ISSN 1022-8594 © 2019 Jahangirnagar University Journal of Science
JUJS Vol. 42, No. 2, pp.45-61

Classical and Robust Correlation Estimates: A Comparative Study

Salma Akter*, Md. Siddiqur Rahman and Mst. Moushumi Pervin
Department of Statistics, Jagannath University, Dhaka, Bangladesh

Abstract

For conducting correlational analysis, one may be faced with statistical problem if the dataset contains outliers and other contaminations. In order to use any correlation method when the dataset contains outliers, one must know the stability of that correlation method. For this reason, it is essential to check the stability of the considered classical correlations and robust correlations so that we can use those correlations without any doubt in our mind. In this paper, several contamination cases have been introduced for checking the stability of the classical and robust correlations that are considered. The performances of robust correlations and classical correlations are compared and also performances of robust correlations among each other are compared through a simulation study and real data examples both for study clean data and contaminated data. Based on simulation study and real data applications, robust correlations have much better performance compared to classical correlations from clean to contaminated data. Among the robust correlations adjusted bivariate winsorized correlation has more stability than any other robust correlations. When some observations in real dataset are replaced by outliers, classical correlation i.e. Pearson correlation is changed drastically and robust correlations remain almost same.

Keywords: Pearson correlation, Point biserial correlation, Spearman's rank correlation, Kendall's tau, Maronna's correlation, univariate winsorized, bivariate winsorized, adjusted bivariate winsorized, quadrant correlation.

Introduction

In data analysis, the association between two variables is often of interest. Correlation analysis is one of the most widely studied techniques in probability and statistics to measure the association. To study the presence or absence of association between variables of interest, we need to use appropriate techniques. A large variety of techniques have been applied in the literature to assess the correlation. Among these techniques, Pearson and point biserial correlations are known as classical correlations and Spearman's rank, Kendall's Tau, Maronna's correlation, univariate

*Author for correspondence e-mail: salmastat@gmail.com

winsorized, bivariate winsorized, adjusted bivariate winsorized and quadrant correlations are known as robust correlations.

Among the classical correlations, Pearson correlation is the most popular measure of association and easy for calculation. A recent publication [1] illustrates that Pearson's correlation coefficient can often offer an effective description of linear association. Another classical correlation i.e. point biserial correlation is based directly on the Pearson correlation coefficient which also shows linear relationship between two variables. However, sometimes classical correlations might not be applicable when the relationship is not linear. Another problem is the presence of outliers and other contaminations in the datasets. Classical correlations are also affected by these outliers, and they could fail to provide reliable estimates. Under this circumstance, robust correlations are suitable to use as they are not much affected by outliers. Therefore, the main focus of our work is to analyze the amount of changes of classical correlations and robust correlations when outliers and other contaminations are present in the datasets. Puth et al. [1] examined the performance of the two rank order correlations (Spearman and Kendall's tau) for describing the strength of association between two continuous variables and found that both rank coefficients provide only slightly better performance than Pearson. In Ma, Xu, Wang and Chen [2] likewise reported that Spearman and Kendall's tau have more advantage over Pearson when one variable is clean and the other corrupted by a tiny fraction of impulsive noise. Abdullah [3] checked the stability of only three correlations i.e. Pearson's, Spearman's and Kendall's tau against a substantial number of outliers. In this study, the stability of several classical and robust correlations has been checked by giving several percentage of outliers through a simulation study. In continuation to the study, present study is also undertaken to compare robust correlations with each other in presence of outliers.

The remainder of this paper is structured as follows. In section 2, we review classical correlations. In section 3, we review robust correlations and provide numerical complexity of different correlations**.** In section 4, we describe a simulation study for comparing the performance of our robust correlations with each other and with the classical ones. In section 5, percentage variation is presented to compare classical and robust correlations in more detail. In section 6, we compare time complexity of all correlations. In section 7, we describe a real data application, and finally we conclude this article by summarizing main findings in this work in section 8.

2. Classical correlations

2.1 Pearson correlation

Pearson correlation [4] is the most popular measure of correlation for measuring the degree or strength of linear relationship between two numerically valued random variables. Suppose X and Y be two variables with n pairs of observations represented by (x_i, y_i); $i = 1, 2, 3, ...n.$. Then the Pearson correlation is defined as

$$r = \frac{Cov\left(X, Y\right)}{S_x S_y}$$

where S_x and S_y are standard deviations of X and Y respectively and $Cov(X, Y)$ is the covariance between X and Y .

2.2 Point biserial correlation

Point biserial correlation [5, 6] is the measurement of association between a dichotomous and a continuous variable. Suppose we have one continuous variable Y and a dichotomous variable X which takes the two values 0 and 1. If we divide the dataset into two groups, group 1 will receive value '1' on X and group 2 will receive value '0' on X . Then we have n pairs of observations represented by (x_i, y_i); $i = 1,2,3,....n.$ Then Point biserial correlation coefficient is defined as [7],

$$r_{pbs} = \left(\frac{\bar{y}_1 - \bar{y}_0}{s_y}\right)\sqrt{pq}$$

where $s_y = \sqrt{\frac{\sum_i^n \left(y_i - \bar{y}\right)^2}{n}}$

$\bar{y}_1$ = mean value on the continuous variable Y for all data points in group 1

$\bar{y}_0$ = mean value on the continuous variable Y for all data points in group 2

p = proportion of group 1 form of the total

$q = 1 - p$

3. Robust correlations

3.1 Spearman's rank correlation

Spearman's rank correlation coefficient [8] is a popular nonparametric rank correlation which equals the Pearson correlation computed from the ranks of observations. The Spearman's rank correlation, denoted by r_S, is defined as ,

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

where d_i is the difference in ranks of the i^{th} element of each random variable considered.

3.2 Kendall's Tau

Let X and Y be joint random variables where $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be a set of n pairs of observations, then any pair of observations (x_i, y_i) and (x_j, y_j) where $i \neq j$, are said to be concordant when $x_i < x_j$ and $y_i < \mathrm{y}_j$ or when $x_i > x_j$ and $y_i > \mathrm{y}_j$, or when $(x_i - x_j)(y_i - y_j) > 0$. They are said to be discordant when $x_i < x_j$ and $y_i > \mathrm{y}_j$, or when $x_i > x_j$ and $y_i < \mathrm{y}_j$ or when $(x_i - x_j)(y_i - y_j) < 0$. If $x_i = x_j$ and $y_i = y_j$, the pair is neither concordant nor discordant. Then the Kendall's tau, denoted by τ [9] can be computed as

$$\tau = \frac{\sum_i^n \sum_j^n \operatorname{sgn}\left(x_i - x_j\right) \operatorname{sgn}\left(\mathrm{y}_i - y_j\right)}{n(n-1)}$$

where

$$\operatorname{sgn}\left(x_i - x_j\right) = \begin{cases} 1 & if \ \left(x_i - x_j\right) > 0 \\ 0 & if \ \left(x_i - x_j\right) = 0 \\ -1 & if \ \left(x_i - x_j\right) < 0 \end{cases}$$

and

$$\operatorname{sgn}\left(y_i - y_j\right) = \begin{cases} 1 & if \ \left(y_i - y_j\right) > 0 \\ 0 & if \ \left(y_i - y_j\right) = 0 \\ -1 & if \ \left(y_i - y_j\right) < 0 \end{cases}$$

3.3 Maronna's correlation

Most available robust correlation estimators which are computed from d -dimensional data are very time consuming [10]. Robust pairwise approaches [11] are sensitive to two-dimensional outliers. Khan et al. [12] considered a simplified version of the bivariate M-estimate. To obtain robustness against two-dimensional structural outliers a bivariate M–estimator is used which is proposed by Maronna [13] as it is computationally efficient.

Maronna's M-estimate of the location vector t *and positive definite symmetric scatter matrix* V *is defined as the solution of the system of equations:*

$$\frac{1}{n}\sum_i u_1\left(d_i\right)\left(z_i - t\right) = 0$$

$$\frac{1}{n}\sum_i u_2\left(d_i^{\,2}\right)\left(z_i - t\right)\left(z_i - t\right)^t = V$$

where $d_i^{\,2} = \left(z_i - t\right)^t V^{-1}\left(z_i - t\right)$, *and* u_1 *and* u_2 *satisfy a set of general assumptions.*

3.4 Univariate Winsorized correlation

The idea of univariate winsorization of the data is introduced by Huber [11] and suggested that classical correlation coefficients can be calculated from these winsorized data. This approach was re–examined by Alqallaf et al. [14] for the estimation of individual elements of a large–dimension correlation matrix. The transformation $u_i = \psi_c((x_i - med(x_i))/mad(x_i)), i = 1,2,\ldots,n$ is obtained for n univariate observations $x_1, x_2, \ldots, x_n$, where $\psi_c(x) = \min(\max(-c, x), c)$ is called Huber score function with a tuning constant c. This tuning constant is chosen by the user, e.g. $c = 2$ or $c = 2.5$. Though the univariate winsorization approach can be computed very rapidly, it does not take into account the orientation of the bivariate data.

3.5 Bivariate Winsorized correlation

Khan et al. [15] proposed a bivariate winsorization to reduce this problem based on an initial robust bivariate correlation matrix R_0, and a corresponding tolerance ellipse. Outliers are shrunken to the border of this ellipse by using the bivariate transformation $u = \min\left(\sqrt{c/D(x)}, 1\right)x$, with $x = (x_1, x_2)^t$. In this transformation $D(x)$ is the Mahalanobis distance based on R_0 where $D(x) = x^t V^{-1} x$. Here c is a tuning constant that was chosen to be $c = 5.99$, the 95% quantile of the χ_2^2 distribution. Then the classical correlation coefficient of u is said as the robustified correlation of $x = (x_1, x_2)^t$.

3.6 Adjusted Bivariate Winsorized correlation

An appropriate initial correlation matrix R_0 is an essential part for bivariate winsorization. Khan et al. [15] proposed an adjusted winsorization method that is more resistant to bivariate outliers. This method considers quadrants relative to the coordinate-wise medians (which are 0 due to the robust standardization of the data) and uses two tuning c_1 and c_2 for performing univariate winsorization of the data. c_1 is used to winsorize the points lying in the two diagonally opposed quadrants that contain most of the standardized data (called the "major quadrants"). c_2 is used to winsorize the remaining data in the other two quadrants. The initial correlation matrix R_0 is obtained by computing the classical correlation matrix of the adjusted winsorized data. Bivariate outliers are handled much better by adjusted winsorization than univariate winsorization. The outliers are shrunken to the boundary of the larger ellipsoid by using bivariate winsorization and thus appropriately down-weighted so that a robust correlation is obtained. Although the initial adjusted winsorization and the resulting bivariate winsorization are not affine-equivariant, they can be computed very rapidly and can appropriately handle bivariate outliers.

3.7 Quadrant correlation

The quadrant correlation coefficient, r_Q [16] is computed by first centering the data by the co-ordinate wise median, then r_Q equals the frequency of observations in the fourth quadrant

$$r_Q = \frac{1}{n}\sum_{i}^{n} sign\left\{\left(x_i - median_j\left(x_j\right)\right)\left(y_i - median_j\left(y_j\right)\right)\right\}$$

Here, the $sign$ function equals 1 for positive arguments and -1 for negative arguments, and $sign(0) = 0$.

Numerical complexity of different correlations

Univariate winsorized correlation and adjusted bivariate winsorized correlation both can be computed in $O(n \log n)$ time. Also, the bivariate winsorized correlation and Maronna's correlation require $O(n \log n)$ time but Maronna's correlation has a larger multiplication factor depending on the number of iterations required. It should be noted here that Spearman's rank correlation can be computed in $O(n \log n)$ time which is the same as adjusted bivariate winsorized correlation. Though Kendall's tau separately examines each of the $\binom{n}{2}$ (order n^2) pairs of bivariate observations, there is an algorithm that can calculate Kendall's tau in $O(n \log n)$ time [17].

4. A Simulation Study

A simulation study is carried out for comparing the performances of classical correlations and robust correlations with each other. More specifically, this study is also carried out in order to compare the performance of robust correlations with classical correlations. To perform the simulation study true Pearson correlations of 0.50, 0.80, 0.90, -0.50, -0.80, -0.90 are considered. For different true Pearson correlation values 10000 datasets each of size 500 from a bivariate standard normal distribution of (X, Y) are generated as clean data. Then the datasets are contaminated by 5%, 10% and 15% outliers. The datasets are contaminated with mean 50 and standard deviation 1 (for X variable) and with mean 100 and standard deviation 1 (for Y variable). For point biserial correlation, the datasets are contaminated only for continuous variable (i.e. X variable). Here, only for point biserial correlation Y variable is considered as dichotomous variable.

For each simulated data set 5% trimmed mean, median absolute deviation (mad) and standard deviation (SD) of the coefficients are recorded. The median absolute deviation (mad) and standard deviation (SD) are shown in the parentheses.

At first the performances of the classical correlations from clean data to contaminated data for different positive true Pearson correlation values are presented in Table 1 (see Appendix). In all the cases, the mean value of Pearson and point biserial correlation is clearly changing with the percentage of outliers increasing, i.e. for true Pearson correlation value 0.90 the mean value of Pearson correlation changes from 0.8998 to 0.9983 (for outliers 5%) and 0.9992(for outliers 15%). Also mean value of point biserial correlation changes from 0.7177 to 0.0633 (for outliers 5%) and 0.0339 (for 15% outliers). It is clear that Pearson correlation and point biserial correlation are immediately affected by outliers and their mean values move away from the values of clean data as the percentage of outliers increases while point biserial correlation changes very rapidly. These results are confirming that the point biserial correlation performs less well than Pearson correlation for contaminated data. In all other true Pearson correlation cases, the conclusions of the simulation results are similar as Table 1. So the results of other true Pearson correlation cases are not included.

In Table 2 (see Appendix) the performance of robust correlations for different percentage of outliers from clean to contaminated data are presented when true Pearson correlation is 0.90. It shows that Spearman's, Kendall’s tau are much less affected by outliers and Quadrant correlation are slightly more stable than Spearman’s correlation and Kendall’s tau. For 5% outliers, mean values of Spearman, Kendall’s tau and Quadrant correlations changes from 0.8904 to 0.9059, 0.7128 to 0.7384, 0.7119 to 0.7269. It is also observed that the more we increase the percentage of outliers, the performance of Spearman and Kendall’s tau is quite poor than Quadrant correlation. Maronna’s correlation and univariate winsorized correlation remain within bounds for 5% of contamination i.e. mean values of these two correlations changes from 0.8993 to 0.8859 and 0.8973 to 0.9187. It is also seen that univariate winsorized correlation being more resistant to outliers among these two correlations when percentage of outliers is increased (mean values of Maronna’s correlation and univariate winsorized correlation changes from 0.8993 to 0.9993 and 0.8973 to 0.9525 for 15% outliers). During the 10% contamination of the data, the mean values of bivariate winsorized and adjusted bivariate winsorized changes from 0.9009 to 0.8970 and from 0.8997 to 0.8943. From Table 2 it is clear that bivariate winsorized and adjusted bivariate winsorized correlation show the better results in the presence of outliers, while adjusted bivariate winsorized correlation gives more stable and reliable results than all other correlations and it seems to be best of all of the ones we considered. In all other true Pearson correlation cases, the conclusions of the simulation results are similar as Table 2. Hence the results of other true Pearson correlation cases are not included.

The performances of the classical correlations and robust correlations from clean to contaminated data are presented in Table 3 (see Appendix) when true Pearson correlation is 0.90. It reports that classical correlations changes rapidly with only 5% of outliers while almost all the robust correlations remain stable i.e. classical correlation i.e. Pearson correlation changes from 0.8998 to 0.9983 while robust correlations i.e. adjusted bivariate winsorized correlation changes from 0.8997 to 0.8946. At the same time, more we increase the percentage of outliers in the data, more the robust correlations perform well while the performance of classical correlations is quite poor. This confirms the non-robustness of the classical correlations and thus we say that robust correlations tend to be quite stable even with 15% of outliers. When we increase outliers from 5% to 15%, the mean value of classical correlation such as Pearson correlation changes from 0.9983 to 0.9992 when mean value of this correlation without outliers is 0.8998. On the other hand, average value of robust correlation i.e. adjusted bivariate winsorized correlation changes from 0.8946 to 0.8971 when mean value of this correlation without outliers is 0.8997. Thus, we clearly observe that the robust correlations are more stable than classical correlations. In all other true Pearson correlation cases, the conclusions of the simulation results are similar as Table 3. Hence the results of other true Pearson correlation cases are not included.

5. Percentage variation

The percentage variation from clean to contaminated data for true Pearson correlation value 0.90 is presented in Table 4 (see Appendix). From the table, it is clear that the performance of Pearson correlation and point biserial correlation are deteriorating with the percentage of outliers increasing. While quadrant correlation, Spearman's rank correlation and Kendall's tau seem to be slightly affected by outliers. A smaller effect is detected for Maronna and univariate winsorized correlation while Maronna's correlation shows behaviour similar to the Pearson correlation for larger amount of outliers. On the other hand, bivariate and adjusted bivariate winsorized correlations are not at all influenced by outliers. Adjusted bivariate winsorized correlation is quite stable even in large proportion of contamination. It is clear, however that classical correlations perform poorly than robust correlations when data is contaminated with outliers. For all other true Pearson correlation value, we find the same results as Table 4. Hence the results of other true Pearson correlation are not included.

6. Numerical complexity of different correlations

Now, we compare the computation times of different correlations. Figure 1 shows the mean cpu times in seconds (based on 100 replicates) for 5 different sample sizes: 10000, 20000, 30000, 40000 and 50000. This plot demonstrates that Pearson correlation, point biserial correlation and quadrant correlation can be computed faster than univariate winsorized correlation. The results in figure 1 also confirms that calculating the adjusted bivariate winsorized correlation requires slightly more time than univariate winsorized correlation. We also see that bivariate winsorized correlation can be computed faster than Spearman's rank correlation and Maronna's correlation and that the time difference increases with sample size. Here, Maronna's correlation is clearly more time consuming. Moreover, Kendall's tau does not appear in the figure as it has huge computational burden.

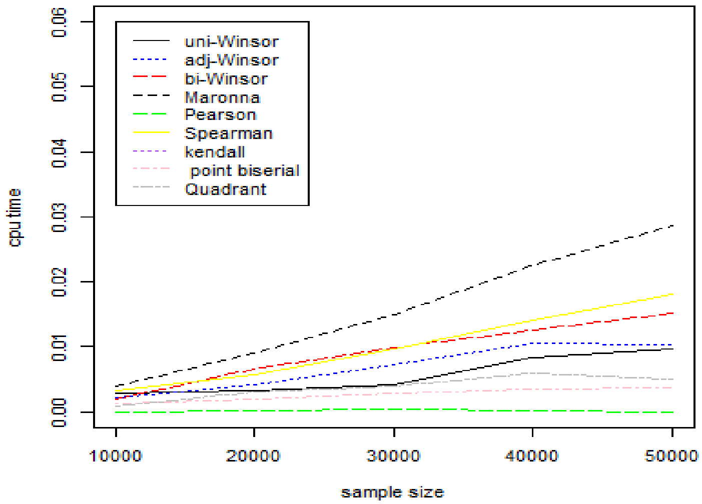

Figure 1: Average computation times for different correlations.

7. Real Data Applications

In this section, a real dataset is used to demonstrate the stability of classical and robust correlations.

Diamonds dataset

This dataset is collected from kaggle.com which was updated by Shivamagrawal (2017). The dataset contains 53940 observations with 10 variables. Here, we only considered two continuous variables. For this reason, we omitted 8 variables from the dataset. Here, the variables which are taken for analysis are carat weight of the diamonds and price of the diamonds.

In this paper our concern is to check the stability of correlations, so there is no dependency and independency relationship between these two variables.

Here, we consider the variable price as X and carat weight as Y. Classical and robust correlations are applied to this dataset and Table 5 represents the resulting values of all correlations.

Now, the data set are contaminated by replacing one small value of variable X (say 230th value 2783) by a large value 278300 and the corresponding value of Y variable (0.52) by a large value 5200. When classical correlations i.e. Pearson and point biserial correlation are applied to the contaminated data, they are immediately affected by outliers. On the other hand, almost all the robust correlations remain stable. For example, values of Pearson correlation changes from 0.9215913 to 0.3025743 after replacing one small value by a large value while robust correlation i.e. univariate winsorized correlation changes from 0.9475939 to 0.9475971. Thus, we clearly observe that the robust correlations are more stable than classical correlations. Among the robust correlations, bivariate winsorized correlation and adjusted bivariate winsorized correlation are more stable as values of both correlation changes from 1 to 0.9999999 where 0.9999999 is almost 1. However, adjusted bivariate winsorized correlation and bivariate winsorized correlation outperform in almost all of the robust correlations. These results are shown in Table 5.

Table 5: Results for diamonds dataset

Correlation		Data	
		Clean	Contaminated
Classical	Pearson Correlation	0.9215913	0.3025743
	Point biserial correlation	0.7697655	0.7396463
Robust	Spearman's rank correlation	0.9628828	0.9628867
	Kendall's tau	0.8341049	0.8341168
	Maronna's correlation	0.9209345	0.9209242
	Univariate Winsorized	0.9475939	0.9475971
	Bivariate Winsorized	1	0.9999999
	Adjusted Bivariate Winsorized	1	0.9999999
	Quadrant correlation	0.889266	0.8893029

8. Conclusion

In this study, the stability of several classical correlations and robust correlations are examined in the presence of outliers or other contamination in the datasets. For this reason a comparison is made among classical correlations and robust correlations using simulated datasets and real datasets. In simulated datasets, we compared the performance of robust correlations and classical correlations by contaminating the simulated datasets. To get the performances of classical and robust correlations in more details percentage variations are compared. Robust correlations have performed much better than classical correlations. As we increased the percentage of outliers in the simulated datasets, the robust correlations have performed much better than the classical correlations. Among the robust correlations we observed that adjusted bivariate winsorized correlation yielded the most efficient correlation in almost all the contamination cases. We also compared the computation times of classical correlations and robust correlations. Though classical correlations required less time for computation but these correlations were unstable and could not give reliable results. On the other hand, robust correlations required slightly more time for computation than classical correlations but they were much more stable than classical correlations. In real datasets, when we replaced some observations by outliers, classical correlations change clearly and give poor results. On the other hand, robust correlations remain stable and give reliable results. From the simulation study and real dataset, it can be summarized that robust correlations are the best performer than classical correlations i.e. robust correlations remained more stable than classical correlations in the contaminated data.

Limitations and Further study

We have seen that the performance of robust correlations is better than classical correlations both in simulation study and real data applications. But, sometimes robust correlations are not sufficiently robust when the percentage of outliers increases in the dataset. In the simulation study, we observed that robust correlations show a behavior similar to the classical correlations for larger amount of outliers. Another drawback of the robust correlations is that most of the robust correlations are time consuming. In this case, we need such robust correlations that have good performance and less numerical complexity at the same time. For this reason, we should try to reduce numerical complexity of the robust correlations.

References

[1] M.T. Puth, M. Neuhäuser and G. D. Ruxton, Effective use of Pearson's product-moment correlation coefficient. *Animal Behaviour*, 2014, vol. 93, pp. 183-189.

[2] R. Ma, W. Xu, Q. Wang and W. Chen, Robustness analysis of three classical correlation coefficients under contaminated Gaussian Model. *Signal Processing*, 2014, vol. 104, pp. 51-58.

[3] M. Abdullah, On a Robust Correlation Coefficient. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 1990, vol.39, pp. 455-460.

[4] K. Pearson, Mathematical contributions to the theory of evolution. III. *Regression, heredity, and panmixia. Philosophical Transactions of the Royal Society, Series A*, 1896, vol. 187, pp. 253–318.

[5] S. Das Gupta, Point biserial correlation coefficient and its generalization. *Psychometrika*, 1960, vol. 25, pp. 393-408.

[6] J. P. Guilford, *Psychometric methods.* (2nd ed.) New York: McGraw-Hill. 1954.

[7] D. Sheskin, Handbook of Parametric and Non Parametric Statistical Procedure (5th ed.). 2011. Boca Raton, FL: CRC Press.

[8] C.E. Spearman, The proof and measurement of association between two things. *American Journal of Psychology*, 1904, vol. 15, pp. 72-101.

[9] M. G. Kendall, A new measure of rank correlation. *Biometrika* , 1938, vol. 30, pp. 81-93.

[10] P. J. Rousseeuw and A. M. Leroy, Robust Regression and Outlier Detection. Wiley-Interscience, New York. 1987.

[11] P. J. Huber, Robust Statistics. Wiley, New York. 1981.

[12] J. A. Khan, S. Van Aelst and R. H. Zamar, Building a robust linear model with forward selection and stepwise procedures. *Computational Statistics of Data Analysis*, 2007(a), 52(1), 239-248.

[13] R.A. Maronna, Robust m-estimators of multivariate location and scatter. *The Annals of Statistics*, 1976, vol.4, pp. 51-67.

[14] F. A. Alqalla, K. P. Konis, R. D. Martin, and R. H. Zamar, Scalable Robust Covariance and Correlation Estimates for Data Mining. *Proceedings of the*

Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, 2002, pp. 14-23.

[15] J. A. Khan, S. Van Aelst and R. H. Zamar, Robust Linear Model Selection Based on Least Angle Regression. *J. Amer. Statist. Assoc*, 2007(b), vol. 102, pp.1289-1299.

[16] F. Mosteller, On some useful inefficient statistics. *The Annals of Mathematical Statistics*, 1946, 17-377.

[17] W.R. Knight, A computer method calculating Kendall's tau with ungrouped data. *Journal of the American Statistical Association*, 1966, 61, 436–439.

Appendix

Table 1: Performance of the classical correlations from clean data to contaminated data for different Positive true Pearson Correlation

True Pearson correlation values	Correlation	Data			
		Clean	Contaminated (outliers)		
			5%	10%	15%
0.50	Pearson	0.4991 (0.0337) (0.0336)	0.9967 (0.0006) (0.0007)	0.9982 (0.0002) (0.0002)	0.9987 (0.0001) (0.0001)
	Point biserial	0.3982 (0.0358) (0.0356)	0.0353 (0.0455) (0.0445)	0.0234 (0.0439) (0.0443)	0.0183 (0.0444) (0.0445)
0.80	Pearson	0.7996 (0.0160) (0.0161)	0.9979 (0.0004) (0.0004)	0.9988 (0.0001) (0.0002)	0.9991 (0.0001) (0.0001)
	Point biserial	0.6377 (0.0217) (0.0217)	0.0565 (0.0456) (0.0448)	0.0384 (0.0447) (0.0444)	0.0303 (0.0436) (0.0445)
0.90	Pearson	0.8998 (0.0084) (0.0085)	0.9983 (0.0003) (0.0003)	0.9990 (0.0001) (0.0001)	0.9992 (0.0001) (0.0001)
	Point biserial	0.7177 (0.0160) (0.0159)	0.0633 (0.0457) (0.0451)	0.0431 (0.0448) (0.0445)	0.0339 (0.0443) (0.0446)

Table 2: Performance of the robust correlations from clean data to contaminated data when Pearson correlation is 0.90

Correlation	Data			
	Clean	Contaminated (outliers)		
		5%	10%	15%
Spearman's	0.8904 (0.0103) (0.0105)	0.9059 (0.0097) (0.0097)	0.9190 (0.0087) (0.0087)	0.9292 (0.0076) (0.0075)
Kendall's tau	0.7128 (0.0132) (0.0135)	0.7384 (0.0134) (0.0133)	0.7575 (0.0124) (0.0124)	0.7702 (0.0112) (0.0112)
Maronna's correlation	0.8993 (0.0086) (0.0086)	0.8859 (0.0074) (0.0075)	0.9406 (0.0497) (0.0397)	0.9993 (0.0001) (0.0027)
Univariate Winsorized	0.8973 (0.0087) (0.0089)	0.9187 (0.0083) (0.0083)	0.9369 (0.0074) (0.0074)	0.9525 (0.0064) (0.0065)
Bivariate Winsorized	0.9009 (0.0085) (0.0087)	0.8967 (0.0088) (0.0088)	0.8970 (0.0088) (0.0089)	0.9003 (0.0089) (0.0091)
Adjusted Bivariate Winsorized	0.8997 (0.0085) (0.0087)	0.8946 (0.0088) (0.0088)	0.8943 (0.0088) (0.0089)	0.8971 (0.0090) (0.0091)
Quadrant	0.7119 (0.0356) (0.0314)	0.7269 (0.0356) (0.0308)	0.7432 (0.0356) (0.0298)	0.7613 (0.0237) (0.0289)

Table 3: Performance of the classical correlations and robust correlations from clean data to contaminated data when true Pearson correlation is 0.90

Correlation		Data			
		Clean	Contaminated (outliers)		
			5%	10%	15%
Classical	Pearson	0.8998 (0.0084) (0.0085)	0.9983 (0.0003) (0.0003)	0.9990 (0.0001) (0.0001)	0.9992 (0.0001) (0.0001)
	Point biserial	0.7177 (0.0160) (0.0159)	0.0633 (0.0457) (0.0451)	0.0431 (0.0448) (0.0445)	0.0339 (0.0443) (0.0446)

Robust	Spearman's	0.8904 (0.0103) (0.0105)	0.9059 (0.0097) (0.0097)	0.9190 (0.0087) (0.0087)	0.9292 (0.0076) (0.0075)
	Kendall's tau	0.7128 (0.0132) (0.0135)	0.7384 (0.0134) (0.0133)	0.7575 (0.0124) (0.0124)	0.7702 (0.0112) (0.0112)
	Maronna's correlation	0.8993 (0.0086) (0.0086)	0.8859 (0.0074) (0.0075)	0.9406 (0.0497) (0.0397)	0.9993 (0.0001) (0.0027)
	Univariate Winsorized	0.8973 (0.0087) (0.0089)	0.9187 (0.0083) (0.0083)	0.9369 (0.0074) (0.0074)	0.9525 (0.0064) (0.0065)
	Bivariate Winsorized	0.9009 (0.0085) (0.0087)	0.8967 (0.0088) (0.0088)	0.8970 (0.0088) (0.0089)	0.9003 (0.0089) (0.0091)
	Adjusted Bivariate Winsorized	0.8997 (0.0085) (0.0087)	0.8946 (0.0088) (0.0088)	0.8943 (0.0088) (0.0089)	0.8971 (0.0090) (0.0091)
	Quadrant	0.7119 (0.0356) (0.0314)	0.7269 (0.0356) (0.0308)	0.7432 (0.0356) (0.0298)	0.7613 (0.0237) (0.0289)

Table 4: Percentage change of correlation from clean to different outliers for true Pearson correlation 0.90

Correlation		5%	10%	15%
Classical	Pearson correlation	10.95	11.02	11.04
	Point biserial correlation	91.18	93.99	95.28
Robust	Spearman's rank correlation	1.74	3.21	4.36
	Kendall's tau	3.59	6.27	8.05
	Maronna's correlation	1.49	4.59	11.12
	Univariate Winsorized	2.38	4.41	6.15
	Bivariate Winsorized	0.47	0.43	0.07
	Adjusted Bivariate Winsorized correlation	0.57	0.60	0.29
	Quadrant correlation	2.11	4.40	6.94