Journal of Science and Technology, 5(1): 17-23, January 2015

PRODUCT M ESTIMATOR OF CORRELATION FOR BIVARIATE DATA : A SIMULATION STUDY

DELUAR JAHAN MOLOY[*1] and JAFAR A. KHAN[2]

[1]Department of Statistics, Mawlana Bhashani Science and Technology University, Tangail; [2]Department of Statistics, Biostatistics and Informatics, University of Dhaka

Abstract

We consider the problem of estimating the correlation coefficient from the bivariate data containing a fraction of outliers. Classical product-moment estimate is affected by the outliers, while existing robust estimates are computationally inefficient. In order to achieve robustness and computational efficiency at the same time, we propose a new robust estimator of correlation, which is called the Product M (PM) correlation estimator. The classical estimator uses non-robust estimators mean and standard deviation as the building blocks. To construct the proposed PM estimator, we replaced these non-robust estimators by robust location M-estimator and MAD. Thus, we developed robust correlation estimator that does not use iterative algorithm. Our simulation studies and real data application show that the proposed PM estimator gives better results in the contaminated data compared to the classical estimator. The performance of our estimate is as good as that of the existing robust estimators. The advantage of our estimator is that it requires less computing time compared to the existing robust estimators.

Keywords: Contaminated data, robustness, iterative algorithm, computational efficiency

Introduction

Real data if contain a fraction of outliers and other contaminations which are difficult to visualize and clean. The classical correlation coefficient, i.e., Pearson's product-moment correlation coefficient *r* is much affected by these outliers and often gives misleading results. We, therefore, turn our attention to robust method of estimation of correlation coefficient. Robust methods are designed to consider the majority of the data, rather than all the data. Therefore, robust methods give reasonable results even when datasets contain a fraction of outliers. A major drawback of the existing robust methods is that they are not computationally suitable, because fitting a robust model is a non-linear optimization problem. In this study, we attempt to develop new correlation estimator for bivariate data that is resistant to outliers and computationally efficient. This estimator is obtained through the robustification of Pearson's *r*. We call this estimator the Product M (PM) estimator $\hat{\rho}_{PM}$ of ρ

In this paper, we present our new robust estimator. We then show the results of a simulation study to compare the performance of our PM estimator with classical *r*, robust

[*] Corresponding author: deluarmoloy@gmail.com

MVE estimator (Rousseeuw, 1985) and Median product (MP) estimator (Shafiullah, 2009; Shafiullah and Khan, 2011)

Product M Correlation Estimator

Pearson's product-moment correlation estimator *r* can be expressed as

$$r = mean\left(Z_x \times Z_y\right)$$

where, $Z_x = \left(x - \bar{x}\right)/s_x$ and $Z_y = \left(y - \bar{y}\right)/s_y$ are the standardized variables, the standardization being done by using classical estimates arithmetic mean and standard deviation. A simple robustification of *r* can be obtained by replacing the mean and standard deviation by the location M-estimator (Huber, 1964) and MAD (Median Absolute Deviation) respectively. Thus, an initial robust estimator, denoted by r_I, is obtained as

$$r_I = M\left(Q_x \times Q_y\right)$$

where, M stands for M-estimator, Q_x and Q_y are robust standardized variables defined as

$$Q_x = \left(x - M\left(x\right)\right)/MAD\left(x\right) \text{ and } Q_y = \left(y - M\left(y\right)\right)/MAD\left(y\right)$$

We consider r_I as an initial robust estimator, because the range of r_I is different that of the classical correlation estimator *r*. This is elaborated below.

When the data follows bivariate normal distribution and there is perfect positive correlation between X and Y (i.e., $\rho = 1$), we have $Q_x = Q_y = Q$. This gives

$$r_I\left(\max\right) = M\left(Q^2\right) \text{ and } r_I\left(\min\right) = -M\left(Q^2\right)$$

where, $Q^2 \sim \chi^2_{(1)}$. The M-estimate of χ^2_1 random variable for sample of size 1000000 (one million) is .5318. Thus we have $-0.5318 \le r_I \le 0.5318$.

In order to obtain the final estimator, i.e., PM correlation estimator $\hat{\rho}_{PM}$, we make a transformation of r_I. First, let us define $\rho_I = \lim_{n\to\infty} r_I$. Since $\rho_I \neq \rho$, we conduct a numerical study of the functional relationship $\rho_I = g\left(\rho\right)$. Table 1 shows the values of ρ_I corresponding values of ρ between 0.00 and 0.99 with interval 0.01.

For each value of ρ in the table 1, we obtained the corresponding value of ρ_I using a bivariate normal sample of size $n = 1$ million. For any negative value of ρ, the value of ρ_I corresponds to the value of $|\rho|$, but with a negative sign.

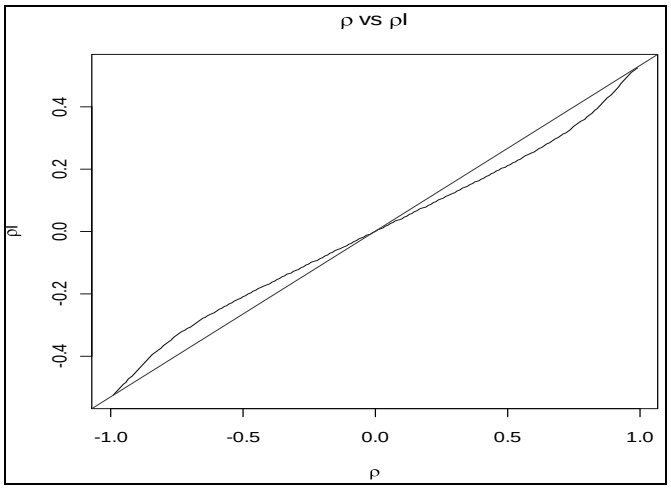

Fig. 1. Relationship between ρ and ρ_I.

The graphical relationship between the values of ρ_I against the values of ρ (Fig. 1). To calculate $\hat{\rho}_I$ for a particular dataset, we first calculate r_I from the data and use Table 1 to make the transformation $\hat{\rho}_{PM} = g^{-1}(r_I)$. For example, if r_I = 0.2322 for a particular dataset, then Table 1 suggests that $\hat{\rho}_{PM}$ = 0.55.

Table 1. Values of ρ_I for different values of ρ.

ρ	.00	0.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.0001	.0042	.0084	.0125	.0167	.0208	.0249	.0289	.0331	.0375
.1	.0416	.0458	.0499	.0541	.0582	.0582	.0666	.0707	.0749	.0790
.2	.0832	.0874	.0915	.0956	.0998	.1040	.1082	.1124	.1165	.1207
.3	.1249	.1290	.1332	.1374	.1415	.1456	.1498	.1540	.1582	.1625
.4	.1667	.1709	.1752	.1795	.1838	.1881	.1924	.1968	.2010	.2054
.5	.2098	.2143	.2187	.2232	.2277	**.2322**	.2367	.2413	.2459	.2505
.6	.2552	.2599	.2648	.2696	.2744	.2794	.2845	.2897	.2949	.3001
.7	.3055	.3109	.3165	.3221	.3280	.3339	.3399	.3461	.3525	.3591
.8	.3658	.3726	.3798	.3872	.3946	.4023	.4102	.4184	.4271	.4361
.9	.4458	.4549	.4640	.4732	.4822	.4907	.4992	.5078	.5160	.5242

We now conduct simulation studies to examine the performance of the proposed PM estimator and compare it with classical estimator *r* and existing robust estimators $\hat{\rho}_{MVE}$ and $\hat{\rho}_{MP}$.

Simulations

We first carried out a simulation to show that Pearson's product-moment correlation estimator *r* is sensitive to outliers while the existing robust MVE (Minimum Volume

Ellipsoid), MPE (Median Product Estimator) and the proposed robust MP estimators are resistant to outliers. For this, we plot the sampling distributions of these estimators for both clean and contaminated data. We then conducted another simulation study to compare the features of $\hat{\rho}_{PM}$, $\hat{\rho}_{MVE}$ and $\hat{\rho}_{MP}$ with respect to standard error, magnitude of bias and CPU time required. We used statistical software R to carry out the simulations.

Robustness of the estimator

In this subsection, we examine the robustness of our proposed estimator along with classical *r* and robust MVE and MP estimators. We generate 200 datasets each of size n = 1000 from bivariate normal distribution with $\rho = 0.5$. For each dataset, we calculated PM correlation estimate $\hat{\rho}_{PM}$ along with classical estimate *r* and existing robust estimates $\hat{\rho}_{MVE}$ and $\hat{\rho}_{MP}$.

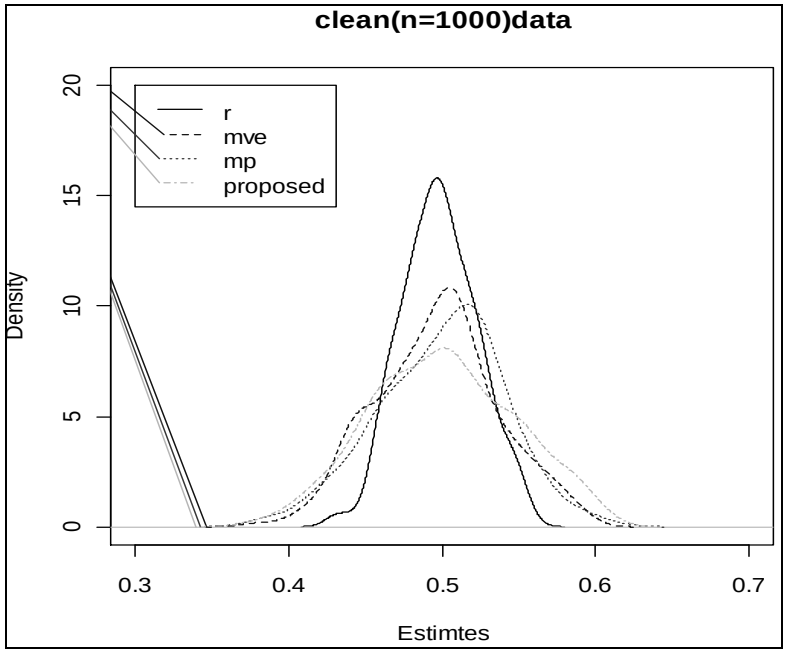

Fig. 2. Sampling distribution of $r, \hat{\rho}_{MVE}, \hat{\rho}_{MP}$ and $\hat{\rho}_{PM}$ for clean data.

Then the datasets are contaminated by replacing a fraction of it by outliers. Each observation of a variable is assigned probability 0.025 of being replaced by a large number. Therefore, the probability that any particular row of the dataset will be contaminated is $1-(1-0.025)^2$, which means approximately 5% of the rows will be contaminated. We then calculated the four estimates again from the contaminated data.

We plotted the sampling distributions of the four estimators for both clean and contaminated data sets. For clean data, figure 2 reveals that all the four estimators give

similar results for clean data. But the classical estimator *r* is the best choice because it has smaller standard error.

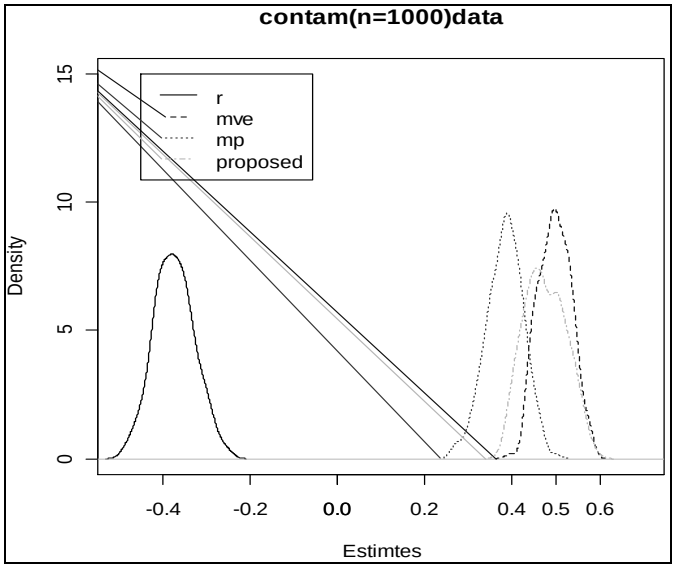

Fig. 3. Sampling distribution of *r*, $\hat{\rho}_{MVE}$, $\hat{\rho}_{MP}$ and $\hat{\rho}_{PM}$ for contaminated data.

The sampling distributions of *r*, $\hat{\rho}_{MVE}$, $\hat{\rho}_{MP}$ and $\hat{\rho}_{PM}$ for contaminated data (Fig. 3). We see that the classical estimator *r* is drastically affected by the outliers, because its density plot does not even include the true parameter $\rho = 0.5$. On the other hand, the robust $\hat{\rho}_{MP}$ estimator does not estimate the parameter properly but $\hat{\rho}_{MVE}$ and $\hat{\rho}_{PM}$ estimators estimate the parameter properly and not affected by the outliers.

$\hat{\rho}_{PM}$ versus $\hat{\rho}_{MVE}$ and $\hat{\rho}_{MP}$

Here we considered three different sample sizes: *n* = 25, *n* = 100, *n* = 400. For each sample size, we generated 200 contaminated datasets form bivariate normal distribution with $\rho = 0.1$, $\rho = 0.5$ and $\rho = 0.9$.

Table 2. Standard error, Magnitude of bias and elapse computing time of MVE, MP and PM estimators

Criteria	ρ	n = 25			n = 100			n = 400		
		MVE	MP	PM	MVE	MP	PM	MVE	MP	PM
Standard error	0.1	.330	.230	.300	.154	.137	.161	.074	.066	.076
	0.5	.267	.225	.252	.120	.125	.130	.056	.065	.070
	0.9	.079	.184	.163	.032	.074	.080	.014	.038	.039
	0.1	.003	.129	.009	.006	.099	.000	.0009	.093	.009
Magnitude of bias	0.5	.018	.246	.055	.008	.147	.029	.0000	.129	.027
	0.9	.010	.296	.153	.003	.158	.089	.0000	.153	.091
Elapse computing time (sec.)	0.1	1.61	0.15	0.44	2.50	0.21	0.75	9.88	0.19	0.72
	0.5	1.66	0.27	0.23	2.60	0.29	0.51	9.64	0.23	0.86
	0.9	1.62	0.13	0.28	2.51	0.27	0.62	9.49	0.24	0.92

Table 2 represents the standard error, magnitude of bias and elapse computing time for $\hat{\rho}_{MVE}$, $\hat{\rho}_{MP}$ and $\hat{\rho}_{PM}$.

We see that standard error and magnitude of bias of our proposed estimator PM are larger than MVE estimator but smaller than MP estimator. On the other hand PM estimator takes less time than existing MVE estimator but more time than MP estimator. Though PM estimator takes more time than MP, PM provides better result than MP.

Application: Motorola vs Market Data

We apply the proposed PM estimator along with classical estimator *r* to Motorola vs. Market data (Adrover, Salibian-Barrera and Zamar, 2002). The response variable (Y) is the difference between the monthly Motorola returns and the returns on 30-day US Treasury bills. The explanatory variable (X) is the difference between the monthly Market returns and the returns on 30-day US Treasury bills.

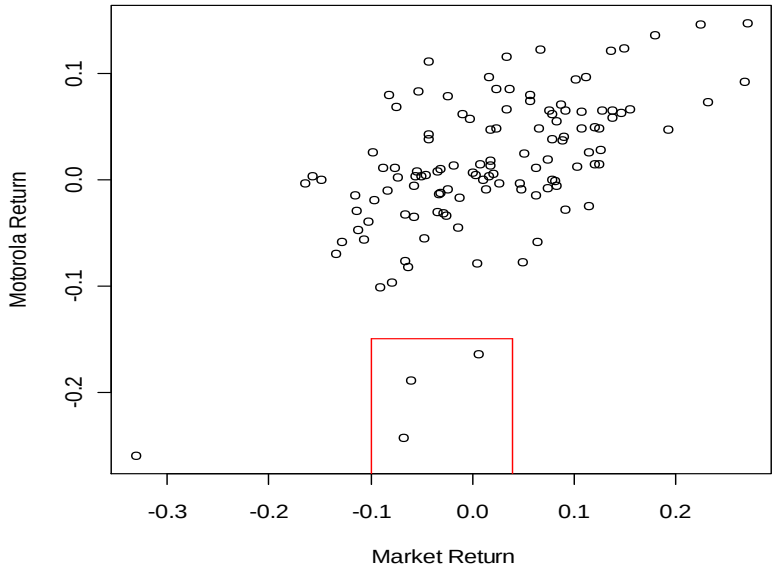

Fig. 4. A scatter plot of motorola return vs. market return.

First, we obtain the two estimates from the original data that contain possible outliers. Then based on the scatter plot (figure 4), we removed three outlying observations from the data and calculate the two estimates again. The results show as below:

Table 3. Different correlation estimates for Motorola vs. Market data

Method of estimation	Estimate of ρ	
	Clean data	Contaminated data
Classical (Pearson's product moment)	0.635	0.596
Proposed robust (MP)	0.610	0.590

Above table reveals that, for the contaminated data, the classical estimate changes from 0.635 to 0.595. On the other hand, our proposed robust estimate shows less sensitivity to outliers (Changes from 0.610 to 0.590).

Conclusion

In this article, we proposed a new robust estimator for bivariate data that does not use iterative algorithm. The proposed Product M (PM) correlation estimator achieves robustness and computational efficiency at the same time. The classical estimator *r* is the mean of the product of two standardized variables. We obtain the initial robust estimator r_I by replacing the mean and standard deviation used in *r* by M-estimator and MAD respectively. Thus, r_I is the M-estimate of the product of two robustly standardized variables. The problem with r_I is that $-0.5318 \le r_I \le 0.5318$, where 0.5318 is the M-estimate of χ_1^2 random variable for sample of size 1000000. Since r_I does not converge to ρ, we denoted the asymptotic value of r_I by ρ_I, and performed an extensive simulation study to explore the relationship between ρ and ρ_I. Based on this numerical study, we made a transformation of r_I and obtained the PM estimator of ρ denoted by $\hat{\rho}_{PM}$. Thus we ensured that $-1 \le \hat{\rho}_{PM} \le 1$, so that $\hat{\rho}_{PM}$ can be compared to classical *r*. The new robust estimator $\hat{\rho}_{PM}$ has much better performance compared to classical *r* in the contaminated data. In the clean data, these two estimators give similar results. When compared to existing robust estimators $\hat{\rho}_{MVE}$ and $\hat{\rho}_{MP}$, the standard error and magnitude of bias of our estimator are slightly larger than those of $\hat{\rho}_{MVE}$ and $\hat{\rho}_{MP}$. Moreover, our proposed estimator requires less time and computationally more suitable.

References

Adrover, J. G., M. Salibian-Barrera and R. H. Zamar, (2002). Globally robust inference for the location and simple linear regression model. *J. Statist. Plan inference*. **119**: 353-375

Huber, P. J., (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*. **35**: 73-101.

Rousseeuw, P. J., (1985). Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications*, Vol. B: 283-297, Reidel, Dordrecht.

Shafiullah, A. Z. M., (2008). *Median-Product Correlation and Regression Estimators: New Robust Estimators for Bivariate Data*. M. S. Thesis, Department of Statistics, Biostatistics & Informatics, University of Dhaka

Shafiullah, A. Z. M. and J. A. Khan, (2011). A new robust correlation estimator for bivariate data. *The Bangladesh Journal of scientific Research.* **24**(2): 97-105