

The Performance of Different Correlation Coefficient under Contaminated Bivariate Data

Bahtiar Jamili Zaini*, Shamshuritawati Sharif

School of Quantitative Sciences, Universiti Utara Malaysia, Malaysia

Received July 31, 2019; Revised September 28, 2019; Accepted February 20, 2020

Copyright©2020 by authors, all rights reserved. Authors agree that this article remains permanently open access under the terms of the Creative Commons Attribution License 4.0 International License

Abstract Bivariate data consists of 2 random variables that are obtained from the same population. The relationship between 2 bivariate data can be measured by correlation coefficient. A correlation coefficient computed from the sample data is used to measure the strength and direction of a linear relationship between 2 variables. However, the classical correlation coefficient results are inadequate in the presence of outliers. Therefore, this study focuses on the performance of different correlation coefficient under contaminated bivariate data to determine the strength of their relationships. We compared the performance of 5 types of correlation, which are classical correlations such as Pearson correlation, Spearman correlation and Kendall's Tau correlation with other robust correlations, such as median correlation and median absolute deviation correlation. Results show that when there is no contamination in data, all 5 correlation methods show a strong relationship between 2 random variables. However, under the condition of data contamination, median absolute deviation correlation denotes a strong relationship compared to other methods.

Keywords Bivariate, Contaminated Data, Correlation, Robust Correlation

1. Introduction

Analysis of bivariate data is a statistical method to investigate the relationship between 2 variables. 2 variables, for example, X and Y , are said to relate when the value assumed by one variable affects the distribution of another variable. For a bivariate datum, the input variable or independent variable, x_i , is plotted on the horizontal axis, and the output variable or dependent variable, y_i , is plotted on the vertical axis. Scatter plot can be used to plot all the ordered pairs (x_i, y_i) of bivariate data on a coordinate axis system and can be used to display the relationship between

these 2 variables. Besides scatter plot, correlation coefficient can be used to measure the relationship between 2 variables [1]. There are several types of correlation coefficients, such as Pearson correlation coefficient, Spearman rank correlation coefficient, Kendall's Tau correlation coefficient, and other robust correlation methods. Correlation coefficient is a simple statistical measure of relationship between 2 random variables. The correlation coefficient computed from the sample data measures the strength and direction of a linear relationship between 2 variables [2]. If there is a strong linear relationship between the variables, the value of correlation coefficient will be close to +1 or -1 depending on the direction of that relationship. However, when there is no linear relationship between the variables or only a weak relationship, the value of correlation coefficient will be close to 0. This correlation coefficient can be used as an indicator to the statistical fit of a regression based on their squared value, frequently known as coefficient of determination, used as a measure of the goodness-of-fit of the [2]–[4].

Furthermore, testing the correlation has become an important subject in multivariate analysis. The importance of testing the correlation has been shown in many areas such as economics, financial market, medicine, and social science [5]–[9]. However, there are issues in testing the correlation matrices using classical statistics when some outlier exists in the data set [10], [11]. In the presence of outliers, the Pearson correlation coefficient results will be inadequate. Therefore, it has reduced the capability to measure the strength of the relationship, damage effects on statistical analysis, increase the variance of error and reduce statistical tests power [12], [13].

To overcome the presence of outliers, robust methods can be used as alternative methods in reducing the influence of outliers [10], [14]–[16]. The aim of robust statistic procedure is to address the data without being affected by the outlier in the data and to ensure the stability of statistical inference under deviations from the assumed

distribution model [17][18]. In multivariate setting, the work started by Huber, which is related to robust methods, has been successfully applied to regression, repeated measures, principal components, canonical correlation, discriminant analysis, dimension reduction and correspondence analysis [15].

2. Methodology

In this study, we focus on the robust correlation coefficient performance with contaminated data to ensure that the hypothesis testing of correlation is conducted effectively. Therefore, the ultimate goal of this research is to identify the best correlation coefficient when involving contaminated data using simulation study.

2.1. Simulation Study

Simulation study is used to compare the performance of all correlation coefficients under 2 situations either with or without contaminated data. Then, R programming version 3.4.2 is used to run the simulation data. This study used correlation values, bias and standard errors to conduct performance analysis among correlation methods [12], [19].

We generated 200 datasets based on the linear relation of: $y_i = 6.0 + 2.0x_i + z_i$ where x_i is pseudo-random numbers from the uniform distribution with $U(0,1)$, z_i is drawn from normal distribution with $N(0,1)$. A 200-dataset is generated based on sample sizes of 20, 30, 50 and 100. For each dataset, we generated contaminated data with a percentage of 10%, 20%, 30%, 40% and 50% of the sample size, respectively. Then, we calculated the correlation coefficient value, bias and standard deviation.

2.2. Nature of Data

Figures 1-6 show the examples of 1 out of 200 datasets about the nature of data where we start from no contaminated data to 50% contaminated data. Figure 1 shows that there is a strong positive linear relationship between 2 variables and there is no contaminated data in this dataset. The simple regression for this dataset is $y = 6.3552 + 1.9414x$. We simulated the dataset in Figure 1 and then added 10% contaminated data in the dataset. Some data appeared to be somewhat distant with other data as illustrated in Figure 2. We continued to simulate the dataset and added another 20% until 50% contaminated data in dataset is achieved. Figures 3 - 6 show more data away from the original dataset when we increased the percentage of contaminated data.

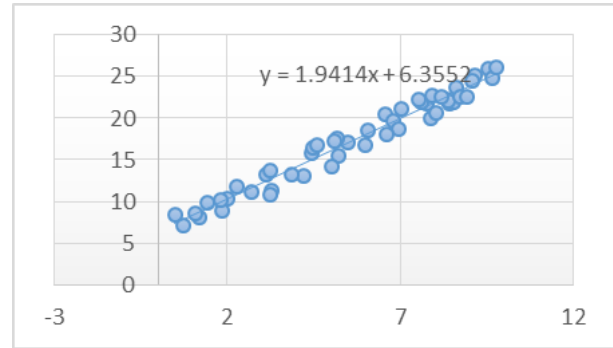


Figure 1. No contamination data

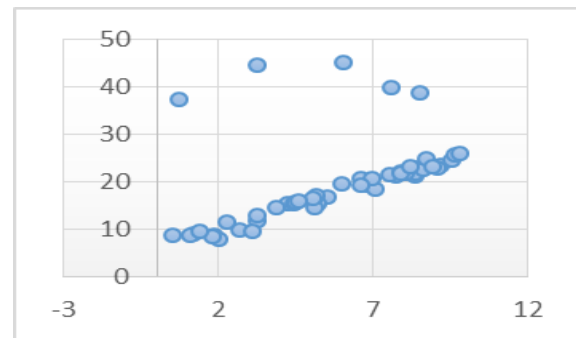


Figure 2. 10% contamination data

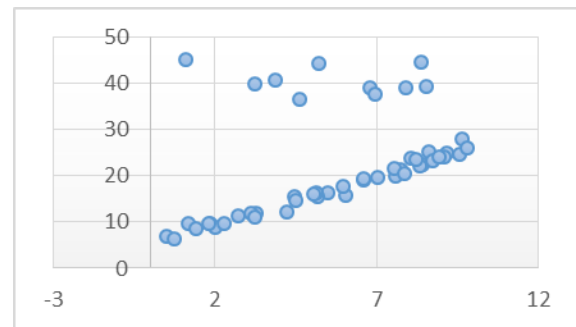


Figure 3. 20% contamination data

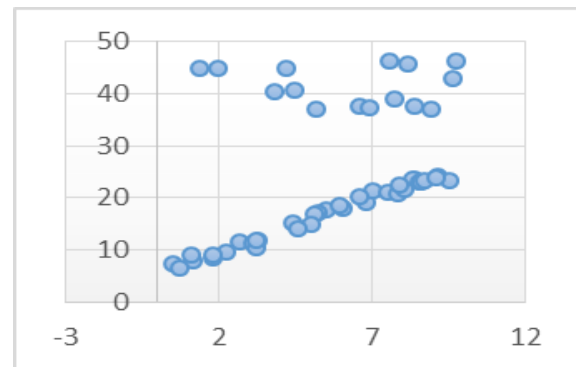


Figure 4. 30% contamination data

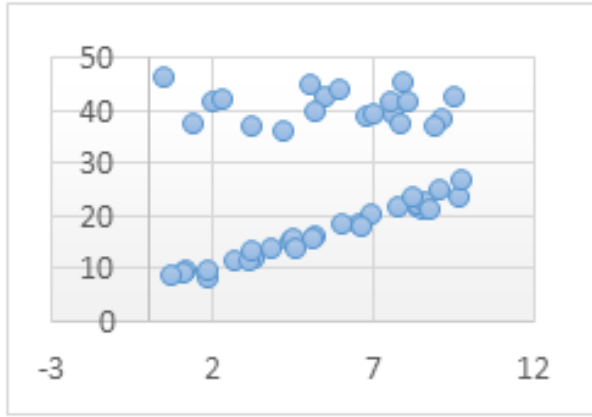


Figure 5. 40% contamination data

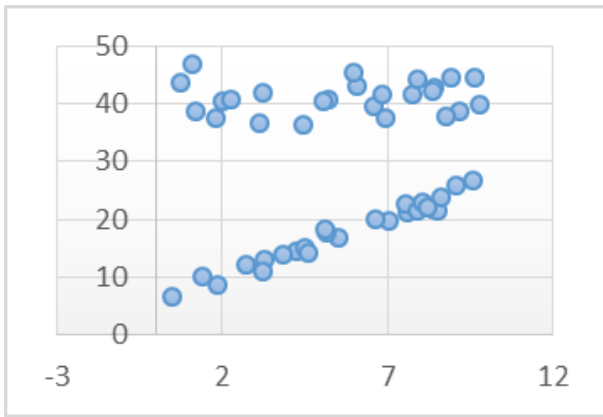


Figure 6. 50% contamination data

2.3. Method of Data Analysis

Generally, we have 6 conditions of contaminated data type (0%, 10%, 20%, 30%, 40%, and 50%), and 4 different sample sizes ($n = 20, 30, 50,$ and 100). Based on that condition, therefore we have 24 sets of 200 generated datasets. The performance evaluation is based on 3 measurements, which are the value of correlation, bias and standard error. 5 correlation coefficients, which are classical correlation, such as Pearson correlation, Spearman correlation and Kendall's Tau correlation with other robust correlations, such as median correlation and median absolute deviation (*MAD*) correlation that will be applied.

2.3.1. Pearson Correlation

There are several types of correlation method to measure relationships among variables, such as Pearson correlation. Pearson correlation coefficient is a classical statistic approach in measuring the relationship between 2 variables, especially for the data that are normally distributed and there is linearity between 2 variables. Let X and Y be drawn from 2 random variables with n sample size, therefore the sample correlation of 2 random variables introduced by Pearson is defined in (1),

$$r_p = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{[\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2]^{1/2}} \quad (1)$$

This Pearson correlation coefficient can be measured when the data is on at least interval scale. However, the Pearson correlation is excessively influenced by outliers, unequal variances, non-normality, and nonlinearity assumptions.

2.3.2. Spearman Correlation

A non-parametric alternative to the Pearson's linear correlation coefficient is Spearman correlation. Usually, the Spearman correlation coefficient can be used when the data is non-normal distributed, measured in ordinal scale, or outliers exist in the data. Technically, the Spearman correlation involved exists ranking of each set of data. Therefore, the Spearman correlation is sometimes called as the Spearman's rank-order correlation. The Spearman's rank-order correlation coefficient measures the strength and direction of association between 2 ranked variables. This correlation coefficient is defined in (2).

$$r_s = 1 - \frac{6 \sum_i^n (d_i)^2}{n(n^2 - 1)} \quad (2)$$

d_i is the difference in paired rankings, $R(X) - R(Y)$, and n is the number of pairs of data. The value of r_s will range from -1 to +1 and will be used in the same manner as the Pearson's linear correlation coefficient, r , was used.

2.3.3. Kendall's Tau Correlation

Similar to the Spearman rank correlation coefficient, Kendall's Tau correlation, τ is based on the ranks of observations. Commonly, the Kendall's Tau correlation is used to measure the degree of correspondence between sets of rankings where the measures are not equidistant. Because it is used with ordinal association data, this correlation is also a non-parametric alternative to the Pearson correlation coefficient. The Kendall's Tau correlation coefficient can be measured in (3).

$$\tau = \frac{n_c - n_d}{n(n-1)/2} \quad (3)$$

where

- n_c is the number of concordant pairs
- n_d is the number of discordant pairs
- n is the number of pairs of data

2.3.4. Median Correlation

Another alternative method that can be used to calculate a correlation between 2 variables is robust correlation approaches. Robust correlation methods could ensure high stability of statistical inference under the deviations from the assumed distribution model [18], [20]. Median correlation is one type of robust correlation coefficient known as the median-product correlation coefficient. As the Pearson correlation refers to the product moment correlation, alternative approaches to calculate the correlation between 2 variables can use median as an

estimator to the mean. This method uses a 2-dimensional sample position determined by the use of median and as an estimator to the mean.

From bivariate data, let (X, Y) pair, we determined the median for X variables, denoted by $Med(x)$, and median for Y variable, denoted by $Med(y)$. Then, the median absolute deviation for variable X denoted by $MAD(x)$ is given by $MAD(x) = Med(|x_i - Med(x)|)$ while the median absolute deviation for variable for variable Y , denoted by $MAD(y)$ is given by $MAD(y) = Med(|y_i - Med(y)|)$. The median covariance for variables X and Y denoted by $Covmed(X, Y)$ is the median determined by given products, which are $Covmed(X, Y) = Med\{(x - Med_x)(y - Med_y)\}$.

In (1), the denominator is replaced by $MAD(x)$ and $MAD(y)$. For the nominator, it is replaced by $Covmed(X, Y)$. Then, the median correlation denoted by r_m , is given by using (4) and (5).

$$r_m = \frac{Covmed(X, Y)}{MAD(x).MAD(y)} \quad (4)$$

or

$$= \frac{Med\{(x-med(x))(y-med(y))\}}{MAD(x).MAD(y)} \quad (5)$$

2.3.5. MAD Correlation

Shevlyakov et al. [21] utilizes the MAD scale estimator to obtain the median correlation coefficient and MAD correlation coefficient. MAD correlation is defined (6) to (8).

$$r_{MAD} = \frac{(MAD^2(u) - MAD^2(v))}{(MAD^2(u) + MAD^2(v))} \quad (6)$$

where

$$u = \frac{x-med(x)}{\sqrt{2}MAD(x)} + \frac{y-med(y)}{\sqrt{2}MAD(y)} \quad (7)$$

$$v = \frac{x-med(x)}{\sqrt{2}MAD(x)} - \frac{Y-med(y)}{\sqrt{2}MAD(y)} \quad (8)$$

3. Results and Discussion

Table 1 shows the results of the performance of Pearson correlation with another robust correlation coefficient. 2

variables have a strong relationship if the value of correlation is close to 1 or -1. For uncontaminated data, which has no outlier, all the correlation methods demonstrate a coefficient value close to original, $\rho = +1$. The Pearson correlation values present the highest values compared to other correlations when the sample size is small, for $n = 20$, $r = 0.986$ and for $n = 30$, $r = 0.985$. For sample size 50 and 100, the MAD correlation gave the highest values $r = 0.99$ for both sample sizes. On the other hand, the Kendall Tau correlation recorded the worst values for all different sample sizes.

The performance of all 5 correlation methods can also be measured by using the average bias and their standard error. Based on 200 generated samples, we calculated the average and standard error of sample correlation. These biases and standard errors for sample correlation are displayed in Table 1. Bias is referred as the difference of average of sample correlation coefficient with true correlation. Therefore, the smaller the value of bias and standard error are, the better the performance of that method is. By referring to Table 1, the Pearson correlation coefficient shows the smallest bias and standard error for small sample size ($n = 20$ and 30) while the MAD correlation coefficient for large sample size ($n = 50$ and 100).

The performances of all 5 correlation methods will be compared with and without contaminated data. Table 2-5 show the correlation coefficient value for all 5 methods for different percentage of contaminated data for sample size, $n = 20, 30, 50$ and 100 respectively. Based on Table 2, for small sample size ($n = 20$), when the contaminated data exist in the dataset, the correlation value for all methods drops compared without contaminated data. It means that the correlation value is highly sensitive to the presence of outliers, which can cause invalid results. For 10% contaminated data existing in the dataset, the MAD correlation gave the highest value of correlation ($r = 0.968$) compared to others. This MAD correlation shows that although outlier exists in the data, the value of correlation is not affected. The value of correlation for the Pearson correlation values ($r = 0.554$) drops more compared to other correlation methods, therefore the Pearson correlation coefficient suffers in the presence of outlier although only 10% outliers exist in the dataset.

Table 1. The Correlation Coefficient without Contaminated Data

Sample Size	Statistic	Correlation Methods				
		Pearson	Spearman	Kendall	Median	MAD
$n = 20$	Mean	0.986	0.971	0.888	0.977	0.981
	Bias	0.014	0.029	0.112	0.023	0.019
	S.E	0.006	0.015	0.036	0.026	0.022
$n = 30$	Mean	0.985	0.978	0.894	0.98	0.984
	Bias	0.015	0.022	0.106	0.02	0.016
	S.E	0.005	0.009	0.026	0.016	0.013
$n = 50$	Mean	0.985	0.984	0.893	0.988	0.99
	Bias	0.015	0.016	0.107	0.012	0.01
	S.E	0.003	0.003	0.011	0.004	0.004
$n = 100$	Mean	0.985	0.984	0.893	0.988	0.99
	Bias	0.015	0.016	0.107	0.012	0.01
	S.E	0.003	0.003	0.011	0.004	0.004

Table 2. The Correlation Coefficient Value for $n = 20$ with different percentage of contaminated data

contaminated data	Correlation Methods				
	Pearson	Spearman	Kendall	Median	MAD
0%	0.986	0.971	0.888	0.977	0.981
10%	0.554	0.791	0.723	0.94	0.968
20%	0.409	0.629	0.572	0.847	0.926
30%	0.315	0.481	0.442	0.625	0.651
40%	0.257	0.354	0.322	0.396	0.342
50%	0.208	0.234	0.215	0.06	0.171

We added another 10% outliers to become 20% outliers in the dataset, the MAD correlation still presents a good value of correlation ($r = 0.926$) compared to others. This MAD correlation shows the best performance compared to other correlation methods although it uses a small sample size. When the data consists of more than 40% contaminated data, the value of correlation for all the methods indicates weak relationships.

The performance of the correlation methods when we increase the number of sample size is presented. Tables 3-5 display the coefficient value of different correlation methods based on different sample sizes, $n = 20, 30, 50$ and 100 , respectively. The performance of large sample size $n = 50$ and $n = 100$, is the same as small sample size, $n = 20$ and $n = 30$. As we can see in Tables 3 - 5, the correlation values for all methods would drop compared to without

outliers in the dataset. The value of MAD correlation still produces the highest correlation coefficients value for different percentage of contaminated data except for 40% contaminated data.

The performance of all 5 correlation methods with contaminated data can also be measured using their average bias and standard error. Based on 200 generated samples, where they contain 10% contaminated data, the value of bias and standard error of correlation methods was calculated as displayed in Table 6. The results show that the MAD correlation has the smallest bias and standard error compared to other methods for all different sample sizes. The median correlations have the second lowest bias and standard error. Meanwhile, the Pearson correlation has the largest bias and standard error for all samples sizes.

Table 3. The Correlation Coefficient Value for $n = 30$ with different percentage of contaminated data

contaminated data	Correlation Methods				
	Pearson	Spearman	Kendall	Median	<i>MAD</i>
0%	0.985	0.978	0.894	0.98	0.984
10%	0.538	0.783	0.714	0.942	0.975
20%	0.398	0.619	0.562	0.843	0.927
30%	0.308	0.47	0.43	0.637	0.657
40%	0.259	0.35	0.318	0.409	0.346
50%	0.211	0.246	0.223	0.096	0.209

Table 4. The Correlation Coefficient Value for $n = 50$ with different percentage of contaminated data

contaminated data	Correlation Methods				
	Pearson	Spearman	Kendall	Median	<i>MAD</i>
0%	0.985	0.984	0.893	0.988	0.99
10%	0.533	0.785	0.714	0.957	0.978
20%	0.409	0.63	0.573	0.867	0.949
30%	0.305	0.467	0.426	0.629	0.655
40%	0.244	0.339	0.311	0.429	0.365
50%	0.224	0.252	0.227	0.079	0.212

Table 5. The Correlation Coefficient Value for $n = 100$ with different percentage of contaminated data

contaminated data	Correlation Methods				
	Pearson	Spearman	Kendall	Median	<i>MAD</i>
0%	0.985	0.984	0.893	0.988	0.99
10%	0.545	0.792	0.719	0.963	0.984
20%	0.399	0.623	0.566	0.873	0.953
30%	0.32	0.48	0.435	0.667	0.688
40%	0.259	0.351	0.318	0.468	0.395
50%	0.203	0.233	0.216	0.044	0.197

Table 6. The Value of Bias and Standard Error of Correlation Methods with 10% Contaminated Data

10% contaminated data		Correlation Methods				
		Pearson	Spearman	Kendall	Median	<i>MAD</i>
$n = 20$	Bias	0.446	0.209	0.277	0.06	0.032
	S.E	0.191	0.121	0.087	0.059	0.039
$n = 30$	Bias	0.462	0.217	0.286	0.058	0.025
	S.E	0.147	0.097	0.071	0.053	0.021
$n = 50$	Bias	0.467	0.215	0.286	0.043	0.022
	S.E	0.122	0.074	0.051	0.034	0.017
$n = 100$	Bias	0.455	0.208	0.281	0.037	0.016
	S.E	0.089	0.057	0.039	0.022	0.008

We continued the performance comparison of correlation methods when the number of contaminated data is increased. Tables 7 – 10 display the results of bias and standard error for 5 correlation methods with different percentage of contaminated data. Based on Table 7, we conclude that the same pattern occurred. The *MAD* correlation has the smallest bias and standard error compared to other methods for all different sample sizes followed by median correlations with the second lowest of bias and standard error. The Pearson correlation still has the largest bias and standard error for all samples sizes.

Table 7. The Value of Bias and Standard Error of Correlation Methods with 20% Contaminated Data

20% contaminated data		Correlation Methods				
		Pearson	Spearman	Kendall	Median	<i>MAD</i>
<i>n</i> =20	Bias	0.591	0.371	0.428	0.153	0.074
	S.E	0.213	0.17	0.121	0.118	0.074
<i>n</i> =30	Bias	0.602	0.381	0.438	0.157	0.073
	S.E	0.161	0.131	0.09	0.105	0.069
<i>n</i> =50	Bias	0.591	0.37	0.427	0.133	0.051
	S.E	0.123	0.099	0.069	0.096	0.037
<i>n</i> =100	Bias	0.601	0.377	0.434	0.127	0.047
	S.E	0.093	0.074	0.05	0.065	0.023

Table 8. The Value of Bias and Standard Error of Correlation Methods with 30% Contaminated Data

30% contaminated data		Correlation Methods				
		Pearson	Spearman	Kendall	Median	<i>MAD</i>
<i>n</i> =20	Bias	0.685	0.519	0.558	0.375	0.349
	S.E	0.205	0.186	0.13	0.212	0.263
<i>n</i> =30	Bias	0.692	0.53	0.57	0.363	0.343
	S.E	0.174	0.154	0.108	0.184	0.217
<i>n</i> =50	Bias	0.695	0.533	0.574	0.371	0.345
	S.E	0.128	0.117	0.081	0.153	0.174
<i>n</i> =100	Bias	0.68	0.52	0.565	0.333	0.312
	S.E	0.091	0.081	0.055	0.104	0.11

Table 9. The Value of Bias and Standard Error of Correlation Methods with 40% Contaminated

40% contaminated data		Correlation Methods				
		Pearson	Spearman	Kendall	Median	<i>MAD</i>
<i>n</i> =20	Bias	0.743	0.646	0.678	0.604	0.658
	S.E	0.232	0.225	0.159	0.313	0.36
<i>n</i> =30	Bias	0.741	0.65	0.682	0.591	0.654
	S.E	0.157	0.151	0.103	0.263	0.305
<i>n</i> =50	Bias	0.756	0.661	0.689	0.571	0.635
	S.E	0.123	0.122	0.084	0.205	0.241
<i>n</i> =100	Bias	0.741	0.649	0.682	0.532	0.605
	S.E	0.079	0.077	0.053	0.15	0.176

Table 10. The Value of Bias and Standard Error of Correlation Methods with 50% Contaminated

50% contaminated data		Correlation Methods				
		Pearson	Spearman	Kendall	Median	<i>MAD</i>
<i>n</i> =20	Bias	0.792	0.766	0.785	0.94	0.829
	S.E	0.232	0.23	0.164	0.401	0.445
<i>n</i> =30	Bias	0.789	0.754	0.777	0.904	0.791
	S.E	0.178	0.176	0.12	0.328	0.353
<i>n</i> =50	Bias	0.776	0.748	0.773	0.921	0.788
	S.E	0.137	0.133	0.091	0.272	0.308
<i>n</i> =100	Bias	0.797	0.767	0.784	0.956	0.803
	S.E	0.092	0.091	0.061	0.208	0.233

For 50% contaminated data existing, all the correlation methods gave a poor result as displayed in Table 10. All correlation methods produced a high value of bias and standard error. The bias is around 0.748 to 0.973. This means that for 50% contaminated data existing, the strength from strong relationship is reduced to no relationships among them.

4. Conclusions

For perfect data with no outlier in the dataset, all 5 correlation methods demonstrated a strong relationship, which is close to +1. However, when outliers exist in the dataset, the *MAD* correlation still gave a strong relationship compared to other methods. This shows that although outliers exist in the data, the value of *MAD* correlation is not affected. The relationship became weak for *MAD* correlation when there is 40% contaminated data. The value of correlation for the Pearson correlation drops more compared to other correlation methods. Therefore, the Pearson correlation coefficient suffers in the presence of outlier. This means that the Pearson correlation value is highly sensitive to the presence of outliers, which can cause invalid results.

Acknowledgements

This research is part of PhD study under Awang Had Salleh Graduate School, Universiti Utara Malaysia, Malaysia.

REFERENCES

- [1] B. Larson, R. and Faber, Elementary Statistics. Picturing the World. Seventh Edition. Pearson Education, 2019.
- [2] A. G. Asuero, A. Sayago, and A. G. González, The correlation coefficient: An overview, *Critical Review in Analytical Chemistry*, vol. 36, no. 1, pp. 41–59, 2006.
- [3] T. Mendenhall, W. and Sincich, Regression Analysis. A Second Course in Statistics. Eighth Edition. Pearson Education, 2019.
- [4] G. C. Montgomery, D. C. and Runger, Applied Statistics and Probability for Engineers. Sixth Edition. John Wiley & Sons, 2013.
- [5] A. Gupta, B. Johnson, and D. Nagar, Testing Equality of Several Correlation Matrices, *Revista Colombiana Estadística*, vol. 36, no. 2, pp. 237–258, 2013.
- [6] M. A. Djauhari and E. T. Herdiani, Monitoring the Stability of Correlation Structure in Drive Rib Production Process: An MSPC Approach, *Open Industrial Manufacturing Engineering Journal*, vol. 1, no. 1, pp. 8–18, 2008.
- [7] N. S. Yusoff and M. A. Djauhari, Regional Annual Fundamental Science Symposium Understanding the Shift of Correlation Structure in Major Currency, in *Regional Annual Fundamental Science Symposium*, 2010.
- [8] M. A. Djauhari and S. L. Gan, Dynamics of Correlation Structure in Stock Market, *Entropy*, vol. 16, no. 1, pp. 455–470, 2014.
- [9] F. Li, Expression and Correlation of miR-124 and miR-126 in Breast Cancer, *Oncology Letters*, vol. 17, no. 6, pp. 5115–5119, 2019.
- [10] R. R. Wilcox, D. A. Granger, and F. Clark, Modern Robust Statistical Methods: Basics with Illustrations Using Psychobiological Data, *Universal Journal of Psychology*, vol. 1, no. 2, pp. 21–31, 2013.
- [11] S. Sharif and T. A. M. Atiany, Testing Several Correlation Matrices using Robust Approach, *Asian Journal of Scientific Research*, vol. 11, no. 1, pp. 84–95, 2017.
- [12] N. A. Ahad, N. A. Zakaria, S. Abdullah, S. S. S. Yahaya, and N. Yusof, Robust Correlation Procedure via Sn Estimator, *Journal of Telecommunication and Computer Engineering*, vol. 10, no. 1–10, pp. 115–118, 2018.
- [13] C. R. Pernet, R. Wilcox, and G. A. Rousselet, Robust Correlation Analyses: False Positive and Power Validation using a New Open Source Matlab Toolbox, *Frontier in Psychology*, vol. 3, no. Jan 2013, 2013.
- [14] A. Z. M. Shafiullah and J. A. Khan, A New Robust Correlation Estimator for Bivariate Data, *Bangladesh Journal of Scientific Research*, vol. 24, no. 2, pp. 97–106, 2011.
- [15] P. M. Yuan, K.-H., & Bentler, Effect of Outliers on Estimators and Tests in Covariance Structure, *British Journal of Mathematical and Statistical Psychology*, vol. 54, no. 1, pp. 161–175, 2001.
- [16] A. R. Yusof, Z. M., Abdullah, S., Yahaya, S.S.S., & Othman, A Robust Alternative to the t – Test., *Modern Applied Science.*, vol. 6, no. 5, 2012.
- [17] R. R. Wilcox, Introduction to Robust Estimation and Hypothesis Testing, 3rd Edition. Oxford Academic Press, 2012.
- [18] G. L. Shevlyakov and P. Smirnov, Robust Estimation of the Correlation Coefficient: An Attempt of Survey, *Austrian Journal of Statistics*, vol. 40, no. 1, pp. 147–156, 2011.
- [19] M. Abdullah, On a Robust Correlation Coefficient, *Journal of Royal Statistical Society. Ser. D (The Statistician)*, vol. 39, no. 4, pp. 455–460, 1990.
- [20] P. J. Huber, John W. Tukey’s Contributions to Robust Statistics, *The Annals of Statistics*, vol. 30, no. 6, pp. 1640–1648, 2019.
- [21] G. L. Shevlyakov, P. O. Smirnov, V. I. Shin, and K. Kim, Asymptotically Minimax Bias Estimation of the Correlation Coefficient for Bivariate Independent Component Distributions, *Journal of Multivariate Analysis*, vol. 111, pp. 59–65, 2012.