

## Master Thesis Report

# Measures of statistical dependence for feature selection 

Computational Study

Mohamad Alshalabi
Supervision: Priyantha Wijayatunga
June 1, 2022

Mått på statistiskt beroende för funktionsval
beräkningsstudie


#### Abstract

The importance of feature selection for statistical and machine learning models derives from their explainability and the ability to explore new relationships, leading to new discoveries. Straightforward feature selection methods measure the dependencies between the potential features and the response variable. This thesis tries to study the selection of features according to a maximal statistical dependency criterion based on generalized Pearson's correlation coefficients, e.g., Wijayatunga's coefficient. I present a framework for feature selection based on these coefficients for high dimensional feature variables. The results are compared to the ones obtained by applying an elastic net regression (for high-dimensional data). The generalized Pearson's correlation coefficient is a metric-based measure where the metric is Hellinger distance. The metric is considered as the distance between probability distributions. The Wijayatunga's coefficient is originally proposed for the discrete case; here, we generalize it for continuous variables by discretization and kernelization. It is interesting to see how discretization work as we discretize the bins finer. The study employs both synthetic and real-world data to illustrate the validity and power of this feature selection process. Moreover, a new method of normalization for mutual information is included. The results show that both measures have considerable potential in detecting associations. The feature selection experiment shows that elastic net regression is superior to our proposed method; nevertheless, more investigation could be done regarding this subject.


## Popular scientific summary

Two random variables could be independent when the values of one variable are not related to those of the other variable's. If they are related mostly, then they are statistically dependent. Therefore, studying the underlying relationship structure of set of variables requires detecting the dependence among the variable using a valid statistical dependence measures. A good measure should be sensitive to the underlying relation. One of the two aims of this thesis is to present a computational proof for two dependence measure.
The second aim is to apply one of the discussed measure in the process of removing redundant variables from big sets of features. The process of removing features from a set is known as feature selection. The selection is needed to explore new relationships. Straightforward feature selection methods measure the statistical dependence between the features and the response variable. Various dependence measures are proposed for feature selection in machine learning and statistical modeling; this includes correlation measures such as Pearson's and Spearman's. The Pearson's correlation works perfectly for linear relationships, while Spearman's correlation performs better for non-linear but monotonic relationships.
This thesis tries to study the selection of the features based on the generalized Pearson's correlation coefficient (Wijayatunga's coefficient). The Wijayatunga's coefficient is a metric-based dependency measure employing a Hellinger distance. The Hellinger distance is a measure that quantifies the similarity between two probability distributions. The Wijayatunga's coefficient is proposed initially for the discrete case. Here, we generalize it for continuous variables. Moreover, a new normalization method for mutual information is presented and tested.
Moreover, comparisons with other measures are conducted. The study employs both synthetic and real-world data to illustrate the validity and power of these new measures. The results show that the new coefficients, namely Wijayatunga's coefficient and normalized mutual information, have high detection properties for non-linear associations, while the results of the feature selection procedure could be investigated further.

## sammanfattning

Vikten av funktionsval för statistiska modeller och maskininlärnings-modeller härrör från deras förklaringsbarhet och förmågan att utforska nya relationer, vilket leder till nya upptäckter. Enkla funktionsvalsmetoder mäter beroenden mellan funktionsoch svarsvariabler. Denna uppsats undersöker funktionsval enligt ett maximalt statistiskt beroendekriterium baserat på generaliserade Pearsons korrelationskoefficienter, t.ex. Wijayatungas koefficient. Denna studie presenterar ett ramverk för funktionsval baserat på Wijayatungas koefficient mellan en högdimensionell slumpvariabel X och en svarsvariabel Y. Resultaten jämförs med de som erhålls genom att tillämpa en elastisk nettoreglering. Den generaliserade Pearsons korrelationskoefficient är ett är ett metriskt M-baserat beroende mått som använder Hellinger-avståndsmätning. Måttet kan betraktas som avståndet mellan sannolikhetsfördelningar. Wijayatungas koefficient är ursprungligen föreslagen för det diskreta fallet; här generaliserar vi det för kontinuerliga variabler genom diskretisering och kernelisering. Det är intressant att se hur diskretisering fungerar när vi diskretiserar facken finare. uppsatsen använder både syntetiska och verkliga data för att illustrera giltigheten och kraften i denna funktionsvalsprocess. Dessutom ingår en ny metod för normalisering av ömsesidig information. Resultaten visar att båda åtgärderna har stor potential för att upptäcka samband. Funktionsvalsexperimentet visar att elastiskt nät är överlägset vår föreslagna metod; Ändå skulle mer undersökningar kunna göras i detta ämne.

## Contents

1 Introduction. ..... 1
2 Previous dependence measure ..... 2
2.1 Pearson's Product-Moment Correlation ..... 2
2.2 Distance Correlation ..... 4
2.3 Entropy and Mutual information ..... 6
2.3.1 A Normalized Mutual Information ..... 7
2.4 Maximal Information Coefficient ..... 8
3 New measures of degree of dependence ..... 8
3.1 Hellinger Distance ..... 9
3.2 Wijayatunga coefficient ..... 10
3.3 A new approach for normalization of the mutual information ..... 11
3.4 Kernel density estimation ..... 12
4 Simulations in R ..... 13
4.1 Dependence measurements ..... 14
4.2 Association models ..... 15
4.3 Noiseless simulations ..... 17
4.3.1 Coefficients average performance. ..... 18
4.3.2 The variability of the coefficients. ..... 19
4.4 The noise effect ..... 20
4.5 The effect of the sample size ..... 22
5 Feature Selection Method ..... 23
5.1 Elastic Net Regularization ..... 23
5.2 Minimal Redundancy Maximal Relevance ..... 25
5.3 Real Life Data Sets ..... 26
6 Conclusion ..... 28
7 References ..... 31

## List of Figures

1 Linear Relationship ..... 3
2 Pearson Correlation ..... 4
3 Distance correlation ..... 5
4 Wijayatunga coefficient ..... 10
5 Normalized Mutual Information ..... 12
6 Association Models ..... 15
$7 \quad$ The effect of adding noise to the linear relationship ..... 20
8 Noise effect on cubic relationships ..... 21
9 Noise effect on different non-linear relationships ..... 22
10 Elastic Net ..... 24

## List of Tables

1 Association functions ..... 16
2 Noiseless simulations' results ..... 18
3 Circular relationship ..... 22
4 Hyperbolic relationship ..... 23
5 Prostate Data ..... 26
6 Elastic net results ..... 27
7 Boston Housing Data ..... 27
8 Elastic net results 2 ..... 28

## 1 Introduction.

According to (Borovkov, 1998), two random variables, say, $X$ and $Y$, can either be dependent when the values of one variable are associated with those of the other, otherwise the variables are independent. The dependence can be deterministic when the variable $X$ is a function of the variable $Y$ or other way around, though the dependence need not be always deterministic. In the latter case, we say that it is a statistical dependence.

In order to dive into the subject of statistical dependence, we need to distinguish between independent and dependent random variables formally. Hastings (1997) defines independence among random variables as follow: The set of random variables, say, $\mathbf{X}=\left\{X_{1}, \ldots, X_{n}\right\}$ said to be mutually independent if, for any sub-collection of them, $\mathbf{X}_{a}=\left\{X_{i}, \ldots, X_{k}\right\}$ where, $\mathbf{X}_{a} \subseteq \mathbf{X}$, we have that

$$
P\left[X_{i} \in \mathcal{X}_{i}, \ldots, X_{k} \in \mathcal{X}_{k}\right]=P\left[X_{i} \in \mathcal{X}_{i},\right] \ldots P\left[X_{k} \in \mathcal{X}_{k}\right]
$$

where $\mathcal{X}_{j}$ is the state space of $X_{j}$. Random variables that don't satisfy this relationship are said to be dependent (Hastings, 1997). To put it in words, a given set of random variables are mutually independent if their joint probability density function is the product of the respective marginal probability density functions. Similarly, this is the caase for the probability mass function for the discrete random variables (Hastings, 1997; Wijayatunga, 2016).

Studying the underlying structure of a set of random variables requires detecting the dependence among them using a valid statistical dependence measure. A valid measure should be sensitive to the underlying relationship (Martínez-Gómez, Richards, and Richards, 2014). Moreover, evaluating and comparing different dependence measures with each other requires considering what reasonable and natural axioms should hold for any measure of dependence (Móri and Székely, 2018).
Even though Rényi (1959) developed seven important properties/axioms for a valid dependence measure, Móri and Székely (2018) reduced the number to four axioms. According to Móri and Székely (2018), for a nonempty set $S$ of pairs of generic random variables $X$ and $Y$ taking values either in Euclidean space or in real, separable Hilbert spaces $H$, then $\operatorname{Corr}(X, Y): S \rightarrow[0 ; 1]$ is called a dependence measure on $S=\mathcal{X} \times \mathcal{Y}$ if the following four axioms are held:

1. $\operatorname{Corr}(X, Y)=0$ if and only if X and Y are independent.
2. $\operatorname{Corr}(X, Y)$ is invariant with respect to all similarity transformations of H ; that is, $\operatorname{Corr}(A * X, B * Y)=\operatorname{Corr}(X, Y)$ where $A, B$ are similarity transformations of H .
3. $\operatorname{Corr}(X, Y)=1$ if and only if $Y=f(X)$ with probability 1 , where $f$ is a similarity transformation of H .
4. $\operatorname{Corr}(X, Y)$ is continuous; that is, if $\left(X_{n}, Y_{n}\right) \in S$ where $n=1,2, \ldots$ such that for some positive constant K we have $E\left(\mid\left(\left.X_{n}\right|^{2}+\mid\left(\left.Y_{n}\right|^{2} \leq K\right)\right.\right.$ and $\left(X_{n}, Y_{n}\right)$ converges weakly to $(X, Y)$ then $\operatorname{Corr}\left(X_{n}, Y_{n}\right) \rightarrow \operatorname{Corr}(X, Y)$.

Similarity of H is defined as a bijection ( $1-1$ correspondence) from H onto itself that multiplies all distances by the same positive real number (scale). Note that similarity transformation is simply change of measurement.
This thesis is a computational study of a new dependence measure, which was introduced in Wijayatunga (2016). This measure is similar to the one proposed in (Granger, Maasoumi and Racine, 2004). The differences between them are discussed later. Furthermore, following Wijayatunga (2016), we introduce a new approach for normalizing the mutual information measure. The new method of normalization seems to be more powerful in detecting associations in comparison to other mutual information based measures.
The thesis is organized as follows. Section 2 covers the previous dependence measures and how they detect dependence. Section 3 is dedicated to the new measure that is based on probability distance called the Hellinger distance. After that, we combine the knowledge provided in Sections 2 and 3 to introduce a new normalized mutual information measure. At the end of Section 2, a small review of Kernel density estimation is provided, which is a vital part of estimating the Wijayatunga coefficient for continuous cases. Section 4 is devoted to a simulation study. Section 5 introduces elastic net regularization; moreover, a real-life data set is included where a comparison between the performance of elastic net and one of the new measures are done. The new measures select features in the data set using backward elimination algorithm. Finally, a section of discussion and conclusions concludes the thesis.

## 2 Previous dependence measure

One can describe correlation as the degree of association between two variables. Generally speaking, the study of interdependence brings the subject of investigating the correlation. Several different kinds of correlation coefficients handle the special characteristics of the variables such as dichotomies, and other measurements of association, deal with nominal and ordinal variables (Asuero, Sayago and González, 2006). In this section, we cover some of the most prominent correlation measures.

### 2.1 Pearson's Product-Moment Correlation

A random variable $X$ can be reduced to a standardized variable through linear transformation where

$$
X^{\prime}=\frac{X-\bar{X}}{\sqrt{\operatorname{Var}(X)}}
$$

The Pearson's correlation coefficient $\rho$ of two random variables $X$ and $Y$ is calculated using the formula $\rho(X, Y)=E\left(X^{\prime}, Y^{\prime}\right)$ (Borovkov, 1998). Let's build the aforementioned coefficient $\rho(X, Y)$ from scratch; first, we define covariance as the expected value of the product of the differences of each random variable from their expected values (Hastings, 1997).

$$
X, Y=\mathrm{E}\left[\left(X-\mu_{X}\right)\left(Y-\mu_{Y}\right)\right]
$$

The $\rho$ is defined as covariance normalization (Hastings, 1997); the following equation presents the Pearson correlation coefficient.

$$
\rho(X, Y)=\mathrm{E}\left[\left(\frac{X-\mu_{X}}{\sigma_{X}}\right)\left(\frac{Y-\mu_{Y}}{\sigma_{Y}}\right)\right]
$$

The different formulas in which Pearson's coefficient could be presented suggest that there are various ways to understand the coefficient (Rodgers and Nicewander, 1988). The geometric representation of $\rho$ is of a great importance for this thesis.
Pearson's coefficient describes perfectly the linear relationship between two variables, as we can see in figure 1 where the coefficient is able to detect the full linearity dependence. At the same time, this accuracy in describing linearity has no power in explaining any other type of correlations. The $\rho$ provides no information about the underlying relationship if the two variables are non-linearly correlated.

Linear Relationship


Figure 1: Small simulated sample with number of observation $n=100$

Calculating $\rho$ requires assumptions regarding the normality of the distribution of the variables (Hastings, 1997; Myers and Well, 2003). Those disadvantages can be overcome by using a non-parametric measure of rank correlation, namely, Spearman's rank correlation coefficient $r$. It's calculated as follows, let's sort the observations of a random variable X ascendingly. The second step is to rank those observations; then, let's rank the observations of the Y variables accordingly. The difference in ranks for each pair represents the main component of the coefficient. The formula for the coefficient is

$$
r_{s}=1-\frac{6 \Sigma d_{i}^{2}}{n\left(n^{2}-1\right)}
$$

where $d_{i}$ is the difference in rank between $X_{i}$ and $Y_{i}$ and $n$ is the number of observations. Spearman's rank coefficient $r$ performs better than $\rho$ in case on monotonic non-linear relationships. Furthermore, as opposed to Pearson's correlation coefficient, Spearman's coefficient can also be used with ordinal variables. A drawback of Spearman's is its limitation
when dealing with a non-monotonic relationship(Myers and Well, 2003). Replacing Pearson's coefficient Spearman's coefficients could solve some of Pearson's drawbacks, yet, the rank coefficient requires monotonic associations (Myers and Well, 2003). Figure 2 illustrates the power and the shortcomings of $\rho$.


Figure 2: An illustration of the correlation coefficient for different type of relationships. Pearson's correlation coefficient is reported above each scatter-plot
This figure is created using an altered version of the code in (Wikimedia Commons, 2022)

### 2.2 Distance Correlation

Distance correlation measures the dependence between random vectors analogous to the product-moment correlation $\rho$; similarly, distance covariance is equivalent to product-moment covariance. These distance measurements are based on certain Euclidean distances between the sample elements rather than the sample moments(Székely, Rizzo, and Bakirov, 2007). There are two main fundamental aspects where the distance correlation generalizes the correlation for all distributions with finite first moments:

1- $R(X, Y)$ is defined for $X$ and $Y$ in arbitrary dimensions.
2- $R(X, Y)=0$ characterizes independence of $X$ and $Y$.

Distance correlation is a metric, which means it satisfies the properties of the true dependence measure. Moreover, in a case of bivariate correlation, R is a function of $\rho$ and $R(X, Y) \leq$ $|\rho(X, Y)|$. In the bivariate normal case, $R$ is a function of $\rho$ and $R(X, Y) \leq|\rho(X, Y)|$ with equality when $\rho= \pm 1$.
The distance covariance performs significantly better against the non-monotone dependencies; at the same time, it holds good power performance in the multivariate normal case compared to the parametric likelihood ratio test(Lyons, 2013). Furthermore, distance cor-
relation can replace the Pearson correlation without requiring normality assumptions to validate the inference. The figure 3 shows how much better distance correlation performs compared to Pearson's $\rho$.
Distance correlation










Figure 3: We see that the distance correlation is detecting the existing of a relationship more accurately than $\rho$. The values above each shape reports the distance correlation coefficient. This figure is created using the an altered version of the code in (Wikimedia Commons, 2022)

The distance covariance between random vectors X and Y with finite first moments is the non-negative number defined as:

$$
d \operatorname{Cov}^{2}(X, Y)=\frac{1}{n^{2}} \sum_{i=1}^{n} \sum_{j=1}^{n} D\left(X_{i}, X_{j}\right) \cdot D\left(Y_{i}, Y_{j}\right)
$$

In the equation above, we replace the inner product of two centered vectors used in Pearson covariance with centered Euclidean distances. Similarly, we define the variance as follow:

$$
\begin{array}{r}
d \operatorname{Var}^{2}(x)=d \operatorname{Cov}^{2}(X, X)=\frac{1}{n^{2}} \sum_{i=1}^{n} D\left(X_{i}, X_{j}\right) \\
d \operatorname{Var}^{2}(Y)=d \operatorname{Cov}^{2}(Y, Y)=\frac{1}{n^{2}} \sum_{i=1}^{n} D\left(Y_{i}, Y_{j}\right)
\end{array}
$$

The distance correlation is calculated using an equation that mimics Pearson's correlation coefficient equation.

$$
d \operatorname{Cor}(X, Y)=\frac{d \operatorname{Cov}(X, X)}{\sqrt{d \operatorname{Var}(x) d \operatorname{Var}(Y)}}
$$

What is left to be done is to calculate an $(N * N)$ distance matrices for each random vector; then, those matrices need to be double centered.

$$
\begin{aligned}
X_{i, j} & =\left|x_{i}-x_{j}\right| \quad \bar{X}_{i}=\frac{1}{n} \sum_{i=1}^{n} x_{i} . \\
\bar{X}_{j} & =\frac{1}{n} \sum_{j=1}^{n} x_{i} . \quad \overline{X_{i, j}}=\frac{1}{n^{2}} \sum_{j=1}^{n} \frac{1}{n} \sum_{i=1}^{n} x_{i, j} \\
D\left(X_{i}, X_{j}\right) & =X_{i, j}-\bar{X}_{i}-\bar{X}_{j}+\overline{X_{i, j}}
\end{aligned}
$$

The Distance correlation measurement suffers from a few drawbacks; the most obvious flaw is the computational cost, where the basic algorithm to compute the test statistic is of order $O\left(n^{2}\right)$ in the sample size as opposed to $O(n)$ for product-moment correlation. Moreover, the test statistic is not distribution-free, even asymptotically. The distribution under the null depends on the underlying distribution of $X$ and $Y$ even as the sample size tends to infinity. Indeed, the distributions are uniformly bounded by a $\chi^{2}$ distribution, which allows for the calculation of a conservative critical value.

### 2.3 Entropy and Mutual information

We define the information regarding the occurrence of an event using the probability of its occurrence. This "self-information" is inversely proportional to the probability of occurrence. Low probability event has high information, i.e., it is surprising. While a high probability event has low information, i.e., we are not surprised if this event happens. (Borovkov, 1998) the information in an event A can be calculated as follow:

$$
I(A)=-\log P(A)
$$

In his groundbreaking paper "A Mathematical Theory of Communication" in 1948, Claude Shannon presented the concept of entropy. According to the information theory, entropy measures the average amount of information needed to represent a possible outcome drawn from a random variable's probability distribution. For a discrete random variable $X$ with the state space $\mathcal{X}=\left\{x_{1}, \ldots, x_{n}\right\}$ and probability mass function $p(X=x)=p(x)$, we define the entropy as follows:

$$
\eta(X)=-\sum_{i=1}^{n} p\left(x_{i}\right) \log _{b} p\left(x_{i}\right)
$$

Where $b$ represents the base of the logarithm used. Common values of $b$ are 2, Euler's number e, and 10 , and the corresponding units of entropy are the bits for $b=2$, nats for $b$ $=\mathrm{e}$, and bans for $\mathrm{b}=10$. We can define the conditional entropy of two discrete variables $X$ and $Y$, which take values $x_{1}, \ldots . x_{k}$ and $y_{1}, \ldots . . y_{l}$ respectively as follow:

$$
\eta(X \mid Y)=-\Sigma_{i, j} p\left(x_{i}, y_{j}\right) \log \frac{p\left(x_{i}, y_{j}\right)}{p\left(y_{j}\right)}
$$

where $p(x, y)$ represents the joint probability distribution, while $p(y)$ is the preserved marginal
of $Y$ and similary for $p(x)$.
The mutual information (MI) of two random variables measures the mutual dependence between the two random variables. The mutual information is linked to the entropy of random variables mentioned above. The mutual information quantifies the deviance between the joint probability distribution of two random variables, $p_{X Y}(x, y)$ and the product of their marginals $p_{X}(x)$ and $p_{Y}(y)$, i.e., the kullback-Leibler deviance between the dependence and independence (Wijayatunga 2016). This concept is connected to our new coefficient, which will be explained later in this Chapter. We calculate the mutual information between two discrete random variables $X$ and $Y$ as a double sum.

$$
M I(X ; Y)=\sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} p_{X Y}(x, y) \log \left(\frac{p_{X Y}(x, y)}{P_{X}(x) P_{Y}(y)}\right)
$$

Similarly, we replace the sums with integrals when dealing with continuous random variables.

$$
M I(X ; Y)=\int_{y \in \mathcal{Y}} \int_{x \in \mathcal{X}} p_{X Y}(x, y) \log \left(\frac{p_{X Y}(x, y)}{p_{X}(x) p_{Y}(y)}\right)
$$

As mentioned above, we can represent this measure in terms of entropy:

$$
\begin{aligned}
& M I(X ; Y)=\eta(X)-\eta(X \mid Y) \\
& M I(X ; Y)=\eta(Y)-\eta(Y \mid X)
\end{aligned}
$$

Therefore, it is symmetric similar to Pearson's, Spearman's and distance correlation. The entropy measure and the mutual information fail to satisfy the metric conditions as they violate either the symmetry property or the triangle inequality property, or both. Those measures are rather measures of divergences but metric distances. (Granger, Maasoumi and Racine, 2004).

### 2.3.1 A Normalized Mutual Information

Overcoming the metric issue mentioned above requires finding a way to normalize the mutual information to bound into the closed interval $[0,1]$. There are many suggested normalization approaches. Proficiency is an information-theoretic measure developed by White, Steingold and Fournelle in 2004. It is a scalar measure, based on mutual information and entropy.

$$
\begin{aligned}
C_{X Y} & =\frac{M I(X ; Y)}{\eta(Y)} \\
C_{Y X} & =\frac{M I(X ; Y)}{\eta(X)}
\end{aligned}
$$

This metric has a shortcoming due to its being non-symmetrical nature where $C_{X Y}=C_{Y X}$ could be violated. A replacement is a symmetrical measure. there different suggested measures. The symmetric uncertainty is a measure developed by Witten and Frank in 2005.

$$
U(X, Y)=2 \frac{M I(X ; Y)}{\eta(X) \eta(Y)}
$$

A new approach to normalization will be discussed later on.

### 2.4 Maximal Information Coefficient

Before moving to the new coefficient, which is the goal of this thesis, we must mention the maximal information coefficient (MIC). The MIC is a maximal information-based nonparametric exploratory statistic that identifies and classifies relationships (Reshef et al., 2011). MIC estimates the relationship between two variables by finding the grid that maximize the resolution on the scatter-plot of those two variables (Reshef et al., 2011). The normalized largest possible mutual information achievable by any x-by-y grid builds up a matrix $M$. The statistic MIC is the maximum value in M (Reshef et al., 2011).

$$
M I C(X, Y)=\max _{n_{x} \times n_{y}} \frac{M I_{n_{x} \times n_{y}}(X ; Y)}{\log \left(\min \left(n_{x}, n_{y}\right)\right)}
$$

$n_{x}$ is the number of bins on $x$-axis and similarly for $n_{y}$. Therefore $n_{x} \times n_{y}$ is a grid over the plotted data so that the mutual information under the grid $n_{x} \times n_{y}$ is represented by $M I_{n_{y} \times n_{y}}(X ; Y)$. The normalization of the maximum value of $M I_{n_{x} \times n_{y}}(X ; Y)$ over all $n_{x} \times n_{y}$ is a result of the following an inequality

$$
0 \leq M I_{n_{x} \times n_{y}}(X ; Y) \leq \min \left(n_{x}, n_{y}\right)
$$

That is, MIC is the normalized maximal mutual information over all possible grids of the data.

## 3 New measures of degree of dependence

The Pearson coefficient is simply a normalized distance between the joint probability distribution of two variables and its distribution when the independence of two variables is assumed, where the normalizing constant is the geometric mean of the two maximal distances. The Pearson coefficient uses a Euclidean-type distance as the distance (Wijayatunga, 2016). This has caused that it can only measure linear associations. Those distributions could be seen as contingency tables representing the frequencies of values for each variable (Fienberg and Gilbert, 1970).
A new dependence measure that is proposed in Wijayatunga (2016) generalizes the Pearson's coefficient using a more general metric distance, namely, Hellinger distance. It utilizes the idea of the distance between a given joint probability distribution of two variables and their joint probability distribution when their independence is assumed. This distance is normalized by the geometric mean of similar distances related to all possible maximal dependencies while preserving only one of the marginals at each time. Many types of dependencies emerge in the case of multinary (discrete) variables; those types of dependencies are not easy to allocate using only a single component or a weighted average of differences. Suppose the weighted average difference is a Euclidean type distance, which can measure only linear dependencies. The inadequacy of those measurements pushes us to find a more suitable distance to measure any non-linear dependencies.

Wijayatunga (2016) proposed using Hellinger metric distance as a distance measure. Therefore, it is argued that it can measure the degree of any type of dependence between two discrete variables. The continuous variables can be discretized appropriately first to measure their dependence. Note that in the discretization process there can be some information loss that should be minimized. However, even though the coefficient is based on a metric distance, it is an open problem to show that resulting coefficient satisfies the metric properties.

### 3.1 Hellinger Distance

Hellinger is an $f$-divergence that quantifies the similarity between two probability distributions. Moreover, Hellinger's type of distance is an $L_{2}$ type of distance. The Hellinger distance is defined as the Hellinger integral (Van der Vaart, 2007). Integral is replaced by the summation for discrete variables. Let $P$ and $Q$ denote two discrete probability measures that their Hellinger distance is defined as

$$
h(P, Q)=\frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{k}\left(\sqrt{p_{i}}-\sqrt{q_{i}}\right)^{2}}
$$

where $p$ is the probability mass function of the probability measure $P$ such that $\sum_{i} p_{i}=1$, and similarly for $q$. This is directly related to the Euclidean norm of the difference of the square root vectors

$$
h(P, Q)=\frac{1}{\sqrt{2}}\left\|\sqrt{p_{i}}-\sqrt{q_{i}}\right\|_{2}
$$

The Hellinger distance forms a bounded metric on the space of probability distributions for given set of random variables. The maximum distance is equal to 1 , which is attainable when P assigns probability zero to every set to which Q assigns a positive probability and vice versa(Van der Vaart, 2007). The Hellinger distance is related to the Bhattacharyya coefficient $B C(P, Q)$ as it can be defined as

$$
h(P, Q)=\sqrt{1-B C(P, Q)}
$$

Hellinger distance satisfies the following properties:

1. $0 \leq h(P, Q) \leq 1$
2. $h(P, Q)=1$ if and only if the measures $P$ and $Q$ are mutually singular, i.e. They share no space.
3. $h(P, Q)=0$ if and only if $P=Q$, i.e they shape the same curve.
4. $h(P, Q)=h(P(x+\alpha), Q(x+\alpha))$ for any constant.
5. $h(P, Q)=h(P(x * \alpha), Q(x * \alpha))$ for any constant $\alpha \neq 0$.

The last two are called the linear invariance properties of the probability metric (Wijayatunga, 2016).

### 3.2 Wijayatunga coefficient

For two discrete variables, $X$ and $Y$, with joint probability measure $P$ the Hellinger distance is metric in their joint probability simplex, i.e., follows the properties mentioned above; a measure of dependency between those variables that is based on Hellinger distance is defined as

$$
\begin{equation*}
\rho^{h}(X, Y)=\frac{h\left(P^{I}, P\right)}{\left.\prod_{P^{X} \in P_{\max }^{X}}\left[h\left(P^{I}, P^{X}\right)\right]^{\frac{1}{P_{\text {max }} \mid}} \prod_{P^{Y} \in P_{\max }^{Y}} h\left(P^{I}, P^{Y}\right)\right]^{\frac{1}{\left|P_{\max }\right|}}} \tag{1}
\end{equation*}
$$

where $P$ is the joint probability measure of $X$ and $Y, P^{I}$ is the joint probability measure when the independence of $X$ and $Y$ is assumed, and $P^{X}$ is the joint probability measure when maximal dependence is assumed while the marginal of $X$ is preserved, similarly, for $P^{Y}, P_{\max }^{X}$ is the set of all probability measures of maximal dependence of $X$ and $Y$ while marginal of $X$ is preserved, and similarly for $P_{\text {max }}^{Y}$. Here, $|A|$ denotes the cardinality of the set $A$. Finally, $h(P ; Q)$ is the (Hellinger) distance metric between two probability measures $P$ and $Q$.


Figure 4: An illustration of the performance of Wijayatunga coefficient in the continuous case for different type of relationships. The function used here employs kernel estimation to calculate the joint probability distribution The coefficient values is reported above each scatter-plot This figure is created using the an altered version of the code in (Wikimedia Commons, 2022)

Wijayatunga generalizes the coefficient for two continuous variables in his discussion on the paper "Sparse graphs using exchangeable random measures" by Caron, F. and Fox, E. in 2017. He states that Hellinger distance requires a normalizing constant of 1. The result is
a measure defines as follows:

$$
\begin{equation*}
\rho^{h}(X, Y)=\left(\frac{1}{2} \iint_{x, y}[\sqrt{f(x, y)}-\sqrt{f(x) f(y)}]^{2} d x d y\right)^{\frac{1}{2}} \tag{2}
\end{equation*}
$$

where $f_{x, y}$ represents the joint probability density function, furthermore, $f_{x}$ and $f_{y}$ are the marginals probability densities.

This function is similar to the one defined in (Granger, Maasoumi and Racine, 2004). The difference can be spotted easily. In the paper mentioned,(Granger, Maasoumi and Racine, 2004) define their measurement as generalization of Bhattacharya-Matusita-Hellinger measure of dependence given by the function:

$$
S_{p}(X, Y)=\frac{1}{2} \iint_{x, y}[\sqrt{f(x, y)}-\sqrt{f(x) f(y)}]^{2} d x d y
$$

This is a metric entropy measure that satisfies distance metric properties. It performs well with non-linear relationships. Both measures aim to estimate the divergence of the joint distribution of two variables from the product of their marginals.

### 3.3 A new approach for normalization of the mutual information

The Wijayatunga coefficient uses the Hellinger distance to examine the statistical separation of two joint probability distributions, one representing the dependence and the other representing the independence. Then, it uses similar type of distance to normalize the measure by considering the space between the product of marginals and the joint probability distribution that portrays maximal dependence while preserving the marginal of $X$ or $Y$. Using the same analogy, for two discrete random variables $X$ and $Y$, we can write a normalized version of their mutual information as follows:

$$
\begin{equation*}
N M I(X, Y)=\frac{M I(X ; Y)}{\left.\prod_{P^{X} \in P_{\text {max }}^{X}}\left[M I_{X}(X ; Y)\right]^{\frac{1}{P_{\text {max }} \mid}} \prod_{P^{Y} \in P_{\text {max }}^{Y}} M I_{Y}(X ; Y)\right]^{\frac{1}{P_{\text {max }}^{Y} \mid}}} \tag{3}
\end{equation*}
$$

where $M I_{X}(X ; Y)$ is the mutual information between $X Y$ calculated when joint probability distribution portrays maximal dependence while preserving the marginal of $X$ and similarly for $M I_{Y}(X ; Y)$. As mentioned before, mutual information quantifies the divergence of joint probability distribution and the product of the marginals of $X$ and $Y$. If a similar divergence between the marginal product and the "jpd" that portrays maximal dependence while preserving the marginal of $X$ or $Y$ is calculated, those divergences could be used to normalize the mutual information.


Figure 5: An illustration of the building blocks of the suggested normalized mutual information.
Upper left grid is the joint probability distribution. Lower left grid is the product of the marginals of $X$ and $Y$.
Upper right is the joint probability distribution that portrays maximal dependence while preserving the marginal of $X$
Lower right is the joint probability distribution that portrays maximal dependence while preserving the marginal of $Y$
Every mutual information in the figure measure the divergence from the product of the marginals.

Figure 5 shows how the different "jpd"s interact; note that there could be more than one $P_{X}(X ; Y)$ as well as $P_{Y}(X ; Y)$. As shown in figure 5 , it is easy to see that this coefficient is computational rather than analytical, i.e., in order to calculate it a computer code is necessary.This is the case for MIC too. Notably, most of the estimation tasks in statistics are increasingly becoming computation nowadays, e.g., high-dimensional regression, etc.

### 3.4 Kernel density estimation

Firstly, we use a histogram to represent the probability distribution of a random variable through approximation of the frequency in each bin. Bins are data structures that allow efficient region queries. Every time a value occurs in a specific bin, the bin's frequency increments by one (Jones, Marron and Sheather, 1996). A bi-dimension histogram is a generalization of the histogram where we expand the bins to a grid of cells; each cell has a size of bandwidth ${ }_{X} \times$ bandwidth $_{Y}$. The bandwidth is a real positive value, which describes the smoothness of a density plot. Furthermore, this number is a free parameter, in which its value cannot be predicted precisely and needs to be estimated; this gives the bandwidth
a strong influence on the results of the estimations (Kroese, Taimre and Botev, 2011). The frequency in each grid cell is the number of data points that their x and y coordinates fall in that cell. Removing the binning grid and estimating the joint probability density function is the goal of kernel density estimation.

The density estimation is the construction of the probability density function of a population using an observed sample (Rosenblatt, 1956). The goal here is to infer the probability density function using a finite sample of data. One of the methods to estimate density is to use the kernel function. Kernel density estimation is a non-parametric way to approximate the probability density function of a random variable (Rosenblatt, 1956). It is a smoothing technique that generalizes a histogram estimation with more robust statistical properties (Kroese, Taimre and Botev, 2011).
Let $\left\{x_{1}, \ldots, x_{n}\right\}$ be a sample of d-variate observed data drawn from a common distribution described by the density function $f$. The kernel density estimate is defined to be

$$
\hat{f}_{H}(X)=\frac{1}{n} \sum_{i=1}^{n} K_{H}\left(x-x_{i}\right)
$$

where
1- $x=x_{1}, x_{2}, \ldots, x_{d}^{T}, x_{i}=x_{i 1}, x_{i 2}, \ldots, x_{i d}^{T}, i=1,2, \ldots . n$ are d-vectors.
2- $H$ is the bandwidth (or smoothing) $d \times d$ matrix which is symmetric and positive definite.
3- $K$ is the kernel function which is symmetric multivariate density.
4- $K_{H}(x)=|H|^{-1 / 2} K\left(H^{-1 / 2} x\right)$
The bandwidth choice is very crucial in creating the kernel estimation. A poor choice of the bandwidth might lead to undesired transformations of the density plot:
1- A small bandwidth leads to under smoothing.
2- A vast bandwidth leads to over smoothing.

There are good reasons to expect that this approximation would be poor. First comes the sensitivity of the kernel-based tests to the choice of bandwidth. Moreover, the asymptotic variance is very poorly estimated in the case of the Hellinger and mutual information measures (Skaug and Tjøstheim 1993). In this study, two different bandwidth selectors are used, the first one is Silverman's rule of thumb (Silverman, 1998), the second one utilizes the SAMSE pilot bandwidths (Duong and Hazelton, 2003; Chacón and Duong, 2009). Bivariate bandwidth selection is a complex problem; the complexity can be overcome by enforcing some constraints on H . These constraints come at the expense of flexibility. The choice of diagonal bandwidth could be adequate in some cases; in other cases, it is more informative to select a full-bandwidth matrix (Duong and Hazelton, 2003).

## 4 Simulations in R

The comparison of different measures of dependence requires testing them under different scenarios. One way to do that is through simulations. An extensive simulation are performed. R programming language has the tools necessary to facilitate the simulations. Two main aspects are tested, the size of the sample and the effect of the Gaussian noise.

### 4.1 Dependence measurements

This thesis aims to examine the performance of the Wijayatunga coefficient and the normalized mutual information in various cases and models. There are no built-in functions for this coefficient; therefore, new functions are written in R. Four different versions of the coefficient are included.
The first one is the representation of the measurement for discrete variables. It produces the measure in equation 1. It can be used for continuous cases due to the inclusion of a discretization method. Across this paper, the discretization is performed using the function discretize from R package infotheo, which uses the equal width binning algorithm. Moreover, the number of bins is equal to $\sqrt{n}$ where $n$ is the number of observations. The function is named W.D.
The second one represents the Wijayatunga coefficient for the continuous variables represented in equation 2. It utilizes the idea of a two-dimensional histogram. We can easily show that this function results from binning the data into fine bins. The function is named W.C.H. The function works by assigning each observation to a square made by the intersection of the intervals on the $X$-axis and $Y$-axis. This differs from the kernel functions, where the kernel tries to smooth the edges of the square for more accuracy in allocating the observations.

The third function serves the same purpose as the previous one. It differs because of the usage of a self-written kernel density estimation, The function is namedW.C.KDE. The function works by estimating the kernel for the joint probability distribution. I use the joint probability distribution to find the marginals and their product. The function suffers due to the choice of bandwidth. The bandwidth is chosen by using a bandwidth selector function named bandwidth.nrd.1; this selector is built upon Silverman's rule of thumb (Silverman, 1998).

The fourth function takes advantage of the new R package ks which includes more advanced kernel density estimation functions and bandwidth selectors. The function is named W.C.KDE.2. While this function produces more accurate results than the previous function, it has three disadvantages. Those flaws are speed, minimal and negligible negative density estimates, and the inability to estimate the "jpd" for full dependent variables because it produces an eigenvalue that is zero or very close to zero. Two of those shortcomings are avoided, and the small negative values within the "jpd" are replaced with zero. To estimate the "jpd" for full dependent variables, a small noise was added to the predictor to remove zero or very close to zero eigenvalues. The results shown in the next figure depict the value of the measurement for different scenarios of association.

Lastly, Two functions are written to investigate the normalization of mutual information; they employs the equation 3. The first of those two functions is a simple direct function; this function discretizes the data and calculates both MI and NMI from an obtained "jpd." The second function uses Laplace smoothing to handle the problem of zero probability estimation

### 4.2 Association models

Studying the behavior of those measures of dependence is done by simulating data sets of $n$ observations of pair of random variables $(X, Y)$ according to the following sixteen models. The models are constructed so that a number of observations $n=1000$ could be fit in a box of $(-1,1) \times(-1,1)$ very often; this helps in the comparison between different level of noise. The figure below shows those models without any noise.


Figure 6: Scatter plots of one simulation from the all 16 models included. The number of observations for each model $\mathrm{n}=1000$ without any noise $\sigma=0$

In the Table 1, a list of functions are listed. Each row depicts one type of association illustrated in the figure above. Two main characters will vary throughout the simulation, the number of observations $n$ and the noise added which has a mean $\mu=0$ and variance $\sigma^{2}$. The value of $\sigma^{2}$ ranges between 0 and 0.09

Association functions

| Type of dependence | X | Y | Comment |
| :---: | :---: | :---: | :---: |
| Independence Case | $x_{i} \sim N(0,0.1)$ | $y_{i} \sim N(0,0.1)$ | $i=1, \ldots . n$ |
| Linear Dependence | $x_{i} \sim N(0,0.1)$ | $y_{i}=x_{i}+\epsilon_{i}$ | $\begin{aligned} & \epsilon_{i} \sim N\left(0, \sigma^{2}\right) \\ & i=1, \ldots . n \end{aligned}$ |
| Cubic Dependence | $x_{i} \sim N(0,0.1)$ | $y_{i}=x_{i}^{3}+1 / 3 x_{i}+\epsilon_{i}$ | $\begin{aligned} & \epsilon_{i} \sim N\left(0, \sigma^{2}\right) \\ & i=1, \ldots . n \end{aligned}$ |
| Quadratic Dependence | $x_{i} \sim N(0,0.1)$ | $y_{i}=x_{i}^{2}-1+\epsilon_{i}$ | $\begin{aligned} & \epsilon_{i} \sim N\left(0, \sigma^{2}\right) \\ & i=1, \ldots . n \end{aligned}$ |
| Sinusoidal <br> Dependence | $x_{i} \sim N(0,0.1)$ | $y_{i}=\sin 9 * x_{i}+\epsilon_{i}$ | $\begin{aligned} & \epsilon_{i} \sim N\left(0, \sigma^{2}\right) \\ & i=1, \ldots . n \end{aligned}$ |
| Cross-Shaped <br> Dependence | $\begin{aligned} x_{i} & \sim N(0,0.1) \\ x_{i} & \sim N\left(0, \frac{\sigma}{4}\right) \end{aligned}$ | $\begin{aligned} y_{i} & \sim N\left(0, \frac{\sigma}{4}\right) \\ y_{i} & \sim N(0,0.1) \end{aligned}$ | $\begin{aligned} i & =0, \ldots . k \\ i & =k+1, \ldots . n \\ k & \sim N(n, 0.5) \end{aligned}$ |
| Circular <br> Dependence | $x_{i}=-l_{i} * \cos k_{i}$ | $y_{i}=-l_{i} * \sin k_{i}$ | $\begin{aligned} & l_{i} \sim N\left(1, \sigma^{2}\right) \\ & k_{i} \sim N(0,1) \\ & i=1, \ldots, n \end{aligned}$ |
| Two Functions | $\begin{aligned} & x_{i}=2 / 3 k_{i}-1 \\ & x_{i}=-\left(-k_{i}\right)^{\cdot 1}+.1 \end{aligned}$ | $\begin{aligned} & y_{i}=x_{i}+\epsilon_{i} \\ & y_{i}=\left(x_{i}-.1\right)^{10}+1+\epsilon_{i} \end{aligned}$ | $\begin{aligned} & k_{i} \geq 0 \\ & k_{i}<0 \\ & k_{i} \sim N(0,1) \\ & \epsilon_{i} \sim N\left(0, \sigma^{2}\right) \\ & i=1, \ldots, n \end{aligned}$ |
| Checkerboard <br> Dependence | $x_{i}=k_{i 0}$ | $y_{i}=k_{i 1}+\epsilon_{i}$ | $\begin{aligned} & \epsilon_{i} \sim n\left(0, \frac{\sigma}{9}\right) \\ & i=1, \ldots . n \\ & k_{i 0}, k_{i 1} \sim N(\mu, \Sigma) \\ & {\left[3 k_{i 0}\right]-\left[3 k_{i 1}\right] \equiv 0} \end{aligned}$ |
| Exponential Function | $x_{i} \sim N(0,0.1)$ | $y_{i}=e^{2 x}+\epsilon_{i}$ | $\begin{aligned} & \epsilon_{i} \sim N\left(0, \sigma^{2}\right) \\ & i=1, \ldots . n \end{aligned}$ |
| Logarithmic Dependence | $x_{i} \sim N(0,0.1)$ | $y_{i}=\log \left(\left\|x_{i}\right\|\right)+\epsilon_{i}$ | $\begin{aligned} & \epsilon_{i} \sim N\left(0, \sigma^{2}\right) \\ & i=1, \ldots . n \end{aligned}$ |
| Non-Linear Dependence | $x_{i} \sim N(0,0.1)$ | $y_{i}=\left\|x_{i}\right\|^{e^{\text {sin }} \text { cos }}+\epsilon_{i}$ | $\begin{aligned} & \epsilon_{i} \sim N\left(0, \sigma^{2}\right) \\ & i=1, \ldots . n \end{aligned}$ |
| Cubic Root <br> Dependence | $x_{i} \sim U_{[a=0, b=1]}$ | $y_{i}=x_{i}^{1 / 3}+\epsilon_{i}$ | $\begin{aligned} & \epsilon_{i} \sim N\left(0, \sigma^{2}\right) \\ & i=1, \ldots . n \end{aligned}$ |
| Hyperbolic Dependence | $x_{i} \sim U_{[a=-1, b=1]}$ | $y_{i}=\left(x_{i}^{2}+\epsilon_{i} * k_{i}\right) * l_{i}$ | $\begin{aligned} i & =1, \ldots, n \\ \epsilon_{i} & \sim N\left(0, \sigma^{2}\right) \\ k_{i} & \sim U_{[a=0, b=1 / 2]} \\ l_{i} & \sim U_{[a=-1, b=1]} \end{aligned}$ |
| Step Dependence | $x_{i k}=k+\epsilon_{i}$ | $y_{i k}=\frac{1}{4} e^{k}+\delta_{i}$ | $\begin{aligned} & \epsilon_{i} \sim N(0, .1) \\ & \delta_{i} \sim N\left(0, \sigma^{2}\right) \\ & i=1, \ldots, \frac{n}{k} \\ & k=1,2,3,4 \end{aligned}$ |
| Cluster <br> Dependence | $x_{i k}=k+\epsilon_{i}$ | $y_{i k}=k+\epsilon_{i}$ | $\begin{aligned} & \epsilon_{i} \sim N(0, .1) \\ & i=1, \ldots ., \frac{n}{k} \\ & k=[-1,+1] \end{aligned}$ |

Table 1: The table includes a short description of each type of association investigated.

### 4.3 Noiseless simulations

The first step to understanding our new measures is to check how they perform in a noiseless simulation. The value of each measure is calculated for each data sets simulated. Two main values are reported for each measure in each model in table 2, the average of 1000 iteration is reported followed by the standard deviation in small brackets. The choice of noiseless models is to ensure a fair comparison between the new coefficients and the more established measurements. The simulation is repeated 1000 times, each time 16 data sets are created for every type of dependence included in the figure( 6). Every data set has $n=1000$ observations with no variance except for the model(1), which represent the independence case. This large number of simulations ensures the production of fair values for the measures and helps avoid any bias. The dependence measures included are 11, they start with Pearson productmoment correlation " $\rho$ ", Spearman's rank-order correlation " $r_{s}$ " and Distance Correlation "D.C". Moreover, mutual information "MI" and maximal information coefficient "MIC" are the two other well-established measures. This thesis aims to investigate the performance of the Wijayatunga coefficient; the four different functions listed in subsection 4.1 are included. Those functions appear directly after the five measures mentioned above. Lastly, two functions that represent the new approach to normalization of the mutual information are included and they come as the last two columns in the table below.

| Model | $\rho$ | $r_{s}$ | D.C | MI | MIC | W.D | $\begin{aligned} & \text { W.C. } \\ & \text { H } \end{aligned}$ | $\begin{aligned} & \text { W.C. } \\ & \text { KDE } \end{aligned}$ | W.C. KDE. 2 | NMI | $\begin{aligned} & \hline \text { NMI. } \\ & 2 \end{aligned}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Independen -ce | $\begin{aligned} & 0.001 \\ & (0.031) \end{aligned}$ | $\begin{aligned} & 0.001 \\ & (0.032) \end{aligned}$ | $\begin{aligned} & 0.056 \\ & (0.011) \end{aligned}$ | $\begin{aligned} & \hline 0.031 \\ & (0.005) \end{aligned}$ | $\begin{aligned} & 0.133 \\ & (0.008) \end{aligned}$ | $\begin{aligned} & 0.142 \\ & (0.022) \end{aligned}$ | $\begin{aligned} & \hline 0.100 \\ & (0.008) \end{aligned}$ | $\begin{aligned} & \hline 0.155 \\ & (0.006) \end{aligned}$ | $\begin{aligned} & 0.055 \\ & (0.006) \end{aligned}$ | $\begin{aligned} & 0.037 \\ & (0.041) \end{aligned}$ | $\begin{aligned} & 0.037 \\ & (0.041) \end{aligned}$ |
| Linear | $\begin{aligned} & \hline 1.000 \\ & (0.000) \end{aligned}$ | $\begin{aligned} & 1.000 \\ & (0.000) \end{aligned}$ | $\begin{aligned} & 1.000 \\ & (0.000) \end{aligned}$ | $\begin{aligned} & 1.765 \\ & (0.069) \end{aligned}$ | $\begin{aligned} & 1.000 \\ & (0.000) \end{aligned}$ | $\begin{aligned} & 1.000 \\ & (0.000) \end{aligned}$ | $\begin{aligned} & \hline 0.742 \\ & (0.011) \end{aligned}$ | $\begin{aligned} & \hline 0.737 \\ & (0.002) \end{aligned}$ | $\begin{aligned} & \hline 0.895 \\ & (0.004) \end{aligned}$ | $\begin{aligned} & 1.000 \\ & (0.000) \end{aligned}$ | $\begin{aligned} & 1.000 \\ & (0.000) \end{aligned}$ |
| Cubic | $\begin{aligned} & 0.934 \\ & (0.013) \end{aligned}$ | $\begin{aligned} & 1.000 \\ & (0.000) \end{aligned}$ | $\begin{aligned} & 0.981 \\ & (0.002) \end{aligned}$ | $\begin{aligned} & 0.777 \\ & (0.130) \end{aligned}$ | $\begin{aligned} & 1.000 \\ & (0.000) \end{aligned}$ | $\begin{aligned} & 0.823 \\ & (0.060) \end{aligned}$ | $\begin{aligned} & 0.492 \\ & (0.049) \end{aligned}$ | $\begin{aligned} & 0.747 \\ & (0.002) \end{aligned}$ | $\begin{aligned} & 0.789 \\ & (0.013) \end{aligned}$ | $\begin{aligned} & 0.922 \\ & (0.073) \end{aligned}$ | $\begin{aligned} & 0.922 \\ & (0.073) \end{aligned}$ |
| Quadratic | $\begin{aligned} & -0.002 \\ & (0.072) \end{aligned}$ | $\begin{aligned} & 0.000 \\ & (0.042) \end{aligned}$ | $\begin{aligned} & 0.542 \\ & (0.007) \end{aligned}$ | $\begin{aligned} & 0.597 \\ & (0.111) \end{aligned}$ | $\begin{aligned} & 1.000 \\ & (0.000) \end{aligned}$ | $\begin{aligned} & 0.687 \\ & (0.087) \end{aligned}$ | $\begin{aligned} & 0.511 \\ & (0.029) \end{aligned}$ | $\begin{aligned} & 0.647 \\ & (0.002) \end{aligned}$ | $\begin{aligned} & 0.583 \\ & (0.006) \end{aligned}$ | $\begin{aligned} & 0.817 \\ & (0.175) \end{aligned}$ | $\begin{aligned} & 0.817 \\ & (0.175) \end{aligned}$ |
| Sinusoidal | $\begin{aligned} & 0.071 \\ & (0.033) \end{aligned}$ | $\begin{gathered} 0.163 \\ (0.036) \end{gathered}$ | $\begin{aligned} & 0.385 \\ & (0.017) \end{aligned}$ | $\begin{aligned} & 0.619 \\ & (0.059) \end{aligned}$ | $\begin{aligned} & 1.000 \\ & (0.000) \end{aligned}$ | $\begin{aligned} & 0.674 \\ & (0.030) \end{aligned}$ | $\begin{gathered} 0.482 \\ (0.026) \end{gathered}$ | $\begin{aligned} & 0.560 \\ & (0.006) \end{aligned}$ | $\begin{gathered} 0.550 \\ (0.006) \end{gathered}$ | $\begin{aligned} & 0.849 \\ & (0.073) \end{aligned}$ | $\begin{aligned} & 0.849 \\ & (0.073) \end{aligned}$ |
| Cross <br> Shaped | $\begin{aligned} & 0.000 \\ & (0.033) \end{aligned}$ | $\begin{aligned} & 0.000 \\ & \\ & (0.036) \end{aligned}$ | 0.328 <br> (0.017) | $\begin{aligned} & 0.232 \\ & \\ & (0.059) \end{aligned}$ | $\begin{aligned} & 0.630 \\ & \\ & (0.000) \end{aligned}$ | $\begin{aligned} & 0.668 \\ & \\ & (0.030) \end{aligned}$ | $\begin{aligned} & 0.293 \\ & \\ & (0.026) \end{aligned}$ | $\begin{aligned} & 0.142 \\ & \\ & (0.006) \end{aligned}$ | $\begin{aligned} & 0.329 \\ & \\ & \\ & \end{aligned}$ | $\begin{aligned} & 0.241 \\ & \\ & (0.073) \end{aligned}$ | $\begin{aligned} & 0.241 \\ & \\ & (0.073) \end{aligned}$ |
| Circular | $\begin{aligned} & 0.000 \\ & (0.034) \end{aligned}$ | $\begin{aligned} & 0.001 \\ & (0.039) \end{aligned}$ | $\begin{aligned} & 0.411 \\ & (0.011) \end{aligned}$ | $\begin{aligned} & \hline 0.755 \\ & (0.016) \end{aligned}$ | $\begin{aligned} & 0.995 \\ & (0.009) \end{aligned}$ | $\begin{aligned} & 0.708 \\ & (0.015) \end{aligned}$ | $\begin{aligned} & 0.642 \\ & (0.004) \end{aligned}$ | $\begin{aligned} & 0.571 \\ & (0.005) \end{aligned}$ | $\begin{aligned} & 0.567 \\ & (0.003) \end{aligned}$ | $\begin{aligned} & 0.811 \\ & (0.011) \end{aligned}$ | $\begin{aligned} & 0.811 \\ & (0.011) \end{aligned}$ |
| Two Functions | $\begin{aligned} & -0.176 \\ & (0.026) \end{aligned}$ | $\begin{aligned} & \hline-0.067 \\ & (0.029) \end{aligned}$ | $\begin{aligned} & \hline 0.461 \\ & (0.018) \end{aligned}$ | $\begin{aligned} & \hline 0.970 \\ & (0.078) \end{aligned}$ | $\begin{aligned} & 1.000 \\ & (0.000) \end{aligned}$ | $\begin{aligned} & 0 . .825 \\ & (0.023) \end{aligned}$ | $\begin{aligned} & \hline 0.565 \\ & (0.024) \end{aligned}$ | $\begin{aligned} & \hline 0.561 \\ & (0.012) \end{aligned}$ | $\begin{aligned} & \hline 0.566 \\ & (0.008) \end{aligned}$ | $\begin{aligned} & \hline 0.915 \\ & (0.045) \end{aligned}$ | $\begin{aligned} & \hline 0.9157 \\ & (0.045) \end{aligned}$ |
| Checkerboard | $\begin{aligned} & 0.058 \\ & (0.040) \end{aligned}$ | $\begin{gathered} 0.226 \\ (0.034) \end{gathered}$ | $\begin{aligned} & 0.281 \\ & (0.018) \end{aligned}$ | $\begin{aligned} & 0.239 \\ & (0.055) \end{aligned}$ | $\begin{aligned} & 0.573 \\ & (0.032) \end{aligned}$ | $\begin{aligned} & 0.456 \\ & (0.052) \end{aligned}$ | $\begin{aligned} & 0.272 \\ & (0.044) \end{aligned}$ | $\begin{aligned} & 0.477 \\ & (0.006) \end{aligned}$ | $\begin{aligned} & 0.358 \\ & (0.007) \end{aligned}$ | $\begin{aligned} & 0.688 \\ & (0.167) \end{aligned}$ | $\begin{aligned} & 0.688 \\ & (0.167) \end{aligned}$ |
| Exponential Function | $\begin{aligned} & 0.904 \\ & (0.013) \end{aligned}$ | $\begin{aligned} & 1.000 \\ & (0.000) \end{aligned}$ | $\begin{aligned} & 0.971 \\ & (0.002) \end{aligned}$ | $\begin{aligned} & 0.863 \\ & (0.145) \end{aligned}$ | $\begin{aligned} & 1.000 \\ & (0.000) \end{aligned}$ | $\begin{gathered} 0.823 \\ (0.069) \end{gathered}$ | $\begin{aligned} & 0.527 \\ & (0.053) \end{aligned}$ | $\begin{aligned} & 0.722 \\ & (0.003) \end{aligned}$ | $\begin{aligned} & 0.792 \\ & (0.010) \end{aligned}$ | $\begin{aligned} & 0.911 \\ & (0.092) \end{aligned}$ | $\begin{aligned} & 0.911 \\ & (0.092) \end{aligned}$ |
| Logarithmic | $\begin{aligned} & \hline 0.001 \\ & (0.032) \end{aligned}$ | $\begin{aligned} & \hline 0.001 \\ & (0.043) \end{aligned}$ | $\begin{aligned} & \hline 0.524 \\ & (0.006) \end{aligned}$ | $\begin{aligned} & \hline 0.744 \\ & (0.090) \end{aligned}$ | $\begin{aligned} & 1.000 \\ & (0.000) \end{aligned}$ | $\begin{aligned} & \hline 0.735 \\ & (0.045) \end{aligned}$ | $\begin{aligned} & \hline 0.496 \\ & (0.037) \end{aligned}$ | $\begin{aligned} & \hline 0.651 \\ & (0.002) \end{aligned}$ | $\begin{aligned} & \hline 0.606 \\ & (0.003) \end{aligned}$ | $\begin{aligned} & \hline 0.846 \\ & (0.092) \end{aligned}$ | $\begin{aligned} & \hline 0.846 \\ & (0.092) \end{aligned}$ |
| Non-Linear | $\begin{aligned} & -0.391 \\ & (0.047) \end{aligned}$ | $\begin{aligned} & -0.312 \\ & (0.039) \end{aligned}$ | $\begin{aligned} & 0.656 \\ & (0.019) \end{aligned}$ | $\begin{aligned} & 1.007 \\ & (0.108) \end{aligned}$ | $\begin{aligned} & 1.000 \\ & (0.000) \end{aligned}$ | $\begin{gathered} 0.798 \\ (0.031) \end{gathered}$ | $\begin{aligned} & 0.570 \\ & (0.022) \end{aligned}$ | $\begin{aligned} & 0.666 \\ & (0.002) \end{aligned}$ | $\begin{aligned} & 0.653 \\ & (0.004) \end{aligned}$ | $\begin{aligned} & 0.855 \\ & (0.047) \end{aligned}$ | $\begin{aligned} & 0.855 \\ & (0.047) \end{aligned}$ |
| Cubic Root | $\begin{aligned} & \hline 0.958 \\ & (0.003) \end{aligned}$ | $\begin{aligned} & \hline 1.000 \\ & (0.000) \end{aligned}$ | $\begin{aligned} & \hline 0.984 \\ & (0.001) \end{aligned}$ | $\begin{aligned} & 1.472 \\ & (0.036) \end{aligned}$ | $\begin{aligned} & 1.000 \\ & (0.000) \end{aligned}$ | $\begin{aligned} & \hline 0.897 \\ & (0.007) \end{aligned}$ | $\begin{aligned} & \hline 0.699 \\ & (0.005) \end{aligned}$ | $\begin{aligned} & \hline 0.712 \\ & (0.002) \end{aligned}$ | $\begin{aligned} & \hline 0.828 \\ & (0.003) \end{aligned}$ | $\begin{aligned} & \hline 0.910 \\ & (0.011) \end{aligned}$ | $\begin{aligned} & \hline 0.910 \\ & (0.011) \end{aligned}$ |
| Hyperbolic | $\begin{aligned} & -0.001 \\ & (0.047) \end{aligned}$ | $\begin{gathered} 0.000 \\ (0.043) \end{gathered}$ | $\begin{aligned} & \hline 0.312 \\ & (0.008) \end{aligned}$ | $\begin{aligned} & 1.212 \\ & (0.019) \end{aligned}$ | $\begin{aligned} & 0.668 \\ & (0.011) \end{aligned}$ | $\begin{aligned} & 0.827 \\ & (0.010) \end{aligned}$ | $\begin{gathered} 0.602 \\ (0.006) \end{gathered}$ | $\begin{aligned} & \hline 0.617 \\ & (0.005) \end{aligned}$ | $\begin{aligned} & 0.554 \\ & (0.003) \end{aligned}$ | $\begin{aligned} & \hline 0.897 \\ & (0.014) \end{aligned}$ | $\begin{aligned} & 0.897 \\ & (0.014) \end{aligned}$ |
| Step | $\begin{aligned} & 0.923 \\ & (0.001) \end{aligned}$ | $\begin{gathered} 0.968 \\ (0.000) \end{gathered}$ | $\begin{aligned} & 0.938 \\ & (0.001) \end{aligned}$ | $\begin{aligned} & 1.040 \\ & (0.000) \end{aligned}$ | $\begin{aligned} & 1.000 \\ & (0.000) \end{aligned}$ | $\begin{gathered} 0.829 \\ (0.031) \end{gathered}$ | $\begin{aligned} & 0.707 \\ & (0.000) \end{aligned}$ | $\begin{aligned} & 0.648 \\ & (0.001) \end{aligned}$ | $\begin{aligned} & 0.705 \\ & (0.001) \end{aligned}$ | $\begin{aligned} & 0.900 \\ & (0.048) \end{aligned}$ | $\begin{aligned} & 0.900 \\ & (0.048) \end{aligned}$ |
| cluster | $\begin{aligned} & \hline 0.000 \\ & (0.004) \end{aligned}$ | $\begin{aligned} & \hline 0.001 \\ & (0.021) \end{aligned}$ | $\begin{aligned} & 0.016 \\ & (0.003) \end{aligned}$ | $\begin{aligned} & 0.009 \\ & (0.003) \end{aligned}$ | $\begin{aligned} & \hline 0.132 \\ & (0.008) \end{aligned}$ | $\begin{aligned} & \hline 0.343 \\ & (0.247) \end{aligned}$ | $\begin{aligned} & \hline 0.060 \\ & (0.008) \end{aligned}$ | $\begin{aligned} & 0.022 \\ & (0.004) \end{aligned}$ | $\begin{aligned} & 0.032 \\ & (0.004) \end{aligned}$ | $\begin{aligned} & 0.385 \\ & (0.349) \end{aligned}$ | $\begin{aligned} & 0.385 \\ & (0.349) \end{aligned}$ |

Table 2: The results of the average and standard deviation values for all dependence measure included in the thesis over 1000 iteration. The iteration is repeated over the 16 association models included in the thesis. The standard deviation is reported in brackets.

### 4.3.1 Coefficients average performance.

The first observation to be spotted is the result of linear dependence. After discretization, the discrete version of the Wijayatunga coefficient, as well as the normalized mutual information, performs perfectly. Additionally, the conventional measures work as suspected. Unfortunately, the main functions, which deal with the continuous version of the Wijayatunga coefficient, don't perform as well as the other functions.

The histogram and the initial kernel density estimation have their limitations in estimating the joint probability distributions. Whereas for the histogram-based function, the full linear association is not entirely detected. Different numbers of bins are used prior to the simulation to check if there is a possibility of observing this full association, the increase
in the number of bins leads to overestimation rather than more accuracy. Moreover, this function can detect the existence of association and the absence of the association in the case of independence. It is worth noting that the highest values of the histogram-based function are recorded for the two linear associations, which are "linear" and "step function."

In the case of the simple kernel function, the results have the same level of precision as the histogram-based function; this is a result of simplifying the bandwidth selector. The simple kernel-based function is incapable of detecting the full dependency. Moreover, this function produces the highest value in the case of independence, in other words, it has the lowest accuracy in reflecting the absence of any association among the measures included in the thesis.

As reported in table 2, the Wijayatunga coefficient function that employs kde function in the ks package performs much better. Adding small noise in the case of full linear dependence lowers the accuracy, yet it has better results than the other two continuous functions. The superiority of the third continuous function persists in all of the results.

After data discretization, the discrete version of the Wijayatunga coefficient predicts the correlation to the same level of accuracy as distance correlation and MIC. Furthermore, the two normalized mutual information yield results with high accuracy. They outperform distance correlation in most cases except for cross-shaped, cubic and cubic root and step functions. The result for those two functions surpasses the MIC performance only for checker-board and hyperbolic dependencies.

The results of applying D.C and MIC align with the findings of previous studies. The kernelbased function, which employs kde function from ks package, matches the performance of D.C function in most cases except for linear and monotonic relationships. MIC has the highest values in case of dependencies, yet, it has been proven that MIC's accuracy decreases very fast when noise is introduced.
The results obtained in the table 2 are an outcome of a simulation study; dependencies with such low noise don't occur in real life. The noiseless simulation makes these results unrealistic. The next step will be to add noise to the data to investigate how the noise would impact the ability to detect associations using different measurements.

### 4.3.2 The variability of the coefficients.

Looking at the variability of Spearman's and Pearson's coefficient in table 2, we see a similar trend across different association. It is worth-noting that for linear associations, those functions has nominal or no variability at all. For non linear relationship, distance correlation has much lower variability than Pearson's and Spearman's coefficients. Looking at the various functions that represent Wijayatunga coefficient for continuous cases shows that the measure has a very small and insignificant variability around its mean. Those functions have a higher variability than the correlation coefficient only in case of linear and monotonic associations.
The discrete measures, which include both Wijayatunga's coefficient and the normalized
mutual information, have higher level of variability.

### 4.4 The noise effect

In this section, only W.C.KDE. 2 will be included as representative for the continuous version of the Wijayatunga coefficient. Moreover, mutual information is excluded as it is not bounded in $[0,1]$ which makes mutual information incomparable with the other measures in this study. Finally, only the simple normalized mutual information is included.
This section aims to investigate the effect of statistical noise on the performance of various dependence measurements. The study calculates the mean of each quantity over 1000 simulated data sets; each data set is fixed to 1000 observations. This simulation process is repeated for ten different levels of statistical noise $\sigma$, which increase incrementally starting from a noiseless level $\sigma=0$ to $\sigma=0.3$.

The effect of adding noise to the linear relationship

$\sigma$

Figure 7

As shown in the figure 7, a noiseless linear relationship is fully detected by all measurements except for the new continuous measure. It has been mentioned earlier that kernels couldn't be obtained when full linear dependency occurs. Furthermore, the mean values of each coefficient decrease as the number of noise increases. Moreover, it is worth noting that this continuous function underperforms compared to all well-established measures that detect
linearity. Only "MIC" performs worse than this continuous function; this aligns with other studies, which question the "MIC" performance. (Kinney and Atwal, 2014) argues that "MIC" under-performs when noise is introduced. The two discrete functions perform at the same accuracy level as D.C and outperform the conventional correlation coefficients.


Figure 8: Noise effect on cubic relationships

Monotonic relationships are easily detected using the most established dependency measures; for instance, the moment product correlation and rank correlation coefficient outperform all other measures. As shown in the Figure 9a, the moment product cannot fully detect those relations in noiseless settings, yet, it deteriorates slower compared to others when the white noise is introduced. The continuous function underperforms corresponding to all other included functions in detecting monotonic relationships.MIC value decrease very fast when noise is introduced. Distance correlation neighbors Spearman coefficient throughout the simulation for monotonic relationship. After discretization, the normalized mutual information is more resilient to the white noise effect in the case of cubic relationship in comparison to the rest of the measures. This normalized MI does not show the same level of resilience in the case of cubic root, Figure 8b.


Figure 9: Noise effect on different non-linear relationships

### 4.5 The effect of the sample size

This section focuses on the effect of the number of observations on the ability of dependency detection. Two different types of association are included, circular and hyperbolic relationships.Noise is added to all associations, where the standard deviation for noise added is $\sigma=0.1$. A thousand iterations for each sample size are created; after that, five different dependency measures are included. The dependency measures included are distance correlation, MIC, Wijayatunga coefficient for both continuous and discrete cases and finally, the new normalized mutual information coefficient.The average of the thousand iterations is calculated and reported in the tables below. For small sample sizes of 5 observations, the discretization method has been altered to separate the data into five levels. This alteration is done so the two discrete measures, the Wijayatunga coefficient and normalized mutual information, could be calculated.

Circular relationship

| Num of obs | D.C | MIC | W.D | W.C.KDE.2 | NMI |
| :--- | :--- | :--- | :--- | :--- | :--- |
| 5 | 0.7359240 | 0.5796736 | 0.6342231 | 0.3079195 | 0.7796115 |
| 10 | 0.5971371 | 0.4297120 | 0.3220150 | 0.2463312 | 0.1785608 |
| 20 | 0.5017117 | 0.7262464 | 0.2020629 | 0.2488002 | 0.0794335 |
| 30 | 0.4671186 | 0.6843075 | 0.4801920 | 0.2624249 | 0.4659884 |
| 100 | 0.3989970 | 0.7436398 | 0.5318263 | 0.3169643 | 0.6389582 |
| 200 | 0.3807809 | 0.7448565 | 0.5695902 | 0.3491187 | 0.7201879 |

Table 3: An illustration of the behaviour of five selected measure of dependence for different number of observations with circular relationship reference and noise $\sigma=0.1$

In table 3,for all included measures except W.C.KDE. 2 , it is easy to detect the correlation for a small sample size ( 5 observations). All values drop when increasing the number of observations to 10, the least affected measure is W.C.KDE. 2 and D.C . The W.C.KDE. 2 starts to increase afterward; on the contrary, D.C keeps decreasing when the number of observations increases. MIC's value decreases initially, but when the sample size hits 100,
enduring values could be observed. Both discrete measurements drop sharply at the beginning; this results from the change in the number of bins when discretizing. When the sample size reaches 30 observations, those two functions improve significantly, and this trend persists afterward for bigger samples.

Hyperbolic relationship

| Num of obs | D.C | MIC | W.D | W.C.KDE. 2 | NMI |
| :--- | :--- | :--- | :--- | :--- | :--- |
| 5 | 0.7608308 | 0.5630077 | 0.4641561 | 0.3732210 | 0.2764109 |
| 10 | 0.6092997 | 0.3433368 | 0.2666443 | 0.2941028 | 0.1911236 |
| 20 | 0.4952303 | 0.4513325 | 0.1492087 | 0.2822829 | 0.0988859 |
| 30 | 0.4383867 | 0.3756586 | 0.5474716 | 0.2865735 | 0.6214799 |
| 100 | 0.3494959 | 0.6034477 | 0.4943629 | 0.3633926 | 0.5770431 |
| 200 | 0.3258794 | 0.6107865 | 0.6671640 | 0.4211862 | 0.7709294 |

Table 4: An illustration of the behaviour of five selected measure of dependence for different number of observations with hyperbolic relationship reference and noise $\sigma=0.1$

Similarly, in table 4, the correlation could be easily detected for the small sample size ( 5 observations); only the normalized mutual information under-performs in this case. W.C.KDE. 2 and D.C behave exactly as they behave in the case of circular correlation. MIC 's values don't pick up their high performance until the sample size reaches 100 . Finally, the values of the discrete measurements move parallelly; moreover, the normalized mutual information produces better results.

## 5 Feature Selection Method

In machine learning literature, the feature selection aims to identify relevant features from a large set of them for prediction tasks, etc. by reducing the prediction error. The idea is to find and eliminate irrelevant features that increase the prediction variance without reducing the bias. In feature selection, redundant features have to be filtered out. There is one main difference between feature selection and feature screening. The goal of feature screening is to reduce the dimensions of the feature space to a smaller size while retaining all relevant features; this is known as the sure screening property. This procedure does not check for interdependency among the explanatory variables. The feature selection is an essential tool even in the case of uncorrelated features. Two different feature selection methods are discussed below. First is the so-called elastic net, then an introduction for "minimal redundancy maximal relevance" feature selection is provided. This method could be wrapped with either forward selection or backward elimination (Hanchuan Peng, Fuhui Long and Ding, 2005).

### 5.1 Elastic Net Regularization

One of the most common practices in statistics is the least-square method for regression parameter estimation. When the covariates are correlated, the least-squares method, although unbiased, suffers from inflated variance; this is because the covariance matrix is nearly singular. The inverse of that matrix is either ill-conditioned or non-existent in the case of a


Figure 10: This figure shows the difference in estimating $\beta$ 's for three different regularization models
singular covariance matrix. Shrinkage methods are used to overcome this issue. There are three different penalization models, ridge regression, Lasso regression and elastic net model. Ridge regression shrinks the covariate coefficients that contribute the least to the model. It introduces an additional term to the least-square objective function; this additional term that is the penalty term, regularizes the value of the coefficients. The $L_{2}$ norm penalty term is used for ridge regression is:

$$
(Y-X \beta)^{T}(Y-X \beta)+\lambda \beta^{T} \beta
$$

where $\lambda$ is the tuning parameter, when $\lambda=0$ the model is the least-square, for any value $\lambda>0$ the ridge estimator produces bias, at the same time, it reduces the variance when the covariance matrix is nearly singular. The choice of $\lambda$ is crucial. The disadvantage of this method is that it maintains all variables within the model.
Least Absolute Shrinkage and Selection Operator (lasso) was proposed by (Tibshirani, 1996). This method shrinks the covariate coefficients to zero through regularizing the model with penalty parameter of $l_{1}$ norm. The lasso estimator minimize the objective function as follows: Even though it is similar to the ridge estimator, it differs in a few aspects. The most notable difference is the ability of the model to render the value of some coefficients to zero. In this sense, the lasso can work as a feature selection method. Furthermore, It is a sparse regression. Three disadvantages come with the lasso; first, in the case of $p>n$ the lasso selects at most n variables before it saturates. Secondly, lasso chooses only one variable from any group of highly correlated variables. When $p>n$, if there are high correlations between predictors, it has been empirically observed that the prediction performance of the lasso is dominated by ridge regression. The elastic net combines both penalties $L_{1}$ and $L_{2}$ to overcome the shortcomings of both models (Zou and Hastie, 2005). It combines the advantages of shrinkage and sparsity. The function is defined as follows:

$$
\min _{\beta, \beta}\left(\frac{1}{2 N} \Sigma_{i=1}^{N}\left(y_{i}-\beta_{0}-X_{i}^{T} \beta\right)^{2}+\lambda P_{\alpha}(\beta)\right)
$$

where

$$
P_{\alpha}(\beta)=\frac{(1-\alpha)}{2}\|\beta\|_{2}^{2}+\alpha\|\beta\|_{1}=\Sigma_{j=1}^{P}\left(\frac{(1-\alpha)}{2} \beta_{j}^{2}+\alpha\left|\beta_{j}\right|\right)
$$

N is the number of observations.
$y_{i}$ is the response at observation i.
$X_{i}$ is data, a vector of p values at observation i.
$\lambda$ is a positive regularization parameter corresponding to one value of Lambda.
The parameters $\beta_{0}$ and $\beta$ are an scalar and a p-vector respectively. Two R packages are employed to run the elastic net regularization. Those packages are caret and glmnet.

### 5.2 Minimal Redundancy Maximal Relevance

Starting with the concept of Max-Dependency, let's denote $X_{1}, \ldots ., X_{p}$ as the set of possible features used to predict an outcome variable $Y$. A dependency measure $D(.,$.$) quantifies$ the correlation between any two random variables. The Max-Dependency system for feature selection involves finding a subset $m$ of the features $X_{i}, \ldots ., X_{i}$, which jointly have the largest dependency with $Y$; in other words, the subset is obtained by solving the following optimization problem

$$
\max _{\{i, \ldots, i\} \subset\{1, \ldots, p\}} D\left(\left\{X_{i}, \ldots, X_{i}\right\}, Y\right)
$$

Solving this optimization problem is infeasible in the case of high dimensional data; the near-optimal solutions are obtained by iterative procedures, where variables are added one at a time. The mentioned algorithm is known as the forward selection. The dependency measure ought to be robust, which is hard to achieve when the sample size is less than the number of features m. alternatively, marginal computations which only include $D\left(X_{k}, Y\right)$ are desired. Those marginals serve as a proxy to Max-Dependency; those marginals are called the Max-Relevance criterion. The optimization problem is written as follows

$$
\max _{\{i, \ldots, i\} \subset\{1, \ldots, p\}} \frac{1}{m} \Sigma_{k=1}^{m} D\left(X_{i}, Y\right)
$$

Finally, when the features are dependent, the above criteria will likely select redundantly features. To mitigate the effect of collinearity, an extra condition could be added, (Hanchuan Peng, Fuhui Long and Ding, 2005) considers the Min-Redundancy condition

$$
\min _{\{i, \ldots, i\} \subset\{1, \ldots, p\}} \frac{1}{m^{2}} \Sigma_{k, l=1}^{m} D\left(X_{i}, X_{i}\right)
$$

Combining both optimization conditions produces what is called "Minimal Redundancy Maximal Relevance"

$$
\max _{\{i, \ldots, i\} \subset\{1, \ldots, p\}} \frac{1}{m} \sum_{k=1}^{m} D\left(X_{i}, Y\right)-\frac{1}{m^{2}} \sum_{k, l=1}^{m} D\left(X_{i}, X_{i}\right)
$$

Forward and backward procedures could be applied to this function.Backward elimination starts with all candidate variables; it tests the deletion of each variable using a chosen model
fit criterion. After each step, it deletes the variable whose loss gives the most statistically insignificant deterioration of the model fit. The algorithm repeats this process until no further variables can be deleted without a statistically significant loss of fit. Forward selection starts with no variables in the model; it tests how the addition of each variable gives the most statistically significant improvement of the model fitness; the process repeats until none improves the model to a statistically significant extent(Da Veiga, 2014).
In the following subsection, W.C.KDE.2is used as measure of distance in "mRMR".

### 5.3 Real Life Data Sets

Two data sets are included in this thesis, the first data set is a biological data set, while the second one touches upon urban planning, crime, and economics. The data set are ordered below according to the number of features in the data-sets.
The first data-set is the "Prostate" data, it is obtained from R package "lasso2"; the data was originally used in (Stamey et al., 1989). It examines the correlation between the level of prostate-specific antigen and a number of clinical measures in men who were about to receive radical prostatectomy. It consists of 9 features and 97 observations. The target variable is log"cancer volume.". The table below provides a glimpse of the data.

| variable abbreviation | Description |
| :--- | :--- |
| lcavol | $\log$ (cancer volume) |
| lweight | $\log$ (prostate weight) |
| age | age |
| lbph | $\log$ (benign prostatic hyperplasia amount) |
| svi | seminal vesicle invasion |
| lcp | $\log$ (capsular penetration) |
| gleason | Gleason score |
| pgg45 | percentage Gleason scores 4 or 5 |
| lpsa | log(prostate specific antigen) |

Table 5: The names of the covariates in the Prostate data-set

Table 6 provides the results of running elastic net regression on this data set; the results suggest that "svi" variable contributes nothing to the model. The Minimal Redundancy Maximal Relevance optimization "mRMR" method produces a different model using backward elimination; with seven predictors, the suggested model excludes "pgg45" instead. The value of " mRMR " for this specific model is 0.1396489 . The comparison of $R^{2}$ is done between the two least square regressions, wherein each, we remove the most redundant variable according to the two feature selection methods. It is shown that the elastic net model produces a higher $R^{2}=0.6973$ value in comparison to "mRMR" where $R^{2}=0.6668$

| variable | importance\% |
| :---: | :---: |
| lpsa | 100 |
| lcp | 61.73 |
| gleason | 33.50 |
| lbph | 14.28 |
| lweight | 5.5 |
| age | 3.7 |
| pgg45 | 0.86 |
| svi | 0 |

Table 6: The variables importance according to the elastic regression for the Prostate dataset
(Harrison and Rubinfeld, 1978) introduces "Boston Housing" data set. This dataset contains information collected by the U.S Census Service concerning housing in the Boston Standard Metropolitan Statistical Area (SMSA) in 1970. The goal is to estimate the willingness to pay for clean air. It contains 506 observations and 14 different variables. The target variable is the median value of owner-occupied homes. Moreover, The independent variables range between structural attributes, neighborhood variables, accessibility variables, and one air pollution variable.

According to (Boston Dataset, 2022), "medv" seems to be censored at 50.00 (corresponding to a median price of $50,000 \$$ ); Censoring is suggested by the fact that the highest median price of exactly $50,000 \$$ is reported in 16 cases, while 15 cases have prices between $40,000 \$$ and $50,000 \$$, with prices rounded to the nearest hundred. (Harrison and Rubinfeld, 1978) does not mention any censoring. The data set is retrieved from "mlbench" package in R.

| variable abbreviation | Description |
| :--- | :--- |
| crim | per capita crime rate by town |
| zn | proportion of residential land zoned for lots over 25,000 sq.ft. |
| indus | proportion of non-retail business acres per town. |
| chas | Charles River dummy variable (1 if tract bounds river; 0 otherwise) |
| nox | nitric oxides concentration (parts per 10 million) |
| rm | average number of rooms per dwelling |
| age | proportion of owner and occupied units built prior to 1940 |
| dis | weighted distances to five Boston employment centres |
| rad | index of accessibility to radial highways |
| tax | full and value property-tax rate per $\$ 10,000$ |
| ptratio | pupil-teacher ratio by town |
| b | $1000(B k-0.63)^{2}$ where bk is the proportion of blacks by town |
| lstat | $\%$ lower status of the population |
| medv | Median value of owner-occupied homes in $\$ 1000$ 's |

Table 7: The names of the covariates in the Boston Housing data-set

It's worth mentioning that the data suffers from some ethical issues, which is mentioned by (The Census users' Guide, 2022) as it states that some of the terminology might be considered inappropriate.

| variable | importance\% |
| :---: | :---: |
| nox | 100 |
| rm | 33.65 |
| chas | 24.32 |
| dis | 9.36 |
| ptratio | 7.15 |
| lstat | 3.94 |
| rad | 1.26 |
| crim | 0.7 |
| indus | 0.29 |
| zn | 0.24 |
| b | 0.04 |
| tax | 0.016 |
| age | 0 |

Table 8: The variables importance according to the elastic regression for the Boston Housing data-set

Similarly, table 8 provides the results of running elastic net regression on the second data set; the results suggest that the "age" variable does not contribute to the model. The Minimal Redundancy Maximal Relevance optimization "mRMR" method produces a different model using backward elimination; with thirteen predictors, the suggested model excludes "rad" instead. The value of " mRMR " for this specific model is 0.1534865 . The comparison of $R^{2}$ is done between the two least square regressions, wherein each, we remove the most redundant variable according to the two feature selection methods. It is shown that the elastic net model produces a higher $R^{2}=0.7406$ value in comparison to " mRMR " where the value of $R^{2}=0.7294$

## 6 Conclusion

This paper investigates various measures of dependence for their suitability of measuring the dependence and application of feature selection in regression modelling. Those measures include Pearson's, Spearman's and distance correlation coefficients, the mutual information and the maximal information coefficient (MIC), and two new dependence measures, namely Wijayatunga's coefficient and normalized mutual information. The investigation is done using simulated data sets. Moreover, a proof of concept for feature selection is conducted using two real-life data sets.

All the conventional dependence measures perform as expected. Pearson's correlation is the best choice for studying linear dependence, given that the variables follow the normal distribution. In case of these assumptions are violated, Spearman's coefficient might be a better choice because it recognizes monotonic dependence. However, even for a nonlinear, monotonic relationship, Pearson's correlation coefficient can be used to check the association direction. Moreover, when adding white noise, the two measures resist more than other measures across all reported relationships. Although, for more complex relationships, those two measures are not affected by the noise because they don't initially detect the relationship; they perceive these relationships as independent variables. Another well-established
measure is mutual information. This measure is incomparable to the other measure as it is not bounded in $[0,1]$, yet, it performs as suspected in noiseless simulations. Its value is negligible only when the variables are independent. Furthermore, it is noteworthy that the cluster association value is smaller than the value of independence for the three measures mentioned above.

Next, both distance correlation and MIC's results support the findings of previous studies. First, distance correlation in noiseless simulations identifies all relationships; furthermore, adding noise affects DC too little compared to the other measures. For linear and monotonic relationships, DC pairs with Pearson's and Spearman's coefficients; the fact that DC generalizes the correlation for all distributions can explain this pairing. When the noise is added to more complex relationships, even though DC matches up the $\rho$ in trend, it has a higher value.The small number of observations does not affect the value of this measure compared to the other measures; in fact, DC surpasses all other measures for small samples.

MIC performs as anticipated across the simulations; this measure surpasses all other measures in soundless simulations. Yet, adding noise deteriorates the value of MIC the most compared to the other measures; this aligns with previous discoveries.

Moving on to the goal of this paper, two different main dependence measures are studied. First, two different approaches to the Wijayatunga coefficient are presented; the first deals with discrete variables, while the second manages the case of continuous variables. The discrete version of Wijayatunga shows good potential in detecting the distance between probability distributions, i.e., the level of dependence among the variables; it produces a high coefficient value for non-monotonic relationships. Moreover, the discrete version of Wijayatunga recovers from the introduction of noise in the case of linear association very fast. It surpasses MIC for higher levels of noise in the linear case. The Wijayatunga coefficient continues to endure the effect of the noise in both monotonic and non-monotonic relationships.

The continuous version of the Wijayatunga coefficient is inspected using three different functions; the first function utilizes the concept of histograms to estimate the distance between the distributions. This function cannot detect perfect linear association; furthermore, the function is sensitive to the number of bins included in the 2D histogram. The second function could be considered a proof-of-principle for the continuous version of the Wijayatunga coefficient using the kernel estimation; this prototype displays the measure's potential; yet, it suffers from accuracy issues as it cannot estimate a proper kernel estimation.

The last of the continuous version of the Wijayatunga coefficient employs the kernel density estimation function in ks R package; this function has much potential, yet, it suffers from two disadvantages. The first drawback of this function is the speed of the function; it takes on average 16 times more in comparison to the previous function; the second drawback is the inability of such kernels to detect full dependencies. This function otherwise detects
all types of associations provided in this study. The noise effect on the value of this coefficient in the monotonic relationships is monotonous; the value decreases proportionally to the increase in the noise amount. This trend changes in the complex relationship case; this function detects the association in hyperbolic and circular cases and resists the noise effect with a higher value than DC. Furthermore, this function is invariant to the number of bins applied. Lastly, this function works for small samples, but its value grows exponentially when the number of observations increases.

Normalized mutual information concludes the simulation section; the function shows considerable potential in detecting associations. It performs well throughout the noiseless simulations; furthermore, it works well in cases of noisy linear and monotonic associations. This function surpasses all other functions when detecting hyperbolic and circular relationships. Lastly, the value of this function is affected by the number of observations.

The last part of the thesis touch upon the feature selection methods; two different methods are included. First comes elastic net regression, which is compared with the "mRMR" method. "mRMR" method in this thesis employs the Wijayatunga coefficient as a measure of dependence. The purpose here is to establish a proof of concept of using this coefficient to eliminate the redundant variables. The attempt here failed to show the effectiveness of this coefficient. The results show that the elastic net performs better than the new coefficient.

To conclude this thesis, two new coefficients are investigated. The first coefficient could be used to evaluate the association between discrete variables; also, an extended version could be used to evaluate the association between continuous variables. This new coefficient shows considerable potential in detecting associations. The second coefficient is a normalization of mutual information; the results suggest that the coefficient can detect various types of associations. Lastly, a minor experiment is done to show the possibility of incorporating the Wijayatunga coefficient in feature selection processes. The results are preliminary and need more investigation.

## 7 References

Asuero, A., Sayago, A. and González, A., 2006. The Correlation Coefficient: An Overview. Critical Reviews in Analytical Chemistry, 36(1), pp.41-59.
Borovkov, A., 1998. Probability theory. Amsterdam: Gordon Breach.
Caron, F. and Fox, E., 2017. Sparse graphs using exchangeable random measures. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79(5), pp.1295-1366.
Chacón, J. and Duong, T., 2009. Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices. TEST, 19(2), pp.375-398.
Cs.toronto.edu. 2022. Boston Dataset. [online] Available at:
[https://www.cs.toronto.edu/ delve/data/boston/bostonDetail.html]
Da Veiga, S., 2014. Global sensitivity analysis with dependence measures. Journal of Statistical Computation and Simulation, 85(7), pp.1283-1305.
Duong, T. and Hazelton, M., 2003. Plug-in bandwidth matrices for bivariate kernel density estimation. Journal of Nonparametric Statistics, 15(1), pp.17-30.
Fienberg, S. and Gilbert, J., 1970. The Geometry of a Two by Two Contingency Table. Journal of the American Statistical Association, 65(330), pp.694-701.
Granger, C., Maasoumi, E. and Racine, J., 2004. A Dependence Metric for Possibly Nonlinear Processes. Journal of Time Series Analysis, 25(5), pp.649-669.
Hanchuan Peng, Fuhui Long and Ding, C., 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8), pp.1226-1238.
Harrison, D. and Rubinfeld, D., 1978. Hedonic housing prices and the demand for clean air. Journal of Environmental Economics and Management, 5(1), pp.81-102.
Hastings, K., 1997. Probability and statistics. 1st ed. Reading, Mass.: Addison-Wesley.
Jones, M., Marron, J. and Sheather, S., 1996. A Brief Survey of Bandwidth Selection for Density Estimation. Journal of the American Statistical Association, 91(433), pp.401-407.
Kinney, J. and Atwal, G., 2014. Equitability, mutual information, and the maximal information coefficient. Proceedings of the National Academy of Sciences, 111(9), pp.3354-3359. Kroese, D., Taimre, T. and Botev, Z., 2011. Handbook of Monte Carlo Methods. John Wiley Sons, Inc.
Lyons, R., 2013. Distance covariance in metric spaces. The Annals of Probability, 41(5).
Martínez-Gómez, E., Richards, M. and Richards, D., 2014. DISTANCE CORRELATION METHODS FOR DISCOVERING ASSOCIATIONS IN LARGE ASTROPHYSICAL DATABASES.

The Astrophysical Journal, 781(1), p. 39.
Myers, J. and Well, A., 2003. Research design and statistical analysis. Mahwah, N.J.: Lawrence Erlbaum Associates.

Móri, T. and Székely, G., 2018. Four simple axioms of dependence measures. Metrika, 82(1), pp.1-16.
Reshef, D., Reshef, Y., Finucane, H., Grossman, S., McVean, G., Turnbaugh, P., Lander, E., Mitzenmacher, M. and Sabeti, P., 2011. Detecting Novel Associations in Large Data Sets. Science, 334(6062), pp.1518-1524.

Rényi, A., 1959. On measures of dependence. Acta Mathematica Academiae Scientiarum Hungaricae, 10(3-4), pp.441-451.
Rodgers, J. and Nicewander, W., 1988. Thirteen Ways to Look at the Correlation Coefficient. The American Statistician, 42(1), p.59.

Rosenblatt, M., 1956. Remarks on Some Nonparametric Estimates of a Density Function. The Annals of Mathematical Statistics, 27(3), pp.832-837.
Shannon, C., 1948. A Mathematical Theory of Communication. Bell System Technical Journal, 27(3), pp.379-423.
Silverman, B., 1998. London: Chapman Hall.
SKAUG, H. and TJØSTHEIM, D., 1993. A nonparametric test of serial independence based on the empirical distribution function. Biometrika, 80(3), pp.591-602.
Stamey, T., Kabalin, J., Mcneal, J., Johnstone, I., Freiha, F., Redwine, E. and Yang, N., 1989. Prostate Specific Antigen in the Diagnosis and Treatment of Adenocarcinoma of the Prostate. II. Radical Prostatectomy Treated Patients. Journal of Urology, 141(5), pp.10761083.

Székely, G., Rizzo, M. and Bakirov, N., 2007. Measuring and testing dependence by correlation of distances. The Annals of Statistics, 35(6)..
The Census users' Guide, U., 2022. 1970 Census: Users' Guide. [online] Census.gov. Available at: [https://www.census.gov/library/publications/1970/dec/users-guide.html]
Tibshirani, R., 1996. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), pp.267-288.
Van der Vaart, A., 2007. Asymptotic statistics. Cambridge: Cambridge University Press.
White, J., Steingold, S. and Fournelle, C., 2004. Performance Metrics for Group-Detection Algorithms.
Wijayatunga, P., 2016. A geometric view on Pearsons correlation coefficient and a generalization of it to non-linear dependencies. RATIO MATHEMATICA, 30.

Commons.wikimedia.org. 2022. Wikimedia Commons. [online] Available at:
[https://commons.wikimedia.org/wiki/File:Correlation ${ }_{e}$ xamples2.svg].
Witten, I. and Frank, E., 2005. Data Mining. San Diego: Elsevier Science Technology Books.
Zou, H. and Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Met

