


METHODOLOGY ARTICLE

Open Access



An introduction to new robust linear and monotonic correlation coefficients

Mohammad Tabatabai^{1*} , Stephanie Bailey¹, Zoran Bursac², Habib Tabatabai³, Derek Wilus¹ and Karan P. Singh⁴

*Correspondence:

mtabatabai@mmc.edu

¹ Meharry Medical College,
Nashville, TN 37208, USA

Full list of author information
is available at the end of the
article

Abstract

Background: The most common measure of association between two continuous variables is the Pearson correlation (Maronna et al. in Safari an OMC. Robust statistics, 2019. <https://login.proxy.bib.uottawa.ca/login?url=https://learning.oreilly.com/library/view/-/9781119214687/?ar&orpq&email=^u>). When outliers are present, Pearson does not accurately measure association and robust measures are needed. This article introduces three new robust measures of correlation: Taba (T), TabWil (TW), and TabWil rank (TWR). The correlation estimators T and TW measure a linear association between two continuous or ordinal variables; whereas TWR measures a monotonic association. The robustness of these proposed measures in comparison with Pearson (P), Spearman (S), Quadrant (Q), Median (M), and Minimum Covariance Determinant (MCD) are examined through simulation. Taba distance is used to analyze genes, and statistical tests were used to identify those genes most significantly associated with Williams Syndrome (WS).

Results: Based on the root mean square error (RMSE) and bias, the three proposed correlation measures are highly competitive when compared to classical measures such as P and S as well as robust measures such as Q, M, and MCD. Our findings indicate TBL2 was the most significant gene among patients diagnosed with WS and had the most significant reduction in gene expression level when compared with control (P value = 6.37E-05).

Conclusions: Overall, when the distribution is bivariate Log-Normal or bivariate Weibull, TWR performs best in terms of bias and T performs best with respect to RMSE. Under the Normal distribution, MCD performs well with respect to bias and RMSE; but TW, TWR, T, S, and P correlations were in close proximity. The identification of TBL2 may serve as a diagnostic tool for WS patients. A **Taba** R package has been developed and is available for use to perform all necessary computations for the proposed methods.

Keywords: Pearson correlation, Spearman correlation, Quadrant correlation, Median correlation, Minimum covariance determinant correlation, Dissimilarity measures, Gene expression, Williams syndrome



Background

Novel measures of correlation that have noticeably improved performance over existing measures can be a fundamental enhancement to understanding data, affecting a broad range of fields. One of the most widely used statistical measures is the correlation coefficient. The choice of correlation and dissimilarity measures is essential in many areas of science including, but not limited to, clustering co-expressed genes, mediation and moderation analysis with structural equation modeling, time series analysis, pattern recognition, autonomous robots, structural engineering, image recognition, graph theoretical algorithms, spatiotemporal trajectory, artificial intelligence, machine learning techniques, classification, principal component analysis, discriminant analysis, and correlation graphs [1–12]. The need for robust techniques is of utmost significance when dealing with high dimensional biological noisy data. Biological bioassay data frequently contain outliers [13]. Therefore, the choice of the metric can considerably affect the analysis results.

Various resistant dissimilarity measures, such as Tukey's biweight estimate proposed by Hardin et al., are available in the literature, however Pearson (P), Spearman (S), and Euclidian dissimilarity measures are the most commonly used techniques in biomedical research [14, 15]. For standardized vectors X and Y with dimensions n , the Euclidean distance d_{Euclid} is related to Pearson distance $d_{Pearson}$ [16] by the following equation:

$$d_{Euclid} = 2\sqrt{n * d_{Pearson}}.$$

The choice of distance measure to assess outliers plays a vital role in determining the outcome of a wide range of applications [17].

A major difficulty in clustering large data is in the usage of an appropriate dissimilarity measure that captures the geometrical characteristics of those data [18]. Shevlyakov and Pavel Smirnov examined the robustness of correlation coefficient estimators under the assumption of normality at various sample sizes [19]. In a simulation study, Winter et al. concluded that the P correlation coefficient is appropriate when the underlying distribution is light-tailed; but, if outliers are present or the underlying distribution is heavy-tailed, then S correlation coefficient should be used [20]. Using a centroid based algorithm, Shirkorshidi et al. concluded that P correlation performs well at high dimensions but not in low dimensions [21]. Robust correlation was identified as a more useful tool in image-guided surgery applications and image registration in radiotherapy [22].

Pearson dissimilarity measure has frequently been used in the assessment of cell-lines using expression levels or sequence variation profiles genome-wide [23]. Yona et al. studied the quality of some dissimilarity measures used in microarray analysis in order to determine the most effective measure(s) for detecting functional links [24]. A robust complementary hierarchical clustering was introduced to guard against genes with outlying expression levels [25]. Moore et al. utilized the correlation coefficient to examine the association between the quality of visually graded chest images and a quantitative assessment of chest phantom images [26]. Several other studies integrated clever bias-reducing techniques such as drawing from the Weibull distribution in analysis, creating new dissimilarity measures with a normalization factor, and testing the performance of logistic and conventional probabilistic hidden variable models when dealing with gene

expression data [27–29], they claimed that these methods helped to mitigate the negative effects of outliers from the data.

The role of DNA methylation in regulating the expression of oncogenes and progression of cancer types also has been found to generate many outliers. A robust correlation coefficient is a vital tool for calculating the correlation between DNA methylation and gene expression in epigenetic studies when outliers are present [30, 31]. The use of an improper correlation can result in a variety of patterns that produce conflicting results regarding gene expression [30]. Nishimura et al. assessed whether the volume of infused crystalloid fluid is correlated to the amount of interstitial fluid leakage during surgery [32], and Kim et al. studied whether opioid growth factor receptor expression is correlated with cell proliferation in cancer cells [33].

Bloch et al. found that improvement in gene clustering can be obtained by applying the Median correlation measure when outliers are present [34]. The choice of dissimilarity measure is essential part of the RNA transcriptome data analysis, which can determine similar genes or tissues, leading to the identification of biomarkers of specific diseases and the discovery of new drug interventions [35].

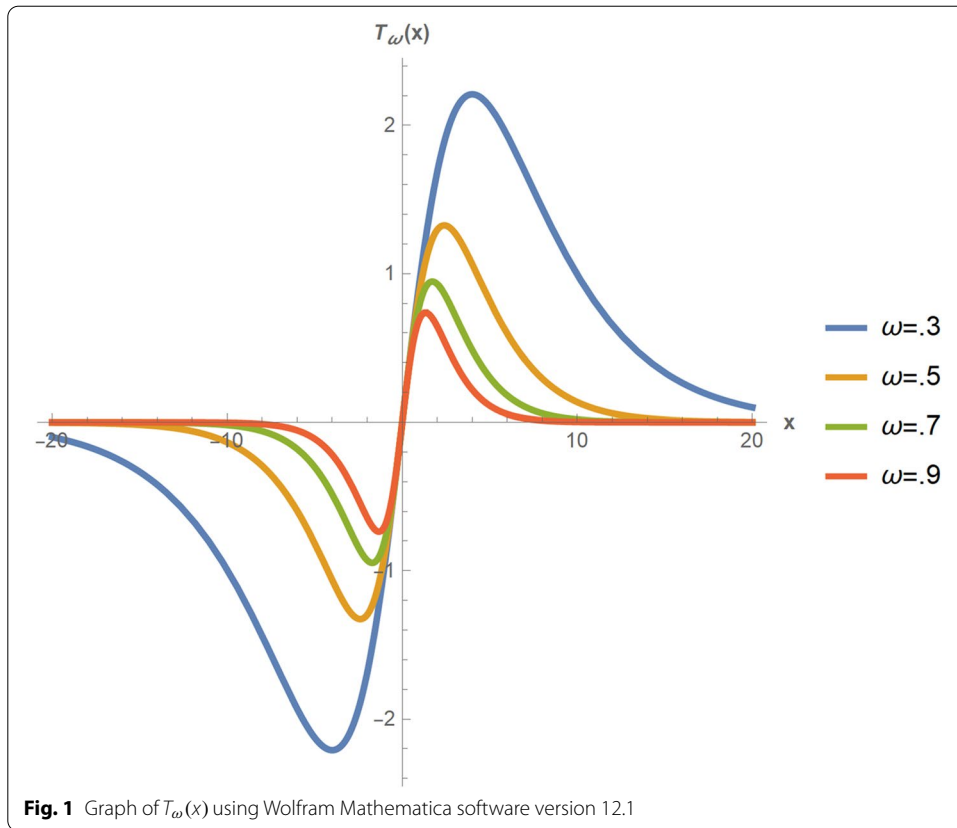
Our simulation results indicate that in the presence of outliers or influential observations, non-robust correlation measures of dissimilarity often result in conclusions that do not represent the true association. We have developed robust linear and monotonic correlation measures capable of giving an accurate estimate of correlation when outliers are present, and reliable estimates when outliers are absent. In this paper, Taba (T), TabWil (TW), and TabWil rank (TWR) correlations are introduced and their robustness are validated by a simulation study in comparison with other widely used correlation estimators.

Methods

Definition 1 The function $T_\omega: \mathbf{R} \rightarrow \mathbf{R}$ is defined as $T_\omega(x) = x * \text{Sech}(\omega * x)$, where Sech is the hyperbolic secant function and ω is the tuning constant. T_ω has the following properties:

- $T_\omega(0) = 0$
- For every real number $x, T_\omega(-x) = -T_\omega(x)$
- For every nonnegative real number $x, T_\omega(x) \geq 0$
- $\frac{d(T_\omega(x))}{d(x)} = 1$ when $x = 0$
- $T_\omega(x) \rightarrow 0$ as $|x| \rightarrow \infty$
- $\frac{d(T_\omega(x))}{d(x)} \rightarrow 0$ as $|x| \rightarrow \infty$
- For every positive real number $k, \frac{T_\omega(kx)}{T_\omega(x)} = k$ as $x \rightarrow 0$
- $T_\omega(x)$ is bounded

Figure 1 depicts the function $T_\omega(x)$ for various values of ω , illustrating the properties mentioned in Definition 1. The value of ω has been calculated using asymptotic efficiency under the assumption of normality [36, 37]. The re-descending property of the function can be seen as $|x|$ approaches infinity. The function T_ω is a bounded influence



function. Due to its properties mentioned in Definition 1, our proposed measures of correlation have high efficiency, a high breakdown point, and will not suffer from masking effects [38, 39].

Robust correlation methods

Taba correlation

For variables X and Y each of size n , we define the Taba robust correlation coefficient r_{Taba} as:

$$r_{Taba}(X, Y) = \frac{\sum_{i=1}^n [T_\omega(C_{1,i}) * T_\omega(C_{2,i})]}{\sqrt{\sum_{i=1}^n [T_\omega(C_{1,i})]^2 * \sum_{i=1}^n [T_\omega(C_{2,i})]^2}},$$

where $C_{1,i} = \frac{x_i - Median(X)}{\hat{\sigma}_{S_n(X)}}$, $C_{2,i} = \frac{y_i - Median(Y)}{\hat{\sigma}_{S_n(Y)}}$. Dispersion measures $\hat{\sigma}_{S_n(X)}$ and $\hat{\sigma}_{S_n(Y)}$ are estimates of the standard deviation for variables X and Y respectively, introduced by Rousseeuw and Croux as a robust scale measurement. Other robust choices such as $\hat{\sigma}_{Q_n(\cdot)}$ are available as an alternative estimate of standard deviation [40, 41]. For the Taba correlation estimator, we set our default value for ω at 0.45 which will give us over 95% in asymptotic efficiency under normality assumptions [36].

TabWil correlation

Let $U = \frac{X - Median(X)}{\hat{\sigma}_{S_n(X)}} + \frac{Y - Median(Y)}{\hat{\sigma}_{S_n(Y)}}$ and $V = \frac{X - Median(X)}{\hat{\sigma}_{S_n(X)}} - \frac{Y - Median(Y)}{\hat{\sigma}_{S_n(Y)}}$.

We define the robust TabWil correlation estimator r_{TabWil} as:

$$r_{TabWil}(X, Y) = \frac{T_\omega(m_1^2) - T_\omega(m_2^2)}{T_\omega(m_1^2) + T_\omega(m_2^2)},$$

where $m_1 = Median(|U|)$ and $m_2 = Median(|V|)$. Both T and TW correlations estimate the linear association between variables X and Y . Figure 2 illustrates the TW correlation coefficient as a function of $Median(|U|)$ and $Median(|V|)$.

TabWil rank correlation

For vectors X and Y , let $R_X = Rank(X)$, $R_Y = Rank(Y)$, where $Rank(X)$ and $Rank(Y)$ refer to the ordinal standing of each element in the vectors X and Y , respectively.

Define

$$D_1 = \frac{R_X - Median(R_X)}{\hat{\sigma}_{S_n(R_X)}} + \frac{R_Y - Median(R_Y)}{\hat{\sigma}_{S_n(R_Y)}} \quad \text{and}$$

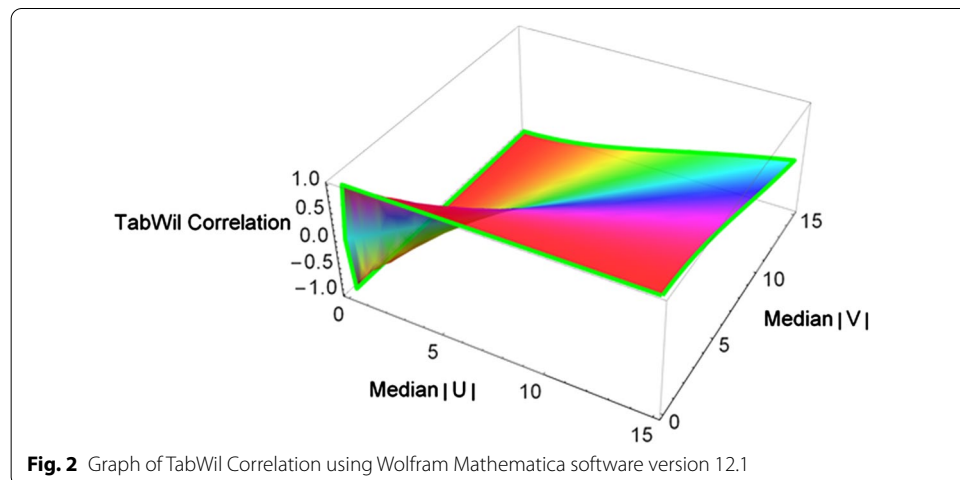
$$D_2 = \frac{R_X - Median(R_X)}{\hat{\sigma}_{S_n(R_X)}} - \frac{R_Y - Median(R_Y)}{\hat{\sigma}_{S_n(R_Y)}}$$

then the robust TabWil rank correlation estimator $r_{TabWilRank}$ is defined as:

$$r_{TabWilRank}(X, Y) = \frac{T_\omega(L_1^2) - T_\omega(L_2^2)}{T_\omega(L_1^2) + T_\omega(L_2^2)},$$

where $L_1 = Median(|D_1|)$ and $L_2 = Median(|D_2|)$ and the default value for ω is 0.05 for both TW and TWR correlations. These values were determined using asymptotic efficiency and outlier tolerance using simulation. There is a trade-off between asymptotic efficiency and outlier tolerance level. In other words, the lower the efficiency, the higher the tolerance level [37, 42].

The TWR correlation estimator measures the monotonic association and direction between two variables X and Y . The TWR correlation can be used with ordinal, interval, or ratio data.



Confidence intervals for proposed Measures

The $(1 - \alpha)100\%$ confidence interval estimator for correlation ρ using any of the three proposed robust measures $(r_{(\cdot)})$ is given by the following lower ($LCL_{(\cdot)}$) and upper ($UCL_{(\cdot)}$) confidence limits:

$$LCL_{(\cdot)} = \text{Tanh} \left(F_{(\cdot)} - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{1 + 0.5r_{(\cdot)}^2}{(n - 3)}} \right)$$

and

$$UCL_{(\cdot)} = \text{Tanh} \left(F_{(\cdot)} + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{1 + 0.5r_{(\cdot)}^2}{(n - 3)}} \right),$$

where $F_{(\cdot)} = \text{ArcTanh}(r_{(\cdot)})$ is the Fisher transformation of robust correlation measure $r_{(\cdot)}$. The symbols $\text{Tanh}(\cdot)$ and $\text{ArcTanh}(\cdot)$ represent the hyperbolic tangent and inverse hyperbolic tangent functions respectively. Bonett and Wright as well as Ruscio studied confidence intervals constructed using $\frac{1+0.5r_{(\cdot)}^2}{(n-3)}$ as an estimate for variance of Fisher transformation. Their results indicate that $(1 - \alpha)100\%$ confidence interval for ρ provide fairly accurate coverage when a robust correlation measure is used [43, 44]. For one sided confidence limits, simply replace $\frac{\alpha}{2}$ by α in the equation for $LCL_{(\cdot)}$ or $UCL_{(\cdot)}$. Alternative methods, such as bootstrapping, are also available for calculating confidence interval estimates [45].

Testing hypothesis for proposed measures

To test the researcher (alternative) hypotheses $H_1 : \rho \neq 0$, $H_1 : \rho > 0$, or $H_1 : \rho < 0$ using any of the three proposed robust correlation measures $r_{(\cdot)}$, one can utilize the test statistic $t_{(n-2)} = r_{(\cdot)} \sqrt{\frac{n-2}{1-0.5r_{(\cdot)}^2}}$ with $n - 2$ degrees of freedom.

Simulation study

The aim of our simulation study is to assess the performance of our proposed methods in comparison with other correlation estimators in the presence and absence of outliers. To achieve this aim,

We have used RStudio version 1.3.1073 utilizing lcmix, robustbase, mvtnorm, Taba, robcor, MethylCapSig, and MultiRNG packages to assess the performance of T, TW, and TWR in comparison with P, S, Quadrant (Q), Median (M), and Minimum Covariance Determinant (MCD) correlation estimators [46, 47]. We generated $m = 5000$ pairs of samples each having size $n = 20, 40, 80, 160, \text{ or } 320$ from one of three distributions: a bivariate normal with mean vector $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$, a bivariate log-normal with mean vector $\begin{pmatrix} 1 \\ 3 \end{pmatrix}$, or a bivariate Weibull with a shape parameter of 1.5. All bivariate distributions had a variance-covariance matrix of the form $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, where the five levels of correlation ρ used in our simulation were set at the 0.0, 0.2, 0.5, 0.7, and 0.9 levels. Random contaminations of our simulated data were generated at the 0%, 5%, and 10% levels. To do this, each iteration randomly drew the appropriate number of observations (based on the

level of contamination) to be corrupted. For each of the selected datapoints, the contaminated datapoints will be equal to the value of the uncontaminated datapoints plus five times the standard deviation of the uncontaminated sample within each iteration (positive shift). The results were similar with a negative shift, but are not shown here. After contamination, the correlation was calculated using each of the eight correlation methods. T, TW, and TWR correlations used tuning constants $\omega = 0.45, 0.05,$ and 0.05 respectively. For comparative purposes, bias and the root mean square errors (RMSE) were calculated for all methods. The bias and the RMSE are defined as:

$$bias = \left| \frac{\sum_{l=1}^m \hat{\rho}_l}{m} - \rho \right|$$

and

$$RMSE = \sqrt{\frac{\sum_{l=1}^m (\hat{\rho}_l - \rho)^2}{m}}.$$

Simulation results

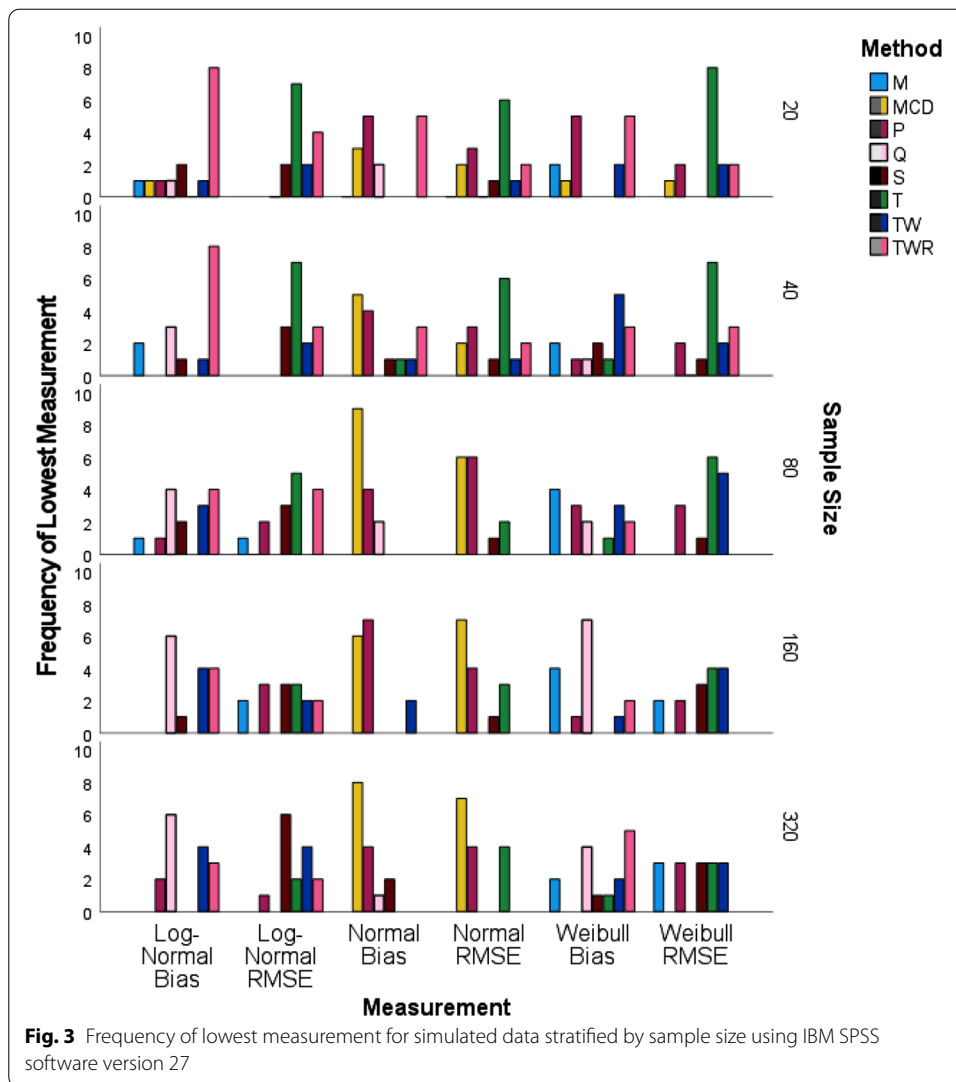
To better understand our simulation results, we ordered the bias and RMSE for each distribution and identified the correlation estimators associated with the ordered results as shown in Additional file 1: Table S1.

Simulation results stratified by sample size

Figure 3 compares the frequency of each correlation method that resulted in having the lowest bias or RMSE in our simulation study, stratified by sample size. For small samples of size 20, TWR or T correlations consistently had the highest frequency of lowest bias and RMSE; tying with P correlation under bias for the bivariate Normal and bivariate Weibull distributions.

For samples of size 40, T correlation uniformly performed best with respect to RMSE under all three bivariate distributions. TWR correlation was shown to have the lowest bias the greatest number of times when the distribution was bivariate Log-Normal, while MCD and TW regularly appeared as having the lowest bias for bivariate Normal and Weibull distributions respectively.

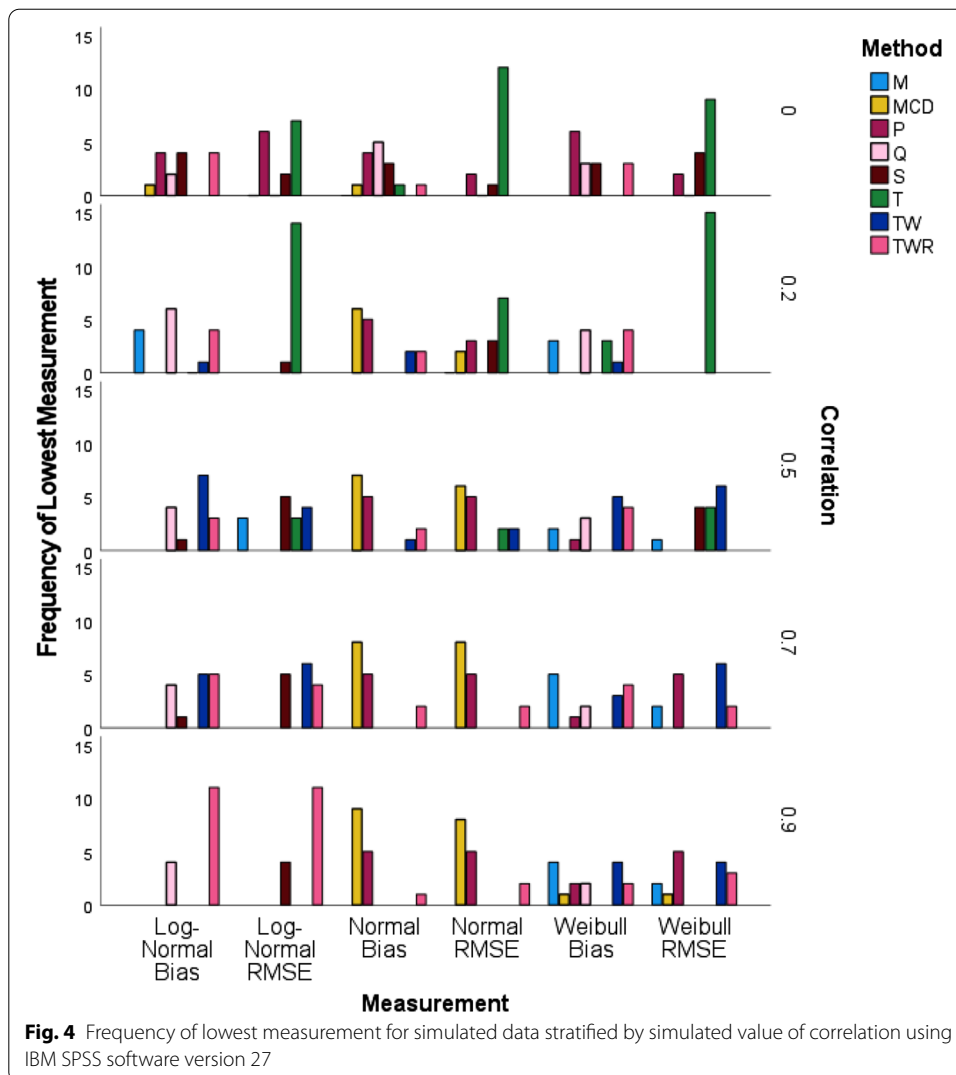
Samples of size 80 or more consistently showed MCD and P correlations having the lowest bias and RMSE under the assumption of a Normal distribution. Assuming the distribution is Weibull, T correlation had the highest frequency of lowest bias or was among methods having the most frequent lowest RMSE. When the distribution was Log-Normal, Q correlation generally had the highest frequency of lowest bias with samples larger than 80, but tied with TWR when the sample size was set to 80. For other non-Normal cases, T correlation performed well with respect to RMSE in Log-Normal distributions, but was overtaken by S and TW correlations when the sample size increased to 320. M correlation performed best with respect to bias when the distribution was bivariate Weibull.



Simulation results stratified by level of correlation (ρ)

Similar to the previous graphic, Fig. 4 depicts the frequency of lowest measurements, this time stratified by the value of correlations. As ρ becomes more positive, the sampling distribution of correlation estimator becomes left skewed. When correlation is set to zero, T correlation performed best in RMSE for all tested bivariate distributions. P correlation performed well overall in terms of bias, but tied with TWR and S correlations when the distribution was bivariate Log-Normal and was overtaken by Q correlation when the distribution was normal.

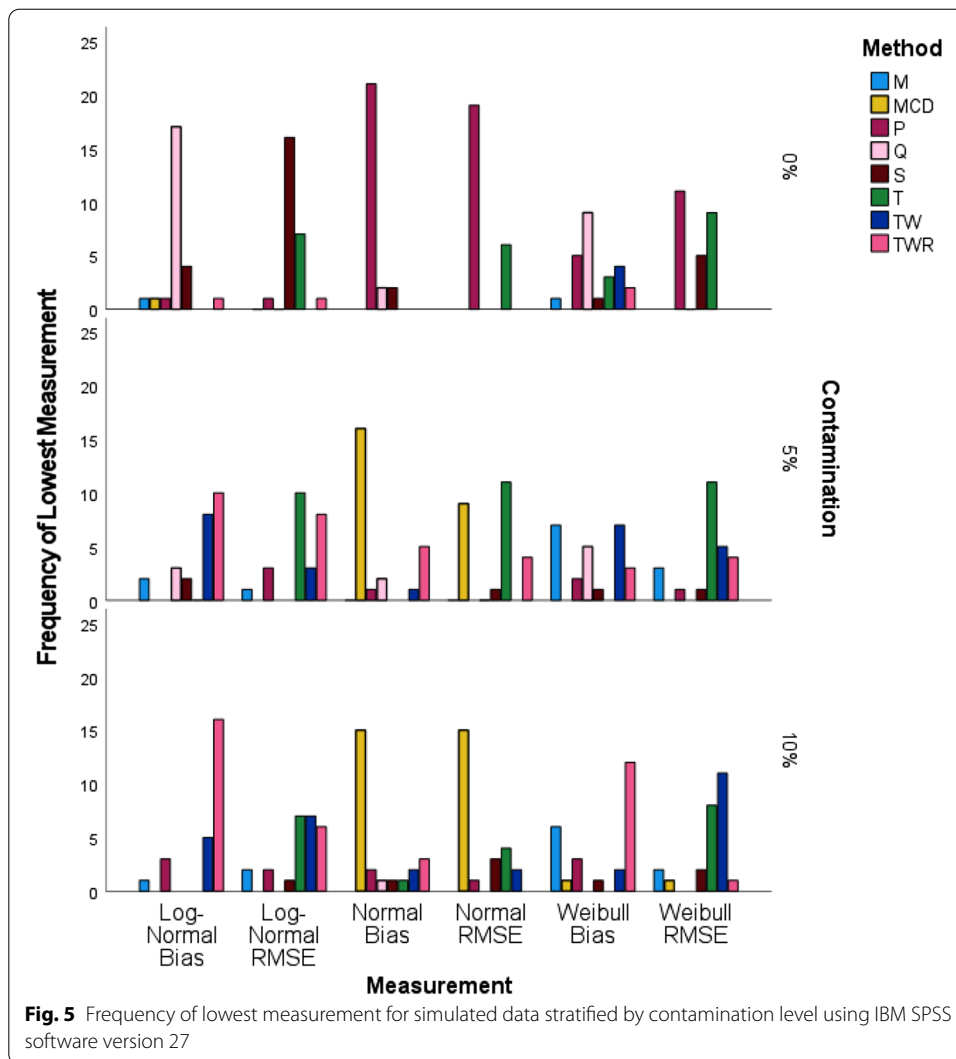
For correlations set at 0.2, T correlation outperformed the other correlation methods for RMSE in all three bivariate distributions. Q performed best with regard to bias in the bivariate Log-Normal, but the MCD performed best when the distribution was bivariate Normal. TWR and Q tied for the best performance in bias when the distribution was bivariate Weibull.



When the correlation level was raised to 0.5, TW performed best in bias for both bivariate Log-Normal and Weibull. MCD performed best with regard to bias and RMSE for bivariate Normal. S and TW performed best in RMSE for bivariate Log-Normal and Weibull distributions, respectively.

At the 0.7 correlation level for the bivariate Normal distribution, MCD performed best with regard to bias and RMSE, but TW performed best in RMSE for both bivariate Weibull and bivariate Log-Normal distributions. M performed best with respect to bias in the bivariate Weibull distribution, while TW and TWR tied for the best bias performance in the bivariate Log-Normal.

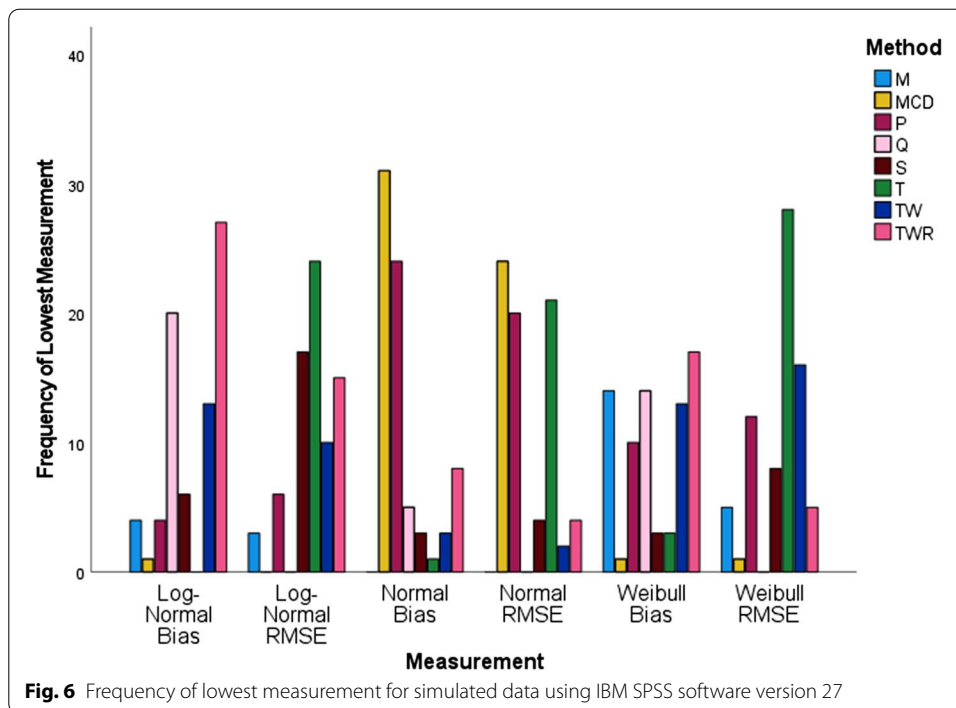
Finally, when the correlation reached the 0.9 level, the best performance in bias and RMSE for both bivariate Log-Normal and Normal belonged to TWR and MCD respectively. TW gave the best performance in terms of RMSE for the bivariate Weibull; TW and M tied for the best performance in bias for the bivariate Weibull distribution.



Simulation results stratified by level of contamination

When the frequency of lowest measurements was stratified by the levels of data contamination, we observed that in the absence of contamination, the best performing bias and RMSE belonged to P correlation. Q correlation had the best performance in bias for both the bivariate Log-Normal and bivariate Weibull distributions. Figure 5 shows that S and P had the best performance with regard to RMSE for bivariate Log-Normal and bivariate Weibull respectively.

At the 5% level of contamination, T had the best performance with regard to RMSE for all three distributions. MCD performed best with regard to bias for the bivariate Normal and TWR had the best performing bias for the bivariate Log-Normal. TW and M tied for the best performance in bias for the bivariate Weibull. Finally, when contamination level reached 10%, MCD performed best in both bias and RMSE for bivariate Normal, while TWR performed best in bias for both the bivariate Log-Normal and bivariate Weibull. TW performed best with respect to RMSE for bivariate Weibull. There was a



tie between TW and T for the best performing RMSE when the distribution was bivariate Log-Normal.

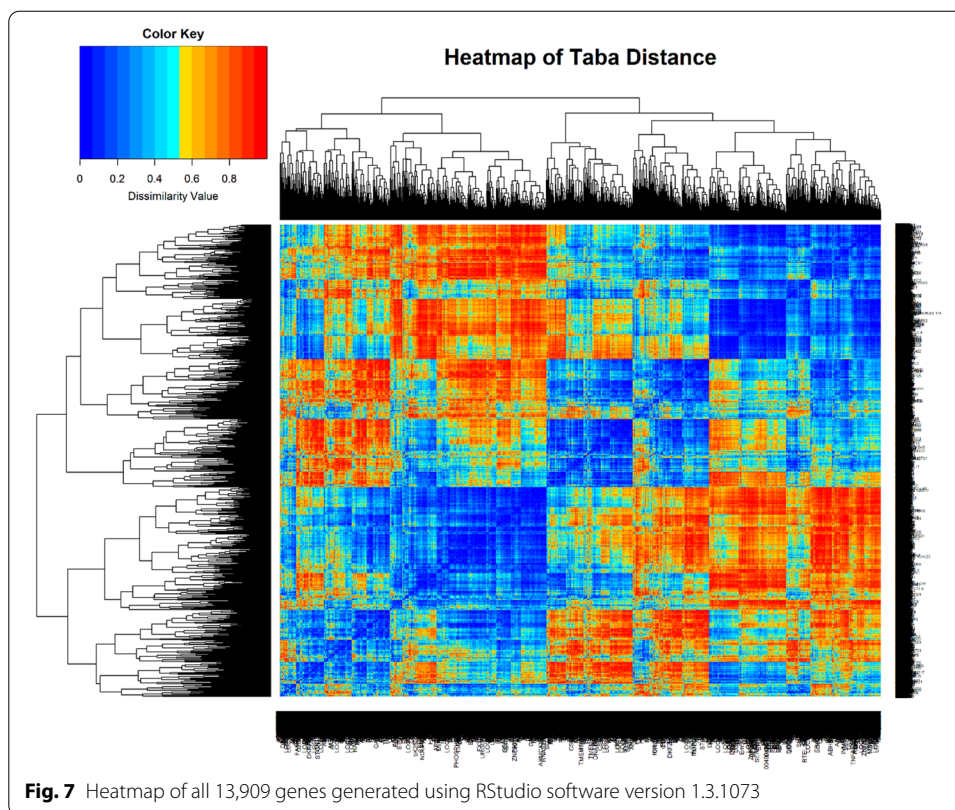
Overall simulation results

Overall, as indicated in Fig. 6, for the bivariate Normal, MCD had the best performance with respect to bias and RMSE, but when the distribution was bivariate Log-Normal or bivariate Weibull, TWR performed best in bias and T had the best performance with respect to RMSE.

Analysis of William syndrome

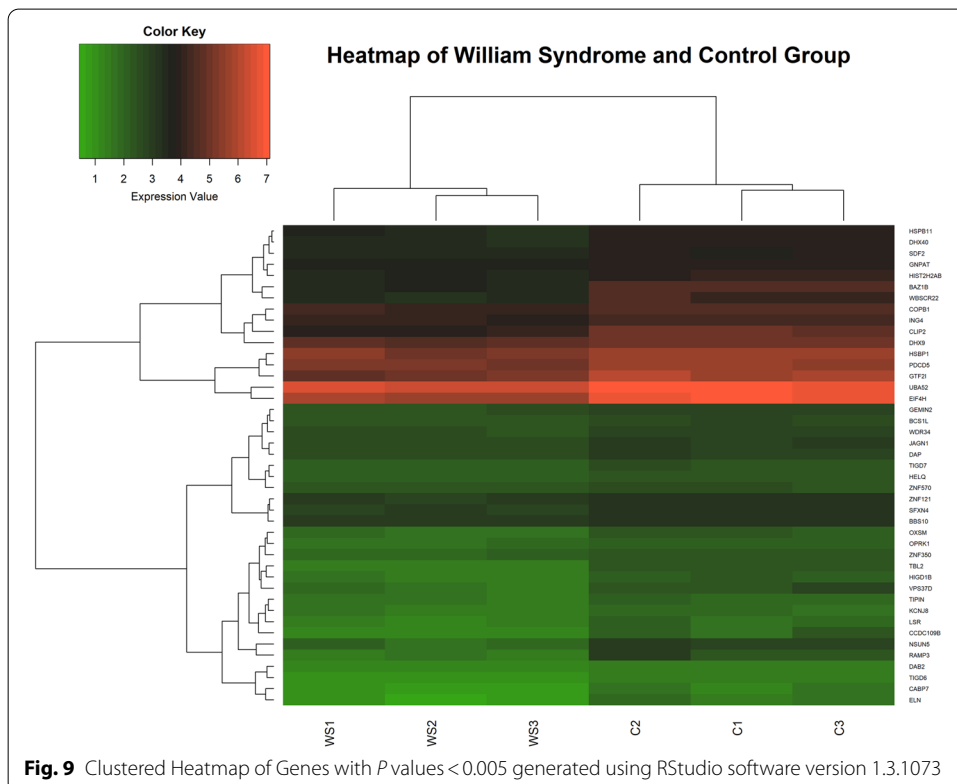
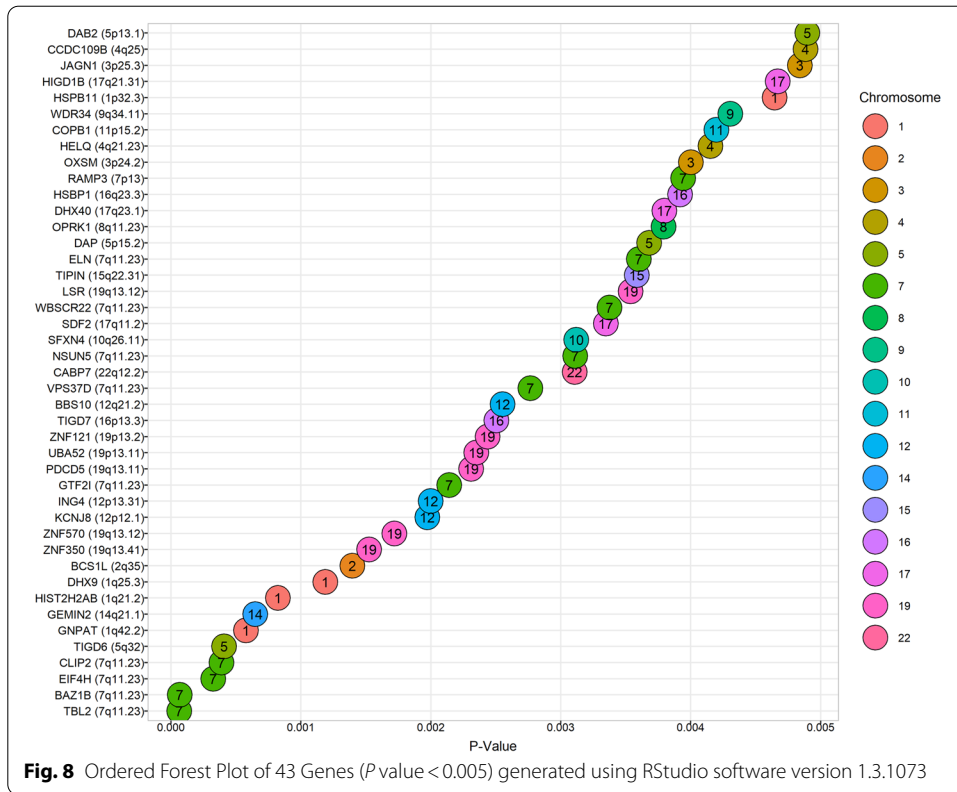
A RNA transcriptome-based dataset of human gene expressions recorded in patients with Williams Syndrome (WS) was obtained from the National Center for Biotechnology Information (NCBI) [48]. Expression levels of genes that appeared in the dataset more than once were combined by averaging expression levels. Any genes with missing values were removed, resulting in a total sample size of 13,909 genes. Tissues were sampled from those with and without WS, each containing three replicated expressions. Silhouette and Elbow graphs were used to determine the optimal number of clusters. Figure 7 shows a visual snapshot of our gene expression data. It is a hierarchical clustering dendrogram for all genes using Taba distance.

After careful examination of data and checking the validity of assumptions, a one-sided t-test was conducted for each of the 13,909 individual genes to determine differences between control and WS groups. Statistical analysis of the data indicated only 43 genes had a P value less than 0.005 when control and WS groups were compared. The most statistically significant reduction in gene expression levels was associated with



transducin beta-like 2 (TBL2) gene (P value = $6.37E-05$). This gene encodes a member of the beta-transducin protein family known to be involved in regulatory functions. This gene is deleted in WS. The 2nd most significant reduction in expression level of the WS group, when compared to control, was observed with the Bromodomain Adjacent to Zinc finger domain, 1B (BAZ1B) gene. BAZ1B plays an important role in neurodevelopment and implicate its haploinsufficiency as a likely contributor to the neurological phenotypes in WS [49]. Eukaryotic translation Initiation Factor 4H (EIF4H) encodes one of the translation initiation factors, which functions to stimulate the initiation of protein synthesis at the level of mRNA utilization. This gene is deleted in WS [50]. The top most significant genes belong to Chromosome 7.

We examined our entire dataset for the presence of outliers and found no significant outliers present. In order to demonstrate the extent of protection of all correlation measures considered in this article against outliers, we selected a random sample of one hundred genes from the set of 13,909, clustered them into two groups using each of the eight correlation estimators, and recorded the genes in each of the two clusters. One of the two groups were randomly selected and 10% of its genes were contaminated. For each selected gene, which consists of six replicates, one replicate was selected at random and contaminated. The contaminated replicate is equal to the value of the uncontaminated replicate plus ten times the standard deviation of the chosen gene prior to contamination. For each correlation measure, genes within both groups were reexamined and compared pre- and post-contamination. T, TW, TWR, and S had a perfect performance and had the same gene clustering results for pre- and post-contamination. Pearson had the



worst performance, misplacing three genes. MCD, M, and Q misplaced only one gene when comparing pre- and post-clustered groups.

Figure 8 shows the ordered forest plot of the top 43 genes having P values less than 0.005. The horizontal axis represents P values and the vertical axis represents genes. The numbers in the forest plot circles represent the chromosome each respective gene belongs to.

Figure 9 illustrates the clustered heatmap using Taba dissimilarity measure for all genes having P values less than 0.005. The hierarchical clustering of samples clearly indicates two groups: the WS and the control. WS samples naturally clustered together and the same was observed for the control samples. The hierarchical clustering of the 43 genes shows at least two clusters. WS group has less intensity in colors when compared to the control group, indicating a significant difference in their expression levels between the two groups. In the cell consisting of WS samples and genes HSPB11, DHX40, SDF2, GNPAT, HIST2H2AB, BAZ1B, and WBSCR22, the most similar pair of genes were HSPB11 and DHX40, which belong to gene class B2 and B3 respectively. As far as we know there is no publication linking these two genes to WS. The next closest gene to these two genes was SDF2. GNPAT and HIST2H2AB were very close in their expression level and BAZ1B and WBSCR22 showed similar expression levels. The next block of genes in the WS category were COPB1, ING4, CLIP2, DHX9, HSBP1, PDCD5, GTF2I, UBA52, and EIF4H. COPB1, ING4, CLIP2, DHX9 and their expression levels showed similar patterns. HSBP1, PDCD5, and GTF2I had similar expression levels. The pair UBA52 and EIF4H were also expressed similarly. The remaining genes had a lower gene expression level when compared with the abovementioned genes.

Taba R package

A statistical R package for calculating proposed robust measures is available (<https://cran.r-project.org/web/packages/Taba/index.html>). This package contains functions that evaluate correlations and their corresponding P values; partial and semi partial correlations; distance (dissimilarity) and P value matrices; as well as estimating generalized partial correlations. For partial, semi-partial, and generalized partial correlations, users will be able to specify the choice of link models such as linear, logistic, and Poisson for each outcome variable. The generalized partial correlation between two variables is similar to partial correlation, but will give the users the opportunity to control for different sets of confounding variables. If the two sets of confounding variables are identical, then the generalized partial correlation will reduce to a partial correlation.

Conclusions

Robustness is a unique quality that not all frequently used measures have. Our work tackles an important issue in the usage of correlation coefficients either directly or indirectly as part of other various disciplines. Although MCD and P typically performed well under normal conditions, not all data follows a Normal distribution, however a majority of gene expression data is not normal [51]. When dealing with small samples from the bivariate Weibull or bivariate Log-Normal distributions the proposed methods are able to more accurately measure association between groups. By using an appropriate robust measure of correlation, one can improve the accuracy of the results and will

enable researchers to better understand the true associations between variables in their models. It is imperative that a robust measure of correlation is used to reduce the severe impact of outliers. Thus, we recommend TabWil and Taba correlation for measuring linear association, and TabWil rank correlation for monotonic association because they are safeguards against the presence of outliers or influential observations.

Overall, MCD performed well based on bias and RMSE when the underlying distribution was bivariate Normal, but TabWil, TabWil rank, Taba, Spearman, and Pearson correlations were in close proximity. When the distribution was bivariate Log-Normal or bivariate Weibull, TabWil rank performed best in terms of bias but Taba performed best with respect to RMSE. Simulation results indicate that the proposed methods are highly robust and capable of determining the dissimilarity in large genomic datasets with thousands of genes, and hundreds of tissues.

Taba robust measure of distance was used to cluster genes using WS gene expression data. When comparing WS with the control group, TBL2 had the most significant reduction in its expression level. The gene TBL2, a possible regulator of the endoplasmic reticulum-resident kinase pathway expressed in a variety of organs such as the heart, skeletal muscle, and several endocrine tissues, can negatively affect nutrient conditions when deleted or under stress [52, 53]. This protein is often deleted in those diagnosed with WS [54, 55]. Other highly correlated genes such as BAZ1B, EIF4H, and CLIP2 are shown to be linked to conditions having similar effects [49, 56, 57].

TBL2 is associated with the eukaryotic 60S ribosomal subunit. This association was endoplasmic reticulum (ER) stress independent, but the TBL2-PERK (PKR-like ER-resident kinase) interaction occurred upon ER stress. This may help in understanding how TBL2 plays a role in the expression of proteins under ER stress [58]. Under ER stress, TBL2 partakes in Activating Transcription Factor 4 (ATF4) translation through its association with mRNA [59]. Furthermore, the deletion of TBL2, along with ER stress or poor nutrient conditions, can lead to impaired ATF4 induction. Thus, TBL2 serves as a potential regulator of the PERK pathway [52]. Due to the fact that haploinsufficiency has been shown for other Beta-Transducin repeat (WD-repeat) containing proteins, hemizygosity of TBL2 may have an impact on some aspects of the WS phenotype [60]. TBL2 has also been known to be highly associated with neurological syndromes [61]. Results suggest that TBL2 along with the 42 most significant genes identified in this study may serve as a diagnostic factor for WS. Future work includes investigating the robustness of the proposed methods in medical imaging and image recognition.

Abbreviations

P: Pearson correlation; S: Spearman correlation; Q: Quadrant correlation; M: Median correlation; MCD: Minimum covariance determinant correlation; T: Taba correlation; TW: TabWil correlation; TWR: TabWil rank correlation; WS: Williams syndrome; RMSE: Root mean square error; TBL2: Transducin Beta like 2; DNA: Deoxyribonucleic acid; Sech: Hyperbolic secant; Tanh: Hyperbolic tangent; ArcTanh: Inverse hyperbolic tangent; NCBI: National Center for Biotechnology Information; RNA: Ribonucleic acid; mRNA: Messenger ribonucleic acid; BAZ1B: Bromodomain Adjacent to Zinc finger domain, 1B; EIF4H: Eukaryotic translation Initiation Factor 4H; HSPB11: Heat Shock Protein Family B (Small) Member 11; DHX40: DEAH-Box Helicase 40; SDF2: Stromal Cell Derived Factor 2; GNPAT: Glyceronephosphate O-Acyltransferase; HIST2H2AB: H2A Clustered Histone 21; BAZ1B: Bromodomain adjacent to zinc finger domain, 1B; WBSCR22: Williams Beuren Syndrome Chromosome Region 22; COPB1: COPI Coat Complex Subunit Beta 1; ING4: Inhibitor of Growth Family Member 4; CLIP2: CAP-Gly Domain Containing Linker Protein 2; DHX9: DExH-Box Helicase 9; HSBP1: Heat Shock Factor Binding Protein 1; PDCD5: Programmed Cell Death 5; GTF2I: General Transcription Factor Ii; UBA52: Ubiquitin Carboxyl Extension Protein 52; EIF4H: Eukaryotic Translation Initiation Factor 4H; ING4: Inhibitor of Growth Family Member 4; CLIP2: CAP-Gly Domain Containing Linker Protein 2; DHX9: DExH-Box Helicase 9; ER: Endoplasmic Reticulum; PERK: PKR-like ER-resident kinase; ATF4: Activating Transcription Factor 4; WD-repeat: Beta-Transducin repeat.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12859-021-04098-4>.

Additional file 1. Simulation results produced using RStudio version 1.3.1073.

Acknowledgements

Not applicable.

Authors' contributions

MT and HT developed the robust methods introduced in the paper. DW performed the simulation on gene expression as well as clustering WS data. MT and DW wrote the manuscript. SB, ZB, HT, and KS contributed to writing and editing the manuscript. Graphics and visual design were performed by DW. All authors read and approved the final manuscript.

Funding

The project has been partially supported by Meharry Medical College RCMI grant (NIH Grant MD007586). This information or content and conclusions are those of the authors and should not be construed as the official position or policy of, nor should any endorsements be inferred by HHS or the US Government. The funder played no role in the design of the study, analysis of the data, or writing the manuscript.

Availability of data and materials

Data can be found on the National Center for Biotechnology Information website (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE128840>). The Taba R Package is available on <https://cran.r-project.org/web/packages/Taba/index.html>.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors have no competing interests.

Author details

¹ Meharry Medical College, Nashville, TN 37208, USA. ² Department of Biostatistics, Florida International University, Miami, FL 33199, USA. ³ Department of Civil and Environmental Engineering, University of Wisconsin Milwaukee, Milwaukee, WI 53211, USA. ⁴ Department of Epidemiology and Biostatistics, University of Texas Health Sciences Center at Tyler, Tyler, TX 75708, USA.

Received: 17 September 2020 Accepted: 22 March 2021

Published online: 31 March 2021

References

- Ben-Dor A, Shamir R, Yakhini Z. Clustering gene expression patterns. *J Comput Biol*. 1999;6:281–97.
- Bezuidenhout CN, Domleo RR. A demonstration of correlation graphs to human body dimensions. *Sci Res Essays*. 2013;9:1273–81.
- Fujita A, Takahashi DY, Balardin JB, Sato JR. Correlation between graphs with an application to brain networks analysis. 2015. [arXiv:1512.06830](https://arxiv.org/abs/1512.06830) [q-bio, stat]. Accessed 12 Jan 2020.
- Iwasaki Y, Kusne AG, Takeuchi I. Comparison of dissimilarity measures for cluster analysis of X-ray diffraction data from combinatorial libraries. *NPJ Comput Mater*. 2017;3:4.
- Jay JJ, Eblen JD, Zhang Y, Benson M, Perkins AD, Saxton AM, et al. A systematic comparison of genome-scale clustering algorithms. *BMC Bioinform*. 2012;13(Suppl 10):S7.
- Lin W-T, Wu Y-C, Cheng A, Chao S-J, Hsu H-M. Engineering properties and correlation analysis of fiber cementitious materials. *Materials*. 2014;7:7423–35.
- Neto AM, Victorino AC, Fantoni I, Zampieri DE, Ferreira JV, Lima DA. Image processing using Pearson's correlation coefficient: applications on autonomous robotics. In: 2013 13th international conference on autonomous robot systems. Lisbon, Portugal: IEEE; 2013. p. 1–6. <https://doi.org/10.1109/Robotica.2013.6623521>.
- Preacher KJ, Zhang Z, Zyphur MJ. Multilevel structural equation models for assessing moderation within and across levels of analysis. *Psychol Methods*. 2016;21:189–205.
- Snape P, Pszczolkowski S, Zafeiriou S, Tzimiropoulos G, Ledig C, Rueckert D. A robust similarity measure for volumetric image registration with outliers. *Image Vis Comput*. 2016;52:97–113.
- Suzuki Y, Hino H, Kotsugi M, Ono K. Automated estimation of materials parameter from X-ray absorption and electron energy-loss spectra with similarity measures. *NPJ Comput Mater*. 2019;5:39.
- Vlachos M, Gunopulos D, Kollios G. Robust similarity measures for mobile object trajectories. In: Proceedings. 13th international workshop on database and expert systems applications. Aix-en-Provence, France: IEEE Comput. Soc.; 2002. p. 721–6. <https://doi.org/10.1109/DEXA.2002.1045983>.

12. Zhao W, Chellappa R, Phillips PJ, Rosenfeld A. Face recognition: a literature survey. *ACM Comput Surv*. 2003;35:399–458.
13. Yellowlees A, Bursa F, Fleetwood KJ, Charlton S, Hirst KJ, Sun R, et al. The appropriateness of robust regression in addressing outliers in an anthrax vaccine potency test. *Bioscience*. 2016;66:63–72.
14. Hardin J, Mitani A, Hicks L, VanKoten B. A robust measure of correlation between two genes on a microarray. *BMC Bioinform*. 2007;8:220.
15. Mukaka MM. Statistics corner: a guide to appropriate use of correlation coefficient in medical research. *Malawi Med J*. 2012;24:69–71.
16. Gentleman R, Ding B, Dudoit S, Ibrahim J. Distance measures in DNA microarray data analysis. In: Gentleman R, Carey VJ, Huber W, Irizarry RA, Dudoit S, editors. *Bioinformatics and computational biology solutions using R and bioconductor*. New York: Springer; 2005. p. 189–208. https://doi.org/10.1007/0-387-29362-0_12.
17. Jaskowiak PA, Campello RJ, Costa IG. On the selection of appropriate distances for gene expression data clustering. *BMC Bioinform*. 2014;15:S2.
18. Guan J, Hsieh F, Koehl P. DCG++: a data-driven metric for geometric pattern recognition. *PLoS ONE*. 2019;14:e0217838.
19. Shevlyakov G, Smirnov P. Robust estimation of the correlation coefficient: an attempt of survey. *Aust J Stat*. 2011;40:10.
20. de Winter JCF, Gosling SD, Potter J. Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: a tutorial using simulations and empirical data. *Psychol Methods*. 2016;21:273–90.
21. Shirshorshidi AS, Aghabozorgi S, Wah TY. A comparison study on similarity and dissimilarity measures in clustering continuous data. *PLoS ONE*. 2015;10:e0144059.
22. Kim J, Fessler JA. Intensity-based image registration using robust correlation coefficients. *IEEE Trans Med Imaging*. 2004;23:1430–44.
23. Mohammad TA, Tsai YS, Ameer S, Chen H-H, Chiu Y-C, Chen Y. CeL-ID: cell line identification using RNA-seq data. *BMC Genomics*. 2019;20:81.
24. Yona G, Dirks W, Rahman S, Lin DM. Effective similarity measures for expression profiles. *Bioinformatics*. 2006;22:1616–22.
25. Badsha MB, Mollah MNH, Jahan N, Kurata H. Robust complementary hierarchical clustering for gene expression data analysis by β -divergence. *J Biosci Bioeng*. 2013;116:397–407.
26. Moore CS, Wood TJ, Beavis AW, Saunderson JR. Correlation of the clinical and physical image quality in chest radiography for average adults with a computed radiography imaging system. *BJR*. 2013;86:20130077.
27. Wang H, Wang Z, Li X, Gong B, Feng L, Zhou Y. A robust approach based on Weibull distribution for clustering gene expression data. *Algorithms Mol Biol*. 2011;6:14.
28. Ray SS, Bandyopadhyay S, Pal SK. Dynamic range-based distance measure for microarray expressions and a fast gene-ordering algorithm. *IEEE Trans Syst Man Cybern B*. 2007;37:742–9.
29. Hasan MN, Rana MM, Begum AA, Rahman M, Mollah MNH. Robust co-clustering to discover toxicogenomic biomarkers and their regulatory doses of chemical compounds using logistic probabilistic hidden variable model. *Front Genet*. 2018;9:516.
30. Spainhour JC, Lim HS, Yi SV, Qiu P. Correlation patterns between DNA methylation and gene expression in the cancer genome atlas. *Cancer Inform*. 2019;18:117693511982877.
31. Córdova-Palomera A, Palma-Gudiel H, Forés-Martos J, Tabarés-Seisdedos R, Fañanás L. Epigenetic outlier profiles in depression: a genome-wide DNA methylation analysis of monozygotic twins. *PLoS ONE*. 2018;13:e0207754.
32. Nishimura A, Tabuchi Y, Kikuchi M, Masuda R, Goto K, Iijima T. The amount of fluid given during surgery that leaks into the interstitium correlates with infused fluid volume and varies widely between patients. *Anesth Anal*. 2016;123:925–32.
33. Kim JY, Ahn HJ, Kim JK, Kim J, Lee SH, Chae HB. Morphine suppresses lung cancer cell proliferation through the interaction with opioid growth factor receptor: an in vitro and human lung tissue study. *Anesth Anal*. 2016;123:1429–36.
34. Bloch KM, Arce GR. Median correlation for the analysis of gene expression data. *Signal Process*. 2003;83:811–23.
35. Liu L, Hawkins DM, Ghosh S, Young SS. Robust singular value decomposition analysis of microarray data. *Proc Natl Acad Sci USA*. 2003;100:13167–72.
36. Rousseeuw PJ, Hubert M. Robust statistics for outlier detection. *WIREs Data Min Knowl Discov*. 2011;1:73–9.
37. Eby W, Li T, Bae S, Singh K. TELBS robust linear regression method. *OAMS*. 2012:65.
38. Maronna R, Martin R, Yohai V, Salibián-Barrera M, Safari an OMC. Robust statistics. 2nd ed. 2019. [https://www.wiley.com/en-us/Robust+Statistics:+Theory+and+Methods+\(with+R\),+2nd+Edition-p-9781119214687](https://www.wiley.com/en-us/Robust+Statistics:+Theory+and+Methods+(with+R),+2nd+Edition-p-9781119214687). Accessed 23 Jan 2020.
39. Shevlyakov G, Morgenthaler S, Shurygin A. Redescending M-estimators. *J Stat Plan Inference*. 2008;138:2906–17.
40. Rousseeuw PJ, Croux C. Alternatives to the median absolute deviation. *J Am Stat Assoc*. 1993;88:1273–83.
41. Croux C, Rousseeuw PJ. Time-efficient algorithms for two highly robust estimators of scale. In: Dodge Y, Whittaker J, editors. *Computational Statistics*. Heidelberg: Springer; 1992. p. 411–28. https://doi.org/10.1007/978-3-662-26811-7_58.
42. Rousseeuw PJ, Leroy AM. Robust regression and outlier detection. Hoboken: Wiley; 2003.
43. Bonett DG, Wright TA. Sample size requirements for estimating pearson, kendall and spearman correlations. *Psychometrika*. 2000;65:23–8.
44. Ruscio J. Constructing confidence intervals for Spearman's rank correlation with ordinal data: a simulation study comparing analytic and bootstrap methods. *J Mod App Stat Meth*. 2008;7:416–34.
45. Bishara AJ, Hittner JB. Confidence intervals for correlations when data are not normal. *Behav Res*. 2017;49:294–309.
46. Raymaekers J, Rousseeuw PJ. Fast robust correlation for high-dimensional data. *Technometrics*. 2019;2019:1–15.
47. Rousseeuw PJ, Driessen KV. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*. 1999;41:212–23.
48. Barak B, Zhang Z, Liu Y, Nir A, Trangle SS, Ennis M, et al. Neuronal deletion of Gtf2i, associated with Williams syndrome, causes behavioral and myelin alterations rescuable by a remyelinating drug. *Nat Neurosci*. 2019;22:700–8.

49. Lalli MA, Jang J, Park J-HC, Wang Y, Guzman E, Zhou H, et al. Haploinsufficiency of BAZ1B contributes to Williams syndrome through transcriptional dysregulation of neurodevelopmental pathways. *Hum Mol Genet.* 2016;25:1294–306.
50. De Cegli R, Iacobacci S, Fedele A, Ballabio A, di Bernardo D. A transcriptomic study of Williams–Beuren syndrome associated genes in mouse embryonic stem cells. *Sci Data.* 2019;6:262.
51. de Torrenté L, Zimmerman S, Suzuki M, Christopheit M, Grealley JM, Mar JC. The shape of gene expression distributions matter: how incorporating distribution shape improves the interpretation of cancer transcriptomic data. *BMC Bioinform.* 2020;21:562.
52. Tsukumo Y, Tsukahara S, Furuno A, Iemura S, Natsume T, Tomida A. TBL2 is a novel PERK-binding protein that modulates stress-signaling and cell survival during endoplasmic reticulum stress. *PLoS ONE.* 2014;9:e112761.
53. Fisch GS. Genetics and genomics of neurobehavioral disorders. Totowa: Humana Press; 2003.
54. TBL2 transducin beta like 2 [Homo sapiens (human)]. National Center for Biotechnology Information; 2020. https://www.ncbi.nlm.nih.gov/gene/26608?_ga=2.241965378.1379159307.1606244325-79102781.1606244325#bibliography.
55. Meng X, Lu X, Li Z, Green ED, Massa H, Trask BJ, et al. Complete physical map of the common deletion region in Williams syndrome and identification and characterization of three novel genes. *Hum Genet.* 1998;103:590–9.
56. Capossela S, Muzio L, Bertolo A, Bianchi V, Dati G, Chaabane L, et al. Growth defects and impaired cognitive-behavioral abilities in mice with knockout for Eif4h, a gene located in the mouse homolog of the Williams–Beuren syndrome critical region. *Am J Pathol.* 2012;180:1121–35.
57. Vandeweyer G, Van der Aa N, Reyniers E, Kooy RF. The contribution of CLIP2 haploinsufficiency to the clinical manifestations of the Williams–Beuren syndrome. *Am J Hum Genet.* 2012;90:1071–8.
58. Tsukumo Y, Tsukahara S, Furuno A, Iemura S, Natsume T, Tomida A. The endoplasmic reticulum-localized protein TBL2 interacts with the 60S ribosomal subunit. *Biochem Biophys Res Commun.* 2015;462:383–8.
59. Tsukumo Y, Tsukahara S, Furuno A, Iemura S, Natsume T, Tomida A. TBL2 associates with ATF4 mRNA via its WD40 domain and regulates its translation during ER stress: TBL2 regulates translation of ATF4 during ER stress. *J Cell Biochem.* 2016;117:500–9.
60. Pérez Jurado LA, Wang Y-K, Francke U, Cruces J. TBL2, a novel transducin family member in the WBS deletion: characterization of the complete sequence, genomic structure, transcriptional variants and the mouse ortholog. *Cytogenet Genome Res.* 1999;86:277–84.
61. Talwar S, Munson PJ, Barb J, Fiuza C, Cintron AP, Logun C, et al. Gene expression profiles of peripheral blood leukocytes after endotoxin challenge in humans. *Physiol Genomics.* 2006;25:203–15.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

