Ensemble Correlation Coefficient

Wejdan Deebani and Nezamoddin N. Kachouie

Department of Mathematical Sciences, Florida Institute of Technology

Abstract

Elements in a sample date are demonstrated based on their characteristics and in turn the characteristics are represented by variables. Identifying the relationship between these variables is crucial for prediction, hypothesis testing, and decision making. The relation between two variables is often quantified using a correlation factor. Once correlation is known it can be used to make predictions. It means when two variables are highly correlated, and if we have observed one variable, we can make a prediction about the other variable. A more accurate prediction will be made where there is strong relationship between variables. Among several correlation factors, Pearson correlation Coefficient has been commonly used. Distance correlation and maximal information coefficient have been introduced recently to address the shortcomings of Pearson correlation coefficient. In this paper, we compare these factors through a set of simulations and combine them to introduce a more robust factor that can be generally used.

Introduction

Data analysis is crucial in almost every field of research such as genomics, economics, physics, medical, social, and political sciences. Identifying associations between/among variables is often required in analysis of large datasets (Hastie, Tibshirani, and Friedman 2002). It is common to have many variables in a dataset and it is difficult to manually examine the relation between each pair of variables (Reshef et al. 2011). It is also difficult to identify the important variables if the correlation among them is not discovered.

There are several different measures to quantify the association between variables in a dataset including Pearson's correlation, Spearman's correlation, distance correlation, maximal information coefficient (MIC), maximal correlation, and mutual information. Some of these correlation measures can only detect linearly correlated data such as the well-known Pearson's correlation while some measures can also detect nonlinear correlation such as maximal correlation and MIC. In addition, some correlation measures can characterize the independence. This means if the correlation score yields a value of zero, one can conclude that the two variables are independent.

Several correlation measures are discussed in this paper. These measures are Pearson's correlation, Spearman's correlation, distance correlation, maximal information coefficient (MIC), and maximal correlation. In addition, we proposed some preliminary ways to combine these coefficients metrics. The idea of combining such metrics is useful especially when the luxury of knowing the underline relationship is not provided. The preliminary ways of combining them that are considered are finding the maximum, mean, and median of these measures. In the method section, the correlation measures that used throughout the paper as well as the ways of combining them are introduced.

Methods

To explore the performance of correlation measures, various measures are selected to closely analyzed them. These correlation measures are Pearson's correlation, Spearman's correlation, distance correlation, maximal correlation, and maximal information coefficient (MIC).

Pearson's Correlation

Pearson's correlation is a measure of strength and direction of the linear relationship between two variables. Its score ranges between -1 and 1, and it describes the degree to which one variable is linearly related to another. Pearson's correlation between two variables X and Y can be written as

$$\rho = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{1}$$

,where $cov(X,Y)$ is the the covariance between X and Y, σ_X is the standard deviation of X, and σ_Y is the standard deviation of Y. So, Pearson's correlation is the covariance between X and Y divided by the product of standard deviation of X and the standard deviation of Y.

Spearman's Correlation

Another measure considered is Spearman's correlation. It assesses how well a monotonic function could describe the relationship between two variables. It is a non-parametric measure of correlation between variables. Spearman's cannot only be used on numerical data but also on any data that can be ranked (Spearman 1904). Suppose that $\mathbf{R} = (R_1, R_2, ..., R_n)$ and $\mathbf{Q} = Q_1, Q_2, ..., Q_n)$ are two vectors

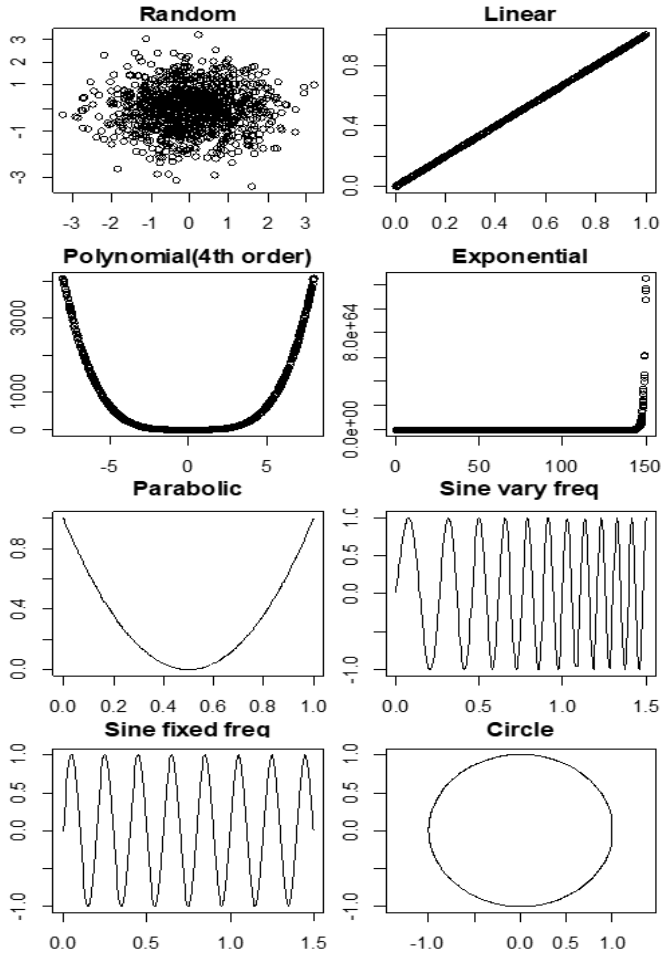

Figure 1: Random, linear, polynomial (fourth order), exponential, parabolic, sinusoidal with varying frequency, sinusoidal with fixed frequency, and circle relationship types for first simulation where there is no noise added to the true signal.

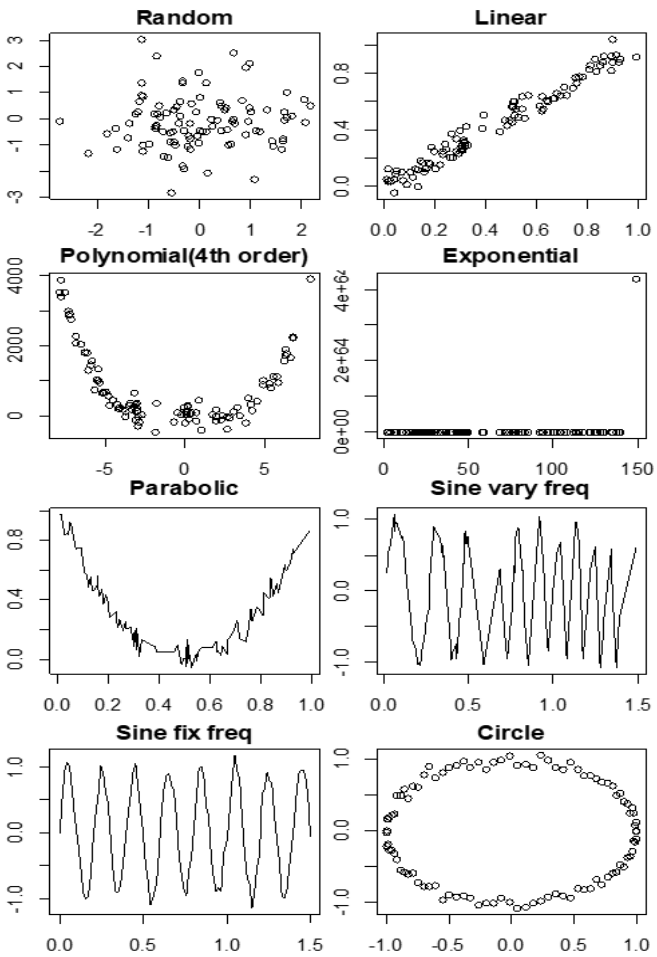

Figure 2: Random, linear, polynomial (fourth order), exponential, parabolic, sinusoidal with varying frequency, sinusoidal with fixed frequency, and circle relationship types for second simulation where the true signal is corrupted with low noise level.

of ranks obtained on a sample of size n, then Spearman's coefficient, r_s, is given by (da Costa 2015):

$$r_s = 1 - \frac{6\sum_{i=1}^{n}(R_i - Q_i)^2}{n^3 - n} \tag{2}$$

Spearman's takes any real values between -1 and 1. If the values in the series increase or decrease together, it will have a positive value. However, if a values of one variable increase as the other decreases, a negative spearman's value is obtained (da Costa 2015). In addition, no assumptions about the frequency distribution of the variables are required in Spearman's correlation, and no assumption about the existing of linear relationship between variables is required (Bolboaca and Jäntschi 2006).

Distance Correlation

Distance correlation, denoted by $\mathcal{R}$, is a measure of dependence between two random vectors X and Y, and its value ranges from zero to one. It uses Euclidean distance in its formula, and it has two main properties: First, distance correlation is zero if and only if the two random vectors are independence, which means that obtaining zero value characterizes the independence between the two vectors. Second, distance correlation can be calculated between two vectors of arbitrary dimensions; for example, $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, where p and q are positive integers. This measure only discovers the linear correlation and cannot discover the nonlinear correlation. The empirical distance correlation $\mathcal{R}_n(X, Y)$ is the square root of the following (Székely et al. 2007)

$$\mathcal{R}_n^2(X,Y) = \begin{cases} \frac{\nu_n^2(X,Y)}{\sqrt{\nu_n^2(X)\nu_n^2(Y)}}, & \nu_n^2(X)\nu_n^2(Y) > 0, \\ 0, & \nu_n^2(X)\nu_n^2(Y) = 0 \end{cases} \tag{3}$$

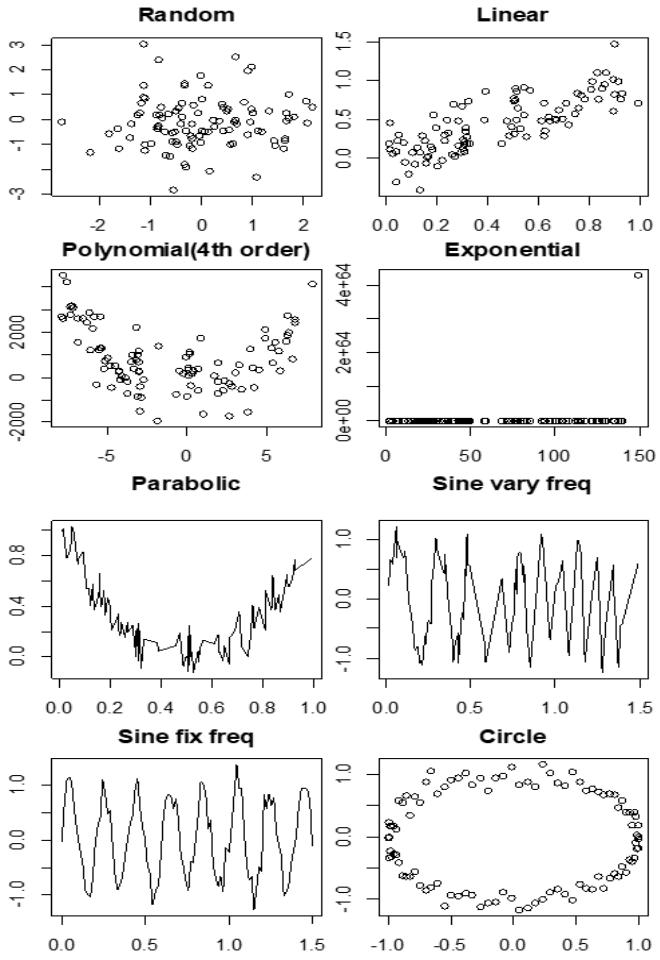

Figure 3: Random, linear, polynomial (fourth order), exponential, parabolic, sinusoidal with varying frequency, sinusoidal with fixed frequency, and circle relationship types for third simulation where true signal is corrupted with medium level of noise.

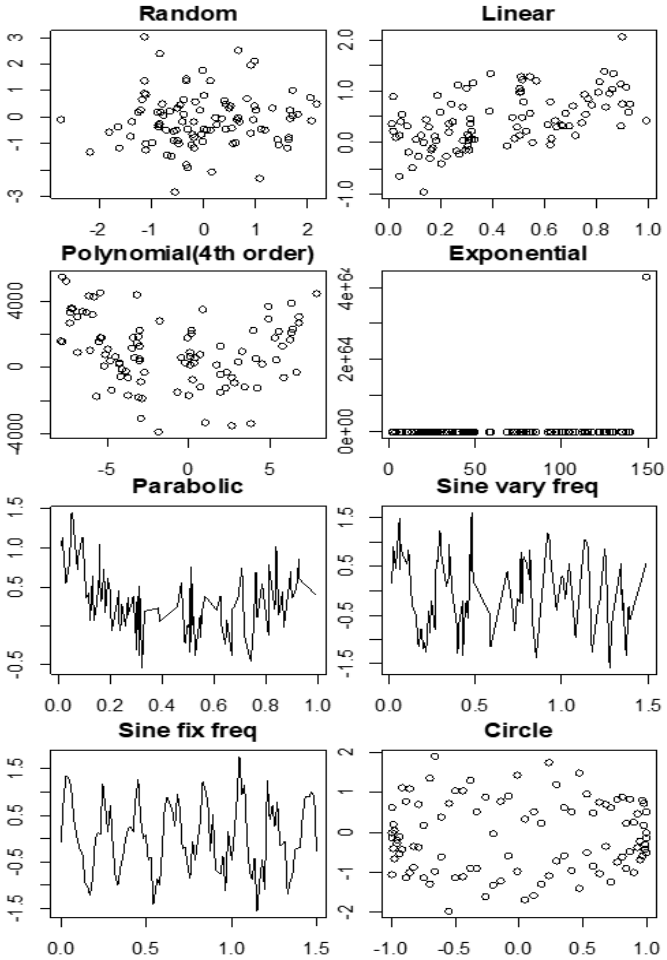

Figure 4: Random, linear, polynomial (fourth order), exponential, parabolic, sinusoidal with varying frequency, sinusoidal with fixed frequency, and circle relationship types for last simulation where true signal is corrupted with high level of noise.

,where

$$\nu_n^2(X,Y) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}B_{kl} \tag{4}$$

$$a_{kl} = |X_k - X_l|_p\,, \qquad \bar{a}_{k.} = \frac{1}{n}\sum_{l=1}^{n} a_{kl},$$

$$\bar{a}_{.l} = \frac{1}{n}\sum_{k=1}^{n} a_{kl}, \qquad \bar{a}_{..} = \frac{1}{n^2}\sum_{k,l=1}^{n} a_{kl},$$

$$A_{kl} = a_{kl} - \bar{a}_{k.} - \bar{a}_{.l} + \bar{a}_{..},$$

Similarly

$$b_{kl} = |Y_k - Y_l|_q\,, \qquad B_{kl} = b_{kl} - \bar{b}_{k.} - \bar{b}_{.l} + \bar{b}_{..},$$

$$k, l = 1, ..., n$$

Also,

$$\nu_n^2(X) = \nu_n^2(X,X) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}^2 \tag{5}$$

$$\nu_n^2(Y) = \nu_n^2(Y,Y) = \frac{1}{n^2}\sum_{k,l=1}^{n} B_{kl}^2 \tag{6}$$

Maximal Correlation

Breiman and Friedman (Breiman and Friedman 1985) defined the Maximal Correlation (MC) between two real valued random variables X and Y as

$$\rho^* = max_{f_1,f_2}\rho(f_1(X), f_2(Y)) \tag{7}$$

where $f_1 : \mathbb{R} \to \mathbb{R}$, and $f_2 : \mathbb{R} \to \mathbb{R}$, are two arbitrary measurable mean-zero functions of X and Y, respectively. So, maximal correlation is an optimization problem that trying to search for transformations of X and Y such

that Pearson's correlation between transformed X and Y is maximized (Nguyen et al. 2014). Maximal correlation does not require assumptions on the data distribution. It can detect non-linear correlations, and it is very efficient and robust to noise.

Maximal Information Coefficient

Another important measure of independence is Maximal Information Coefficient (MIC). MIC takes value between zero and one, and it has two main properties: Generality and equitability. Generality means that with sufficiently large sample size, the statistic should capture a wide range of association such as linear, exponential, or periodic. Equitability means that MIC gives similar scores to equally noisy relationships regardless the type of relationships. In addition, with probability approaching 1 as sample size grows, MIC gives scores of one to all noiseless functional relationships and gives scores that tend to 0 to statistically independent variables. An advantage of MIC is the ability to catch non-linear associations as well as linear associations (Reshef et al. 2011). It has no assumption about the distribution of the measured data. According to Szkely and Rizzo, MIC has simple computing formula, it applies to sample sizes $n \geqslant 2$. MIC of two vectors x and y is defined as follows (Zhang et al. 2014)

$$MIC = max\left\{\frac{I(x,y)}{log_2 min\left\{n_x, n_y\right\}}\right\} \tag{8}$$

, where

$$\begin{aligned} I(x,y) &= H(x) + H(y) - H(x,y) \\ &= \sum_{i=1}^{n_x} p(x_i) log_2 \frac{1}{p(x_i)} + \sum_{j=1}^{n_y} p(y_j) log_2 \frac{1}{p(y_j)} \\ &\quad - \sum_{i=1}^{n_x}\sum_{j=1}^{n_y} p(x_i, y_j) log_2 \frac{1}{p(x_i, y_j)} \end{aligned} \tag{9}$$

and n_x and n_y are the number of bins of the partition of the $x - axis$ and $y - axis$, respectively. Also, $n_x.n_y < B(n)$, $B(n) = n^{0.6}$

Ensemble Correlation Coefficient

Different correlation measures perform differently when they are applied to different data sets. Some of them detect linear relations while other detect non-linear relations. In fact, the nonlinear correlation measures themselves, such as the maximal correlation and MIC, perform differently and give different scores with non-linear relationships. If one knew that the relationship between a pair of variables is monotonic, he/she would choose Spearman's coefficient since it is designed to detect such a relation. On the other hand, in real application, the underline relationship between a pair of variables is unknown, and one may face difficulty to decide which correlation method should be trusted. Consequently, selecting the suitable coefficient to detect the association could be challenging. Therefore, applying a wide range of correlation coefficients and ensembling the scores could result in an improved or more robust score for different data sets. In this work, we address this issue and propose some statistics that involve the joint contribution of several correlation methods and can be used to ensemble their results.

After obtaining the aforementioned correlation measures, we propose the use of all coefficients in an ensemble metric that is calculated as the maximum, mean, or median of the coefficients. It is important to notice that distance correlation, maximal correlation, and MIC values range between 0 and 1; however, both Pearson's and Spearman's correlations range between -1 and 1. Consequently, in order to find the max, mean, and median, we square Pearson's and Spearman's correlations to obtain values that range between 0 and 1 for all correlation measures. We have conducted several simulations to explore these correlation measures and aggregate them to a single representing value as follows.

Simulation

To explore different correlation measures, four separate simulations were performed. The number of iterations for each simulation is 100. All simulations compute Pearson's correlation, Spearman's correlation, distance correlation, maximal correlation, and maximal information coefficient (MIC) for the following relationship types similar to (Reshef et al. 2011): Random, linear, polynomial of fourth order, exponential, parabolic, sinusoidal with varying frequency, sinusoidal with fixed frequency, and circle. We have extended the simulations in (Reshef et al. 2011) by adding several different scenarios with additive noise levels. For more discussion in the comparison between Pearson's correlation, distance correlation, and MIC regarding their power under several relationships, please see (Simon and Tibshirani 2014). They highlighted that MIC does not always perform best comparing with Pearson's and distance correlation. This was achieved by obtaining the power of these three measures with different relationship types and additive noise.

The first simulation we have is performed on noiseless true signals with aforementioned relationships while the second, third, and fourth simulations are performed on true signal corrupted with low (about 5%), medium (about 20%), and high (about 40%) level of noise respectively. Moreover, in each of the four simulations, we find maximum, mean, and median as our preliminary way to combine the aforementioned measures.

Results

The simulated data is generated for variety of relationship types under four different cases: First case does not include noise, second case includes low level noise, third case includes medium level noise, last case includes high level noise. Then, Pearson's correlation, Spearman's correlation, distance correlation, maximal correlation, and maximal information coefficient are applied as well as finding the maximum, mean, and median of these statistics.

The first simulation deals with noiseless data generated for various relationship types, and it computes the scores of several statistics, which are Pearson's correlation, Spearman's correlation, distance correlation, maximal correlation, and maximal information coefficient. One can visualize

Table 1: Scores of Pearson's correlation, Spearman's correlation, distance correlation, maximal correlation, maximal information coefficient, max, mean, and median correlation measures given to various noiseless functional relationships

Relationship Type	Pearson	Spearman	Distance Correlation	Maximal Correlation	MIC	Max	Mean	Median
Random	0.03	0.04	0.06	0.07	0.13	0.13	0.07	0.06
Linear	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Polynomial (4th order)	0.06	0.04	0.47	1.00	1.00	1.00	0.51	0.47
Exponential	0.22	1.00	0.29	1.00	1.00	1.00	0.70	1.00
Parabolic	0.05	0.04	0.50	1.00	1.00	1.00	0.52	0.50
Sinusoidal (varying freq)	-0.06	-0.05	0.10	0.78	1.00	1.00	0.38	0.10
Sinusoidal (fixed freq	8.17E-18	-1.11E-04	0.09	0.99	1.00	1.00	0.42	0.09
Circle	2.15E-19	1.49E-03	0.20	1.00	0.68	1.00	0.38	0.20

Table 2: Scores of Pearson's correlation, Spearman's correlation, distance correlation, maximal correlation, maximal information coefficient, max, mean, and median correlation measures given to various low noisy functional relationships

Relationship Type	Pearson	Spearman	Distance Correlation	Maximal Correlation	MIC	Max	Mean	Median
Random	0.06	0.08	0.15	0.18	0.31	0.31	0.16	0.15
Linear	0.98	0.98	0.98	0.99	1.00	1.00	0.99	0.98
Polynomial (4th order)	-0.23	-0.26	0.48	0.99	0.85	0.99	0.49	0.48
Exponential	0.20	0.89	0.20	1.00	1.00	1.00	0.66	0.89
Parabolic	-0.21	-0.28	0.51	0.99	0.98	0.99	0.52	0.51
Sinusoidal (varying freq)	-0.07	-0.07	0.14	0.71	0.53	0.71	0.28	0.14
Sinusoidal (fixed freq)	-3.44E-04	-1.68E-04	0.10	0.94	0.38	0.94	0.28	0.10
Circle	-4.37E-04	-2.26E-03	0.20	0.99	0.58	0.99	0.35	0.20

these noiseless relationship types in figure 1. Table 1 shows the scores obtained from the first simulation. All correlation measures assign low scores to random relationship as expected and appeared in the first row of table 1. It is clear that Pearson's correlation assigns score of one to perfectly linear relationship since there is no noise considered. However, Pearson's correlation gives low scores to the rest of relationships since they are nonlinear correlation. Spearman's correlation gives score of one to linear, and exponential relationships since Spearman's correlation assesses monotonic relationships whether linear or nonlinear. It gives low scores to the rest of relationship types because they do not present a monotonic behavior. As Pearson's correlation, distance correlation results in high scores with linear relationships while results in low scores with the rest of relationships since distance correlation does not catch nonlinear relationships. Additionally, considering the circle relationship, both Pearson's and distance correlations have low scores, but clearly distance correlation has a very much higher value (0.20) with circle relation than Pearson's correlation ($2.15e^{-19}$). Maximal correlation assigns high scores to all relationship types except of course the random relation. Maximal information coefficient (MIC) assign scores of one to linear, polynomial of fourth order, exponential, parabolic, sinusoidal with varying frequency, and sinusoidal with fixed frequency relationships since these scores are obtained from noiseless relationship, and the similarities in these scores in the case where there are no noise added proves the equitability property of MIC that Reshef introduced (Reshef et al. 2011). The score that is obtained by MIC for the circle relationship is 0.68 which is not equal to one for this noiseless true relationship (figure 1). As a result, MIC does not satisfy the equitability property with circle relation. This was also proven by Kinney and Atwal (Kinney and Atwal 2014). It is important to point out that in circle relationship, maximal correlation (MC) outperforms MIC (MC=1.00 , MIC=0.68) although the circle relation is noiseless. The maximum of all correlation measures in all relationship except random relation is one since these are true noiseless relations. The mean of all correlation measures is larger than the median except for exponential relation. This is because in exponential relation, three of the measures receive scores of ones, and the other two receive low scores, impacting the mean to be affected by the small values while the median is 1.

The second simulation deals also with same relationships corrupted with low noise. These corrupted relationships can be seen in figure 2. It should be noted that the added noise to the exponential relationship is not visible because of high amplitude of the signal (y-axis). Table 2 shows all correlation scores as well as maximum, mean, and median of the correlation measures computed for different relationships. All correlation measures have high scores of about 0.98 for linear relationship. For polynomial of order four and parabolic relationships, only maximal correlation and MIC perform well, since they can detect nonlinear relationship, and maximal correlation is slightly higher

Table 3: Scores of Pearson's correlation, Spearman's correlation, distance correlation, maximal correlation, maximal information coefficient, max, mean, and median correlation measures given to various medium noisy functional relationships

Relationship Type	Pearson	Spearman	Distance Correlation	Maximal Correlation	MIC	Max	Mean	Median
Random	0.06	0.08	0.15	0.18	0.31	0.31	0.16	0.15
Linear	0.82	0.82	0.80	0.84	0.71	0.84	0.80	0.82
Polynomial (4th order)	-0.19	-0.22	0.39	0.86	0.49	0.86	0.36	0.39
Exponential	0.20	0.89	0.20	1.00	1.00	1.00	0.66	0.89
Parabolic	-0.21	-0.27	0.49	0.96	0.86	0.96	0.48	0.49
Sinusoidal (varying freq)	-0.07	-0.06	0.15	0.68	0.49	0.68	0.26	0.15
Sinusoidal (fixed freq)	-6.98E-04	2.35E-05	0.11	0.92	0.39	0.92	0.28	0.11
Circle	-8.61E-04	-2.53E-03	0.19	0.97	0.53	0.97	0.34	0.19

Table 4: Scores of Pearson's correlation, Spearman's correlation, distance correlation, maximal correlation, maximal information coefficient, max, mean, and median correlation measures given to various high noisy functional relationships

Relationship Type	Pearson	Spearman	Distance Correlation	Maximal Correlation	MIC	Max	Mean	Median
Random	0.06	0.08	0.15	0.18	0.31	0.31	0.16	0.15
Linear	0.58	0.57	0.56	0.61	0.43	0.61	0.55	0.57
Polynomial (4th order)	-0.14	-0.17	0.29	0.63	0.34	0.63	0.26	0.29
Exponential	0.20	0.89	0.20	1.00	1.00	1.00	0.66	0.89
Parabolic	-0.16	-0.21	0.37	0.75	0.48	0.75	0.34	0.37
Sinusoidal (varying frequency)	-0.07	-0.06	0.16	0.58	0.41	0.58	0.23	0.16
Sinusoidal (fixed frequency)	-1.34E-03	-7.46E-04	0.13	0.77	0.37	0.77	0.26	0.13
Circle	-2.92E-03	-5.50E-03	0.16	0.60	0.24	0.60	0.20	0.16

than MIC. For exponential relationship, maximal correlation, MIC, and Spearman's correlation identify the relation with high scores. However, in other nonlinear relationships, sine wave with varying frequency, sine wave with fixed frequency, and circle, maximal correlation clearly outperforms MIC. For example, in circle relationship, maximal correlation score is 0.99 while MIC score is 0.58. Moreover, the proposed ensemble correlation measure yields a score that robustly demonstrates the relationship for all different cases. Mean correlation is better than median correlation factor in sinusoidal and circle relations, but the median is better than the mean correlation for exponential. The mean and median correlation factors are about the same for other relationships.

The last two simulations consider adding medium and high noise levels to the relationships as demonstrated in figure 3 and figure 4, respectively. The corresponding correlation scores are shown in table 3 and table 4, respectively. Also, as shown in table 3 and table 4, comparing between distance correlation and MIC under noisy linear relation shows that distance correlation performs better than MIC. This agrees with what Simon and Tibshirani (Simon and Tibshirani 2014) have commented regarding the power of these two measures. Their experimental analysis include computing the power of three correlation methods, two of which are distance correlation, and MIC for several relations, one of which is linear. They have shown that in linear relation, distance correlation has higher power than MIC as the noise level increases. The Proposed ensemble factor obtains a high score of 0.61 by averaging several correlation measures such as distance correlation and MIC.

Conclusions

For prediction and decision making purposes, we need to discover the relationship between different variables in a dataset. Often, the relation between two variables is quantified and represented by a correlation coefficient. The correlation can then be used to make predictions and informed decisions. Strong relationship between variables can help with more accurate predictions. There are several correlation coefficients such Pearson's correlation, distance correlation, and maximal information coefficients. Each of which can better identify a specific type of relationship between two variables. We have studied them in this work and proposed different ways to ensemble them to a single representative coefficient. It is desirable to have a single factor capable of explaining the relationship between variables in a dataset. Although, the proposed aggregation methods are preliminary ways to combine different correlation coefficients, this work provides some insights about an interesting idea. Our future research will be conducted to develop a more sophisticated method to aggregate different correlation coefficients.

References

Bolboaca, S.-D., and Jäntschi, L. 2006. Pearson versus spearman, kendall's tau correlation analysis on

structure-activity relationships of biologic active compounds. *Leonardo Journal of Sciences* 5(9):179–200.

Breiman, L., and Friedman, J. H. 1985. Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association* 80(391):580–598.

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

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

Kinney, J. B., and Atwal, G. S. 2014. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences* 111(9):3354–3359.

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

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

Simon, N., and Tibshirani, R. 2014. Comment on" detecting novel associations in large data sets" by reshef et al, science dec 16, 2011. *arXiv preprint arXiv:1401.7645*.

Spearman, C. 1904. The proof and measurement of association between two things. *The American journal of psychology* 15(1):72–101.

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

Zhang, Y.; Jia, S.; Huang, H.; Qiu, J.; and Zhou, C. 2014. A novel algorithm for the precise calculation of the maximal information coefficient. *Scientific reports* 4.