Hassan S. URAIBI, PhD
College of Administration &Economics
University of Al-Qadisiyah, IRAQ
E-mail: hssn.sami1@gmail.com, hassan.uraibi@qu.edu.iq
Professor Habshah MIDI, PhD
Faculty of Science, University Putra Malaysia
E-mail: habshah@science.upm.edu.my

ON ROBUST BIVARIATE AND MULTIVARIATE CORRELATION COEFFICIENT

Abstract. *The main purpose of this paper is to formulate a robust correlation coefficient for high dimensional data in the presence of multivariate outliers. The proposed method is compared with the existing robust bivariate correlation based on Adjusted Winsorization data and the well-known Pearson's correlation coefficient. The performance of our proposed method is investigated using artificial data and simulation study. An important implication of these findings is that the robust correlation based on RFCH estimator is more reliable and more efficient than the existing methods in all type of contamination scenarios.*

Keywords: *Robust Correlation, RFCH, Winsorization, Outliers, MCD, MVE.*

1. Introduction

Let ρ be the correlation coefficient between two random variables (X,Y) and let $(x_1,y_1),\dots,(x_n,y_n)$ be n observations from a bivariate normal distribution. The Pearson's product moment correlation coefficient which is widely used for estimating ρ is defined as,

$$\hat{\rho} = \frac{\mathrm{COV}(x,y)}{\sqrt{\mathrm{Var}(x)}\sqrt{\mathrm{Var}(y)}} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}} \quad (1)$$

where $\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$ and $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$ are the sample mean.

The density function of bivariate normal distribution where $\hat{\rho}$ is the maximum likelihood estimator ofρ can be written as follows,

$$f(x,y;\mu_1,\mu_2,\sigma_1,\sigma_2,\rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{(1-\rho^2)}} exp\left\{-\frac{1}{2\sqrt{(1-\rho^2)}}\right.$$

$$\left.\times\frac{(x-\mu_1)^2}{\sigma_1^2} + \frac{(y-\mu_2)^2}{\sigma_2^2} - 2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}\right\} \quad (2)$$

DOI: 10.24818/18423264/53.2.19.13

where μ_1 and μ_2 are the parameters of mean, σ_1 and σ_2 are standard deviations of the random variables X and Y respectively (Shevlyakov and Smirnov,2011). t is well known that sample mean and sample standard deviation are easily affected by outliers or other contamination. In this situation $\hat{\rho}$ or r becomes nonrobust when outliers may occur in either x_i or y_i or in both (x_i, y_i). As such, contamination model should be described using mixture normal densities, where outliers are distributed from another distribution. It is called normal mixture model (Marrona et al. ,2006) where the distribution of outliers is normal but it differs from the normal distribution of the majority of data. Tukey's gross error model (Tukey,1960) described such model based on sample correlation coefficient as follows,

$$f(x,y) = (1-\epsilon)N(x,y;0,0,1,1,\rho) + \epsilon\, N(x,y;0,0,k,k,\acute{\rho}) \quad (3)$$

where $k > 1$, $\mathrm{sgn}(\acute{\rho}) = -\mathrm{sgn}(\rho)$ and ϵ is positive number such that $(0 \leq \epsilon < 0.50)$.

Since the sample correlation does not exactly comes from the bivariate normal distribution, but from a normal multivariate model, the second part of Equation (3) would effect on the accuracy of the estimation of correlation coefficient. This is due to the fact that the mean and standard deviation estimation of good data in the first part of Equation (3) are combined with their counterparts in the second part. Thus, the sample correlation coefficient of model (3) is biased especially, when the $\acute{\rho}$ is significant and differ from their counterparts in the first part of Equation (3) (Shevlyakov and Smirnov,2011). Consequently, the presence of outliers in a data destroys the estimated value of ρ of good data and may change its sign (Gnanadesikan and Kettenring, 1972; Devlin, Gnanadesikan, and Kettenring, 1981). On the solution considered to overcome this issue, the robustness literature shows a variety of approaches of robust correlations based on robust variance-covariance matrix. The commonly used robust multivariate location and scatter matrices are the Minimum Covariance Determinant (MCD) and Minimum Volume Ellipsoid (MVE) that were introduced by Rousseeuw (1984) and Rousseeuw (1985), respectively.

Unfortunately, the MCD and the MVE are not feasible option for high dimensional data due to their time consuming procedures, even though the Fast-MCD is used (Rousseeuw and Van Driessen ,1999), see also Khan et al. (2007a) and Khan et al. (2007b). However, some publications addressed this issue and proposed using the approach of robust univariate correlation (see, Alqallaf et al.,2002) or bivariate correlation that will be discussed later. Khan et al. (2007b) proposed Adjusted Winsorization correlation which reduces the computation time as a robust bivariate correlation to overcome the problem of high dimensional data and bivariate outliers (outliers that are present in two predictors). Khan et al. (2007b) pointed out that the Adjusted Winsorization correlation yields very poor results in the presence of multivariate outliers. Olive and Hawkins (2010) proposed Reweighted Fast

DOI: 10.24818/18423264/53.2.19.13

Consistent and High breakdown (RFCH) estimator which is faster than the Fast MCD. To the best our knowledge, this estimator has not been used in the development of robust correlation coefficient in the presence of multivariate outliers (outliers that are present in more than two predictors). Therefore, the objective of this paper is to propose robust multivariate correlation matrix based on RFCH estimator which is more robust and less time consuming than the Adjusted Winsorized correlation. This paper is organized to present the bivariate correlation based on Adjusted Winsorization data in Section 2. The Section 3 describes the robust multivariate correlation matrix which relies on the RFCH estimator. Section 4 and Section 5 illustrate numerical example and simulation study to assess the performance of the RFCH correlation coefficient.

2. Adjusted Winsorization Correlation

Khan et al. (2007b) pointed out that the Pearson's correlation coefficients can be calculated from Winsorized data based on the approach of univariateWinsorization of the data which was introduced by Huber (1981). Alqallaf et al. (2002) re-studied Huber (1981) approach to estimate the individual elements of high dimensional

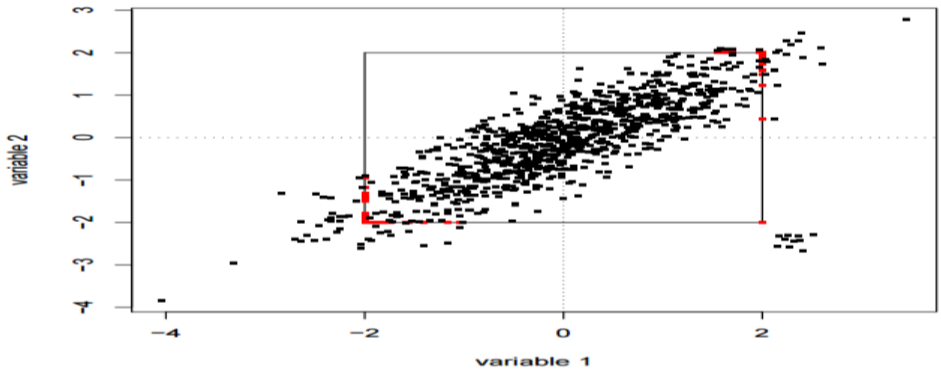

Figure 1: Winsorized univariate data with tuning constant ($c = 2$) to find the robust correlation estimates.

Correlation matrix. This idea starts by standardizing the dependent variable y_i and independent variable x_i robustly as follows:

$$X_i = \frac{x_i - med(x_i)}{MAD(x_i)} \text{ and } Y_i = \frac{y_i - med(y_i)}{MAD(y_i)}$$

where $i = 1,2, \ldots, n$.

They first determined tuning constant, c equals to 2. Observations that having absolute value greater than c, are transformed using Huber score function that are defined as follows,

$$\psi_c(X_i) = \begin{Bmatrix} c\,\text{sign}(X_i) & if & |X_i| > c \\ X_i & & O.W \end{Bmatrix} \quad (4)$$

and

DOI: 10.24818/18423264/53.2.19.13

$$\psi_c(Y_i) = \begin{Bmatrix} c\ \text{sign}(Y_i) & if & |Y_i| > c \\ Y_i & & O.W \end{Bmatrix} \quad (5)$$

By using Figure 1, Khan et al. (2007b) illustrated that the univariateWinsorization approach brings the outlying observations to the boundary of a $2c \times 2c$ square. The plot shows the effect of the correlation outliers at the bottom right corner which are shrunken to the corner$(2, -2)$, and thus are left almost unchanged. To remedy this problem, a robust correlation estimator derived from pairwise affine equivariant covariance estimator, such as bivariate M-estimator can be used (Maronna,1976), DGK , OGK (Maronna and Zamar,2002). Khan et al. (2007b) pointed out that this method is not fast enough with high dimensional data. Therefore, they suggested the approach of Adjusted Winsorization as bivariate Winsorization approach of the data which is based on an initial tolerance ellipse and corresponding to robust bivariate correlation matrix. The transformation shrinks the outliers to the border of this ellipse. This approach can be summarized as follows,

Step1. Determine different tuning constants c for different quadrants which is called the Adjusted Winsorization: Let $\alpha_i = \psi_c(X_i) * \psi_c(Y_i)$, then n_1 is the subsample of $\alpha_k$where $\alpha_k = (\alpha_i < 0)$ and $k = 1,2,..,m$, and $m < n$. The remaining α_i which is equivalent to $(n - m)$ is the second subsample, denoted as n_2. If $(n_1 < n_2)$, the new tuning parameter $c_1 = c \times \sqrt{h}$ where $h = n_1/n_2$ and then re-calculate the transformed variables using Equations (4) and (5), with new constant c_1. Hence, the transformation includes only the outliers inn_1. When $n_1 > n_2$ the new tuning parameter is $c_1 = c \times \sqrt{h}$ where $h = n_2/n_1$. The Huber univariateWinsorization correlation is given as follows,

$$\hat{\rho}[\psi_c(Y_i), \psi_c(X_i)] = \frac{Cov[\psi_c(Y_i), \psi_c(X_i)]}{\sigma_{\psi_c(Y_i)} \cdot \sigma_{\psi_c(X_i)}} \quad (6)$$

Since $\sigma_{\psi_c(Y_i)} = \sigma_{\psi_c(X_i)} = 1$, the Equation (6) is then written as follows,

$$\hat{\rho}[\psi_c(Y_i), \psi_c(X_i)] = Cov[\psi_c(Y_i), \psi_c(X_i)] \quad (7)$$

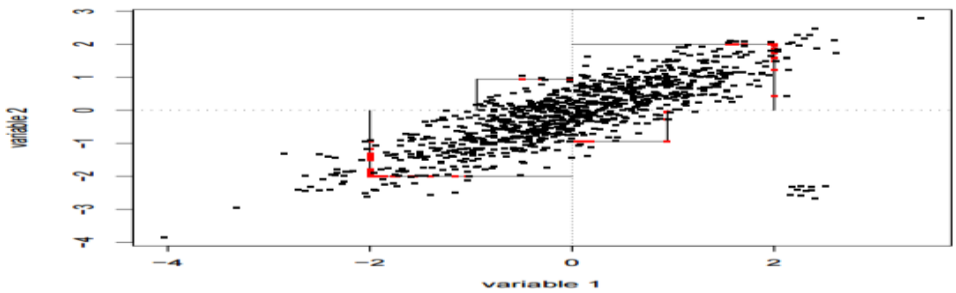

Figure 2: Adjusted Winsorization data with tuning constant ($c = 2$) and $c_1 = \sqrt{hc}$ for calculating the initial robust correlation estimate.

DOI: 10.24818/18423264/53.2.19.13

Figure 2 illustrates the limitations of the Adjusted Winsorization for finding the initial robust correlation estimate. It is obvious that the outlying points are illustrated to the corners of squares.

Step 2 Outliers transformation:

Suppose that φ is a 2×2 matrix of $Cov[\psi_c(Y_i), \psi_c(X_i)]$, such that,

$$\varphi = \begin{bmatrix} 1 & Cov[\psi_c(Y_i), \psi_c(X_i)] \\ Cov[\psi_c(X_i), \psi_c(Y_i)] & 1 \end{bmatrix} \quad (8)$$

and let $Z = (Y_i \quad X_i)$, and the $D(Z) = diag\,(Z\varphi^{-1}Z')$ is the Mahalanobis distance. The data points that has $D(Z) > \chi^2_{(0.95,2)}$ are multiplied by $\sqrt{\chi^2_{(0.95,2)}/D(Z)}$, so that the $\tilde{Z}$ is the final transformed data which is used to find the correlation $D(Z)$.

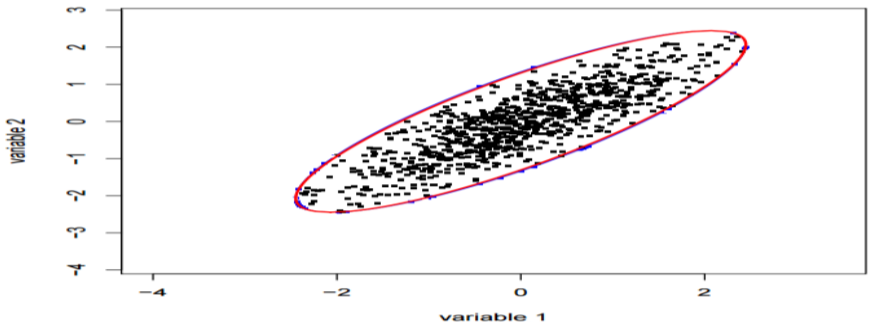

Figure 3: Bivariate Winsorization tolerance ellipses for clean (smaller ellipse) and contaminated (larger ellipse) data.

Figure 3 shows the tolerance ellipses used for bivariate Winsorization of both the full dataset of Figure 1 and the clean dataset after shrinkage of outliers. The bivariate Winsorization tolerance ellipse of clean data is slightly smaller than the tolerance ellipse of contaminated data and the outliers are shrunken toward the boundary of the larger ellipse. Note that the ellipses connect points of equal robust Mahalanobis distance which is based on the coordinatewise median and initial robust bivariate correlation matrix. The bivariate Winsorized correlation coefficient can be computed from the contaminated data, the points outside the largest ellipse are shrunken towards the boundary of that ellipse.Khan et al. (2007 a) concluded that in spite of the initial adjusted Winsorization and the resulting bivariate Winsorization are not affine equivariant, they are very fast to compute and appropriately handle correlation outliers The estimates of adjusted Winsorized correlation are consistent under certain regular conditions, provided that the location and scale estimates are consistent. Appendix A summarizes the algorithm of adjusted-Winsorization.

3. The correlation based on Reweighted Fast Consistent and High breakdown estimator (RFCH).

The concentrating algorithm assumes that the normality assumption for a linear regression is violated due to outliers or other contamination. The RFCH algorithm

DOI: 10.24818/18423264/53.2.19.13

is employed to clean the data. This procedure uses the DGK (Devlin et al., 1981) and Median Ball (MB) (Olive and Hawkins, 2008), These algorithms are summarized as follows.Suppose the matrix X is a combination of the response vector y^* and the covariates matrixX^*.

(i) The DGK Algorithm

Step 1: Begin by computing the classical estimator $(\bar{\mathrm{x}}, \mathrm{cov})$ of the original dataset to give the initial or starting point $(T_{0,Start}, C_{0,Start})$, and find the initial Mahalanobis distance.

$$D_{0,DGK} = \sqrt{(X - T_{0,Start})^t (C_{0,Start})^{-1} (X - T_{0,Start})} \quad (9)$$

Step 2: Arrange the initial Mahalanobis distances in increasing order to compute their median $Med_{0,DGK} = Median(D_{0,DGK})$. Those observations in the original dataset whose Mahalanobis distances are less than the median of all the Mahalanobis distances will be in the remaining set (half dataset) and will be denoted by $\tilde{X}_{1,DGK} = \{X_{ij} : D_{0,DGK} \leq Med_{0,DGK}\},\ i = 1, \ldots, p,\ \ j = 1,2, \ldots, m.$

Step 3: Let $C_{0,DGK}$ be equal to $C_{0,Start}$ where $C_{0,Start}$ is the variance-covariance matrix of the original data. Calculate the average and the variance-covariance estimators of $\tilde{X}_{1,DGK}$to get the first attractor $(T_{1,DGK}, C_{1,DGK})$.

Step 4: If the diagonal elements of $C_{1,DGK}$ are equal to $C_{0,Start}$ then stop the algorithm. Otherwise, repeat Steps 1-3 until convergence, to get the final attractor $(T_{k,DGK}, C_{k,DGK})$ and $\tilde{X}_{k,DGK}$, where k is the convergence step.

(ii) The Median Ball (MB) Algorithm

Step 1: Suppose the initial variance-covariance matrix $C_{0,Start} = diag\,(p)$ of the identity matrix, and that *Med* is the median vector of the matrix X. Then the Mahalanobis distance based on the median is defined as follows:

$$D_{0,MB} = \sqrt{(X - \mathrm{MED(X)})^t (C_{0,Start})^{-1} (X - \mathrm{MED(X)})} \quad (10)$$

Step 2: The location criterion cut-off point is the median of $D_{0,MB}$, and is denoted by$luct$,

$$luct = Med_{0,MB} = Median(D_{0,MB}) \quad (11)$$

where $luct \neq 0.5$. The cut-off point should be the quantile of $D_{0,MB}$ whose probability equals 0.5. For the concentration of X, find the half dataset with only non-outlying observations whose Mahalanobis distances are less than or equal to the median:

$$\tilde{X}_{1,MB} = \{X_{ij} : D_{0,MB} \leq Med_{0,MB}\},\ i = 1, \ldots, p,\ \ j = 1,2, \ldots, m \quad (12)$$

Step 3: Compute the average and the variance-covariance matrix of $\tilde{X}_{1,MB}$.

Step 4: For more concentrations, compute the Mahalanobis distances again, and repeat Steps (1-3) until convergence at the final attractor $(T_{k,MB}, C_{5,MB})$ and $\tilde{X}_{k,MB}$ where k is the convergence step.

(iii) The Reweighted Fast and Consistent High Breakdown (RFCH) Algorithm

DOI: 10.24818/18423264/53.2.19.13

Olive and Hawkins (2010) developed the MB estimator by adding the location criterion or cut-off point to select the attractor, and proposed the so-called Fast Consistent and High breakdown (FCH) estimator. Olive and Hawkins (2010) noted that the FCH estimator uses the attractors with the smallest determinant.

Step 1: Following the same approach as Olive and Hawkins (2010), define the final attractors as follows:

$$T_{FCH} = \begin{cases} T_{K,DGK} & if \ \sqrt{|C_{K,DGK}|} < \sqrt{|C_{K,MB}|} \\ T_{K,MB} & Otherwise \end{cases} \quad (13)$$

and

$$C_{FCH} = \begin{cases} \frac{MED\left(D_i^2((T_{K,DGK},C_{K,DGK}))\right)}{\chi^2_{(p,0.5)}} \times C_{K,DGK}, & if \ \sqrt{|C_{K,DGK}|} < \sqrt{|C_{K,DGK}|} \\ \frac{MED\left(D_i^2((T_{K,MB},C_{K,MB}))\right)}{\chi^2_{(p,0.5)}} \times C_{K,MB}, & Otherwise \end{cases} \quad (14)$$

where $\chi^2_{(p,0.5)}$ is the 50th percentile of a chi-square distribution with p degrees of freedom. According to Theorem 1 (Olive and Hawkins, 2010) as long as the start (T_k, C_k) is a consistent estimator of either $(T_{k,DGK}, s_0 C_{k,DGK})$ or $(T_{k,MB}, s_1 C_{k,MB})$, the FCH attractor is a consistent estimator of $(T_{k,FCH}, a\, C_{k,FCH})$, where

$$s_0 = \frac{MED\left(D_i^2(T_{k,DGK},C_{k,DGK})\right)}{\chi^2_{(p,0.5)}} \text{ and } s_1 = \frac{MED\left(D_i^2(T_{k,MB},C_{k,MB})\right)}{\chi^2_{(p,0.5)}} \text{ are positive constants,}$$

and $a = s_0$ or $a = s_1$ based on the criterion cut-off point.

Step 2: Obtain the Reweighted FCH attractors by isolating the observation with $D_i^2(T_{\mathrm{FCH}}, C_{\mathrm{FCH}}) \leq \chi^2_{(p,0.975)}$, and using the classical estimator to obtain $(T_{1,\mathrm{FCH}}, C_{1,\mathrm{FCH}})$ from:

$$\tilde{X}_{1,\mathrm{FCH}} = \left\{X_{ij} : D_i^2(T_{\mathrm{FCH}}, C_{\mathrm{FCH}}) \leq \chi^2_{(p,0.975)}\right\},\ i = 1, \ldots, p,\ j = 1,2, \ldots, m \quad (15)$$

Compute the new cut-off point as $\frac{MED\left(D_i^2(T_{1,\mathrm{RFCH}},C_{1,\mathrm{FCH}})\right)}{\chi^2_{(p,0.5)}}$. The new variance-covariance matrix is:

$$C_{2,\mathrm{FCH}} = \frac{MED\left(D_i^2(T_{1,\mathrm{FCH}},C_{1,\mathrm{FCH}})\right)}{\chi^2_{(p,0.5)}} \times C_{1,\mathrm{FCH}} \quad (16)$$

Step 3: Repeat steps (1-2) with the new cut-off point until convergence, to get the final attractors $(T_{\mathrm{RFCH}}, C_{\mathrm{RFCH}})$ and $\tilde{X}_{\mathrm{RFCH}}$.

Upon convergence, the RFCH produced the final $(T_{\mathrm{RFCH}}, C_{\mathrm{FCH}})$ estimators which is $\sqrt{n}$ consistent according to Olive and Hawkins (2010). The RFCH correlation matrix is then formulated based on the C_{RFCH} defined as follows:

DOI: 10.24818/18423264/53.2.19.13

$$\hat{\rho}_{RFCH} = \begin{bmatrix} 1 & \frac{c_{RFCH}^{(1,2)}}{\sqrt{c_{RFCH}^{(1,1)} c_{RFCH}^{(2,2)}}} & \cdots & \frac{c_{RFCH}^{(1,p+1)}}{\sqrt{c_{RFCH}^{(1,1)} c_{RFCH}^{(p+1,p+1)}}} \\ \frac{c_{RFCH}^{(2,1)}}{\sqrt{c_{RFCH}^{(1,1)} c_{RFCH}^{(2,2)}}} & 1 & \cdots & \frac{c_{RFCH}^{(2,p+1)}}{\sqrt{c_{RFCH}^{(2,2)} c_{RFCH}^{(p+1,p+1)}}} \\ \vdots & \vdots & & \\ \frac{c_{RFCH}^{(p+1,1)}}{\sqrt{c_{RFCH}^{(1,1)} c_{RFCH}^{(p+1,p+1)}}} & \frac{c_{RFCH}^{(p+1,2)}}{\sqrt{c_{RFCH}^{(2,2)} c_{RFCH}^{(p+1,p+1)}}} & \ddots \; \vdots & \\ & & \cdots \; 1 & \end{bmatrix} \quad (17)$$

The computation of the RFCH is illustrated using the stack-loss data (available in R-software) as follows: Firstly, the DGK algorithm is employed. The classical estimator $(\mathrm{T}_{1,\mathrm{D}}, \mathrm{C}_{1,\mathrm{D}}) = (\bar{\mathrm{X}}, \mathrm{S})$ is used as an initial estimator to get the final attractor of DGK estimator$(T_{K,D}, C_{K,D})$. The DGK is an iterative algorithm for location and dispersion estimator which may converge in five steps.

Consider,$T_{1,D} = [0.43, 0.41, 0.37, -0.61\ ]$, $C_{1,D} = \begin{pmatrix} 2.94 & 2.44 & 1.60 & 0.83 \\ 2.44 & 2.39 & 1.29 & 0.93 \\ 1.60 & 1.29 & 1.14 & 0.50 \\ 0.83 & 0.93 & 0.50 & 1.45 \end{pmatrix}$

are the classical estimators of robust standardize stack loss data (available in R software), The values of MD of the first trials (k=1) is presented in Table (1). The second classical estimator $(T_{2,D}, C_{2,D})$ is computed based on half of the original data, check whether $MD_D(1)$ is less than the median of $MD_D(1)$. This procedure is repeated five times (k=1,2,3,4,5) to obtain the DGK estimators,

$$T_{5,D} = [-0.28, -0.06, -0.51, -0.37], C_{5,D} = \begin{pmatrix} 0.39 & 0.44 & 0.26 & 0.20 \\ 0.44 & 0.56 & 0.22 & 0.40 \\ 0.26 & 0.22 & 0.50 & 0.04 \\ 0.20 & 0.40 & 0.04 & 0.99 \end{pmatrix}.$$

Table (1) shows that the MD of the five trials of DGK algorithm for stakloss data. It can be obtained from the results of the estimators which converged after four steps.

Secondly, the Median Ball Algorithm is employed.

The MB estimator $(T_{K,M}, C_{K,M})$ uses $(\mathrm{T}_{1,\mathrm{M}}, \mathrm{C}_{1,\mathrm{M}}) = (\mathrm{MED(X)}, I_{\mathrm{p}})$ as a start, where $\mathrm{MED(X)}$ is the coordinate-wise median and I_{p} is the identity matrix.

DOI: 10.24818/18423264/53.2.19.13

$T_{1,M} = [0,0,0,0]$, $C_{1,M} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$,

It is obvious that the MB estimator converged after two steps which yields the consistent estimator $(T_{K,M}, C_{K,M})$.

$T_{K,M} = [-0.31, -0.15, -0.18, -0.14]$,

$$C_{K,M} = \begin{pmatrix} 0.36 & 0.36 & 0.31 & -0.05 \\ 0.36 & 0.44 & 0.27 & -0.02 \\ 0.31 & 0.27 & 0.55 & 0.04 \\ -0.05 & -0.02 & 0.04 & 0.71 \end{pmatrix}.$$

Table (2) shows that the MD of the five trials of MB algorithm for stakloss data. It can be obtained from the results of the estimators which converged after three steps.If the DGK location estimator $T_{K,D}$ has a greater Euclidean distance than $\mathrm{MED(X)}$, FCH uses the MB attractor. The FCH uses the smallest determinant as the location criterion to choose the attractor if:

$$\left\|\mathrm{T_{K,D}} - \mathrm{MED(X)}\right\| \leq \mathrm{MED}\left(\mathrm{MD_i}\left(\mathrm{MED(X)}, I_p\right)\right) \quad (18)$$

where $\mathrm{MD_i}$ is the square root of the mahalanobis distance.

Hence, $\left\|\mathrm{T_{K,D}} - \mathrm{MED(X)}\right\| = 0.35$ which is less than $\mathrm{MED}\left(\mathrm{D_i}\left(\mathrm{MED(X)}, I_p\right)\right) = 4.14$, therefore the first condition is satisfied and if the $\sqrt{|\mathrm{C_{K,D}}|} < \sqrt{|\mathrm{C_{K,M}}|}$ the FCH attractor is $(\mathrm{T}_{k,\mathrm{D}}, \mathrm{C}_{k,\mathrm{D}})$, otherwise $(\mathrm{T}_{k,\mathrm{M}}, \mathrm{C}_{k,\mathrm{M}})$. In our example of stackloss data, $\sqrt{|\mathrm{C_{K,D}}|} = 0.05$ which is less than $\sqrt{|\mathrm{C_{K,M}}|} = 0.07$, so that FCH attractor is $(\mathrm{T}_{k,\mathrm{D}}, \mathrm{C}_{k,\mathrm{D}})$.Finally in the third step, the Reweighted Fast and Consistent, High breakdown (RFCH) algorithm is employed.

Let (T_A, C_A) be the attractor used, then the FCH estimator is $T_{FCH} = T_A$ and weighted C_A as follows:

$$C_{FCH} = \frac{MED\left(MD^2((T_A, C_A))\right)}{\chi^2_{(p,0.5)}} \times C_A \quad (19)$$

where $\chi^2_{(\mathrm{p},0.5)}$ is the 50$^{\text{th}}$ percentile of a chi-square distribution with p degrees of freedom. Refer to our example, the weighted $\mathrm{C}_{k,\mathrm{D}}$ requires finding the constant c which is equals to $\frac{\mathrm{MED}\left(MD^2(\mathrm{T}_{k,\mathrm{D}}\mathrm{C}_{k,\mathrm{D}})\right)}{\chi^2_{(\mathrm{p},0.5)}}$, where $\chi^2_{(3,0.5)} = 3.36$ and the $MD^2(\mathrm{T}_{k,\mathrm{D}}, \mathrm{C}_{k,\mathrm{D}})$ is already computed in Table (1) where k=5 in DGK algorithm. As a result of that c is (1.87), then

DOI: 10.24818/18423264/53.2.19.13

$$C_{FCH} = 1.87 \times \begin{pmatrix} 0.39 & 0.44 & 0.26 & 0.20 \\ 0.44 & 0.56 & 0.22 & 0.40 \\ 0.26 & 0.22 & 0.50 & 0.04 \\ 0.20 & 0.40 & 0.04 & 0.99 \end{pmatrix} = \begin{pmatrix} 1.06 & 0.42 & 0.76 & 0.83 \\ 0.42 & 0.94 & 0.08 & 0.48 \\ 0.76 & 0.08 & 1.86 & 0.38 \\ 0.83 & 0.48 & 0.38 & 0.76 \end{pmatrix}$$

and

$T_{FCH} = [-0.31, -0.15, -0.18, -0.14]$

Reweighted (T_{FCH}, C_{FCH}) estimator need to find $\mathrm{MD}^2(T_{FCH}, C_{FCH})$ with cutoff point equivalent to $\chi^2_{(p,0.975)}$. The classical estimator $(\widehat{T_{FCH_1}}, \widehat{C_{FCH_1}})$ is obtained from n_1 cases of original data which anolog $\mathrm{MD}^2(T_{FCH}, C_{FCH}) \leq \chi^2_{(p,0.975)}$. So that

$$\widehat{C_{FCH}} = \frac{MED\left(MD^2(\widehat{T_{FCH_1}}, \widehat{C_{FCH_1}})\right)}{\chi^2_{(p,0.5)}} \times \widehat{C_{FCH_1}} \qquad (20)$$

Numerically, $\mathrm{MD}^2(T_{FCH}, C_{FCH})$ exhibited in Table (3) with cut-off point $\chi^2_{(p,0.975)}$ equals to 14.11. The classical estimators of n_1 cases with their $\mathrm{MD}^2[\mathrm{Re}(1)] \leq$ 14.11 is,

$$\widehat{T_{FCH_1}} = [-0.29, 0.05, -0.30, -0.30]$$

$$\widehat{C_{FCH_1}} = \begin{pmatrix} 0.70 & 0.42 & 0.69 & 0.60 \\ 0.42 & 0.66 & 0.39 & 0.45 \\ 0.69 & 0.39 & 1.82 & 0.54 \\ 0.60 & 0.45 & 0.54 & 0.55 \end{pmatrix}$$

The first reweighted of $\widehat{C_{FCH_1}}$ is equivalent to multiplying $\widehat{C_{FCH_1}}$by the constant $c = \frac{MED\left(MD^2(\widehat{T_{FCH_1}}, \widehat{C_{FCH_1}})\right)}{\chi^2_{(p,0.5)}}$.

$$\widehat{C_{FCH}} = 1.40 \times \begin{pmatrix} 0.70 & 0.42 & 0.69 & 0.60 \\ 0.42 & 0.66 & 0.39 & 0.45 \\ 0.69 & 0.39 & 1.82 & 0.54 \\ 0.60 & 0.45 & 0.54 & 0.55 \end{pmatrix} = \begin{pmatrix} 0.98 & 0.59 & 0.97 & 0.83 \\ 059 & 0.93 & 0.55 & 0.63 \\ 0.97 & 0.55 & 2.54 & 0.76 \\ 0.83 & 0.63 & 0.76 & 0.78 \end{pmatrix}$$

Reweighted $\widehat{C_{FCH}}$ through finding the classical estimator $(\widehat{T_{FCH_2}}, \widehat{C_{FCH_2}})$ of n_2 cases which poses $\mathrm{MD}^2(\widehat{\mathrm{T_{FCH_1}}}, \widehat{\mathrm{C_{FCH}}}) \leq \chi^2_{(p,0.975)}$. Finally, the RFCH estimator equals to

$$\widehat{C_{RFCH}} = \frac{MED\left(MD^2(\widehat{T_{FCH_2}}, \widehat{C_{FCH_2}})\right)}{\chi^2_{(p,0.5)}} \times \widehat{C_{FCH_2}} \qquad (21)$$

Table 1, Table 2 and Table 3 exhibit the DGK, MB and RFCH algorithms respectively. The values in parenthesis are the cutoff points and Re is the Reweighted RFCH.

DOI: 10.24818/18423264/53.2.19.13

Table 1. The DGK algorithm outputs of Mahalanobis distances at each iteration for stack loss data.

k=1 *(3.23)*	*k=2* *(4.61)*	*k=3* *(6.29)*	*k=4* *(6.29)*	*k=5* *(6.29)*
6.25	**95.42**	**157.42**	**157.42**	**157.42**
5.82	**45.82**	**50.50**	**50.50**	**50.50**
4.86	**77.70**	**140.80**	**140.80**	**140.80**
5.25	**69.28**	**175.37**	**175.37**	**175.37**
0.42	1.99	2.02	2.02	2.02
1.61	3.01	3.13	3.13	3.13
4.07	**7.07**	**11.32**	**11.32**	**11.32**
3.65	**9.04**	**19.06**	**19.06**	**19.06**
2.96	3.06	2.85	2.85	2.85
3.22	3.90	**6.29**	**6.29**	**6.29**
2.93	4.12	3.15	3.15	3.15
4.55	4.38	2.71	2.71	2.71
2.43	6.47	26.29	26.29	26.29
3.16	4.61	4.40	4.40	4.40
3.48	**6.82**	**10.19**	**10.19**	**10.19**
1.76	3.43	3.29	3.29	3.29
7.54	**9.49**	**17.20**	**17.20**	**17.20**
2.28	2.75	3.60	3.60	3.60
2.57	2.92	2.84	2.84	2.84
0.87	3.72	5.69	5.69	5.69
10.59	**48.66**	**145.58**	**145.58**	**145.58**

The classical estimators of n_2 cases with their $\mathrm{MD}^2[\mathrm{Re}(2)] \leq 14.11$ are presented in Table 3 as follows,

$$\widehat{T_{FCH_2}} = [-2.74\,e-01, 6.94e-18, -3.51e-01, -3.27e-01]$$

$$\widehat{C_{FCH_2}} = \begin{pmatrix} 0.66 & 0.38 & 0.63 & 0.55 \\ 0.38 & 0.65 & 0.40 & 0.44 \\ 0.63 & 0.40 & 1.74 & 0.52 \\ 0.55 & 0.40 & 0.53 & 0.53 \end{pmatrix}$$

The second reweighted of $\widehat{C_{FCH_2}}$ is equivalent to multiplying $\widehat{C_{FCH_2}}$ by the constant $c_2 = \frac{MED\left(MD^2(\widehat{T_{FCH_2}}, \widehat{C_{FCH_2}})\right)}{\chi^2_{(p,0.5)}} = 1.24.$

The final estimate of location and variance covariance matrix for RFCH estimates are as follows: Location

DOI: 10.24818/18423264/53.2.19.13

estimates, $\widehat{T_{RFCH}} = [-2.74\,e-01, 6.94e-18, -3.51e-01, -3.27e-01]$ and variance covariance matrix

$$\widehat{C_{RFCH}} = 1.24 \times \begin{pmatrix} 0.66 & 0.38 & 0.63 & 0.55 \\ 0.38 & 0.65 & 0.40 & 0.44 \\ 0.63 & 0.40 & 1.74 & 0.52 \\ 0.55 & 0.40 & 0.53 & 0.53 \end{pmatrix} = \begin{pmatrix} 0.82 & 0.47 & 0.78 & 0.68 \\ 047 & 0.81 & 0.50 & 0.55 \\ 0.78 & 0.50 & 2.16 & 0.65 \\ 0.68 & 0.55 & 0.65 & 0.65 \end{pmatrix}$$

Hence, the RFCH correlation matrix is given by

$$\hat{\rho}_{RFCH} = \begin{bmatrix} 1 & 0.58 & 0.59 & 0.94 \\ 0.58 & 1 & 0.38 & 0.75 \\ 0.59 & 0.38 & 1 & 0.55 \\ 0.94 & 0.75 & 0.55 & 1 \end{bmatrix}$$

Table 2. The MB algorithm outputs of Mahalanobis distances at each iteration for stack loss data.

k=1 *(4.14)*	*k=2* *(6.49)*	*k=3* *(6.49)*
40.27	**111.28**	**111.28**
33.15	**53.46**	**53.46**
25.28	**89.74**	**89.74**
7.08	**68.40**	**68.40**
1.17	2.00	2.00
1.73	2.82	2.82
4.55	**7.63**	**7.63**
4.81	**9.70**	**9.70**
1.02	3.35	3.35
2.96	3.61	3.61
0.69	2.66	2.66
1.19	2.84	2.84
2.17	**6.49**	**6.49**
2.19	4.13	4.13
3.87	3.20	3.20
4.14	3.94	3.94
14.70	**22.01**	**22.01**
6.56	**9.27**	**9.27**
5.32	**9.03**	**9.03**
1.38	4.26	4.26
4.90	**47.08**	**47.08**

DOI: 10.24818/18423264/53.2.19.13

Table 3. The RFCH algorithm outputs of Mahalanobis distances at each iteration for stack loss data.

Re(1) (11.14)	Re(2) (11.14)	Final MD
84.02	**83.78**	**105.21**
26.96	**31.88**	**47.18**
75.14	**70.21**	**84.65**
93.60	**77.15**	**76.31**
1.08	1.33	1.70
1.67	2.05	2.52
6.04	2.21	6.67
10.17	3.36	4.68
1.52	1.72	2.48
3.36	3.16	4.18
1.68	1.97	3.02
1.45	2.75	3.74
14.03	9.64	6.67
2.35	4.29	4.68
5.44	3.65	5.01
1.76	2.17	3.25
9.18	4,23	5.66
1.93	1.46	2.17
1.52	1.87	2.75
3.04	3.80	4.73
77.70	**76.44**	**62.84**

4. A Simulation Study

To evaluate the performance of our proposed correlation coefficient based on RFCH, we extended the simulation study of Abdullah (1990) to be suitable for multivariate analysis. We generate 50 sample size of clean data according to the following linear regression model,

$$y_i = 2.0 + 1.0\,X_1 + 1.0\,X_2 + 1.0\,X_3 + \varphi\sigma \qquad (22)$$

where $X_1 \sim N(5.0,1)$, $X_2 \sim N(10.0,1)$, $X_3 \sim N(15.0,1)$, the standard deviation of errors σ is selected to make signal to noise is $\varphi = 3$.

We use robust scale for all data as follows

$$\hat{X} = \frac{X - Med(X)}{MAD(X)}, \quad \hat{y} = \frac{y - Med(y)}{MAD(y)}$$

Then the Pearson's correlation coefficients, $\hat{\rho}_{(x_1,y)}$,$\hat{\rho}_{(x_2,y)}$ and $\hat{\rho}_{(x_3,y)}$ are computed. The $\hat{\rho}_{(x_1,y)}$,$\hat{\rho}_{(x_2,y)}$ and $\hat{\rho}_{(x_3,y)}$ are considered as the target regression

DOI: 10.24818/18423264/53.2.19.13

coefficients, since in the orthogonal design $\hat{\rho}_{(x_1,y)} = \widehat{\beta_1}, \hat{\rho}_{(x_2,y)} = \widehat{\beta_2}$ and $\hat{\rho}_{(x_3,y)} = \widehat{\beta_3}$ (see Appendix B) . Hence, we assume that $\rho = \hat{\rho}$ to simulate the normal density function $N(X, Y; 0,0,1,1, \rho)$.

The contamination of the data is constructed to study the effect of univariate outlier (the outliers that are present only in one predictor), bivariate outliers (outliers that are present in two predictors) and multivariate outliers (outliers that are present in more than two predictors) as follows,

1- Delete the first five good observations of y_i and replaced it with Vertical Outliers (VO's) which are generated from another normal distribution with mean equals to 2.0 and standard deviation equals to 5.0.
2- Five Leverage Points (LP's) are generated from uniform distribution on [5,10] to replace the first five good observations in x_1.
3- Both VO's and LP's are present in the data. The outliers are generated similar to 1 and 2 contaminations.
4- In addition to five leverage points generated as in 2, the first 10% good observations of x_2 are replaced by LP's which are generated from uniform distribution on [5, 10] to create bivariate LP's.
5- Multivariate outliers are created by generating the same process as in 1 and 4.

The Pearson's ,the Adjusted Winsorization and the RFCH correlations were then applied to the data. For each simulated data sets, there is 10000 replications. Over all 10000 replications, the average of Pearson's, Adj.Winso.corandRFCH.cor correlation coefficients are computed for each dataset which is denoted as $\overline{\hat{\rho}_{(x_j,y)}}$ where $j = 1,2,3$. Then, we find the of absolute value of $\left|\overline{\hat{\rho}_{(x_j,y)}} - \rho_{(x_j,y)}\right|$ which is denoted as Bias. The Root Mean Square Error (*RMSE*) of the correlation coefficients replications is computed using the following formula,

$$RMSE = \sqrt{\left(\overline{\hat{\rho}_{(x_j,y)}} - \rho_{(x_j,y)}\right)^2 + Var\left(\hat{\rho}_{(X_j,Y)}\right)}$$

where

$$Var\left(\hat{\rho}_{(X_j,Y)}\right) = \frac{\sum_{b=1}^{B}\left(\hat{\rho}_{(X_j,Y)} - \overline{\hat{\rho}_{(x_j,y)}}\right)^2}{B-1}$$

and B is the number of replications.

The Standard Error of $\hat{\rho}_{(X_j,Y)}$ which is denoted as SE is the square root of $Var\left(\hat{\rho}_{(X_j,Y)}\right)$. Table 4 presents the bias, SE and RMSE of the $\hat{\rho}_{(x_1,y)}$, $\hat{\rho}_{(x_2,y)}$ and $\hat{\rho}_{(x_3,y)}$ for Pearson's, Adj.Winso.corandRFCH.cor. The best method is the one that produces the lowest bias, SE and RMSE. We can see from Table 4 that the three methods are fairly close when there are no outliers in the data. We see that the values of Bias1, Bias2 ,Bias3, SE1,SE2,SE3 and RMSE1, RMSE2, RMSE3 of Pearson's are immediately affected by the presence of bivariate and multivariate

DOI: 10.24818/18423264/53.2.19.13

outliers (case 3,4,5). The Adj.Winso.cor is better than the Pearson's for all type of outliers, but cannot outperform the RFCH.cor. On the other hand, the RFCH.cor is not much affected by outliers. For all type of outliers, the values of Bias1, Bias2 and Bias3, SE1,SE2,SE3 and RMSE2, RMSE3 of the RFCH.cor remain almost the same as when there is no outlier in the data. It is evident from the results, that for bivariate outliers, the Adj.Winso.cor is good only for the correlation of X_2 and the response y (RMSE= 0.78) and X_3 and the response variable y $(RMSE = 0.72)$. However, for multivariate outliers (case 5) all RMSEs for Adj.Winso.cor≥ 0.80. It is interesting to note that the RFCH.cor consistently has the smallest biases, SEs and RMSEs followed by Adj.Winso.cor and Pearson. The results of simulation study are remarkably consistent in all outlier's cases.

The highest performance of RFCH.cor is shown with Case3, Case4 and Case5 where simulated data having vertical outliers jointed with univariate and bivariate outliers. Due to the fact that Adj.Winso.cor is dealing with bivariate correlation, the presence of vertical outliers in Y- direction and univariate outliers in X-direction is considered a bivariate outlier. This conception can be extended to consider the multivariate outliers when the X- direction having bivariate outliers.

It is evident that the RFCH.cor estimates with all contamination cases are consistent and more stable than Pearson's and Adj.Winso.cor which failed to be resistance to the presence of outliers.

5. Hawkins BraduKass (HBK) Data

The performance of our proposed RFCH correlation estimator is further assess using Hawkins, Bradu and Kass (1984) data. This is an artificial three-predictors data set containing 75 observations with 10 outliers in both of the spaces (LP's) [cases 1-10], 4 outliers in the X-space [cases 11-14] (see Hawkins et al., 1984; Habshah et al. , 2009; Rousseeuw and Leroy,1987). Most of the single case deletion identification methods fail to identify the outliers in Y-space though some of them detected cases 11-14 as outliers in the Y-space. These data are considered in order to compare the performance of the Pearson's and Adj.Winso.cor with our proposed correlation. All data are scaled and Pearson's , Adj.Winso.cor and RFCH.cor are then computed. We assumed that the estimated correlation coefficient of the cases 15-75, $\hat{\rho}_{(x_j y)}$ (when the first 14 high leverage points are removed from the data) is the target coefficient, for $j = 1,2,3$. Hence, the correlation coefficient for the scaled data equivalents to the regression coefficients (see Appendix B). Therefore, the absolute value of the bias between sample correlation coefficient and the target correlation is considered as the criterion to evaluate the performance of all the three methods. The correlation method which shows the lowest bias is the best method. In addition to this comparison criterion, the total amount of correlation coefficients (regression coefficients) that explain the variability in Y is also considered as another comparison criterion. We assumed that the summation of target correlation is the threshold. For our data, the threshold

DOI: 10.24818/18423264/53.2.19.13

equals to (-0.080). The value of (-0.080) is obtained from the data when outliers are removed from the data. The summation of these coefficients represents the total contribution of all independent variables in Y. The absolute value of the difference between the estimated correlations (coefficients) and the target correlation and the summation of the target correlation are used as the criterion to determine the best correlation method. Another criterion to choose the best method of correlation is the one in which the summation of coefficients is close to the threshold.

Table 5 exhibits the sample correlation, the absolute value of the differences; $\left|\hat{\rho}(X_1,Y)-\rho_{(x_1,y)}\right|$, $\left|\hat{\rho}(X_2,Y)-\rho_{(x_2,y)}\right|$ and $\left|\hat{\rho}(X_3,Y)-\rho_{(x_3,y)}\right|$ and the summation of the target correlations. The results show that the Pearson's and Adj.Winso.cor coefficients are much affected by outliers. The RFCH.cor coefficients are more accurate than the Pearson's and Adj.Winso.cor coefficients because it has the lowest values of $\left|\hat{\rho}(X_1,Y)-\rho_{(x_1,y)}\right|$, $\left|\hat{\rho}(X_2,Y)-\rho_{(x_2,y)}\right|$ and $\left|\hat{\rho}(X_3,Y)-\rho_{(x_3,y)}\right|$ and yields perfect total amount of RFCH.cor coefficients which is (-0.080) which is exactly equals to the sum of the target correlations.

6. Conclusion

The main focus of this study was to formulate a fast and more efficient alternative correlation coefficients between variables for high dimensional data in the presence of multivariate outliers. We have compared its performance with several correlations coefficients. The widely used Pearson's correlation is not robust in the presence of outliers. The adjusted Winsorization correlations is put forward to remedy this problem. Nonetheless, the adjusted Winsorization correlation is only robust to bivariate and not robust to multivariate outliers. We propose RFCH correlations in this regard. The numerical example and simulation study revealed that our proposed robust correlation based on RFCH is more resistant to univariate, bivariate and multivariate outliers, irrespective of the outliers scenarios and percentage of outliers in the dataset. It is more consistent and more robust than the adjusted Winsorization correlation.

REFERENCES

[1]**Abdullah B. Mokhtar (1991),** ***On a Robust Correlation Coefficient.*** *Journal of the Royal Statistical Society. Series D (The Statistician) Vol. 39, No. 4 (1990), pp. 455-460;*

[2]**Alqallaf, F. A., Konis, K. P., Martin, R D. and Zamar, R. H. (2002),** ***Scalable Robust Covariance and Correlation Estimates for Data Mining***. Paper presented at the Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining;

DOI: 10.24818/18423264/53.2.19.13

[3]**Devlin, S. J, Gnanadesikan, R. and Kettenring, J. R. (1981),** ***Robust Estimation of Dispersion Matrices and Principal Components***. *Journal of the American Statistical Association, 76*(374), 354-362;
[4]**Gnanadesikan, R. and Kettenring, J. R. (1972),** ***Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data***. *Biometrics*, 81-124;
[5]**Huber, P. J. (1981),** ***Robust Statistics***. New York: *John Wiley and Sons*, Inc.;
[6]**Khan, J. A., Van Aelst, S. and Zamar, R. H. (2007a),** ***Building a Robust Linear Model with Forward Selection and Stepwise Procedures.****Computational Statistics and Data Analysis, 52*(1), 239-248;
[7]**Khan, J. A., Van Aelst, S. and Zamar, R. H. (2007b),** ***Robust Linear Model Selection Based on Least Angle Regression****. Journal of the American Statistical Association,* 102(480), 1289-1299;
[8]**Maronna, R. A., Martin, R. D. and Yohai, V. J. (2006),** ***Robust Statistics: Theory and Methods***: *John Wiley and Sons*, Ltd.;
[9]**Olive, D. J, and Hawkins, D. M. (2008),** ***High Breakdown Multivariate Estimators.*** *Preprint, see (www. math. siu. edu/olive/preprints. htm)*;
[10]**Olive, D. J. and Hawkins, D. M. (2010),** ***Robust Multivariate Location and Dispersion****.Preprint, see (www. math. siu. edu/olive/preprints. htm)*;
[11]**Rousseeuw, P. J. and Driessen, K. V. (1999),** ***A Fast Algorithm for the Minimum Covariance Determinant Estimator***. *Technometrics, 41*(3), 212-223;
[12]**Rousseeuw, P. J. (1985),** ***Multivariate Estimation with High Breakdown Point.*** *Mathematical statistics and applications, 8*, 283-297;
[13]**Rousseeuw, P.J. and Yohai, V. J. (1984),** ***Robust Regression by Means of S-Estimators Robust and Nonlinear Time Series Analysis****, Lecture Notes in Statistics. Heidelberg, Germany: Springer-Verlag*;
[14]**Shevlyakov, G. and Smirnov, P. (2011),** ***Robust Estimation of the Correlation Coefficient: An Attempt of Survey.****Austrian Journal of Statistics, 40(1and2), 147-156*;
[15]**Tukey, J. W. (1960),** ***A Survey of Sampling from Contaminated Distributions.*** **In I. Olkin156**.*Austrian Journal of Statistics*, Vol. 40 (2011), No. 1 and 2, 147–156.

DOI: 10.24818/18423264/53.2.19.13

Table 4. The bias, standard error and root mean square errors of Pearson's, Adj.Winso and RFCH.cor

Case	Method	Bias1	Bias2	Bias3	SE1	SE2	SE3	RMSE1	RMSE2	RMSE3
Without outliers	Pearson's	0.72	0.71	0.71	0.11	0.16	0.16	0.73	0.73	0.73
	Adj.Winso.cor	0.71	0.72	0.71	0.13	0.11	0.13	0.73	0.73	0.72
	RFCH.cor	0.73	0.73	0.73	0.11	0.16	0.16	0.75	0.75	0.75
Case1 (5 VO's)	Pearson's	0.85	0.86	0.86	0.17	0.10	0.10	0.87	0.87	0.87
	Adj.Winso	0.67	0.77	0.93	0.20	0.16	0.21	0.70	0.78	0.95
	RFCH.cor	0.73	0.73	0.74	0.19	0.19	0.15	0.75	0.75	0.76
Case2 (5 LP's in X_1)	Pearson's	0.91	0.71	0.72	0.15	0.16	0.11	0.92	0.73	0.73
	Adj.Winso.cor	0.91	0.72	0.71	0.04	0.11	0.08	0.91	0.73	0.72
	RFCH.cor	0.73	0.73	0.73	0.16	0.16	0.13	0.75	0.74	0.75
Case 3 (5 VO's and 5 LP's in X_1)	Pearson's	0.81	0.85	0.85	0.32	0.17	0.17	0.87	0.86	0.87
	Adj.Winso.cor	0.77	0.77	0.93	0.22	0.13	0.21	0.80	0.78	0.95
	RFCH.cor	0.73	0.73	0.73	0.16	0.16	0.16	0.75	0.75	0.75
Case 4 (5 LP's in X_1 and 5 LP's in X_2)	Pearson's	0.91	0.90	0.71	0.15	0.14	0.16	0.92	0.91	0.72
	Adj.Winso.cor	0.90	0.89	0.71	0.14	0.09	0.13	0.91	0.90	0.72
	RFCH.cor	0.73	0.73	0.73	0.16	0.16	0.13	0.75	0.75	0.75
Case 5 (5 VO's, 5 LP's in X_1 and 5 LP's in X_2)	Pearson's	0.80	0.94	0.85	0.35	0.33	0.17	0.87	0.99	0.87
	Adj.Winso.cor	0.77	0.86	0.93	0.22	0.20	0.19	0.80	0.88	0.95
	RFCH.cor	0.73	0.72	0.72	0.13	0.18	0.18	0.74	0.74	0.74

DOI: 10.24818/18423264/53.2.19.13

Table 5. Sample correlation coefficient, bias and summation of Sample correlation coefficient of Pearson's, Adj.Winso.cor and RFCH.cor for HBK (1984) artificial data.

	Pearson's	Adj.Winso.cor	RFCH.cor
$\hat{\rho}(X_1,Y)$	0.743	0.883	0.098
$\hat{\rho}(X_2,Y)$	0.708	0.92	0.003
$\hat{\rho}(X_3,Y)$	0.757	0.938	-0.181
$\left\|\hat{\rho}(X_1,Y)-\rho_{(x_1,y)}\right\|$	0.823	0.963	0.178
$\left\|\hat{\rho}(X_2,Y)-\rho_{(x_2,y)}\right\|$	0.789	1.001	0.083
$\left\|\hat{\rho}(X_3,Y)-\rho_{(x_3,y)}\right\|$	0.937	1.018	0.101
$M=\sum_{j=1}^{3}\hat{\rho}(X_j,Y)$	2.208	2.742	-0.080
$\lvert M-(-0.08)\rvert$	2.288	2.822	6.96E-05

DOI: 10.24818/18423264/53.2.19.13