12-2020

Comparative Evaluation of Statistical Dependence Measures

Eman Abdel Rahman Ibrahim
University of Arkansas, Fayetteville

Citation

Ibrahim, E. A. (2020). Comparative Evaluation of Statistical Dependence Measures. *Graduate Theses and Dissertations* Retrieved from https://scholarworks.uark.edu/etd/3903

Comparative Evaluation of Statistical Dependence Measures

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Statistics and Analytics

by

Eman Abdel Rahman Ibrahim
University of Arkansas
Bachelor of Arts in Mathematics, 2015

December 2020
University of Arkansas

This thesis is approved for recommendation to the Graduate Council.

Qingyang Zhang, PhD.
Thesis Director

Mark E. Arnold, PhD.
Committee Member

Giovanni Petris, PhD.
Committee Member

Abstract

Measuring and testing dependence between random variables is of great importance in many scientific fields. In the case of linearly correlated variables, Pearson's correlation coefficient is a commonly used measure of the correlation strength. In the case of nonlinear correlation, several innovative measures have been proposed, such as distance-based correlation, rank-based correlations, and information theory-based correlation. This thesis focuses on the statistical comparison of several important correlations, including Spearman's correlation, mutual information, maximal information coefficient, biweight midcorrelation, distance correlation, and copula correlation, under various simulation settings such as correlative patterns and the level of random noise. Furthermore, we apply those correlations with the overall best performance to a real genomic data set, to study the co-expression between genes in serous ovarian cancer.

Keywords: *Pearson's correlation, copula correlation, distance correlation, maximal information coefficient correlation, mutual information, and Spearman's correlation*

Acknowledgment

“Nothing is more beautiful than the smile that has struggled through tears.” Success often seemed possible, but a faint dream. However, with perseverance and the support of family and friends, I am more than grateful to say, that faint dream is now my reality; I have reached the finish line. I have always thought of myself as a self-motivated woman with a motive to succeed, but the foundation created by the love of my family is what keeps me strong.

First and foremost, I would like to thank God Almighty for giving me the strength, knowledge, ability, and opportunity to undertake this research study and to persevere and complete it satisfactorily. Without His blessings, this achievement would not have been possible.

I am deeply indebted to my life partner, my husband Yousef Ibrahim, for putting up with my long hours sitting in the office and for providing guidance and a sounding board when needed. My husband encourages me and has always said there is a light at the end of the tunnel and I just have to keep one foot in front of the other to eventually see it. I am deeply appreciative of his fostering. I also want to thank my supportive children, Waleed, Somaya, Ammar, Hamzah, and Muhammad who endured my long sleepless nights while studying.

Also, my parents are my heroes and inspiration. My parents desired nothing more than a brighter future for me. I was urged to strive higher and be one of the top scholars. Even though my father is not with me today, I want to thank him for being the cornerstone of any success in my career. I am also grateful to my mother. She has been the greatest source of support, love, and tenderness of which I am deeply thankful. Her constant smiles and prayers created a strong foothold for me to become the person I am today. Mom, I want you to know you are a blessing, and it’s so heart-warming to have your support. I am very proud to have you as my mother.

And, I am so thankful to my sisters and brothers Marwan, Emad, Manal, Maysa’a, and Hadeel Almasri who sent their inspirations to lighten my heart and remind me that challenges are “mind over matter.”

Furthermore, I am thankful to the Department of Mathematical Sciences and its entire staff for all the considerate guidance. I would like to express my sincere gratitude to my instructors in the Mathematics Department who assured me I was capable, asking me to eliminate the word “impossible” from my vocabulary.

A special thanks to Dr. Qingyang Zhang, my wonderful advisor and mentor, for the optimism, reassurance, thoughtful guidance, and recommendations on this thesis. Dr. Qingyang Zhang, you are an amazing person, and I value all your consistent support and patience which cannot be underestimated. I would like to express my deepest appreciation to my committee members; Dr. Giovanni Petris and Dr. Mark Arnold for serving on my master thesis discussion. I have benefited extremely from their insight, expertise, and knowledge of Statistics and Mathematics.

Last but not least, I want to express my gratitude to all of my friends. I am indebted to all the phone calls through my ups and downs and for the uplifting moments that kept me positive and more capable of finishing my master’s degree and looking for a career, not a job.

I am honored that all of these amazing people were here to witness my journey of success. The completion of my thesis would not have been possible without their support and nurturing.

Thank you all for believing in me. I can stand in front of you all and say, nothing is impossible to achieve; dreams can come true. Thank you.

Table of Contents

Chapter 1

Introduction

In many scientific studies, it is of great importance to measure and test the dependence between random variables. Therefore, a powerful statistical dependence measure is essential. Accurate quantification of the correlation between two variables can help make predictions, and in general, when the correlation (linear or nonlinear) is stronger, a more precise prediction can be made. Measuring the dependence between random variables is an effective way to identify their directional movement with each other (Wang et al., 2015), e.g., the hourly electricity consumption vs the hourly temperature, height vs weight, the time spent on marketing business vs the number of new customers, the prices of certain crop products and the available supply of such products.

Pearson's correlation coefficient is the most widely used measure of linear dependence, because of its simplicity and nice statistical properties. It is defined as the quotient of the covariance with the product of their standard deviations. Pearson's correlation is always between -1 and 1, where -1 and 1 indicate a perfect linear relation while 0 indicates no linear relation. Mathematically, if two random variables are independent, they must be uncorrelated, and the coefficient of correlation must be zero. However, two variables are uncorrelated does not necessarily mean they are independent (Wang et al., .2015). In 1895, Karl Pearson proposed the product-moment correlation coefficient, which still serves as the basis of many correlative analyses. However, the major limitation of Pearson's correlation is that it can only measure the linear relation, i.e., it is not sufficient for statistical independence test due to the existence of nonlinear associations. During the past decades, many important measures have been developed targeting different types of associations. This thesis aims to compare some of these important correlative measures, including Spearman's correlation, mutual information (MI), maximal information coefficient (MIC), biweight midcorrelation (bicor), distance correlation (dcor) and copula correlation (Ccor), under various simulation settings such as sample size, correlative patterns (linear or nonlinear relationship) and the level of random noise. Our simulations show that some of these measures such as Spearman's correlations can detect linear and nonlinear monotonic relationships. Some methods, including distance correlation, MIC, and MI, can also identify certain non-monotonic relationships.

1.1 Statistical independence

Two random variables are said to be independent if the outcome of one random variable does not affect the conditional probability of the other. In other words, if the two random variables are independent, one does not affect the distribution of the other. The detection of dependence relies on a measure that is sensitive to the true underlying relation (Martínez-Gómez, Richards & Richards, 2014). In testing the statistical independence between two random variables, the combination of different correlation measures can provide more insights about the underlying association (Zhang, Qi & Ma., 2011).

Suppose we have two continuous random variables X and Y with probability density functions f(x) and f(y) and cumulative distribution functions F(x) and F(y), respectively. Given that the combined random variable (X, Y) exists, the two random variables are said to be independent if their joint density function is equal to the product of their marginal densities, or equivalently the joint cumulative distribution function equals the product of their respective cumulative distribution functions, i.e.,

$$F(x,y) = F(x)F(y), \text{ or } \mathrm{f}(x,y) = \mathrm{f}(x)\mathrm{f}(y)$$

for any x and y in the sampling space.

For discrete or categorical variables, statistical independence is defined in a similar way, but using probability mass function instead of the probability density function.

1.2 Measure of linear dependence

We begin with notations and concepts. Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a random sample of size n from random variables X and Y. The hypothesis testing of dependence between X and Y can be formulated as follows: the null hypothesis is that there is no association between two variables, against the alternative hypothesis that there is an association between two variables, i.e.,

$$\text{H}_0\text{: } X \text{ and } Y \text{ are independent,}$$

$$\text{H}_1\text{: } X \text{ and } Y \text{ are dependent.}$$

By comparing the p-value with the pre-specified significance level α, one may reject or accept the null hypothesis. The statistical hypothesis test is formulated as follows:

$$H_0 : F(X,Y) = F(X)F(Y),$$

$$H_a : F(X,Y) \neq F(X)F(Y),$$

where $F(X), F(Y)$ represents are the cumulative distribution functions of the random variable X and Y, and $F(X,Y)$ is the joint cumulative distribution function of X and Y.

Let X and Y be two univariate random variables, with expectations E(X) and E(Y). Let Var(X) denote the variance of the random variable X, then

$$Var(X) = E(X^2) - \left(E(X)\right)^2$$

and the covariance between X and Y is

$$Cov\,(X,Y) = E(XY) - E(X)E(Y).$$

Pearson’s correlation coefficient between X and Y is defined as

$$\rho_{xy} = \frac{Cov(X,Y)}{\sqrt{Var(x)}\sqrt{Var(Y)}},$$

where it can be easily seen that the Pearson's correlation is a rescaled version of covariance between X and Y (scaled by the product of the standard deviation of X and the standard deviation of Y). Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a random sample of size n, then the sample estimate of Pearson's correlation between X and Y is

$$\frac{\sum_{i=1}^{n}(x_i-\tilde{x})(y_i-\tilde{y})}{\sqrt{\sum_{i=1}^{n}(x_j-\tilde{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\tilde{y})^2}},$$

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

1.3 Linear and nonlinear relations

The association between two random variables can be classified into two categories: linear and nonlinear. In many applications, the nonlinear relationship is equally important as the linear relationship (Ding and Li, 2015). The nonlinear relations can be further classified into monotonic nonlinear relations and non-monotonic nonlinear relations. It is well known that for a monotonic relationship, Spearman's correlation coefficient would be an appropriate measure of association. For non-monotonic relations, however, the detection and measure can be very challenging, and it can be very difficult to decide which method is the most suitable one. Therefore, it is of great interest to test these measures under different correlative patterns.

Figure 1.3.1 gives some examples of correlative patterns, where it can be seen that in several nonlinear especially non-monotonic nonlinear settings, the prevailing Pearson's correlation completely fails to measure the association.

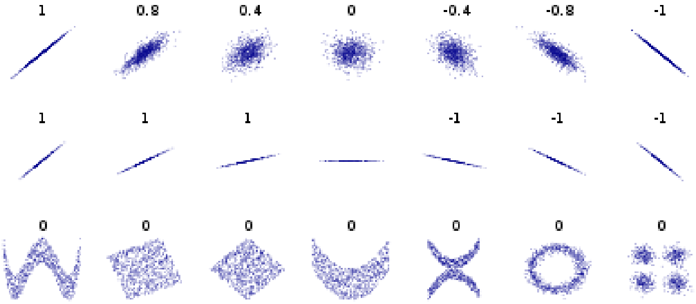
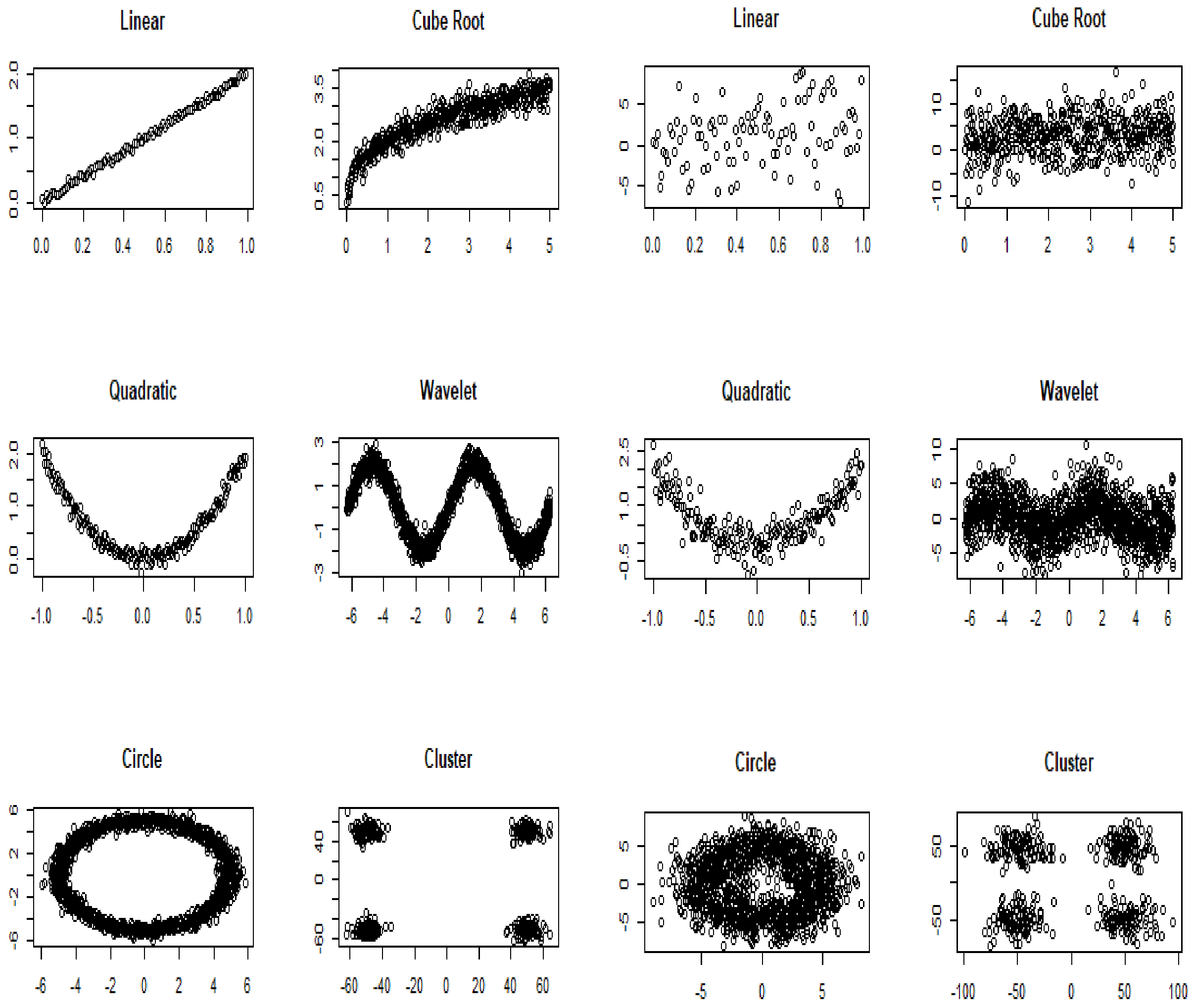


Figure 1.3.1 A small Pearson's correlation coefficient does not indicate independence or weak dependence, as the variables may have a nonlinear relationship.

In this thesis, we aim to compare the statistical performance (in terms of both correlation strength and significance) of six different measures, including Spearman's correlation, mutual information, maximal information coefficient, biweight midcorrelation, distance correlation, and copula correlation, under many different simulation settings, such as linear, cube root, quadratic, wavelet, circle, and cluster (Figure 1.3.2).


(a) Low level of noise

(b) High level of noise

Figure 1.3.2 Correlative patterns such as linear, cube root, quadratic, wavelet, circle, and cluster with different levels of noise.

Table 1.3.1 below lists all the six correlative patterns with equations that we used for simulation studies. It should be noted that all the noise term follows a normal distribution with mean 0 and variance that will be varied in different settings.

Table 1.3.1
Simulation settings considered in this work

Setting	Equation	Domain
Linear	$y = 2x + \varepsilon$	$0 < x < 1$
Cube Root	$y = 20x^{1/3} + \varepsilon$	$0 < x < 1$
Quadratic	$y = 2x^2 + \varepsilon$	$-1 < x < 1$
Wavelet	$y = 2 sin\, x\ + \varepsilon$	$-2\pi < x < 2\pi$
Circle	$X = (5 + \varepsilon_x) \cos\theta$, $y = (5 + \varepsilon_y) \sin\theta$ where ε_x and ε_y are independent	$0 < \theta < 2\pi$
Cluster	$x_1 = -50 + \varepsilon_{x_1}, y_1 = 50 + \varepsilon_{y_1}$ $x_2 = 50 + \varepsilon_{x_2}, y_2 = 50 + \varepsilon_{y_2}$ $x_3 = -50 + \varepsilon_{x_3},\ y_3 = -50 + \varepsilon_{y_3}$ $x_4 = 50 + \varepsilon_{x_4},\ y_4 = -50 + \varepsilon_{y_4}$ where $\varepsilon_{x_1}, \varepsilon_{x_2}, \varepsilon_{x_3}$ and ε_{x_4} are independent, $\varepsilon_{y_1}, \varepsilon_{y_2}, \varepsilon_{y_3}$ and ε_{y_4} are independent	

Chapter 2

Methodology

In this section, we review the definitions and statistical properties of the six selected measures.

2.1 Spearman's rank based correlation

Spearman's correlation coefficient is defined as the correlation of ranks. It is designed to measure the monotonic relation between two variables. Spearman's correlation can be used on both continuous and ordinal categorical data. Similar to Pearson's correlation, Spearman's correlation is always between -1 and 1. It is a negative value if one variable increases as the other decreases (da Costa, 2015). However, unlike Pearson's correlation coefficient, Spearman's correlation does not rely on the normal assumption (Bolboaca & Jantschi. 2006). Let $X = (x_1, \dots, x_n)$ and $Y = (\mathrm{y}_1, \dots, \mathrm{y}_n)$ be a random sample of size n, Spearman's correlation $\boldsymbol{r_s}$ is defined as follows

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

where n is the total number of samples of two variables, and for each random variable, the rank difference of the *i*

2.2 Mutual information

$(X) \geq 0$ is nonnegative. The above definition of entropy extends to a pair of random variables (X, Y) with joint probability density function f (x, y). We define the joint entropy of (X, Y) as H $(X, Y) = -E \ \log f\ (X, Y)$.

Let X and Y be the two random variables with marginal probability density functions as f_1(X) and $f_2(Y)$, respectively. With given Y, the conditional density function of X is $f(x, y)/f_2(y)$ and the conditional entropy is

$$\mathrm{H}\ (X|Y) = -E \ \log \frac{f(X,Y)}{f_2(Y)}$$

Mutual information $I(X, Y)$ calculates the amount of information gained from one random variable (Figure 2.2.1).

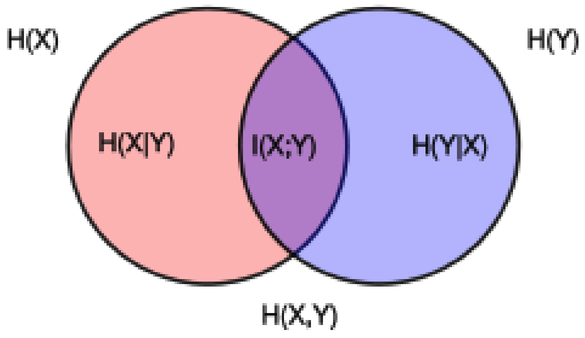


Figure 2.2.1 Venn diagram showing the relationships between MI and entropies (Wikipedia,2019).

$$\begin{aligned} I(X,\ Y) &= H(X) - H(X|Y) \\ &= H(X) + H(Y) - H(X,\ Y) \\ &= \textstyle\sum_x P(\chi)\, log\left(\frac{1}{P(\chi)}\right) + \sum_y P(y)\, log\left(\frac{1}{P(y)}\right) + \sum_{x,y} P\ (x, y) log P(x, y) \\ &= \textstyle\sum_{x,y} P(\chi, y)\, log\left(\frac{1}{P(\chi)}\right) + \sum_{x,y} P(x, y)\, log\left(\frac{1}{P(y)}\right) + \sum_{x,y} P\ (x, y) log P(x, y) \\ &= \textstyle\sum_{x,y} P\ (x, y)\ log\left(\frac{P(x,y)}{P(x)P(y)}\right) \end{aligned}$$

Mutual information (MI) measures the amount of information in units (bits). For discrete random variables with joint probability mass function P (x, y), the MI is defined as

$$I(X,Y) = \sum_{y \in Y} \sum_{\varkappa \in X} P(\chi, y) \log(\frac{P(\chi,y)}{P(\chi)P(y)}).$$

For continuous random variables with joint probability density function f (x, y), the MI can be defined as

$$I(X,Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(X,Y) \log \frac{f(X,Y)}{f_1(X) f_2(y)} dxdy.$$

An equivalent way of defining $I(X,Y)$ between the two variables X and Y is

$$I(X,Y) = H(X) + H(Y) - H(X,Y),$$

where $H(X), H(Y)$ are the entropies of X and Y, and $H(X,Y)$ is the joint entropy between X and Y. The term entropy measures the uncertainty of a random variable.

The entropy and mutual information are related through the following derivation

$$I(X,Y) = \mathrm{E}log \left(\frac{1}{f_1(\mathrm{X})} . \frac{f(X,Y)}{f_2(Y)}\right)$$

$$= \mathrm{E}\ (-log f_1(X) + log \frac{f(X,Y)}{f_2(Y)})$$

$$= -\mathrm{E}log\ f_1(X) + \mathrm{E}log \frac{f(X,Y)}{f_2(Y)}$$

$$= H(X) + H(Y) - H(X,Y).$$

Since H (X, Y) is symmetric, it follows that $I(X, Y) = I(Y, X)$. Hence, the difference in uncertainty about X given knowledge of Y equals the difference in uncertainty about Y given knowledge of X (Kinney & Atwal, 2014). When X and Y are independent, their mutual information is zero. In other words,

$$P(X,Y) = P(X)P(Y)\ or\ \log(\frac{P(X,Y)}{P(X)P(Y)}) = \log 1 = 0.$$

In the case that the two variables are identical, or functionally related, then the information of X reveals everything about Y, and the entropy of the random variable become equivalent to the mutual information, $I(X,Y) = H(X) = H(Y)$

2.3 Maximal information coefficient

Another popular dependence measure is the maximal information coefficient (MIC). Reshef et al. (2011) introduced the notion of maximal information coefficient which could potentially measure both linear and non-linear relationships between variables. Tang et al. (2014) stated the MIC can be useful in the large datasets to measure the associations between the thousands of variable pairs. As it takes values between 0 and 1, MIC could not reflect the directional movement. There are two fundamental properties of MIC, including equitability and generality. Generality indicates that the statistic must capture a wider variety of associations, such as periodic, exponential, or linear, with an adequately larger sample size. Equitability shows that MIC provides similar scores for similarly noisy relationships, irrespective of what type of the relation is.

As the sample size goes to infinity, MIC almost surely gives score of 1 to every functional relationship and gives score of 0 to statistically independent variables. There is not any parametric or distributional assumption in the MIC. MIC is defined by Reshef et al. as the maximum taken over all x-by-y grids G up to a given grid resolution, $\left\{\frac{I\,(\chi,\mathrm{y})}{\log_2 min\{n_\chi,n_\mathrm{y}\}}\right\}$ based on the empirical probability distribution over the boxes of a grid G. For two random variables X and Y having sample n ≥ 2, the MIC is defined as follows

$$\mathrm{MIC} = \max\left\{\frac{I\,(x\,,\mathrm{y})}{\log_2 min\{n_x,n_\mathrm{y}\}}\right\} \ ,$$

where $I(x,\mathrm{y}\,) = H(x) + H(\mathrm{y}\,) - H(x,\mathrm{y}\,)$, i.e.,

$$I(\chi,\mathrm{y}\,) = \sum_{i=1}^{n_\chi}\mathcal{P}\,(\chi_i)\log_2\frac{1}{\mathcal{P}(\chi_i)} + \sum_{i=1}^{n_\mathrm{y}}\mathcal{P}\,(\mathrm{y}_i)\log_2\frac{1}{\mathcal{P}(\mathrm{y}_i)} - \sum_{i=1}^{n_\chi}\sum_{i=1}^{n_\mathrm{y}}\mathcal{P}\,(\chi_i,\mathrm{y}_i)\log_2\frac{1}{\mathcal{P}(\chi_i,\mathrm{y}_i)}$$

where, n_x and n_y represents the bins between the partition of the axes. $n_x \,.\, n_y < B(n), B(n) = n^{0.6}$. Nguyen et al. (2014) pointed out the maximal correlation does not require assumptions on the distribution of data. It appears robust and very efficient, and it can also detect nonlinear correlation.

2.4 Biweight midcorrelation

Biweight midcorrelation (bicor) is based on the measure of similarity between variables. There are two major advantages for bicor. First, the calculation is straightforward, consisting of some simple steps such as the calculation of median. Second, it is more robust to outliers comparing to other measures such as Spearman's correlation (Yuan et al., 2013).

To define the biweight midcorrelation (bicor) of two numeric vectors $x = (x_1, x_2, \ldots x_n)$ and $y = (y_1, y_2, \ldots y_n)$, we must define $a_{i\,,}\, b_i$ with $i = 1,2, \ldots, n,$ where $med(x)$ is the median and $mad(x)$ is the absolute median deviation of x:

$$a_i = \frac{x_i - med(x)}{9mad(x)}$$

Similarly, we define b_i, where $med(y)$ is the median and $mad(y)$ is the absolute median deviation of y:

$$b_i = \frac{y_i - med(y)}{9mad(y)}$$

where $med(x)$ is the median and $mad(x)$ is the absolute median deviation,

$$mad(x) = med(|x_i - med(x)|)$$

These equations are used to define weight, m_i. For X, the weight is defined as

$$m_i^{(x)} = (1 - a_i^2)^2 I(1 - |a_i|),$$

where I is the identity function. Yuan et al. (2013) mentioned that the indicator is 1 when $I(1 - |a_i|) > 0$ and is 0 when $I(1 - |a_i|) \leq 0$. Using the definition of weight to normalize so that the sum of the weights is 1

$$\widetilde{x_i} = \frac{(x_i - med(x))m_i^{(x)}}{\sqrt{\sum_{j=1}^{n}[(x_j - med(x))m_j^{(x)}]^2}}, \qquad \widetilde{y_i} = \frac{(y_i - med(y))m_i^{(y)}}{\sqrt{\sum_{j=1}^{n}[(y_j - med(y))m_j^{(y)}]^2}}$$

$$bicor(x,y) = \frac{\sum_{i=1}^{n}(x_i - med(x))m_i^{(x)}(y_i - med(y))m_i^{(y)}}{\sqrt{\sum_{i=1}^{n}[(x_j - med(x))m_j^{(x)}]^2}\sqrt{\sum_{k=1}^{n}[(y_k - med(y))m_k^{(y)}]^2}}$$

Biweight midcorrelation has many successful applications, for instance, gene co-expression analysis and gene community (clique) detection(Zeng et al., 2013). To study gene co-expression, DNA microarray data have been widely used. Genes and their protein products tend to work in cooperation rather than in isolation. However, most of the existing studies focused on single gene or single type of genetic data and overlooked the interactions between genes and other factors. Maxim clique concept was used to further look into the Signaling pathways involving multiple genes or biomarkers. The most commonly used correlation is Pearson correlation. Other proposed approaches include biweight midcorrelation and half-thresholding strategy. Being more robust to outliers, the biweight midcorrelation has a whip hand over Pearson correlation plus experiments on simulated datasets have proven it to have better performance (Zeng et al., 2013).

2.5 Distance correlation

Distance correlation is a novel measure of dependence between two sets of random variables of arbitrary dimension. The distance correlation between two random vectors X and Y (Székely, Rizzo & Bakirov, 2007) is described as a rescaled distance covariance (same as Pearson's correlation in spirit)

$$dCor(X,Y) = dCov(X,Y)/\sqrt{dCor(X,X)dCor(Y,Y)}$$

where the squared distance covariance is defined as $dCov^2(X,Y) = Cov(\| x_1 - x_2 \|, \| y_1 - y_2 \|) - 2\,Cov(\| x_1 - x_2 \|, \| y_1 - y_2 \|)$, and a natural estimator of $dCov^2(X,Y)\ is$

$\widehat{dCov}^2(X,Y) = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{A_{ij}B_{ij}}{n^2}$,

where $A_{ij} = a_{ij} - \bar{a}_\iota - \bar{a}_J + \bar{a}$ and $B_{ij} = b_{ij} - \bar{b}_\iota - \bar{b}_J + \bar{b}$, if we let $a_{ij} = \left\|X_i - X_j\right\|$,

$\bar{a}_\iota = \sum_{k=1}^{n}\sum_{l=1}^{n}\frac{\|X_R - X_i\|}{n}$, $\bar{a}_J = \sum_{l=1}^{n}\frac{\left\|X_l - X_j\right\|}{n}$, $\bar{a} = \sum_{k=1}^{n}\frac{\|X_l - X_k\|}{n^2}$, let $b_{ij} = \left\|Y_i - Y_j\right\|$,

$\bar{b}_\iota = \sum_{k=1}^{n}\frac{\|Y_R - Y_i\|}{n}$, $\bar{b}_J = \sum_{l=1}^{n}\frac{\left\|Y_l - Y_j\right\|}{n}$, $\bar{b} = \sum_{k=1}^{n}\sum_{l=1}^{n}\frac{\|Y_l - Y_k\|}{n^2}$. The estimate of distance correlation $\widehat{dCor}(X,Y) = \frac{\widehat{dCov}(X,Y)}{\sqrt{\widehat{dCov}(X,X)\widehat{dCov}(Y,Y)}}$.

Two remarkable properties of distance correlation are

1. $0 \le dCor\,(X,Y) \le 1$: In comparison to negative Pearson's correlation, this is always positive.
2. $dCor\,(X,Y) = 0$ if and only if X and Y are independent.

2.6 Copula correlation

Copula correlation is a dependence measure of the deterministic relationship using hidden uniform noise. The copula function for any random vector $\mathrm{X}_1, \mathrm{X}_2, \ldots\ldots \mathrm{X}_n$ is defined as

$$F(x_1, x_2, \ldots\ldots x_n) = C(F_1(x_1), F_2(x_2), \ldots F_n(x_n)),$$

where F stands for the joint cumulative distribution function and $F_1(x_1), F_2(x_2), \ldots F_n(x_n)$ are the marginal cumulative distribution function. By Sklar's theorem (Sklar (1959)), one can decompose the joint distribution function into the copula form of its marginals. Moreover, the joint density is

$$f(x_1, x_2, \ldots\ldots x_n) = f_1(x_1) * \ldots * f_n(x_n) C(F_1(x_1), F_2(x_2), \ldots, F_n(x_n)).$$

Given that F_i and C are differentiable, $C = \partial^n \frac{C}{(\partial F_1 \ldots \partial F_n)}$. Under the limited scenario, the joint probability density function is the product of the copula density and the marginal densities. For example, if the *i* random variables X_i's are independent, then $C = 1$ and $f(x_1, x_2, \dots\dots x_n) = f_1(x_1) * \dots * f_n(x_n)$. Clemen and Reilly (1999) state that the n-dimensional joint distribution function *F* has two components (1) copula function, and (2) marginal distribution function. Let $X = (X_1, X_2, \cdot\cdot\cdot, X_n)$ be a random vector with distribution function *F*, and *Y* be uniformly distributed on (0, 1) and independent of *X*. We know that $U_i = F_i(X_i, Y)$ is uniformly distributed on (0, 1), therefore $X_i = F_i^{-1}(U_i)$. If we let the copula C be the distribution function of $U = (U_1, U_2, \cdot\cdot\cdot, U_n)$, then we have

$$F(X) = P(X \leq x)$$

$$= P(F_i^{-1}(Ui) \leq x_i, 1 \leq i \leq n)$$

$$= P(Ui \leq Fi(x_i), 1 \leq i \leq n)$$

$$= C(F_1(x_1), \cdot\cdot\cdot, F_n(x_n)).$$

This implies that *C* is the copula of *F*. Conveniently, a joint distribution function *F*(x,y) can be written in terms of the marginal distribution functions $F_X(x)$ and $F_Y(y)$ for the random variable X and Y using the relation $F(x,y) = C(F_X(x), F_Y(y))$. Hence, the copula function $C(u, v)$ can be written as

$$C(u,v) = F(FX^{-1}(u), FY^{-1}(v)),$$

and immediately it follows that

$$C(F_x(x), F_y(y)) = F(FX^{-1}(F_x(x)), FY^{-1}(F_y(y))) = F(x,y).$$

For calculating copula distance between the copula density c (x,y) and the independence copula density by using L_p distance, $CD_\alpha = \iint |c(x,y) - 1|^\alpha dxdy$, $\alpha > 0$. CD_2 is the Pearson's $\emptyset^2$ with its scaled version being $\emptyset$cor = $\sqrt{CD/(1 + CD_2)}$. Particularly, the copula correlation is a scale version of CD_1 as Ccor = $\frac{1}{2} CD_1 = \frac{1}{2} \iint |\, c(x,y) - 1| dxdy$.

Chapter 3

Simulation studies and real data application

In this section, we compare all the six dependence measures in terms of the statistical power using under various simulation settings, including Spearman's correlation, mutual information, maximal information coefficient, biweight midcorrelation, distance correlation, and copula correlation. A real genomic application is also provided. For a complete picture about how these measures work in different correlative patterns, we considered linear, cube root, quadratic, wavelet, circle, and cluster settings.

3.1 Simulated studies

We conducted simulation studies with the inclusion of the noise. The purpose of including the additive noise is to increase randomness and to test the robustness of the correlation measures. We considered both relatively low and high levels of additive noise. The R-packages for our implementation include pspearman, minerva, wgcna, energy, copula, and infotheo. For all settings, the sample size is fixed at 80.

3.1.1 Spearman's correlation

We used Fisher's method to transform Spearman's correlation coefficient to a z value

$$z = \frac{1}{2}\, ln\, (\frac{1+\ p}{1-\ p}),$$

where p is the Spearman's rank correlation coefficient. It can be proved that z asymptotically follows a normal distribution with mean 0.

The two R packages used for this analysis are infotheo and pspearman. Averages of the resulting p-values were summarized. Figure 3.1.1 and Figure 3.1.2 illustrate the result for all six patterns with different levels of noise. The results were based on 80 samples.

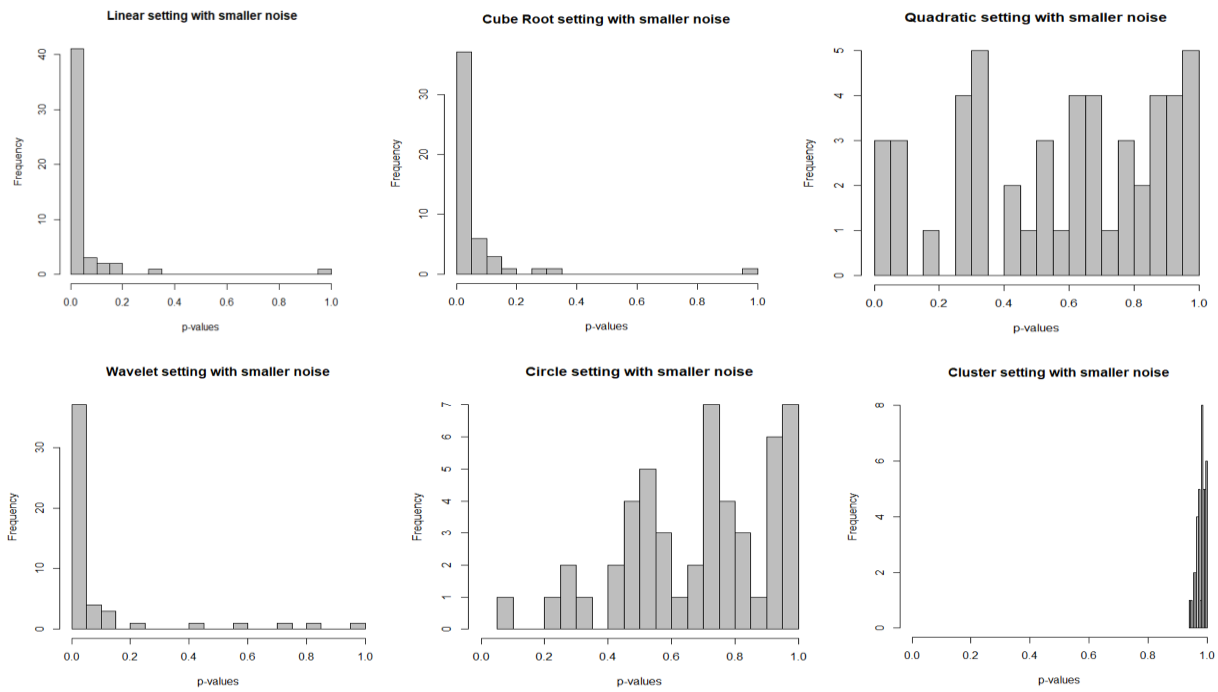


Figure 3.1.1 Spearman's rank correlation for linear, cube root, quadratic, wavelet, circle, and cluster settings with smaller noise.

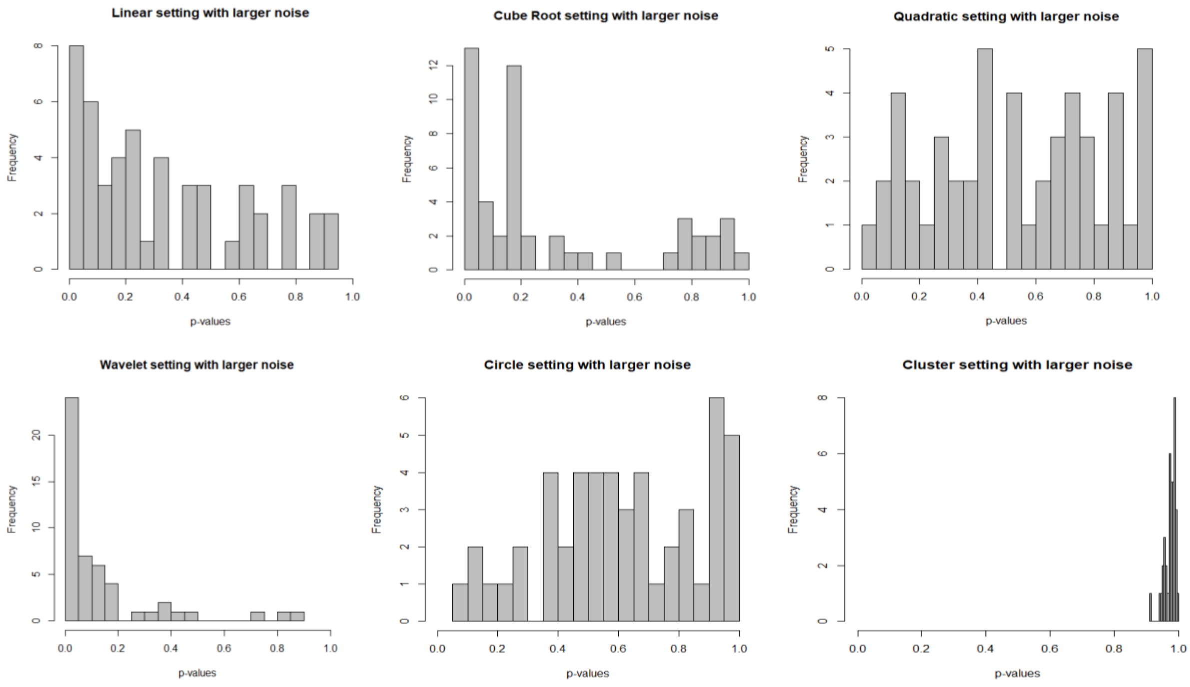


Figure 3.1.2 Spearman's rank correlation for linear, cube root, quadratic, wavelet, circle, and cluster settings with larger noise. `

Table 3.1.1 showed the empirical statistical power and the average *p*-value.

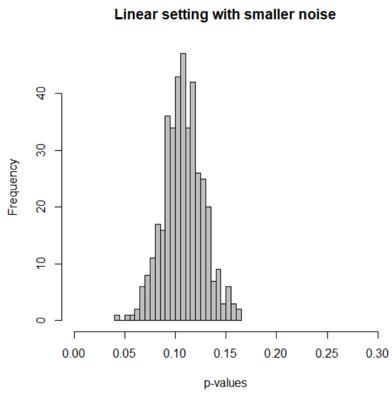
Table 3.1.1
Spearman correlation method

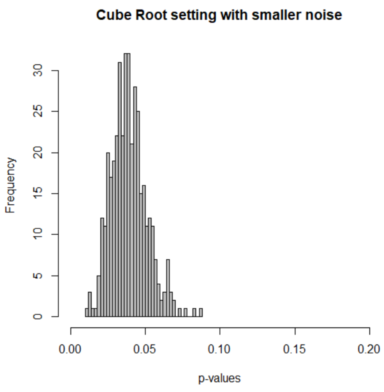
Relationship	Smaller Noise		Larger Noise	
	Empirical power	Mean *p*-value	Empirical power	Mean *p*-value
Linear	0.825	0.049	0.213	0.333
Cube Root	0.787	0.093	0.254	0.288
Quadratic	0.013	0.607	0.038	0.537
Wavelet	0.788	0.097	0.388	0.141
Circle	0.0	0.686	0.0	0.610
Cluster	0.0	0.973	0.0	0.974

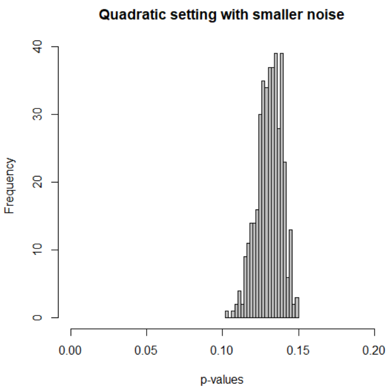
3.1.2 Mutual information

Mutual information (MI) is a measure of information quantity shared between two random variables. Figure 3.1.3 and Figure 3.1.4 show the result for linear, cube root, quadratic, wavelet, circle, and cluster setting with mutual information under different levels of noise. Similar to the Spearman's correlation, the MI can be converted to z value by Fisher's *z* transformation for independence test. The continuous data are discretized to compute entropy.

Figure 3.1.3 shows the distribution of the *p*-values with smaller noise. It is apparent that the variability of an estimate is significantly lower.







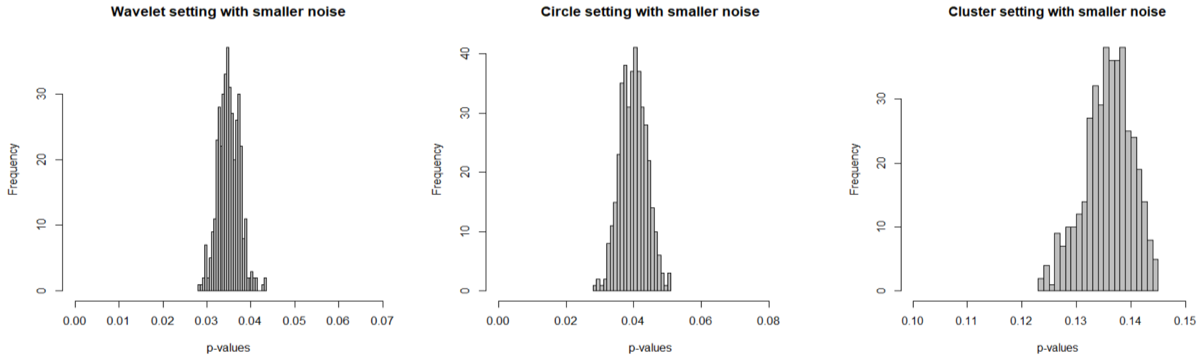


Figure 3.1.3 Mutual information for linear, cube root, quadratic, wavelet, circle, and cluster settings with smaller noise.

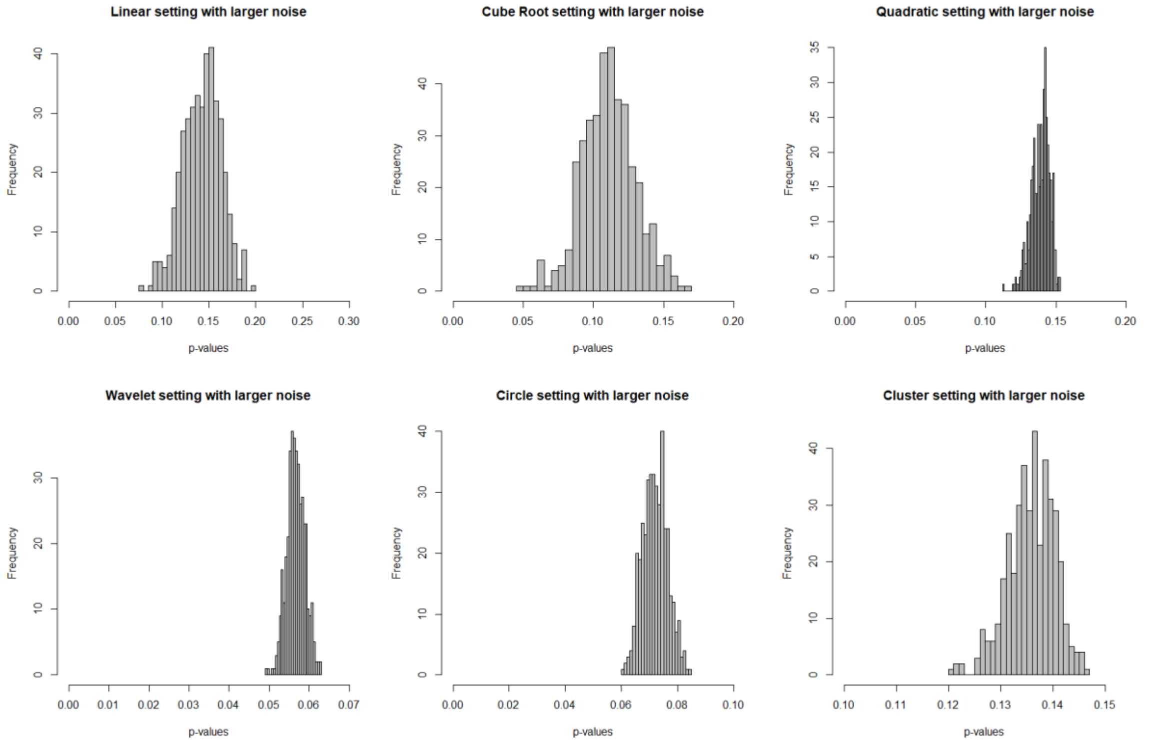


Figure 3.1.4 Mutual information for linear, cube root, quadratic, wavelet, circle, and cluster settings with larger noise.

Figure 3.1.4 shows the simulation results for the same model structure but with larger noise level. Table 3.1.2 shows the statistical result of empirical power and the average *p*-value with different number bins (nbins), where it can be seen that the mutual information works well for cube root, wavelet and circle settings with nbins=1/3 (see table 3.1.2 (a)). However, the mutual information fails to detect any linear dependence, which is the most common setting with nbins = ¼ and ½ shown in Table (b) and Table (c) respectively. Mutual information strongly depends on the choice of nbins. Therefore, MI method is very unstable for continuous data.

Table 3.1.2

(a) Mutual information (nbins=1/3)

Relationship	Smaller Noise		Larger Noise	
	Empirical power	Mean *p*-value	Empirical power	Mean *p*-value
Linear	0.0	0.108	0.0	0.142
Cube Root	0.790	0.039	0.003	0.111
Quadratic	0.0	0.134	0.0	0.138
Wavelet	1.0	0.035	0.005	0.057
Circle	0.988	0.049	0.0	0.072
Cluster	0.0	0.136	0.0	0.136

(b) Mutual information (nbins=1/4)

Relationship	Smaller Noise		Larger Noise	
	Empirical power	Mean *p*-value	Empirical power	Mean *p*-value
Linear	0.003	0.110	0.0	0.145
Cube Root	0.835	0.039	0.003	0.111
Quadratic	0.008	0.064	0.0	0.069
Wavelet	1.0	0.0	1.0	0.0
Circle	1.0	0.0	1.0	0.0
Cluster	1.0	0.011	1.0	0.011

(c) Mutual information (nbins=1/2)

Relationship	Smaller Noise		Larger Noise	
	Empirical power	Mean *p*-value	Empirical power	Mean *p*-value
Linear	0.0	0.243	0.0	0.255
Cube Root	0.0	0.197	0.0	0.239
Quadratic	0.0	0.353	0.0	0.359
Wavelet	0.0	0.224	0.0	0.277
Circle	0.0	0.283	0.0	0.340
Cluster	0.0	0.395	0.0	0.395

3.1.3 Maximal information coefficient

Our next simulation is for maximal information coefficient (MIC). MIC method measures the linear and nonlinear relationships between two continuous variables. The simulated random samples are produced for different patterns and settings. An R package (Minerva) is used to calculate the *p*-value in six different models with sample size n = 80. For sample size n, the bin (alpha) equals 0.6 where B(n) = n^{α} search-grid size. The “infotheo” package utilized several entropy estimators to implement various measures of information theory. The software package requires discretization of continuous data, and computes MIC across all grids. The results of MIC's correlation for six settings with smaller noise obtained from the first simulation are presented in Figure 3.1.5

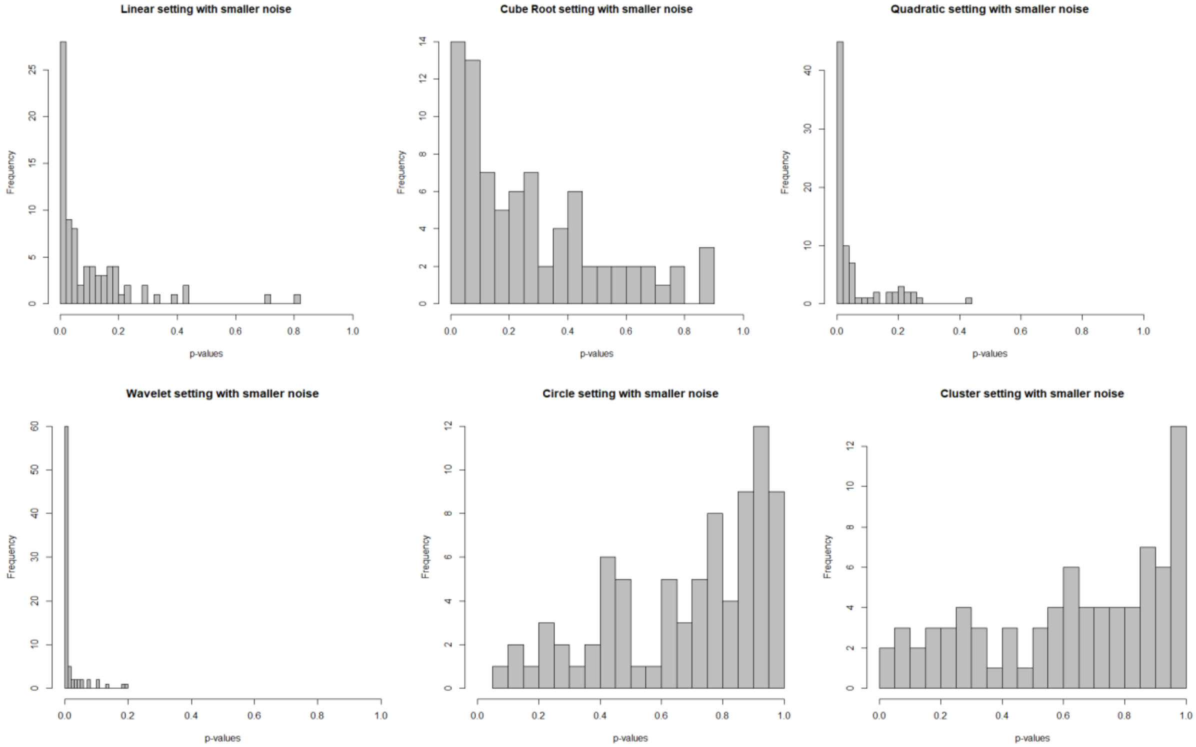


Figure 3.1.5 Maximal information coefficient’s method for linear, cube root, quadratic, wavelet, circle, and cluster settings with smaller noise.

The average *p*-value ranges from 0.012 to 0.690. Observably, most settings did not show a good strength of the dependence linear or nonlinear relationship within a noise-free environment. Figure 3.1.6 shows MIC's correlation for six settings with a larger noise. The mean of all *p*- values are high, and hence, MIC performs poorly to detect measure dependence for linear and nonlinear relationships.

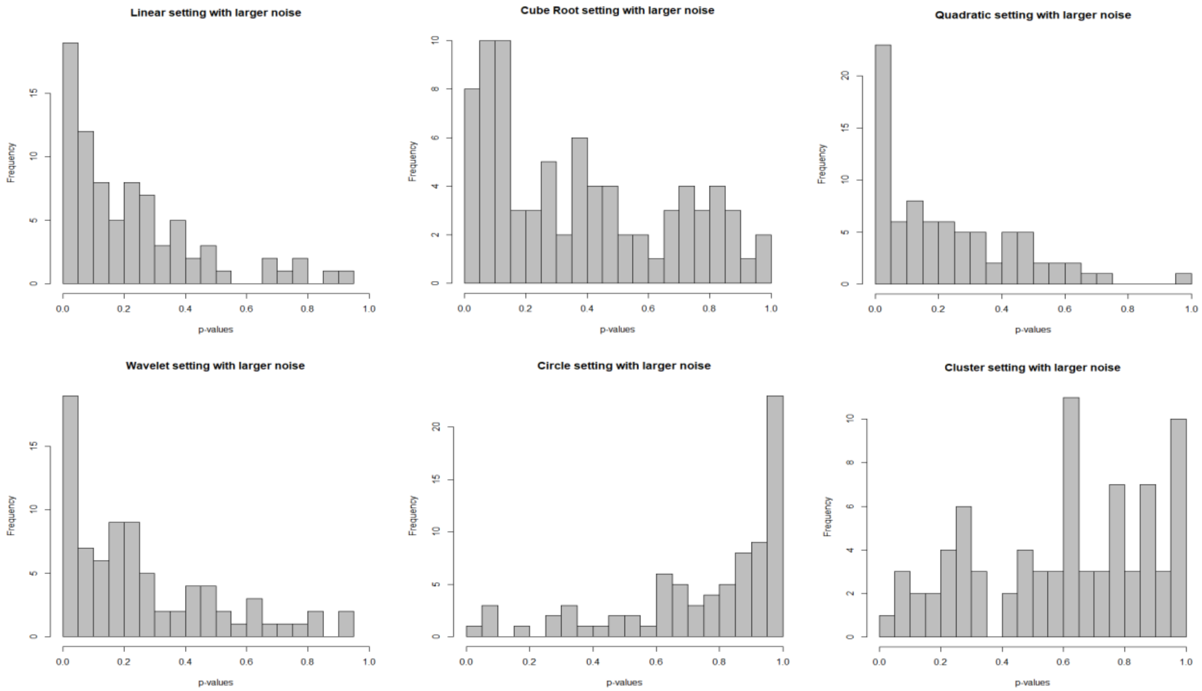


Figure 3.1.6 Maximal information coefficient’s method for linear, cube root, quadratic, wavelet, circle, and cluster settings with larger noise.

Table 3.1.3 shows the empirical statistical power and the average *p*-value.

Table 3.1.3
Maximal information coefficient

Relationship	Smaller Noise		Larger Noise	
	Empirical power	Mean *p*-value	Empirical power	Mean *p*-value
Linear	0.548	0.065	0.225	0.209
Cube Root	0.225	0.279	0.087	0.406
Quadratic	0.713	0.065	0.20	0.254
Wavelet	0.875	0.012	0.188	0.322
Circle	0.0	0.690	0.013	0.707
Cluster	0.038	0.479	0.025	0.596

3.1.4 Biweight midcorrelation

Biweight midcorrelation (bicor) is median based, which reduces sensitivity to outliers. Consequently, the results of simulations prove that the bicor performs better in identifying the uncertainty in the dependent variable when the independent variable is observed. The graphs presented using the bicor method demonstrates a measure of the similarity levels. However, the statistical method depends on the R-language, which interprets multiple data and variables. The package components used for bicor are (BiocManager) and the library (WGCNA).

The R package WGCNA includes functions corAndPvalue and bicorAndPvalue that calculate correlations of matrices and their associated Student *p* -values efficiently and accurately (Langfelder and Horvath 2008).

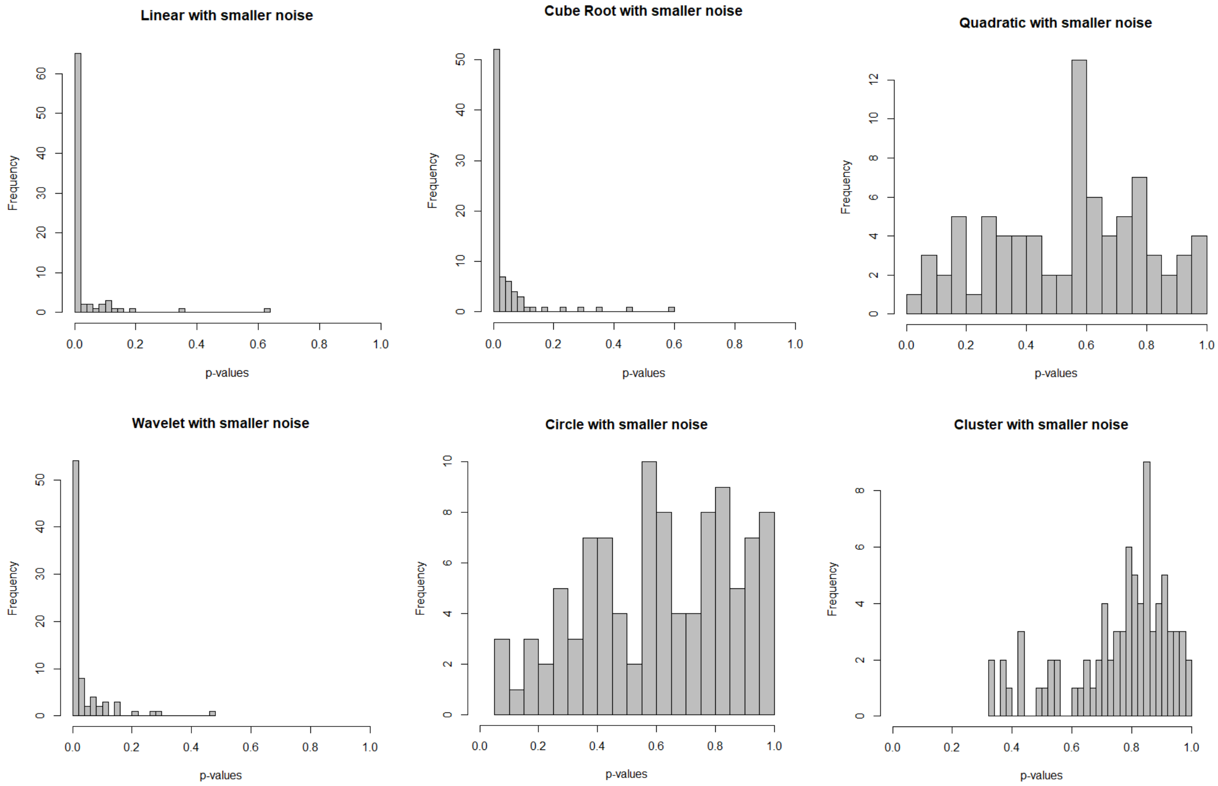


Figure 3.1.7 Biweight midcorrelation's method for linear, cube root, quadratic, wavelet, circle, and cluster settings with smaller noise.

The two main parameters considered to generate the average method in the formula are median pseudo ranks and weight pseudo ranks. Some distributions in Figure 3.1.7 indicate the obtainability of strong association for linear and nonlinear relationship among smaller noise models. For example, linear, cube root, and wavelet have mean p- values less than statistical significance level, while quadratic, circle, and cluster do not detect measure of dependence since the mean p-value is greater than 5%. Whereas the distributions of large noise models do not perform good quality in this case as shown in Figure 3.1.8. Circle and cluster represent the highest p-value than the other models.

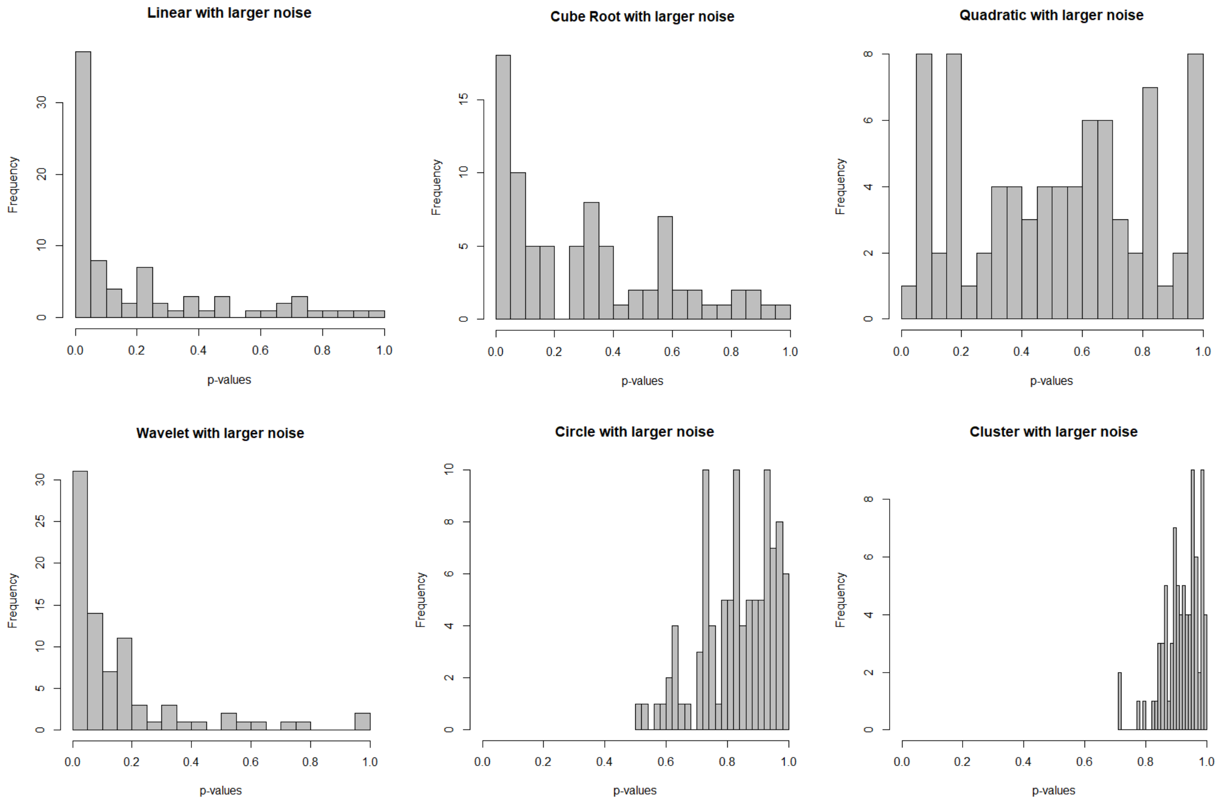


Figure 3.1.8 Biweight midcorrelation’s method for linear, cube root, quadratic, wavelet, circle, and cluster settings with larger noise.

The summary of *p*-values is presented in Table 3.1.4, where it can be seen that biweight midcorrelation is able to provide quality results for linear cube root and wavelet.

Table 3.1.4
Biweight midcorrelation

Relationship	Smaller Noise		Larger Noise	
	Empirical power	Mean *p*-value	Empirical power	Mean *p*-value
Linear	0.813	0.029	0.200	0.207
Cube Root	0.715	0.045	0.150	0.306
Quadratic	0.006	0.543	0.059	0.514
Wavelet	0.744	0.039	0.188	0.157
Circle	0.0	0.607	0.0	0.832
Cluster	0.0	0.758	0.0	0.917

3.1.5 Distance correlation

The following simulation considered is the distance correlation (dcor) measure. It is equivalent to product-moment covariance and correlation. The test for dependence relationships is considered for different settings with distance correlation and the two different levels of noise. The samples were randomly generated from the normal distribution with sample size, $n = 80$. The result was tested for the significance level of 5%. The dcor R-package helps in analyzing the multivariate data. The correlation process applies to both the larger noise and the smaller noise data sets, depending on the distance. For the dcor package library (energy) was used to derive the codes in the R-Program with a distance correlation test. 5000 permutations were considered to get a more accurate result.

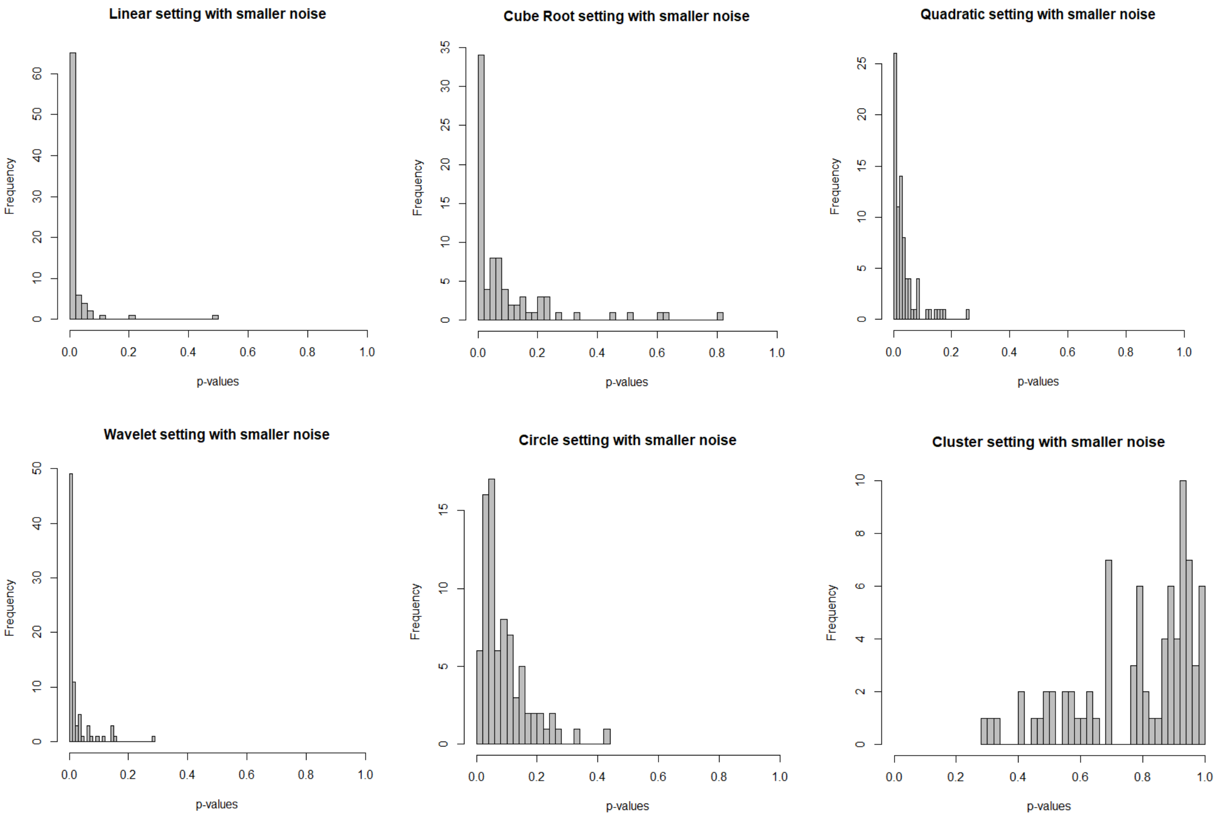


Figure 3.1.9 Distance's correlation for linear, cube root, quadratic, wavelet, circle, and cluster settings with smaller noise.

Figure 3.1.9 summarizes the simulation results for linear, cube root, quadratic, wavelet, circle, and cluster setting with distance correlation smaller noise. It appears that the wavelet model would get perfect strength of that dependence within the variety of noise, likewise the linear function. Figure 3.1.10 illustrates the results of six functions for larger noise. The result demonstrates that none of the settings identify the dependence test.

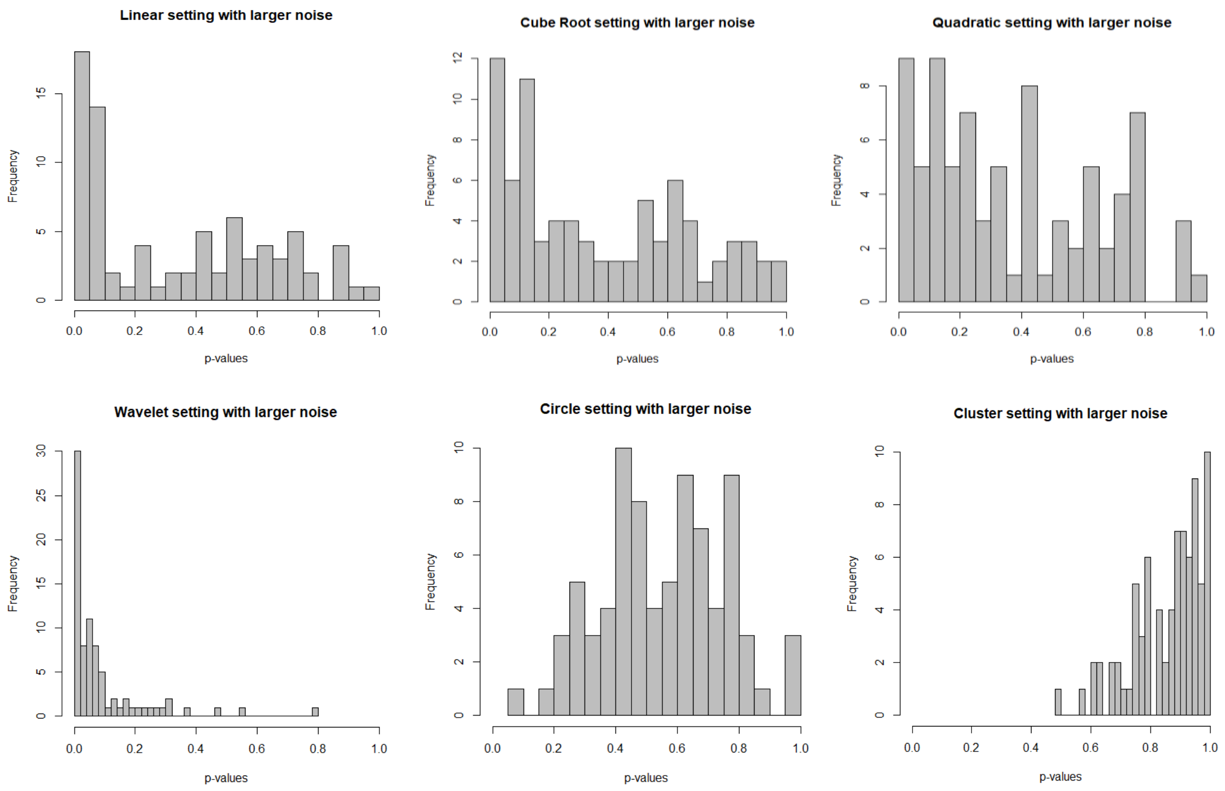


Figure 3.1.10 Distance's correlation for linear, cube root, quadratic, wavelet, circle, and cluster settings with larger noise.

Table 3.1.5. illustrates the empirical statistical power and the average *p*-value, where it can be seen that the distance measure is sensitive to linear cube root, quadratic and wavelet dependence.

Table 3.1.5
Distance correlation

Relationship	Smaller Noise		Larger Noise	
	Empirical power	Mean *p*-value	Empirical power	Mean *p*-value
Linear	0.863	0.019	0.213	0.344
Cube Root	0.740	0.099	0.150	0.443
Quadratic	0.751	0.037	0.101	0.326
Wavelet	0.963	0.011	0.550	0.073
Circle	0.541	0.073	0.0	0.442
Cluster	0.0	0.414	0.0	0.449

3.1.6 Copula correlation

The copula cluster is an R-package for the implementation of the clustered algorithm. The copula function found data sets for the complex multivariate dependence to produce the process. The normal distributed data with sample size *n* = 80 and a significance level of 5% was considered for simulation. The number of permutations considered during the simulation is 1000.

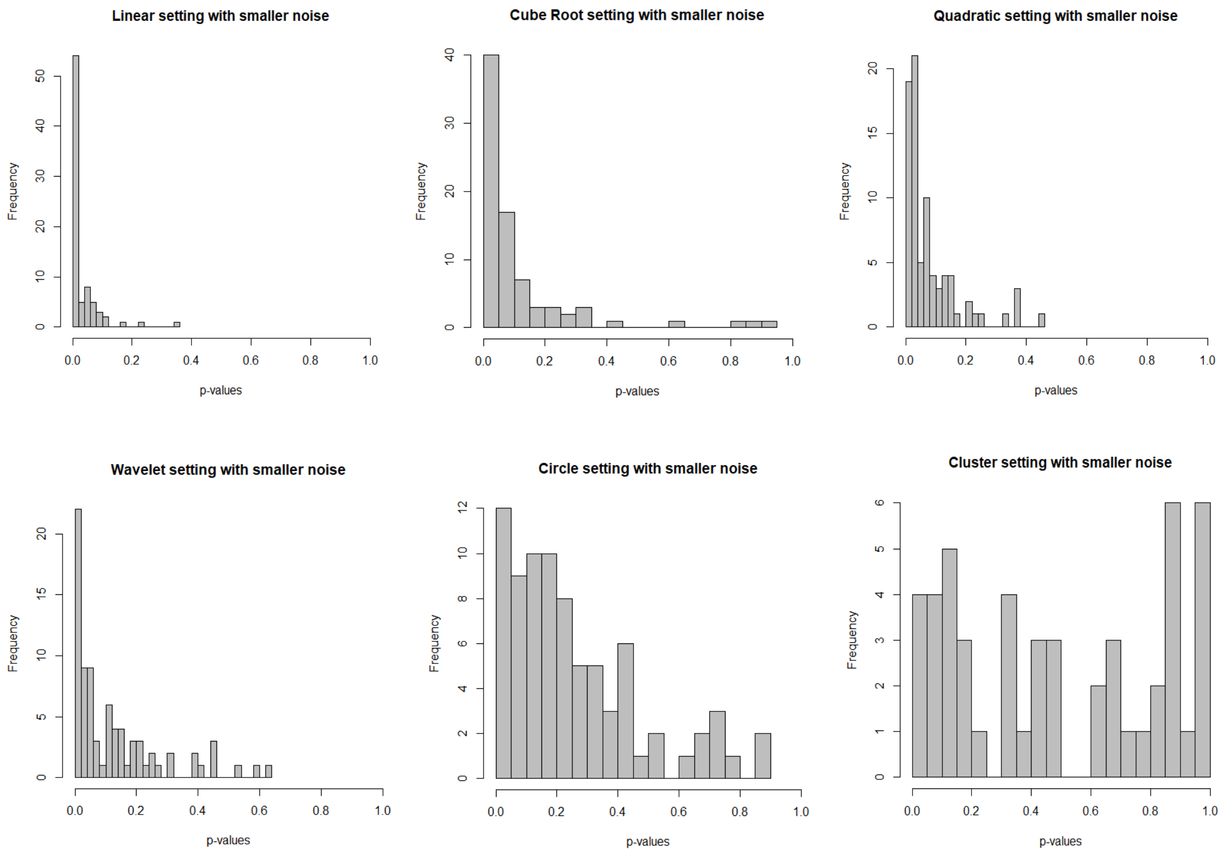


Figure 3.1.11 Copula correlation for linear, cube root, quadratic, wavelet, circle, and cluster settings with smaller noise.

Figure 3.1.11 illustrates the results for linear, polynomial, quadratic, wavelet, circle, and cluster setting with Ccor smaller noise generated from a normal distribution. Monotonic relationships are common when interpretation depends on the copula correlation method. The quadratic, cluster, and circle setting graphs indicate a variety of parameters that affect the data analysis to attain the spectrum range. Copula correlation gives a high *p*-value of the setting since they are nonlinear relationships. The copula correlation displays the function of both smaller and significant noise data types. Ccor analysis depends on the linear bivariate relationship.

Figure 3.1.12 summarizes the *p*-values for larger noise.

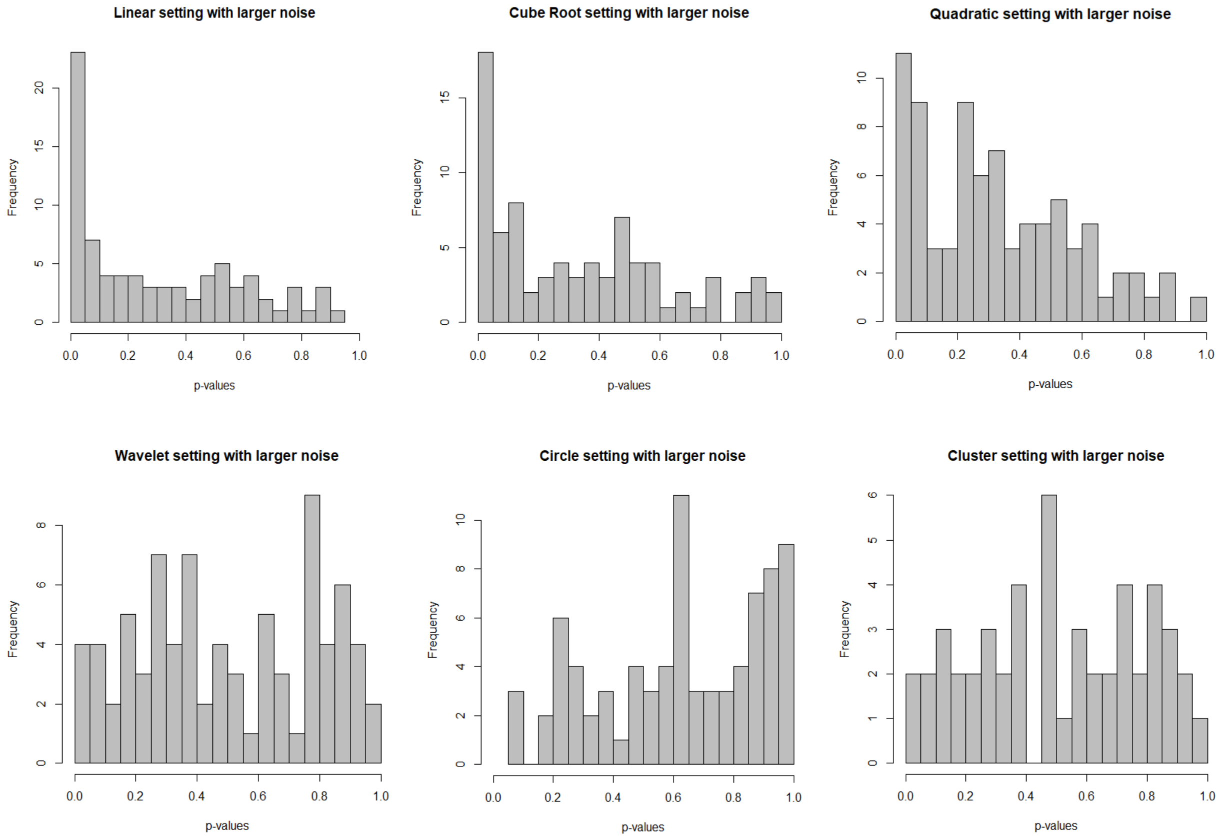


Figure 3.1.12 Copula correlation for linear, cube root, quadratic, wavelet, circle, and cluster settings with larger noise.

Table 3.1.6 shows the empirical statistical power and the average *p*-value, where it can be seen that the copula correlation has satisfactory performance only for linear setting.

Table 3.1.6
copula correlation

Relationship	Smaller Noise		Larger Noise	
	Empirical power	Mean *p*-value	Empirical power	Mean *p*-value
Linear	0.789	0.029	0.213	0.306
Cube Root	0.462	0.082	0.150	0.334
Quadratic	0.338	0.128	0.075	0.500
Wavelet	0.138	0.269	0.050	0.532
Circle	0.175	0.257	0.0	0.625
Cluster	0.101	0.497	0.060	0.505

Table 3.1.7 summarizes the overall performance of each measure.

Table 3.1.7
Simulation performance of different settings

Measures	Simulation settings with overall satisfactory performance
Spearman's rank correlation	Linear, Cube Root, Wavelet
Mutual information	Cube Root, Wavelet, Circle
Maximal information coefficient	Quadratic, Wavelet
Biweight midcorrelation	Linear, Cube Root, Wavelet
Distance correlation	Linear, Cube Root, Quadratic, Wavelet, Circle
Copula correlation	Linear

The above results show that Spearman's correlation, biweight midcorrelation and distance correlation have overall satisfactory performance for linear and nonlinear relationships.

3.2 A genomic application

In this part, we applied some selected measures to a dataset from the Cancer Genome Atlas (TCGA), pre-processed by Zhang et al. (2014). The dataset contained the expression level of 245 cancer-related genes from 150 samples. The analysis focuses on the detection of co-expressed genes using three measures that have overall good performance from simulation studies, including Spearman’s rank, distance correlation and biweight midcorrelation.

Gene co-expression analysis has been widely applied for molecular biology research, especially for the systems-level or pathway-level studies. In general, the functions in isolation of genes and their protein products do not perform. The functions perform jointly and in cooperation. Tremendous research efforts have been made to clarify the molecular basis of the initiation and progression of ovarian cancer. However, most of those studies have concentrated on a single gene or a specific type of data, which in return may not identify the complex mechanisms of cancer formation by neglecting to detect the interactions of different genetic and epigenetic factors (Zhang et al., 2014). In practice, temporal changes in gene expression require more complex detection methods than simple correlation measures that may result in complex association patterns. For example, the effect of regulation may lead to time-lagged associations and interactions local to a subset of samples.

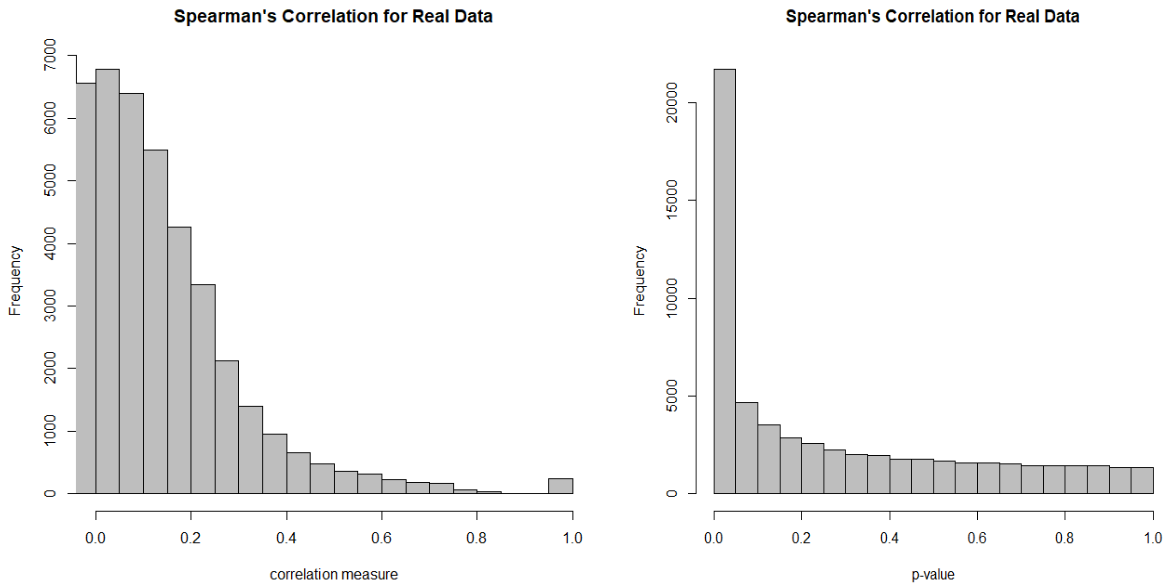


Figure 3.2.1 Histogram of TCGA ovarian cancer data using Spearman's method including the correlation measure (left panel), and *p*-value (right panel), from 5000 replications

Figure 3.2.1 summarized the Spearman's rank correlation for more than 20,000 pairs of genes that are significantly associated ($p < 0.05$).

For dcor, the energy package was used with index =1, which is the exponent on Euclidean distance. Euclidean distance $\|x_i - x_j\|^d$, where $0 < d < 2$ to compute distance correlation and *p*-value. Figure 3.2.2 shows the distribution of correlation and *p*-value by using dcor method. We found more than 25,000 pairs of genes having *p*-value less than 5%.

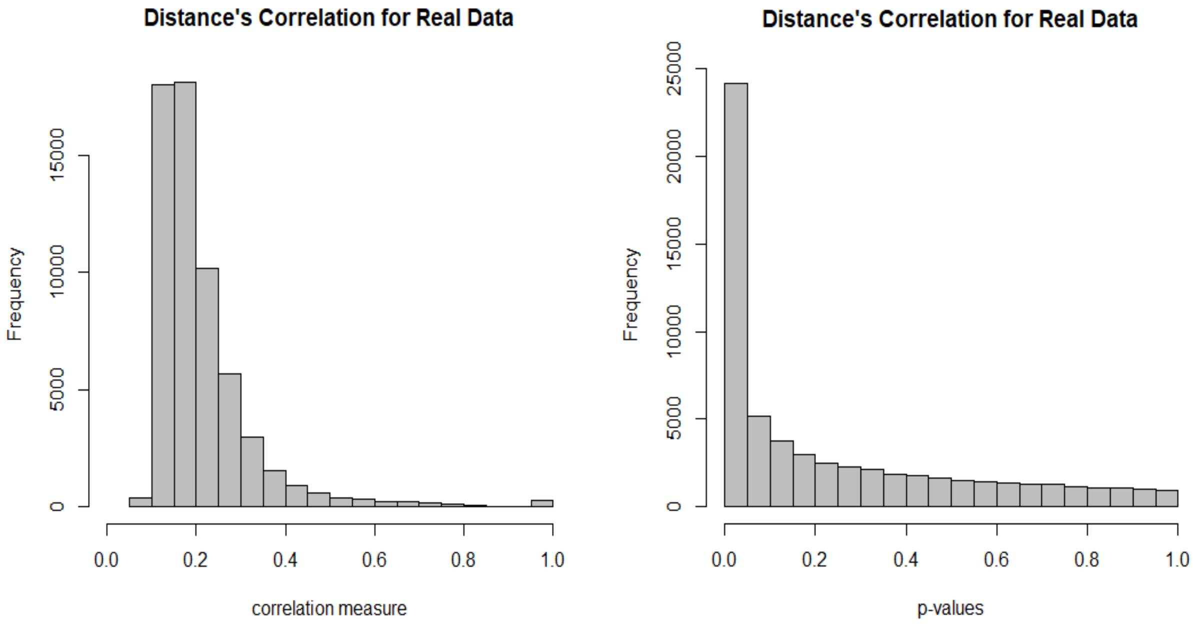


Figure 3.2.2 Histogram of TCGA ovarian cancer data using distance method including the correlation measure (left side), and *p*-value (right side), with 5000 number of replications

Finally, the co-expression of all gene pairs were measured by biweight midcorrelation measure. It concentrates on the media-based analysis, which diminished sensitivity towards the outliers. To compute the biweight midcorrelation (bicor) between pairs of genes, the WGCNA package was used to compute correlation measure and *p*-value. In the histogram which are demonstrated in Figure 3.2.3, it is noticeable that there is a significant correlation for more than 20,000 pairs after replication while the correlation measure shows a strong correlation considering that the majority of the gene pairs are dependent.

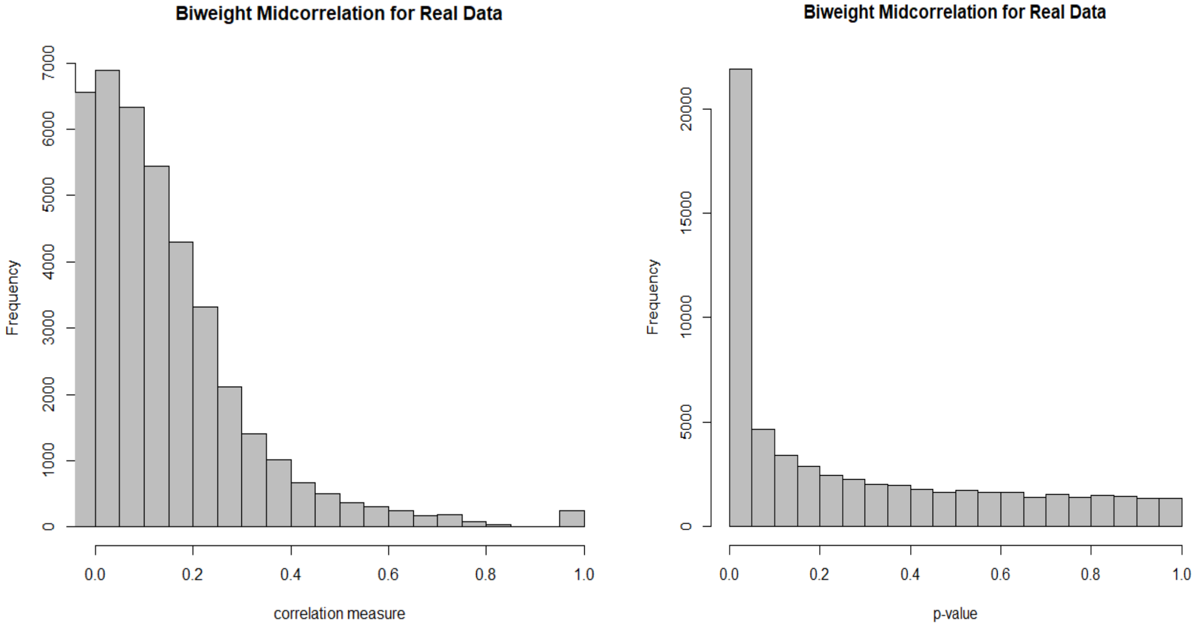
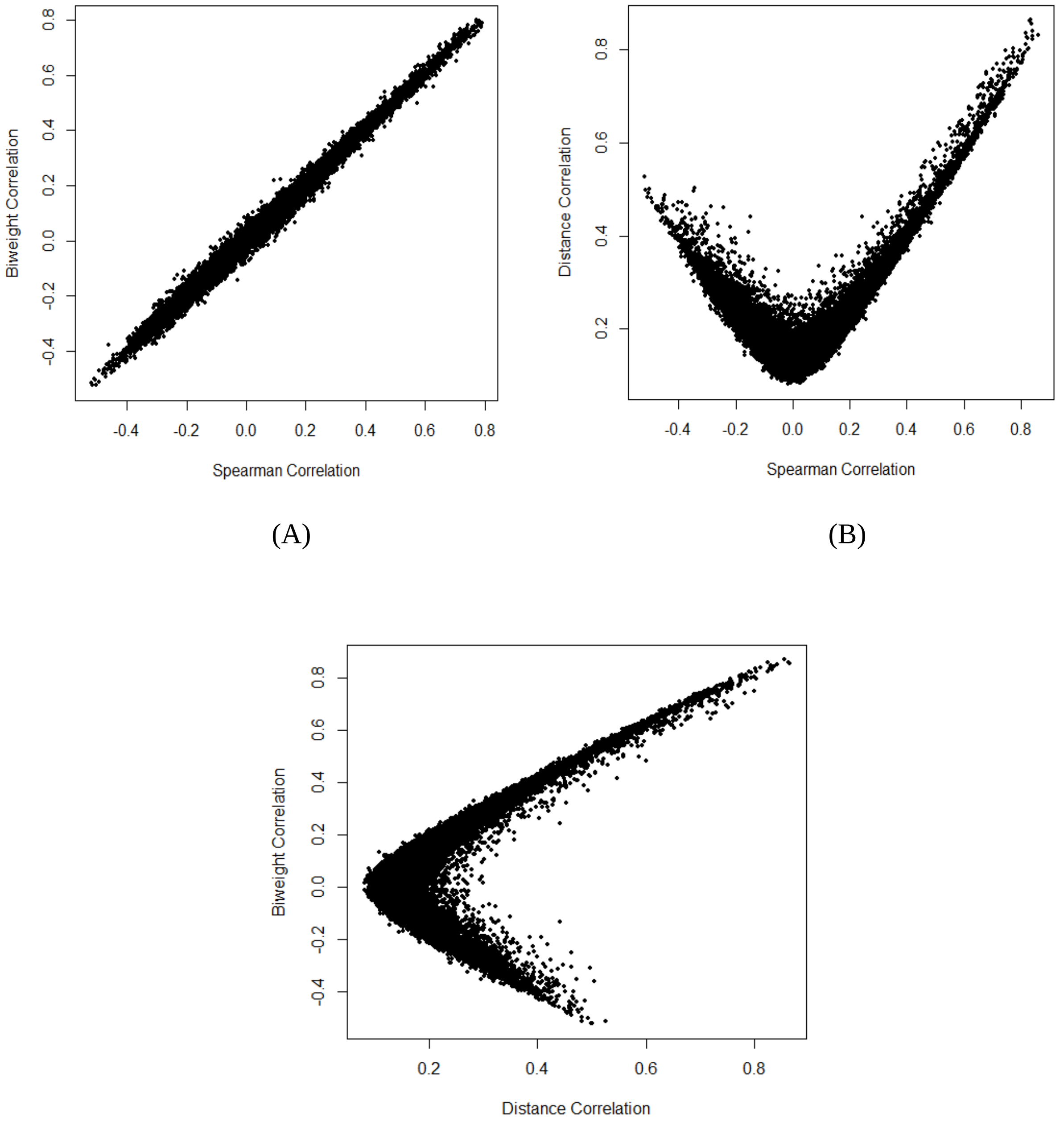


Figure 3.2.3 Histogram of TCGA ovarian cancer data using biweight midcorrelation including the correlation measure (left panel), and *p*-value (right panel).

We then investigated the consistency between the three measures. The figure 3.2.4 below show the agreement between each pair of measures: (A) Spearman's correlation vs biweight midcorrelation; (B) Spearman's correlation vs distance correlation; (C) Biweight midcorrelation vs distance correlation.



(A)

(B)

(C)

Figure 3.2.4 Comparison of correlations: (A) Spearman's correlation coefficient vs. biweight midcorrelation, (B) Spearman's correlation coefficient vs. distance correlation, and (C) distance correlation vs. biweight midcorrelation.

As can be seen from Figure 3.2.4, for the majority of co-expressed gene pairs, especially those with strong co-expression, all three measures are similar. Table 3.2.1 presents in 6 pairs of strongly correlated genes as examples. Our findings confirm some recent reports that the majority of co-expressed genes are linear or monotonic nonlinear.

Table 3.2.1
Examples of co-expressed gene pairs by Spearman's rank correlation, biweight midcorrelation and distance correlation

Gene pairs (*i*, *j*)	Spearman correlation	Biweight midcorrelation	Distance correlation
(42,188)	0.8083	0.8224	0.7891
(47,199)	0.7917	0.8119	0.7766
(88,244)	0.8190	0.8356	0.8282
(89,244)	0.7735	0.7987	0.7932
(190,235)	0.8397	0.8585	0.8233
(196,214)	0.8031	0.8002	0.7753

Chapter 4

Conclusions

In many scientific domains, it is essential to identify and measure different types of associative relations between variables from experimental or observational data. The relationship between two variables is often characterized by some type of correlation coefficient, which can be utilized for further decision-making and predictions. Pearson's correlation coefficient is popular as a measure of strength of the relationship between two variables. The procedure, however, is limited to linear associations and is excessively sensitive to outliers. To measure nonlinear-type relations, a number of correlation measures have been recently developed, including distance correlation, MIC, mutual information, etc. In this work, we conduct an extensive simulation study to systematically compare these measures in various settings. Based on our simulation result, Spearman's correlation, biweight midcorrelation and distance correlation have better statistical performance overall. They can be robust alternatives to other statistical measures, especially when the underlying relation is nonlinear. The mutual information does not work well in linear settings, and the performance depends on discretization for continuous data.

4.1 Discussion

We would like to point out that all the dependence measures considered in this thesis have certain drawbacks. For instance, it is known that MIC depends on a user-defined parameter, namely B(n). Also, the computational cost of MIC increases exponentially as the number of data points gets larger; therefore, it is not suitable for large-scale datasets. Additionally, as pointed out by Simon and Tibshirani (2014), MIC may not work well in the presence of substantial noise.

Kinney and Atwal (2014) also noted that MIC is not equitable, and the MIC values might not be affected by variable noise for specific relationships.

Although mutual information is a popular measure of nonlinear or combinatorial dependence between two variables, it has been pointed out that the estimate of MI measure could be challenging for small datasets due to the discretization and number of bins. In addition, MI does not satisfy the criterion of equitability (equitability is a criterion that the statistic should give similar scores to equally noisy relationships of different types). Thus, it is not a reliable method for continuous data.

MIC and distance correlation are two promising measures for nonlinear relations. Simon and Tibshirani (2014) state that in many cases, distance correlation exhibits more statistical power than the MIC. It can also be seen in our analysis that even with a small sample size, the distance correlation has satisfactory performance at a different level of noise. Copula correlation could potentially capture the complete dependence structure inherent in variables (Xi et al., 2014). However, the copula-based methods are analytically complex and difficult to interpret, and fitting the parameters of a copula is a challenging statistical problem.

The distance correlation and biweight midcorrelation have overall satisfactory performance for most of the correlative patterns, with affordable computational cost and good robustness to outliers. However, there is still a need to find or develop a measure that is interpretable and sensitive to both linear and nonlinear, monotonic and non-monotonic relations.

4.2 Future work

There are several directions that we would like to explore in the future. First, we will incorporate some additional measures recently developed to measure nonlinear relations, to name a few, the projection correlation (Zhu et al., 2017), and multiscale graph correlation (MGC, Shen, Priebe & Vogelstein, 2019).

Second, we may extend the evaluation of correlation measures from univariate variables to multivariate variables or random vectors of arbitrary dimensions. Compared to model-based exploration such as multiple linear regression and principal component analysis, the correlation method is model-free and does not rely on any assumption on the model structures. Also, categorical variables are commonly seen in many scientific studies. Further analysis can be conducted by comparing the correlation measures for the association between categorical variables or even the association between a categorical variable (either ordinal or nominal) and a continuous variable.

Third, we may test all correlation measures on other, real datasets. For instance, it will be interesting to apply distance correlation to some genomic datasets to identify nonlinearly correlated biomarkers or biological pathways. Such analyses may shed new light to the complex relations between many different types of biological factors.

References

Bolboaca, S.-D., and Jantschi, L., 2006. Pearson ¨ versus Spearman, Kendall's tau correlation analysis on structure-activity relationships of biologic active compounds. Leonardo Journal of Sciences 5(9):179–200.

Chang, Y., Li, Y., Ding, A., & Dy, J. G. (2016). A robust-equitable copula dependence measure for feature selection. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016*.

Clemen, R., and Reilly, T. (1999). "correlations and copulas for Decision and Risk Analysis," Management Science, Vol. 45(2)

Da Costa, J. P., 2015. Rankings and Preferences: New Results in Weighted Correlation and Weighted Principal Component Analysis with Applications. Springer.

Deebani, W., & Kachouie, N. N. (2018). Ensemble correlation coefficient. *International Symposium on Artificial Intelligence and Mathematics, ISAIM 2018*. https://doi.org/10.1007/978-3-319-55895-0_17

Ding, A., and Li, Y. (2015). copula correlation: An Equitable Dependence Measure and Extension of Pearson's correlation. arXiv:1312.7214

Fisher, L. D., & van Belle, G. (1993). Biostatistics: A Methodology for the Health Sciences. *John Wiley and Sons Ltd*, New York, United States 1993.

Hastie, T.; Tibshirani, R.; and Friedman, J. 2002. The elements of statistical learning: Data mining, inference, and prediction. Biometrics.

Kinney, J. B., & Atwal, G. S. (2014). Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1309933111

Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*. https://doi.org/10.1186/1471-2105-9-559

Li, D. X. (2000). On default correlation: A copula function approach. *The Journal of Fixed Income*, *9*(4), 43-54.

Martinez-Gomez, E., Richards, M., T. & Richards, D., T. (2014). distance correlation methods for discovering associations in large astrophysical databases. *The Astrophysical Journal*, 781 (1)

Nguyen, H. V.; Muller, E.; Vreeken, J.; Efros, P.; and B ¨ ohm, ¨ K. 2014. Multivariate maximal correlation analysis. In Proceedings of the 31st International Conference on Machine Learning (ICML-14), 775–783.

Reshef, D. N.; Reshef, Y. A.; Finucane, H. K.; Grossman, S. R.; McVean, G.; Turnbaugh, P. J.; Lander, E. S.; Mitzenmacher, M.; and Sabeti, P. C. 2011. Detecting novel associations in large data sets. Science 334(6062):1518–1524.

Rüschendorf, L. (2009). On the distributional transform, Sklar's theorem, and the empirical copula process.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(1):379–423,623–656.

Shen, C., Priebe, C., & Vogelstein, J., (2019) From Distance Correlation to Multiscale Graph Correlation, Journal of the American Statistical Association, 115:529, 280-291, DOI: 10.1080/01621459.2018.1543125

Simon, N., & Tibshirani, R. (2014). Comment on “Detecting Novel Associations In Large Data Sets” by Reshef Et Al, Science Dec 16, 2011. *Science*. http://arxiv.org/abs/1401.7645

Sklar, A. 1959. Fonctions de Re´partition a` n Dimensions et Leurs Marges. Publications de l’Institut Statistique de l’Universite´ de. Paris, 8 229–231.

Spearman, C. (2010). The proof and measurement of association between two things. *International journal of epidemiology*, *39*(5), 1137-1150.

Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The annals of statistics*, *35*(6), 2769-2794.

Tang, D., Wang, M., Zheng, W.,& Wang, H.(2014). RapidMic: Rapid Computation of the maximal Information Coefficient. Evol Bioinform Online. 10: 11–16. DOI: 10.4137/EBO.S13121

Wang, Y X., Liu, K., Elizabeth, T., Rotter, J., Medina, M., Waterman, M., Huang, H., (2018). Generalized correlation measure using count statistics for gene expression data with ordered samples. *Bioinformatics*, 34(4), 617–624. https://doi.org/10.1093/bioinformatics/btx641

Wang, Y., Li, Y., Cao, H. et al. (2015). Efficient test for nonlinear dependence of two continuous variables. BMC Bioinformatics 16: 260.

Xi, Z., Jing, R., Wang, P., & Hu, C. (2014). A copula-based sampling method for data-driven prognostics. *Reliability Engineering and System Safety*. https://doi.org/10.1016/j.ress.2014.06.014

Yuan, L., Sha, W., Sun, ZL., &Zheng, CH., (2013). biweight midcorrelation-Based Gene Differential Coexpression Analysis and Its Application to Type II Diabetes. ICIC 2013. Communications in Computer and Information Science, vol 375. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-39678-6_14

Zeng, C., Yuan, L., Sha, W. & Sun, Z. (2013). Gen differential coexpression analysis based on biweight correlation and maximum clique.

Zhang, Q., Burdette, J., Wang, J., (2014). Integrative network analysis of TCGA data for ovarian cancer. *BMC Systems Biology*. 8:1338. DOI 10.1186/s12918-014-0136-9

Zhang, Z. Qi, Y. & Ma, X.(2011). Asymptotic independence of correlation coefficients with application to the testing hypothesis of independence. 5: 342–372. doi: 10.1214/11-EJS610

Zhu, L., Xu, K., Li, R., & Zhong, W. (2017). Projection correlation between two random vectors. *Biometrika*, *104*(4), 829–843. https://doi.org/10.1093/biomet/asx043