new robustness weights for location parameter is simple and easy to implement. Suppose we have a random sample $X_1, \cdots, X_n$ from $N(\mu, \sigma^2)$. Then most observations should lie in a range $[\mu-2\sigma, \mu+2\sigma]$. The distance between observations tends to be affected by the standard deviation of the population, which means that the distance between observations would have a certain pattern for given population. If an observation does not follow the pattern, then this observation is likely to be located outside of the range, so it can be considered as an outlier.

To make these idea more precise, we proceed as follow. For each i, let $d(X_{ij}) = \{|X_i - X_j|,\ j=1,\cdots,i-1,i+1,\cdots,n\}$. Then $d(X_{ij})$ would have $n-1$ different values whose distribution primarily depends on the population standard deviation. Suppose X_k is the only outlier among the observations, then $d(X_{kj})$ should be much larger than the other $d(X_{ij}),\ i\neq k$. However, since $d(X_{ij})$ consists of $n-1$ numbers, their median will be used for determining the weights of X_i.

The robustness weights for each observation is defined as follows:

1. For each $i=1,\cdots,n$, compute $d(X_{ij})$.
2. Let m_i be the median of $d(X_{ij})$.
3. Define the robustness weights for X_i by

$$r_i = B\left(\frac{m_i}{ks}\right) \tag{5}$$

where $s=$ median$\{m_i,\ i=1,\cdots,n\}$ and the value k plays same role as the turing parameter in Huber and Tukey estimator. B is the biweight function,

$$B(x) = \begin{cases} (1-x^2)^2 & \text{if } |x| < 1 \\ 0 & \textit{otherwise.} \end{cases} \tag{6}$$

Using r_i, we can compute the robust weighted average by

$$\sum_{i=1}^{n} w_i X_i, \quad \text{where } w_i = \frac{r_i}{\sum_{j=1}^{n} r_j}\ . \tag{7}$$

We refer to this estimator as the WMMD (Weighted Mean based on Median Distance). In steps 1 and 2, we compute $d(X_{ij})$, the absolute differences from X_i to the other observations and then compute m_i, the median value of these differences. In step 3, we compare m_i with m_j, $j\neq i$ and if m_i is large relatively to others, then X_i would have small weight. We determine the robustness weights using the bisquare function which is known to have good properties for robust estimation problem (Cleveland, 1979).

3. Comparison of the Performance

In this section, we provide numerical evidence of the effectiveness of WMMD in reducing the influence of outliers. We compare the performance of WMMD with the median, the trimmed mean, Huber estimator, and Tukey biweight estimator.

As a true population, we only considered the population which has a symmetric distribution. The reason for this is that estimators of location are best understood when a distribution's natural candidates for location all nearly coincide (e.g. mean, median, mode). We considered the following 5 distributions which have heavier tails than normal distribution:

1. $t(3)$
2. Cauchy distribution with location parameter 0 and scale parameter s with $s=1,5$
3. Contaminated normal (CN) distribution; $F(x) = 0.9\Phi(x) + 0.1\Phi\left(\frac{x}{\sigma}\right)$ with $\sigma = 10, 30$ where $\Phi(x)$ is the cumulative distribution function for $N(0,1)$

We need to choose the turning parameters for Huber estimator, Tukey estimator, and WMMD. Smaller values of the turning parameters make these estimators more resistant to outliers, but less efficient for the normal case. For Huber and Tukey estimator, we used the default values of S-Plus, which are 1.45 and 5, respectively. For the fair comparison, we need to choose the turing parameter of WMMD in such a way as WMMD produce almost identical performance with both Huber and Tukey estimator in the case of $N(0,1)$, and $k=5.5$ turned out to satisfy such criterion.

We considered 10% trimmed mean and 20% trimmed mean and they are denoted by T(10) and T(20), respectively.

The sample sizes considered were $n = 10, 20, \cdots, 90, 100$. The performance of each estimator was measured by Monte Carlo MSE over 1000 Monte Carlo simulation samples and the results are reported in Table 1 to Table 5. Each table shows the relative efficiency of WMMD to the other estimators, which is defined by the ratio of Monte Carlo MSE of the other estimator to that of WMMD. The larger number indicates the better performance of WMMD.

Table 1 : The relative efficiency of WMMD for Cauchy distribution with location parameter 0 and scale parameter 1.

n	Tukey	Huber	T(10)	T(20)	Median
10	1.0663	1.3010	3.7453	1.3570	0.8012
20	1.0443	1.2764	2.3764	1.1194	0.8883
30	1.0603	1.2168	1.9779	1.0330	0.8895
40	1.0524	1.3714	1.9724	1.1616	0.9031
50	1.0724	1.3955	2.0377	1.1338	0.9085
60	1.0754	1.3592	1.9371	1.1284	0.9427
70	1.0872	1.3768	1.9304	1.0936	0.8971
80	1.0853	1.4115	1.9956	1.1316	0.9762
90	1.0803	1.4372	2.0113	1.1413	0.8981
100	1.0879	1.4418	1.9929	1.1401	0.9697

Table 2 : The relative efficiency of WMMD for t(3)

n	Tukey	Huber	T(10)	T(20)	Median
10	1.0285	1.0007	1.0490	0.9715	1.0241
20	1.0116	1.0139	1.0456	0.9770	1.1447
30	1.0184	0.9978	1.0087	0.9545	1.1452
40	1.0124	1.0169	1.0373	0.9646	1.2002
50	1.0456	1.0596	1.0655	1.0133	1.1485
60	1.0278	1.0472	1.0595	0.9768	1.1515
70	1.0293	1.0665	1.0793	1.0065	1.1502
80	1.0223	1.0522	1.0535	1.0131	1.1969
90	1.0307	1.0355	1.0376	0.9807	1.1692
100	1.0126	1.0301	1.0364	0.9935	1.2512

Table 3 : The relative efficiency of WMMD for Cauchy distribution with location parameter 0 and scale parameter 5.

n	Tukey	Huber	T(10)	T(20)	Median
10	0.9805	1.2873	5.1422	1.4310	0.8092
20	1.0351	1.3542	2.5805	1.1813	0.8580
30	1.0208	1.3300	2.0916	1.1255	0.8847
40	1.0533	1.3710	2.0492	1.1377	0.8903
50	1.0672	1.3689	1.9666	1.1126	0.8870
60	1.0701	1.3734	2.0111	1.1178	0.9062
70	1.0747	1.3029	1.8574	1.0588	0.9142
80	1.0711	1.3252	1.7756	1.0691	0.8768
90	1.0735	1.3784	1.9449	1.1080	0.9595
100	1.0892	1.3032	1.7732	1.0493	0.8456

Table 4 : The relative efficiency of WMMD for CN distribution with $\sigma = 30$

n	Tukey	Huber	T(10)	T(20)	Median
10	1.0501	1.2441	1.2157	1.2081	1.4027
20	0.9940	1.1929	1.1941	1.1803	1.4949
30	0.9998	1.3625	1.3553	1.2793	1.6027
40	0.9898	1.2581	1.2530	1.2050	1.4581
50	0.9853	1.3301	1.3328	1.2835	1.5966
60	0.9868	1.2895	1.2914	1.2478	1.6081
70	0.9767	1.2155	1.2105	1.2030	1.5292
80	0.9823	1.3263	1.3201	1.2783	1.6206
90	0.9770	1.2136	1.2119	1.2133	1.6098
100	0.9829	1.2981	1.2880	1.2594	1.7215

Table 5 : The relative efficiency of WMMD for CN distribution with $\sigma = 10$

n	Tukey	Huber	T(10)	T(20)	Median
10	1.0382	1.1417	1.1133	1.1083	1.3246
20	0.9960	1.1333	1.1175	1.1228	1.4430
30	0.9952	1.1694	1.1452	1.1391	1.3830
40	0.9833	1.1533	1.1439	1.1588	1.5269
50	0.9808	1.1746	1.1750	1.1890	1.5125
60	0.9761	1.1379	1.1412	1.1435	1.5349
70	1.0030	1.1726	1.1586	1.1174	1.3543
80	0.9800	1.1858	1.1766	1.1830	1.4547
90	0.9746	1.1855	1.1824	1.1775	1.5009
100	0.9838	1.1542	1.1470	1.1576	1.4471

The performance of the median heavily depends on the type of the underlying distribution. There is no doubt that the median is the best location estimator for Cauchy distribution, but shows very poor performance at t(3) and CN distributions. Since we hardly know the exact form of the undelying distribution in practice, the median is not a good choice in a practical point of view. The performance of the trimmed means also depends on the type of the underlying distribution. Besides, they show very poor results at Cauchy and CN distributions. It is very clear that WMMD is better than Huber for all distributions here. It is also clear that WMMD is better than Tukey for Cauchy distribution, but Tukey and WMMD yield very comparable results for t(3) and CN distributions.

Simulation results indicate that WMMD is a competitive method. Moreover, WMMD does not require iteration and generalizes easily to higher dimensions, and these are another attractive points of WMMD over other estimators.

4. Conclusion

In recent years, the size of data set which we usually handle is enormous, so a lot of outliers could be included in data set. Therefore the robust procedures that automatically handle outliers becomes very importance issue.

For local regression problem, Park (2003) proposes the idea of determining robustness weights using weighted median distance between observations and compares the performance with lowess (Cleveland, 1979). It turns out that his proposed method is more appropriate for heavy contamination. In this paper, we expanded the idea of Park (2003) to the univariate location parameter estimation problem. we proposed the robustness weights for the weighted mean in the univariate case and compared its finite sample properties with several existing methods. The asymptotic behavior of the proposed method is not derived in this paper, and is left as a further research topic.

References

[1] Cleveland, W. S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association.* **74**, 829-836.

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

[3] Hoaglin, D. C., Mosteller, F., and Tukey, J. W. (1976). *Understanding Robust and Exploratory Data Analysis.* John Wiley & Sons, New York.

[4] Park, D. (2004). Robustness Weight by Weighted Median Distance. To appear in *Computational Statistics.*

[5] Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection.* John Wiley & Sons, New York.

[6] Venables, W. N. and Ripley, B. D. (1999). *Modern Applied Statistics with S-PLUS.* Springer-Verlag, New York.

[Received September 2003, Accepted April 2004]