8-2023

A Comparative Study of Techniques for Non-monotonic Dependence with Emphasis on Sensitivity to Sample Size, Noise Level and Computational Attributes

Fariha Tasnim
University of Arkansas, Fayetteville

A Comparative Study of Techniques for Non-monotonic Dependence with Emphasis on Sensitivity to Sample Size, Noise Level and Computational Attributes

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

by

Fariha Tasnim
University of Dhaka
Bachelor of Science in Statistics, 2012
University of Dhaka
Master of Science in Statistics, 2014

August 2023
University of Arkansas

This thesis is approved for recommendation to the Graduate Council.

Avishek Chakraborty, Ph.D.
Thesis Chair

Mark E. Arnold, Ph.D.
Committee Member

Qingyang Zhang, Ph.D.
Committee Member

Abstract

Evaluating association between variables is often of interest by many researchers. To serve this purpose, different association measures have been developed. However, type of relation between variables affects the degree of relationship. Hence, detection of the relationship between variables is germane to measuring the correlation coefficient. With that mindset, here we explored six non-monotonic measure of association techniques and compared them with three classical approaches. Due to inconsistency in definition and range of different techniques, it is not feasible to compare the correlation estimates as their nature of variability differ. Therefore, we used permutation test based on Monte Carlo approximation for testing independence. We paired the correlation estimates along with the p-value for deciding strength of association between variables. At first, we explored and compared the association measures on twelve distinct simulation models under diverse scenario segregated by three sample sizes and three noise levels. In addition, we assessed the computational time and peak memory (RAM) usage during computations for each of these methods. Next, we applied all methods on Cape Floristic Region (CFR) data to capture association between four pairs of variable combinations with varying sample sizes. Overall, we found that, increasing sample size improves the performance of correlation measure. For any type of relationship, non-monotone measures were consistent in capturing association for large sample size. With respect to time and peak memory usage, we found two of the methods were not efficient.

Acknowledgements

I would like to acknowledge and give my warmest thanks to my supervisor Dr. Avishek Chakraborty. His guidance and advice carried me through all the stages of writing this thesis. I would also like to express my gratitude to my committee members Dr. Mark E. Arnold and Dr. Qingyang Zhang for their support and suggestions.

I appreciate my professors, office staffs and peers for the wonderful time I have had at the Department of Mathematical Sciences, University of Arkansas. Finally, I would like to give special thanks to my family as a whole for their continuous support and understanding.

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Importance of Measures of Association Studies

An association between two or more numeric variables is a measurement that gauges the relationship between the variables (Haug, 2023). If the variables are associated, knowledge about one variable provides information about the other variable (Samuel and Okey, 2015). For instance, in clinical trials, measure of association is used to understand the impact of newly manufactured drugs on patients. If the recommended dose of the testing drug improves/deteriorates the health of the patient, then there exists positive/negative association. On the other hand, if the medicine does not improve/deteriorate the existing health condition then there is no association of the drug on health improvement.

Measuring association between variables is considered as the 'sine qua non' of scientific research which has become a veritable instrument to experts in various fields (Merlo and Lynch, 2010). It has wide application from market forecasting in economics to general social behavior in sociology (Samuel and Okey, 2015). It is most widely used in medical science, namely, in the areas of epidemiology and psychology to quantify relationships between exposures and diseases or behaviors (Haug, 2023).

1.2 Types of Association

1.2.1 Linear and Non-linear Association

Depending on pattern, association can be of different types. When the relation between two variables tends to be approximately straight line then that is known as linear association which is specifically referred to as correlation. Correlation can be positive or negative based on sign. For example, during summer ice cream consumption tends to be higher since people consume more ice cream when it's hot out. So, there exists a positive correlation between the temperature and ice cream consumption. Besides, increase in elevation causes decrease in pressure. As the pressure decreases,the temperature decreases. Hence, there is a negative

correlation between altitude and temperature. In contrast, if two variables exhibit other type of relationship pattern, e.g., sinusoids, exponential etc. then that is known as non-linear association. Consider the association between side of a square and it's area. The area of a square can be obtained by taking the square if it's side. It follows that, side of a square and its area has non-linear quadratic relation.

1.2.2 Monotone and Non-monotone Association

Based on the direction of two associated variables, their relation can be classified as monotone and non-monotone association. If increasing the value of one variable either increase or decrease the value of other variable i.e., change in one variable due to increase in another variable always occurs in one direction then the relation between those two variables is monotonic association. In lieu, if the value of one variable increases, then the value of other variable may sometimes increase but can sometimes decrease as well, then they are said to share a non-monotonic association. Linear association or correlation are monotonic association. However, non-linear association can be monotone or non-monotone. Such as, when two variables share exponential association that is non-linear monotonic association. Again, if the variables share quadratic/sinusoid association, then their association is non-linear and non-monotone.

1.3 Classical Measures of Association

There are various devices to quantify association between two numeric variables in a sample. The common measures developed, calculate some kind of correlation coefficient. To name a few, Pearson's product moment correlation, Spearman's rank correlation, Kendall's rank correlation are some commonly used correlation or association measures. A brief description of these correlation metrics given below.

Suppose,$\{x_1, \ldots, x_n\}$ are n i.i.d samples of continuous random variable X. We define another continuous variable Y taking on values $\{y_1, \ldots, y_n\}$ expressed in terms of X. We

consider X as the independent variable and Y as the dependent variable. Here, we will assume that there are no ties in the observations of X and Y. Let, $\mathrm{R}(x_i)$ be the rank of x_i and $\mathrm{R}(y_i)$ be the rank of y_i, so that, $\mathrm{R}(x_i) = \sum_{j=1}^{n} \mathbb{I}\left\{x_{(j)} \leq x_{(i)}\right\}$ and $\mathrm{R}(y_i) = \sum_{j=1}^{n} \mathbb{I}\left\{y_{(j)} \leq y_{(i)}\right\}$ respectively.

1.3.1 Pearson's Product-moment Correlation

Pearson's product-moment correlation is a measure of linear association. Theoretically, a best fitted line is considered using the values of two variables that goes through the expected values. The correlation coefficient then calculates the distance of actual variable values from their respective expected values (Rodgers and Nicewander, 1988). Pearson's correlation coefficient is represented by ρ and is defined mathematically as the covariance of the two variables, normalized by the square root of their variances.

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X) \cdot \mathrm{Var}(Y)}} \tag{1}$$

The range of the coefficient is (-1, +1). Here, $\rho < 0$ indicates negative correlation; $\rho > 0$ indicates positive correlation and 0 indicates no linear association between X and Y.

Correlation measures based on rank is the alternative to address the parametric limitations of Pearson's correlation. Rank correlation coefficients uses ranks of the data rather than the actual observed values (Huang, 2010). Kendall's rank correlation and Spearman's rank correlation are most frequently used rank correlation metrics.

1.3.2 Spearman's Rank Correlation

Spearman's rank correlation is a measure of monotonic relationship. It is appropriate when one or both variables are continuous but skewed or ordinal. Hence, Spearman's rank correlation coefficient is less sensitive to outliers. It can be defined as Pearson's correlation based on the ranks of the variables in place of actual observations (Myers and Well, 2003). So, we

can define the correlation coefficient ρ_s as,

$$\rho_s = \frac{\mathrm{Cov}(\mathrm{R}\,(x_i)\,, \mathrm{R}\,(y_i))}{\sqrt{\mathrm{Var}(\mathrm{R}\,(x_i)) \cdot \mathrm{Var}(\mathrm{R}\,(y_i))}} \tag{2}$$

In the absence of ties, ρ_s can be expressed as,

$$\rho_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} (\mathrm{R}\,(x_i) - \mathrm{R}\,(y_i))^2}{n(n^2-1)} \tag{3}$$

Here, ρ_s can vary within the range, $-1 \leq \rho_s \leq +1$. Positive value of correlation coefficient indicates rank of X and Y are similar and negative correlation coefficient occurs when observations of X and Y have dissimilar ranks. If the correlation coefficient is 0, it means there is no monotonic relationship between variables.

1.3.3 Kendall's Rank Correlation

Kendall's rank correlation is another non-parametric measure of association based on the direction of order of X and Y. Like Spearman's rank correlation, Kendall's tau correlation coefficient is measured for ordinal or continuous data.

For any two pairs of observations (x_i, y_i) and (x_j, y_j) of the variables X and Y, where $i < j$, if (x_i–x_j) and (y_i–y_j) have the same sign, the pairs are called concordant pairs; whereas if they have opposite signs, the pairs are known as discordant pairs (Nelsen, 1999). In absense of ties, the coefficient has the formula,

$$\tau = \frac{(\text{Number of concordant pairs}) - (\text{Number of discordant pairs})}{\binom{n}{2}} \tag{4}$$

Here, $\binom{n}{2} = \frac{n(n-1)}{2}$ is the total number of pairs of observations

Kendall's τ coefficient is obtained by normalizing the symmetric difference. Hence, it takes values between $[-1, +1]$ (Abdi, 2007). When order of one variable is the exact opposite of the other variable resulting largest possible distance then τ is 1. When both variables

have identical order then τ becomes $+1$ corresponding to the smallest possible distance. Therefore, $\tau > 0$ when observations of X and Y are more likely to be concordant than discordant, and there is a positive relationship between X and Y. In contrast, $\tau < 0$ means a negative relationship between X and Y (Huang, 2010).

1.4 Problems with Monotonic Association Measures

Pearson's correlation coefficient aims to measure the linear association between variables. It is a parametric measure used for data in normal or approximately normal distribution. This coefficient is sensitive to outliers. This is not an appropriate measure while dealing with non-linear and non-monotone associations. Both Spearman's rank correlation and Kendall's rank correlation are designed to address monotonic non-linear association. Hence, these applications may return zero correlation even if there is non-monotonic association between X and Y. So, they are not appropriate when dealing with non-monotonic association.

In the study of measure of association, often there are situations where relationship between two variables is non-monotone. For example, consider about 'aging curve' in sports. Athletes' ability increases with age and then decreases. James and Zminda (1988) showed that, baseball players on average perform best around age 27. Young players increase in ability as they get stronger and learn how to play. However, their ability decreases as they become older because they get slower and have more difficulty recovering from injuries. Hence, the ability of players shows non-monotonic association with age having one curve for $x < 27$ and another curve for $x > 27$. Pearson's correlation, Spearman's rank correlation or Kendall's rank correlation fails to measure such association.

Several measures of dependence have appeared in literature over time that claimed to capture monotone as well as non-monotonic association between two continuous variables. The main objective of this thesis is to make detailed evaluations of six such proposed methods and determine the optimal measures for monotonic or non-monotonic associations. Although, correlation typically refer to linear association only, all the methods we reviewed stated non-

linear/non-monotone associations as correlation. Therefore, to maintain consistency with the methods under comparison we would mention any type of association as correlation. A comparison of these six correlation coefficients would be done with three classical methods proposed by Pearson, Spearman, and Kendall. For this purpose, permutation test based on Monte Carlo simulation study has been carried out under twelve different relation types between X and Y. In addition, we would explore the differences of computational time and memory usage by each of these methods. Alongside the performance of correlation measures and their independence test on simulation models, we would apply the methods to compute and compare correlations between four pairs of variables from a real life data.

2 Methodology

In research and industry, one essential aspect is to measure the strength of dependence between two variables. Therefore, it is important to have a suitable coefficient that works as a measure of dependence. Although classical methods are well utilized, they have some limitations. As discussed in previous chapter, they fail to capture non-monotonic association. Over time, several different options have been introduced by different school of thoughts to serve this purpose. Here, we focus on six such methods,

- Hoeffding's D correlation based on joint cumulative distribution functions and ranks (Hoeffding, 1994)
- Optimal correlation obtained by transformed function of variables based on maximal correlation coefficient (Breiman and Friedman, 1985)
- Distance correlation by measuring the pairwise distances (Székely et al., 2007)
- A modified Kendall's correlation measure based on joint cumulative distribution functions and ranks (Bergsma and Dassios, 2014)
- A recent correlation coefficient which is simple in terms of calculation (Chatterjee, 2021)
- An improved method of Chatterjee's correlation (Lin and Han, 2023)

2.1 Correlation Coefficient

We would adopt the definitions of the coefficients of measuring association of these above approaches. However, for purpose of this study, we would modify some computational procedures. We would consider variables X and Y as discussed in chapter 1 (Introduction). A brief discussion on the methods is given below.

2.1.1 Hoeffding's D Correlation

Most conventional alternative measure of the classical correlation approaches is Hoeffding's D correlation that can capture monotonic as well as non-monotonic association. Hoeffding's D correlation coefficient is a non-parametric measure that depends on the rank order of the observations. Theoretically, it calculates the distance between the joint distribution and the product of the marginal distributions of two variables (Hoeffding, 1994). Mathematically, we can define the D statistic as,

$$D = \frac{A - 2(n-2)B + (n-2)(n-3)C}{n(n-1)(n-2)(n-3)(n-4)} \tag{5}$$

Here, $A = \sum_{i=1}^{n} ([R(x_i) - 1][R(x_i) - 2][R(y_i) - 1][R(y_i) - 2])$,

$B = \sum_{i=1}^{n} ([R(x_i) - 2][R(y_i) - 2]c_i)$ and, $C = \sum_{i=1}^{n} (c_i(c_i - 1))$

Also, c_i is the number of bivariate observations of X and Y such that for any (x_j, y_j), $x_j \leq x_i$ and $y_j \leq y_i$ (Wilding and Mudholkar, 2008).

Range of D is $(-\frac{1}{60}, \frac{1}{30})$. For convention, range of D can be expressed as, $-0.5 \leq D \leq 1$, which is 30 times the original D statistic (Harrell Jr, 2023). Therefore, we use the following formula to compute D statistic (Hollander et al., 2013),

$$D = 30 \cdot \frac{A - 2(n-2)B + (n-2)(n-3)C}{n(n-1)(n-2)(n-3)(n-4)} \tag{6}$$

The signs of the coefficient have no interpretation because it identifies non-monotonic relationships. It's value 1 means complete dependence between X and Y whereas 0 means X and Y are independent. The larger the value of D, the more dependent are X and Y.

2.1.2 Maximal Correlation Based on Transformation of Variables

A non-parametric optimization procedure was developed by Breiman and Friedman (1985). The method involves iteration using bivariate conditional expectations until optimum trans-

formation of X and Y is obtained. The transformed functions of X and Y denoted as $\phi(X)$ and $\theta(Y)$ respectively would provide maximal correlation between the variables.

The process starts with setting up initial values $\phi_0(X) = X$ and $\theta_0(Y) = Y/||Y||$ where the norm $||.||$ is defined as $||.|| = \sqrt{E(.)^2}$. Then, we compute the objective function,

$$e^2(\theta, \phi) = E[\theta_0(Y) - \phi_0(X)]^2 \tag{7}$$

Next, we define $\phi_1(X) = E[\theta_0(Y)|X]$ and $\theta_1(Y) = E[\phi_1(X)|Y]/||E[[\phi_1(X)|Y]||$. Here, $E[\theta_0(Y)|X]$ is obtained by fitting local linear regression for K^{th} nearest neighbour (KNN) considering $\theta_0(Y)$ as response and X as predictor. We determined the best value of K by applying cross-validation in each run separately, using the *caret* R-package (Kuhn and Max, 2008). Similarly $E[\phi_1(X)|Y]$ is computed by considering $\phi_1(X)$ as response and Y as predictor by KNN. Then, again we compute the objective function using equation (7) for given $\phi(X)$ and $\theta(Y)$.

The process is repeated until, between two successive steps, the minimization function fails to decrease by more than a pre-specified threshold δ. In each step, we would replace old $\phi(X)$ and $\theta(Y)$ by their new values respectively and minimize the objective function until the optimal solution for $\phi(X)$ and $\theta(Y)$ are obtained. This iterative optimization system is called Alternating Conditional Expectations (ACE). Algorithm 1 summarizes the ACE optimization process.

Correlation between the transformed functions is obtained by taking Pearson product-moment correlation between $\phi(X)$ and $\theta(Y)$. This correlation is the maximal correlation between X and Y. Therefore, the maximal correlation ρ^*, obtained by transformed function is,

$$\rho^* = \rho(\phi(X), \theta(Y)) = \frac{\text{Cov}(\phi(X), \theta(Y)}{\sqrt{\text{Var}(\phi(X)) \cdot \text{Var}(\theta(Y))}} \tag{8}$$

The maximal correlation obtained by ACE algorithm varies between 0 to 1. The correlation

coefficient value become zero when $\phi(X)$ and $\theta(Y)$ are independent of each other,i.e., there is no relation between them. Whereas, a correlation value 1 indicates complete dependence.

Algorithm 1: ACE algorithm to obtain optimum functions of X and Y

Input : Data containing information on X and Y, stopping threshold δ

Output: $\phi(X)$ and $\theta(Y)$ as optimum functions of X and Y

Initialization: Set, $\phi_0(X) = X$ and, $\theta_0(Y) = Y/||Y||$ where, $||.|| = \sqrt{E(.)^2}$

Compute, $e_0^2(\theta, \phi) = E[\theta_0(Y) - \phi_0(X)]^2$

At step t: Define, $\phi_t(X) = E[\theta_{(t-1)}(Y)|X]$ by fitting local linear regression for K^{th} nearest neighbour (value of K obtained by cross-validation for each run) considering $\theta_{(t-1)}(Y)$ as response and X as predictor.

Also define, $\theta_t(Y) = E[\phi_{(t)}(X)|Y]/||E[[\phi_{(t)}(X)|Y]||$ by fitting local linear regression for K^{th} nearest neighbour (value of K obtained by cross-validation for each run) considering $\phi_{(t)}(X)$ as response and Y as predictor.

Compute, $e_t^2(\theta, \phi) = E[\theta_t(Y) - \phi_t(X)]^2$

If $e_{(t-1)}^2(\theta, \phi) - e_{(t)}^2(\theta, \phi) > \delta$, run Step (t+1) as above.

Else, Exit.

2.1.3 Distance Correlation

Székely et al. (2007) initiated the idea of distance correlation. Consider the joint distribution of X and Y. Now, for any two pairs of observations (x_i, y_i) and (x_j, y_j), define,

$$A_{ij} = a_{ij} - \bar{a_{i.}} - \bar{a_{.j}} + \bar{a_{..}} \tag{9}$$

Where, a_{ij} is the absolute distance between x_i and x_j, i.e., $a_{ij} = |x_i - x_j|$,

$\bar{a_{i.}} = \frac{1}{n}\sum_{j=1}^{n} a_{ij}$,$\bar{a_{.j}} = \frac{1}{n}\sum_{i=1}^{n} a_{ij}$, and, $\bar{a_{..}} = \frac{1}{n^2}\sum_{i,j=1}^{n} a_{ij}$

Define B_{ij} in similar way for Y considering, $b_{ij} = |y_i - y_j|$ as the absolute distance between y_i and y_j. Now, the distance correlation $\mathcal{R}(X,Y)$ is defined as,

$$\mathcal{R}(X,Y) = \sqrt{\mathcal{R}^2(X,Y)} = \begin{cases} \frac{\nu(X,Y)}{\sqrt{\nu(X,X)\cdot\nu(Y,Y)}} & \nu(X,X)\cdot\nu(Y,Y) > 0 \\ 0 & \nu(X,X)\cdot\nu(Y,Y) = 0 \end{cases} \tag{10}$$

Here, $\nu_n(X,Y)$ is the non-negative distance covariance. It is defined as,

$$\nu(X,Y) = \sqrt{\frac{1}{n^2}\sum_{i,j=1}^{n} A_{ij}\cdot B_{ij}} \tag{11}$$

Likewise, $\nu(X,X)$ and $\nu(Y,Y)$ are defined as, $\nu(X,X) = \sqrt{\frac{1}{n^2}\sum_{i,j=1}^{n} A_{ij}^2}$ and,
$\nu(Y,Y) = \sqrt{\frac{1}{n^2}\sum_{i,j=1}^{n} B_{ij}^2}$

Distance correlation is analogous to Pearson's product-moment correlation. It can vary within the range, $0 \leq \mathcal{R} \leq 1$, being 0 only if X and Y are independent.

2.1.4 A Modification of Kendall's τ

Bergsma and Dassios (2014) derived a natural extension of Kendall's rank correlation. The empirical expression of kendall's correlation coefficient is,

$$\tau = \frac{1}{n^2}\sum_{i,j=1}^{n} \text{sign}(x_i - x_j)\cdot\text{sign}(y_i - y_j) \tag{12}$$

Where, $\text{sign}(x_i - x_j)$ and $\text{sign}(y_i - y_j)$ are signs of (x_i–x_j) and (y_i–y_j) respectively. Now, for any z_1, z_2, z_3 and z_4, they defined a function,

$$s(z_1,z_2,z_3,z_4) = \text{sign}(|z_1 - z_2|^2 + |z_3 - z_4|^2 - |z_1 - z_3|^2 - |z_2 - z_4|^2) \tag{13}$$

and expressed equation (12) in terms of X an Y as,

$$\tau^2 = \frac{1}{n^4} \sum_{i,j,k,l=1}^{n} s(x_i, x_j, x_k, x_l) \cdot s(y_i, y_j, y_k, y_l) \tag{14}$$

They proposed an improved function based on absolute differences,

$$a(z_1, z_2, z_3, z_4) = \text{sign}(|z_1 - z_2| + |z_3 - z_4| - |z_1 - z_3| - |z_2 - z_4|) \tag{15}$$

Finally derived τ^*, the modified τ coefficient,

$$\tau^*(X, Y) = \frac{1}{n^4} \sum_{i,j,k,l=1}^{n} a(x_i, x_j, x_k, x_l) a(y_i, y_j, y_k, y_l) \tag{16}$$

Following Cauchy–Schwartz inequality, normalized version of τ^* is,

$$\tau_b^* = \frac{\tau^*(X, Y)}{\sqrt{\tau^*(X, X) \cdot \tau^*(Y, Y)}} \tag{17}$$

Here, $\tau^*(X, X) = \frac{1}{n^4} \sum_{i,j,k,l=1}^{n} a(x_i, x_j, x_k, x_l) a(x_i, x_j, x_k, x_l)$. Likewise, we can define $\tau^*(Y, Y)$ for Y.

τ_b^* takes on values between 0 and 1. Any value within this range indicates presence association whereas 0 indicates no association between X and Y.

2.1.5 Chatterjee's Rank-based Approach

Chatterjee (2021) proposed a simple correlation metric which can measure the strength of relationship between two variables. Akin to Pearson's correlation coefficient, this coefficient, expressed in the form of ξ_n, approaches its maximum value subject to one variable comes close to a noiseless function of the other. However, the correlation coefficient is not symmetric in X and Y. So, $\xi_n(X, Y)$ assesses whether Y is a function of X. In contrast, to explore whether X is a function of Y, we need to compute $\xi_n(Y, X)$.

To compute $\xi_n(X,Y)$, rearrange data as, $(x_{(1)},y_{(1)}),...,(x_{(n)},y_{(n)})$ such that, $x_{(1)} \leq ... \leq x_{(n)}$. Let, r_i be the rank of $y_{(i)}$ such that, for any j, $y_{(j)} \leq y_{(i)}$. Also let, l_i be the number of j for which $Y_{(j)} \geq Y_{(i)}$. Now, the correlation coefficient is defined as,

$$\xi_n(X,Y) = 1 - \frac{n \cdot \sum_{i=1}^{n-1} |r_{i+1} - r_i|}{2 \cdot \sum_{i=1}^{n-1} l_i \cdot (n - l_i)} \tag{18}$$

On the contrary, to compute $\xi_n(Y,X)$ we rearrange the data as $(x_{(1)},y_{(1)}),...,(x_{(n)},y_{(n)})$ where, $y_{(1)} \leq ... \leq y_{(n)}$ and assume that r_i be the rank of $x_{(i)}$ such that, for any j, $x_{(j)} \leq x_{(i)}$. Then, using right side of equation (18), we can evaluate $\xi_n(Y,X)$.

When $n \longrightarrow \infty$ Range of $\xi_n(X,Y)$ lies between $[0,1]$. However, for finite n, the maximum and minimum possible values of $\xi_n(X,Y)$ are $\frac{(n-2)}{(n+1)}$ and $-\frac{1}{2} + \mathcal{O}(\frac{1}{n})$, respectively. So, the lower bound is approximately -0.5 and, the upper bound approaches 1 as the sample size n gets larger. Hence, we can consider the range of $\xi_n(X,Y)$ as $[-0.5,1]$.

Although simple in calculation, this correlation measure has some limitations when it comes to testing independence. It lacks power (Chatterjee, 2021; Shi et al., 2022), has a slower critical detection boundary (Auddy et al., 2021). Hence, we focus on an improved version of Chatterjee's correlation coefficient introduced by Lin and Han.

2.1.6 A Modification of Chatterjee's Measure

Lin and Han (2023) incorporated M many right nearest neighbors into Chatterjee's correlation coefficient formula and proposed a reconstructed correlation coefficient.

Let, M be the number of right nearest neighbor. Order data, $(x_{(1)},y_{(1)}),...,(x_{(n)},y_{(n)})$ such that, $x_{(1)} \leq ... \leq x_{(n)}$. Let, r_i be the rank of $y_{(i)}$ and $r_{j_m(i)}$ be the rank of j^{th} nearest neighbor of i^{th} observation (ordered based on X) of Y. Now, the correlation coefficient has the following formula,

$$\xi_{n,M}(X,Y) = -2 + \frac{6 \cdot \sum_{i=1}^{n} \sum_{m=1}^{M} \min(r_i, r_{j_m(i)})}{(n+1)[nM + M(M+1)/4]} \tag{19}$$

Like Chatterjee's correlation, this correlation is asymmetric. Therefore, to compute $\xi_{n,M}(Y,X)$, one needs to order the dataset $(x_{(1)},y_{(1)}),...,(x_{(n)},y_{(n)})$ such that, $y_{(1)} \leq ... \leq y_{(n)}$. Consider, r_i be the rank of $x_{(i)}$ and $r_{j_m(i)}$ be the rank of j^{th} nearest neighbor of i^{th} observation (ordered based on Y) of X. Then using the right side of equation (19) compute the correlation coefficient between (Y,X).

Range of $\xi_{n,M}(X,Y)$ is between $[-0.5,1]$ up to a bias of order $\frac{M}{n}$ for finite sample.

Table 1 summarizes the methods for calculating correlation coefficients and the corresponding ranges of those methods.

Table 1: List of range of all correlation measures under comparison

Type of Correlation	Method/ Reference	Notation	Range
Linear/Monotone associations	Pearson	ρ	$[-1,+1]$
	Spearman	ρ_s	$[-1,+1]$
	Kendall	τ	$[-1,+1]$
Non-monotone associations	Hoeffding	D	$[-0.5,+1]$
	Breiman and Friedman	ρ^*	$[0,+1]$
	Distance	$\mathcal{R}$	$[0,+1]$
	Bergsma and Dassios	τ_b^*	$[0,+1]$
	Chatterjee	ξ_n	$[-0.5,+1]$
	Lin and Han	$\xi_{n,M}$	$[-0.5,+1]$

A few points to be noted,

- Unlike linear/monotone relationships where sign of the measure is important to understand the direction of dependence, they do not have any equivalent meaning for non-monotone relationships since in case of non-monotonic association there is no concept of positive or negative association. Only the strength of relationship can be assessed by any association measures that can capture non-monotonic relationships. So, any correlation estimate closure to the lower bound means weak non-monotone association

whereas, any estimate close to upper bound of correlation measure indicates strong non-monotone association.

- The correlation estimate of Y as a function of X and X as a function of Y are different for Chatterjee's rank based correlation (Chatterjee, 2021) and modified version of Chatterjee's correlation (Lin and Han, 2023). Both of this methods takes into account the right nearest neighbour while calculation of correlation which makes the correlation metrics asymmetric. Furthermore, maximal correlation (Breiman and Friedman, 1985) deals with iterative transformations of the functions of X and Y, it also produces different correlation although the estimates are nearly same. But, the number of iteration is different for correlation of Y as a function of X and X as a function of Y. So, for these three measures we would report correlation for both Y as a function of X and X as a function of Y.

- The maximal correlation is the correlation between the transformed functions of X and Y instead of the actual variables (Breiman and Friedman, 1985). In contrast to other methods which involve fixed number of computational steps depending on sample size, computing maximal correlation requires convergence of an algorithm. Hence, for datasets with identical sample sizes (such as the original data and it's permuted versions) computation time for maximal correlation can be different due to different number of iterations required for the convergence of the algorithm.

2.2 Test of Association

Correlation coefficients are used to measure the strength of association between two variables. However, the measures obtained are based on sample data. It is often difficult to compare or assess values of different measures on same sample as the measures have different definition and range. So, their nature of variability differ. Therefore, for meaningful interpretation and comparison a uniform criterion should be used. One such criterion could be test of

independence so that the p-value of the test can be used for deciding strength of association comparison. There are many tests available, such as, asymptotic tests, permutation tests etc. Here, we would use permutation test based on Monte Carlo approximation to compare all the correlation measures under consideration. Permutations tests are widely used in literature for testing independence of variables (Ludbrook, 1994). An exact permutation test is time consuming for moderately large sample size. Therefore, we would use resampling testing method, also known as Permutation test based on Monte Carlo approximation to compare all the correlation measures under consideration.

2.2.1 Permutation Test

For any correlation measure between two variables X and Y, we are interested in testing the hypothesis,

H_o : X and Y are independent, i.e., there is no association between them

H_a : They are dependent, i.e., there is association between X and Y

A permutation test rejects the null hypothesis in favor of independence for large values of the observed correlation coefficient for given bivariate distribution (DiCiccio and Romano, 2017). The test is carried out as follows,

Let, n = Sample size; T = Number of permutations; α = Level of significance; and, Cor(X,Y) = Observed correlation coefficient between X and Y

For $t = 1, ..., T$; let $(i_{t1}, ..., i_{tn})$ be a random permutation of the values of variable Y, and $Cor_t(X, Y)$ be the correlation coefficient value of the t^{th} resample. For computation of Monte Carlo permutation p-value based on T resamples, one needs to take into account whether the sign of the measure is important in deciding about it's significance as discussed after table 1. In that case, one needs to compute a two-sided p-value as,

$$p\text{-}value = \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}[|Cor_t(X,Y)| > |Cor(X,Y)|] \tag{20}$$

Otherwise, a one sided p-value should be computed as,

$$p\text{-}value = \frac{1}{T}\sum_{t=1}^{T}\mathbb{I}[Cor_t(X,Y) > Cor(X,Y)] \tag{21}$$

Now, for any predetermined α, we compare the p-value obtained by permutation test with the level of significance (α) and decide on independence between X and Y.

- If $p\text{-}value < \alpha$, there exists statistically significant evidence in favor of H_a. So, reject H_0 and accept H_a. Hence, conclude that, X and Y are dependent.
- If $p\text{-}value \geq \alpha$, there does not exist statistically significant evidence in favor of H_a. So, reject H_a and do not reject H_0. So, conclude, X and Y are independent.

3 Simulation Study

In this chapter we present correlation from a bivariate dataset under different settings. First, we experimented with different simulation functions that result in different kind of relationships between the variables. We also explored how the presence of noise, to different extent, influences these correlation measures. Finally, we used datasets with varying number of observations to understand how sample size influences performance of these methods.

3.1 Simulation Setting

For one of the two variables X, we simulate it's observations from, $X \sim Unif(-1, 1)$ in all settings. Table 2 summarizes the simulation functions in terms of X. Graphical representation of the functions is given in figure 1.

Then, we generate the variable Y as: $Y = f(X) + \epsilon$ where $f(X)$ is chosen from the 12 simulation functions in table 2 and $\epsilon \sim N(0, \sigma^2)$ is the amount of noise in data. ϵ is chosen using one of the three noise levels of σ^2,

- **Zero noise:** $\epsilon = 0$
- **Low noise:** $\epsilon \sim N(0, 0.01^2)$
- **High noise:** $\epsilon \sim N(0, 0.1^2)$

For models S_{11} and S_{12}, noise were included inside the exponent function.

3.2 Simulation Results

For each simulation models, we showed the plot of data and tables containing the correlation estimates between variables X and Y along with p-value from the permutation tests. The p-values, reported in tables 3 to 14 are based on 10,000 permutations for all methods except maximal correlation. For maximal correlation, we reduced the number of permutations to 3,000 due to the convergence issue described in previous chapter. We set the stopping

Table 2: Functions used to generate simulation models

Index	Pattern type	Simulation functions
S_1	Linear	$f(X) = X$
S_2	Quadratic	$f(X) = X^2$
S_3	Cosinusoid	$f(X) = cos(2\pi X)$
S_4	W-shaped	$f(X) = \begin{cases} \lvert X + 0.5\rvert & \text{if } X < 0 \\ \lvert X - 0.5\rvert & \text{if } X \geq 0 \end{cases}$
S_5	Bump	$f(X) = \begin{cases} 2X + 1 & \text{if } -1 \leq X \leq 0 \\ -2X + 1 & \text{if } 0 \leq X \leq 1 \end{cases}$
S_6	Zig-zag	$f(X) = \begin{cases} 2.99X + 1.99 & \text{if } -1 \leq X \leq -0.33 \\ -2.99X + 0.01 & \text{if } -0.33 \leq X \leq 0.34 \\ 3.03X - 2.03 & \text{if } 0.34 \leq X \leq 1 \end{cases}$
S_7	Double-bump	$f(X) = \begin{cases} 4X + 3 & \text{if } -1 \leq X \leq -0.5 \\ -4X - 1 & \text{if } -0.5 \leq X \leq 0 \\ 4X - 1 & \text{if } 0 \leq X \leq 0.5 \\ -4X + 3 & \text{if } 0.5 \leq X \leq 1 \end{cases}$
S_8	Cross	$f(X) = \begin{cases} X \text{ if } (f(X) \leq 0 \cup X \leq 0) \cap (f(X) \geq 0 \cup X \geq 0) \\ -X \text{ if } (f(X) \geq 0 \cup X \leq 0) \cap (f(X) \leq 0 \cup X \geq 0) \end{cases}$
S_9	Box	$f(X) = \begin{cases} -X - 1 \text{ if } (f(X) \leq 0 \cup X \leq 0) \\ X + 1 \text{ if } (f(X) \geq 0 \cup X \leq 0) \\ X - 1 \text{ if } (f(X) \leq 0 \cup X \geq 0) \\ -X + 1 \text{ if } (f(X) \geq 0 \cup X \geq 0) \end{cases}$
S_{10}	Parallel	$f(X) = \begin{cases} -X - 1 \text{ if } X \leq 0 \\ -X + 1 \text{ if } X \geq 0 \end{cases}$
S_{11}	Exponent of cube	$f(X) = exp(X^3)$
S_{12}	Exponent of sinusoid	$f(X) = exp(sin(2\pi X)$

threshold, $\delta = 10^{-6}$ for all implementations of maximal correlation in this thesis. As noted in Chapter 2, three of the methods produced different results when the labels of X and Y are switched. In the following tables and associated discussion, we report two columns of results for each one of those methods that correspond to the two different labels. We used the notation (X,Y) and (Y,X) to separate the two sets of labels. The former implies Y is modeled as a function of X and the latter implies the exact opposite.

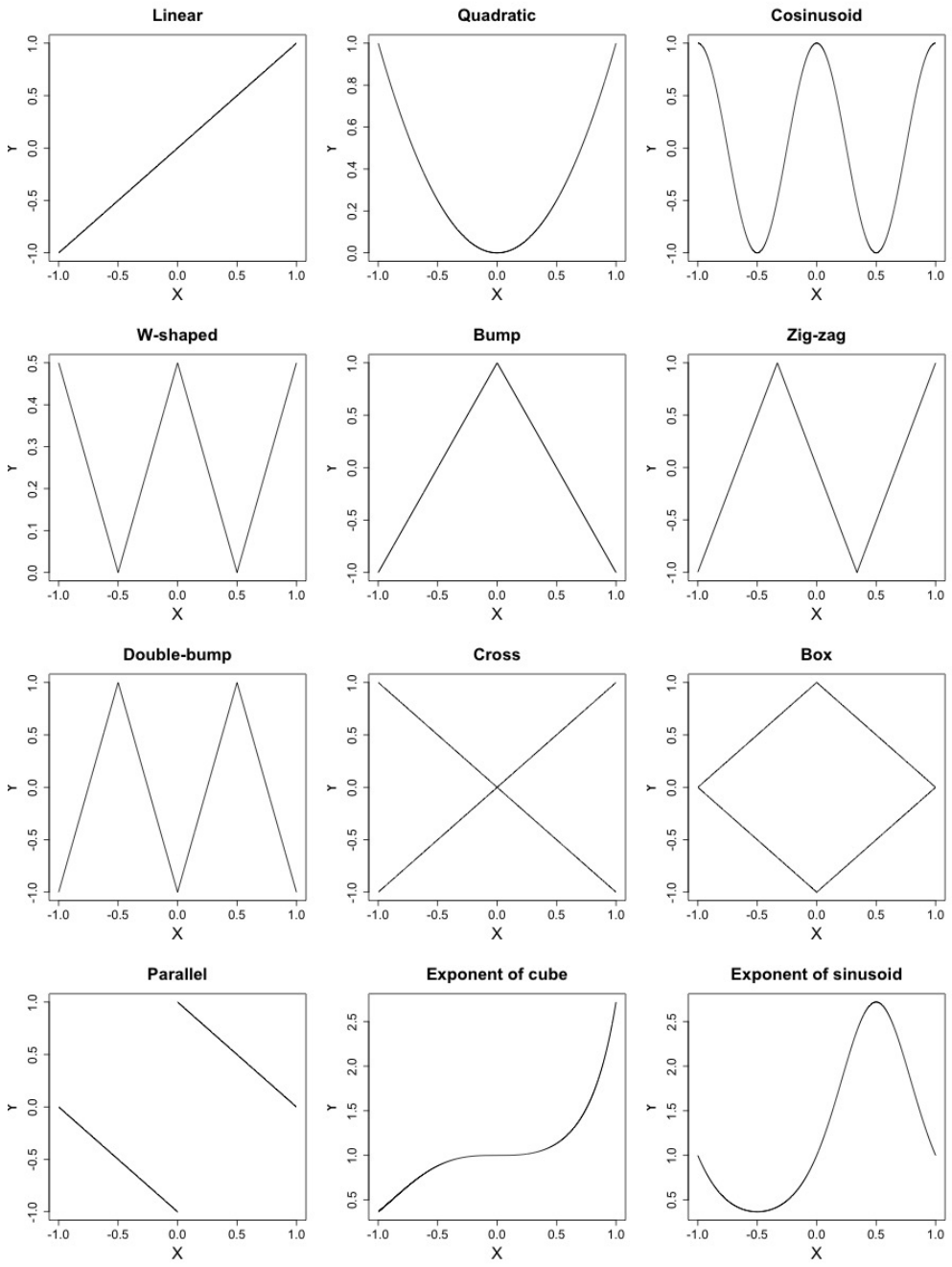

Figure 1: Plot of the functions of Simulation models

As for choice of M in (Lin and Han, 2023), the authors used values of M as exponents of sample size within $(0, 1)$. Choosing the exponent 0 makes M = 1, choosing the exponent $\frac{1}{2}$ makes M = 5 and 10 in case of sample sizes 25 and 100, respectively. So, we explored three values: 1, 5, 10. However, during computation we found that the correlation estimate

at M = 1 is approximately same which is also reported by (Lin and Han, 2023). So, while reporting we dropped the estimates for M = 1 and presented the outcome for M = 5 an 10.

3.2.1 S_1: Linear

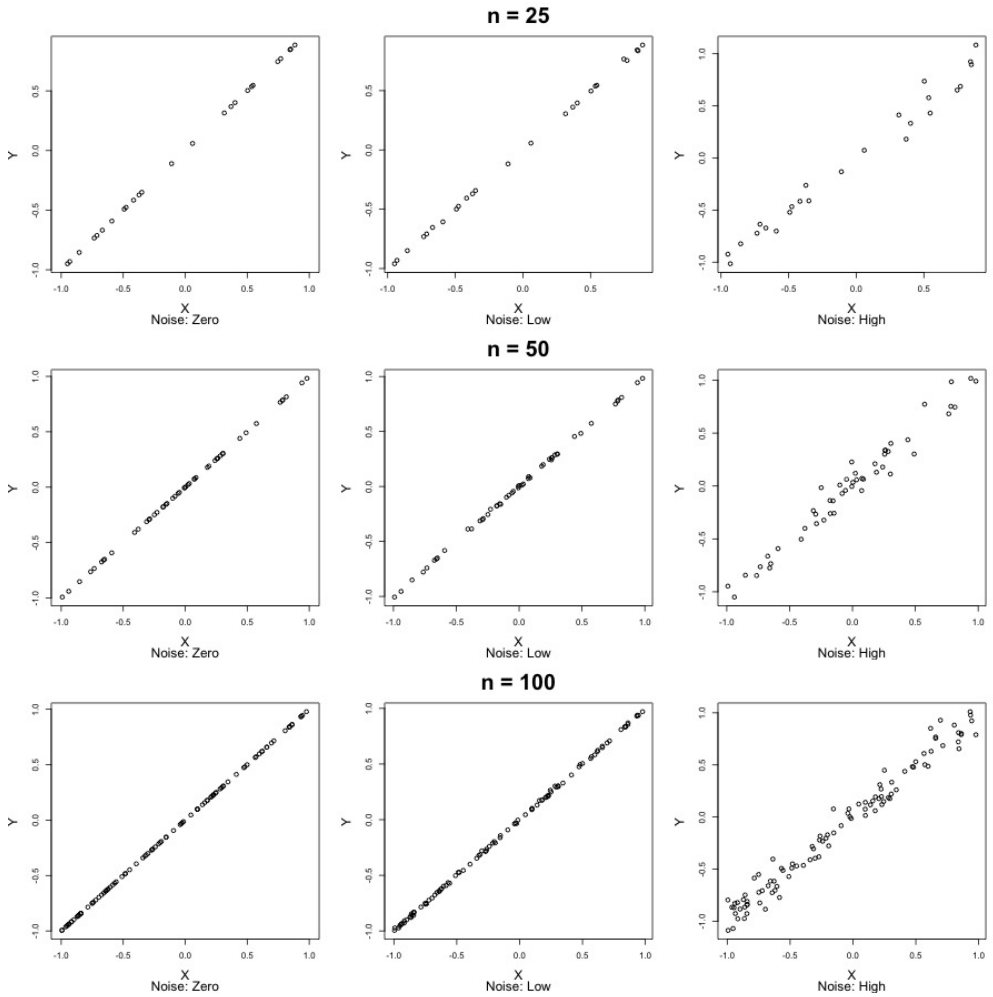

Figure 2: Plot of Simulation models, S_1: Linear

In this model X and Y share linear relation. From Table 3 we observe that, all correlation measures except ξ_n and $\xi_{n,M}$ captured the linear relation perfectly. Even in zero noise, these two methods failed to detect perfect correlation. As the value of M increased for $\xi_{n,M}$, the amount of correlation decreased. For all methods, increasing sample size improved the correlation estimate while adding noise to the model decreased the association.

Table 3: Correlation estimate and p-value for simulation model, S_1: Linear

Sample size	Noise level	Correlation estimate and p-value													
		ρ	ρ_s	τ	D	ρ^*		$\mathcal{R}$	τ_b^*	ξ_n		$\xi_{n,M}$			
						(X,Y)	(Y,X)			(X,Y)	(Y,X)	(X,Y)		(Y,X)	
												M=5	M=10	M=5	M=10
n = 25	Zero	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.885 (0.000)	0.885 (0.000)	0.830 (0.000)	0.703 (0.000)	0.830 (0.000)	0.703 (0.000)
	Low	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.974 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.865 (0.000)	0.875 (0.000)	0.875 (0.000)	0.828 (0.000)	0.702 (0.000)	0.828 (0.000)	0.702 (0.000)
	High	0.981 (0.000)	0.947 (0.000)	0.847 (0.000)	0.988 (0.000)	0.988 (0.000)	0.980 (0.000)	0.979 (0.000)	0.749 (0.000)	0.755 (0.000)	0.764 (0.000)	0.792 (0.000)	0.684 (0.000)	0.792 (0.000)	0.684 (0.000)
n = 50	Zero	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.941 (0.000)	0.941 (0.000)	0.913 (0.000)	0.844 (0.000)	0.913 (0.000)	0.844 (0.000)
	Low	1.000 (0.000)	0.999 (0.000)	0.989 (0.000)	0.978 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.876 (0.000)	0.929 (0.000)	0.929 (0.000)	0.910 (0.000)	0.842 (0.000)	0.910 (0.000)	0.842 (0.000)
	High	0.980 (0.000)	0.975 (0.000)	0.876 (0.000)	.709 (0.000)	0.989 (0.000)	0.989 (0.000)	0.978 (0.000)	0.716 (0.000)	0.756 (0.000)	0.785 (0.000)	0.816 (0.000)	0.786 (0.000)	0.814 (0.000)	0.786 (0.000)
n = 100	Zero	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.970 (0.000)	0.970 (0.000)	0.956 (0.000)	0.920 (0.000)	0.956 (0.000)	0.920 (0.000)
	Low	1.000 (0.000)	1.000 (0.000)	0.991 (0.000)	0.955 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.864 (0.000)	0.948 (0.000)	0.948 (0.000)	0.948 (0.000)	0.916 (0.000)	0.948 (0.000)	0.916 (0.000)
	High	0.985 (0.000)	0.986 (0.000)	0.899 (0.000)	.722 (0.000)	0.988 (0.000)	0.987 (0.000)	0.978 (0.000)	0.724 (0.000)	0.819 (0.000)	0.812 (0.000)	0.840 (0.000)	0.840 (0.000)	0.840 (0.000)	0.840 (0.000)

3.2.2 S_2: Quadratic

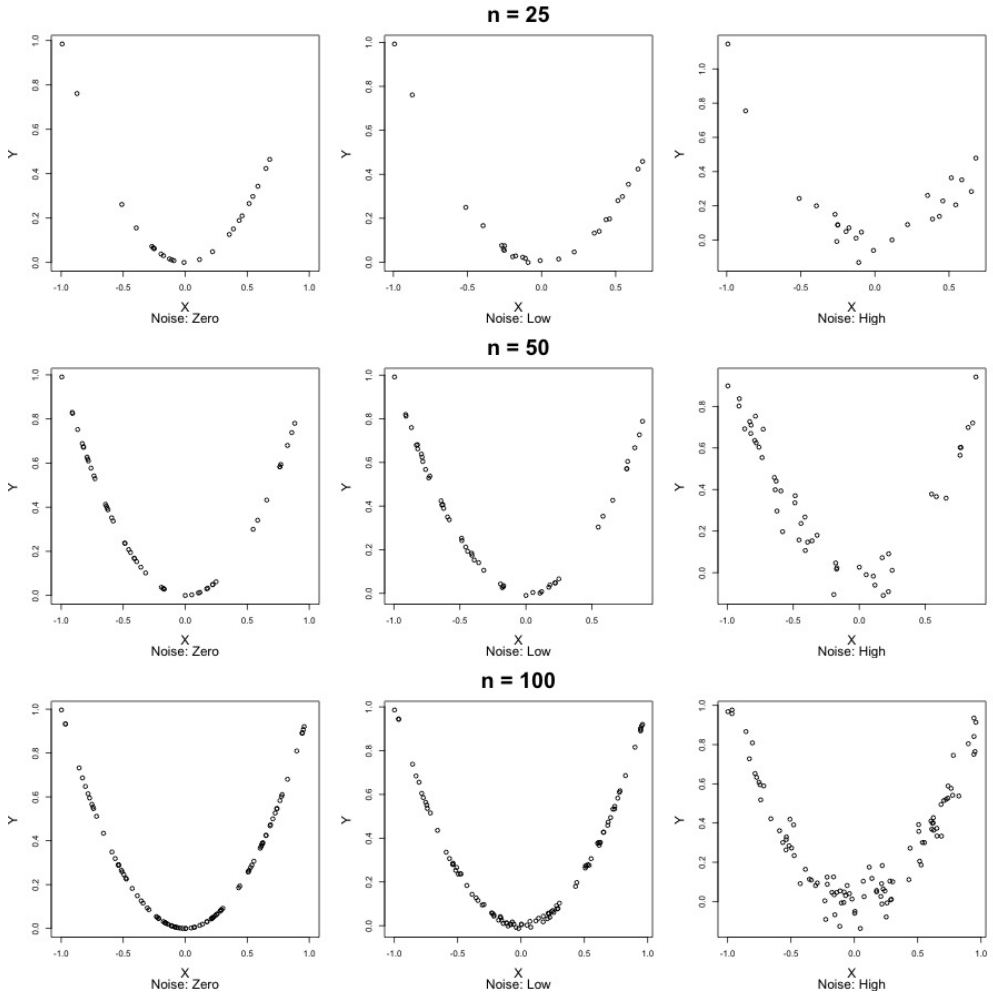

Figure 3: Plot of Simulation models, S_2: Quadratic

Here, Y is a quadratic function of X. Therefore, as X varies from -1 to 1, Y takes on values from 0 to 1. As evident from table 4, three classical methods failed to capture any significant association while all non-monotonic measures captured significant correlation. Increasing sample size improved the correlation estimate for ξ_n and $\xi_{n,M}$ while the correlation decreased for D, $\mathcal{R}$, and τ_b^*. As the value of M increased from 5 to 10 for $\xi_{n,M}$, the amount of correlation decreased. Adding noise to the model decreased association for all methods.

Table 4: Correlation estimate and p-value for simulation model, S_2: Quadratic

Sample size	Noise level	Correlation estimate and p-value													
		ρ	ρ_s	τ	D	ρ^*		$\mathcal{R}$	τ_b^*	ξ_n		$\xi_{n,M}$			
						(X,Y)	(Y,X)			(X,Y)	(Y,X)	(X,Y)		(Y,X)	
												M=5	M=10	M=5	M=10
n = 25	Zero	-0.168 (0.415)	-0.281 (0.179)	-0.240 (0.085)	0.293 (0.000)	0.999 (0.000)	0.999 (0.000)	0.647 (0.000)	0.296 (0.000)	0.788 (0.000)	0.337 (0.002)	0.421 (0.000)	0.236 (0.011)	0.306 (0.001)	0.262 (0.005)
	Low	-0.184 (0.382)	-0.269 (0.192)	-0.220 (0.122)	0.280 (0.000)	0.999 (0.000)	0.999 (0.000)	0.648 (0.001)	0.274 (0.000)	0.760 (0.000)	0.351 (0.001)	0.421 (0.000)	0.239 (0.009)	0.306 (0.001)	0.260 (0.006)
	High	-0.301 (0.143)	-0.388 (0.057)	-0.280 (0.043)	0.172 (0.001)	0.974 (0.000)	0.974 (0.000)	0.604 (0.003)	0.158 (0.005)	0.563 (0.000)	0.255 (0.018)	0.353 (0.000)	0.223 (0.014)	0.243 (0.005)	0.237 (0.011)
n = 50	Zero	-0.200 (0.166)	-0.110 (0.449)	-0.045 (0.643)	0.253 (0.000)	1.000 (0.000)	1.000 (0.000)	0.508 (0.000)	0.207 (0.000)	0.884 (0.000)	0.181 (0.019)	0.607 (0.000)	0.356 (0.000)	0.140 (0.008)	0.109 (0.025)
	Low	-0.202 (0.153)	-0.110 (0.449)	-0.038 (0.700)	0.230 (0.000)	0.999 (0.000)	0.999 (0.000)	0.508 (0.000)	0.189 (0.000)	0.862 (0.000)	0.234 (0.004)	0.601 (0.000)	0.352 (0.000)	0.144 (0.006)	0.112 (0.026)
	High	-0.183 (0.200)	-0.071 (0.622)	-0.033 (0.727)	0.092 (0.000)	0.969 (0.000)	0.969 (0.000)	0.459 (0.001)	0.092 (0.002)	0.545 (0.000)	-0.125 (0.922)	0.402 (0.000)	0.223 (0.000)	0.037 (0.216)	0.040 (0.198)
n = 100	Zero	-0.095 (0.350)	-0.082 (0.417)	-0.061 (0.375)	0.256 (0.000)	1.000 (0.000)	1.000 (0.000)	0.502 (0.000)	0.225 (0.000)	0.941 (0.000)	0.205 (0.000)	0.790 (0.000)	0.638 (0.000)	0.240 (0.000)	0.200 (0.000)
	Low	-0.093 (0.352)	-0.075 (0.455)	-0.060 (0.377)	0.236 (0.000)	1.000 (0.000)	1.000 (0.000)	0.500 (0.000)	0.216 (0.000)	0.909 (0.000)	0.350 (0.000)	0.778 (0.000)	0.633 (0.000)	0.214 (0.000)	0.191 (0.000)
	High	-0.168 (0.096)	-0.134 (0.184)	-0.099 (0.140)	0.115 (0.000)	0.967 (0.000)	0.966 (0.000)	0.471 (0.000)	0.132 (0.000)	0.637 (0.000)	0.040 (0.259)	0.562 (0.000)	0.483 (0.000)	0.110 (0.002)	0.094 (0.005)

3.2.3 S_3: Cosinusoid

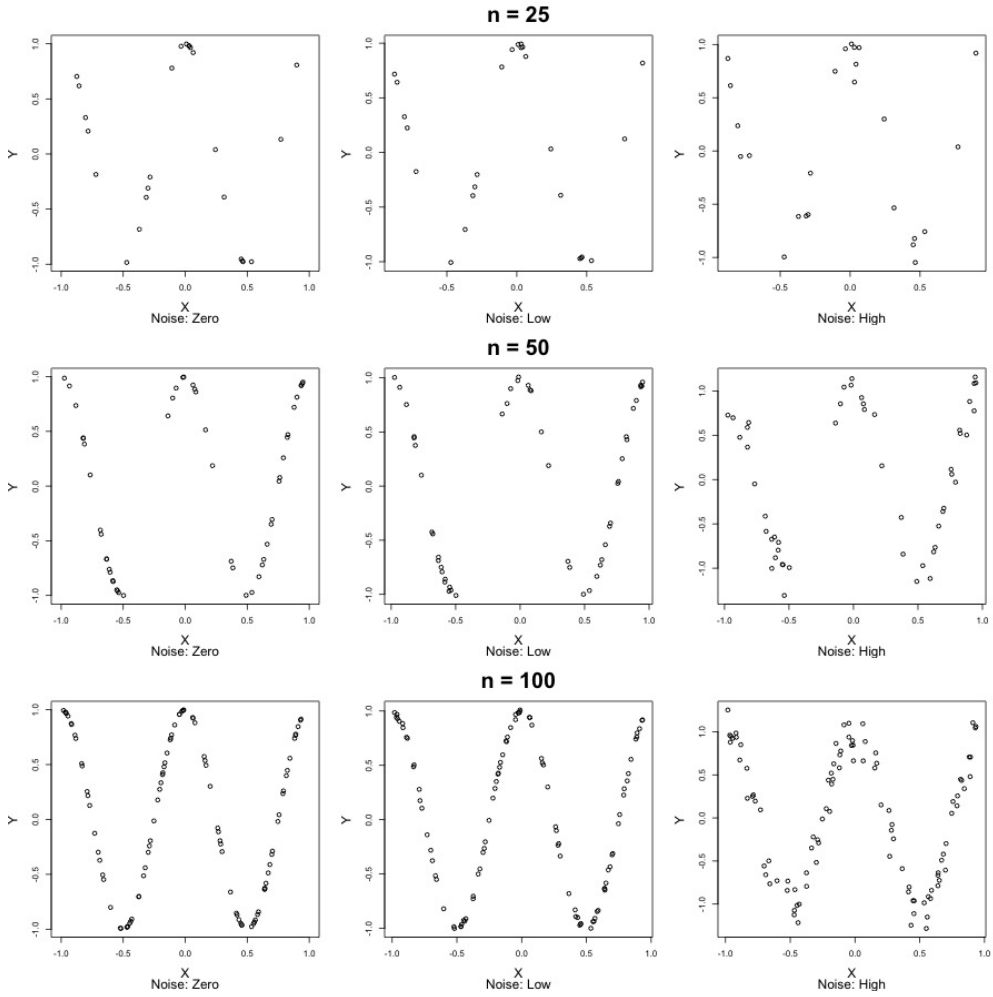

Figure 4: Plot of Simulation models, S_3: Cosinusoid

Table 5 summarizes the correlation estimates and p-value when Y is a cosine function of X. D and τ_b^* captured negligible correlation while the correlations for $\mathcal{R}$ was weak. ξ_n and $\xi_{n,M}$ for M = 5 could measure moderate correlation for all sample sizes.For M = 10, $\xi_{n,M}$ produced weak correlation. In all methods, significant association were obtained when sample size increased.

Table 5: Correlation estimate and p-value for simulation model, S_3: Cosinusoid

Sample size	Noise level	Correlation estimate and p-value													
		ρ	ρ_s	τ	D	ρ^*		$\mathcal{R}$	τ_b^*	ξ_n		$\xi_{n,M}$			
						(X,Y)	(Y,X)			(X,Y)	(Y,X)	(X,Y)		(Y,X)	
												M=5	M=10	M=5	M=10
n = 25	Zero	0.151 (0.468)	0.012 (0.953)	-0.067 (0.630)	0.036 (0.083)	0.751 (0.006)	0.751 (0.009)	0.326 (0.300)	0.035 (0.136)	0.601 (0.000)	0.159 (0.091)	0.153 (0.041)	0.056 (0.253)	0.064 (0.198)	0.080 (0.172)
	Low	0.136 (0.513)	0.024 (0.906)	-0.053 (0.686)	0.037 (0.068)	0.756 (0.006)	0.754 (0.005)	0.327 (0.284)	0.049 (0.091)	0.606 (0.000)	0.216 (0.035)	0.168 (0.026)	0.072 (0.201)	0.071 (0.176)	0.099 (0.131)
	High	0.024 (0.909)	-0.076 (0.722)	-0.093 (0.509)	0.026 (0.118)	0.687 (0.042)	0.711 (0.032)	0.346 (0.248)	0.038 (0.134)	0.548 (0.000)	-0.063 (0.677)	0.170 (0.026)	0.031 (0.330)	0.069 (0.177)	0.060 (0.230)
n = 50	Zero	-0.158 (0.265)	-0.157 (0.276)	-0.138 (0.154)	0.071 (0.001)	0.990 (0.000)	0.989 (0.000)	0.307 (0.065)	0.069 (0.007)	0.784 (0.000)	-0.030 (0.625)	0.380 (0.000)	0.091 (0.045)	0.131 (0.009)	0.118 (0.021)
	Low	-0.160 (0.269)	-0.163 (0.263)	-0.143 (0.141)	0.073 (0.001)	0.989 (0.000)	0.990 (0.000)	0.306 (0.061)	0.070 (0.007)	0.773 (0.000)	0.040 (0.321)	0.374 (0.000)	0.087 (0.054)	0.094 (0.043)	0.099 (0.037)
	High	-0.148 (0.305)	-0.123 (0.384)	-0.073 (0.462)	0.027 (0.028)	0.953 (0.000)	0.953 (0.000)	0.311 (0.069)	0.042 (0.032)	0.625 (0.000)	0.124 (0.081)	0.335 (0.000)	0.076 (0.076)	0.063 (0.105)	0.083 (0.064)
n = 100	Zero	-0.088 (0.386)	-0.074 (0.455)	-0.047 (0.492)	0.075 (0.000)	0.999 (0.000)	0.999 (0.000)	0.341 (0.001)	0.079 (0.000)	0.886 (0.000)	0.068 (0.141)	0.642 (0.000)	0.403 (0.000)	0.111 (0.001)	0.064 (0.025)
	Low	-0.086 (0.398)	-0.070 (0.485)	-0.046 (0.513)	0.057 (0.000)	0.998 (0.000)	0.998 (0.000)	0.288 (0.006)	0.049 (0.002)	0.875 (0.000)	0.003 (0.475)	0.633 (0.000)	0.391 (0.000)	0.037 (0.117)	0.020 (0.234)
	High	-0.153 (0.128)	-0.144 (0.149)	-0.093 (0.170)	0.035 (0.001)	0.972 (0.000)	0.966 (0.000)	0.280 (0.009)	0.037 (0.004)	0.696 (0.000)	-0.074 (0.884)	0.552 (0.000)	0.361 (0.000)	-0.004 (0.527)	-0.011 (0.628)

3.2.4 S_4: W-shaped

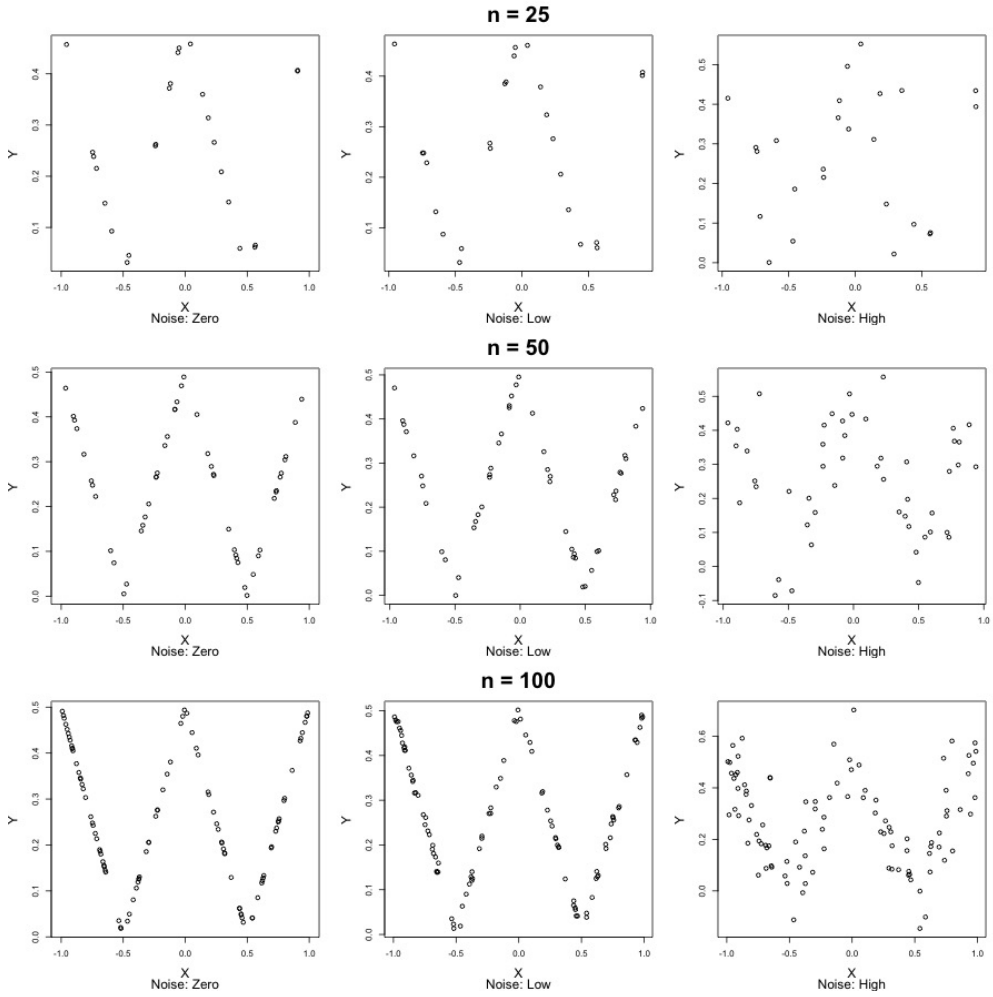

Figure 5: Plot of Simulation models, S_4: W-shaped

When X and Y share a W-shaped relation, D, $\mathcal{R}$ and τ_b^* provide negligible significant correlation which did not have much impact of increased sample size. M= 5 gave relatively large correlation than M = 10 for $\xi_{n,M}$. ξ_n generated moderate significant correlation but we could not identify any pattern due to increased sample size.

Table 6: Correlation estimate and p-value for simulation model, S_4: W-shaped

Sample size	Noise level	Correlation estimate and p-value													
		ρ	ρ_s	τ	D	ρ^*		$\mathcal{R}$	τ_b^*	ξ_n		$\xi_{n,M}$			
						(X,Y)	(Y,X)			(X,Y)	(Y,X)	(X,Y)		(Y,X)	
												M=5	M=10	M=5	M=10
n = 25	Zero	0.173 (0.402)	0.012 (0.950)	-0.067 (0.635)	0.036 (0.074)	0.758 (0.006)	0.752 (0.011)	0.338 (0.284)	0.057 (0.084)	0.601 (0.000)	0.159 (0.091)	0.153 (0.038)	0.056 (0.247)	0.064 (0.191)	0.080 (0.181)
	Low	0.133 (0.523)	0.025 (0.903)	-0.040 (0.762)	0.037 (0.068)	0.760 (0.007)	0.766 (0.003)	0.340 (0.282)	0.063 (0.071)	0.587 (0.000)	-0.014 (0.539)	0.154 (0.039)	0.067 (0.218)	0.073 (0.173)	0.091 (0.153)
	High	-0.090 (0.661)	-0.166 (0.428)	-0.140 (0.323)	0.026 (0.117)	0.389 (0.654)	0.502 (0.356)	0.386 (0.165)	0.088 (0.031)	0.495 (0.000)	0.207 (0.041)	0.207 (0.015)	0.035 (0.319)	0.052 (0.239)	0.054 (0.253)
n = 50	Zero	-0.134 (0.353)	-0.157 (0.269)	-0.138 (0.158)	0.071 (0.001)	0.989 (0.000)	0.989 (0.000)	0.319 (0.057)	0.076 (0.007)	0.784 (0.000)	-0.030 (0.631)	0.380 (0.000)	0.091 (0.049)	0.131 (0.011)	0.118 (0.019)
	Low	-0.138 (0.335)	-0.157 (0.275)	-0.133 (0.169)	0.068 (0.001)	0.989 (0.000)	0.989 (0.000)	0.317 (0.053)	0.068 (0.008)	0.766 (0.000)	0.047 (0.289)	0.374 (0.000)	0.088 (0.056)	0.132 (0.010)	0.094 (0.045)
	High	-0.104 (0.473)	-0.088 (0.539)	-0.033 (0.726)	0.004 (0.225)	0.886 (0.000)	0.837 (0.001)	0.315 (0.066)	0.004 (0.301)	0.414 (0.000)	0.193 (0.013)	0.247 (0.000)	0.049 (0.159)	0.065 (0.103)	0.079 (0.069)
n = 100	Zero	-0.080 (0.431)	-0.074 (0.451)	-0.047 (0.481)	0.033 (0.003)	1.000 (0.000)	1.000 (0.000)	0.328 (0.002)	0.079 (0.000)	0.435 (0.000)	-0.041 (0.740)	0.362 (0.000)	0.278 (0.000)	0.068 (0.025)	0.079 (0.010)
	Low	-0.075 (0.457)	-0.068 (0.500)	-0.042 (0.532)	0.056 (0.000)	0.996 (0.000)	0.996 (0.000)	0.294 (0.007)	0.048 (0.001)	0.868 (0.000)	-0.055 (0.805)	0.631 (0.000)	0.389 (0.000)	0.005 (0.402)	0.008 (0.367)
	High	-0.210 (0.035)	-0.188 (0.062)	-0.118 (0.080)	0.013 (0.025)	0.843 (0.000)	0.850 (0.000)	0.248 (0.033)	0.018 (0.047)	0.418 (0.000)	0.006 (0.453)	0.333 (0.000)	0.220 (0.000)	-0.024 (0.778)	-0.013 (0.649)

3.2.5 S_5: Bump

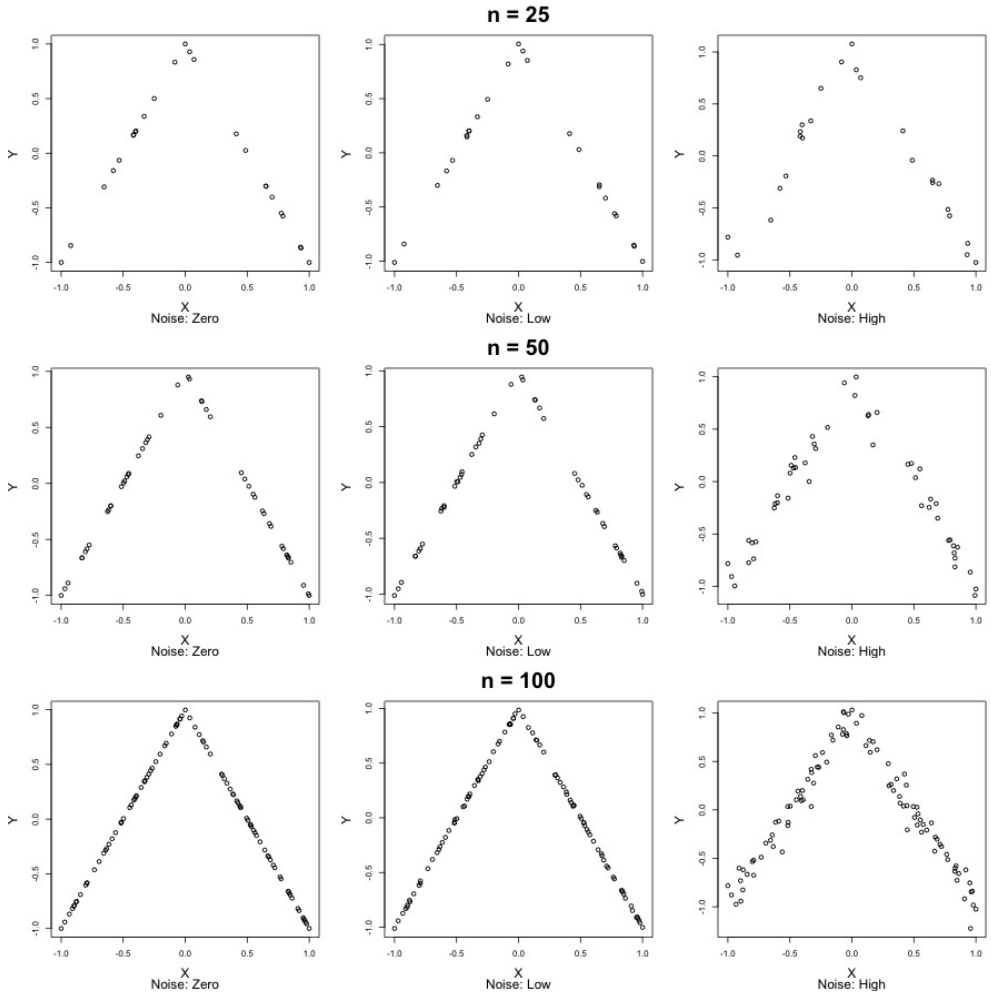

Figure 6: Plot of Simulation models, S_5: Bump

From table 7 we see, D and τ_b^* both correlations are significantly weak. However, increasing sample size increased the correlation for D whereas decreased for τ_b^*. $\mathcal{R}$ and ξ_n correlations were moderate when sample size was 25. Increasing sample sizes reduced the correlation coefficient value of $\mathcal{R}$ while improved the correlation coefficient value of ξ_n. $\xi_{n,M}$ produced weak and negligible correlation for M = 5 and M = 10 respectively. Correlation for both choice of M increased by increasing sample size.

Table 7: Correlation estimate and p-value for simulation model, S_5: Bump

Sample size	Noise level	Correlation estimate and p-value													
		ρ	ρ_s	τ	D	ρ^*		$\mathcal{R}$	τ_b^*	ξ_n		$\xi_{n,M}$			
						(X,Y)	(Y,X)			(X,Y)	(Y,X)	(X,Y)		(Y,X)	
												M=5	M=10	M=5	M=10
n = 25	Zero	-0.216 (0.293)	-0.271 (0.186)	-0.223 (0.110)	0.208 (0.000)	1.000 (0.000)	1.000 (0.000)	0.622 (0.001)	0.249 (0.003)	0.779 (0.000)	0.154 (0.100)	0.357 (0.001)	0.095 (0.138)	0.194 (0.016)	0.108 (0.117)
	Low	-0.218 (0.294)	-0.262 (0.205)	-0.220 (0.115)	0.186 (0.000)	1.000 (0.000)	1.000 (0.000)	0.624 (0.001)	0.195 (0.005)	0.755 (0.000)	0.260 (0.014)	0.349 (0.000)	0.091 (0.149)	0.156 (0.035)	0.098 (0.136)
	High	-0.238 (0.253)	-0.275 (0.175)	-0.227 (0.114)	0.191 (0.000)	0.980 (0.000)	0.980 (0.000)	0.617 (0.001)	0.217 (0.003)	0.736 (0.000)	0.356 (0.002)	0.348 (0.000)	0.085 (0.166)	0.114 (0.081)	0.094 (0.144)
n = 50	Zero	0.121 (0.400)	0.059 (0.684)	0.023 (0.808)	0.226 (0.000)	1.000 (0.000)	1.000 (0.000)	0.506 (0.000)	0.193 (0.000)	0.885 (0.000)	0.149 (0.044)	0.605 (0.000)	0.348 (0.000)	0.167 (0.002)	0.114 (0.021)
	Low	0.123 (0.396)	0.057 (0.687)	0.025 (0.793)	0.218 (0.000)	1.000 (0.000)	1.000 (0.000)	0.507 (0.000)	0.167 (0.000)	0.876 (0.000)	0.114 (0.094)	0.604 (0.000)	0.349 (0.000)	0.148 (0.007)	0.112 (0.021)
	High	0.100 (0.487)	0.045 (0.763)	0.011 (0.904)	0.163 (0.000)	0.986 (0.000)	0.985 (0.000)	0.500 (0.001)	0.149 (0.001)	0.707 (0.000)	0.008 (0.462)	0.554 (0.000)	0.342 (0.000)	0.102 (0.032)	0.113 (0.023)
n = 100	Zero	0.025 (0.796)	-0.008 (0.938)	-0.030 (0.684)	0.245 (0.000)	1.000 (0.000)	1.000 (0.000)	0.501 (0.000)	0.226 (0.000)	0.941 (0.000)	0.157 (0.006)	0.790 (0.000)	0.632 (0.000)	0.220 (0.000)	0.179 (0.000)
	Low	0.025 (0.799)	-0.003 (0.975)	-0.027 (0.684)	0.237 (0.000)	1.000 (0.000)	1.000 (0.000)	0.501 (0.000)	0.197 (0.000)	0.932 (0.000)	0.220 (0.000)	0.788 (0.000)	0.631 (0.000)	0.217 (0.000)	0.177 (0.000)
	High	0.027 (0.789)	0.003 (0.974)	-0.031 (0.647)	0.180 (0.000)	0.988 (0.000)	0.982 (0.000)	0.494 (0.000)	0.169 (0.000)	0.816 (0.000)	0.279 (0.000)	0.731 (0.000)	0.601 (0.000)	0.231 (0.000)	0.192 (0.000)

3.2.6 S_6: Zig-zag

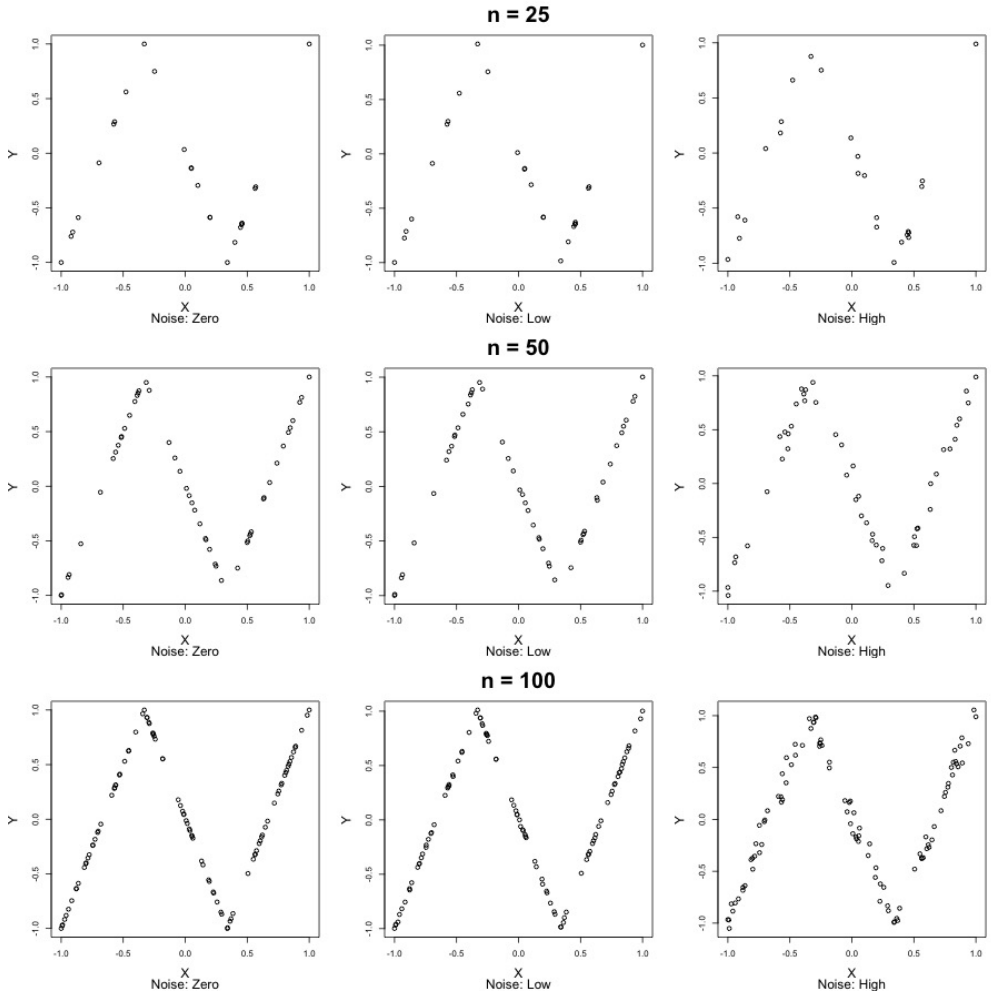

Figure 7: Plot of Simulation models, S_6: Zig-zag

Table 8 exhibit that, correlations between the variables were quite negligible for D and τ_b^*. $\mathcal{R}$ and $\xi_{n,M}$ provide weak significant correlation. $\mathcal{R}$ showed no pattern based on increased sample size. In contrast, $\xi_{n,M}$ increased by increasing sample size but decreased by increasing value of M. Strong correlations were obtained by ξ_n than other methods and had an increasing correlation pattern due to increase of sample.

Table 8: Correlation estimate and p-value for simulation model, S_6: Zig-zag

Sample size	Noise level	Correlation estimate and p-value													
		ρ	ρ_s	τ	D	ρ^*		$\mathcal{R}$	τ_b^*	ξ_n		$\xi_{n,M}$			
						(X,Y)	(Y,X)			(X,Y)	(Y,X)	(X,Y)		(Y,X)	
												M=5	M=10	M=5	M=10
n = 25	Zero	0.051 (0.811)	-0.008 (0.968)	0.013 (0.913)	0.064 (0.025)	0.857 (0.000)	0.860 (0.000)	0.389 (0.135)	0.051 (0.135)	0.668 (0.000)	-0.077 (0.727)	0.337 (0.001)	0.057 (0.250)	0.123 (0.066)	0.100 (0.133)
	Low	0.51 (0.810)	-0.008 (0.970)	0.020 (0.875)	0.122 (0.003)	0.857 (0.000)	0.855 (0.000)	0.536 (0.011)	0.101 (0.040)	0.663 (0.000)	0.106 (0.186)	0.400 (0.000)	0.246 (0.009)	0.235 (0.007)	0.293 (0.003)
	High	0.026 (0.906)	0.005 (0.979)	0.013 (0.904)	0.046 (0.054)	0.853 (0.000)	0.858 (0.000)	0.391 (0.129)	0.068 (0.083)	0.582 (0.000)	0.048 (0.340)	0.348 (0.000)	0.082 (0.170)	0.151 (0.041)	0.095 (0.139)
n = 50	Zero	0.495 (0.001)	0.467 (0.000)	0.383 (0.000)	0.111 (0.000)	0.992 (0.00)	0.992 (0.000)	0.405 (0.006)	0.114 (0.002)	0.831 (0.000)	0.197 (0.012)	0.621 (0.000)	0.372 (0.000)	0.220 (0.000)	0.152 (0.006)
	Low	0.495 (0.000)	0.466 (0.001)	0.381 (0.000)	0.109 (0.000)	0.992 (0.000)	0.992 (0.000)	0.405 (0.007)	0.098 (0.003)	0.831 (0.000)	0.226 (0.005)	0.620 (0.000)	0.370 (0.000)	0.209 (0.001)	0.137 (0.010)
	High	0.485 (0.001)	0.433 (0.002)	0.327 (0.001)	0.083 (0.000)	0.982 (0.000)	0.983 (0.000)	0.397 (0.009)	0.094 (0.005)	0.743 (0.000)	0.259 (0.002)	0.599 (0.000)	0.361 (0.000)	0.169 (0.002)	0.111 (0.022)
n = 100	Zero	0.269 (0.006)	0.261 (0.009)	0.201 (0.004)	0.116 (0.000)	1.000 (0.000)	1.000 (0.000)	0.398 (0.000)	0.133 (0.000)	0.912 (0.000)	-0.014 (0.578)	0.791 (0.000)	0.636 (0.000)	0.113 (0.001)	0.103 (0.002)
	Low	0.270 (0.007)	0.260 (0.009)	0.202 (0.002)	0.115 (0.000)	1.000 (0.000)	1.000 (0.000)	0.399 (0.000)	0.128 (0.000)	0.909 (0.000)	0.030 (0.309)	0.790 (0.000)	0.635 (0.000)	0.097 (0.004)	0.101 (0.002)
	High	0.273 (0.007)	0.264 (0.009)	0.195 (0.004)	0.086 (0.000)	0.991 (0.000)	0.991 (0.000)	0.386 (0.000)	0.109 (0.000)	0.817 (0.000)	0.071 (0.126)	0.759 (0.000)	0.621 (0.000)	0.100 (0.004)	0.099 (0.003)

3.2.7 S_7: Double-bump

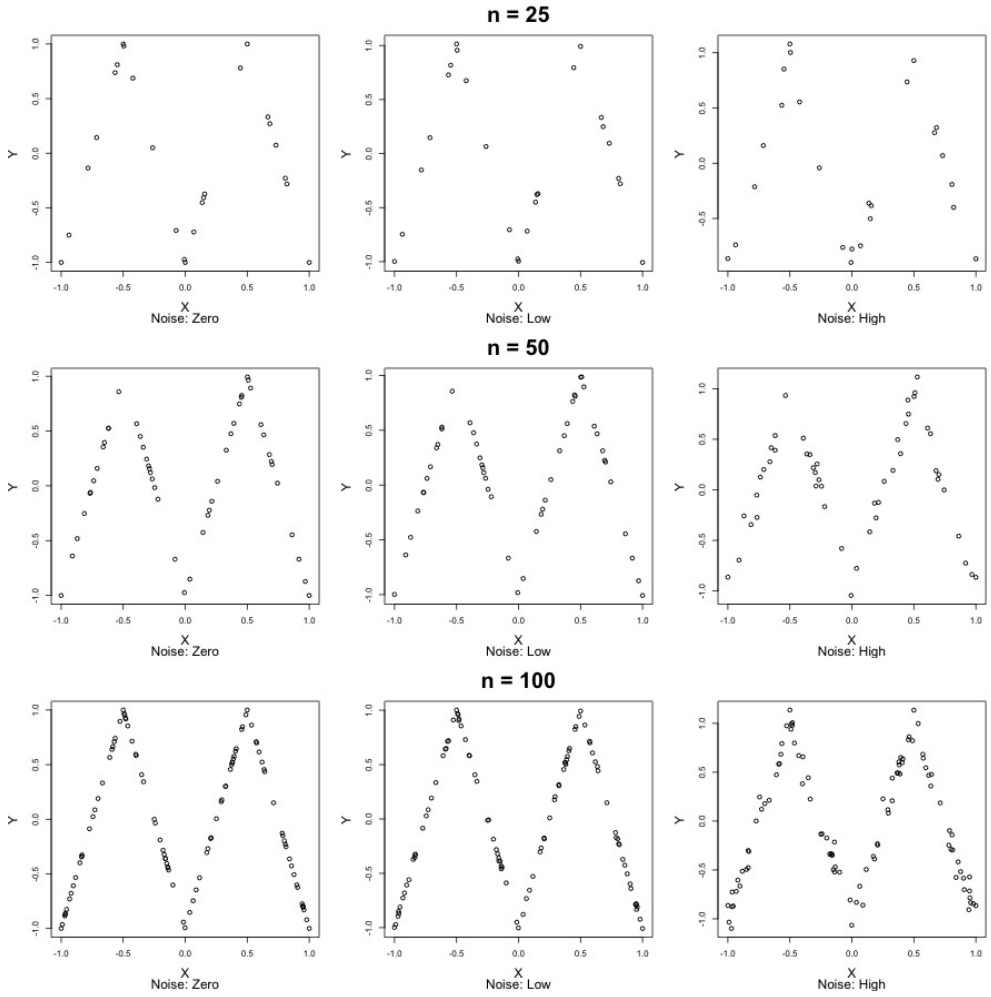

Figure 8: Plot of Simulation models, S_7: Double-bump

From table 9 we observe that, D and τ_b^* correlations were negligible for all samples. $\mathcal{R}$ correlation were weak. Similarly, $\xi_{n,M}$ provide weak correlation which increased by increasing sample size. Increasing M from 5 to 10 reduced the correlation to a noticeable amount. ξ_n produced moderate correlation as opposed to other methods and had an increasing correlation pattern due to increase of sample.

Table 9: Correlation estimate and p-value for simulation model, S_7: Double-bump

Sample size	Noise level	Correlation estimate and p-value													
		ρ	ρ_s	τ	D	ρ^*		$\mathcal{R}$	τ_b^*	ξ_n		$\xi_{n,M}$			
						(X,Y)	(Y,X)			(X,Y)	(Y,X)	(X,Y)		(Y,X)	
												M=5	M=10	M=5	M=10
n = 25	Zero	0.045 (0.833)	0.036 (0.863)	0.040 (0.761)	0.018 (0.175)	0.861 (0.002)	0.860 (0.001)	0.361 (0.190)	0.014 (0.297)	0.577 (0.000)	0.087 (0.226)	-0.028 (0.610)	-0.119 (0.938)	-0.001 (0.450)	-0.050 (0.678)
	Low	0.044 (0.834)	0.042 (0.831)	0.047 (0.736)	0.017 (0.171)	0.859 (0.001)	0.863 (0.000)	0.359 (0.197)	0.021 (0.242)	0.563 (0.000)	0.038 (0.366)	-0.037 (0.639)	-0.125 (0.942)	-0.075 (0.829)	-0.067 (0.765)
	High	0.029 (0.895)	0.030 (0.877)	0.047 (0.727)	0.004 (0.311)	0.828 (0.002)	0.830 (0.002)	0.355 (0.220)	0.005 (0.398)	0.481 (0.000)	-0.029 (0.575)	0.005 (0.428)	-0.063 (0.733)	-0.028 (0.590)	-0.010 (0.504)
n = 50	Zero	-0.127 (0.370)	-0.138 (0.329)	-0.098 (0.312)	0.043 (0.007)	0.981 (0.000)	0.981 (0.000)	0.283 (0.118)	0.023 (0.138)	0.774 (0.000)	-0.019 (0.578)	0.371 (0.000)	0.050 (0.161)	0.059 (0.119)	0.070 (0.091)
	Low	-0.127 (0.379)	-0.140 (0.333)	0.097 (0.318)	0.043 (0.007)	0.980 (0.000)	0.980 (0.000)	0.284 (0.114)	0.031 (0.082)	0.771 (0.000)	0.046 (0.297)	0.374 (0.000)	0.050 (0.153)	0.057 (0.122)	0.069 (0.095)
	High	-0.155 (0.278)	-0.176 (0.223)	-0.122 (0.216)	0.035 (0.013)	0.972 (0.000)	0.972 (0.000)	0.285 (0.118)	0.024 (0.123)	0.699 (0.000)	0.053 (0.267)	0.369 (0.000)	0.063 (0.115)	0.054 (0.130)	0.098 (0.040)
n = 100	Zero	0.093 (0.356)	0.106 (0.294)	0.085 (0.213)	0.050 (0.000)	0.998 (0.000)	0.998 (0.000)	0.252 (0.022)	0.039 (0.013)	0.883 (0.000)	-0.032 (0.691)	0.635 (0.000)	0.395 (0.000)	0.027 (0.189)	0.024 (0.206)
	Low	0.094 (0.352)	0.105 (0.298)	0.085 (0.206)	0.049 (0.000)	0.998 (0.000)	0.998 (0.000)	0.252 (0.023)	0.033 (0.016)	0.879 (0.000)	0.014 (0.401)	0.634 (0.000)	0.394 (0.000)	0.007 (0.394)	0.013 (0.295)
	High	0.096 (0.338)	0.097 (0.336)	0.075 (0.262)	0.037 (0.001)	0.983 (0.000)	0.983 (0.000)	0.245 (0.032)	0.028 (0.024)	0.803 (0.000)	0.063 (0.162)	0.605 (0.000)	0.381 (0.000)	-0.040 (0.911)	-0.011 (0.621)

3.2.8 S_8: Cross

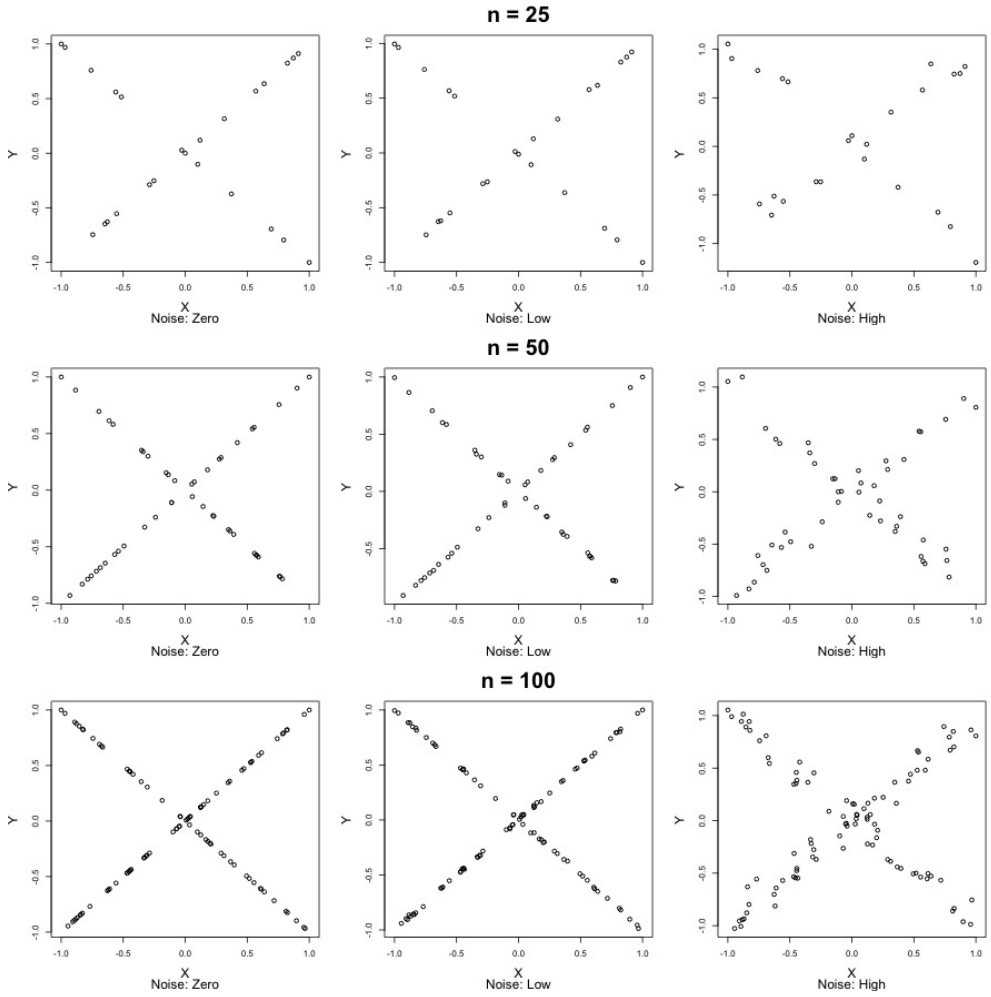

Figure 9: Plot of Simulation models, S_8: Cross

All the methods are consistent in capturing correlation when X and Y share cross pattern relationship. The association captured by each method are significantly small which suggests the association between X and Y is weak/negligible. For all methods, increasing sample size increased the significance level. Increasing M did not have any impact on correlation pattern of $\xi_{n,M}$. Correlation obtained by ξ_n decreased by increasing sample size.

Table 10: Correlation estimate and p-value for simulation model, S_8: Cross

Sample size	Noise level	Correlation estimate and p-value													
		ρ	ρ_s	τ	D	ρ^*		$\mathcal{R}$	τ_b^*	ξ_n		$\xi_{n,M}$			
						(X,Y)	(Y,X)			(X,Y)	(Y,X)	(X,Y)		(Y,X)	
												M=5	M=10	M=5	M=10
n = 25	Zero	-0.038 (0.855)	-0.028 (0.898)	-0.027 (0.833)	0.018 (0.162)	0.988 (0.000)	0.988 (0.000)	0.440 (0.065)	0.069 (0.093)	0.260 (0.015)	0.010 (0.457)	0.282 (0.002)	0.222 (0.016)	0.325 (0.001)	0.220 (0.015)
	Low	-0.040 (0.843)	-0.039 (0.844)	-0.047 (0.718)	0.017 (0.171)	0.989 (0.000)	0.989 (0.000)	0.359 (0.197)	0.021 (0.242)	0.563 (0.000)	0.038 (0.366)	-0.037 (0.639)	-0.125 (0.942)	-0.075 (0.829)	-0.067 (0.765)
	High	-0.071 (0.734)	-0.072 (0.733)	-0.047 (0.724)	0.004 (0.311)	0.928 (0.000)	0.928 (0.000)	0.355 (0.220)	0.005 (0.398)	0.481 (0.000)	-0.029 (0.575)	0.005 (0.428)	-0.063 (0.733)	-0.028 (0.590)	-0.010 (0.504)
n = 50	Zero	-0.170 (0.231)	-0.160 (0.262)	-0.144 (0.141)	0.051 (0.005)	1.000 (0.000)	1.000 (0.000)	0.319 (0.057)	0.032 (0.089)	0.238 (0.003)	0.173 (0.025)	0.093 (0.040)	0.093 (0.045)	0.010 (0.032)	0.090 (0.046)
	Low	-0.167 (0.252)	-0.158 (0.278)	-0.144 (0.137)	0.047 (0.004)	0.999 (0.000)	0.999 (0.000)	0.319 (0.050)	0.024 (0.120)	0.223 (0.005)	0.262 (0.001)	0.093 (0.041)	0.093 (0.045)	0.114 (0.017)	0.091 (0.046)
	High	-0.194 (0.180)	-0.204 (0.153)	-0.146 (0.134)	0.033 (0.017)	0.962 (0.000)	0.962 (0.000)	0.311 (0.064)	0.023 (0.130)	0.192 (0.013)	0.068 (0.215)	0.073 (0.076)	0.088 (0.051)	0.145 (0.006)	0.088 (0.054)
n = 100	Zero	0.000 (0.999)	-0.001 (0.992)	0.025 (0.721)	0.051 (0.000)	1.000 (0.000)	1.000 (0.000)	0.296 (0.005)	0.059 (0.003)	0.189 (0.002)	0.023 (0.356)	0.188 (0.000)	0.200 (0.000)	0.070 (0.000)	0.206 (0.000)
	Low	0.001 (0.993)	-0.007 (0.944)	0.021 (0.752)	0.046 (0.000)	0.999 (0.000)	0.999 (0.000)	0.297 (0.006)	0.062 (0.001)	0.182 (0.003)	0.061 (0.161)	0.187 (0.000)	0.199 (0.000)	0.182 (0.000)	0.208 (0.000)
	High	0.003 (0.977)	0.002 (0.977)	0.013 (0.850)	0.026 (0.003)	0.944 (0.000)	0.945 (0.000)	0.286 (0.008)	0.033 (0.016)	0.109 (0.041)	0.109 (0.042)	0.143 (0.000)	0.172 (0.000)	0.118 (0.001)	0.133 (0.000)

3.2.9 S_9: Box

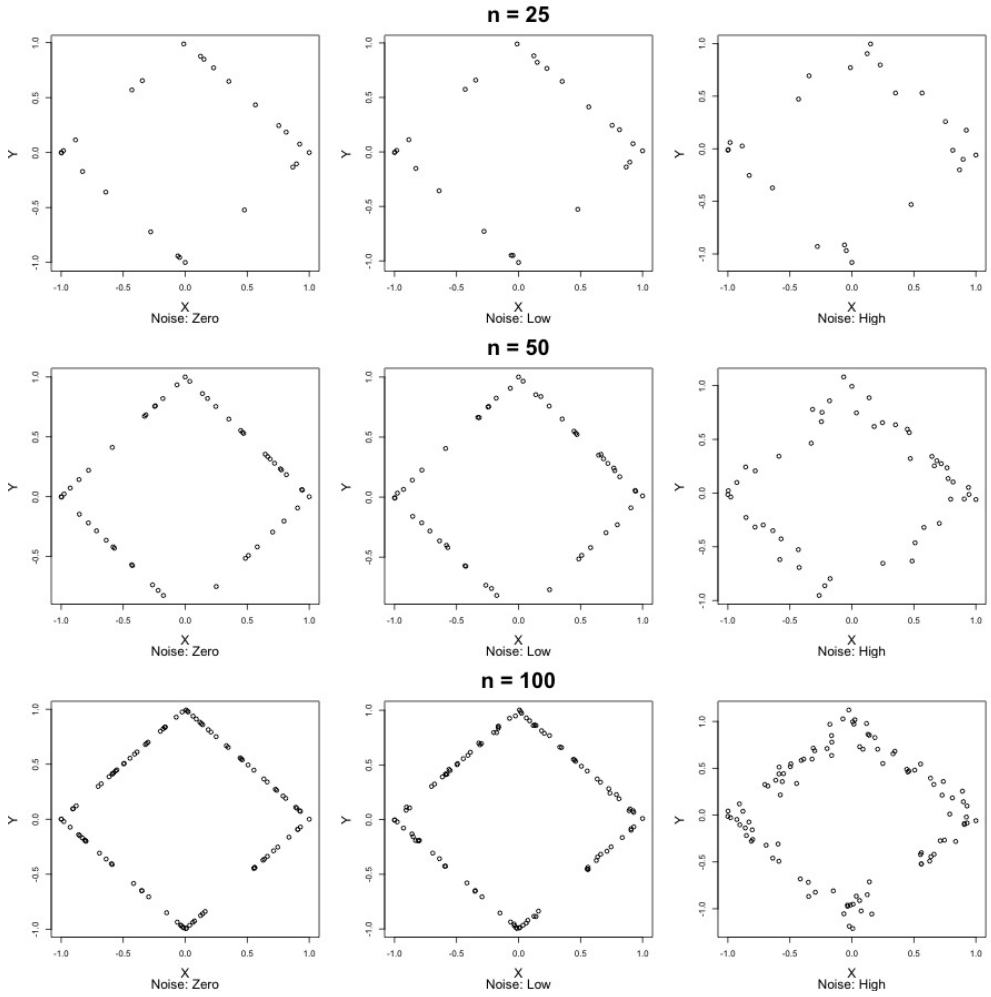

Figure 10: Plot of Simulation models, S_9: Box

Similar to previous model, table 11 suggest that in model S_9 there was very weak association between X and Y. All methods were consistent in capturing significant correlation. Although, increasing sample size did not show any improvement in increasing the correlation for any method.

Table 11: Correlation estimate and p-value for simulation model, S_9: Box

Sample size	Noise level	Correlation estimate and p-value													
		ρ	ρ_s	τ	D	ρ^*		$\mathcal{R}$	τ_b^*	ξ_n		$\xi_{n,M}$			
						(X,Y)	(Y,X)			(X,Y)	(Y,X)	(X,Y)		(Y,X)	
												M=5	M=10	M=5	M=10
n = 25	Zero	-0.059 (0.773)	0.013 (0.949)	0.030 (0.826)	0.052 (0.038)	0.986 (0.000)	0.986 (0.000)	0.402 (0.107)	0.073 (0.089)	0.240 (0.024)	0.216 (0.036)	0.034 (0.297)	-0.015 (0.528)	0.036 (0.297)	0.001 (0.462)
	Low	-0.060 (0.776)	0.010 (0.959)	0.020 (0.874)	0.051 (0.043)	0.984 (0.000)	0.984 (0.000)	0.403 (0.109)	0.090 (0.053)	0.236 (0.028)	0.216 (0.032)	0.038 (0.284)	-0.010 (0.503)	0.036 (0.289)	0.001 (0.458)
	High	-0.072 (0.729)	-0.015 (0.945)	-0.020 (0.873)	0.029 (0.101)	0.957 (0.000)	0.957 (0.000)	0.386 (0.139)	0.049 (0.134)	0.240 (0.023)	0.211 (0.041)	0.039 (0.274)	-0.004 (0.481)	0.046 (0.255)	0.008 (0.420)
n = 50	Zero	-0.063 (0.658)	-0.056 (0.698)	0.000 (0.992)	0.056 (0.003)	0.999 (0.000)	0.999 (0.000)	0.289 (0.102)	0.042 (0.059)	0.205 (0.009)	0.230 (0.003)	0.115 (0.020)	0.026 (0.274)	0.116 (0.019)	-0.001 (0.466)
	Low	-0.065 (0.662)	-0.053 (0.716)	0.006 (0.954)	0.053 (0.004)	0.998 (0.000)	0.998 (0.000)	0.290 (0.100)	0.031 (0.090)	0.202 (0.012)	0.233 (0.003)	0.115 (0.018)	0.027 (0.276)	0.112 (0.021)	0.001 (0.457)
	High	-0.087 (0.542)	-0.094 (0.517)	-0.058 (0.559)	0.028 (0.029)	0.951 (0.000)	0.951 (0.000)	0.258 (0.182)	0.022 (0.129)	0.168 (0.028)	0.053 (0.270)	0.089 (0.046)	0.020 (0.318)	0.051 (0.150)	-0.005 (0.499)
n = 100	Zero	0.070 (0.494)	0.064 (0.526)	0.000 (0.995)	0.058 (0.000)	0.999 (0.000)	0.997 (0.000)	0.279 (0.009)	0.054 (0.004)	0.244 (0.000)	0.034 (0.291)	0.160 (0.000)	0.112 (0.001)	0.145 (0.000)	0.096 (0.003)
	Low	0.072 (0.474)	0.064 (0.523)	0.000 (0.993)	0.055 (0.000)	0.999 (0.000)	0.993 (0.000)	0.278 (0.008)	0.045 (0.006)	0.237 (0.000)	0.053 (0.195)	0.157 (0.000)	0.112 (0.001)	0.136 (0.000)	0.103 (0.001)
	High	0.068 (0.498)	0.065 (0.524)	0.020 (0.762)	0.032 (0.001)	0.963 (0.000)	0.965 (0.000)	0.269 (0.014)	0.031 (0.017)	0.156 (0.007)	0.218 (0.000)	0.116 (0.002)	0.091 (0.004)	0.093 (0.005)	0.081 (0.008)

3.2.10 S_{10}: Parallel

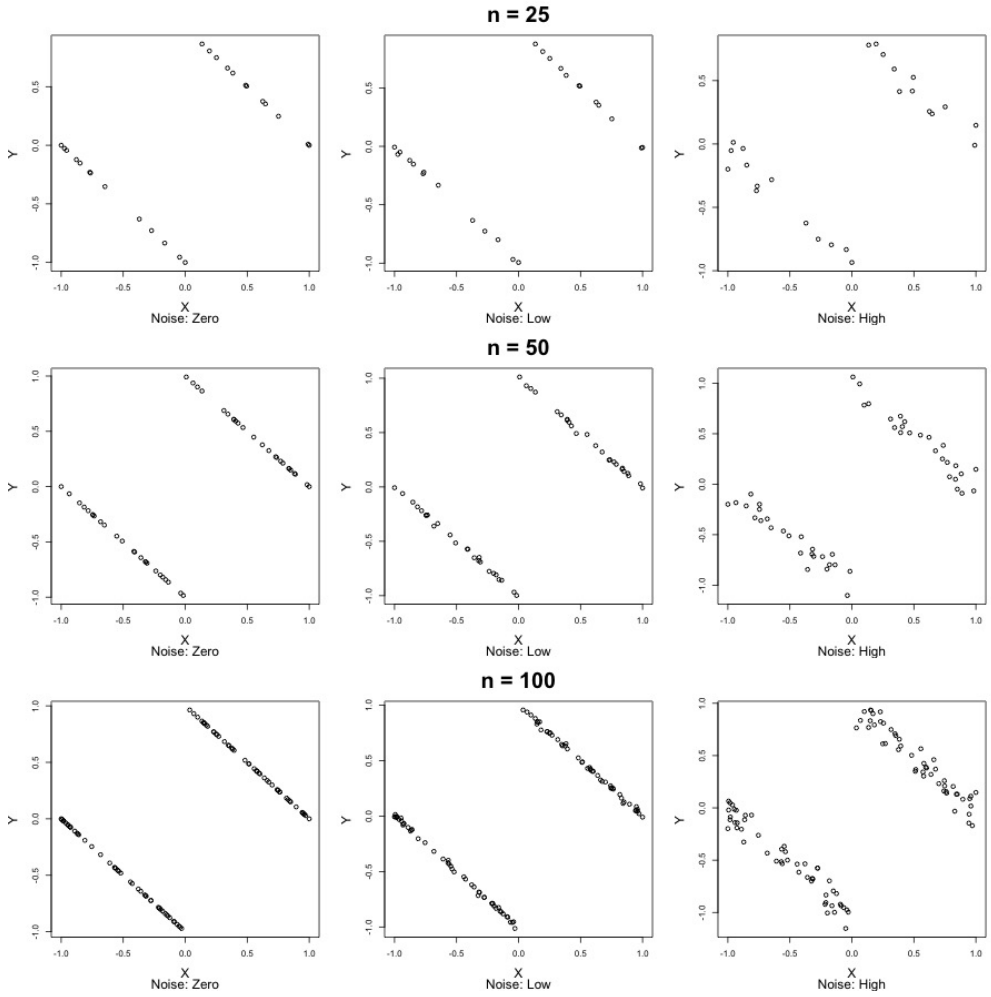

Figure 11: Plot of Simulation models, S_{10}: Parallel

From table 12 we see that, D were not consistent with other method in capturing correlation. Other non-monotonic measures captured moderate to strong correlation while Hoeffding's D correlation was negligible. Increasing sample increased the correlation coefficient for $\mathcal{R}$, ξ_n and $\xi_{n,M}$. The rate of increase in correlation due to increase in sample was slower for $\mathcal{R}$ as compared to ξ_n and $\xi_{n,M}$. Increased sample did not change the correlations obtained by τ_b^*.

Table 12: Correlation estimate and p-value for simulation model, S_{10}: Parallel

Sample size	Noise level	Correlation estimate and p-value													
		ρ	ρ_s	τ	D	ρ^*		$\mathcal{R}$	τ_b^*	ξ_n		$\xi_{n,M}$			
						(X,Y)	(Y,X)			(X,Y)	(Y,X)	(X,Y)		(Y,X)	
												M=5	M=10	M=5	M=10
n = 25	Zero	0.414 (0.042)	0.491 (0.014)	0.037 (0.779)	0.040 (0.069)	0.993 (0.000)	0.993 (0.000)	0.694 (0.001)	0.378 (0.000)	0.769 (0.000)	0.774 (0.000)	0.360 (0.000)	0.191 (0.028)	0.369 (0.000)	0.200 (0.025)
	Low	0.417 (0.038)	0.502 (0.012)	0.053 (0.694)	0.036 (0.079)	0.992 (0.000)	0.993 (0.000)	0.694 (0.000)	0.373 (0.000)	0.764 (0.000)	0.760 (0.000)	0.369 (0.000)	0.200 (0.022)	0.369 (0.000)	0.200 (0.020)
	High	0.396 (0.047)	0.480 (0.017)	0.147 (0.299)	-0.002 (0.419)	0.965 (0.000)	0.965 (0.000)	0.677 (0.001)	0.344 (0.000)	0.649 (0.000)	0.423 (0.000)	0.341 (0.000)	0.199 (0.023)	0.332 (0.001)	0.212 (0.018)
n = 50	Zero	0.571 (0.000)	0.498 (0.000)	0.020 (0.831)	0.050 (0.004)	1.000 (0.000)	1.000 (0.000)	0.687 (0.000)	0.444 (0.000)	0.882 (0.000)	0.884 (0.000)	0.618 (0.000)	0.400 (0.000)	0.620 (0.000)	0.402 (0.000)
	Low	0.572 (0.000)	0.497 (0.000)	0.030 (0.755)	0.047 (0.004)	0.999 (0.000)	0.999 (0.000)	0.684 (0.000)	0.395 (0.000)	0.869 (0.000)	0.764 (0.000)	0.614 (0.000)	0.398 (0.000)	0.601 (0.000)	0.395 (0.000)
	High	0.562 (0.000)	0.503 (0.000)	0.136 (0.163)	0.010 (0.143)	0.972 (0.000)	0.973 (0.000)	0.645 (0.000)	0.338 (0.000)	0.694 (0.000)	0.521 (0.000)	0.531 (0.000)	0.371 (0.000)	0.460 (0.000)	0.350 (0.000)
n = 100	Zero	0.474 (0.000)	0.500 (0.000)	0.010 (0.886)	0.056 (0.000)	1.000 (0.000)	1.000 (0.000)	0.720 (0.000)	0.396 (0.000)	0.941 (0.000)	0.941 (0.000)	0.792 (0.000)	0.646 (0.000)	0.793 (0.000)	0.646 (0.000)
	Low	0.475 (0.000)	0.499 (0.000)	0.021 (0.762)	0.049 (0.000)	1.000 (0.000)	1.000 (0.000)	0.719 (0.000)	0.363 (0.000)	0.923 (0.000)	0.867 (0.000)	0.786 (0.000)	0.642 (0.000)	0.768 (0.000)	0.636 (0.000)
	High	0.460 (0.000)	0.482 (0.000)	0.106 (0.121)	0.020 (0.008)	0.977 (0.000)	0.967 (0.000)	0.716 (0.000)	0.333 (0.000)	0.805 (0.000)	0.703 (0.000)	0.712 (0.000)	0.606 (0.000)	0.666 (0.000)	0.583 (0.000)

3.2.11 S_{11}: Exponent of cube

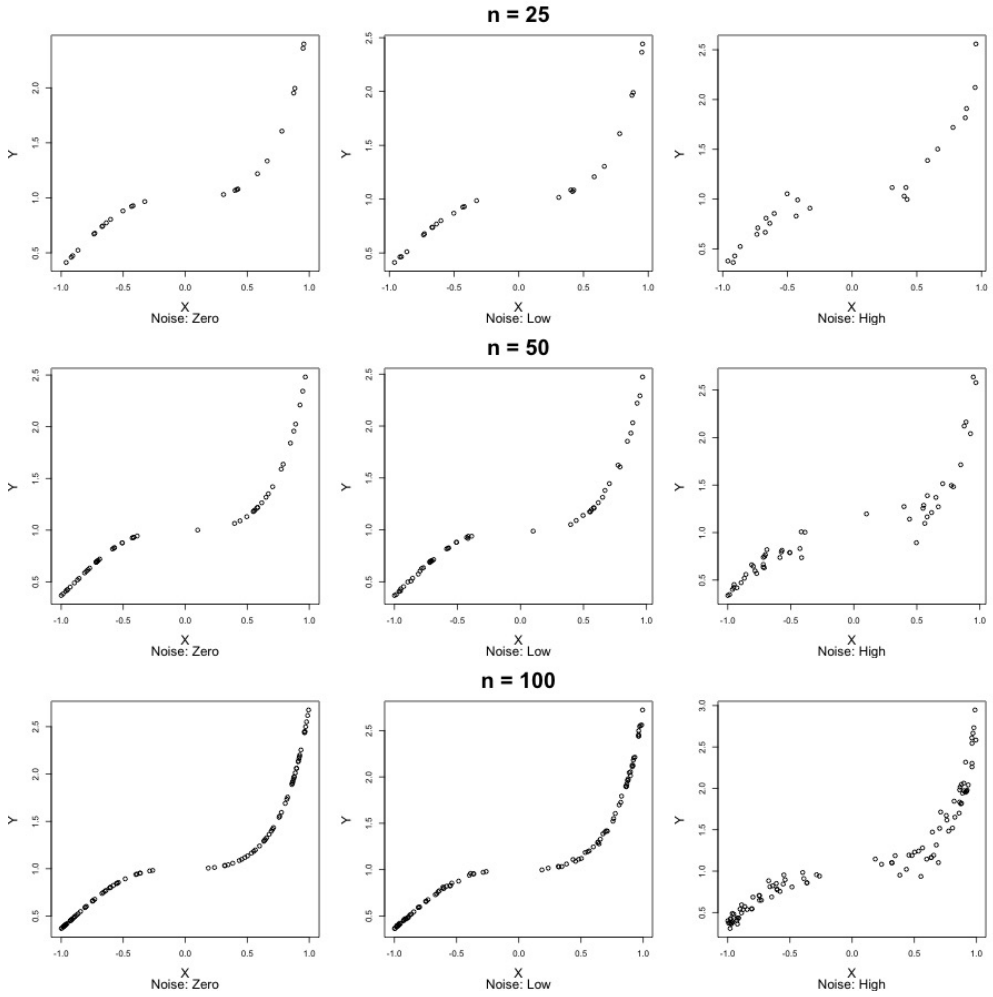

Figure 12: Plot of Simulation models, S_{11}: Exponent of cube

From figure 12 we observe that, X and Y share monotonic relation. All the methods including the classical approaches were consistent in capturing very strong significant correlation between X and Y.

Table 13: Correlation estimate and p-value for simulation model, S_{11}: Exponent of cube

Sample size	Noise level	Correlation estimate and p-value													
		ρ	ρ_s	τ	D	ρ^*		$\mathcal{R}$	τ_b^*	ξ_n		$\xi_{n,M}$			
						(X,Y)	(Y,X)			(X,Y)	(Y,X)	(X,Y)		(Y,X)	
												M=5	M=10	M=5	M=10
n = 25	Zero	0.915 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.995 (0.000)	0.993 (0.000)	0.962 (0.000)	0.814 (0.000)	0.884 (0.000)	0.884 (0.000)	0.830 (0.000)	0.703 (0.000)	0.830 (0.000)	0.703 (0.000)
	Low	0.915 (0.000)	0.998 (0.000)	0.987 (0.000)	1.000 (0.000)	0.995 (0.000)	0.993 (0.000)	0.962 (0.000)	0.821 (0.000)	0.885 (0.000)	0.885 (0.000)	0.830 (0.000)	0.703 (0.000)	0.830 (0.000)	0.703 (0.000)
	High	0.901 (0.000)	0.979 (0.000)	0.900 (0.000)	0.774 (0.000)	0.983 (0.000)	0.987 (0.000)	0.962 (0.000)	0.716 (0.000)	0.769 (0.000)	0.750 (0.000)	0.787 (0.000)	0.682 (0.000)	0.787 (0.000)	0.682 (0.000)
n = 50	Zero	0.915 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.939 (0.000)	0.816 (0.000)	0.941 (0.000)	0.941 (0.000)	0.913 (0.000)	0.844 (0.000)	0.913 (0.000)	0.844 (0.000)
	Low	0.914 (0.000)	1.000 (0.000)	0.997 (0.000)	0.971 (0.000)	1.000 (0.000)	0.999 (0.000)	0.940 (0.000)	0.795 (0.000)	0.926 (0.000)	0.926 (0.000)	0.909 (0.000)	0.842 (0.000)	0.909 (0.000)	0.842 (0.000)
	High	0.887 (0.000)	0.975 (0.000)	0.871 (0.000)	0.775 (0.000)	0.987 (0.000)	0.987 (0.000)	0.908 (0.000)	0.692 (0.000)	0.821 (0.000)	0.810 (0.000)	0.850 (0.000)	0.811 (0.000)	0.850 (0.000)	0.811 (0.000)
n = 100	Zero	0.918 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.931 (0.000)	0.817 (0.000)	0.970 (0.000)	0.970 (0.000)	0.956 (0.000)	0.920 (0.000)	0.956 (0.000)	0.920 (0.000)
	Low	0.917 (0.000)	0.999 (0.000)	0.990 (0.000)	0.958 (0.000)	1.000 (0.000)	1.000 (0.000)	0.930 (0.000)	0.794 (0.000)	0.950 (0.000)	0.951 (0.000)	0.948 (0.000)	0.916 (0.000)	0.948 (0.000)	0.916 (0.000)
	High	0.905 (0.000)	0.981 (0.000)	0.881 (0.000)	0.757 (0.000)	0.987 (0.000)	0.989 (0.000)	0.913 (0.000)	0.689 (0.000)	0.850 (0.000)	0.848 (0.000)	0.871 (0.000)	0.862 (0.000)	0.869 (0.000)	0.861 (0.000)

3.2.12 S_{12}: Exponent of sinusoid

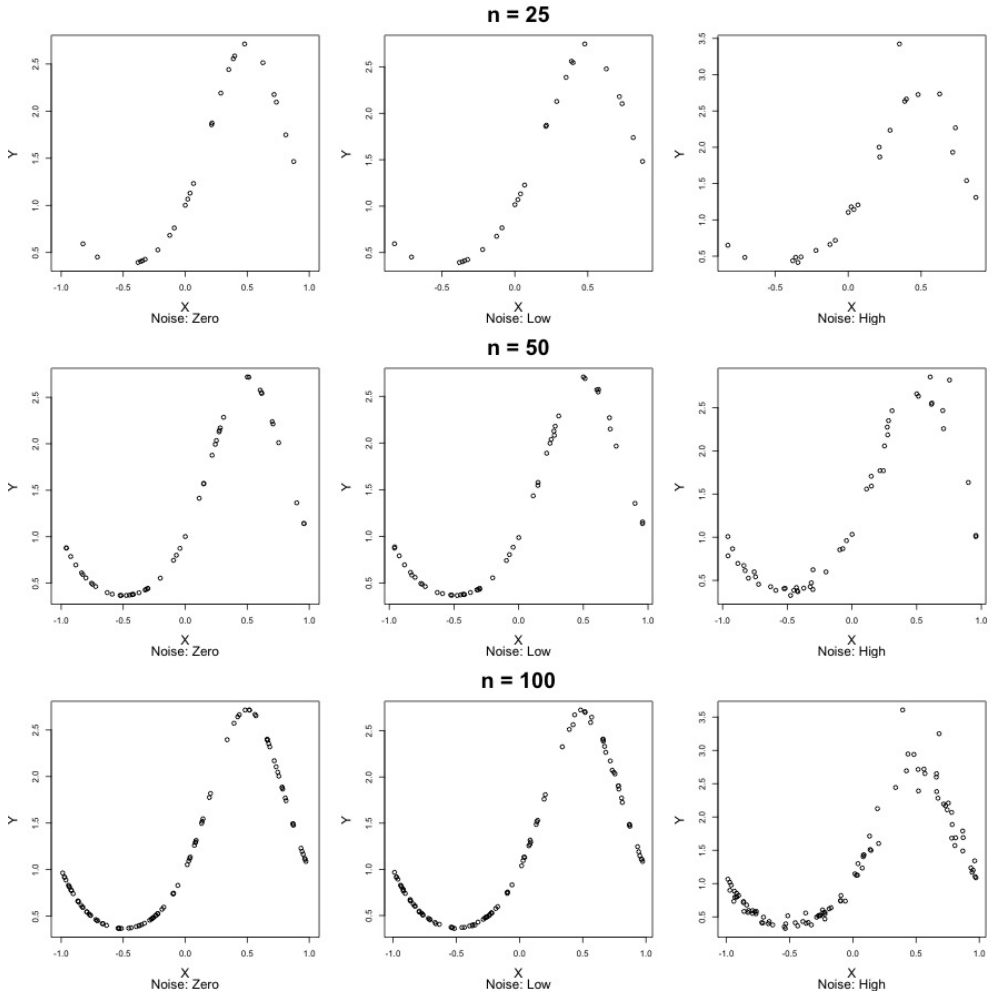

Figure 13: Plot of Simulation models, S_{12}: Exponent of sinusoid

Although figure 13 refers that X and Y share non-monotonic relation but ρ, ρ_s produced strong significant correlation. $\mathcal{R}$ and ξ_n correlations were significantly strong as well. τ, D, τ_b^* and $\xi_{n,M}$ methods had moderate correlations. But, $\xi_{n,M}$ increased as the sample size increased and M = 5 had better correlation than M = 10.

Table 14: Correlation estimate and p-value for simulation model, S_{12}: Exponent of sinusoid

Sample size	Noise level	Correlation estimate and p-value													
		ρ	ρ_s	τ	D	ρ^*		$\mathcal{R}$	τ_b^*	ξ_n		$\xi_{n,M}$			
						(X,Y)	(Y,X)			(X,Y)	(Y,X)	(X,Y)		(Y,X)	
												M=5	M=10	M=5	M=10
n = 25	Zero	0.873 (0.000)	0.747 (0.000)	0.480 (0.000)	0.432 (0.000)	0.993 (0.000)	0.993 (0.000)	0.892 (0.000)	0.501 (0.000)	0.784 (0.000)	0.534 (0.000)	0.485 (0.000)	0.417 (0.000)	0.494 (0.000)	0.419 (0.000)
	Low	0.872 (0.000)	0.749 (0.000)	0.050 (0.000)	0.416 (0.000)	0.993 (0.000)	0.993 (0.000)	0.892 (0.000)	0.490 (0.000)	0.755 (0.000)	0.423 (0.000)	0.487 (0.000)	0.422 (0.000)	0.503 (0.000)	0.427 (0.000)
	High	0.845 (0.000)	0.747 (0.000)	0.533 (0.000)	0.263 (0.000)	0.988 (0.000)	0.987 (0.000)	0.905 (0.000)	0.385 (0.000)	0.490 (0.000)	0.490 (0.000)	0.449 (0.000)	0.391 (0.000)	0.437 (0.000)	0.389 (0.000)
n = 50	Zero	0.820 (0.000)	0.758 (0.000)	0.535 (0.000)	0.391 (0.000)	0.998 (0.000)	0.995 (0.000)	0.860 (0.000)	0.454 (0.000)	0.885 (0.000)	0.586 (0.000)	0.634 (0.000)	0.475 (0.000)	0.513 (0.000)	0.453 (0.000)
	Low	0.819 (0.000)	0.758 (0.000)	0.531 (0.000)	0.372 (0.000)	0.998 (0.000)	0.998 (0.000)	0.861 (0.000)	0.438 (0.000)	0.873 (0.000)	0.631 (0.000)	0.631 (0.000)	0.472 (0.000)	0.512 (0.000)	0.451 (0.000)
	High	0.793 (0.000)	0.739 (0.000)	0.505 (0.000)	0.335 (0.000)	0.988 (0.000)	0.987 (0.000)	0.838 (0.000)	0.436 (0.000)	0.723 (0.000)	0.414 (0.000)	0.575 (0.000)	0.473 (0.000)	0.456 (0.000)	0.454 (0.000)
n = 100	Zero	0.788 (0.000)	0.747 (0.000)	0.505 (0.000)	0.480 (0.000)	1.000 (0.000)	1.000 (0.000)	0.877 (0.000)	0.534 (0.000)	0.941 (0.000)	0.587 (0.000)	0.805 (0.000)	0.682 (0.000)	0.587 (0.000)	0.576 (0.000)
	Low	0.788 (0.000)	0.746 (0.000)	0.501 (0.000)	0.466 (0.000)	1.000 (0.000)	1.000 (0.000)	0.877 (0.000)	0.535 (0.000)	0.926 (0.000)	0.596 (0.000)	0.801 (0.000)	0.680 (0.000)	0.586 (0.000)	0.571 (0.000)
	High	0.753 (0.000)	0.742 (0.000)	0.505 (0.000)	0.394 (0.000)	0.993 (0.000)	0.993 (0.000)	0.875 (0.000)	0.483 (0.000)	0.785 (0.000)	0.538 (0.000)	0.732 (0.000)	0.651 (0.000)	0.584 (0.000)	0.559 (0.000)

For all simulation models, maximal correlation ρ^* generated strong significant correlation between the transformed functions of variables. Increasing sample size and adding noise did not have any impact in increasing correlation.

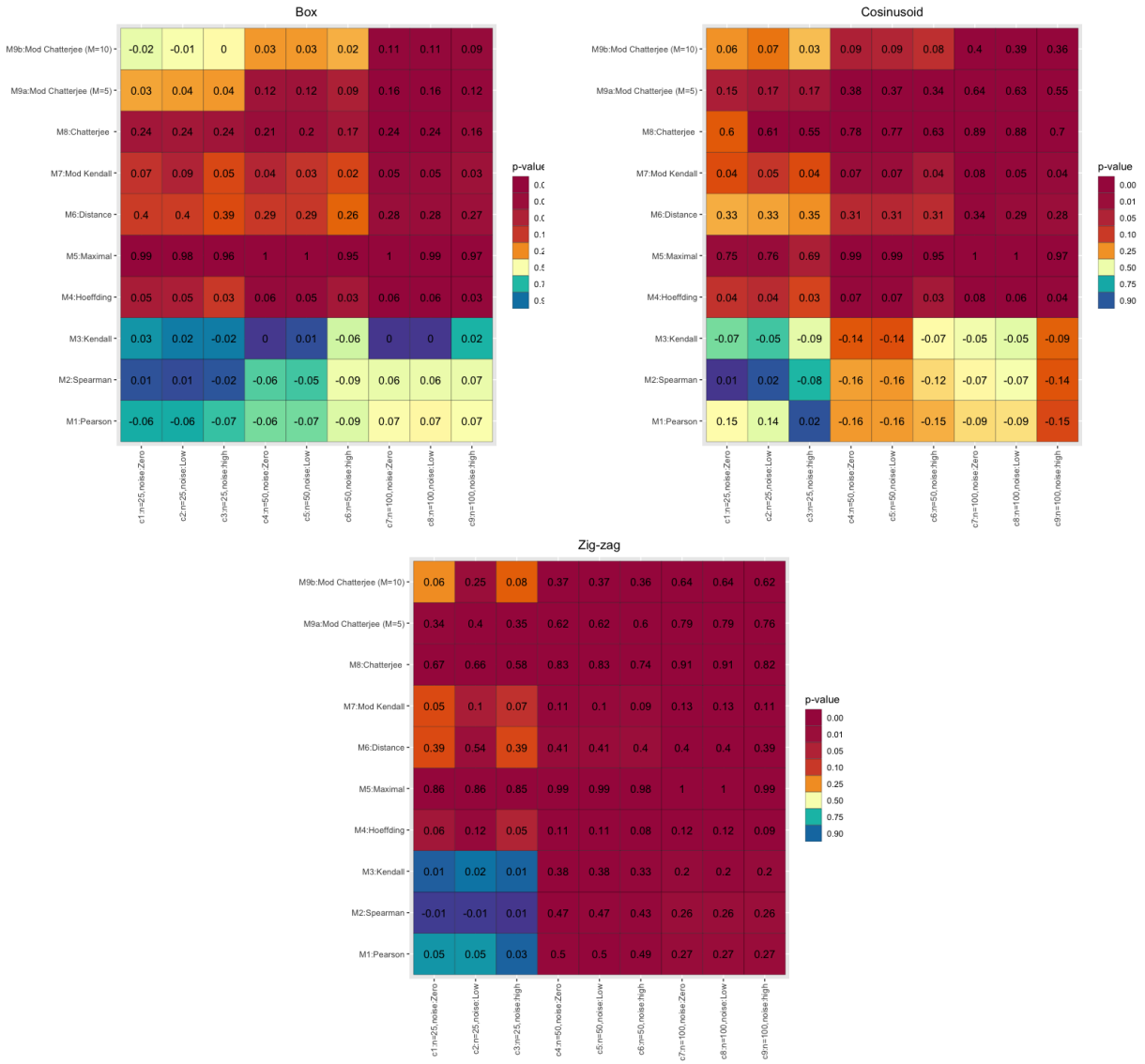

Figure 14: Plot of comparison of methods across three simulation models

Overall, different non-monotonic measure of associations are not consistent in capturing the correlation under different simulation models 14. Increasing sample size improves the performance of correlation measure 15. We found Chatterjee's rank-based correlation showing consistent pattern for most non-monotonic relationship. For this simulation study, We

observed that, while working with $\xi_{n,M}$, M = 5 gave higher correlation than M = 10. For most of the simulation models, ξ_n and ρ^* gave significantly high correlation as compared to other methods. Adding noise mostly decreased the correlation estimates although there were few exceptions.

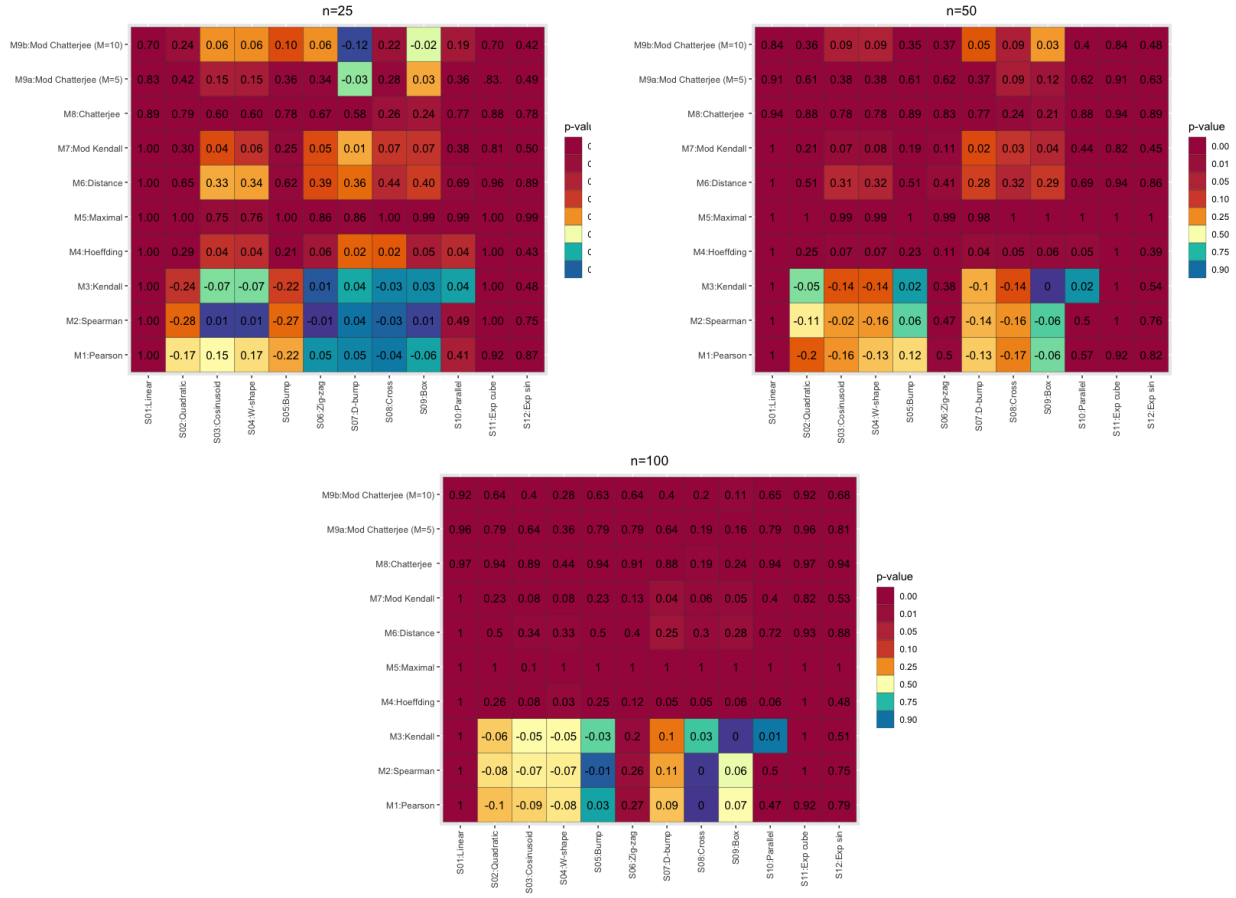

Figure 15: Plot of comparison of methods across three sample size

3.2.13 Time and Memory Used During Analysis

Now we report the computational aspects of these methods, in terms of run time and peak memory usage, for different sample sizes and simulation models. Here, the computational time is reported in seconds and peak RAM used is presented as Mebibytes(MiB). The values reported here are obtained by using the R-package *peakRAM* (Quinn, 2017) which returns the elapsed time of executing the method and the maximum amount of RAM allocated at

any point during that implementation.

Since maximal correlation is the only convergence based algorithm in this pool, the computation time of all other methods are not expected to significantly vary based on simulation model as long as sample size stays same. So, for each sample size, we took the average of run time and memory usage over all simulation models except maximal correlation.

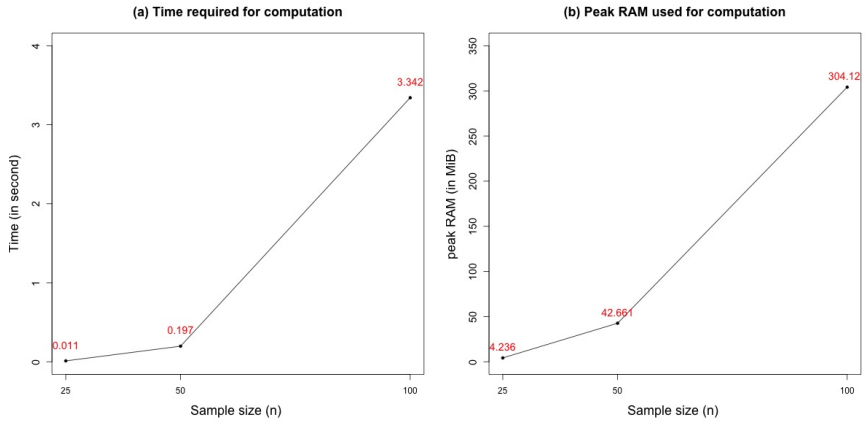

Figure 16: Plot of time and memory usage for τ_b^*

We start by reporting modified τ correlation,i.e., τ_b^*. Figure 16(a) shows the change in time required for computation of τ_b^* for different sample size. Computational time for n = 25 and n = 50 were less than 1 second. When the sample size increased to 100 then computation of τ_b^* required 3.342 seconds on average. Figure 16(b) shows how peak memory use varied by sample. The pattern is similar to time used for computation. With sample size 25, on an average τ_b^* computation required small amount of peak RAM (4.236 MiB). However, increasing the sample size to 50 increased peak RAM use by approximately 10 times (42.661 MiB). For 100 samples τ_b^* computation requires 304.125 MiB on average which is approximately 7 times the memory requirement for n = 50.

Next, we present the peak memory usage of distance correlation ($\mathcal{R}$) and Hoeffding's D correlation. In case of peak memory usage for computation of distance correlation, on average it took 0.106 MiB to 1.4 MiB as the sample size increased from 25 to 100. Figure 17(a) showed the change in peak RAM used during computation of distance correlation for

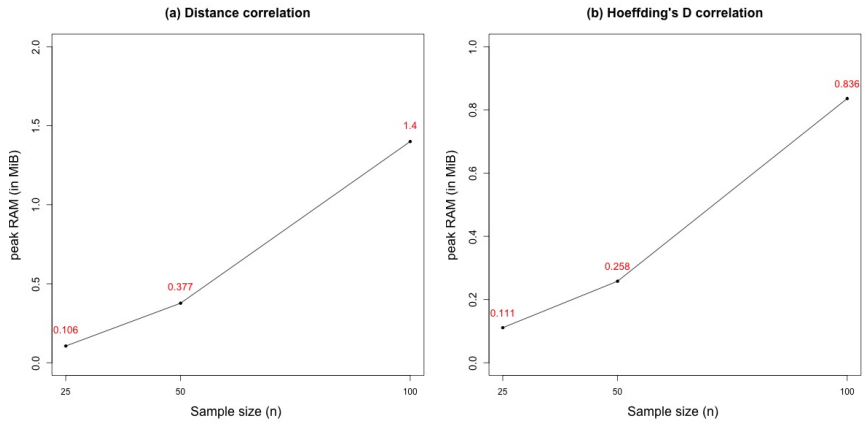

Figure 17: Plot of memory usage for $\mathcal{R}$ and D

three sample sizes. Likewise, computation of Hoeffding's D required peak RAM usage of 0.111 MiB for n = 25 and for n = 100 the peak RAM usage was 0.836 MiB. Figure 17(b) showed the changes occurred due to change in sample size.

For maximal correlation, one important quantity to observe is, the number of iterations required for convergence. However, in our analysis we did not find any specific pattern for the number of iterations with respect to sample size or noise level. For simulation model S_2 (Quadratic), when the sample size was 25, 40 iteration were required for convergence. However, as the sample size increased from 25 to 50 and 50 to 100, the number of iterations changed from 40 to 34 and 34 to 8 respectively. Again, for S_5 (Bump), for n = 50, it required 10 iteration to converge at zero noise level, which increased to 56 for low noise level and then dropped to 5 for high noise level. However, for n = 25 and n = 100, iteration number decreased for increased noise level. Furthermore, for S_{10} (Parallel), increasing sample size increased number of iteration for zero and low noise level and decreased for high noise level. In the contrary, for n = 25 and n = 50, iteration number decreased from zero to high noise level gradually. Whereas, for n = 100, iteration number increased and then sharply decreased for adding noise level. Another interesting feature is that, for the same dataset, switching the labels of X and Y can possibly lead to significantly different number of iterations. In case of S_8 (Cross), for n = 100 and zero noise level, it required 109 iterations for Y as a function of X and 61 iterations for X as a function of Y. Whereas for n = 25 and 50, it required less

itration number for Y as a function of X than X as a function of Y. Peak memory usage showed a very mild increase due to increase in sample size, going upto 18 MiB from 15 MiB when sample size changed from 25 to 100. We note that, these numbers implied that the maximal correlation required much higher peak memory than other methods except τ_b^*.

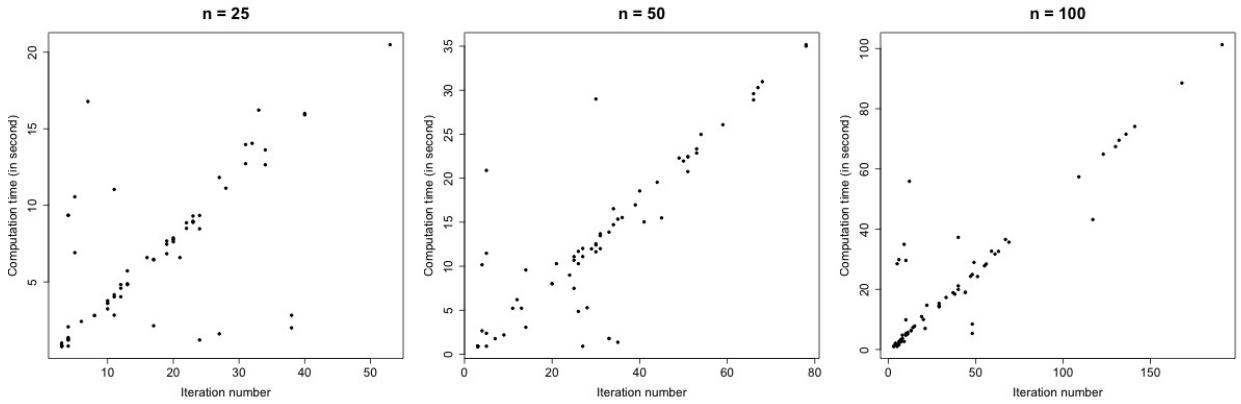

Figure 18: Scatter plot of computation time and iteration number for ρ^* by sample size

Figure 18 shows the relation of computational time and iteration number for convergence of maximal correlation algorithm for different samples. The computational time and iteration number required for convergence of maximal correlation varied by different simulation models. We observed that, (i) for fixed value of n,computational time showed a strong linear increasing pattern with respect to number of iterations and (ii) both computational time and iteration number increased due to increase in sample size. However, figure 18 also shows, for some instances of a longer computational time associated with smaller number of iterations or vice-versa. This is most likely due to the use of the algorithm to choose the best value of K, as described in section 2.1.2, that can take varying amount of time based on the observations.

All other methods required less than 0.1 MiB peak RAM usage for computation irrespective of sample size. In addition, except maximal correlation and τ_b^*, other methods required computational time less than one second. We did not find any pattern in time and memory usage due to change of sample size for these methods. Therefore, we did not report them in this thesis.

4 Application on Real Data

In previous section we checked the performance of all correlation measures on different simulation models. In this section we would focus on assessing these correlations on real data.

4.1 Description of Data

The real-world dataset that we considered here comes from Cape Floristic Region (CFR) of South Africa, a known hotspot for biodiversity research (Gelfand et al., 2006; Myers et al., 2000; Rebelo, 2002). The environmental and topographical characteristics across CFR are available at the South African Atlas of Hydrology and Climatology (Schulze, 1997) as GIS raster layers with a minimum pixel resolution of 1 minute latitude by 1 minute longitude.

For our analysis, we randomly selected 200 pixels in CFR and specifically considered four such characteristics: altitude (mean elevation within a pixel), roughness (range of elevation within a pixel), maximum summer temperature and minimum winter temperature. The units of altitude and roughness are Meters and the units of temperature values are Celsius. All four variables in this data had tied observations. So, we considered each pair separately and removed all tied observations from each variable within that pair. Hence, from the initial data, we had four subsets of data where each subset had two variables with no tied observations. Since, the number of tied observations in each subset were different for different variables, the number of observations in each subset after removing ties were different. The four pairs of variables exhibit diverse kind of pairwise linear association, as seen in figure 19,

- Moderately positive for altitude and roughness
- Strong negative for altitude and minimum winter temperature
- Weak negative for the summer maximum and winter minimum temperatures
- Negligible for altitude and summer maximum temperature

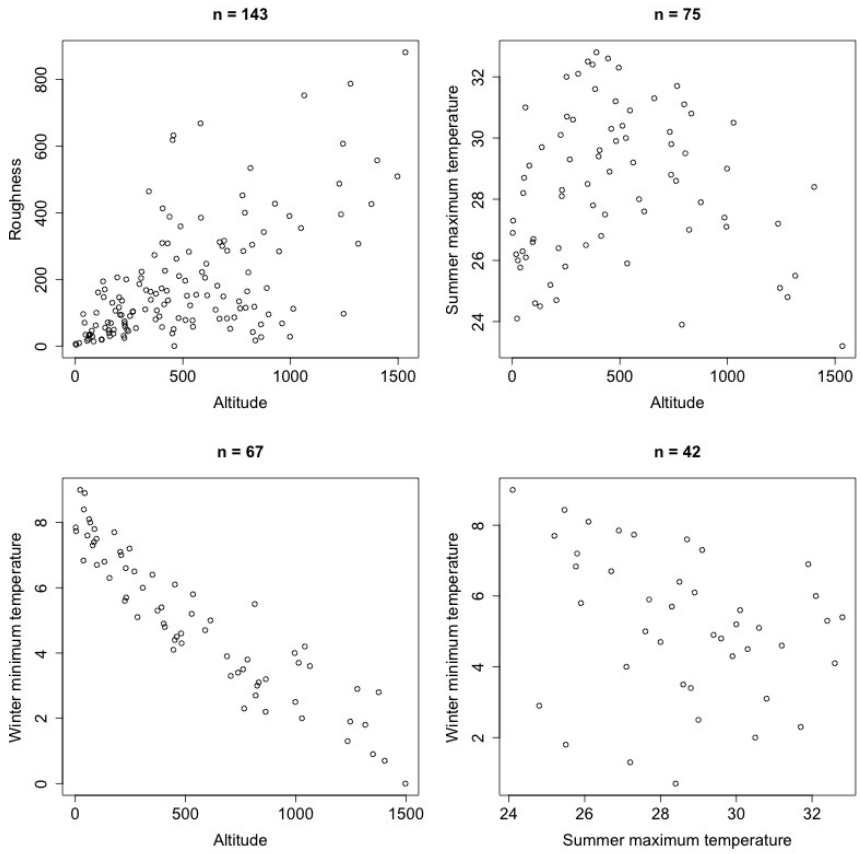

Figure 19: Scatterplots of selected variable pairs from CFR dataset

Our goal in this analysis was to understand to what extent the general measures of correlations, (a) retain presence of any significant linear association in a pair, (b) identify existence of non-linear or non-monotonic dependence and, (c) agree with each other with respect to strength of dependence within a pair.

4.2 Results

Here, we present the results obtained by application of all correlation measures on the CFR data. Similar to simulation analysis, we used same procedure to obtain correlation and permutation test for testing independence between variables.

Table 15 summarizes the correlation estimates and p-value for all four pairs of variables. From the table we observe that Pearson correlation ρ retained significant strong linear correlation for altitude and roughness, moderate linear association for altitude and winter min-

Table 15: Correlation estimate and p-value for CFR data

Methods			Correlation estimate and p-value			
			(A[a],R[b])	(A,S[c])	(A,W[d])	(S,W)
			n=143	n=75	n=67	n=42
ρ			0.596 (0.000)	-0.031 (0.794)	-0.927 (0.000)	-0.291 (0.062)
ρ_s			0.557 (0.000)	0.151 (0.193)	-0.942 (0.000)	-0.317 (0.044)
τ			0.409 (0.000)	0.093 (0.238)	-0.794 (0.000)	-0.233 (0.032)
D			0.112 (0.000)	0.024 (0.012)	0.541 (0.000)	0.039 (0.019)
ρ^*	(X,Y)		0.685 (0.001)	0.594 (0.030)	0.916 (0.000)	0.671 (0.033)
	(Y,X)		0.684 (0.000)	0.579 (0.055)	0.915 (0.000)	0.674 (0.029)
$\mathcal{R}$			0.547 (0.000)	0.325 (0.008)	0.922 (0.000)	0.413 (0.015)
τ_b^*			0.176 (0.000)	0.059 (0.003)	0.635 (0.000)	0.083 (0.008)
ξ_n	(X,Y)		0.297 (0.000)	0.184 (0.005)	0.665 (0.000)	-0.111 (0.877)
	(Y,X)		0.257 (0.000)	0.020 (0.389)	0.654 (0.000)	0.166 (0.038)
$\xi_{n,M}$	(X,Y)	M=5	0.277 (0.000)	0.139 (0.001)	0.492 (0.000)	0.003 (0.435)
		M=10	0.283 (0.000)	0.114 (0.006)	0.339 (0.000)	-0.027 (0.656)
	(Y,X)	M=5	0.227 (0.000)	0.030 (0.214)	0.481 (0.000)	0.030 (0.270)
		M=10	0.255 (0.000)	0.027 (0.216)	0.335 (0.000)	-0.004 (0.483)

[a]A: Altitude of terrain
[b]R: Roughness of terrain
[c]S: Summer maximum temperature
[d]W: Winter minimum temperature

imum temperature and marginally significant weak association for summer maximum and winter minimum temperatures.

In capturing association between different pairs of variables that share diverse type of

linear association, ρ_s and $\mathcal{R}$ were almost consistent with ρ when the variables share strong, moderate or weak linear association. While the correlation between altitude and summer maximum temperature was negligible and insignificant for ρ, ρ_s and τ; $\mathcal{R}$ showed significant weak association and Hoeffding's D and τ_b^* produced significant negligible association between the pair.

As opposed to simulation study, where maximal correlation ρ^* always obtained very strong correlation between transformed variables, here, we see it managed to obtain strong correlation only when the variables themselves share strong linear association. In all other cases maximal correlation generated moderate correlation.

In case of CFR data, performing time and memory analysis is not that meaningful since we worked with a single sample size and did not add any noise level. So, here we did not present computational aspects of CFR data analysis.

5 Conclusion

5.1 Discussion

In this thesis, we explored six non-monotonic correlation measures focusing on different rank based measures, maximal correlation and distance based correlation for diverse type of relations between variables. In addition, we explored three classical approaches of correlation measure, Pearson correlation, Spearman's rank correlation and Kendall's τ.

We observed that, Pearson correlation, Spearman's rank correlation and Kendall's τ showed significant association when the relation between variables were linear and/or monotonic. In case of non-monotonic pattern, they produced insignificant estimates.

All methods except modified τ and maximal correlation require negligible computational time and peak RAM usage. Both computational time and peak RAM usage had cubic increase as the sample size increases for modified τ. Out of all methods, modified τ occupied most peak RAM usage.

Although maximal correlation captured strong significant correlation between variables for all simulation models. We should keep in mind that, maximal correlation is the association between transformed functions of original variables. So, it should be used when there is scope to consider functions of actual variables rather than the variables themselves. We found that, maximal correlation for Y as a function of X and X as a function of Y were approximately same but their iteration number for convergence and computational time varied. Furthermore, due to the issue of convergence, computational time and peak RAM usage of maximal correlation is high.

Chatterjee's rank based correlation (ξ_n) produced significant correlation for most simulation models exhibiting non-monotone association. In contrast, it did not show similar performance for CFR data. The modified Chatterjee's correlation ($\xi_{n,M}$) varied by the choice of M. We observed, larger correlation was obtained for smaller choice of M. Moreover, both (ξ_n) and ($\xi_{n,M}$) methods produced different correlation estimate based on the direction

of relationship between variables. In such case, one can compute both correlation for Y as a function of X and X as a function of Y and consider maximum between the two estimates. This procedure was followed by Lin and Han (2023) for testing independence.

To summarize, different non-monotonic measure of associations were not consistent in capturing correlation for different type of relationship pattern between variables. Chatterjee's rank based correlation and maximal correlation captured strong association in most of the cases. The fact that for some simulation models, all methods had very low correlation estimates indicate that there was a random relationship between the variables. In such case, it is understood that there is no linear, monotone or non-monotone association.

5.2 Extension

There were few points that were beyond the scope of this work. First, we assumed there were no ties in data. In reality, presence of ties in data is quite frequent. Using rank based correlation measures would create problem as observations with same values would have same rank. In that case, applying such methods require separate treatment. Some methods like Spearman's rank correlation and Hoeffding's D measure suggested to rearrange the data at first. Then use average of the rank for observations with same rank (Daniel, 1990; Hoeffding, 1994). Kendall's τ and Chatterjee's rank based correlation proposed separate formula to handle ties (Chatterjee, 2021; Daniel, 1990). One can compare the performances of these measures in case of tied observations. Second, we worked with one dimensional variables only. Some methods, such as, maximal correlation, distance correlation proposed ways to handle multi-dimensional variables (Breiman and Friedman, 1985; Székely et al., 2007). This could be another way of exploring that what happens when the variables are multi-dimensional. Furthermore, we assumed that both variables are numeric. Breiman and Friedman (1985) developed algorithm in case of categorical data. There could be other methods in literature for assessing relation between categorical or ordinal variables. Those could be studied as an extension of this work. Finally, to ensure coherency in all meth-

ods, we limited our computation to permutation tests for testing independence. However, permutation tests are not efficient as they are time consuming for large number of samples (Chatterjee, 2021; Christensen and Zabriskie, 2022). There were alternative tests suggested by some of the proposed methods (Chatterjee, 2021; Lin and Han, 2023). In addition, many other alternative independence tests were also proposed which we did not cover in this thesis.

5.3 Scope of Further Study

So far we studied nine correlation measures that capture linear, monotone and/or non-monotone association. Over time many proposals were introduced for this purpose. There could be other methods which proposed more efficient tests or association measures to capture any type of association between variables. To name a few, Blum et al. (1961) suggested a modification of Hoeffding's D metric. Heller et al. (2013) introduced a test of dependence based on pairwise distances between random vectors of any dimension. Pfister et al. (2018) proposed a Kernel based test to identify joint independence. Wang et al. (2017) developed G-squared statistic for multi-dimensional joint distributions which is identical to the square of the Pearson correlation coefficient, R-squared. Kraskov et al. (2004) proposed two classes of improved estimators for mutual information based on entropy estimates from k-nearest neighbour distances. Reshef et al. (2011) defined the maximal information coefficient (MIC) which belongs to a larger class of maximal information-based non-parametric exploration (MINE) statistics for identifying and classifying relationships.

A R-code

A.1 Functions of correlation and permutation test

```
library(ggplot2)
library(lattice)
library(caret)
```

```
# Method 1: Pearson's product-moment correlation #
```

```
rho <- function (y,x) {
  rho = cov(x,y)/(sd(x)*sd(y))
  return(rho)
}
pearsoncor.test <- function (x,y) {
oboutput = rho(y,x)
y_permute = replicate(MC, sample(y))
output = apply(y_permute,2,rho,x=x)
MCpvalue <- (sum(ifelse(abs(output)>abs(oboutput),1,0)))/MC
return (c(cor=oboutput, pvalue=MCpvalue))
}
```

```
# Method 2: Spearman's rank correlation #
```

```
srho <- function (y,x) {
  rank.x <- rank(x)
  rank.y <- rank(y)
  d <- rank.x - rank.y
```

```
  srho = 1 - ((6* sum(d^2))/(n*(n^2-1)))
  return(srho)
}
spearman.test <- function(x,y) {
  oboutput = srho(y,x)
  y_permute = replicate(MC, sample(y))
  output = apply(y_permute,2,srho,x=x)
  MCpvalue <- (sum(ifelse(abs(output)>abs(oboutput),1,0)))/MC
  return(c(cor=oboutput, pvalue=MCpvalue))
}
```

```
# Method 3: Kendall's tau #
```

```
kend <- function(y,x) {
  a <- sign(outer(x,x,'-'))
  b <- sign(outer(y,y,'-'))
  kend <- sum(a*b)/(n*(n-1))
  return(kend)
}
kendall.test <- function(x,y) {
  oboutput = kend(y,x)
  y_permute = replicate(MC, sample(y))
  output = apply(y_permute,2,kend,x=x)
  MCpvalue <- (sum(ifelse(abs(output)>abs(oboutput),1,0)))/MC
  return(c(cor=oboutput, pvalue=MCpvalue))
}
```

```
# Method 4: Hoeffding's D #
```

```
hoeffding <- function(y,x) {
 a = sign(outer(x,x,'-'))
 b = sign(outer(y,y,'-'))
 sum_sign = a+b
 c = apply(sum_sign,1,function(x) {length(which(x==2))})
 rx <- rank(x)
 ry <- rank(y)
 A = sum((rx-1)*(rx-2)*(ry-1)*(ry-2))
 B = sum((rx-2)*(ry-2)*c)
 C = sum(c*(c-1))
 D = (30*(A-(2*(n-2)*B)+((n-2)*(n-3)*C)))/
      (n*(n-1)*(n-2)*(n-3)*(n-4))
 return(D)
}
hoeffding.test <- function(x,y) {
  oboutput = hoeffding(y,x)
  y_permute = replicate(MC, sample(y))
  output = apply(y_permute,2,hoeffding,x=x)
  MCpvalue <- (sum(ifelse(output>oboutput,1,0)))/MC
  return(c(cor=oboutput, pvalue=MCpvalue))
}
```

```
# Method 5: Maximal correlation #
```

```
# Smooth function #
smooth<-function(x,y) {
  x1<-sort(x)
  y1<-y[order(x)]
  TrainData <- matrix(x1,ncol=1)
  colnames(TrainData)<-"x1"
  TrainClasses <- matrix(y1,ncol=1)
  colnames(TrainClasses)<-"y1"
  traindat<-data.frame(x1=x1,y1=y1)
  lmFit <- train(y1 ~ . + x1,data=traindat,
                 method = "knn",
                 preProcess = c("center", "scale"),
                 trControl = trainControl(method = "cv"))
  neighbor<-lmFit$bestTune
  #neighbor <- K
  out<-array(0,length(x))
  for (z in 1:length(x)) {
    z1<-rank(x)[z]
    pos<- max(z1-neighbor,1):min(z1+neighbor,n) #K=20
    xbar<-mean(x1[pos])
    ybar<-mean(y1[pos])
    beta<-cov(x1[pos],y1[pos])/var(x1[pos])
    alpha<-ybar-beta*xbar
    smooth<-alpha+beta*x1[z1]
    out[z]<-smooth
  }
  return(out)
```

```
}
# Norm function #
norm_y<-function(a){
  sqrt(mean(a^2))
}
# Coefficient function #
rhostar <- function (y,x) {
  phi<- x
  theta<- y/norm_y(y)
  error<-c()
  error[1]<-mean((theta-phi)^2)
  phix<-x
  phiy<-theta
  phi<-smooth(phix,phiy)
  thetax<-y
  thetay<-phi
  thetanum<-smooth(thetax,thetay)
  theta<-thetanum/norm_y(thetanum)
  error[2]<-mean((theta-phi)^2)
  count = 2
  while((error[count-1] - error[count]) > 10^-6){
    count = count + 1
    phix<-x
    phiy<-theta
    phi<-smooth(phix,phiy)
    thetax<-y
    thetay<-phi
```

```
    thetanum<-smooth(thetax,thetay)
    theta<-thetanum/norm_y(thetanum)
    error[count]<-mean((theta-phi)^2)
  }
  out <-cor(phi,theta)
  return(c(out,count))
}
rhostar.test <- function(x,y) {
oboutput = rhostar(y,x)
y_permute = replicate(MC, sample(y))
output = apply(y_permute,2,rhostar,x=x)
MCpvalue <- (sum(ifelse(output[1,]>oboutput[1],1,0)))/MC
return (c(cor=oboutput[1], obcount=oboutput[2],
pvalue=MCpvalue, count=output[2,]))
}
```

```
# Method 6: Modified tau #
```

```
# when number of combinations is smaller than number of permutations
signfun <- function (index,z) {
  az <- sign(abs(z[index[1]]-z[index[2]])+abs(z[index[3]]-z[index[4]])
  -abs(z[index[1]]-z[index[3]])-abs(z[index[2]]-z[index[4]]))
  return (az)
}
# when number of combinations is larger than number of permutations
signfun_comb_vec <- function (z,comb)
{
```

```
  az <- sign(abs(z[comb[1,]]-z[comb[2,]])+abs(z[comb[3,]]-z[comb[4,]])
  -abs(z[comb[1,]]-z[comb[3,]])-abs(z[comb[2,]]-z[comb[4,]]))
  return (az)
}
signfun_comb_mat <- function (z,comb)
{
  taustar.z <- sign(abs(z[comb[1,]]-z[comb[2,]])
  +abs(z[comb[3,]]-z[comb[4,]])-abs(z[comb[1,]]
  -z[comb[3,]])-abs(z[comb[2,]]-z[comb[4,]]))
  out = sum(taustar.z*taustar.x)/(n^4)
  rm(taustar.z)
  return (out)
}
taustar<- function(x,y) {
  comb <- combn(n,4)
  taustar.x <- signfun_comb_vec(z=x,comb=comb)
  taustar.y <- signfun_comb_vec(z=y,comb=comb)
  obtaustar <- (1/n^4)*sum(taustar.y*taustar.x)
  obtaustar.x <- (1/n^4)*sum(taustar.x*taustar.x)
  obtaustar.y <- (1/n^4)*sum(taustar.y*taustar.y)
  taub <- obtaustar/sqrt(obtaustar.x*obtaustar.y)
  return(cor=taub)
}
#when number of combinations is smaller than number of permutations
taustar.test1 <- function(y,x) {
 comb <- combn(n,4)
 taustar.x <- apply(comb,2,signfun,z=x)
```

```
 taustar.y <- apply(comb,2,signfun,z=y)
 obtaustar <- (1/n^4)*sum(taustar.y*taustar.x)
 obtaustar.x <- (1/n^4)*sum(taustar.x*taustar.x)
 obtaustar.y <- (1/n^4)*sum(taustar.y*taustar.y)
 taub <- obtaustar/sqrt(obtaustar.x*obtaustar.y)
 y_permute = replicate(MC, sample(y))
 MCtaustar = apply(y_permute,2,signfun,comb=comb)
 MCpvalue <- (sum(ifelse(MCtaustar>obtaustar,1,0)))/MC
return(c(cor=taub, pvalue=MCpvalue))
}
# when number of combinations is larger than number of permutations
taustar.test2 <- function(x,y) {
 comb <- combn(n,4)
 taustar.x <- signfun_comb_vec(z=x,comb=comb)
 taustar.y <- signfun_comb_vec(z=y,comb=comb)
 obtaustar <- (1/n^4)*sum(taustar.y*taustar.x)
 obtaustar.x <- (1/n^4)*sum(taustar.x*taustar.x)
 obtaustar.y <- (1/n^4)*sum(taustar.y*taustar.y)
 taub <- obtaustar/sqrt(obtaustar.x*obtaustar.y)
 y_permute = replicate(MC, sample(y))
 MCtaustar = apply(y_permute,2,signfun_comb_mat,comb=comb)
 MCpvalue <- (sum(ifelse(MCtaustar>obtaustar,1,0)))/MC
 return(c(cor=taub, pvalue=MCpvalue))
}
```

```
# Method 7: Distance correlation #
```

```
dist_corr <- function(y,x) {
  a = as.matrix(dist(x))
  b = as.matrix(dist(y))
  A = a - matrix(rowMeans(a),nrow=nrow(a),ncol=ncol(a),byrow=F)
  -matrix(colMeans(a),nrow=nrow(a),ncol=ncol(a),byrow=T) + mean(a)
  B = b - matrix(rowMeans(b),nrow=nrow(b),ncol=ncol(b),byrow=F)
  -matrix(colMeans(b),nrow=nrow(b),ncol=ncol(b),byrow=T) + mean(b)
  n = nrow(a)
  dist_cov = sqrt(sum(A*B)/(n^2))
  dist_var_x = sqrt(sum(A*A)/(n^2))
  dist_var_y = sqrt(sum(B*B)/(n^2))
  DistCorr = dist_cov/sqrt(dist_var_x*dist_var_y)
  return(DistCorr)
}
distance_cor.test <- function (x,y) {
  oboutput = dist_corr(y,x)
  y_permute = replicate(MC, sample(y))
  output = apply(y_permute,2,dist_corr,x=x)
  MCpvalue <- (sum(ifelse(output>oboutput,1,0)))/MC
  return (c(cor=oboutput, pvalue=MCpvalue))
}
```

 # Method 8: Chatterjee's Correlation #

```
xi_cor <- function(y,x) {
  n = length(x)
  datamat <- cbind(x,y)
```

```
  datamat_order_by_x <- datamat[order(x),]
  r <- rank(datamat_order_by_x[,2], ties.method = "max")
  rankdiffsum <- sum(abs(diff(r)))
  z <- 1-((3*rankdiffsum)/(n^2-1))
  return(z)
}
xicor.test <- function (x,y) {
  oboutput = xi_cor(y,x)
  y_permute = replicate(MC, sample(y))
  output = apply(y_permute,2,xi_cor,x=x)
  MCpvalue <- (sum(ifelse(output>oboutput,1,0)))/MC
  return (c(cor=oboutput, pvalue=MCpvalue))
}
```

 # Method 9: Improved Chatterjee's Correlation #

```
lhcor <- function(y,x) {
  dat <-data.frame(x,y)
  dat = dat[order(x),]
  yrank <-rank(dat[,2])
  n <- length(y)
  out = 0
  for (j in 1:M) {
    out = out+sum(pmin(yrank[1:(n-j)],yrank[(j+1):n]))
    + sum(yrank[(n-j+1):n])
  }
  lhcor = -2 + ((6*out)/((n+1)*((n*M)+(M*(M+1))/4)))
```

```
  return(lhcor)
}
lhcor.test <- function (x,y) {
  oboutput = lhcor(y,x)
  y_permute = replicate(MC, sample(y))
  output = apply(y_permute,2,lhcor,x=x)
  MCpvalue <- (sum(ifelse(output>oboutput,1,0)))/MC
  return (c(cor=oboutput, pvalue=MCpvalue))
}
```

A.2 Output extraction

```
library(peakRAM)
# Define X and Y
n = length(x) # Sample size
MC =    # Mention number of permutation here
PearCor = pearsoncor.test(x,y)
PearCor_TM = peakRAM(rho(y,x))
SpearCor = spearman.test(x,y)
SpearCor_TM = peakRAM(srho(y,x))
KendCor = kendall.test(x,y)
KendCor_TM = peakRAM(kend(y,x))
HoeffdingCor = hoeffding.test(x,y)
HoeffdingCor_TM = peakRAM(hoeffding(y,x))
RhostarCor = rhostar.test(x,y)
RhostarCorR = rhostar.test(y,x)
RhostarCor_TM = peakRAM(rhostar(y,x))
RhostarCorR_TM = peakRAM(rhostar(x,y))
```

```
TaustarCor = taustar.test2(x,y) #Pick function based on
                                #combination and permutation number
TaustarCor_TM = peakRAM(taustar(x,y))
DistCor = distance_cor.test(x,y)
DistCor_TM = peakRAM(dist_corr(x,y))
XiCor = xicor.test(x,y)
XiCor_TM = peakRAM(zai_wo_tie(x,y))
XiCorR = xicor.test(y,x)
XiCorR_TM = peakRAM(zai_wo_tie(y,x))
M =  #Mention M here
LHCor = lhcor.test(x,y)
LHCor_TM = peakRAM(lhcor(y,x))
LHCor_R = lhcor.test(y,x)
LHCor_R_TM = peakRAM(lhcor(x,y))
```

References

Abdi, H. (2007), *Kendall rank correlation*, Thousand Oaks (CA): Sage.

Auddy, A., Deb, N., and Nandy, S. (2021), "Exact Detection Thresholds for Chatterjee's Correlation," *arXiv preprint arXiv:2104.15140*.

Bergsma, W. and Dassios, A. (2014), "A consistent test of independence based on a sign covariance related to Kendall's tau," .

Blum, J. R., Kiefer, J., and Rosenblatt, M. (1961), *Distribution free tests of independence based on the sample distribution function*, Sandia Corporation.

Breiman, L. and Friedman, J. H. (1985), "Estimating optimal transformations for multiple regression and correlation," *Journal of the American statistical Association*, 80, 580–598.

Chatterjee, S. (2021), "A new coefficient of correlation," *Journal of the American Statistical Association*, 116, 2009–2022.

Christensen, W. F. and Zabriskie, B. N. (2022), "When your permutation test is doomed to fail," *The American Statistician*, 76, 53–63.

Daniel, W. W. (1990), *Applied nonparametric statistics, PWS-Kent*, Inc., Boston, Massachusetts.

DiCiccio, C. J. and Romano, J. P. (2017), "Robust permutation tests for correlation and regression coefficients," *Journal of the American Statistical Association*, 112, 1211–1220.

Gelfand, A. E., Holder, M., Latimer, A., Lewis, P. O., Rebelo, A. G., Silander Jr, J. A., and Wu, S. (2006), "Explaining species distribution patterns through hierarchical modeling," .

Harrell Jr, F. E. (2023), *Hmisc: Harrell Miscellaneous*, R package version 5.0-1.

Haug, M. G. (2023), *measure of association*, Encyclopedia Britannica.

Heller, R., Heller, Y., and Gorfine, M. (2013), "A consistent multivariate test of association based on ranks of distances," *Biometrika*, 100, 503–510.

Hoeffding, W. (1994), "A non-parametric test of independence," *The Collected Works of Wassily Hoeffding*, pp. 214–226.

Hollander, M., Wolfe, D. A., and Chicken, E. (2013), *Nonparametric statistical methods*, John Wiley & Sons.

Huang, G.-H. (2010), "Measure of Association," in *International Encyclopedia of Education (Third Edition)*, eds. P. Peterson, E. Baker, and B. McGaw, pp. 260–263, Elsevier, third edition edn.

James, B. and Zminda, D. (1988), *Bill James Presents the great american baseball stat book*, Villard Books.

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004), "Estimating mutual information," *Physical review E*, 69, 066138.

Kuhn and Max (2008), "Building Predictive Models in R Using the caret Package," *Journal of Statistical Software*, 28, 1–26.

Lin, Z. and Han, F. (2023), "On boosting the power of Chatterjee's rank correlation," *Biometrika*, 110, 283–299.

Ludbrook, J. (1994), "Advantages of permutation (randomization) tests in clinical and experimental pharmacology and physiology," *Clinical and experimental pharmacology and physiology*, 21, 673–686.

Merlo, J. and Lynch, K. (2010), *Encyclopedia of Research Design*, Sage.

Myers, J. L. and Well, A. D. (2003), *Research Design Statistical Analysis*, Psychology Press, 2nd edition edn.

Myers, N., Mittermeier, R. A., Mittermeier, C. G., Da Fonseca, G. A., and Kent, J. (2000), "Biodiversity hotspots for conservation priorities," *Nature*, 403, 853–858.

Nelsen, R. B. (1999), *An introduction to copulas. Properties and applications*, vol. 139 of *Lect. Notes Stat.*, New York, NY: Springer.

Pfister, N., Bühlmann, P., Schölkopf, B., and Peters, J. (2018), "Kernel-based tests for joint independence," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80, 5–31.

Quinn, T. (2017), *peakRAM: Monitor the Total and Peak RAM Used by an Expression or Function*, R package version 1.0.2.

Rebelo, A. (2002), "The state of plants in the Cape Flora," in *Proceedings of a Conference Held at the Rosebank Hotel in Johannesburg (GH Verdoorn and J. Le Roux, eds.)*, vol. 18.

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, 1518–1524.

Rodgers, J. L. and Nicewander, W. A. (1988), "Thirteen ways to look at the correlation coefficient," *American statistician*, pp. 59–66.

Samuel, M. and Okey, L. E. (2015), "The relevance and significance of correlation in social science research," *International Journal of Sociology and Anthropology Research*, 1, 22–28.

Schulze, R. (1997), "South African Atlas of Agrohydrology and Climatology (Water Research Commission Report TT82/96)," *Pretoria, South Africa*.

Shi, H., Drton, M., and Han, F. (2022), "On the power of Chatterjee's rank correlation," *Biometrika*, 109, 317–333.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007), "Measuring and testing dependence by correlation of distances," *The Annals of Statistics*, 35, 2769–2794.

Wang, X., Jiang, B., and Liu, J. S. (2017), "Generalized R-squared for detecting dependence," *Biometrika*, 104, 129–139.

Wilding, G. E. and Mudholkar, G. S. (2008), "Empirical approximations for Hoeffding's test of bivariate independence using two Weibull extensions," *Statistical Methodology*, 5, 160–170.