Systematically Exploring Associations among Multivariate Data

Lifeng Zhang
School of Information, Renmin University of China
59, Zhongguancun Street, Haidian
Beijing, P.R.China, 100872
l.zhang@ruc.edu.cn

Abstract

Detecting relationships among multivariate data is often of great importance in the analysis of high-dimensional data sets, and has received growing attention for decades from both academic and industrial fields. In this study, we propose a statistical tool named the neighbor correlation coefficient (*nCor*), which is based on a new idea that measures the local continuity of the reordered data points to quantify the strength of the global association between variables. With sufficient sample size, the new method is able to capture a wide range of functional relationship, whether it is linear or nonlinear, bivariate or multivariate, main effect or interaction. The score of *nCor* roughly approximates the coefficient of determination (R^2) of the data which implies the proportion of variance in one variable that is predictable from one or more other variables. On this basis, three *nCor* based statistics are also proposed here to further characterize the intra and inter structures of the associations from the aspects of nonlinearity, interaction effect, and variable redundancy. The mechanisms of these measures are proved in theory and demonstrated with numerical analyses.

Introduction

Identifying relationships among variables is one of the most critical issues in data analysis and interpretation (Altman and Krzywinski 2015) with a wide range of applications in diverse fields from data science to neuroscience. Nowadays, however, a large data set may contain a vast number of variable pairs and combinations that are difficult to be examined manually (Reshef et al. 2011). Association measures can be used to quickly find out the significant associations scattered in thousands or even millions of potential relationships without modelling the relationships explicitly, and thereby provide valuable knowledge and promising pointers for future study.

Consider a data sample $\{(\mathbf{x}_{(t)}, y_{(t)})|1\leq t\leq N\}$ that is observed from an underlying functional relationship expressed as follows.

$$y = f(\mathbf{x}) + e = \sum_{x_i\in\mathbf{x}} g_i(x_i) + \sum_{\mathbf{x}_i\subseteq\mathbf{x}} h_i(\mathbf{x}_i) + e \qquad (1)$$

where $y\in\mathbb{R}$, $\mathbf{x} = (x_i|1\leq i\leq M)\in\mathbb{R}^M$, $e\in\mathbb{R}$, and $M\geq 2$ respectively denote dependent variable, multiple independent variables, subset of $\mathbf{x}$, and additive noise. $f(\cdot)$, $g_i(\cdot)$ and $h_i(\cdot)$ denote the underlying function, main effect and interaction effect respectively.

If $f(\cdot)$ is linear in which all $g_i(\cdot)$ are linear and all $h_i(\cdot)$ are null, the Pearson correlation coefficient should be a perfect measure of how much fluctuation in one variable can be explained by another variable (R^2) (Altman and Krzywinski 2015). If $f(\cdot)$ is nonlinear, the traditional correlation test is no longer sufficient. Developing concise and efficient nonlinear association detection methodologies has been a challenging research and received wide attention for over a half century.

Some of the previous approaches have been developed based on the theory of mutual information (MI) and partitioning (binning) techniques. A typical one is the maximal information coefficient (MIC), which is the state-of-the-art association measure that has been extensively evaluated recently (Reshef et al. 2011; Reshef et al. 2018). This kind of methods use partitioning as a means to apply MI on continuous random variables based on the idea that, if an association exists between two variables, then a grid can be drawn on the scatterplot that partitions the data to encapsulate that relationship. Similarly, Heller and Gorfine's SDDP adopted summation or maximization aggregation of the scores over all partitions of a fixed size to estimate the MI (Heller et al. 2016). In addition, other techniques, such as kernel density estimation (KDE), k-nearest neighbor distances (kNN), and nonlinear correlation information entropy (NICE), also can be used to compute the score of MI as a dependence measure (Moon, Rajagopalan, and Lall 1995; Darbellay and Vajda 1999; Kraskov, Stogbauer, and Grassberger 2004; Wang, Shen, and Zhang 2005).

Another method named distance correlation (*dCor*) has a compact representation analogous to the Pearson correlation coefficient, however is calculated based on certain Euclidean distances between sample elements (Székely, Rizzo, and Bakirov 2007; Székely and Rizzo 2009). *dCor* can be viewed as a special case of kernel based method (Sejdinovic et al. 2013), which is a kind of more general statistic defined on reproducing kernel Hilbert spaces (Gretton et al. 2008; Gretton and Gyorfi 2010). Empirical studies (Reshef et al. 2015; Reshef et al. 2018) showed that *dCor* also achieved excellent performance in some situations. In order to construct

a distribution free test, Székely and Rizzo (2009) considered using the ranks of each random variable instead of the actual values in computing *dCor*, and Heller, Heller, and Gorfine (2013) introduced an association test based on the cross-classification of the distances from center points.

From the 1980s, a number of higher order correlation measures have been proposed to construct concise nonlinear model validity tests (Aguirre 1995; Billings and Zhu 1995; Mao and Billings 2000) for system identification. Zhang, Zhu, and Longden (2007) and Zhu, Zhang, and Longden (2007) introduced a set of first order correlation functions, named omni-directional cross-correlation functions (ODCCF) by considering the symmetrical properties of nonlinear relationship. Other extensions of the ordinary correlation test include the Spearman rank correlation coefficient, Kendall coefficient of concordance (Kendall 1938), maximal correlation (Rényi 1959; Breiman and Friedman 1985), principal curve based methods (Hastie and Stuetzle 1989; Delicado 2001; Delicado and Smrekar 2009), randomized dependence coefficient (RDC) (Lopez-Paz, Hennig, and Scholkopf 2013), and nonlinear spectral correlation (Liu, Sohn, and Jeon 2017), which all capture a certain range of nonlinear relationships.

Nevertheless, these approaches still cannot effectively detect associations in a satisfactory manner under every condition, since they are incapable of equitability estimating the R^2 of the relationships and show strong preference for some types of nonlinear functions (Reshef et al. 2018). In addition, the overwhelming majority of the existing methods are designed for pairwise association detection, and thus unable to detect interaction effects. If only main effects exist among multiple variables, pairwise test should be sufficient since the influence of the variables is separable in such situation. In contrast, interaction effect describes a situation in which the simultaneous influence of two or more independent variables is not additive. In the real world, actually, many bivariate relationships appear to be insignificant or non-functional, but can in fact be explained by interaction effects. Due to the complexity of interactions, sometimes, there is no any trend, principle curve, or particular pattern identifiable in the pairwise tests, and even various fitting or transformation would be of no avail. Therefore, whenever interactions occur all the bivariate analysis techniques tend to be less effective. Although *dCor* and some MI estimators can handle multivariate data, they are still incapable of distinguishing interactions from main effects in all cases.

In the present study, we propose a new method named the neighbor correlation coefficient (*nCor*) to detect the relationships among data sequences in both the bivariate and multivariate cases with the following properties. (i) With sufficient sample size, the method could capture a wide range of functional relationships including not only various bivariate functional forms such as exponential or periodic, but also multivariate associations and in particular interaction effects. (ii) The method roughly measures the association strength (R^2) of the data that have the same statistical power increasing with sample size for whatever functional relationship. (iii) The method can be used to further characterize and distinguish the inter and intra structure of the detected relationship from the aspects of nonlinearity, interactivity, and variable redundancy. Finally, *nCor* differs from the previous approaches in that it detects associations by measuring the local continuity of the concomitants obtained from data reordering, rather than partitioning the scatterplots, estimating the probability distributions, or computing with pairwise distances (and reproducing kernel Hilbert spaces). For this reason, in the new method, the independent variables are only used for reordering data points, but not involved in the computation of correlation scores at all. This provides an alternative way to assess the relationships among multivariate data.

The rest of this study is organized as follows. In the next two sections, *nCor* and three *nCor* based statistics are proposed. Subsequently, empirical studies are performed to evaluate the effectiveness of the new statistics and make comparisons with previous approaches. In the last Section, conclusions are drawn to summarize the study. To reduce the length of the study, more analyses and experimental studies are enclosed in supplementary material.

Neighbor correlation coefficient (*nCor*)

To simplify the proofs, without losing generality, throughout this study x is assumed to be continuous and uniformly distributed, and $f(\cdot)$, $g_i(\cdot)$ and $h_i(\cdot)$ are all assumed to be continuous functions. Supplementary material gives the theoretical proofs of all the lemmas and theorems, as well as the empirical proofs on the robustness of the new method against data distribution and function continuity. Actually, the new method exhibits almost same performance under different data distributions when the sample size is sufficiently large (In Supplementary Material, we tested 8 continuous and discrete distributions including uniform, normal, exponential, and bimodal).

nCor for bivariate data

Consider a set of paired data $\{(x_{(t)}, y_{(t)})|1\leq t\leq N\}$. To detect the potential relationship $y = g(x) + e$, sample points need to be rearranged initially in an increasing order of the independent variable. The concepts of order statistics and concomitants are given as follows (David and Nagaraja 2003).

Order statistics: Sorting independent variable data with respect to its values to obtain a new sequence denoted by $x_{(1:N)}\leq x_{(2:N)}\leq\cdots\leq x_{(N:N)}$, where $x_{(k:N)}$ is known as the k-th order statistic of $\{x_{(t)}\}$. Let $\{n_{(k)}|1\leq k\leq N\}$ be the reordering permutation, that is, if $n_{(k)} = t$ then $x_{(k:N)} = x_{(t)}$.

Lemma 1. *Let $x_{(1)}, x_{(2)}, \cdots, x_{(N)}$ be a sample of a random variable which is continuous and uniformly distributed on $[a, b]$. Let $\Delta x_{(k:N)}$ be the difference between two neighboring order statistics of $\{x_{(t)}\}$ which can be derived as*

$$\Delta x_{(k:N)} = x_{(k+1:N)} - x_{(k:N)} \tag{2}$$

Then, it holds that

$$\lim_{N\to\infty} \Delta x_{(k:N)} = 0,\ \forall 1\leq k\leq N-1 \tag{3}$$

Concomitants: Rearranging dependent variable data in accordance with $\{n_{(k)}\}$ to yield a new sequence $y_{[1:N]}, y_{[2:N]}, \cdots, y_{[N:N]}$, where $y_{[k:N]}$ is known as the k-th concomitant, and defined as $y_{[k:N]} = y_{(t)}$ when $n_{(k)} = t$.

Lemma 2. *Let* $\{(x_{(t)}, y_{(t)})|1\leq t\leq N\}$ *be observed from a noise free continuous relationship* $y = g(x)$*, where* $x\in[a, b]$ *is a uniformly distributed random variable. Let* $\Delta y_{[k:N]}$ *be the difference between two neighboring concomitants which can be derived as*

$$\Delta y_{[k:N]} = y_{[k+1:N]} - y_{[k:N]} \tag{4}$$

Let Δy *denote the sequence of* $\{\Delta y_{[k:N]}\}$*. If* N *is sufficiently large, then* $\mathrm{Var}(\Delta y) < \mathrm{Var}(y)$*. In addition,*

$$\lim_{N\to\infty} \Delta y_{[k:N]} = 0,\ \forall 1\leq k\leq N-1 \tag{5}$$

Figure 1 shows the scatterplots of four typical data relationships that when $\Delta x_{(k:N)}$ is sufficiently small (due to large N), the amplitude of $\Delta y_{[k:N]}$ should be much smaller than that of $y_{(t)}(y_{[k:N]})$.

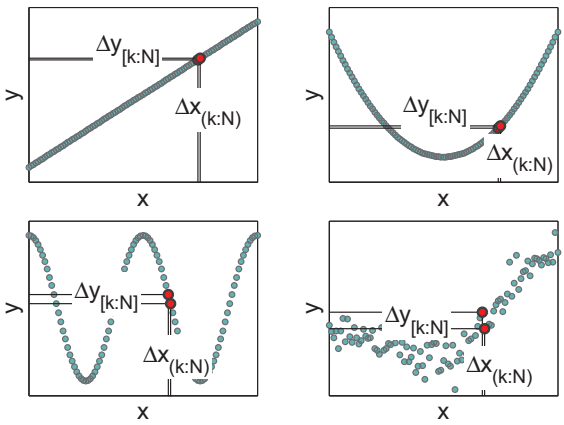


Figure 1: The scatterplots of four bivariate associations.

nCor measures how much knowing the independent variables determines the value of the dependent variable based on the idea that, if a continuous functional relationship exists, the data points which are very similar in the independent variables should also have similar values in the dependent variable. In such a situation, $y_{[k:N]}$ will exhibit a positive correlation with $y_{[k+1:N]} = y_{[k:N]} + \Delta y_{[k:N]}$. Pairwise neighboring concomitants, then, are used to compute *nCor* by means of the product-moment correlation coefficient as below.

Definition 1. *neighbor correlation coefficient (*nCor*). Let* (y', y'') *denote the paired sequences of the neighboring concomitants where* $y' = \{y_{[k:N]}|1 \leq k \leq N-1\}$ *and* $y'' = \{y_{[k+1:N]}|1 \leq k \leq N-1\}$*.* $nCor$ *is defined as*

$$nCor(x, y) = \frac{\mathrm{Cov}(y', y'')}{\sqrt{\mathrm{Var}(y')\mathrm{Var}(y'')}} \tag{6}$$

where $\mathrm{Cov}(\cdot)$ *and* $\mathrm{Var}(\cdot)$ *denote covariance and variance operators respectively.* $nCor$ *when applied to a sample can be calculated as (7).*

Theorem 1. *Let* $\{(\boldsymbol{x}_{(t)}, y_{(t)})\}$ *(*$|\boldsymbol{x}| \geq 1$*) be a data sample that is observed from random variables* $(\boldsymbol{x}, y)$*. If* $(\boldsymbol{x}, y)$ *are independent, then the expectation of the correlation coefficient has* $nCor(\boldsymbol{x}, y) = 0$*. When applied to a sample, a hypothesis test rejects the null hypothesis of independence if*

$$|nCor(\boldsymbol{x}, y)| > \tanh\left(\Phi^{-1}(1-\alpha/2)/\sqrt{(N-4)}\right) \tag{8}$$

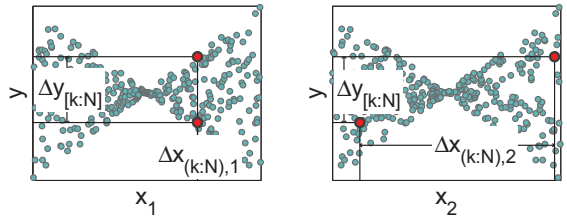
where $\Phi(\cdot)$ *denotes the standard normal cumulative distribution function and* α *is the significance level of the test.*

Theorem 2. *Let* $\{(x_{(t),i}, y_{(t)})\}$ *be paired data that is observed from the relationship as defined in* (1)*, and each* $x_i \in \boldsymbol{x}$ *is uniformly distributed on* $[a, b]$*. If a main effect* $g_i(x_i)$ *exists, with sufficient* N*, the correlation coefficient has* $nCor(x, y) > 0$*. In addition,*

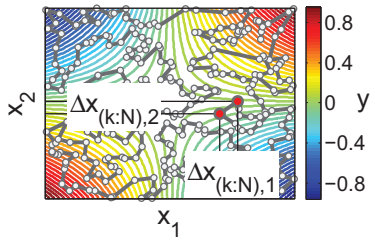
$$\lim_{N\to\infty} nCor(x, y) = \mathrm{Var}(g(x))/\mathrm{Var}(y) \tag{9}$$

nCor for multivariate data

When considering the relationship among three or more variables, an interaction effect may arise that the association between each of the interacting variables and the dependent variable depends on the values of the other interacting variable(s). Figure 2(a) clearly suggests that when an interaction occurs, the value of $y_{(t)}$ is unpredictable if only $x_{(t),1}$ or $x_{(t),2}$ is known. Although $\Delta x_{(k:N),1}$ is small, $\Delta y_{[k:N]}$ could be large due to the potentially large value of $\Delta x_{(k:N),2}$. In this case, the aforementioned order statistics based data rearranging is no longer sufficient.



(a)



(b)

Figure 2: The scatterplots (a) and contour map (b) of $y = x_1x_2$ with 100 sample points.

To address this problem, we convert the data reordering process to a travelling salesman problem (TSP) by considering the reordering permutation as a short route that visits each sample point in the multi-dimensional independent variable space exactly once. Then, $\{n_{(k)}\}$ obtained from solving TSP are applied to generate concomitants for computing $nCor$. Figure 2(b) shows that although a short route cannot make each $\{x_{(t),i}\}$ be rearranged in an ascending or descending

$$nCor(x,y) = \frac{(N-1)\sum_{k=1}^{N-1} y_{[k:N]}y_{[k+1:N]} - \sum_{k=1}^{N-1} y_{[k:N]} \sum_{k=1}^{N-1} y_{[k+1:N]}}{\sqrt{(N-1)\sum_{k=1}^{N-1} y_{[k:N]}^2 - \big(\sum_{k=1}^{N-1} y_{[k:N]}\big)^2}\sqrt{(N-1)\sum_{k=1}^{N-1} y_{[k+1:N]}^2 - \big(\sum_{k=1}^{N-1} y_{[k+1:N]}\big)^2}} \quad (7)$$

order, it still ensures not only a small distance between each pair of connected data points in the space of **x** (which is defined as $\lambda_{n_{(k)}n_{(k+1)}} = \|\mathbf{x}_{(k:N)} - \mathbf{x}_{(k+1:N)}\|$), but also a small difference in y ($\Delta y_{[k:N]}$).

Lemma 3. *Let $\{(\boldsymbol{x}_{(t)}, y_{(t)})|1{\le}t{\le}N\}$ be observed from a noise free continuous relationship $y = f(\boldsymbol{x})$ where each $x_i \in \boldsymbol{x}$ is uniformly distributed on $[a,b]$. Let $\lambda_{n^*_{(k)}n^*_{(k+1)}}$ be obtained from the optimum reordering permutation that is defined as*

$$\{n^*_{(k)}\} = \underset{n_{(1)},\cdots,n_{(N)}}{\arg\min} \Big(\sum_{k=1}^{N} \lambda_{n_{(k)}n_{(k+1)}} + \lambda_{n_{(1)}n_{(N)}}\Big) \quad (10)$$

Then, it holds that

$$\lim_{N\to\infty} \Delta y_{[k:N]} = 0,\ \forall 1{\le}t{\le}N \quad (11)$$

Theorem 3. *Let $\{(\boldsymbol{x}_{(t)}, y_{(t)})\}$ be a data sample that is observed from the relationship as defined in (1), and each $x_i \in \boldsymbol{x}$ is uniformly distributed on $[a,b]$. Suppose $nCor(\boldsymbol{x},y)$ is calculated based on the optimum $\{n^*_{(k)}\}$. Then,*

$$\begin{aligned}\lim_{N\to\infty} nCor(\boldsymbol{x},y) &= \mathrm{Var}(f(\boldsymbol{x}))/\mathrm{Var}(y)\\ &= \sum_{x_i\in\boldsymbol{x}} \frac{\mathrm{Var}(g_i(x_i))}{\mathrm{Var}(y)} + \sum_{|\boldsymbol{x}_i\subseteq\boldsymbol{x}|\ge 2} \frac{\mathrm{Var}(h_i(\boldsymbol{x}_i))}{\mathrm{Var}(y)}\end{aligned} \quad (12)$$

In this study, we adopt the nearest neighbor (NN) algorithm (Algorithm 1) to solve TSP (Gutin and Punnen 2007), which is simple and can quickly yield a short route of sufficient quality to satisfy the needs of association detection. (i) The time complexity of NN algorithm is $O(N^2)$ which means that the computational time of $nCor(\mathbf{x},y)$ mainly depends on N but not on M. (ii) $nCor$ is able to cope with high dimensional **x** without a large expense of computational cost. With increasing M, however, the power of $nCor$ on approximating R^2 decreases little by little, since with fixed N the data points become more sparse in a higher dimensional space. (iii) $nCor$ is robust to non-optimal reordering, and it is not sensitive on any particular order of the data points. Even though a TSP route is comparatively bad, the majority of the connected data points are still close to each other in **x** space, that is enough for computing $nCor$.

Three *nCor* based association measures

To further characterize the inter and intra structure of the detected associations, three *nCor* based statistics are also proposed in this study.

Algorithm 1 NN algorithm based data reordering.

Input: Euclidean distance matrix of sample data $\{\mathbf{x}_{(t)}\}$, denoted by $[\lambda_{pq}]^{N\times N}$ where $\lambda_{pq} = \|\mathbf{x}_{(p)} - \mathbf{x}_{(q)}\|$;
Output: concomitants $\{y_{[k:N]}|1{\le}k{\le}N\}$;
Start on data point $t{\leftarrow}1$ as the current data point, set $n_{(1)}{\leftarrow}1$ and $y_{[1:N]}{\leftarrow}y_{(t)}$;
for $k{\leftarrow}1$ to $N-1$ **do**
  Find out the shortest distance connecting the current data point t and an unvisited data point $i \notin \{n_{(1)},\cdots,n_{(k)}\}$ that $i^*{\leftarrow}\arg\min_i \lambda_{it}$;
  Move the current data point to $t{\leftarrow}i^*$, set $n_{(k+1)}{\leftarrow}i^*$ and $y_{[k+1:N]}{\leftarrow}y_{(i^*)}$;
end for

Definition 2. *The coefficient of interaction (COI)*

$$\begin{aligned}COI(\boldsymbol{x},y) =& nCor(\boldsymbol{x},y) - \sum_{x_i\in\boldsymbol{x}} \max\big(0, nCor(x_i,y)\big)\\ &- \sum_{|\boldsymbol{x}_i\subset\boldsymbol{x}|\ge 2} \max\big(0, COI(\boldsymbol{x}_i,y)\big)\end{aligned} \quad (13)$$

where $2{\le}|\boldsymbol{x}_i|{<}|\boldsymbol{x}|$.

$COI(\mathbf{x},y)$ ($COI(x_1\cdots x_M,y)$) is a measure of the strength of the interaction effect exactly in terms of **x**, and by Theorem 3 it holds that $\lim_{N\to\infty} COI(\mathbf{x},y) = \mathrm{Var}(h(\mathbf{x}))/\mathrm{Var}(y)$.

Remark 1. *COI and* nCor *can be used together to distinguish interaction from main effect. (i) If $nCor(x_i,y)$ is significant, then a main effect exists between x_i and y. The stronger the main effect is the larger the $nCor$ value will be. (ii) If $nCor(\boldsymbol{x}_i,y)$ is significant and $COI(\boldsymbol{x},y){>}0$, then an interaction may exists. The stronger the interaction effect is the larger the COI value will be.*

Definition 3. *The coefficient of nonlinearity (CON)*

$$CON(x,y) = nCor(x,y) - Cor^2(x,y) \quad (14)$$

where $Cor(x,y)$ denotes the Pearson correlation coefficient.

$CON(x,y)$ is a measure of the nonlinearity of main effect, which indicates the strength of the nonlinear part of a bivariate association. CON is defined similar as the measure of nonlinearity $MIC - \rho^2$ in (Reshef et al. 2011).

Remark 2. *CON and Cor can be used together to distinguish linear and nonlinear associations. Consider a significant $nCor(x,y)$. (i) $CON(x,y){\le}0$ indicates that only a linear association exists. (ii) $CON(x,y){>}0$ indicates that the association is nonlinear. The stronger the nonlinear effect is the larger the CON value will be. (iii) An insignificant*

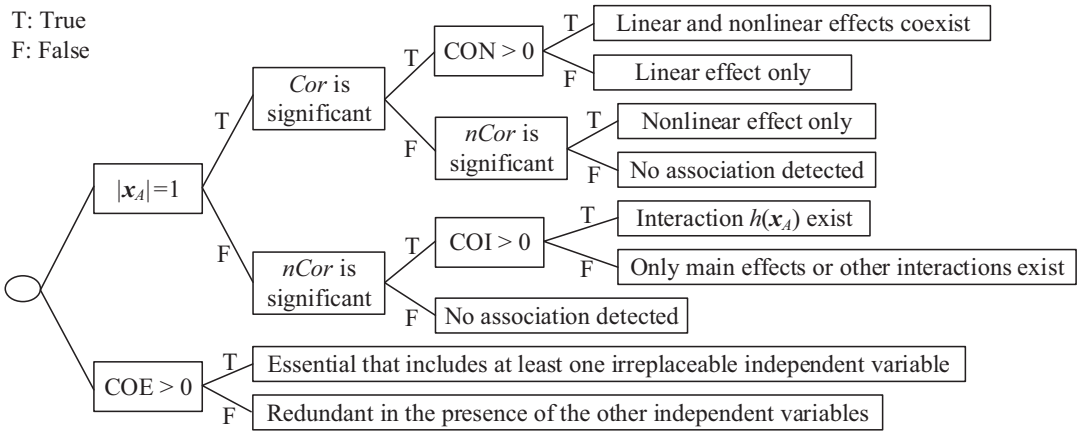


Figure 3: Diagnose and characterize various associations by using $nCor$ and the three $nCor$ based statistics

$Cor(x, y)$ *indicates that there only exists a nonlinear effect such that a linear model will completely fail to capture the underlying relationship.*

Definition 4. *The coefficient of essentialness (COE)*

$$COE(\boldsymbol{x}_s, y) = nCor(\boldsymbol{x}, y) - \max\left(0, nCor(\boldsymbol{x}\backslash\boldsymbol{x}_s, y)\right) \tag{15}$$

where $\boldsymbol{x}_s \subset \boldsymbol{x}$*, and* $\boldsymbol{x}\backslash\boldsymbol{x}_s = \{x_i | \forall x_i \in \boldsymbol{x}, x_i \notin \boldsymbol{x}_s\}$.

$COE(\mathbf{x}_s, y)$ ($COE(x_{s_1} \cdots x_{s_m}, y)$) implies whether or not $\mathbf{x}_s$ is redundant in the presence of $\mathbf{x}\backslash\mathbf{x}_s$. Its role is similar as that of the partial correlation coefficient in the linear case and conditional MI (CMI) which is the MI of two variables conditioned to a third one (Fleuret 2004; Sato et al. 2006; Runge 2018).

Remark 3. *COE can be used to detect if a subset of independent variables is essential in analyzing the dependent variable. Consider a significant* $nCor(\boldsymbol{x}, y)$ *and a subset* $\boldsymbol{x}_s$*. (i)* $COE(\boldsymbol{x}_s, y) > 0$ *indicates that* $\boldsymbol{x}_s$ *is essential in that it contains at least one irreplaceable independent variable that must be involved in model construction. (ii) Generally, the more essential* $\boldsymbol{x}_s$ *of same size is, the larger value the COE test will yield, and thus needs to be given a higher priority in analyzing* y*.*

Summarily, Fig. 3 depicts the decision tree that represents how to diagnose and characterize various associations arising from a subset of independent variables $1 \leq |\mathbf{x}_s \subseteq \mathbf{x}| \leq M$ by the use of Cor, $nCor$ and the three $nCor$ based statistics.

Empirical studies

In this section, a set of simulation examples and a real-world data set are employed to illustrate the effectiveness of the new method. Supplementary material gives the detailed experimental settings, as well as more empirical demonstrations of $nCor$ on detecting associations and approximating R^2 under different data distributions, sample sizes, non-optimum data reordering, and independent variable numbers.

Simulation experiments and comparative analysis

Six simulated examples were performed here for comparison purposes. Table 1 presents the underlying associations that cover a wide range of functional forms including parabolic, exponential, periodic, cross term, mixture function, and even classification problem (step function).

Table 1: The six simulated examples (|**g**| and |**h**| respectively denote the numbers of main effects $g(\cdot)$ and interactions $h(\cdot)$ occuring in each $f(\cdot)$).

Underlying functions	|**g**|	|**h**|
$y_1 = \begin{cases} sin(10x_2), & x_1 \geq 0 \\ sin(10x_2 + 2), & x_1 < 0 \end{cases}$	1	1
$y_2 = x_1x_2 - 0.7x_2x_3 + 3x_1x_2x_3$	0	3
$y_3 = x_1x_2^2 - 3x_1 + cos(20x_3) + e$	2	1
$\begin{cases} y_4 = cos(20x_1x_2)x_2 + 0.5\exp(x_2) + e \\ x_3 = 0.667x_2 + 0.333u \end{cases}$	1	1
(x_1, x_2, y_5): two-spirals problem	0	1
(x_1, x_2, y_6): noisy two-spirals problem	0	1

In examples 1 to 4, we considered three random independent variables with uniform distribution and amplitude range from -1 to 1. In examples 3 and 4, a normally distributed random noise with zero mean and variance of 0.25 was applied to increase the difficulties of association detection. All the data sequences for the first four examples were generated with length of 1000. In example 4, collinearity occurred between x_2 and x_3, and u was set to be a random variable having an identical distribution of x_i. Examples 5, called two-spirals problem, is a benchmark task for nonlinear classification, which consists of two spirals each with 200 samples in a 2-D space. In example 6, each independent variable of the two-spirals problem was corrupted by a normally distributed additive noise with zero mean and variance of 1×10^{-4}. In addition, 40 randomly selected samples (10%) were wrongly categorized that led to a noisy dependent variable.

Here, we compared *nCor* with MIC, *dCor*, kNN based MI, CODCF, and the RDC whose outstanding performances have been extensively demonstrated (Reshef et al. 2018). To make comparisons easier, the MI values were re-scaled to the range $[0, 1]$ (Gelfand and Yaglom 1957; Lange and

Table 2: Association detection by using previous methods (significant scores are marked with underlines).

	$MIC(x,y)$			$CODCF(x,y)$			$RDC(\mathbf{x},y)$						
	x_1	x_2	x_3	x_1	x_2	x_3	x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3	$x_1x_2x_3$
y_1	0.134	0.462	0.134	-0.032	-0.096	-0.001	0.107	0.244	0.095	0.260	0.149	0.212	0.255
y_2	0.180	0.229	0.231	0.384	0.479	0.353	0.478	0.532	0.503	0.693	0.730	0.714	0.976
y_3	0.412	0.131	0.268	-0.659	-0.360	0.039	0.664	0.390	0.118	0.809	0.673	0.382	0.814
y_4	0.133	0.242	0.226	0.009	0.441	0.392	0.098	0.477	0.414	0.486	0.424	0.483	0.491
y_5	0.222	0.207		0.098	0.006		0.012	0.001		0.020			
y_6	0.196	0.183		0.080	0.009		0.011	0.004		0.023			
	dCor(**x**,*y*)						$r^2_{MI}(\mathbf{x},y) = 1 - exp(-2MI(\mathbf{x},y))$						
	x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3	$x_1x_2x_3$	x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3	$x_1x_2x_3$
y_1	0.054	0.114	0.043	0.106	0.062	0.079	0.089	0	0.992	0	0.862	0.001	0.548	0.522
y_2	0.210	0.213	0.262	0.279	0.267	0.290	0.342	0.296	0.510	0.312	0.791	0.725	0.796	0.978
y_3	0.624	0.106	0.070	0.558	0.525	0.092	0.498	0.487	0.163	0.207	0.643	0.546	0.256	0.572
y_4	0.052	0.405	0.360	0.342	0.263	0.400	0.361	0.089	0.260	0.212	0.461	0.205	0.326	0.383
y_5	0.091	0.034		0.081				0.025	0.090		0.747			
y_6	0.079	0.049		0.079				0.001	0.030		0.463			

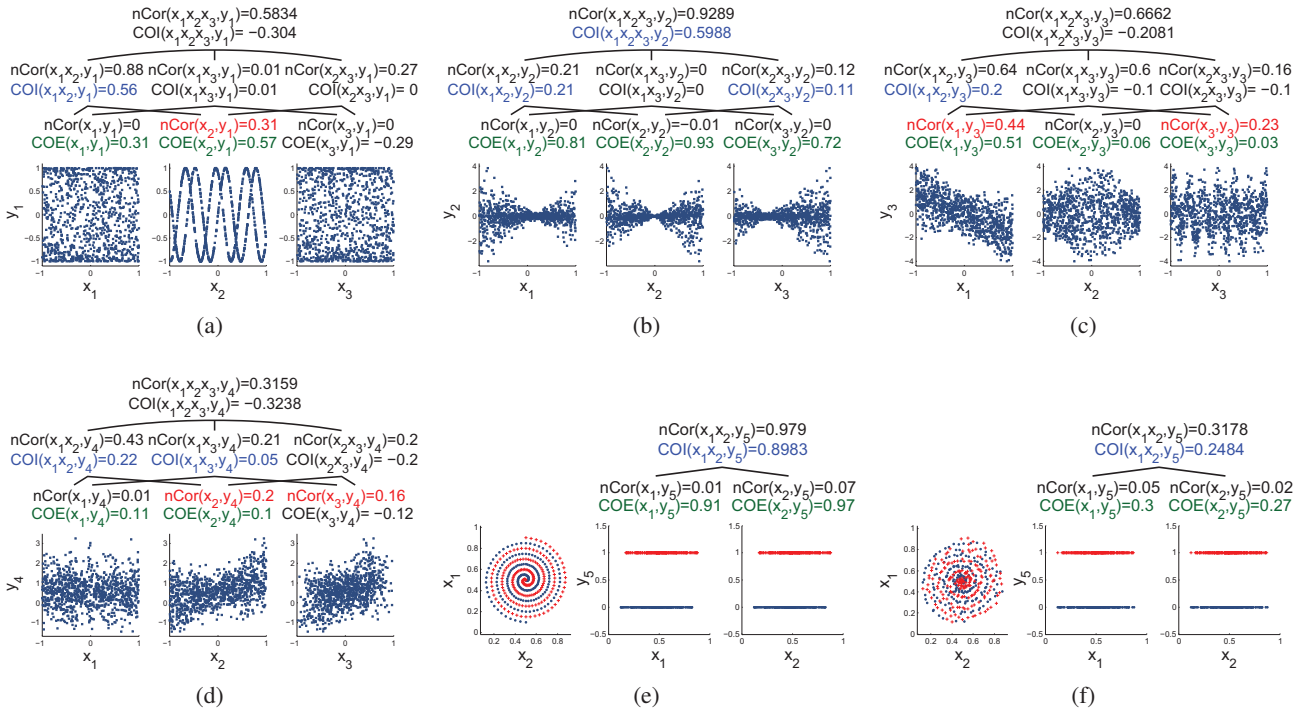


Figure 4: The scatterplots and association detection for the six examples. According to the properties of the new statistics, significant main effects (*nCor*), detected interaction effects (COI) and essential independent variables (COE) are marked as red, blue, and green respectively. *nCor* were tested for significance using Fisher's transformation with the same confidence limits ($\alpha = 5\%$) that are ± 0.062 for examples 1-4 and ± 0.098 for the last two.

Grubmuller 2005). The hypothesis test introduced in (Zhang, Zhu, and Longden 2007; Reshef et al. 2011; Székely, Rizzo, and Bakirov 2007; Lopez-Paz, Hennig, and Scholkopf 2013) was used to detect significant associations at 95% confidence level ($\alpha = 0.05$). For MIC and other MI measures, the empirical confidence limits were obtained through using 1000 surrogate sets of random data (Reshef et al. 2018).

Tables 2 presents the experimental results. (i) None of these measures can capture every underlying relationship with satisfactory results, and particularly the association between x_1 and y_1 which is missed out by all the methods except RDC. Generally, the multivariate measures achieve

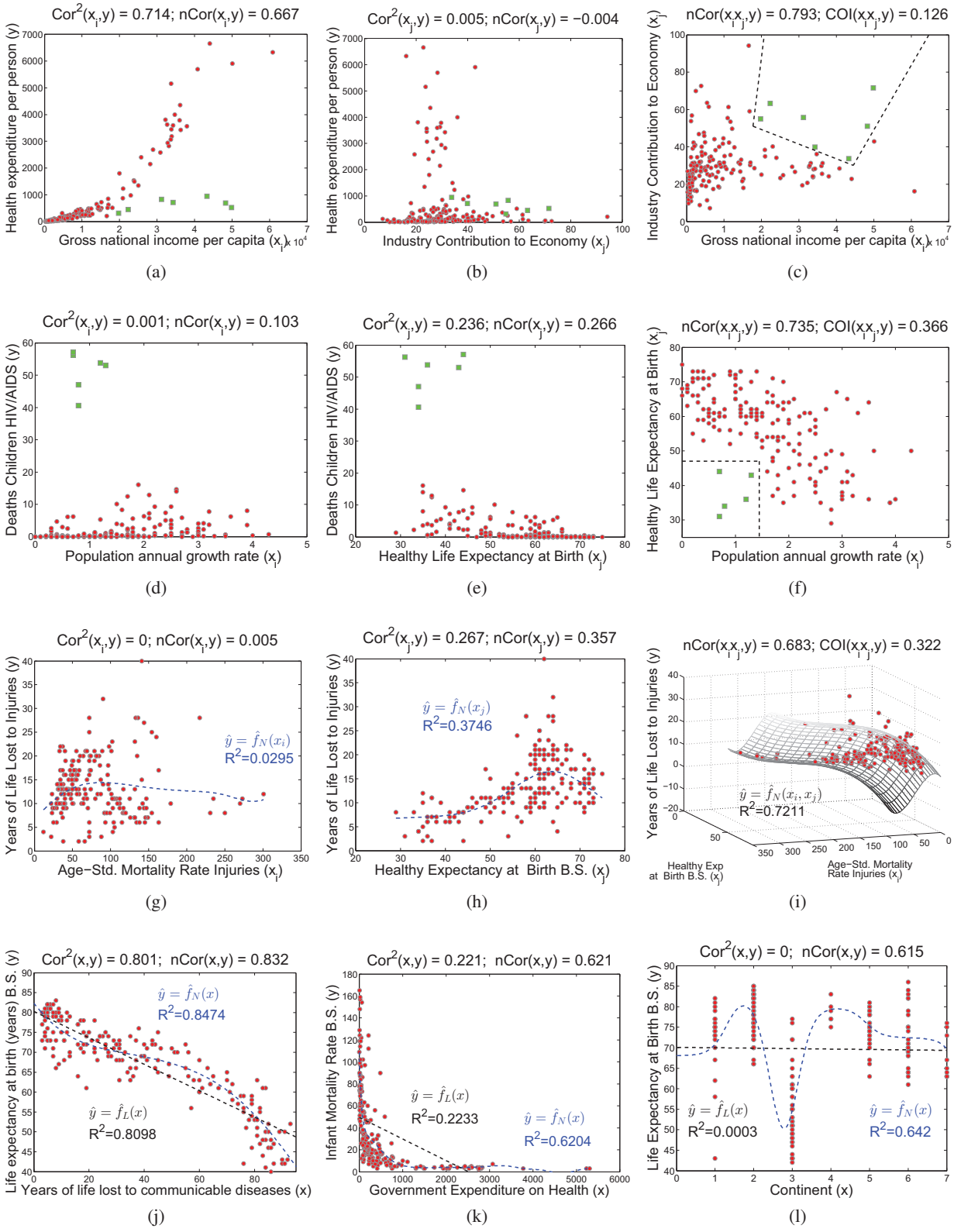


Figure 5: Six typical examples of the associations detected by *nCor* and COI, including three interactions (a-i) and three main effects (j-l). In (g-l), $\hat{f}_L(\cdot)$ and $\hat{f}_N(\cdot)$ respectively indicate that the line or curve (surface), and R^2 are obtained through using linear regression or ANN.

better performance than the bivariate ones. By using these methods, however, there are still some missed detections. (ii) Despite exceeding the confidence interval, these measures cannot assign appropriate scores to correctly state the importance of each independent variable to predicting or classifying the dependent variable. For instance, the values of $CODCF(x_2, y_1)$, $RDC(x_1x_2, y_1)$, $MIC(x_1, y_5)$, and $MIC(x_2, y_5)$ just slightly exceed the confidence limits, whereas the corresponding variables are strongly associated. $dCor(x_2, y_1) = 0.11$ and $r^2_{MI}(x_2, y_1) = 0.99$, however, the real R^2 is in fact about 0.3. (iii) Although RDC, $dCor$ and MI can handle both bivariate and multivariate data, they still cannot be used to distinguish between interactions and main effects. The detection results of the two measures are sometimes ambiguous and confusing so that it is quite difficult to properly discern interaction effects from these values. For example, $r^2_{MI}(x_2, y_1) > r^2_{MI}(x_1x_2, y_1)$, but there is a strong interaction $h(x_1, x_2)$, and y_1 only can be predicted properly by using x_1 and x_2 simultaneously. In example 2, $r^2_{MI}(x_1x_2, y_2)$, $r^2_{MI}(x_1x_3, y_2)$, and $r^2_{MI}(x_2x_3, y_2)$ yield very similar values, and especially $MI(x_1x_3, y_2) > MI(x_1, y_2) + MI(x_3, y_2)$, but actually an exact interaction $h(x_1, x_3)$ does not exist. In addition, $RDC(x_1x_3, y_2) > RDC(x_2x_3, y_2) > RDC(x_1x_2, y_2)$, however, y_2 precisely contains terms x_1x_2 and x_2x_3.

Figure 3 shows the scatterplots of the six examples and the correlation detection results. As shown in the figures, the scatterplots display a variety of patterns. Especially between y_1, y_4 and x_1, there is even no any pattern that can be discovered by visual inspection of the scatterplots, since the data points just look like completely random. In contrast to the existing methods, *nCor* base statistics successfully detect the underlying relationships without any missed or false judgement. By COI test, all the interaction effects are precisely discerned from the associations along with a rough assessment of the effect strength. In example 2, $COI(x_1x_2x_3, y_2)/COI(x_1x_2, y_2) = 2.9$ and $COI(x_2x_3, y_2)/COI(x_1x_2, y_2) = 0.52$ are approximately equal to the theoretical values of the corresponding variance ratios which can be derived as $\mathrm{Var}(3x_1x_2x_3)/\mathrm{Var}(x_1x_2) = 3$ and $\mathrm{Var}(0.7x_2x_3)/\mathrm{Var}(x_1x_2) = 0.49$. Moreover, a negative $COE(x_3, y_4)$ indicates that x_3 is redundant. That is to say, although $nCor(x_3, y_4)$ yields a considerable value, x_3 is not essential to analyzing y_4 since the predictive information carried by x_3 is fully overlapping with x_2.

Association detection in large data set

nCor based statistics were used to explore a real-world data set that consists of 357 social, economic, health, and political indicators for 202 countries around the world for the time period from 1960 through 2005. It was originally collected from the World Health Organization (WHO) and partner organizations (Rosling 2008; W.H.O. 2009). By the new method, we detected a huge number of interesting associations including both nonlinear main effects and interactions. For more statistical results see supplementary material.

Figure 5 shows six typical associations detected by the new measures. To confirm that *nCor* is an effective estimate of R^2, linear regression and feedforward artificial neural network (ANN) were implemented to identify the detected relationships to obtain the real R^2 of the data. (i) Fig. 5 (a) depicts a superposition of two relationships which has been studied previously (Reshef et al. 2011) that, most data points obey a steeper trend, and the others obey a less steep trend. Obviously, it is impossible to separate the two trends of health expenditure (y) when considering the national income (x_i) alone. By COI test, we found another indicator, called industry contribution to economy (x_j), which does not directly affect but interactively influences y (Figs. 5 (b, c)). When looking at y in the space of x_i and x_j, the less steep minority of points can be precisely separated from the others by three lines. (ii) Similarly, Figs. 5 (d-f) show another example which is even more persuasive. The COI test detect a strong interaction effect implying the fact that neither a low population growth rate (x_i) nor a short healthy life expectancy (x_j) is unique to the counties with extremely high deaths among children due to HIV/AIDS (the outliers of y), but a combination of the two is. (iii) The third example is an association consisting of both main and interaction effects. Fig. 5 (g-i) show the relationships among the three variables, as well as the curves, surface, and the R^2 obtained from the best fitted ANNs (10 ANNs was trained for each case). By means of *nCor*, we can not only accurately reveal the composition of the association, but also properly foretell the R^2 of the data. (iv) Figs. 5 (j-l) show three pairwise associations which are diagnosed respectively as weak, strong, and only nonlinear effects, and then confirmed by linear regression and ANN. Fig. 5 (l) suggests that even with a qualitative independent variable, *nCor* still exhibits an excellent detection power.

Conclusion

Data-driven research is becoming increasingly popular in fields as varied as biology, physics, political science, and economics. In such kind of studies, association detection is one of the most critical issues, and may provide a lot of valuable insight into large and complex data sets that is otherwise difficult to obtain. *nCor* inherits the merits of the Person correlation coefficient in the linear case, but is generally applicable to measuring all types of functional relationships. The three *nCor* based statistics can be used to distinguish and characterize the associations from the aspects of nonlinearity, interactivity, and variable redundancy. These measures, as illustrated in the empirical studies, are simple but powerful, and may have a wide range of applications from quick association detection to various data analysis and interpretation.

References

Aguirre, L. A. 1995. A nonlinear correlation function for selecting the delay time in dynamical reconstructions. *Physics Letters A* 203(2-3):88–94.

Altman, N., and Krzywinski, M. 2015. Points of significance: Association, correlation and causation. *Nature Methods* 12(10):899–900.

Billings, S. A., and Zhu, Q. M. 1995. Model validation tests for multivariable nonlinear models including neural networks. *International Journal of Control* 62(4):749–766.

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

Darbellay, G. A., and Vajda, I. 1999. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory* 45(4):1315–1321.

David, H. A., and Nagaraja, H. N. 2003. *Order Statistics, Third Edition*. New Jersey: John Wiley and Sons.

Delicado, P., and Smrekar, M. 2009. Measuring non-linear dependence for two random variables distributed along a curve. *Statistics and Computing* 19(3):255–269.

Delicado, P. 2001. Another look at principal curves and surfaces. *Journal of Multivariate Analysis* 77(1):84–116.

Fleuret, F. 2004. Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research* 5:1531–1555.

Gelfand, I. M., and Yaglom, A. M. 1957. Calculation of the amount of information about a random function contained in another such function. *American Mathematical Society Translations: Series 2* 12(1):199–236.

Gretton, A., and Gyorfi, L. 2010. Consistent nonparametric tests of independence. *Journal of Machine Learning Research* 11(3):1391–1423.

Gretton, A.; Fukumizu, K.; Teo, C.; Song, L.; Scholkopf, B.; and Smola, A. 2008. A kernel statistical test of independence. In *Advances in neural information processing systems (NIPS) 20*, 585–592.

Gutin, G., and Punnen, A. P. 2007. *The traveling salesman problem and its variations*. Boston: Springer.

Hastie, T., and Stuetzle, W. 1989. Principal curves. *Journal of the American Statistical Association* 84:502–516.

Heller, R.; Heller, Y.; Kaufman, S.; Brill, B.; and Gorfine, M. 2016. Consistent distribution-free k-sample and independence tests for univariate random variables. *Journal of Machine Learning Research* 17:1–54.

Heller, R.; Heller, Y.; and Gorfine, M. 2013. A consistent multivariate test of association based on ranks of distances. *Biometrika* 100(2):503–510.

Kendall, M. 1938. A new measure of rank correlation. *Biometrika* 30:81–93.

Kraskov, A.; Stogbauer, H.; and Grassberger, P. 2004. Estimating mutual information. *Physical Review E* 69:066138.

Lange, O. F., and Grubmuller, H. 2005. Generalized correlation for biomolecular dynamics. *Proteins Structure Function and Bioinformatics* 62(4):1053–1061.

Liu, P.; Sohn, H.; and Jeon, I. 2017. Nonlinear spectral correlation for fatigue crack detection under noisy environments. *Journal of Sound and Vibration* 400:305–316.

Lopez-Paz, D.; Hennig, P.; and Scholkopf, B. 2013. The randomized dependence coefficient. In *Advances in neural information processing systems (NIPS) 27*, 1–8.

Mao, K. Z., and Billings, S. A. 2000. Multi-directional model validity tests for non-linear system identification. *International Journal of Control* 73(2):132–143.

Moon, Y.; Rajagopalan, B.; and Lall, U. 1995. Estimation of mutual information using kernel density estimators. *Phys Rev E* 52(3):2318–2321.

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.

Reshef, Y. A.; Reshef, D. N.; Finucane, H. K.; Sabeti, P. C.; and Mitzenmacher, M. M. 2015. Measuring dependence powerfully and equitably. *Journal of Machine Learning Research* 17(1):7406–7468.

Reshef, D.; Reshef, Y.; Sabeti, P.; and Mitzenmacher, M. 2018. An empirical study of the maximal and total information coefficients and leading measures of dependence. *The Annals of Applied Statistics* 12:123–155.

Rényi, A. 1959. On measures of dependence. *Acta Mathematica Hungarica* 10:441–451.

Rosling, H. 2008. Indicators in gapminder world. http://www.gapminder.org/gapminder-world/indicators-in-gapminder-world/.

Runge, J. 2018. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 84, 938–947.

Sato, T.; Yamanishi, Y.; Horimoto, K.; Kanehisa, M.; and Toh, H. 2006. Partial correlation coefficient between distance matrices as a new indicator of protein-cprotein interactions. *Bioinformatics* 22(20):2488–2492.

Sejdinovic, D.; Sriperumbudur, B.; Gretton, A.; and Fukumizu, K. 2013. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics* 41:2263–2291.

Székely, G., and Rizzo, M. 2009. Brownian distance covariance. *The Annals of Applied Statistics* 3:1236–1265.

Székely, G.; Rizzo, M.; and Bakirov, N. 2007. Measuring and testing dependence by correlation of distances. *The Annals of Statistics* 35:2769–2794.

Wang, Q.; Shen, Y.; and Zhang, J. Q. 2005. A nonlinear correlation measure for multivariable data set. *Physica D* 200:287–295.

W.H.O. 2009. WHO statistical information system (WHOSIS). http://www.who.int/whosis/en/.

Zhang, L. F.; Zhu, Q. M.; and Longden, A. 2007. A set of novel correlation tests for nonlinear system variables. *International Journal of Systems Science* 38(1):47–60.

Zhu, Q. M.; Zhang, L. F.; and Longden, A. 2007. Development of omni-directional correlation functions for nonlinear model validation. *Automatica* 43(9):1519–1531.