A Comparative Simulation Study on the Performance of Nonmonotonic Tests in Psychology

Perez Alonso, Andres

Faculteit der Sociale Wetenschappen

A Comparative Simulation Study on the Performance of Nonmonotonic Tests in Psychology

Andres Felipe Perez Alonso

Master's Thesis Psychology,

Methodology and Statistics Unit, Institute of Psychology

Faculty of Social and Behavioral Sciences, Leiden University

Date: April 2021

Student number: 2556715

Supervisor: Dr. Julian D. Karch

Abstract

The identification of the dependence between variables is a common task in Psychology. The most common approaches to this task are Pearson's, Spearman's, and Kendall correlation coefficients. A clear limitation of those coefficients is that they can only identify linear and monotonic relationships. In recent years, several methods to identify nonmonotonic associations have been developed. Nevertheless, there is not a clear answer to which method should be used when facing unknown conditions in research as is often the case in Psychology. In this study, we aimed to identify which dependence test performs the best under conditions that can be found in psychological research. **Method:** A simulation was performed to compare nine dependence tests through hypothesis testing. The conditions assessed were sample size, type of relationship, and noise. Three approaches were employed to summarize the statistical power and analyse the results: complete class, average power, and Maximin. **Results:** There was not a uniformly most powerful test across all conditions. However, several nonmonotonic tests presented a good performance in terms of power for most conditions. Moreover, Mutual Information (MI) estimated through Kernel Density Estimation with the Epanechnikov kernel and the Sheater-Jones plugin bandwidth outperformed all other methods in terms of the analysis approaches of this study. **Conclusion:** For the evaluated conditions we recommend the use of MI estimated through the defined settings. Nevertheless, other modern tests should not be immediately discarded as their difference in performance with MI is small and could be due to the design of this specific simulation.

Keywords: dependence tests, correlation, nonmonotonic associations, Monte-Carlo simulation, hypothesis testing.

Table of Content

Introduction

Finding and describing the dependence between two variables is a recurrent task in Psychology. The most well-known methods to do this include Pearson's product-moment correlation (Galton, 1889; Pearson, 1920), Kendall rank correlation coefficient (Kendall, 1938), and Spearman's rank correlation coefficient (Spearman, 1904). The first one can identify linear dependence between variables while the latter two can also identify nonlinear monotonic relationships.

Nevertheless, variable associations in Psychology are not limited only to linear monotonic relationships. In general, nonlinear dependences have been described in several areas of Psychology, such as Cognitive Science, Social Psychology, Organizational Psychology, and Clinical Psychology (Guastello, 2001; Guastello et al., 2009). More specifically, various nonlinear and nonmonotonic types of associations between two variables have been identified. For instance, quadratic dependences, resembling an inverted U, have been found when studying the relationship between stress and performance (Westman & Eden, 1996). Exponential relationships have been suggested in learning research where the learning rate starts improving rapidly and then flattens after a certain amount of trials (Leibowitz et al., 2010). Moreover, human behaviors over time often present cyclic patterns that can be captured by a sine or cosine term in longitudinal psychological data. (Verboon & Leontjevas, 2018). Furthermore, the association between two variables can be affected by the presence of a moderator variable. This interaction effect can lead to cross associations that resemble simultaneous negative and positive linear relationships in the shape of an 'X' (Corning, 2002; Frazier et al., 2004). Therefore, the Pearson's, Kendall's, and Spearman's correlation coefficients are often not adequate to be used in psychological research.

Nonlinearity has been studied in statistics in different ways. Diagnosing nonlinearity and approaching it with nonparametric regression has been one of the first ways to address this problem (Andersen, 2009; Fox, 2016). Moreover, James et al. (2013) describe some of the methods to account for nonlinearities from the statistical learning point of view: polynomial regression, step functions, regression and smoothing splines, local regression, and generalized additive models are some of the techniques used to approach nonlinearity when aiming to predict or explain the association between two variables. However, most of these methods require prior knowledge from the researcher

about the relationship between the variables and are used to model specific and desired associations in the data.

Considering that not all investigation in Psychology desires to confirm or model a specific type of relationship, but that in some cases the exploration of the data and identification of associations through hypothesis testing are the aims of the research, dependence measures similar to Pearson's or Kendall coefficients gain relevance. In order to assess dependences in a broader sense, and not only addressing linear and monotonic associations, different methods have been created in the past decades. One of the first approaches to solve this problem was Hoeffding's D (Hoeffding, 1948), but in recent years, a large number of alternatives have appeared. Some of the most well documented of such alternatives are Distance Correlation (dCor; Szekely et al., 2007), Heller–Heller–Gorfine measure (HHG; Heller et al., 2012), Mutual Information (MI; Cover & Thomas, 2006), Hilbert-Schmidt Independence Criterion (HSIC; Gretton et al., 2005), and the Maximal Information Coefficient (MIC; Reshef et al., 2011). All the previously mentioned measures test the hypothesis of statistical independence between variables, that is, whether there is a relationship between them regardless of the type of association they have. Formally, there is statistical independence when the distribution of one variable is the same regardless of the values and distribution of the second variable (H_0). In contrast, there is dependence if the change in levels of one of them is followed by changes in the other (H_1).

Aiming to find the best hypothesis test for detecting dependence, these new methods have been extensively evaluated in several investigations (de Siqueira Santos et al., 2014; Ding & Li, 2015; Kinney & Atwal, 2014; Makonnen, 2019; Reshef et al., 2011). The most recurring methodology to approach this problem has been the implementation of simulations where the different methods are evaluated under several conditions such as sample size or type of association. In order to assess the performance of the tests under these different conditions, the type I error rate and power level have been the most popular approaches and have reported promising results. Dependence tests such as dCor, HHG, and MI have not only identified functional and non-functional relationships, but also presented high power levels across different sample sizes and noise levels while maintaining a satisfactory type I error rate (de Siqueira Santos et al., 2014; Kinney & Atwal, 2014; Makonnen, 2019).

Nevertheless, no test was the most powerful across all conditions. On the contrary, previous investigations indicated mixed results where different tests performed better than other tests under specific conditions. On one hand, the MI and MIC tests outperformed all other methods in cyclic relationships (de Siqueira Santos et al., 2014; Reshef et al, 2011) while Pearson's correlation coefficient had the best performance in linear associations (Clark, 2013; Ding & Li, 2015; Makonnen, 2019). On the other hand, some of such methods have performed worse than others depending on the noise levels. The HHG and dCor measures have shown dominance over MIC in several simulations when the noise levels are high (Gorfine et al., 2012; Simon & Tibshirani, 2014).

Following these results, some researchers have followed different strategies to recommend the use of particular tests in specific conditions. For example, de Siquiera Santos et al. (2014) provided a decision tree for which each node was divided considering the level of several conditions (e.g., large or small sample size; functional or non-functional relationship, among others). Moreover, Clark (2013) recommended the use of new tests instead of the classic methods due to its ability to assess a broader type of relationships, and Makonnen (2019) concluded that dCor should always be used when the sample size is large and MI when the underlying association is cyclic. However, although these recommendations provide general guidelines for the selection of the correct method, it is still unclear which test should be used when the type of relationship or noise levels are not known a priori as is most often the case in Psychology.

Thus, in search of simplifying the selection problem of dependence measures in psychological research, this study aims to:

1. Establish assessment methods that summarize the performance of each test across the different conditions.
2. Identify which dependence test performs best according to these methods to provide a general recommendation for future investigation in Psychology.

This will be a step towards the reduction of the selection problem when facing a situation where the underlying conditions of the relationship are not known by the researcher.

Following previous investigations, this study will implement a Monte-Carlo simulation in which the performance of nine different dependence tests will be compared, namely: (1) Pearson's product-moment correlation coefficient, (2) Spearman's rank

correlation coefficient, (3) Kendall rank correlation coefficient, (4) HHG measure, (5) Hoeffding's D, (6) Distance Correlation, (7) Mutual Information, (8) Hilbert-Schmidt Independence Criterion, and (9) a combination between HHG and Pearson's correlation with a Bonferroni correction. Furthermore, as the identification of dependence and hypothesis testing are the scope of this research, the comparison will be done through power level and type I error rate under several conditions that can be found when doing research in Psychology.

The structure of the remainder of this document entails a methods section, in which the nine dependence measures will be described, and the approaches to the assessment of the performance of the tests will be explained. Later, a results section will present the outcomes of the simulation for each assessment approach. And lastly, a final section will discuss the results in light of previous research and will be followed by the conclusions and recommendations for future research on this topic.

Method

Statistical Independence

All of the nine methods mentioned above test the dependence between variables, which can be described by the following two hypotheses:

1. H_0: independence of variables
2. H_1: dependence between variables

The null hypothesis of independence of two random variables X and Y can be mathematically expressed by the following equality:

$$F_{X,Y}(x, y) = F_X(x)F_Y(y) \quad (1)$$

where $F_{X,Y}(x, y)$ can either be the joint cumulative distribution function, the joint probability density function, or the joint characteristic function of the combined variable (x, y). Likewise, $F_X(x)$ and $F_Y(y)$ are the marginal cumulative distribution functions, marginal probability density functions, or marginal characteristic functions of the variables x and y. To provide context, the characteristic functions can be briefly defined as transformations of the probability density functions that can take values in the complex numbers (for more details please see DasGupta [2011] and Székely et al. [2007]).

Dependence tests

Pearson's product-moment correlation coefficient

Pearson's correlation (Galton, 1889; Pearson, 1920) is the most commonly used method to measure the association between two variables in Psychology. It can only identify linear relationships between continuous variables and assumes that both variables are normally distributed. Moreover, it can take any value between -1 and 1, for which a coefficient of -1 represents a perfect linear negative association, and 1 shows a perfect linear positive association. Mathematically, it is the standardized form of the covariance between two variables x and y. First, let us define the covariance as a measure of the linear joint variability of those two variables. It can be expressed as:

$$cov(x, y) = E[(X - E[X])(Y - E[Y])] \tag{2}$$

where $E[\,]$ stands for the expected value of the variables, which in this case can be understood as their means (Rice, 2007). That is, the covariance measures the joint variability of two variables in terms of their joint deviation from their respective means. Now, let us define Pearson's correlation for the population as follows:

$$\rho(x, y) = \frac{cov(x, y)}{\sigma_x \sigma_y} \tag{3}$$

for which σ_x and σ_y are the standard deviations of the variables x and y. Please note that in case of perfect linear dependence $cov(x, y) = \sigma_x \sigma_y$ and $\rho(x, y) = 1$. Please remember that the covariance measures the joint variability in terms of the deviation of the variables from their means.

Furthermore, it is also possible to redefine the covariance as follows:

$$cov(x, y) = E[XY] - E[X]E[Y] \tag{4}$$

Considering this definition and following the null hypothesis of independence where $E[XY] = E[X]E[Y]$, the right side of Equation (4) would be equal to 0, thus, $cov(x, y) = 0$ and $\rho(x, y) = 0$. Therefore, Pearson's correlation coefficient value will be 0 in case the variables are independent. However, it is relevant to mention that this only holds for linear dependences as the covariance only measures the linear joint variability.

Finally, Pearson's correlation coefficient for a sample is computed as follows:

$$r_{Pearson}(x,y) = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}} \quad (5)$$

where $\bar{x}$ and $\bar{y}$ represent the mean value of those two random variables. Moreover, the hypothesis testing for Pearson's correlation coefficient is calculated assuming a t-distribution with $n-2$ degrees of freedom under the null hypothesis of linear independence. The t-statistic is calculated as follows:

$$t = \frac{r(x,y)\sqrt{n-2}}{\sqrt{1 - r(x,y)^2}} \quad (6)$$

Spearman's rank correlation coefficient

Spearman's correlation coefficient (Spearman, 1904) uses the same formula as Pearson's correlation coefficient; however, it is not applied to the raw scores, but to the rank transformation of the values of each variable. That is, the elements of each variable are given a rank in increasing order from 1 to N values in the variable. This transformation allows Spearman's correlation coefficient to identify not only linear relationships but also nonlinear monotonic associations. It can also be calculated for ordinal variables, not only continuous ones. For the population it is mathematically expressed as:

$$r_{Spearman}(rg_x, rg_y) = \frac{cov(rg_x, rg_y)}{\sigma_{rgx}\, \sigma_{rgy}} \quad (7)$$

where rg_x and rg_y are the rank transformations of variables x and y, and σ_{rg} represent the standard deviation of the corresponding variable. Furthermore, for a sample it is defined as:

$$r_{Spearman}(rg_x, rg_y) = \frac{\sum(rg_x - \overline{rg}_x)(rg_y - \overline{rg}_y)}{\sqrt{\sum(rg_x - \overline{rg}_x)\sum(rg_y - \overline{rg}_y)^2}} \quad (8)$$

for which $\overline{rg}_x$ and $\overline{rg}_y$ represent the mean of the corresponding ranks. Moreover, if there are not tied ranks, Spearman's rank correlation coefficient can also be calculated as follows:

$$r_{Spearman}(rg_x, rg_y) = 1 - \frac{6\sum d^2}{n(n^2-1)} \quad (9)$$

for which d are the differences between the corresponding ranks of the values from each variable. Spearman's correlation coefficient is similar to Pearson's correlation in the values it can assume. It ranges between -1 and 1, being -1 a perfect negative monotonic relationship, and 1 a perfect positive monotonic association. Finally, its hypothesis testing assumes a t-distribution under the null hypothesis of independence, and the t-statistic is calculated in the same way as in Pearson's correlation.

Kendall rank correlation coefficient

Similar to Spearman's correlation, Kendall rank correlation (Kendall, 1938) is a measure of dependence between variables that applies a rank transformation to the values of two variables and can take values from -1 to 1. However, the calculation of the coefficient is different, it can be expressed as follows:

$$\tau(x, y) = \frac{(number\ of\ concordant\ pairs) - (number\ of\ discordant\ pairs)}{0.5n(n-1)} \quad (10)$$

where for observations (x_i, y_i) and (x_j, y_j) a pair is considered concordant when $x_i > x_j$ and $y_i > y_j$ or when $x_i < x_j$ and $y_i < y_j$. On the other hand, a pair is considered discordant when $x_i > x_j$ and $y_i < y_j$ or when $x_i < x_j$ and $y_i > y_j$. In the same way as Spearman's correlation, Kendall rank correlation offers several advantages compared to Pearson's correlation coefficient. First, it can not only identify linear relationships, but also nonlinear monotonic associations. Second, it can be calculated for ordinal variables. Third, it does not need to assume that the variables are normally distributed.

Finally, the usual way to perform the hypothesis testing for Kendall rank correlation coefficient is through an approximation to a normal distribution where $\tau(x, y)$ is transformed to Z values as can be seen below:

$$Z = \frac{\tau(x, y)}{\sigma} \quad (11)$$

for which σ is the standard deviation of τ and can be defined as $\sigma = \sqrt{\frac{2(2n+5)}{9n(n-1)}}$. Kendall rank correlation coefficient, Pearson's correlation coefficient, and Spearman's rank correlation coefficient were calculated using the 'cor.test' function from the R package 'base' (R Core Team, 2019).

Heller-Heller-Gorfine measure

HHG measure was proposed by Heller et al. (2012) as an alternative to consistently test dependence between variables regardless of their type of association. It is based on the Euclidean or norm distances of two variables x and y. Let that distance be $d(.,.)$ for either x or y, so that the distance between points i and j can be represented by $d(x_i, x_j)$ or $d(y_i, y_j)$. Now, let us consider observations $k = (1, \dots, n)$ where $k \neq i$ and $k \neq j$. HHG creates a cross-classification table with elements $A_{11}\ A_{12}\ A_{21}\ A_{22}$ where:

$$A_{11(i,j)} = \sum_{\substack{k=1 \\ k \neq i \\ k \neq j}}^{n} I\left\{d(x_i, x_k) \leq d(x_i, x_j)\right\} I\left\{d(y_i, y_k) \leq d(y_i, y_j)\right\} \quad (12)$$

$$A_{12(i,j)} = \sum_{\substack{k=1 \\ k \neq i \\ k \neq j}}^{n} I\left\{d(x_i, x_k) \leq d(x_i, x_j)\right\} I\left\{d(y_i, y_k) > d(y_i, y_j)\right\} \quad (13)$$

$$A_{21(i,j)} = \sum_{\substack{k=1 \\ k \neq i \\ k \neq j}}^{n} I\left\{d(x_i, x_k) > d(x_i, x_j)\right\} I\left\{d(y_i, y_k) \leq d(y_i, y_j)\right\} \quad (14)$$

$$A_{22(i,j)} = \sum_{\substack{k=1 \\ k \neq i \\ k \neq j}}^{n} I\left\{d(x_i, x_k) > d(x_i, x_j)\right\} I\left\{d(y_i, y_k) > d(y_i, y_j)\right\} \quad (15)$$

for which $I(.)$ is an indicator function that assigns a 1 if the statement inside is true, and a 0 if it is false. These A elements are then used to calculate $A_{1.(i,j)}$, $A_{2.(i,j)}$, $A_{.1(i,j)}$, and $A_{.2(i,j)}$ that represent the following: $A_{m.(i,j)} = A_{m1(i,j)} + A_{m2(i,j)}$ and $A_{.m(i,j)} = A_{1m(i,j)} + A_{2m(i,j)}$ where $m = 1, 2$. A summary regarding all A elements can be seen in the cross-classification table in Table 1. Considering the previous information, the authors calculate the chi-squared statistic for 2 x 2 contingency tables as follows:

$$S(i,j) = \frac{(N-2)\{A_{12}(i,j)A_{21}(i,j)A_{11}(i,j)A_{22}(i,j)\}^2}{A_{1.}(i,j)\ A_{2.}(i,j)\ A_{.1}(i,j)\ A_{.2}(i,j)} \quad (16)$$

Finally, to test whether the two variables present an association, the authors suggest using a permutation test to obtain a p-value from the statistic T, which is defined as:

$$T = \sum_{i=1}^{n} \sum_{j=1\ j \neq i}^{n} S(i,j) \quad (17)$$

In summary, the HHG test looks into the number of times the distances of x and y values coincide and differ in order to identify if there is a relationship between the two variables. Afterward, this information is inserted in a 2 x 2 contingency table (see Table 1) that is used to compute several chi-square test statistics $S(i,j)$. Lastly, these $S(i,j)$ are summed to obtain a final T statistic that can be used to compute a permutation test. The HHG was calculated using the 'hhg.test' function from the R package 'HHG' provided by Brill & Kaufman (2019), which is based on the earlier implementation by Heller et al. (2012).

Table 1

Cross-Classification Table for A Elements

	$d(y_i, y_k) \leq d(y_i, y_j)$	$d(y_i, y_k) > d(y_i, y_j)$	Row Totals
$d(x_i, x_k) \leq d(x_i, x_j)$	$A_{11(i,j)}$	$A_{12(i,j)}$	$A_{1.(i,j)}$
$d(x_i, x_k) > d(x_i, x_j)$	$A_{21(i,j)}$	$A_{22(i,j)}$	$A_{2.(i,j)}$
Column Totals	$A_{.1(i,j)}$	$A_{.2(i,j)}$	

Hoeffding's D

Hoeffding's D is a non-parametric test of independence developed by Hoeffding (1948) and was one of the first approaches to test the nonmonotonic associations between variables regardless of their type of relationship. It is mathematically defined by Heller et al. (2016) as:

$$D = \iint N\{\hat{F}_{XY}(x,y) - \hat{F}_X(x)\hat{F}_Y(y)\}^2 d\hat{F}_{XY}(x,y) \tag{18}$$

where $\hat{F}$ represents either the joint or the marginal empirical cumulative distribution functions of all x and y in the population. It is important to mention that $D = 0$ in case the null hypothesis of independence $\hat{F}_{XY}(x,y) = \hat{F}_X(x)\hat{F}_Y(y)$ is true, that is, Hoeffding's D will be 0 if and only if x and y are independent. Furthermore, Hoeffding (1948) describes its empirical calculation based on the ranks of the two variables x and y. Let us consider those ranks as a and b respectively for both variables, and c as the points where both x and y are smaller than the value of observation i; c is expressed as follows:

$$c_i = \sum_{j=1}^{n} \varphi(x_j, x_i)\varphi(y_j, y_i) \tag{19}$$

where $\varphi(x_j, x_i) = 1$ when $x_j < x_i$ and $\varphi(x_j, x_i) = 0$ if $x_j \geq x_i$. And where $\varphi(y_j, y_i)$ is similarly defined. That is, c_i is the sum of the number of observations when both $x_j < x_i$ and $y_j < y_i$. Considering the previous information, the original author defined the following statements:

$$A = \sum_{i=1}^{n} (a_i - 1)(a_i - 2)(b_i - 1)(b_i - 2) \tag{20}$$

$$B = \sum_{i=1}^{n} (a_i - 2)(b_i - 2)\, c_i \tag{21}$$

$$C = \sum_{i=1}^{n} c_i(c_i - 1) \tag{22}$$

A, B, and C are then used to define Hoeffing's D in the mathematical expression below:

$$D(x, y) = \frac{A - 2(n-2)B + (n-2)(n-3)C}{n(n-1)(n-2)(n-3)(n-4)} \tag{23}$$

Moreover, the author defines a *D*-test of independence for a desired α in the following inequality:

$$P\{D(x, y) > \rho_n\} \leq \alpha \tag{24}$$

where P is the probability distribution under the null hypothesis of independence, and ρ_n is the smallest number that satisfies the inequality. Its value is decided with the following formula:

$$\rho_n = \frac{1}{30}\sqrt{\frac{2(n^2 + 5n - 32)}{9n(n-1)(n-3)(n-4)\alpha}} \tag{25}$$

Finally, the H_0 null hypothesis of independence between variables x and y is rejected if and only if $D(x, y) > \rho_n$. Hoeffding's D was calculated through the 'hoeffd' function from the 'Hmisc' R package developed by Harrell (2020).

Distance Correlation

Distance correlation was developed by Székely et al. (2007) to measure dependence between two variables x and y based on Euclidean distances between sample elements. First, it transforms the joint probability density function $F_{X,Y}(x, y)$ and the

marginal probability density functions $F_X(x)$ and $F_Y(y)$ into joint and marginal characteristic functions that assume values in the complex numbers. Please note that a complex number can be described in the form $z = x + iy$ where $i = \sqrt{-1}$, x is the real part of the complex number, and y is the imaginary part of the complex number (DasGupta, 2011). Furthermore, Karr (1993) define the characteristic function of a random variable x as the expected value of e^{itX} in the following expression:

$$f_X(t) = E\left[e^{itX}\right] = \int_{-\infty}^{\infty} e^{itX} F_X(x)\, dx \tag{26}$$

for which i is the imaginary part, and t is the real part of the characteristic function. Considering these transformations, Székely et al. (2007) proceed to first define the distance covariance as:

$$dCov = \frac{1}{c^2} \iint \frac{|f_{X,Y}(t,s) - f_X(t) f_Y(s)|^2}{|t|^2 |s|^2}\, dt\, ds \tag{27}$$

where t and s are the real part of the characteristic functions of the variables x and y, and c is a constant. Please note that based on this definition, and under the null hypothesis of independence $f_{X,Y}(t,s) = f_X(t) f_Y(s)$ the distance covariance $dCov = 0$ when the product of the marginal characteristic functions is equal to the joint characteristic function. Furthermore, these characteristic functions involve several advantages compared to the estimation of probability density functions, such as less computational burden when making computations with random variables. For further information regarding characteristic functions, the readers are referred to the original investigation of Székely et al. (2007).

After transforming the functions, the authors compute the Distance Correlation in several steps. First, let us consider the variables $A_{kl} = a_{kl} - \bar{a}_{k.} - \bar{a}_{.l} + \bar{a}$ and $B_{kl} = b_{kl} - \bar{b}_{k.} - \bar{b}_{.l} + \bar{b}$. For which all elements in A_{kl} are described as follows (please note that $|\,.\,|$ stands for Euclidean distance):

$$a_{kl} = |x_k - x_l| \tag{28}$$

$$\bar{a}_{k.} = \frac{1}{n} \sum_{l=1}^{n} a_{kl} \tag{29}$$

$$\bar{a}_{.l} = \frac{1}{n}\sum_{k=1}^{n} a_{kl} \tag{30}$$

$$\bar{a} = \frac{1}{n^2}\sum_{k,l=1}^{n} a_{kl} \tag{31}$$

That is, a_{kl} is the Euclidean distance between k and l for variable x; $\bar{a}_{k.}$ is the mean of k row; $\bar{a}_{.l}$ is the mean of l column; and $\bar{a}$ is the mean of the whole distance matrix. All elements in B_{kl} are similarly defined for variable y. Second, and considering the previous definitions, the empirical distance covariance, and the respective distance variance of two random variables x and y are calculated as can be seen below:

$$dCov(X,Y) = \sqrt{\frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}B_{kl}} \tag{32}$$

$$dVar(X) = dCov(X,X) = \sqrt{\frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}} \tag{33}$$

$$dVar(Y) = dCov(Y,Y) = \sqrt{\frac{1}{n^2}\sum_{k,l=1}^{n} B_{kl}} \tag{34}$$

Finally, Distance Correlation is the standardized value of distance covariance and is defined as:

$$dCor(x,y) = \begin{cases} \dfrac{dCov(x,y)}{\sqrt{dVar(x)dVar(y)}}, & dVar(x)dVar(y) > 0 \\ 0, & dVar(x)dVar(y) = 0 \end{cases} \tag{35}$$

The values that $dCor(x,y)$ can take will vary between 0 and 1, and will be 0 if and only if x and y are independent. The authors recommend performing hypothesis testing for Distance Correlation through a permutation test. In this study, it was calculated

through the permutation test from the 'dcor.test' function of the R package 'energy' provided by Rizzo & Szekely (2019).

Mutual Information

Mutual Information is originally a method from the Information Theory that can be defined as the reduction in the uncertainty of a variable x due to the knowledge of a variable y (Cover & Thomas, 2006). Let us consider a joint probability mass function $p(x, y)$ and marginal probability mass functions $p(x)$ and $p(y)$; mutual information $I(x, y)$ is then mathematically expressed as:

$$I(x, y) = E\left(log\frac{p(x, y)}{p(x)p(y)}\right) \tag{36}$$

Which for continuous random variables can be redefined considering the joint probability density function $f(x, y)$ and the marginal probability density functions $f(x)$ and $f(y)$ as follows:

$$I(x, y) = E\left(log\frac{p(x, y)}{p(x)p(y)}\right) = \iint f(x, y)log\frac{f(x, y)}{f(x)f(y)}dx\,dy \tag{37}$$

Please note that $\frac{f(x,y)}{f(x)f(y)} = 1$ under the previously defined null hypothesis of independence $f(x, y) = f(x)f(y)$. And, that $log(1) = 0$ if we consider a logarithm of base 2. Thus, $I(x, y) = 0$ when variables x and y are independent and the product of their respective marginal density functions is equal to its joint density function.

After stating the definition of Mutual Information, the only problem left is the estimation of the density functions. Several methods have been proposed to solve this problem, going from count and histograms for discrete data, to K-Nearest-Neighbors, Bayesian approaches, and kernel estimation for continuous data (Gencaga et al., 2014; Kraskov et al., 2004; Paninski, 2003; Wand & Jones, 1995). The estimation of Mutual Information can greatly vary depending on the selection of the approach used to address this problem. And, although methods such as K-Nearest-Neighbors have shown good results for Mutual Information (Ding & Li, 2015; Kinney & Atwal, 2014), this study used the Kernel Density Estimation (KDE) approach as it is one of the most common choices when computing MI (Kinney & Atwal, 2014; Makonnen, 2019; Reshef et al., 2011), and has shown a good performance compared to other estimation methods (Khan et al, 2007).

The use of KDE to estimate a density function $f(x)$ requires the selection of two fundamental parameters: the type of kernel K, and the bandwidth h which defines the kernel smoothing (Wand & Jones, 1995). Both of these parameters are in the definition of the estimated density function $\hat{f}(x)$ for a random sample $X_1, \dots, X_n$ provided by Park & Marron (1990) that can be seen below:

$$\hat{f}(x) = n^{-1} \sum_{i=1}^{n} h^{-1} K\{ h^{-1}(x - X_i)\} \tag{38}$$

There are several kernels (Gaussian, Triangular, Box, among others) and several algorithms for bandwidth calculation that have shown different and positive results in various situations (Makonnen, 2019; Harpole et al., 2014; Reshef et al., 2011). Nevertheless, the selected K and h for this study were the Epanechnikov kernel and the Sheater-Jones plug-in (SJDP) algorithm respectively. First, Wand & Jones (1995) showed that the Epanechnikov kernel is the most efficient compared to other options such as Biweight, Triweight, Normal, and Triangular kernels, even when it is only a trivial difference. Second, Harpole et al. (2014) compared eight different algorithms commonly used to calculate the bandwidth and found that in general SJPD outperforms all other methods. Moreover, the Epanechnikov kernel is the default selection in the function 'cmi' from the 'mpmi' R package (Pardy, 2020) which was the one used to compute MI in this study, and SJPD is also the default option in the 'dpik' function from the 'KernSmooth' package (Wand, 2019) that was used to estimate the bandwidth for MI. For further information regarding the computation of SJPD, the readers are referred to the original investigation from Sheater & Jones (1991) and the plug-in section from Wand & Jones (1995). Finally, although the R function that will be used reports Z statistics for hypothesis testing of mutual information, this study will use a permutation test to assess the dependence between variables due to the recommendation of the author of the package (Pardy, 2020).

Hilbert-Schmidt Independence Criterion

A brief description of the definition of HSIC will be presented in this section. However, for a detailed explanation regarding all its concepts, the readers are referred to the original investigation and its extensions (Gretton et al., 2005; Gretton et al., 2007; Pfister et al., 2018). The HSIC was originally developed by Gretton et al. (2005) as a new

independence test for two random variables based on the cross-covariance operators between elements of Reproducing Kernel Hilbert Spaces (RKHS), which can be briefly defined as spaces that contain functions.

To define HSIC, let us start by considering an RKHS $\mathcal{F}$ that contains all real-valued functions of a variable x, and an RKHS $\mathcal{G}$ that contains all real-valued functions of a variable y. HSIC is based on the cross-covariance between the elements of $\mathcal{F}$ and $\mathcal{G}$, that can be defined as:

$$cov\big(f(x), g(y)\big) = \mathbf{E}_{x,y}[f(x)g(y)] - \mathbf{E}_x[f(x)]\mathbf{E}_y[g(y)] \tag{39}$$

for which f and g are all possible functions that are part of $\mathcal{F}$ and $\mathcal{G}$. Please note that under the null hypothesis of independence $\mathbf{E}_{x,y}[f(x)g(y)] = \mathbf{E}_{\mathrm{x}}[f(x)]\mathbf{E}_y[g(y)]$, the right side of the Equation (39) would be equal to 0, thus, $cov\big(f(x), g(y)\big) = 0$. As HSIC is based on this cross-covariance, then $HSIC = 0$ if and only if variables x and y are independent. For more information regarding the theoretical definition of HSIC for the population, the readers are referred to Gretton et al. (2005) and Pfister et al. (2018).

Now, let us define the empirical calculation of HSIC. Its estimator can be described as $HSIC(Z, \mathcal{F}, \mathcal{G})$ where $Z = \{(x_1, y_1), \dots, (x_m, y_m)\}$, that is, Z is a series of m observations drawn from the joint distribution $F_{xy}(x, y)$ of a sample of the variables x and y. Considering this, $HSIC(Z, \mathcal{F}, \mathcal{G})$ is defined as follows:

$$HSIC(Z, \mathcal{F}, \mathcal{G}) = (m-1)^{-2}\mathbf{tr}(KHLH) \tag{40}$$

where H, K and L are matrices of dimensions $m \times \mathrm{m}$; H is a centering matrix, $K_{ij} = k(x_i, x_j)$, and $L_{ij} = l(y_i, y_j)$. That is, K and L are matrices containing the kernel functions k and l of the variables x and y. Please note that if either K or L are already centered then it is no longer necessary to add the matrix H and it is possible to rewrite $\mathbf{tr}(KHLH)$ simply as $\mathbf{tr}(KL)$.

Furthermore, as HSIC involves the use of kernels, similar to Mutual Information it also requires the selection of the type of kernel. For this study, we chose the Gaussian kernel following what is recommended by Pfister et al. (2018) who obtained satisfactory results after testing HSIC under several conditions using it. Moreover, it is also the default option in the ‘dhsic.test’ function from the ‘dHSIC' R package provided by Pfister &

Peters (2019) which was used to calculate HSIC in this research. It is important to note that this function specifically calculates the d-variable Hilbert-Schmidt Independence Criterion (dHSIC) developed by Pfister et al. (2018) as an extension of the original HSIC to test independence for more than two variables. Nevertheless, as we are only interested in studying independence between two variables, the extension is not relevant for this study.

Finally, after calculating the HSIC value, a hypothesis test should be performed. Based on Pfister et al. (2018) this can be carried out in three different ways: a permutation test, a bootstrap test, or a gamma approximation. The same authors evaluated these three procedures in several simulations and concluded that the permutation test was the most consistent method. Therefore, HSIC hypothesis testing was performed using a permutation test.

Pearson and HHG with a Bonferroni correction

Previous simulation studies have found that HHG outperforms most of the other methods in several conditions such as non-functional relationships or high noise levels, and it is at least as good as most of the modern tests at low noise levels and functional associations. However, when the underlying dependence between the two variables was linear it was considerably less powerful compared to other methods in general, and almost always when compared to Pearson's correlation coefficient specifically (de Siqueira Santos et al., 2014; Ding & Li, 2015; Gorfine et al., 2012).

Considering those previous results, this study added a new alternative that takes into account both HHG and Pearson's coefficient when performing the hypothesis testing of dependence between variables. The procedure was simple and involved the computation of both methods. First, the two measures were carried out and only the one with the minimum p-value was selected. Second, as we were doing multiple hypothesis testing for this new alternative, it was possible to incur the multiple comparison problem. This problem can be defined as the increase of the probability of finding at least one statistically significant result when the number of simultaneous tests being carried out increases. In other words, the more statistical tests are performed, the more likely is to incur wrong inferences (Goldman, 2008). To address this issue, a Bonferroni correction was applied by multiplying the selected p-value by the number of simultaneous tests,

which in this case was 2. This alternative will be referred to as HHG-Pearson for the remainder of this document.

Permutation Test

For four out of the nine methods, namely Pearson, Kendall, Spearman, and Hoeffding's D, the hypothesis testing was performed using the traditional approach assuming a parametric null distribution that was already implemented in its corresponding R packages. On the other hand, HHG, HSIC, dCor, and MI required the use of a permutation test as a non-parametric approach. The remaining method is the HHG-Pearson application with a Bonferroni correction; thus, the Pearson's test used the traditional approach, and HHG a permutation test. Before describing the details of the simulation, it is necessary to define what is a permutation test.

A permutation test is a non-parametric approach to hypothesis testing usually used when there is no way of knowing the null distribution under the null hypothesis of a test statistic (Rice & Lumley, 2008). In these cases, the permutation test will allow us to perform hypothesis testing by estimating the sample distribution under the null hypothesis. For dependence methods like the ones in this study, the null distribution is the one of independence between two variables. Therefore, let us consider two variables x and y that share some kind of association, and a paired data (x_i, y_i) for which the variable y_i will be randomly shuffled to simulate a situation of no relationship. Thus, creating a new paired data (x_i, y_i^*) where y_i^* is the randomly shuffled y_i that simulates a situation of no dependence. This procedure is done for every observation i and will be repeated a certain number of times. These repetitions are the permutations that give its name to the permutation test, and for every permutation, the test statistic is calculated. After estimating the sample null distribution, the original test statistic from the paired data (x_i, y_i) is compared to all the test statistics calculated for every permutation of the paired data (x_i, y_i^*) to observe the probability of occurrence of our original test statistic under the null hypothesis.

The number of repetitions can vary, but usually, the minimum is 1000. In this study, we used 10000 permutations for the tests that require this procedure, namely, HHG, HSIC, dCor, and MI. However, as HHG's computational burden increases exponentially with the sample size, and due to limitations in computational power, the number of permutations for HHG was 1000 when the sample size was greater than 100.

Simulation study

In order to compare the methods described in previous sections, a Monte-Carlo simulation was carried out. The simulation assessed the performance of the nine tests in several conditions that can be found while doing research in Psychology. For the current study, the factors considered and evaluated were sample size, type of relationship, and noise level. The design of the simulation can be summarized as follows:

1. Sample size (8 levels): 10, 20, 35, 50, 75, 100, 150 and 500
2. Type of relationship (5 levels): linear, exponential, quadratic, sine wave, and non-functional cross.
3. Noise (3 levels): low, medium, high.

For a total of 8*5*3 = 120 design cells that were repeated 1000 times. All of the nine tests were evaluated through all design cells. The details for each factor and its selected levels are described below.

Sample Size

The levels for this factor were selected based on the most commonly present sample sizes in Psychology published papers. Marszalhk et al. (2011) described the sample sizes used in Psychology research in four journals for the past three decades. The sample sizes varied from 1 to 45000, however, the most common values ranged from 10 to 140. Furthermore, Kuhberger et al. (2014) randomly selected 1000 published papers in different areas of psychological research and found that the most frequent values for the sample size varied between 1 and 100. Thus, eight levels were selected: 10, 20, 35, 50, 75, 100, 150 and 500. The chosen levels focused on sample sizes below 100 as these values were the most common in psychological research.

Type of Relationship

Four functional and one non-functional relationships between variables were considered for the simulation: linear, exponential, quadratic, sine wave, and non-functional cross. On the one hand, the exponential relationship was defined as a negative function considering that it can be easily found in the literature (Cajueiro, 2006; Leibowitz et al., 2010). Moreover, the quadratic relationship was established as an Inverted-U as it has been used to define different dependences between variables in Psychology such as the virtue-satisfaction or stress-performance relationships (Grant &

Schwartz, 2011; Westman & Eden, 1996). On the other hand, although sine-wave and cross relationships can be found in Psychology (Corning, 2002; Verboon & Leontjevas, 2018), they were mainly included as they are common relationships used to assess the performance of dependence tests (de Siquiera Santos et al., 2014; Ding & Li, 2015; Kinney & Atwal, 2014; Reshef et al., 2011). An example for all associations can be seen in Figure 1 and its respective functions can be seen in Equations (41) – (45). Please note they include an error term ε that will be explained in the following section.

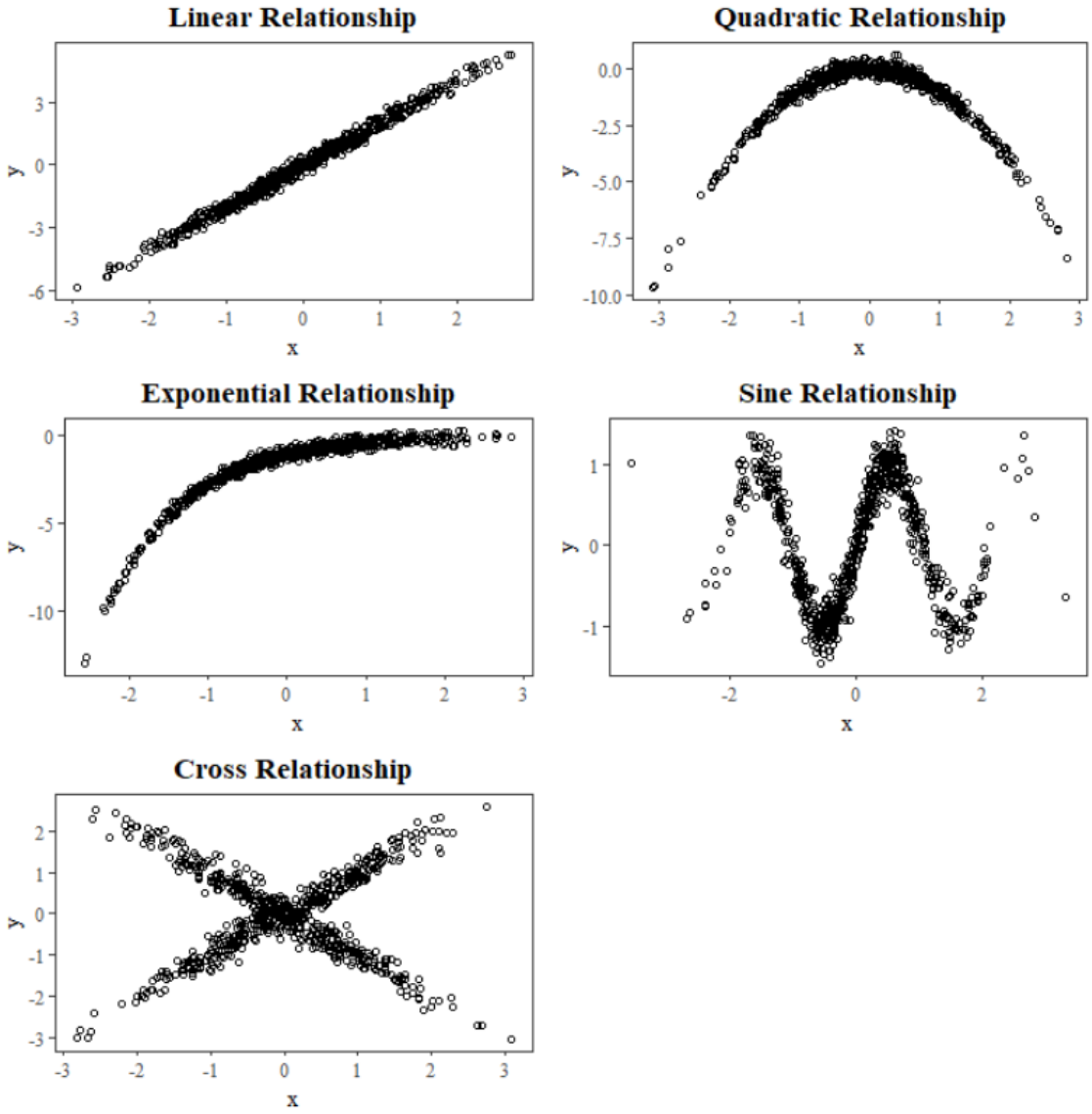

Figure 1. Five types of association evaluated in the simulation

Linear:	$y = (2x) + \epsilon$	(41)
Quadratic:	$y = (-x^2) + \epsilon$	(42)
Exponential:	$y = (-e^{-x}) + \epsilon$	(43)
Sine:	$y = \sin(3x) + \epsilon$	(44)
Cross:	$y = (\pm\, x) + \epsilon$	(45)

Noise Level

As a perfect relationship between variables in actual research is not plausible, an error term ϵ was added to the created Y variable. The error term ϵ was generated with a normal distribution $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon)$ where σ_ϵ stands for the standard deviation of ϵ. This σ_ϵ was defined by the noise level. Moreover, the noise levels were selected based on the proportion of variance of Y that was explained by X. That is, we decided to use the R^2 statistic to define the noise of the associations. It is important to note that although R^2 is often understood in Psychology as the proportion of variance of Y explained by X in a linear model, for this study it was considered as the proportion of variance explained by the true model regardless of whether it was linear or nonlinear. A formal definition of how R^2 was understood in this study can be seen below:

$$R^2 = 1 - \frac{\sigma_\epsilon^2}{\sigma_Y^2} \tag{46}$$

for which σ_ϵ^2 is the error variance of the error term ϵ (unexplained variance), and σ_Y^2 is the total variance of the Y variable. For more information regarding the definition of the R^2 statistic for nonlinear associations please see Doksum & Samarov (1995).

The R^2 values were specified following the literature in social research, and three levels were defined (Cohen, 1988; Hair et al, 2015). A low noise level was represented by an R^2 of 0.75, a medium level by an R^2 of 0.45, and a high level of noise by an R^2 of 0.1. An optimization procedure was performed to find the best σ_ϵ that led to the desired R^2 value. The final selected values for σ_ϵ varied for each type of relationship, but all of them led to the same three levels of R^2 for four out of the five associations. The remaining association was the cross non-functional dependence for which this procedure was not carried out as the expected R^2 for a perfect relationship of this type is zero. For this specific case, σ_ϵ was selected considering the typically used values in similar simulations

(Makonnen, 2019; Simon & Tibshirani, 2011). For details on the values of the noise levels please see Appendix A.

Assessment of the Methods

The evaluation of the performance of the methods was approached in two steps. First, by hypothesis testing via type I error rate and empirical power. Second, by summarizing the results of empirical power through three different approaches: complete class analysis, average power, and Maximin. Please note that as no difference between the tests was expected in type I error rate, the main comparison in the simulation was performed in terms of the empirical power and the summary approaches employed in this study.

Evaluating Hypothesis Test

Type I Error Rate. The type I error rate evaluated how often each method rejected the null hypothesis of independence between variables when there was truly no association between X and Y. This was done in two steps. First, by performing 1000 simulations for all sample sizes for which each method calculated the p-value with a set of independent data. Second, calculating the average number of p-values below the significance level. The defined significance level was 0.05 as it is one of the standard values used in psychological research.

Empirical Power. The empirical power can be defined as the probability of the test to reject the null hypothesis of independence when there was truly an association between X and Y. For this study, it was assessed for all conditions in the simulation and was calculated by estimating the average number of p-values that were lower than the defined significance level of 0.05. That is, we calculated how often the methods rejected the null hypothesis when the alternative hypothesis was true.

Summarizing Power

Considering that it is essentially impossible for a test to have the best performance (e.g., highest power) in all possible situations (Lehman & Romano, 2005), the performance of the tests over all conditions in the simulation was summarized to ease the comparison of the nine dependence measures. First, a complete class analysis was performed to reduce the number of viable methods based on their power. Second, the

average power was calculated to obtain a unique value to compare the tests. Third, a Maximin analysis was carried out.

Complete Class. A complete class procedure was carried out. It aims to reduce a decision problem (e.g., selecting the best dependence measure) without loss of relevant information. In other words, this procedure will decrease the number of viable tests to use in the future by selecting only the ones that perform the best out of all methods. To understand this procedure, let us consider the empirical power of each dependence test as a decision rule denominated as $power$. One of these methods can be eliminated from the decision problem if it is dominated by another $power'$ from a different method, so that:

$$power' \geq power \quad \text{for all conditions}$$

$$power' > power \quad \text{for some conditions}$$

In this case, $power$ is dominated by the decision rule $power'$, and the method corresponding to $power$ is now considered inadmissible in a complete class of dependence measures. Let us name this complete class as C. A method can only be admissible in C if such dominating $power'$ does not exist. A class is complete when for every decision rule $power$ not in class C, there is a $power'$ dominating it in C (Lehman & Romano, 2005). All methods outside C should not be considered anymore as there is another method inside C that is at least as good as the ones outside of it. As this research will not consider all existent dependence tests, the resulting C would only be a complete subclass. On the other hand, this procedure will reduce the number of possible dependence measures that can be considered for future research in Psychology. Finally, taking into account that it is exceedingly difficult to dominate a test across all simulated conditions, a small level of tolerance was considered for this analysis. Tolerance in this context can be defined as adding a small value close to 0 to the power of one of the tests such that a small margin of what is accepted as dominance is added to the comparison. To understand this better let us consider the tolerance level as tol, and let us redefine the previous conditions of dominance as:

$$power' + tol \geq power \quad \text{for all conditions}$$

$$power' > power \quad \text{for some conditions}$$

Average Power. The second step of the summary was performed by taking the average power of each method across all conditions. Thus, it was decided to take the

average power across the more likely unknown conditions (noise level, and type of association) for every certainly known condition (sample size) in actual research. In this way, a decision can be made based on a known factor when facing a research problem.

Maximin. The Maximin approach is another way to address the problem of deciding which test to use when there is not a unique measure with the best performance. It mainly consists of choosing the test with the maximum power in the worst-case scenario. In other words, the Maximin approach can be defined as considering the minimum power of each test and then finding the maximum among those minimums. It is important to note that the test with this maximum power must also maintain a low type I error rate (Lehman & Romano, 2005). In the same way as the average power, the Maximin was applied for the conditions that will be certainly known in actual psychological research. Thus, this study will report the Maximin results for every sample size considered in the simulation.

Results

The results of the simulation study will be presented in this section. First, a general overview of the hypothesis testing outcomes will be introduced through a table of the type I error rate. This overview shows how each dependence test behaved under the null hypothesis of independence. Second, a description of the results from each analysis approach to summarizing power will be presented in its corresponding subsections. Due to space constrains not all the power results of all conditions from the simulation will be described in this document. For the complete results the reader is referred to Appendix B.

Hypothesis Testing

Type I Error Rate

The results from the simulation for which both X and Y variables were generated at random for each evaluated sample size can be seen in Table 2. As the table shows, all the methods maintained the expected type I error rate under the specified significance level for most of the evaluated sample sizes. Only Hoeffding's D presented high type I error rates for small sample sizes ($N < 75$) and for the largest sample size ($N=500$).

Table 2

Type I Error Rate of the Dependence Measures for Each Sample Size

N	Pearson	Spear	Kend	Hoeffding	HHG	dCor	MI	HSIC	HHG-Pearson
10	.045	.054	.045	.115	.054	.049	.049	.054	.042
20	.046	.036	.038	.071	.043	.037	.043	.047	.035
35	.050	.045	.045	.063	.055	.044	.049	.048	.053
50	.057	.058	.053	.064	.050	.059	.055	.047	.048
75	.047	.049	.046	.054	.052	.041	.047	.050	.047
100	.040	.049	.049	.049	.041	.040	.052	.041	.030
150	.045	.041	.041	.047	.050	.046	.046	.056	.043
500	.047	.058	.058	.063	.059	.059	.043	.062	.042

Note. N = Sample size, Pearson = Pearson's correlation coefficient, Spear = Spearman's correlation coefficient, Kend = Kendall correlation coefficient, Hoeffding = Hoeffding's D, HHG = Heller-Heller-Gorfine Measure, dCor = Distance Correlation, MI = Mutual Information, HSIC = Hilbert-Schmidt Independence Criterion, HHG-Pearson = Combined HHG and Pearson with Bonferroni correction.

Summarizing Power

Complete Class

In order to reduce the number of viable tests that should be included in the complete subclass of dependence measures, the dominance relationships between the nine methods were analysed. As expected, there was not any test that dominated all other tests in terms of empirical power across all the simulated conditions. Nevertheless, Figure 2 shows the few dominance relationships that were found between the tests considering a small level of tolerance of 0.01. First, Distance Correlation presented dominance over two of the classical methods: Spearman's and Kendall correlation coefficients. Second, Mutual Information showed dominance over HSIC across all the simulated conditions. Considering these results, the dominated tests, namely, Spearman's coefficient, Kendall coefficient, and HSIC must not be included in the complete subclass of dependence tests. Thus, these three methods were not taken into account for the following analyses.

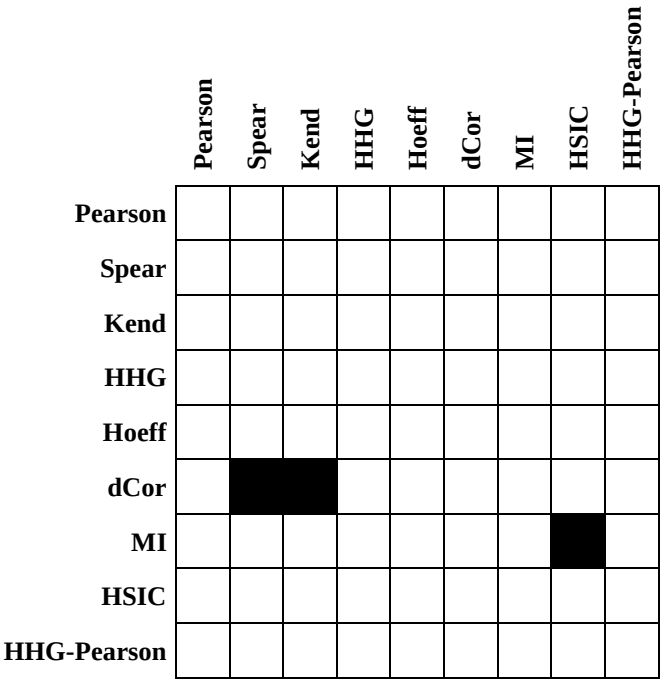

Figure 2. Dominance relationships between the tests. The black cells indicate that the test in the corresponding row dominates the test in the corresponding column across all the simulated conditions.

Average Power

Considering only the remaining six methods included in the complete subclass from the previous section, the average power across noise and type of association was calculated for all the simulated sample sizes. The outcomes can be seen in Table 3. The main result of this analysis was that Mutual Information had the highest average power for all sample sizes, except for the smallest one (N = 10), for which HHG-Pearson showed the maximum average power. It is important to note that Hoeffding's D average power for small sample sizes (N < 75) was not considered for this analysis due to its high type I error rate.

Table 3

Average Empirical Power for Each Sample Size, and Test with the Maximum Mean Power.

N	Pearson	Hoeffding	HHG	dCor	MI	HHG-Pearson	Max Test
10	.277	.288*	.232	.280	.272	.283	HHG-Pearson
20	.374	.384*	.446	.429	.503	.479	MI
35	.425	.503*	.629	.561	.666	.656	MI
50	.455	.593*	.728	.668	.764	.753	MI
75	.481	.702	.812	.774	.832	.831	MI
100	.489	.756	.860	.824	.876	.874	MI
150	.496	.810	.914	.876	.925	.913	MI
500	.508	.965*	.994	.987	.996	.987	MI

Note. * indicates the results were not considered for this analysis due to the high type I error rate of the method. N = Sample size, Pearson = Pearson's correlation coefficient, Hoeffding = Hoeffding's D, HHG = Heller-Heller-Gorfine Measure, dCor = Distance Correlation, MI = Mutual Information, HHG-Pearson = Combined HHG and Pearson with Bonferroni correction, Max Test = Test with the maximum average power for each sample size.

Moreover, the average power of all methods for each sample size can be seen in Figure 3. Three main outcomes are clearly visible in the figure. First, it is possible to see that the differences between HHG, MI, dCor, and HHG-Pearson were small. Second, there was a clear order in terms of performance for the six methods. It is as follows: MI, HHG-Pearson, HHG, dCor, Hoeffding's D, and Pearson's correlation. Lastly, it is clear that the higher the sample size, the higher the power for all nonmonotonic dependence tests. All of them seem to continuously approach a power of 1 when the sample size increase.

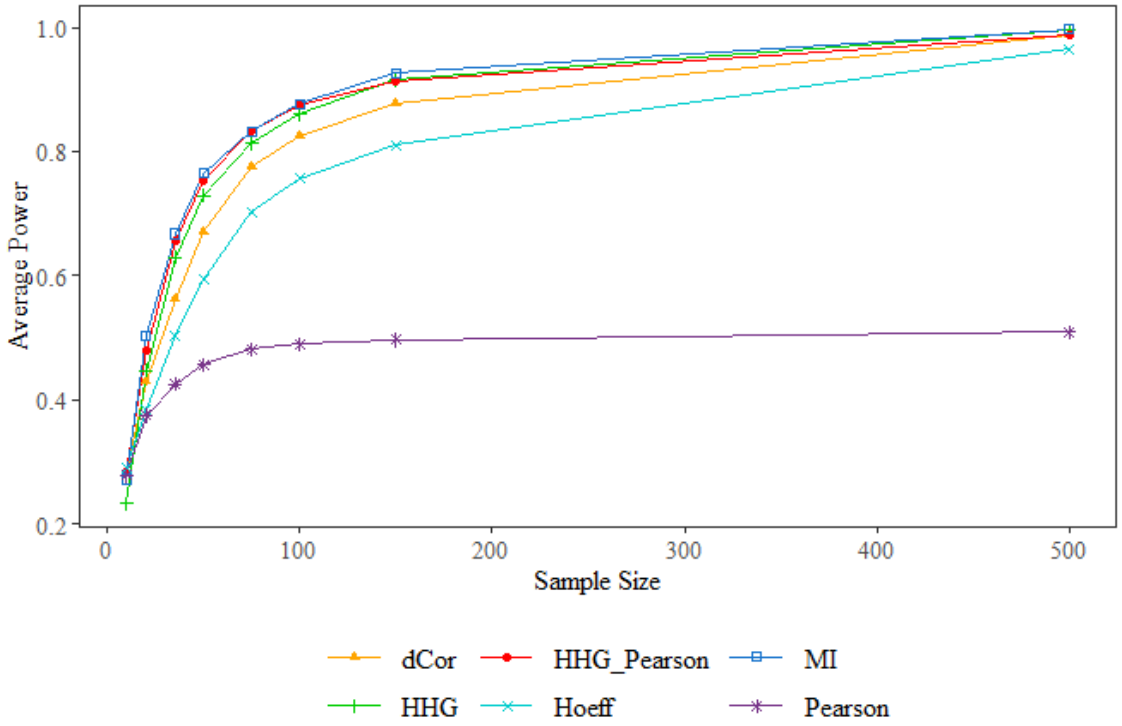

Figure 3. Average power across type of relationship and noise level for each one of the sample sizes in the simulation. Each line corresponds to one of the six dependence tests considered in the analysis. dCor = Distance Correlation, HHG-Pearson = Combined HHG and Pearson with Bonferroni correction, HHG = Heller-Heller-Gorfine, Hoeff = Hoeffding's D, MI = Mutual Information, Pearson = Pearson's correlation.

Maximin

Initially, and similar to the average power, the Maximin analysis was going to be performed for the six remaining methods. However, it was decided to also remove Hoeffding's D from the analysis due to its high type I error rates for the sample sizes where N<75. Thus, Maximin was carried out only for the five remaining dependence tests. Table 4 shows the minimum power across noise level and the type of association of the five remaining methods for each one of the sample sizes. The most important result from this analysis was that Mutual Information presented the best performance compared to the other methods in the worst-case scenario of minimum power. This was the case for all the simulated sample sizes.

Table 4

Minimum Power for Each Sample Size, and Test with the Maximum over those Minimums.

N	Pearson	HHG	dCor	MI	HHG-Pearson	Max Test
10	.036	.051	.047	.054	.040	MI
20	.051	.070	.057	.076	.049	MI
35	.049	.080	.073	.095	.061	MI
50	.054	.121	.101	.144	.098	MI
75	.059	.132	.098	.171	.090	MI
100	.063	.176	.117	.231	.124	MI
150	.046	.297	.179	.412	.206	MI
500	.062	.924	.819	.961	.811	MI

Note. N = Sample size, Pearson = Pearson's correlation coefficient, HHG = Heller-Heller-Gorfine Measure, dCor = Distance Correlation, MI = Mutual Information, HHG-Pearson = Combined HHG and Pearson with Bonferroni correction, Max Test = Test with the maximum minimum power (Maximin) for each sample size.

Moreover, and similar to the average power analysis, two different outcomes can be seen in Figure 4 which shows the minimum power for each one of the samples sizes. First, the difference of performance between some of the methods was small. More specifically, the differences between MI and HHG, and the difference between HHG-Pearson and dCor were small. Second, in terms of performance a clear order can be identified from the figure: MI, HHG, HHG-Pearson, dCor, and Pearson's correlation. This order is similar to the one from the average power analysis.

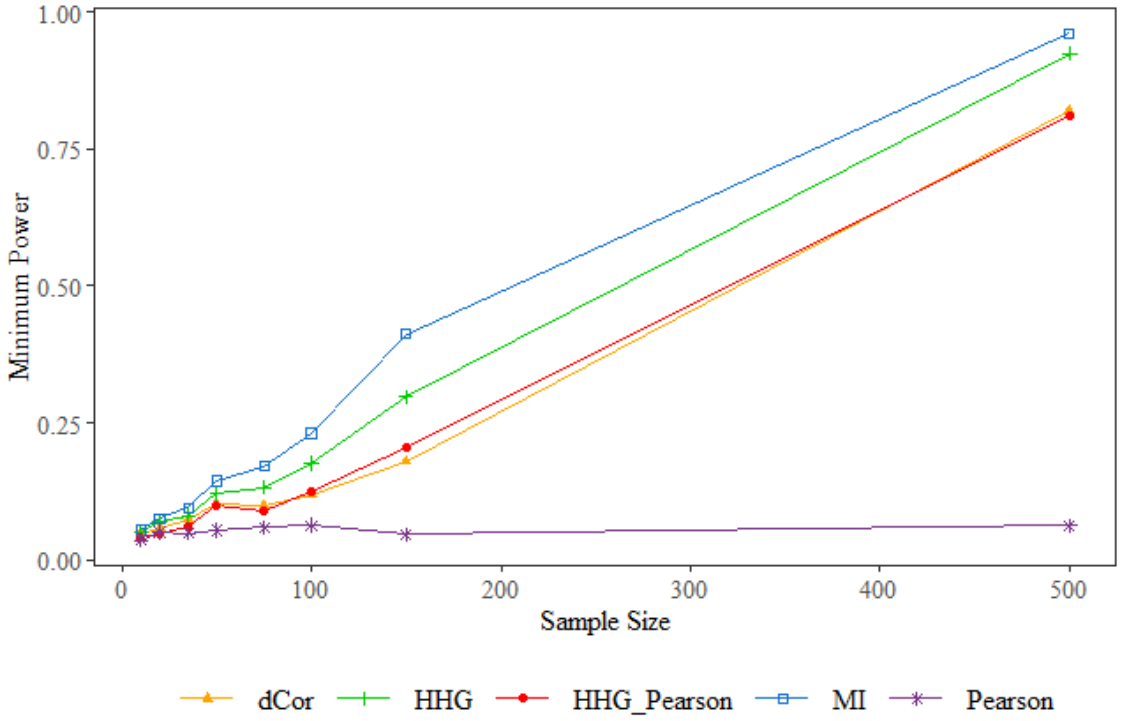

Figure 4. Minimum power across type of relationship and noise level for each one of the sample sizes in the simulation. Each line corresponds to one of the six dependence tests considered in the analysis. dCor = Distance Correlation, HHG-Pearson = Combined HHG and Pearson with Bonferroni correction, HHG = Heller-Heller-Gorfine, MI = Mutual Information, Pearson = Pearson's correlation.

Furthermore, Table 5 shows the details of the conditions involved in the minimum power scenario of each method by sample size. It is clear that the most challenging situations involved the presence of a sine wave relationship and a high noise level regardless of the sample size. The sine wave association was present in all the worst-case scenarios for all methods in the Maximin analysis, and the high noise level was present in all scenarios for HHG, dCor, MI, and HHG-Pearson. The only different conditions for this analysis were found in the minimum power of Pearson's correlation coefficient. More specifically, low and medium noise levels were present in the worst-case scenario of this test for the sample sizes of 35, 50, 75, and 100.

Table 5

Conditions Involved in the Scenario of Minimum Power of All Dependence Tests by Sample Size.

N	Pearson	HHG	dCor	MI	HHG-Pearson
10	SW - HN	SW - HN	SW - HN	SW - HN	SW - HN
20	SW - HN	SW - HN	SW - HN	SW - HN	SW - HN
35	SW - MN	SW - HN	SW - HN	SW - HN	SW - HN
50	SW - LN	SW - HN	SW - HN	SW - HN	SW - HN
75	SW - MN	SW - HN	SW - HN	SW - HN	SW - HN
100	SW - LN	SW - HN	SW - HN	SW - HN	SW - HN
150	SW - HN	SW - HN	SW - HN	SW - HN	SW - HN
500	SW - HN	SW - HN	SW - HN	SW - HN	SW - HN

Note. SW = Sine wave association, HN = High noise, MN = Medium noise, LN = Low noise, N = Sample size, Pearson = Pearson's correlation coefficient, HHG = Heller-Heller-Gorfine Measure, dCor = Distance Correlation, MI = Mutual Information, HHG-Pearson = Combined HHG and Pearson with Bonferroni correction.

Discussion

This study aimed to identify the best performing dependence measure in a simulation under several conditions. More specifically, nine dependence measures were assessed through hypothesis testing under 120 possible scenarios that combined three different conditions that can be found while doing psychological research: sample size, type of relationship, and noise level. The empirical power results were summarized and analyzed by three different approaches, namely, complete class, average power, and Maximin analyses.

Regarding our first analysis, the complete class results showed two main outcomes. First, Distance Correlation dominated two of the conventional dependence tests, namely, Spearman's correlation coefficient and Kendall correlation coefficient. These results are in line with previous research. For instance, Makonnen (2019) found that Distance Correlation consistently presented a higher empirical power than Spearman's and Kendall correlation coefficients. Moreover, other investigations concluded that Distance Correlation should be used in most situations while Spearman's and Kendall coefficients should not be used or should only be applied in specific conditions (Clark, 2013; de Siquiera Santos et al., 2014). Second, Mutual Information

dominated HSIC across all conditions. In summary, these results suggest that Spearman's coefficient, Kendall coefficient, and HSIC should not be used when testing for independence between two variables in psychological research. Although is important to note that these dominance relationships between tests involved a small tolerance level that increased the margin of what was accepted as dominance in this research.

Furthermore, the mean power analysis showed that on average Mutual Information outperformed all other methods in most conditions, except for the smallest sample size, for which HHG-Pearson showed the best average power. Moreover, Mutual Information also outperformed the other measures in the comparison based on the Maximin analysis, that is, Mutual Information had the best performance in the worst-case scenario of every method in the simulation. For the Maximin results is relevant to mention that the worst-case scenario by sample size in this simulation was almost the same for all the methods in this analysis. Therefore, these outcomes are only providing information regarding a few specific situations from the whole 120 possible scenarios. These specific situations almost always involved a sine wave association and a high noise level regardless of the sample size.

In short, even though there was not a single method that had the highest power in all conditions, Mutual Information outperformed all other methods considering the average power and the worst-case scenario. These results, alongside the dominance of MI over HSIC, are not consistent with results in previous research. For instance, de Siquiera Santos et al. (2014) compared eight dependence tests in a simulation and their results placed Mutual Information as one of the less powerful tests regardless of the type of association of the variables. Moreover, Makonnen (2019) performed a simulation with 320 possible scenarios and found that Hoeffding's D and dCor showed a better performance compared to Mutual Information in most cases. The only condition in which Mutual Information outperformed other methods was in the sine wave association. Other investigations have also shown that Mutual Information does not usually present the best performance in simulations under different conditions (Ding & Li, 2015; Reshef et al., 2011).

The unexpected and surprising results from this study can be explained by the variability of Mutual Information calculation depending on the approach used to estimate the probability density function of the variables (Khan et al., 2007). All investigations mentioned before employed different estimators when computing Mutual Information.

On one hand, investigations from Ding & Li (2015), and Reshef et al. (2011) employed a K-Nearest-Neighbor estimator while de Siquiera Santos et al. (2014) used a histogram approach to estimate the probability functions. On the other hand, Makonnen (2019) not only used the same method as this investigation (KDE) but also the same kernel (Epanechnikov). However, the author employed Silverman's rule of thumb algorithm to calculate the bandwidth of the kernel while this research used the SJDP approach.

In order to further investigate the difference between previous investigations and this research, a second simulation was performed to compare the Mutual Information estimators from de Siquiera Santos et al. (2014), Makonnen (2019), and this study. The conditions of the simulation were similar to the main simulation of this research, however, as this second simulation is out of the scope of this study its details will not be described in this document (for further information please see Appendix C). However, according to a complete class, average power, and Maximin analyses, its results placed this study's Mutual Information estimation as the method with the best performance compared to the approaches from the previously mentioned authors, followed by Makonnen (2019) estimation method, and de Siquiera Santos et al. (2014) approach. These results are consistent with Khan et al. (2007) who found that KDE outperforms other methods used to estimate Mutual Information in small sample sizes. Moreover, the results are also in line with the research from Harpole et al. (2014), in which SJDP outperformed all other bandwidth algorithms used when performing KDE. In summary, it seems clear that the selection of the estimation method can greatly affect the results of Mutual Information calculation. Moreover, these results suggest that Mutual Information estimated according to the settings of this study outperforms other estimation methods such as histogram count or KDE with a different bandwidth algorithm. However, this simulation did not include other estimation methods such as KNN or the Bayesian approach that should be considered when deciding how to address this problem.

Aside from the unexpected performance of the Mutual Information test, the results also showed that in general HHG, dCor, and HHG-Pearson are also presenting a satisfactory power level for almost all sample sizes. These results are in line with most of the outcomes of previous research. For instance, Ding & Li (2015) compared several dependence tests and found that dCor and HHG presented good performance for almost all conditions. Furthermore, several investigations have placed dCor and HHG between the best-performing methods regardless of the specific conditions used to assess them

(Clark, 2013; de Siquiera Santos et al., 2014; Gorfine et al., 2012; Makonnen, 2019; Simon & Tibshirani, 2014). Finally, it is relevant to take into account that all results from this study must be carefully considered keeping in mind the limitations that are discussed in the next section.

Limitations and Recommendations

Several limitations were present in the process of this research. First, the main problem was the deficiency of computational power while performing the simulation, this led to different issues. For instance, initially, a permutation test was considered for hypothesis testing in all methods to compare them in the same conditions. Nevertheless, due to this computational limitation, it was not feasible to perform a permutation test for all methods. However, it is also important to note that we took into account that a permutation test is not usually used when calculating the p-value of the classic correlation coefficients. A second issue related to the computational power was the reduction in the number of permutations for HHG from 10000 permutations to 1000 for sample sizes greater than 100. Moreover, only 1000 simulations were computed in this study due to its computational burden, a more accurate result could have been obtained with a greater number of simulations.

Second, taking into account that not all conditions in the simulation occur with the same frequency in actual research, we aimed to perform a weighted average of the power for which the weight of each type of relationship was going to be decided considering how often these dependencies appear in psychological research. Nevertheless, it resulted impossible to set a weight for each type of association based on real data as no information was found regarding the probability of occurrence of an underlying relationship in psychological research.

Third, a common limitation of simulation studies is that the results cannot be generalizable beyond the conditions that were simulated. In this case, only 120 scenarios from an almost infinite combination of possible conditions were simulated. However, is relevant to mention that the selected three conditions and their corresponding levels were chosen considering real possible scenarios in actual psychological research. Thus, the results are at least relevant to these possible situations that can be found while doing a psychological investigation.

For future research related to this topic, we recommend to further investigate the different possible estimation methods of Mutual Information as the choice of different settings can lead to greatly different results. Furthermore, this study only assessed the performance of the dependence measures through hypothesis testing, that is, we evaluated if the different measures could identify an association. However, for future investigation, it would be interesting to also assess the quantity of the relationship. In other words, we recommend evaluating if the methods can not only identify but also correctly measure how strong is the association. Finally, another recommendation for future research is the assessment of more methods created to measure associations between variables such as the Copula Correlation (Ding & Lin, 2015), the Copula-based Kernel Dependency Measure (Poczos et al., 2012), or the Randomized Correlation Coefficient (Lopez-Paz et al., 2013).

Conclusion

In conclusion, this study suggests that there is not a single method to test dependence between variables that has the highest power in all conditions. Nevertheless, we recommend the use of Mutual Information estimated through KDE with the Epanechnikov kernel and the SJDP bandwidth algorithm. This method outperformed all other dependence measures assessed in this study in the Maximin and the average power analyses regardless of the sample size. Moreover, it also outperformed two different types of Mutual Information. On the other hand, methods such as HHG, dCor, and the new alternative HHG-Pearson should not be immediately discarded. These three methods presented a good performance in almost all conditions and their difference in power compared to Mutual Information is small. These differences can be due to the specific design of the simulation of this study.

Code Accessibility

All the R code used to perform the simulation in this study can be found in https://github.com/AndresFPA/DependSim

References

Andersen, R. (2009). Nonparametric methods for modeling nonlinearity in regression analysis. *Annual Review of Sociology*, *35*(1), 67–85. https://doi.org/10.1146/annurev.soc.34.040507.134631

Brill, B., & Kaufman, S. (2019). HHG: Heller-Heller-Gorfine tests of independence and equality of distributions. (R package version 2.3.2) [Computer software]. The Comprehensive R Archive Network. Available from https://CRAN.R-project.org/package=HHG

Cajueiro, D. O. (2006). A note on the relevance of the q-exponential function in the context of intertemporal choices. *Physica A: Statistical Mechanics and Its Applications*, *364*, 385–388. https://doi.org/10.1016/j.physa.2005.08.056

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). L. Erlbaum Associates.

Corning, A. F. (2002). Self-esteem as a moderator between perceived discrimination and psychological distress among women. *Journal of Counseling Psychology*, *49*(1), 117–126. https://doi.org/10.1037/0022-0167.49.1.117

Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory* (2nd ed.). Wiley-Interscience.

DasGupta, A. (2011). *Probability for statistics and machine learning fundamentals and advanced topics.* Springer.

Ding, A. A., & Li, Y. (2015). Copula Correlation: An equitable dependence measure and extension of Pearson's correlation. *ArXiv:1312.7214 [Stat]*. http://arxiv.org/abs/1312.7214

Doksum, K., & Samarov, A. (1995). Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression. *The Annals of Statistics*, *23*(5), 1443-1473. https://doi.org/10.1214/aos/1176324307

Fox, J. (2016). *Applied regression analysis and generalized linear models* (3rd ed.). SAGE.

Frazier, P. A., Tix, A. P., & Barron, K. E. (2004). Testing moderator and mediator effects in counseling psychology research. *Journal of Counseling Psychology, 51*(1), 115–134. https://doi.org/10.1037/0022-0167.51.1.115

Gencaga, D., Malakar, N. K., & Lary, D. J. (2014). Survey on the estimation of mutual information methods as a measure of dependency versus correlation analysis. *AIP Conference Proceedings, 1636, 80,* 80–87. https://doi.org/10.1063/1.4903714

Goldman, M. (2008). *Why is multiple testing a problem?* https://www.stat.berkeley.edu/~mgoldman/Section0402.pdf

Gorfine, M., Heller, R., & Heller, Y. (2012). Comment on "Detecting Novel Associations in Large Data Sets." http://www.math.tau.ac.il/~ruheller/Papers/science6.pdf

Grant, A. M., & Schwartz, B. (2011). Too much of a good thing: The challenge and opportunity of the inverted U. *Perspectives on Psychological Science, 6*(1), 61–76. https://doi.org/10.1177/1745691610393523

Gretton, A., Bousquet, O., Smola, A. & Scholkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. *Algorithmic Learning Theory, 3734,* 63-77. https://doi.org/10.1007/11564089_7

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Scholkopf, B. & Smola, A. J. (2007). A kernel statistical test of independence. *Neural Information Processing Systems*. 585–592.

Hair, J. F., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM: Indeed a silver bullet. *Journal of Marketing Theory and Practice, 19*(2), 139–152. https://doi.org/10.2753/MTP1069-6679190202

Harpole, J. K., Woods, C. M., Rodebaugh, T. L., Levinson, C. A., & Lenze, E. J. (2014). How bandwidth selection algorithms impact exploratory data analysis using kernel density estimation. *Psychological Methods, 19*(3), 428–443. https://doi.org/10.1037/a0036850

Harrell, F. (2020). Hmisc: Harrell Miscellaneous. (R package version 4.4-2) [Computer software]. The Comprehensive R Archive Network. Available from https://CRAN.R-project.org/package=Hmisc

Hayfield, T., & Racine, J. S. (2008). Nonparametric econometrics: The np package () [Computer software]. The Comprehensive R Archive Network. Available from http://www.jstatsoft.org/v27/i05/

Heller, R., Heller, Y., & Gorfine, M. (2012). A consistent multivariate test of association based on ranks of distances. *Biometrika*, *100*(2), 503–510. https://doi.org/10.1093/biomet/ass070

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Karr, A. F. (1993). *Probability*. Springer New York.

Khan, S., Bandyopadhyay, S., Ganguly, A. R., Saigal, S., Erickson, D. J., Protopopescu, V., & Ostrouchov, G. (2007). Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. *Physical Review E*, *76*(2), 026209. https://doi.org/10.1103/PhysRevE.76.026209

Kinney, J. B., & Atwal, G. S. (2014). Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, *111*(9), 3354–3359. https://doi.org/10.1073/pnas.1309933111

Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical Review E*, *69*(6), 066138. https://doi.org/10.1103/PhysRevE.69.066138

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PLoS ONE*, *9*(9), e105825. https://doi.org/10.1371/journal.pone.0105825

Leibowitz, N., Baum, B., Enden, G., & Karniel, A. (2010). The exponential learning equation as a function of successful trials results in sigmoid performance. *Journal of Mathematical Psychology*, *54*(3), 338–340. https://doi.org/10.1016/j.jmp.2010.01.006

Lopez-Paz, D., Hennig, P., & Schölkopf, B. (2013). The Randomized Dependence Coefficient. *ArXiv:1304.7717 [Stat]*. http://arxiv.org/abs/1304.7717

Makonnen, S. (2019). *Comparing the performance of non-monotone dependence measures in psychology: A simulation study on the power of mutual information and distance correlation* [Master's Thesis, Leiden University]. Leiden University Student Repository. https://hdl.handle.net/1887/85357

Marszalhk, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills, 112*(2), 331-348. DOI 10.2466/03.11.PMS.n2.2.331-348

Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Computation*, *15*(6), 1191–1253. https://doi.org/10.1162/089976603321780272

Pardy (2020). mpmi: Mixed-Pair mutual information estimators. (R package version 0.43.1) [Computer software]. The Comprehensive R Archive Network. Available from https://CRAN.R-project.org/package=mpmi

Park, B. U., & Marron, J. S. (1990). Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, *85*(409), 66–72. https://doi.org/10.1080/01621459.1990.10475307

Pfister, N., Bühlmann, P., Schölkopf, B., & Peters, J. (2018). Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *80*(1), 5–31. https://doi.org/10.1111/rssb.12235

Pfister, N., & Peters, J. (2019). dHSIC: Independence testing via Hilbert-Schmidt independence criterion (R package version 2.1) [Computer software]. The Comprehensive R Archive Network. Available from https://CRAN.R-project.org/package=dHSIC

Poczos, B., Ghahramani, Z., & Schneider, J. (2012). Copula-based kernel dependency measures. In *International Conference on Machine Learning*. Edinburgh, Scotland

R Core Team (2019). R: A language and environment for statistical computing. [Computer software]. The Comprehensive R Archive Network. Available from https://www.R-project.org/

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., & Sabeti, P. C. (2011). Detecting novel associations in large data sets. *Science*, *334*(6062), 1518–1524. https://doi.org/10.1126/science.1205438

Rice, J. (2006). *Mathematical statistics and data analysis* (3rd ed.). Thomson Brooks/Cole.

Rice, K., & Lumley, T. (2008). *Permutation tests* [PowerPoint slides]. https://faculty.washington.edu/kenrice/sisg/SISG-08-06.pdf

Rizzo, M., & Szekely, G. (2019). Energy: E-statistics: multivariate inference via the energy of data (R package version 1.7-7) [Computer software]. The Comprehensive R Archive Network. Available from https://CRAN.R-project.org/package=energy

Sheather, S. J., & Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, *53*(3), 683–690. https://doi.org/10.1111/j.2517-6161.1991.tb01857.x

Simon, N., & Tibshirani, R. (2014). Comment on "Detecting Novel Associations in Large Data Sets" by Reshef et al, Science Dec 16, 2011. *ArXiv:1401.7645 [Stat]*. http://arxiv.org/abs/1401.7645

Verboon, P., & Leontjevas, R. (2018). Analyzing cyclic patterns in psychological data: A tutorial. *The Quantitative Methods for Psychology*, *14*(4), 218–234. https://doi.org/10.20982/tqmp.14.4.p218

Wand, M. (2019). KernSmooth: Functions for kernel smoothing (R package version 2.23-16) [Computer software]. The Comprehensive R Archive Network. Available from https://CRAN.R-project.org/package=KernSmooth

Wand, M., & Jones, M. (1995). *Kernel Smoothing* (1st ed.). Chapman & Hall.

Westman, M., & Eden, D. (1996). The Inverted-U relationship between stress and performance: A field study. *Work & Stress*, *10*(2), 165–173. https://doi.org/10.1080/02678379608256795

Appendix A

Noise level selection

The noise level for each type of relationship was decided in a process of several steps that can be seen below:

1. The desired R^2 values were defined taking into account literature in social research (Cohen, 1988; Hair et al, 2015).
    a. Low noise level: R^2 of 0.75.
    b. Medium noise level: R^2 of 0.45.
    c. High noise level: R^2 of 0.1.
2. Using the 'optimize' function from the 'stats' R package (R Core Team, 2019) an optimization procedure was performed. This procedure consisted of finding the optimal σ_ϵ that led to the desired R^2 in an iterative process. This was done in the following way:
    a. Generate random data for an X variable (sample size 5000).
    b. Generate a Y variable following a specific type of relationship with X (linear, quadratic, exponential, sine wave, or cross).
    c. Select a σ_ϵ value in a small range from 0 to 6.
    d. Add noise to the Y variable with the selected σ_ϵ value as the standard deviation of the noise.
    e. Calculate the R^2 value of this noisy relationship.
    f. Compute the square difference between the calculated and the desired R^2 values. That is, calculating $(current\ R^2 - desired\ R^2)^2$.
    g. If the computed difference is smaller than a previously defined tolerance level (0.001 in this study) then the process is stopped and the σ_ϵ value is selected as the final noise level.
    h. If the computed difference is not smaller than the tolerance level, then the procedure is performed again from the top. The selection of a new σ_ϵ value (step c) is carried out following a golden section search optimization method as this is the default method in the 'optimize' function.

This procedure was done for each type of relationship and each noise level. The R^2 values were calculated using the 'npregbw' function from the 'np' package (Hayfield

& Racine, 2008) which is based on the definition from Doksum & Samarov (1995). The final noise levels in σ_ϵ values can be seen in Table A1 below. Please note that the noise level of the cross-relationship was selected only using the typically employed values in similar simulations.

Table A1

Noise levels in σ_ϵ values for all types of relationships.

Relationship	Low noise	Medium noise	High noise
Linear	1.163181	2.310038	4.58325
Quadratic	0.856069	1.539105	3.99750
Exponential	1.212181	2.332178	4.58359
Sine wave	0.410357	0.767012	2.29179
Cross	0.3	0.6	1

Appendix B

Table B1

Complete results of the simulation for all 120 simulated scenarios

Relationship	N	Noise	Pear	Kend	Spear	HHG	HSIC	Hoeff	dCor	MI	HHG-Pear
linear	10	Low	0.953	0.876	0.881	0.741	0.824	0.857	0.939	0.816	0.924
linear	10	Medium	0.607	0.485	0.510	0.291	0.359	0.552	0.572	0.363	0.508
linear	10	High	0.208	0.174	0.185	0.093	0.113	0.261	0.189	0.123	0.139
linear	20	Low	1.000	0.999	0.998	0.988	0.989	0.998	0.998	0.993	1.000
linear	20	Medium	0.902	0.847	0.859	0.606	0.628	0.819	0.880	0.677	0.861
linear	20	High	0.434	0.376	0.394	0.191	0.208	0.398	0.396	0.230	0.340
linear	35	Low	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
linear	35	Medium	0.993	0.989	0.988	0.879	0.877	0.976	0.988	0.911	0.987
linear	35	High	0.676	0.618	0.627	0.312	0.310	0.586	0.619	0.344	0.592
linear	50	Low	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
linear	50	Medium	1.000	1.000	1.000	0.972	0.964	0.999	1.000	0.987	1.000
linear	50	High	0.839	0.790	0.793	0.462	0.445	0.749	0.792	0.512	0.775
linear	75	Low	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
linear	75	Medium	1.000	1.000	1.000	0.999	0.999	1.000	1.000	1.000	1.000
linear	75	High	0.965	0.948	0.951	0.679	0.657	0.913	0.941	0.678	0.933
linear	100	Low	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
linear	100	Medium	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
linear	100	High	0.979	0.973	0.974	0.813	0.786	0.962	0.971	0.832	0.969
linear	150	Low	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
linear	150	Medium	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
linear	150	High	0.999	0.998	0.998	0.933	0.921	0.993	0.997	0.913	0.998
linear	500	Low	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
linear	500	Medium	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
linear	500	High	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
quadratic	10	Low	0.249	0.120	0.105	0.313	0.330	0.249	0.318	0.362	0.326
quadratic	10	Medium	0.163	0.093	0.093	0.180	0.170	0.193	0.180	0.211	0.188
quadratic	10	High	0.075	0.044	0.038	0.067	0.066	0.099	0.080	0.077	0.072
quadratic	20	Low	0.302	0.131	0.110	0.718	0.645	0.393	0.696	0.830	0.672
quadratic	20	Medium	0.191	0.088	0.081	0.356	0.341	0.194	0.348	0.503	0.353
quadratic	20	High	0.104	0.063	0.062	0.112	0.098	0.121	0.116	0.151	0.098
quadratic	35	Low	0.296	0.138	0.108	0.955	0.907	0.716	0.918	0.982	0.937
quadratic	35	Medium	0.215	0.099	0.091	0.676	0.576	0.344	0.601	0.773	0.616
quadratic	35	High	0.091	0.069	0.069	0.179	0.152	0.110	0.134	0.195	0.138
quadratic	50	Low	0.305	0.133	0.108	0.994	0.983	0.914	0.995	0.999	0.994
quadratic	50	Medium	0.233	0.106	0.089	0.857	0.754	0.505	0.813	0.923	0.809
quadratic	50	High	0.097	0.070	0.062	0.231	0.189	0.130	0.188	0.295	0.199
quadratic	75	Low	0.340	0.153	0.123	0.999	0.999	0.990	0.999	1.000	0.999

Relationship	N	Noise	Pear	Kend	Spear	HHG	HSIC	Hoeff	dCor	MI	HHG-Pear
quadratic	75	Medium	0.232	0.113	0.090	0.971	0.913	0.772	0.950	0.988	0.949
quadratic	75	High	0.085	0.064	0.060	0.349	0.257	0.162	0.268	0.397	0.280
quadratic	100	Low	0.303	0.142	0.112	1.000	1.000	1.000	1.000	1.000	1.000
quadratic	100	Medium	0.213	0.102	0.086	0.998	0.980	0.914	0.992	0.999	0.991
quadratic	100	High	0.114	0.066	0.064	0.473	0.346	0.182	0.342	0.497	0.380
quadratic	150	Low	0.312	0.162	0.129	1.000	1.000	1.000	1.000	1.000	1.000
quadratic	150	Medium	0.229	0.091	0.081	1.000	0.998	0.998	1.000	1.000	1.000
quadratic	150	High	0.116	0.063	0.063	0.706	0.525	0.325	0.569	0.722	0.596
quadratic	500	Low	0.344	0.135	0.105	1.000	1.000	1.000	1.000	1.000	1.000
quadratic	500	Medium	0.247	0.105	0.089	1.000	1.000	1.000	1.000	1.000	1.000
quadratic	500	High	0.110	0.080	0.078	1.000	0.988	0.947	0.996	0.994	0.996
exponential	10	Low	0.642	0.487	0.481	0.346	0.411	0.538	0.631	0.413	0.577
exponential	10	Medium	0.377	0.257	0.254	0.173	0.196	0.328	0.337	0.208	0.295
exponential	10	High	0.151	0.109	0.108	0.088	0.109	0.175	0.147	0.111	0.125
exponential	20	Low	0.929	0.868	0.859	0.747	0.722	0.866	0.915	0.789	0.899
exponential	20	Medium	0.645	0.500	0.511	0.336	0.340	0.503	0.605	0.445	0.572
exponential	20	High	0.320	0.207	0.207	0.136	0.133	0.240	0.271	0.164	0.253
exponential	35	Low	0.997	0.980	0.983	0.946	0.923	0.976	0.990	0.971	0.988
exponential	35	Medium	0.868	0.786	0.778	0.615	0.557	0.759	0.854	0.711	0.836
exponential	35	High	0.488	0.370	0.357	0.219	0.192	0.356	0.421	0.305	0.405
exponential	50	Low	0.999	0.998	0.999	0.994	0.984	0.996	0.998	0.995	0.999
exponential	50	Medium	0.964	0.921	0.921	0.798	0.753	0.912	0.959	0.870	0.947
exponential	50	High	0.612	0.487	0.480	0.336	0.277	0.459	0.548	0.409	0.542
exponential	75	Low	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	1.000
exponential	75	Medium	0.990	0.974	0.973	0.928	0.871	0.974	0.985	0.945	0.984
exponential	75	High	0.792	0.700	0.694	0.475	0.397	0.643	0.738	0.544	0.734
exponential	100	Low	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
exponential	100	Medium	0.999	0.997	0.997	0.980	0.947	0.993	1.000	0.990	1.000
exponential	100	High	0.895	0.807	0.802	0.607	0.504	0.777	0.845	0.690	0.858
exponential	150	Low	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
exponential	150	Medium	1.000	1.000	1.000	1.000	0.995	1.000	1.000	0.998	1.000
exponential	150	High	0.962	0.914	0.916	0.814	0.715	0.902	0.950	0.862	0.949
exponential	500	Low	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
exponential	500	Medium	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
exponential	500	High	1.000	1.000	1.000	1.000	0.998	1.000	1.000	0.999	1.000
sine wave	10	Low	0.068	0.076	0.082	0.190	0.235	0.232	0.110	0.251	0.142
sine wave	10	Medium	0.067	0.058	0.068	0.099	0.146	0.188	0.091	0.138	0.073
sine wave	10	High	0.036	0.040	0.049	0.051	0.057	0.113	0.047	0.054	0.040
sine wave	20	Low	0.053	0.090	0.085	0.567	0.482	0.407	0.230	0.513	0.406
sine wave	20	Medium	0.052	0.074	0.084	0.230	0.233	0.217	0.139	0.249	0.171
sine wave	20	High	0.051	0.050	0.048	0.070	0.085	0.086	0.057	0.076	0.049

Relationship	N	Noise	Pear	Kend	Spear	HHG	HSIC	Hoeff	dCor	MI	HHG-Pear
sine wave	35	Low	0.056	0.137	0.121	0.924	0.755	0.720	0.491	0.779	0.853
sine wave	35	Medium	0.049	0.095	0.087	0.467	0.460	0.308	0.233	0.513	0.350
sine wave	35	High	0.049	0.054	0.056	0.080	0.088	0.097	0.073	0.095	0.061
sine wave	50	Low	0.054	0.141	0.129	0.997	0.923	0.924	0.780	0.949	0.989
sine wave	50	Medium	0.065	0.113	0.106	0.706	0.652	0.484	0.402	0.732	0.580
sine wave	50	High	0.065	0.082	0.084	0.121	0.127	0.127	0.101	0.144	0.098
sine wave	75	Low	0.084	0.187	0.172	1.000	0.996	0.998	0.993	0.997	1.000
sine wave	75	Medium	0.059	0.135	0.140	0.930	0.862	0.765	0.703	0.940	0.864
sine wave	75	High	0.061	0.065	0.068	0.132	0.153	0.115	0.098	0.171	0.090
sine wave	100	Low	0.063	0.226	0.217	1.000	1.000	1.000	1.000	1.000	1.000
sine wave	100	Medium	0.067	0.152	0.150	0.987	0.959	0.931	0.928	0.990	0.975
sine wave	100	High	0.065	0.072	0.073	0.176	0.179	0.145	0.117	0.231	0.124
sine wave	150	Low	0.096	0.308	0.289	1.000	1.000	1.000	1.000	1.000	1.000
sine wave	150	Medium	0.070	0.202	0.206	1.000	1.000	0.999	0.997	1.000	1.000
sine wave	150	High	0.046	0.062	0.067	0.297	0.296	0.203	0.179	0.412	0.206
sine wave	500	Low	0.145	0.719	0.681	1.000	1.000	1.000	1.000	1.000	1.000
sine wave	500	Medium	0.116	0.512	0.493	1.000	1.000	1.000	1.000	1.000	1.000
sine wave	500	High	0.062	0.117	0.115	0.924	0.914	0.792	0.819	0.961	0.811
cross	10	Low	0.246	0.165	0.134	0.484	0.449	0.230	0.258	0.512	0.443
cross	10	Medium	0.198	0.108	0.099	0.243	0.239	0.172	0.187	0.291	0.262
cross	10	High	0.129	0.096	0.088	0.128	0.126	0.147	0.117	0.152	0.132
cross	20	Low	0.263	0.195	0.146	0.898	0.824	0.232	0.402	0.960	0.844
cross	20	Medium	0.214	0.142	0.116	0.529	0.464	0.178	0.248	0.659	0.467
cross	20	High	0.158	0.105	0.090	0.215	0.181	0.117	0.143	0.314	0.207
cross	35	Low	0.243	0.193	0.125	0.999	0.992	0.317	0.597	1.000	0.998
cross	35	Medium	0.210	0.150	0.114	0.837	0.742	0.184	0.347	0.932	0.771
cross	35	High	0.147	0.093	0.070	0.359	0.270	0.106	0.163	0.493	0.311
cross	50	Low	0.243	0.179	0.132	1.000	1.000	0.414	0.810	1.000	1.000
cross	50	Medium	0.201	0.137	0.097	0.956	0.909	0.174	0.444	0.991	0.921
cross	50	High	0.160	0.099	0.084	0.510	0.380	0.113	0.199	0.660	0.443
cross	75	Low	0.254	0.204	0.139	1.000	1.000	0.770	0.984	1.000	1.000
cross	75	Medium	0.210	0.164	0.114	0.998	0.989	0.284	0.684	1.000	0.995
cross	75	High	0.149	0.121	0.099	0.729	0.567	0.149	0.267	0.829	0.648
cross	100	Low	0.263	0.206	0.145	1.000	1.000	0.974	0.998	1.000	1.000
cross	100	Medium	0.209	0.167	0.122	1.000	0.999	0.339	0.851	1.000	1.000
cross	100	High	0.172	0.115	0.097	0.866	0.737	0.132	0.319	0.920	0.817
cross	150	Low	0.230	0.195	0.130	1.000	1.000	1.000	1.000	1.000	1.000
cross	150	Medium	0.220	0.152	0.110	1.000	1.000	0.576	0.988	1.000	1.000
cross	150	High	0.163	0.114	0.096	0.974	0.918	0.156	0.471	0.978	0.959
cross	500	Low	0.234	0.184	0.127	1.000	1.000	1.000	1.000	1.000	1.000
cross	500	Medium	0.210	0.154	0.114	1.000	1.000	1.000	1.000	1.000	1.000

Relationship	N	Noise	Pear	Kend	Spear	HHG	HSIC	Hoeff	dCor	MI	HHG-Pear
cross	500	High	0.165	0.102	0.083	1.000	1.000	0.739	0.998	1.000	1.000

Appendix C

Mutual Information Simulation

To further investigate the differences of three different ways to estimate Mutual Information, a Monte-Carlo simulation was carried out. The differences between the approaches to estimate Mutual Information are summarized as follows:

1. Mutual Information from this study: KDE to estimate the density functions.
    - Epanechnikov kernel.
    - SJPD algorithm to calculate the bandwidth.
2. Mutual Information (Makonnen, 2019): KDE to estimate the density functions.
    - Epanechnikov kernel.
    - Silverman's rule of thumb algorithm to calculate the bandwidth.
3. Mutual Information (de Siquiera Santos et al., 2014): histogram count approach to estimate the density functions.

For simplicity, Mutual Information from this study will be referred to as MI, Makonnen (2019) Mutual Information will be referred to as Makonnen's MI, and de Siquiera Santos et al. (2014) Mutual Information as Siquiera's MI. The simulation assessed the performance of the three methods in several conditions that can be found while doing research in Psychology. These conditions are the same as the ones described in the main document. The design of the simulation can be summarized as follows:

4. Sample size (8 levels): 10, 20, 35, 50, 75, 100, 150 and 500
5. Type of relationship (5 levels): linear, exponential, quadratic, sine wave, and non-functional cross.
6. Noise (3 levels): low, medium, high.

For a total of 8*5*3 = 120 design cells that were repeated 1000 times. The three tests were evaluated through all design cells.

Results

Complete Class

First, the dominance relationships between the three tests were analysed. The only dominance found was MI over Siquiera's MI (this can be seen in Figure 1C). It is

important to mention that in this case no tolerance level was added in the analysis. In other words, MI maintained a higher power than Siquiera's MI across all conditions. For the sake of the comparison, Siquiera's MI was not removed from the following analyses.

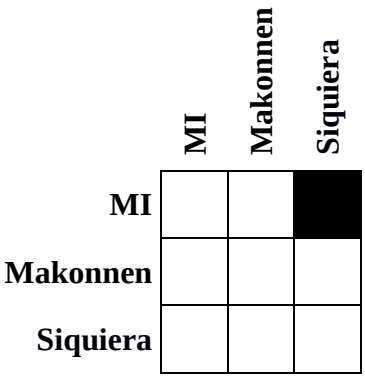

Figure 1C. Dominance relationships between the tests. The black cells indicate that the test in the corresponding row dominates the test in the corresponding column across all the simulated conditions.

Average Power

The average power across the type of relationship and noise level for each sample size can be seen in Table 1C. As expected, MI had the highest average power for all sample sizes. Aside from this, on one hand, Makonnen's MI had, on average, outperformed Siquiera's MI for small sample sizes ($N < 75$). On the other hand, Siquiera's MI had a better performance than Makonnen's MI when $N \geq 75$.

Table 1C

Average empirical power of three MI for each sample size, and the test with the maximum mean power.

N	MI	Makonnen	Siquiera	Max Test
10	.270	.170	.109	MI
20	.503	.328	.264	MI
35	.665	.483	.439	MI
50	.761	.583	.579	MI
75	.832	.683	.690	MI
100	.876	.743	.761	MI

150	.925	.806	.831	MI
500	.997	.957	.982	MI

Note. N = Sample size, MI = Mutual Information, Makonnen = Makonnen (2019) Mutual Information, Siquiera = de Siquiera Santos et al. (2014) Mutual Information. Max Test = Test with the maximum average power for each sample size.

Maximin

Table 2C shows the minimum power of the three Mutual Information estimates across the type of relationship and noise level for each sample size. Similar to the average power results, MI presented the best performance for all of the sample sizes. However, in this case, Makonnen's MI tied with MI in the smallest sample size (N=10). Moreover, Makonnen's MI also had the best performance than Siquiera's MI for five out of the eight simulated sample sizes.

Table 2C

The minimum power of five methods for each sample size, and test with the maximum over those minimums.

N	MI	Makonnen	Siquiera	Max Test
10	.053	.053	.042	MI and Makonnen
20	.078	.068	.056	MI
35	.095	.088	.076	MI
50	.135	.096	.094	MI
75	.167	.107	.117	MI
100	.229	.168	.147	MI
150	.418	.210	.233	MI
500	.965	.654	.776	MI

Note. N = Sample size, MI = Mutual Information, Makonnen = Makonnen (2019) Mutual Information, Siquiera = de Siquiera Santos et al. (2014) Mutual Information. Max Test = Test with the maximum average power for each sample size.

Furthermore, Table 3C shows the scenario for which each method presented its minimum power. The combination of a sine wave association and high noise conditions

appeared in almost all sample sizes for the three methods. The only exception was N = 20 for Makonnen's MI, for which the involved conditions were the quadratic relationship and high noise.

Table 3C

Conditions involved in the situation of minimum power of all dependence tests by sample size.

N	MI	Makonnen	Siquiera
10	SW-HN	SW-HN	SW-HN
20	SW-HN	Q-HN	SW-HN
35	SW-HN	SW-HN	SW-HN
50	SW-HN	SW-HN	SW-HN
75	SW-HN	SW-HN	SW-HN
100	SW-HN	SW-HN	SW-HN
150	SW-HN	SW-HN	SW-HN
500	SW-HN	SW-HN	SW-HN

Note. SW = Sine wave association, Q = Quadratic association, HN = High noise, N = Sample size, MI = Mutual Information, Makonnen = Makonnen (2019) Mutual Information, Siquiera = de Siquiera Santos et al. (2014) Mutual Information.

Finally, Table 4C shows the complete results of the simulation for each one of the 120 possible scenarios that were simulated. The results are organized by type of relationship, sample size, and noise level.

Table 4C

Complete results of the simulation for all 120 simulated scenarios.

Relationship	N	Noise	MI	Makonnen	Siquiera
linear	10	Low	0.816	0.444	0.308
linear	10	Medium	0.364	0.154	0.116
linear	10	High	0.123	0.081	0.047
linear	20	Low	0.995	0.814	0.721
linear	20	Medium	0.679	0.265	0.233
linear	20	High	0.231	0.090	0.077
linear	35	Low	1.000	0.985	0.964
linear	35	Medium	0.916	0.475	0.425
linear	35	High	0.349	0.140	0.123

Relationship	N	Noise	MI	Makonnen	Siquiera
linear	50	Low	1.000	0.999	0.999
linear	50	Medium	0.981	0.692	0.691
linear	50	High	0.506	0.155	0.173
linear	75	Low	1.000	1.000	1.000
linear	75	Medium	1.000	0.879	0.895
linear	75	High	0.674	0.255	0.307
linear	100	Low	1.000	1.000	1.000
linear	100	Medium	1.000	0.963	0.983
linear	100	High	0.833	0.384	0.411
linear	150	Low	1.000	1.000	1.000
linear	150	Medium	1.000	0.998	1.000
linear	150	High	0.906	0.465	0.581
linear	500	Low	1.000	1.000	1.000
linear	500	Medium	1.000	1.000	1.000
linear	500	High	1.000	0.983	1.000
quadratic	10	Low	0.365	0.201	0.134
quadratic	10	Medium	0.202	0.124	0.092
quadratic	10	High	0.076	0.063	0.049
quadratic	20	Low	0.834	0.433	0.401
quadratic	20	Medium	0.515	0.192	0.194
quadratic	20	High	0.154	0.068	0.066
quadratic	35	Low	0.982	0.750	0.729
quadratic	35	Medium	0.767	0.331	0.374
quadratic	35	High	0.190	0.090	0.080
quadratic	50	Low	0.999	0.916	0.939
quadratic	50	Medium	0.924	0.508	0.594
quadratic	50	High	0.294	0.107	0.107
quadratic	75	Low	1.000	0.994	0.997
quadratic	75	Medium	0.988	0.764	0.803
quadratic	75	High	0.396	0.130	0.161
quadratic	100	Low	1.000	0.998	1.000
quadratic	100	Medium	0.999	0.887	0.946
quadratic	100	High	0.493	0.171	0.209
quadratic	150	Low	1.000	1.000	1.000
quadratic	150	Medium	1.000	0.985	0.997
quadratic	150	High	0.714	0.265	0.357
quadratic	500	Low	1.000	1.000	1.000
quadratic	500	Medium	1.000	1.000	1.000
quadratic	500	High	0.994	0.795	0.960
exponential	10	Low	0.407	0.177	0.151
exponential	10	Medium	0.208	0.099	0.065

Relationship	N	Noise	MI	Makonnen	Siquiera
exponential	10	High	0.109	0.071	0.055
exponential	20	Low	0.790	0.419	0.388
exponential	20	Medium	0.441	0.165	0.167
exponential	20	High	0.162	0.079	0.090
exponential	35	Low	0.968	0.668	0.704
exponential	35	Medium	0.705	0.286	0.287
exponential	35	High	0.308	0.107	0.103
exponential	50	Low	0.996	0.842	0.907
exponential	50	Medium	0.867	0.425	0.520
exponential	50	High	0.412	0.148	0.168
exponential	75	Low	1.000	0.964	0.987
exponential	75	Medium	0.955	0.589	0.687
exponential	75	High	0.545	0.216	0.226
exponential	100	Low	1.000	0.996	0.999
exponential	100	Medium	0.992	0.730	0.858
exponential	100	High	0.688	0.250	0.352
exponential	150	Low	1.000	1.000	1.000
exponential	150	Medium	0.999	0.914	0.967
exponential	150	High	0.859	0.413	0.522
exponential	500	Low	1.000	1.000	1.000
exponential	500	Medium	1.000	1.000	1.000
exponential	500	High	0.999	0.927	0.996
Sine wave	10	Low	0.245	0.283	0.162
Sine wave	10	Medium	0.135	0.141	0.094
Sine wave	10	High	0.053	0.053	0.042
Sine wave	20	Low	0.510	0.690	0.409
Sine wave	20	Medium	0.245	0.250	0.197
Sine wave	20	High	0.078	0.068	0.056
Sine wave	35	Low	0.774	0.956	0.759
Sine wave	35	Medium	0.507	0.493	0.353
Sine wave	35	High	0.095	0.088	0.076
Sine wave	50	Low	0.946	0.999	0.921
Sine wave	50	Medium	0.725	0.697	0.564
Sine wave	50	High	0.135	0.096	0.094
Sine wave	75	Low	0.998	1.000	0.998
Sine wave	75	Medium	0.942	0.909	0.806
Sine wave	75	High	0.167	0.107	0.117
Sine wave	100	Low	1.000	1.000	0.999
Sine wave	100	Medium	0.991	0.982	0.929
Sine wave	100	High	0.229	0.168	0.147
Sine wave	150	Low	1.000	1.000	1.000

Relationship	N	Noise	MI	Makonnen	Siquiera
Sine wave	150	Medium	1.000	0.999	0.996
Sine wave	150	High	0.418	0.210	0.233
Sine wave	500	Low	1.000	1.000	1.000
Sine wave	500	Medium	1.000	1.000	1.000
Sine wave	500	High	0.965	0.654	0.776
cross	10	Low	0.522	0.398	0.187
cross	10	Medium	0.286	0.172	0.092
cross	10	High	0.151	0.094	0.049
cross	20	Low	0.955	0.854	0.605
cross	20	Medium	0.656	0.372	0.259
cross	20	High	0.314	0.163	0.104
cross	35	Low	1.000	0.991	0.932
cross	35	Medium	0.928	0.646	0.486
cross	35	High	0.487	0.246	0.193
cross	50	Low	1.000	1.000	0.997
cross	50	Medium	0.992	0.830	0.739
cross	50	High	0.650	0.332	0.286
cross	75	Low	1.000	1.000	1.000
cross	75	Medium	1.000	0.967	0.935
cross	75	High	0.826	0.479	0.440
cross	100	Low	1.000	1.000	1.000
cross	100	Medium	1.000	0.994	0.981
cross	100	High	0.916	0.636	0.608
cross	150	Low	1.000	1.000	1.000
cross	150	Medium	1.000	1.000	1.000
cross	150	High	0.979	0.853	0.816
cross	500	Low	1.000	1.000	1.000
cross	500	Medium	1.000	1.000	1.000
cross	500	High	1.000	0.999	1.000