Bangladesh J. Sci. Res. 24(2): 97-106, 2011 (December)

A NEW ROBUST CORRELATION ESTIMATOR FOR BIVARIATE DATA

A. Z. M. Shafiullah* and Jafar A. Khan

Department of Statistics, Biostatistics & Informatic
University of Dhaka, Dhaka- 1000, Bangladesh

Abstract

The problem of calculating the correlation estimate from bivariate data containing a fraction of outliers has been considered in this paper. Classical product-moment estimate is affected by outliers, while 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. We call our estimator Median-Product (MP) correlation estimator. The classical estimator of correlation uses non-robust estimator mean and standard deviation as the building blocks. To construct the proposed MP estimator, we replaced these non-robust estimators by their robust counterpart median and MAD (Median Absolute Deviation from median). Thus, we developed robust estimator of correlation that does not use iterative algorithm. Our simulation studies and real data application show that the proposed MP estimator of correlation gives better results in the contaminated data compared to the classical estimator. The performance of our estimator is similar to that of the existing robust estimators. The advantage of our estimator is that it requires less computing time compared to the existing robust estimators.

Key words: Outliers, robustness, iterative algorithm, computational efficiency, median product

Introduction

Real datasets usually contain a fraction of outliers and other contaminations. The classical correlation coefficient, i.e., Pearson's product-moment correlation coefficient r is much affected by these outliers and often gives misleading results. Robust methods are designed to consider the *majority* of the data rather than *all* the data. Therefore, robust methods give reasonable results even when data contain a fraction of outliers. Several robust correlation estimators are available in the literature. Stahel (1981) and Donoho (1982) proposed the Stahel-Donoho estimator of multivariate location and scatter. This estimator is the weighted mean vector and covariance matrix, where the weight assigned to a point decreases as the distance of the point from the estimated center increases. The Minimum Volume Ellipsoid (MVE) estimator (Rousseeuw and Leroy 1987 and Rousseeuw and van Zomeren 1990) searches among all ellipsoids containing half of the

* Corresponding author. E-mail: stalin_dustat@yahoo.com

data to obtain the ellipsoid with the minimum volume. The mean vector and covariance matrix are then calculated from the points in this minimum-volume ellipsoid. Marazzi (1993) rescales the covariance matrix to obtain consistency at the multivariate normal model. The S-estimate (Davies 1987) minimizes an M-scale of the distances of the points from the estimated center. Minimum Covariance Determinant (MCD) estimator (Rousseuw and Vandriessen 1999 and Maronna *et al.* 2006) searches for half the data that has the smallest "trimmed scale". A major drawback of existing robust methods is that they are not computationally suitable, because fitting a robust model is a nonlinear optimization problem. In this study we propose a new robust correlation estimator (MP) for bivariate data that is resistant estimator $\hat{\rho}_{MP}$ of ρ and this estimator achieves robustness and computational efficiency at the same time. In the following section we present our new robust estimator. Then we show the results of simulation study to compare the performance of our MP estimator with classical r and robust MVE estimators. We apply the proposed estimator to a real data set and write the summary of this article.

Median product correlation estimator

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

$$r = mean(Z_x \times Z_y) \tag{1}$$

where $Z_x = (x - \bar{x})/s_x$ and $Z_y = (y - \bar{y})/s_y$ are the standardized variables, the standardization being done by using the classical estimates means and standard deviations. A simple robustification of r can be performed by replacing these non-robust building blocks of r by their robust counterparts. Thus, an initial robust estimator, denoted by r_M, is obtained as

$$r_M = median(Q_x \times Q_y), \tag{2}$$

where Q_x and Q_y are robustly standardized variables defined as

$$Q_x = (x - median(x))/MAD(x) \text{ and } Q_y = (y - median(y))/MAD(y).$$

We are considering r_M as an "initial" robust estimator, because the range of r_M is different from that of the classical correlation estimator r. This is elaborated below. When the data follow 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

$$\max r_M = median(Q^2) \text{ and } \min r_M = -median(Q^2),$$

where $Q^2 \sim \chi_1^2$. The median of χ_1^2 random variable is 0.4549. Thus, we have

$$-0.4549 \leq r_M \leq 0.4549.$$

In order to obtain the final estimator, i.e., MP correlation estimator $\hat{\rho}_{MP}$, we make a transformation of r_M. First, let us define $\rho_M = \lim_{n\to\infty} r_M$. Since, $\rho_M \neq \rho$, we conducted a numerical study of the functional relationship $\rho_M = g(\rho)$.

Table 1. ρ_M **for different values of** ρ

ρ	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0014	0.0029	0.0044	0.0067	0.0085	0.0110	0.0127	0.0151	0.0172
0.1	0.0196	0.0225	0.0248	0.0274	0.0305	0.0336	0.0355	0.0384	0.0416	0.0437
0.2	0.0477	0.0509	0.0545	0.0576	0.0605	0.0637	0.0669	0.0705	0.0748	0.0774
0.3	0.0813	0.0856	0.0888	0.0926	0.0962	0.0999	0.1040	0.1075	0.1129	0.1159
0.4	0.1199	0.1246	0.1299	0.1335	0.1363	0.1406	0.1455	0.1508	0.1548	0.1592
0.5	0.1633	0.1681	0.1737	0.1778	0.1827	0.1866	0.1920	0.1974	0.2023	0.2079
0.6	0.2114	0.2166	0.2221	0.2277	0.2321	0.2384	0.2419	0.2489	0.2547	0.2583
0.7	0.2643	0.2697	0.2758	0.2821	0.2875	0.2934	0.2996	0.3056	0.3106	0.3166
0.8	0.3227	0.3285	0.3349	0.3414	0.3479	0.3526	0.3595	0.3674	0.3717	0.3784
0.9	0.3857	0.3922	0.3984	0.4062	0.4130	0.4202	0.4270	0.4328	0.4401	0.4477

Table 1 shows the asymptotic values of ρ_M for different values of ρ. The first column and first row of the table give the first and second places of decimal for the values of ρ. For example, the third value of first column (0.2) and forth value of first row (0.03) altogether gives the value of $\rho = 0.23$, and the corresponding entry in the table gives the value of $\rho_M = 0.0576$. They measure the same degree of association. The table includes the values of ρ_M for every corresponding values of ρ between 0.00 and 0.99 with an increment of 0.01. For constructing the table, we consider each value of ρ and generated bivaraite data of size $n = 1$ million from normal population. Then we obtained the corresponding value of ρ_M using (1) as asymptotic value of *r*. For negative values of ρ, the values of ρ_M corresponds to the value of $|\rho|$, but with a negative sign.

Figure 1 plots the values of ρ_M against the values of ρ.

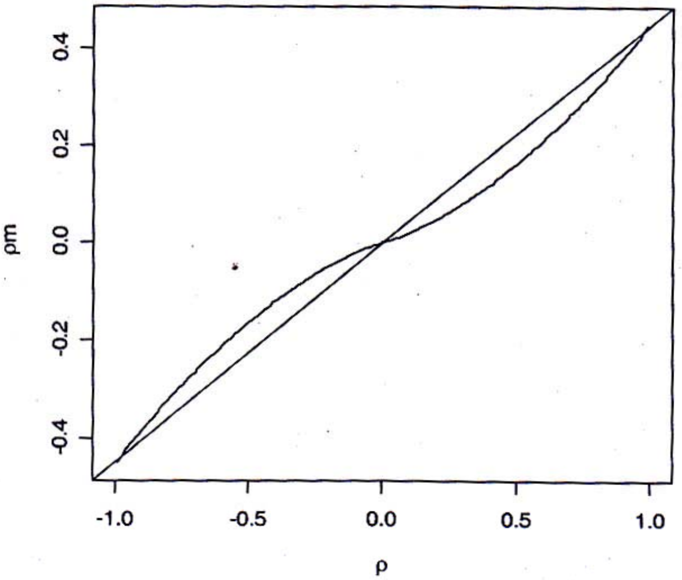

Figure 1. Relationship between ρ and ρ_M.

To calculate $\hat{\rho}_{MP}$ for a particular dataset, we first calculate r_M from the data, and use Table 1 to make the transformation $\hat{\rho}_{MP} = g^{-1}(r_M)$. For example, if $r_M = 0.3857$ for a particular data set, then Table 1 suggests that $\hat{\rho}_{MP} = 0.90$.We now conduct simulation studies to examine the performance of the proposed MP estimator and compare it with classical estimator *r* and existing robust estimator $\hat{\rho}_{MVE}$.

Simulation

In order to justify the performance of the proposed estimator $\hat{\rho}_{MP}$, we conduct extensive simulation studies. 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 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 $\hat{\rho}_{MP}$ and $\hat{\rho}_{MVE}$ with respect to standard error, magnitude of bias and CPU time required. We used R to carry out the simulations.

Robustness of the estimators

We generated 200 datasets each of size $n = 1000$ from bivariate normal distribution with $\rho = 0.5$. For each dataset, we calculated MP correlation estimate $\hat{\rho}_{MP}$ along with classical estimate r and the existing robust estimate $\hat{\rho}_{MVE}$. Then, the data are contaminated by replacing a fraction of observations on X and Y 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 will be contaminated is $1-(1-0.025)^2$, which means that approximately 5% of the rows will be contaminated. We then calculated the three estimates again from the contaminated data. We plotted the sampling distributions of the three estimators for clean data sets and contaminated data sets. Figure 2 reveals that all the three estimators give similar results for clean data, though the classical estimator r has smaller standard error. Figure 3 shows the sampling distributions of r, $\hat{\rho}_{MVE}$ and $\hat{\rho}_{MP}$ for contaminated data.

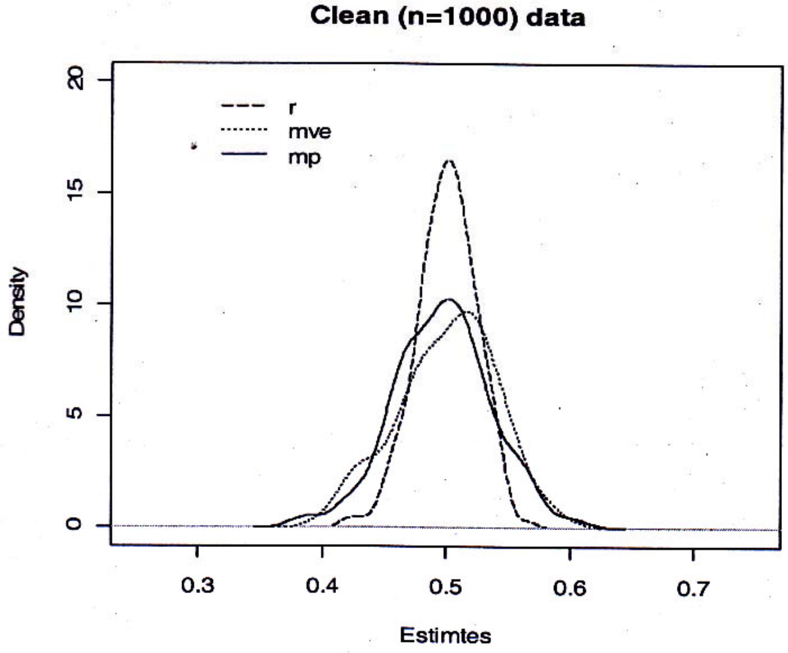

Figure 2. Sampling distributions of r, $\hat{\rho}_{MVE}$ and $\hat{\rho}_{MP}$ for clean data.

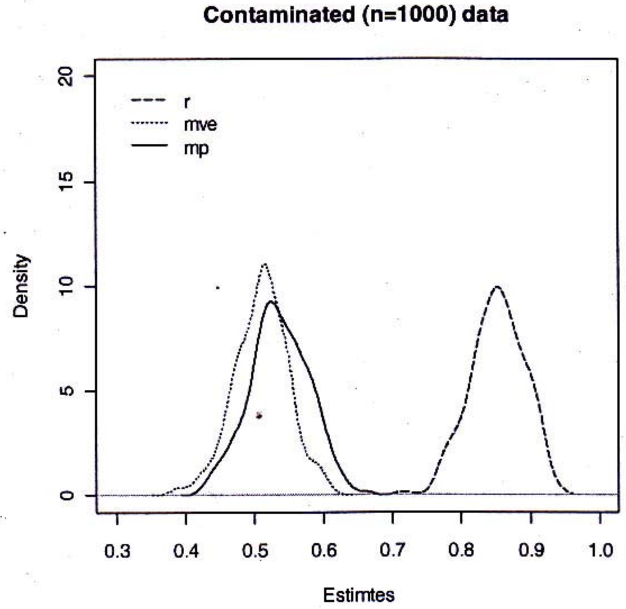

Figure 3. Sampling distributions of *r*, $\hat{\rho}_{MVE}$ and $\hat{\rho}_{MP}$ for contaminated data.

We observe that the classical estimator *r* is seriously affected by the contaminations, because its density plot does not even include the true parameter $\rho = 0.5$. On the other hand, the robust estimators $\hat{\rho}_{MP}$ and $\hat{\rho}_{MVE}$ are not affected by the outliers.

$\hat{\rho}_{MP}$ *versus* $\hat{\rho}_{MVE}$

We considered three different sample sizes: $n = 25, n = 100$ and $n = 400$. For each sample size, we generated 200 datasets from bivariate normal distribution with $\rho = 0.1, 0.5$ and 0.9. For each generated dataset, we calculated $\hat{\rho}_{MP}$ and $\hat{\rho}_{MVE}$. Table 2 presents the standard errors, average magnitude of bias and CPU time required for these two estimators. We observe that the existing MVE estimator has smaller standard error and magnitudes of bias for small and large values of ρ. However, for moderate value of ρ, the proposed MP estimator has less bias compared to MVE. Moreover, MP requires much less CPU time. Also, the standard error of all these estimates decrease as sample size increases, while the bias does not change with sample size.

Table 2. Simulation results (Average of estimates, standard errors, average biases and CPU times of MVE and MP estimators).

Criteria	ρ	*n*=25		*n*=100		*n*=400	
		MVE	MP	MVE	MP	MVE	MP
Average of Estimates	0.1	0.112	0.166	0.101	0.156	0.105	0.167
	0.5	0.543	0.518	0.504	0.528	0.511	0.527
	0.9	0.919	0.946	0.900	0.902	0.901	0.908
Standard error	0.1	0.293	0.233	0.156	0.114	0.069	0.072
	0.5	0.282	0.254	0.115	0.123	0.058	0.066
	0.9	0.078	0.139	0.028	0.064	0.014	0.037
Average bias	0.1	0.012	0.066	0.001	0.056	0.005	0.067
	0.5	0.043	0.018	0.004	0.028	0.011	0.027
	0.9	0.019	0.046	0.000	0.002	0.001	0.008
CPU time	0.1	6.54	0.80	8.67	1.20	23.79	1.18
	0.5	5.75	1.01	8.82	1.18	23.74	1.17
	0.9	5.71	1.05	8.52	1.13	23.70	1.57

Application: Motorola vs. Market Data

We applied the proposed MP estimator along with classical r to Motoroala vs. Market data (Adrover *et al.* 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. First, we obtained 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 calculated the two estimates again. The results are shown in Table 3.

Table 3. Different correlation estimates for Motorola data.

Estimation	Clean data	Contaminated data
Classical	0.63	0.59
Proposed MP	0.66	0.66

The classical estimator *r* gives much different results for the original and cleaned data, while the proposed MP estimator gives same results. This shows that MP estimator is not affected by outliers.

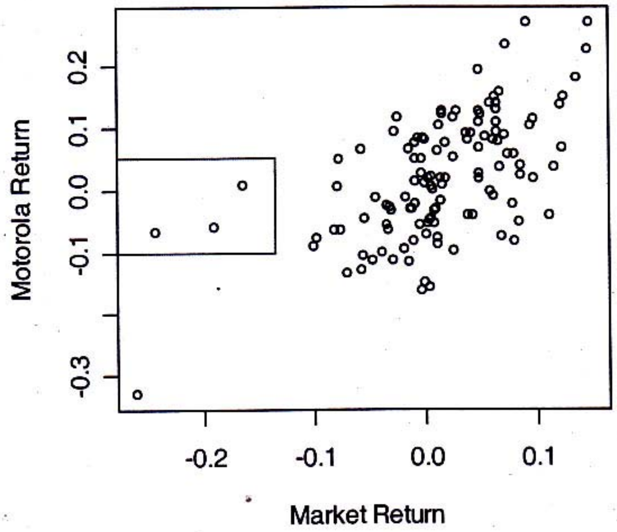

Figure 4. Motorola return vs. Market return.

Summary

In this article, we proposed a new robust estimator for bivariate data that does not use iterative algorithm. The proposed Median-Product (MP) 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 obtained initial robust estimator r_M by replacing the means and standard deviations used in r by median and MAD. Thus, r_M is the median of the product of two robustly standardized variables. The problem with r_M is that $-0.4549 \le r_M \le 0.4549$, where 0.4549 is the median of χ_1^2 random variable. We denoted the asymptotic value of r_M by ρ_M, and performed a simulation study to explore the relationship between ρ and ρ_M. Based on this numerical study, we made a transformation of r_M and obtained the MP estimator of ρ denoted by $\hat{\rho}_{MP}$.

The new robust estimator $\hat{\rho}_{MP}$ has much better performance compared to classical r in the contaminated data. When compared to existing robust estimator $\hat{\rho}_{MVE}$, the standard error of our estimator is comparable to that of $\hat{\rho}_{MVE}$. Though the bias of our estimator is slightly greater than that of MVE, the proposed MP estimator is computationally more suitable.

References

Adrover, J., M. Salibian-Barrera and R. Zamar. 2004. Globally robust inference for the location and simple linear regression models. *Journal of statistical planning and inference*, **119:** 353-375.

Davies, P. L. 1987. Asymptotic behavior of S-estimators of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, **15**: 1269-1292.

Donoho, D. L. 1982. Breakdown properties of multivariate location estimators. Ph.D. Qualifying paper, Harvard University.

Marazzi, A. 1993. Algorithms, Routines and S Functions for Robust Statistics. Wadsworth and Books/cole.

Maronna, R. A., R. D. Martin and V. J. Yohai. 2006. Robust Statistics, Theory andMethods. Wiley, England, 33-188.

Rousseeuw, P. J. and B. C. Vandriessen. 1999. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41**: 212-223.

Rousseeuw, P. J. and A. M. Leroy. 1987. Robust Regression and Outlier Detection.Wiley.

Rousseeuw, P. J., and B. C. Vanzomeren. 1990. Unmasking multivariate outliers andleverage points, *Journal of the American Statistical Association*, 85: 633-639.

Stahel, W. A. 1981. Breakdown of covariance estimators, Research Report 31, Fachgruppe fur Statistik, ETH. Zurich.

(Manuscript received on 10 September, 2011; revised on 10 January)