Breakdown Analysis of Pearson Correlation Coefficient and Robust Correlation Methods

Friday Zinzendoff Okwonu^1^2, Bolaji Laro Asaju^3^ and Festus Irimisode Arunaye^2^

^1^School of Quantitative Sciences, College of Arts and Sciences, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia
^2^Department of Mathematics, Faculty of Science, Delta State University, Abraka.
^3^Department of Computer Science, Federal University Wukari, Wukari , Taraba State, Nigeria

Abstract: This paper discussed plug in robust procedure to robustify the Pearson product moment correlation coefficient (PPMCC). The mean of PPMCC is highly susceptible to influential observations hence the PPMCC is not robust against data set that contains substantial amount of influential observations. This study focused on robust plug in techniques with high breakdown points. The performance of these techniques are compared using real and simulated data set. The comparative analysis indicates different degrees of robustness and breakdown based on the percentage of contamination and data modification.

Keywords: Correlation coefficient; Robust mean; Mahalanobis distance; Influential observations

1. Introduction
The standard computational correlation coefficient technique was advanced by Bravais in 1846 and illustrated by Pearson in 1896 [1]. In the forgoing, the method was named after Karl Pearson. Pearson called the method “Product moment” hence it is known as Pearson product moment correlation coefficient. This method describes relationship between two variables [2]. However, the two variable data points determine the direction or sign of the correlation coefficient (r) value. The population correlation coefficient is often denoted by rho (ρ) and it’s sample estimate is simply (r), its values ranges from plus one to minus one (r = ±1). The performance of the Pearson correlation technique is affected by influential observations. A single influential observation can reduce strong positive correlation value to relatively strong correlation value [3-5]. In [6] the correlation coefficient value for the income and blood pressure data set (n = 120) was 0.88 and a single influential observation was introduced to the income data, the correlation value reduced to approximately 0.7.
The Pearson’s product moment correlation coefficient (PPMCC) is susceptible to influential observations [7]. Experimentally, influential observations introduced into the data set can reduce the correlation value and change the sign from positive to negative and otherwise. This paper focused on robust techniques to robustify the PPMCC against influential observations. Procedure for transforming influential observations has been discussed in [8]. Influential observation can be present in the dependent or independent variable or in some cases in both variables. The focus of this paper is to robustify the Pearson product moment correlation coefficient using two robustification techniques based on plug-in. The first robust procedure apply weight function and delete the influential observations while the other transforms the influential observations into inliers. The breakdown points of these methods are also investigated.

The reminder of this paper is described as follows. Section 2 describes the Pearson and Spearman methods while section 3 contains the proposed robust methods. Results and conclusions follow in section 4 and section 5.

2. Conventional Methods

2.1 Pearson’s Product Moment Correlation Coefficient (PPMCC)

Let \( x_j, y_j, j = 1, 2, \ldots, k \) be the observed bivariate data points. Then the Pearson product moment correlation coefficient [9] is stated mathematically as

\[
r_{\text{ppmcc}} = \frac{\sum_{j=1}^{k} (x_j - \bar{x})(y_j - \bar{y})}{\sqrt{\sum_{j=1}^{k} (x_j - \bar{x})^2 \sum_{j=1}^{k} (y_j - \bar{y})^2}}
\]

where \( \bar{x} = \frac{\sum_{j=1}^{k} x_j}{k} \) and \( \bar{y} = \frac{\sum_{j=1}^{k} y_j}{k} \) are the sample means. Note that \((x_j, y_j, j = 1, \ldots, n)\) are data points from two variables assumed to be normally distributed with parameters \( \mu_x, \mu_y, \delta^2_x, \delta^2_y, \rho \) that is,

\[
N\left(x, y, \mu_x, \mu_y, \delta^2_x, \delta^2_y, \rho\right) = \frac{1}{2\pi \sqrt{\delta^2_x \delta^2_y (1-\rho^2)}} \exp\left\{-\frac{1}{2(1-\rho^2)} \left[ \frac{(x-\mu_x)^2}{\delta^2_x} - 2\rho \frac{(x-\mu_x)(y-\mu_y)}{\delta^2_x \delta^2_y} + \frac{(y-\mu_y)^2}{\delta^2_y} \right] \right\}
\]

The mean and variance of \( x \) are denoted by \( \mu_x, \delta^2_x \) and the mean and variance of \( y \) are denoted by \( \mu_y, \delta^2_y \) [10-11]. The correlation coefficient between the two variables is denoted by \( \rho \). From Equation(1), \( r_{\text{ppmcc}} \) is the sample estimate of \( \rho \). The mean of the bivariate data set is not robust against outliers as such the value of the correlation coefficient may lead to bias analysis.

2.2 Spearman Rank Correlation Coefficient (SRCC)

This is a nonparametric correlation coefficient denoted by \( r_{SRCC} \). This technique focused on the ranks of the data set [1]. This method can be stated as

\[
r_{SRCC} = 1 - \frac{6 \sum_{j=1}^{k} (r_{yj} - r_{xj})^2}{k(k^2 - 1)} = 1 - \frac{6\omega^2}{k(k^2 - 1)}
\]
Where $r_{yj}, r_{xj}$ denotes the ranks of the bivariate data points or simply dependent and independent variables.

3. Robust methods

The concept of robust statistics was coined from early research work reported in [12]. The basic concept of robust procedure is to reduce the influence of outliers either by deleting the data point or weight the data set by using different weighting schemes [13-14]. This section focusses on robust correlation coefficients based on weighting function and transformation of outliers.

Definition 3. 1(Breakdown point)

The breakdown point is the minimum proportion of influential observations(outliers) that can cause an estimator to take arbitrary values.

3.1 Weighted moment correlation coefficient (WMCC)

Consider $y_j = \beta_0 + \sum_{i=1}^{k} \hat{\beta_i} x_{ji}$

Such that it minimizes $\text{med}_{j} r_j^2$ where

$$r_j = y_j - \sum_{i=1}^{l} \hat{\beta_i} x_{ji} \tag{3}$$

Equation (3) is the residual of the simple linear regression. For the purpose of weighting, the following definition [15] is vital

$$\min_{\beta} \sum_{j=1}^{k} w_{j} r_{j}^2$$

The weight function $w_j$ is defined as

$$w_j = \begin{cases} 1, & \text{abs}\left( \frac{r_j}{\Delta} \right) \leq \frac{s}{2} \\ 0, & \text{abs}\left( \frac{r_j}{\Delta} \right) > \frac{s}{2} \end{cases} \tag{4}$$

where $\Delta = 1.4826 \left( 1 + \frac{s}{(k-1)} \right) \text{sqrt}(\text{med}_{j} r_j^2)$.

Based on equation (4) the weighted moment correlation coefficient (WMCC) is computed as

$$r_{\text{wmcc}} = \frac{\sum_{j=1}^{k} (w_j y_j - \bar{y}_{wj}) (w_j x_j - \bar{x}_{wj})}{\sqrt{\sum_{j=1}^{k} (w_j x_j - \bar{x}_{wj})^2 \sum_{j=1}^{k} (w_j y_j - \bar{y}_{wj})^2}} \tag{5}$$

Where $\bar{x}_{wj} = \frac{\sum_{j=1}^{k} w_j x_j}{\sum_{j=1}^{k} w_j}$ and $\bar{y}_{wj} = \frac{\sum_{j=1}^{k} w_j y_j}{\sum_{j=1}^{k} w_j}$ are the weighted mean respectively.
3.2 Weighted Mean Based on Mahalanobis Distance

The Mahalanobis distance is well studied and the distance is defined as

\[ M_c = \sqrt{(x_j - \bar{x}_p)S^{-1}(y_j - \bar{y}_p)}, \]  

(6)

where \( S = \frac{1}{k-1} \sum_{i=1}^{k} (x_i - \bar{x}) (y_i - \bar{y}) \), and \( M_c = \sqrt{\chi^2_\beta} \), \( \beta = 0.975 \), and \( \chi^2_\beta \) comes from the Chi square distribution[10]. The weight function based on the data set can be computed as

\[ w_{mj} = \begin{cases} 1, & \text{if } M_c \leq \sqrt{\chi^2_\beta}, \text{(inliers)} \\ 0, & \text{if } M_c \geq \sqrt{\chi^2_\beta}, \text{(outliers)} \end{cases} \]  

(7)

Equation (7) help to detect good date points(inliers) and outliers by assigning one and zero to the data points based on this, the weighted mean using the Mahalanobis weight function is computed as

\[ \bar{x}_{wmj} = \frac{\sum_{j=1}^{k} w_{mj} x_j}{\sum_{j=1}^{k} w_{mj}}, \quad \text{and} \quad \bar{y}_{wmj} = \frac{\sum_{j=1}^{k} w_{mj} y_j}{\sum_{j=1}^{k} w_{mj}}, \]  

(8)

The above procedure discards the outliers and compute the weighted mean based on the good data points. The weighted Mahalanobis product moment correlation coefficient is computed as follows:

\[ r_{mah} = \frac{\sum_{j=1}^{k} (w_{mj} x_j - \bar{x}_{wmj})(w_{mj} y_j - \bar{y}_{wmj})}{\sqrt{\sum_{j=1}^{k} (w_{mj} x_j - \bar{x}_{wmj})^2} \sqrt{\sum_{j=1}^{k} (w_{mj} y_j - \bar{y}_{wmj})^2}}. \]  

(9)

3.3 Modified Weighted Approach

Recall Equation (7) which identifies and delete outliers, this equation is modified such that the weight function can identify both the inliers and outliers. However, this approach does not delete the outliers but transform the data point from outliers to inliers. The weight function is defined as follows [16]

\[ W_{mij} = \begin{cases} 1, & \text{if } M_k \leq \sqrt{\chi^2_\beta} \beta, \\ 0 + \left(1 - \frac{\alpha p}{n+p} \right), & \text{if } M_k > \sqrt{\chi^2_\beta} \beta, \alpha = 1.5 \end{cases} \]  

(10)

Applying Equation (10), the reweighed means can be computed as follows:

\[ \tilde{x}_{wmj} = \frac{\sum_{j=1}^{k} W_{mij} x_j}{\sum_{j=1}^{k} W_{mij}}, \quad \text{and} \quad \tilde{y}_{wmj} = \frac{\sum_{j=1}^{k} W_{mij} y_j}{\sum_{j=1}^{k} W_{mij}}. \]  

This procedure involves two stages of weighting, the first identifies the outliers and the second stage use the information from stage one to transform the outliers to inliers while other data points are intact. The weight assigned to the outliers depends on the sample size(n) and the dimension(p) of the data set [16]. The modified weighted correlation coefficient is given as
The minimum covariance determinant (MCD) estimator is a well-known high breakdown robust estimator of location proposed by Rousseeuw and Van Driessen [17-20]. The MCD is affine equivariant. This plug-in technique has been applied extensively in diverse statistical methods [21]. Extensive discussion of the MCD is reported in [17-19,21-22]. To enhance the efficiency and performance of the MCD, the MCD estimator is reweighted based on the Mahalanobis distance concept. Similar weighting procedure as in Equation (7) is performed. The reweighted means based on the MCD technique are computed as
\[
\bar{x}_{rmcd} = \frac{\sum_{j=1}^{k} w(d_j)x_j}{\sum_{j=1}^{k} w(d_j)} \quad \text{and} \quad \bar{y}_{rmcd} = \frac{\sum_{j=1}^{k} w(d_j)y_j}{\sum_{j=1}^{k} w(d_j)}
\]
Where \( w(d_j) \) denoted the reweighted function similar to Equation (7). The robust, high breakdown and affine equivariant correlation coefficient is stated as
\[
r_{mcd} = \frac{\sum_{j=1}^{k} (w_{rmcd}x_j - \bar{x}_{rmcd})(w_{rmcd}y_j - \bar{y}_{rmcd})}{\sqrt{\sum_{j=1}^{k} (w_{rmcd}x_j - \bar{x}_{rmcd})^2 \sum_{j=1}^{k} (w_{rmcd}y_j - \bar{y}_{rmcd})^2}}
\]

4. Results and Discussions
The data collection consists of two real data set obtain via google forms survey; the first data set consist of 40 respondents on time and cost of grab transport services from source to destinations within Malaysia. The second data set consist of sixty respondents on income and blood pressure. They willingly gave out the information required for this study.

For the first category, the time in minutes and cost (RM) was obtained. The objective is to investigate if time and cost correlate positively or negatively. The data set was applied to the different methods described above. Table 1 contains the different correlation values.

| Methods | \( r_{ppmcc} \) | \( r_{scc} \) | \( r_{wncc} \) | \( r_{mah} \) | \( r_w \) | \( r_{mcd} \) |
|---------|----------------|----------------|----------------|------------|--------|--------|
| Correlation values | 0.94 | 0.93 | 0.94 | 0.67 | 0.94 | 0.94 |

Based on the transport data set, all the methods performed comparable except the correlation approached based on Mahalanobis distance.

The second category of data set was collected to investigate whether income and blood pressure correlate positively or otherwise. Table 2 contain the different correlation values obtained based on the methods.
From Table 2, we observed the different values with respect to blood pressure and income, the respective method showed different performance rating. In general, the performance is comparable except for two methods. Having investigated these methods based on two real data set, we want to determine how robust these methods are when the data set contains outliers. To achieve this, we apply the contaminated normal model [23-25]

\[ x = (1 - \varepsilon)N(0,1) + \varepsilon N(\mu, \sigma) \]  

(13)

This model is used to obtain the dependent classification as follows

\[ y = 2.0 + 1.0x + u \]  

(14)

The choice of the mean and standard deviation in Equation (13) is user dependent depending on the degrees of contamination. The sample size for this study is \( n = 40 \). For consistency and stability of results, it will be replicated 1000 times. The contamination percentage ranges from \( \varepsilon = 10\% \), \( \varepsilon = 20\% \), \( \varepsilon = 30\% \), \( \varepsilon = 40\% \), \( \varepsilon = 50\% \). The mean of contamination is \( \mu = 3 \) and standard deviation of contamination is \( \sigma = 3 \). The results based on the simulation is contained in Figure 1.

![Comparative analysis of breakdown points of the conventional and robust correlation methods](image_url)

**Figure 1.** Breakdown points of conventional and robust correlation methods

Based on Figure 1 and the contamination model used in this study, the Spearman correlation is robust, the other methods start breaking-down at 30% contamination, this implies that these methods can only
withstands 30% outliers whereas the Spearman method can resist up to 40%. The Mahalanobis approach is not robust based on the contamination parameters used in this study.

5. Conclusion

From the analysis based on the real data set, the grab transportation system has strong positive correlation in-terms of the two variables of interest studied. These methods showed that income and blood pressure correlate positively. For the simulated data set, for zero percent contamination all the techniques have strong positive correlation except the Mahalanobis approach which remain consistently weak while other techniques have minimum 30% breakdown values and 40% maximum breakdown values. In this study, the proposed techniques performed comparable based on the data set used with the well-known robust techniques in-terms of outlier resistance and breakdown.

References
[1] Spearman, C. (1904). The proof and measurement of association between two things, 15,72-101. The American Journal of Psychology, 100(3/4), special centennial issue (Autumn -Winter,1987), 441-471. DOI: 10.2307/1422689. https://www.jstor.org/stable/1422689.
[2] Niven, E. B. and Deutsch, C.V. (2012). Calculating a robust coefficient and quantifying its uncertainty. Computers & Geosciences 40:1-9. doi.10.1016/j.cageo.2011.06.021.
[3] Abdullah, M.B., (1990). On a robust correlation coefficient. The statistician 39,455-460.
[4] Shevlyakov, G. L., (1997). On robust estimation of a correlation coefficient. Journal of Mathematical Science 83(3),434-438.
[5] Jeongate,K and Fessler, J.A(2004).Intensity based image registration using robust correlation coefficient. IEEE Transactions on medical imaging,23(11):1430-1444.
[6] Okwonu, F.Z., Zahayu Md Y., Apanapudor, J.S., Hasnum, N.I. and Abidin, N.A.Z. The Effects of Income Level on Body Mass Index and Blood Pressure Based on Shewart Control Charts Analysis, Submitted, 2020.
[7] Shafiullah, A.Z.M. and Khan, J.A. (2011). A new robust correlation estimator for bivariate data. Bangladesh Journal of Scientific Research 24(2):97-106.https://doi.org/10.3329/bjsr.v24i2.10766.
[8] Gideon, R. A. and Hollister, R.A. (1987). A rank correlation coefficient resistant to outliers. Journal of the American Statistical Association 82(398), 656-666.
[9] Pearson,K.(1920). Notes on the history of correlation. Biometrika 13(1),25-45.
[10] Shevlyakov, G. and Smirnov, P. (2011). Robust estimation of the correlation coefficient: an attempt of survey. Austrian Journal of Statistics (1&2),147-156.
[11] Kendall, M.G. and Stuart, A. (1963). The advanced theory of statistics. Inference and relationship. Vol. 2, Charles Griffin & Company, London.
[12] Box, G.E.P. (1953). Non-normality and test on variances, Biometrika,40 (3/4):318-335. https://www.jstor.org/stable/pdf/2333350.pdf
[13] Rousseeuw, P. J. and Zomeren, B.C. (1990). Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association 85(411),633-639.
[14] Campbell, N. A., (1980). Robust procedures in multivariate analysis 1: robust covariance estimation. Applied Statistics 29(3), 231-237.
[15] Rousseeuw, P.J. and Leroy, A. M. (1987). Robust regression outlier detection. https://onlinelibrary.wiley.com/doi/pdf/10.1002/0471725382.fmatter.
[16] Okwonu, F.Z. and Zahayu, M. Y. (2020). Performance analysis of robust locations estimators. ICSE Conference, 29th Jan-30th Jan. Kalatain Kota Baru, Malaysia.
[17] Hubert, M. and Debruyne, M. (2010). Minimum covariance determinant. Computational Statistics, 2,36-43.

[18] Rousseeuw, P. J., and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. Technometrics, 41(3): 212-223.

[19] Croux, C. and Haesbroeck, G. (1999). Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. Journal of Multivariate Analysis, 71,161-190.

[20] Maronna, R.A. Martin, R.D. and Yohai, V.J. (2006). Robust statistics, theory and methods, Wiley, England,33-188.

[21] Okwonu, F.Z.(2013). Several robust techniques in two groups unbiased linear classification. PhD thesis, Universiti Sains Malaysia.

[22] Pison, G., Van Aelst, S. and Willems, G. (2002). Small sample corrections for LTS and MCD. Metrika, 55,111-123.

[23] Tukey, J.W. (1960). A survey of sampling from contaminated distributions, in Contribution to Probability and Statistics, Olkin1. Ed,Stanford, Stanford University Press,448-485. https://catalog.princeton.edu/catalog/4462259

[24] Ahad, N.A.,Abdollah, S., Zakaria, N.A.and Yahaya, S.S.S.(2017). Median based robust correlation coefficient. Proceedings of the 13th IMT-GT International Conference on Mathematics, Statistics and their Applications. AIP Conference Proceeding, 050002-1-050002-5. http://doi.org/10.1063/1.5012221.

[25] Okwonu, F.Z. and Othman, A.R.,(2012). A model classification technique for linear discriminant analysis for two groups. International Journal of Computer Science Issues,9(3):125-128