Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

An intuitive general rank-based correlation coefficient

Divya PANDOVE[‡1], Shivani GOEL[2], Rinkle RANI[3]

[1]*Research Lab, Computer Science and Engineering Department, Thapar University, Patiala 147004, India*

[2]*Department of Computer Science Engineering, School of Engineering and Applied Sciences, Bennett University, Greater Noida 201310, India*

[3]*Computer Science and Engineering Department, Thapar University, Patiala 147004, India*

E-mail: dpandove@gmail.com; shigo108@yahoo.co.in; raggarwal@thapar.edu

Received Sept. 21, 2016; Revision accepted Jan. 14, 2017; Crosschecked May 15, 2018

Abstract: Correlation analysis is an effective mechanism for studying patterns in data and making predictions. Many interesting discoveries have been made by formulating correlations in seemingly unrelated data. We propose an algorithm to quantify the theory of correlations and to give an intuitive, more accurate correlation coefficient. We propose a predictive metric to calculate correlations between paired values, known as the general rank-based correlation coefficient. It fulfills the five basic criteria of a predictive metric: independence from sample size, value between −1 and 1, measuring the degree of monotonicity, insensitivity to outliers, and intuitive demonstration. Furthermore, the metric has been validated by performing experiments using a real-time dataset and random number simulations. Mathematical derivations of the proposed equations have also been provided. We have compared it to Spearman's rank correlation coefficient. The comparison results show that the proposed metric fares better than the existing metric on all the predictive metric criteria.

Key words: General rank-based correlation coefficient; Multivariate analysis; Predictive metric; Spearman's rank correlation coefficient

https://doi.org/10.1631/FITEE.1601549 **CLC number:** TP301

1 Introduction

The practice of finding correlations between data points has been prevalent for a very long time. This concept was introduced by Sir Francis Galton, who noticed a relationship between the height of a man and the length of his forearm (Hauke and Kossowski, 2011). The only thing that has changed is the volume of data and the availability of better tools for analysis and data storage. As a result, correlations in data emerge more rapidly, without incurring much cost. Today, the focus of methods for finding correlations in data is centered on the 'what' part of the data but not the 'why' part. This means that to successfully find correlations in a given dataset, it is not important to understand all the underlying variables of the system and their relationships (the 'why' part of data). We just need to find the variables that help in making predictions (the 'what' part of data). The correlations in these variables help us find patterns in data. The 'why' aspect of data might be interesting and appealing to the human mind, but it does not generate valuable insights into relationships between data points. Instead of focusing on finding the cause-and-effect relationships in a particular data collection, we try to find patterns and correlations. This helps us visualize links in data that have not been seen before. The premise of this approach is that causality can rarely be proven (Didelez and Pigeot, 2001). Though correlations play an important part in analyzing small datasets, they truly shine when dealing with large data. Nowadays, experts are

‡ Corresponding author
ORCID: Divya PANDOVE, http://orcid.org/0000-0001-8694-1538

developing necessary tools to identify and compare nonlinear correlations. The techniques of analysis are being aided and enhanced by fast-growing novel approaches and software that can extract non-causal relationships in data (Reshef et al., 2011).

The value of a correlation coefficient implies a dependency relationship between two data values. If the value of correlation is high, it means that when one datum value changes, the likelihood of the other one changing is also very high. A weak correlation implies that there might be little or no effect of one variable on the other; i.e., there is little or no dependency between the two variables. A correlation coefficient is used to express the degree of dependence between the two variables. Fig. 1 shows the interpretation of various values of a correlation coefficient. These values lie between -1 and 1.

With poor correlations, there is low probability of similarity between objects. However, if the correlation is strong, then the probability of similarity is quite high. To fully use the power of correlations, there is a need to understand the relationship between the two variables under consideration. This can be explained with the help of an example. Say, in a dataset there are two variables: x and y. Here, x represents students' marks and y gives the number of hours studied by them. To understand the relationship between x and y, we need to find a correlation between them. The degree of dependency between the two variables will help us make various predictions. Even if we are not able to predict the course of one variable, studying the other can help us predict the future (Liao et al., 2015b).

In this study, we propose a predictive metric, describing a general rank-based correlation coefficient (GRCC) that is based on calculating rank distances between all the observations. It quantifies the concept of correlations as discussed above, and we provide a mathematical proof for the proposed scheme, along with simulations to verify the proposed metric.

The focus of the paper is on introducing and validating a novel way of looking at correlations, and showing improvement in an existing metric of correlation calculation. The increasing role of correlation analysis in data science is the basic motivation of this study. We have studied in detail various case studies and research papers that have successfully used correlation analysis to discover hidden patterns in data. In all of the literature, we could not find any general correlation metric that is intuitive in nature. Most of the work is done in a particular domain, dealing with a particular type of data. Our work deals with this research gap. We propose a rank-based metric, which is more intuitive and compatible with the way the human brain perceives correlations. GRCC is a novel metric that fulfills the five basic criteria of a predictive metric: independence from sample size, value between -1 and 1, measuring the degree of monotonicity, insensitivity to outliers, and intuitive demonstration. This metric is an algorithm instead of a stand-alone formula, making it one of the few metrics defined by an algorithm. The hierarchy of need and motivation for GRCC is given in Fig. 2. The diagram shows that the need for GRCC comes from the shortcomings of traditional metrics like Spearman's rank correlation coefficient. These metrics, combined with the domains and datasets in which work has been done, bring to light the research gap in this field. The motivation for GRCC comes from a need to develop a general metric for correlation analysis that meets all the criteria for a predictive metric and can be used in varied domains.

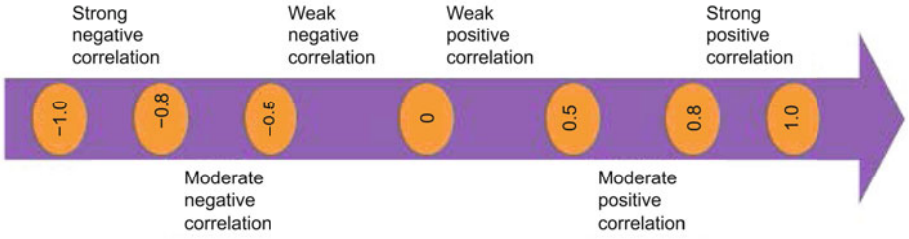

Fig. 1 Interpretation of a value of a correlation coefficient

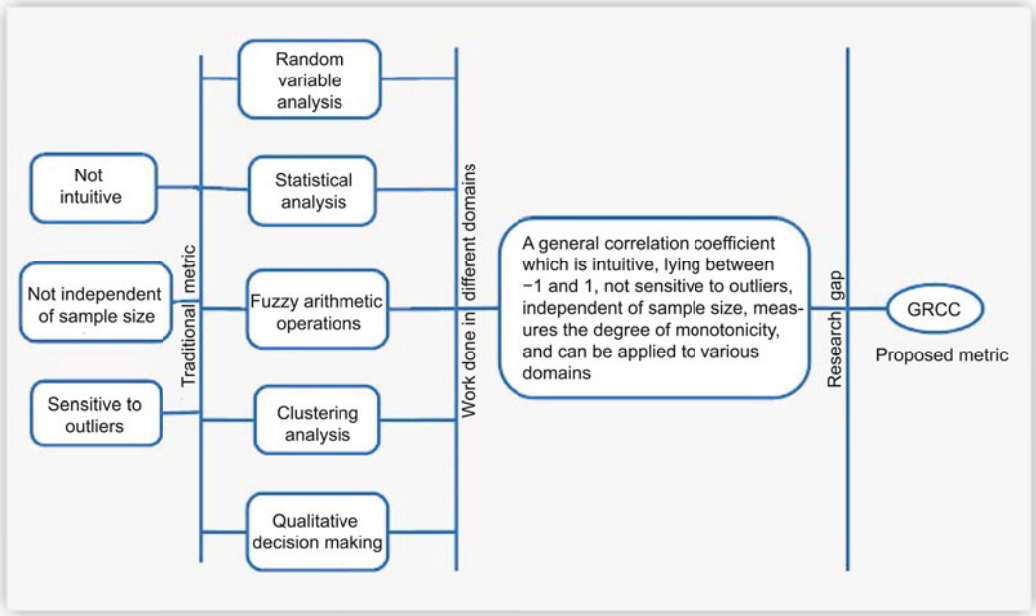

Fig. 2 Hierarchy of need and motivation for GRCC

2 Related work

When working with huge amounts of data, establishing correlations among data points brings out a clearer picture of the nature of data at hand. The weak and strong correlations in data values help us capture the present and the future. There have been many successful instances where correlations in data points have been used for understanding available but seemingly irrelevant data, and carving out results that have been amazingly accurate.

Amazon has been a pioneer in the field of big data analysis. Initially, it collected information regarding the books sold on its website and used it to make recommendations to specific customers. This was done by finding similarities in the customer database, known as user-to-user collaborative filtering (Ritala et al., 2014). Later, it started finding correlations among products themselves. This association technique came to be known as item-to-item collaborative filtering (Linden et al., 2003). The mathematical determination of the data (finding correlations) in advance increased the efficiency of the system many-fold and cut across product categories (Chen et al., 2012). Walmart, one of the largest retailers in the world, had analyzed the data that it had collected over the years, and discovered interesting correlations that helped them develop important business strategies (Sen et al., 2006). In 2011, Fair Isaac Corporation (FICO) developed the medical adherence score to determine how often people consume their prescribed medication. Correlations were found in variables that may sound irrelevant, such as marital status, how long they have stayed at an address, and if they own a car. These relationships helped them predict which patients were most likely to forget to take their medication, and can send them reminders (Volpone et al., 2015). If properly analyzed, data can reveal extraordinary insights into the medical condition of a patient. The predictive analytics algorithms when applied on the data on premature babies in intensive care units have been able to predict onset of infections. The data show correlation instead of causality for detection and prevention of many deadly diseases (McGregor, 2013). Google used its enormous amount of data to predict the spread of H1N1 virus and the winter flu, by looking for a correlation between the frequency of a certain query and the spread of flu over a given time and space. It used frequently searched terms as a proxy for flu and ran 450 million varied models, and found that 45 search queries had strong correlation between their predictions and official figures when used in a mathematical model (Ginsberg et al., 2009). In the year 2000, the Sloan Digital Sky Survey began collecting astronomical data. A telescope

was located in New Mexico. In the initial weeks, it gathered more astronomical data than ever before. By 2010, the data in the archives of the survey were huge, reaching almost 140 terabytes. However, its successor, installed in 2016, acquires this quantity of data every five days. Such huge amounts of data can only make sense by using correlational non-causal analysis (Davenport et al., 2013). In the big-data age, where the datasets are too big and the area under consideration is too complex, the hypotheses are no longer driven by trial and error. Earlier they were based on developing abstract ideas concerning a phenomenon and collecting data to verify those theories. This was usually done by selecting a proxy to study the relationship between two variables and verifying how good the proxy was by using correlation analysis. Now, sophisticated computational analysis helps identify the most optimum proxy for a given problem (Kitano, 2002). This means that instead of a hypothesis-driven approach, we use a data driven approach.

Correlation measures have been used by many researchers to analyze data, taken from different kinds of datasets. A lot of work in this area has been done on fuzzy datasets and their variants. Murthy et al. (1985) expressed a need for a measure of correlation between two fuzzy membership functions. They focused on studying properties possessed by a correlation measure. Then they defined the measure of correlation based on these properties. Chaudhuri and Bhattacharya (2001) used Spearman's rank correlation coefficient to find correlation between two fuzzy sets. The members of the sets have to be ranked according to fuzzy membership values. They proposed a membership value-based fuzzy correlation measure. Liu and Kao (2002) worked on a correlation coefficient of random numbers. They developed a methodology for calculating a correlation coefficient whose value is a fuzzy number. Hong (2006) strictly focused on giving an exact solution to the fuzzy correlation coefficient, without the aid of programming. Hung (2001) focused on developing a method for calculating a correlation coefficient of intuitionistic fuzzy sets by deploying mathematical statistics. He proposed a formula that gave not only strength but also polarity of the relationship between the datasets. He also extended the research to interval-valued intuitionistic fuzzy sets, i.e., the values of these sets lying in the interval $[-1, 1]$. Mitchell (2004) defined a correlation coefficient between two intuitionistic fuzzy sets, and assumed that an ensemble of ordinary fuzzy sets could be interpreted as an intuitionistic fuzzy set. Chen et al. (2013) have worked on correlation coefficients of hesitant fuzzy sets and their applications to clustering analysis. The hesitant fuzzy sets allocate degree of membership to an element in a set. This means that an element can be represented as a member of a set with several possible values. An element can be a member of more than one set. More recently, Liao et al. (2015a) worked on hesitant fuzzy linguistic term sets (HFLTSs). These sets represent hesitant qualitative information in decision making. The authors proposed HTFLSs based on traditional correlation coefficients of fuzzy sets, intuitionistic fuzzy sets, and hesitant fuzzy sets. Given that HFLTSs have different weights, weighted correlation coefficients and ordered weighted correlation coefficients have also been investigated.

Distance correlation has also been applied to random variables by Huo and Székely (2016). They proved that distance correlation can be implemented by an $O(n \log n)$ algorithm that is comparable to other computationally efficient algorithms. Kong et al. (2012) examined mortality and lifestyle factors that run in families. They used distance correlation to measure pairwise differences between related and random people. Li et al. (2012) have also worked on life sciences using distance correlation. They dealt with very high dimensional data, and developed an independent screening procedure based on distance correlation. Székely and Rizzo (2012) emphasized the uniqueness of distance correlation. Their work was extended by Lyons (2013) to a general metric space. A comparative analysis of the correlation measures used by researchers on different datasets is given in Table 1.

This section briefly discusses the diverse data domains that employ correlation analysis to make sense of the existing data. It should be highlighted that data points are to be understood in terms of correlations among them rather than causal relationships. The approach followed is either the detection of a proxy to perform correlation analysis, or the use of correlations, sensor data, and previous results to perform predictive analysis. All of these case studies and research findings have laid the foundation for an approach that can quantify the process of correlation

Table 1 Comparative analysis of correlation measures used on different datasets

Reference	Type of dataset	Correlation measure	Application domain	Application examples
Chaudhuri and Bhattacharya (2001)	Fuzzy	Membership value-based fuzzy correlation measure	Random variables	Image processing
Hung (2001)	Intuitionistic and interval-valued intuitionistic fuzzy	Correlation coefficient of intuitionistic fuzzy sets	Mathematical statistics	–
Liu and Kao (2002)	Fuzzy	Fuzzy correlation coefficient	Statistical analysis, fuzzy environment	Correlation between technology and management in 15 machinery firms in Taiwan, China
Mitchell (2004)	Intuitionistic fuzzy and normal fuzzy	Intuitively satisfying correlation coefficient	Fuzzy environment	–
Hong (2006)	L-R fuzzy numbers	Fuzzy correlation coefficient	T_{W}-based fuzzy arithmetic operations	–
Kong et al. (2012)	Real data from subpopulation of the Beaver Dam Eye Study	Distance correlation	Mortality and lifestyle factors in families and random people	Relationships between multiple clusters of variables with real-valued attributes
Li et al. (2012)	Very high dimensional data	Distance correlation	Diverse scientific fields	Screening features in very high dimensional data
Chen et al. (2013)	Hesitant fuzzy sets, interval-valued hesitant fuzzy sets	Correlation coefficient of hesitant fuzzy sets	Clustering analysis of correlation coefficients of hesitant fuzzy sets	Software evaluation, classification, and assessment of business risk failure
Liao et al. (2015a)	Hesitant fuzzy linguistic term	Correlation coefficient, weighted correlation coefficients, ordered weighted correlation coefficients	Qualitative decision making	Traditional Chinese medical diagnosis
Huo and Székely (2016)	Synthetic datasets, random variables	Distance correlation	Applications where statistical dependence needs to be calculated	Feature screening in ultra high dimensional data analysis

determination in a given dataset. This approach is proposed and verified in the following sections.

3 Problem formulation

In today's information-centric world, the focus is on analyzing huge amounts of data and bringing out meaningful patterns in them. This involves computation of correlations among thousands of data points. These computations are performed on paired observations that could be in the form of data buckets or time series. Such an analysis follows a non-causal approach. Causality can be defined as one factor contributing to the development of a variable, and its removal affects the frequency of that variable. For example, cigarette smoke has been known to be a contaminated substance that leads to increased rates of different types of cancers and heart and respiratory diseases. It is not necessary to precisely identify which component in the smoke is the primary culprit before introducing preventive measures (Deufemia et al., 2014). There is a need to refine the existing correlation determining factors and to come up with an approach that would factor in the various problems encountered in large datasets. These are the problems of dimensionality, outliers, coefficients showing fake correlations that do not exist, etc. In this study, the focus is on subsets of observations that have the same multivariate features. There is a need to perform multivariate feature selection and identification of predictive set of metrics that best suits our purpose. The predictive metric should fulfill five main criteria (Granville, 2014):

1. It should be independent of the sample size to enable comparison across datasets of various sizes.

2. It must lie between -1 and 1, with 0

meaning no correlation. This is similar to existing correlation measures to obtain back compatibility.

3. It should be general such that it can measure the degree of monotonicity (X grows with Y), rather than linearity ($X = aY$). This means that it should be more general than the traditional correlation measures, but must not be as general as the distance correlation which is equal to 0 if and only if X and Y are independent.

4. It should have no sensitivity to outliers so that it is robust.

5. It should be intuitive so that it translates how the human brain perceives correlations.

The criteria for a predictive metric are represented in Fig. 3.

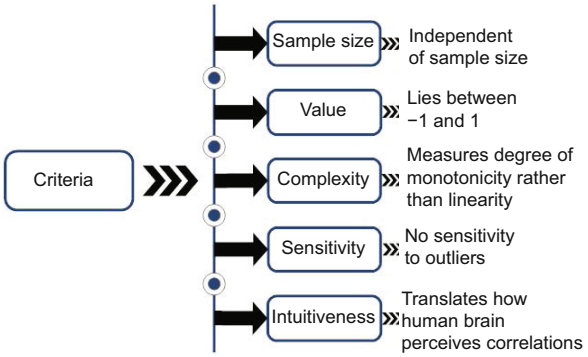

Fig. 3 Criteria for a predictive metric

4 Proposed approach: GRCC

We propose a more general measure. We start by considering Spearman's rank correlation coefficient, which is given by Xiao et al. (2015):

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}, \tag{1}$$

where $d_i = x_i - y_i$ is the difference between ranks of observations. This coefficient lies between -1 and 1, which satisfies the second criterion of the five main criteria mentioned previously.

Now, we develop a new general rank-based correlation coefficient that is defined by an algorithm. Our aim is to develop a correlation coefficient that satisfies all the conditions of a predictive metric. Like all the existing metrics, the proposed approach will also work on finding correlations between a pair of observations. We assume that out of the two observations $\boldsymbol{x}$ and $\boldsymbol{y}$, $\boldsymbol{x}$ is ordered. The metric is then defined by a series of steps. The algorithm for GRCC is given in Algorithm 1. The metric combines basic principles of correlation analysis and algorithm design to address the research gap in the existing literature. Most of the fundamentals in this metric are the same as Spearman's rank correlation coefficient, but we have introduced elements to make GRCC more general and intuitive. GRCC is denoted by g and is governed by a parameter c, which is the prior distribution. We consider this algorithm only at $c = 1$ and $c = 2$. This metric is symmetric, meaning that it will not change even if $\boldsymbol{x}$ and $\boldsymbol{y}$ are swapped. Like Spearman's rank correlation, if the order of $\boldsymbol{y}$ is reversed, then only the sign of correlation will change. Also, for every value of $c > 0$, the value of g will always lie between -1 and 1. When the value of c is equal to 2, it becomes very sensitive to outliers. The value of g at $c = 1$ is the most well-rounded solution. In the algorithm, the rank distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ is calculated in a, and between $\boldsymbol{x}$ and reverse order of $\boldsymbol{y}$ in b. The smallest value between a and b helps determine the sign of the correlation. The value of the denominator d is most crucial to determine the exact value of g. Three ways are mentioned in the algorithm, and any one of them can be used to determine the value of d, depending on factors such as the number of data points and type of data. Diagrammatically, the ways to select the denominator d have been described in Fig. 4.

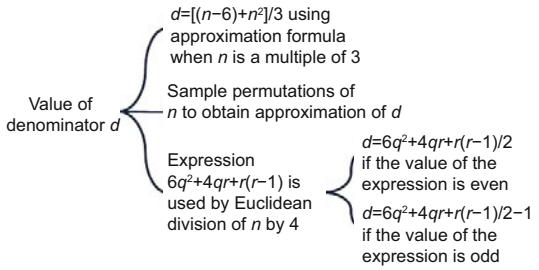

Fig. 4 Ways to select the value of denominator d

The value of the denominator d can be selected in any of the following ways:

1. An approximation formula that is most accurate when n is a multiple of 3 is given by

$$d = \frac{(n-6)+n^2}{3}.$$

2. Sample permutations of n to obtain an approximation of d.

Algorithm 1 General rank-based correlation coefficient (GRCC)

Input: a parameter c, which is a constant governing the value of the correlation coefficient; number of terms n
Output: The data in one of the two variables of the paired observations are sorted

while $i \leq n$ **do**
  $a = \sum_{i=1}^{n} [x(i) - y(i)]^c$ // Rank distance between all the values of variables $x(i)$ and $y(i)$
end while
while $i \leq n$ **do**
  $b = \sum_{i=1}^{n} [x(i) - (n-1-y(i))]^c$ // Rank distance between all the values of variables $x(i)$ and $n-1-y(i)$
end while
Compute the minimum of a and b, i.e., $\min(a, b)$
Determine sign(s) of the coefficient based on the following conditions:
if $a > b$ **then**
  $s = 1$
else if $a < b$ **then**
  $s = -1$
else if $a = b$ **then**
  $s = 0$
end if
The general rank-based coefficient $g = s[1 - \min(a, b)/d]$

3. A specific formula of d can be used by performing Euclidian division of n by 4, and forming an equation: $n = 4q + r$, $0 \leq r < 4$, where q is the quotient and r is the remainder. Now the expression can be given as $6q^2+4qr+r(r-1)$. If the value of this expression is even, then $d = 6q^2+4qr+r(r-1)/2$; if this value is odd, then $d = 6q^2+4qr+r(r-1)/2-1$. This equation is more suited for even numbers.

The traditional correlations are very robust when it comes to small datasets. However, as the number of observations keeps increasing, they slowly converge to 0. If the positioning of the observations under $\boldsymbol{X}$ or $\boldsymbol{Y}$ is disturbed, the correlation tends to reverse its sign. Hence, a general correlation coefficient is needed which is free from these defects and fulfills the five criteria mentioned previously.

The cases of the denominator are discussed as follows:

1. $d = [(n-6) + n^2]/3$ is the equation that is most appropriate when n is a multiple of 3. This is shown by the denominator of this equation. The complexity of d when using this equation is $O(n^2)$. However, this equation is restricted by the factor 3, and thus a more general equation is needed.

2. There is a brute force method to compute the value of d, i.e., to generate all random permutations from 0 to $n-1$ for n terms. This method can produce the most accurate value for d, but its computational complexity is as high as $O(n!)$. We give a general algorithm for generating random numbers using permutations in Algorithm 2. The best way to use this method is to sample enough random permutations to obtain an approximation of d.

Algorithm 2 Generating random numbers using permutations

Input: number of data points, n; an initialized array $\boldsymbol{X}_{n!\times n} = \boldsymbol{0}$
Output: d

while $i \leq n!$ **do**
  Generate the i^{th} permutation using perm(i, n)
  Generate perm(i) as a random number on $(0, 1, \cdots, n-1)$
  Repeat the above step until perm(i) is different from perm(0), perm(1), $\cdots$, perm$(n-1)$
  Append perm(i) at location $3 \times (i-1)$ in $\boldsymbol{X}$
end while
Use this array for calculation of d by using sampling and approximation

3. $d = 6q^2 + 4qr + r(r-1)/2$ can be derived by considering the following two functions:

$$u = \text{sum}|x(j) - y^c(j)|, \tag{2}$$

$$v = \text{sum}|y(j) - z^c(j)|, \tag{3}$$

where $z(j) = n - 1 - y(j)$. Now function $t(\boldsymbol{x})$ can be defined as

$$t(\boldsymbol{x}) = \min(|u|, |v|), \tag{4}$$

where $|x|$ is the absolute value of x. Now we compute $d(n)$, which is defined as the maximum of $t(\boldsymbol{x})$ computed for all the permutations of $\boldsymbol{x}$ from 0 to $n-1$. If q is the quotient and r is the remainder for the Euclidean division (Gratton and Kolotilin, 2015) of

n by 4, then we have

$$n = 4q + r. \tag{5}$$

Adding Eqs. (2) and (3), we have

$$c(\boldsymbol{x}) = \text{sum}_i f(x(i), y(i)), \tag{6}$$

where $f(a,b) = |b-a| + |b-(n-1-a)|$. Considering the function f, its value can be given as

$$f(a,b) = \begin{cases} 2a - (n-1), & b \geq a, b \geq n-1-a, \\ 2b - (n-1), & b \leq a, b \geq n-1-a, \\ -2a + (n-1), & b \geq a, b \leq n-1-a, \\ -2b + (n-1), & b \leq a, b \leq n-1-a. \end{cases}$$

Now we compute the values of $f(a,b)$ for various conditions for a particular value of $p \in [0, n-1]$. The values of a and b are varied from p to $n-1-p$ such that one of the following conditions is true:

(1) $p \leq a \leq n-1-p$ and $b = p$;
(2) $p \leq a \leq n-1-p$ and $b = n-1-p$;
(3) $p \leq b \leq n-1-p$ and $a = p$;
(4) $p \leq b \leq n-1-p$ and $a = n-1-p$.

These conditions are based on threshold values for the computation. Next, we compute the maximum value of $c(\boldsymbol{x})$ over all the permutations $x_i \in [0, n-1]$. Because q is the quotient of n divided by 4 and r is the remainder, the corresponding value of c_n is

$$\begin{aligned} c_n &= \sum_{p=0}^{q} (n-1-2p) + r(n-1-q) \\ &= q(n-1) - 2\sum_{p=0}^{q}(p) + r(n-1-q) \\ &= q(n-1) - 2\frac{q(q-1)}{2} + r(n-1-q) \\ &= qn - q - q^2 + q + r(n-1-q) \\ &= q(4q+r) - q^2 + r(4q+r-1-q) \\ &= 4q^2 + qr - q^2 + 4rq + r^2 - r - rq \\ &= 3q^2 + 4qr + r(r-1). \end{aligned} \tag{7}$$

Because we are considering the divisibility factor of 4, let us consider the cases when n is even. Therefore, the value of q is substituted as $q = 2k$. Then the equation is reduced to

$$c_n = 12k^2 + 8kr + r(r-1). \tag{8}$$

Now, $d(n) = c_n/2$; therefore, $d(n)$ is

$$d(n) = 6k^2 + 4kr + r(r-1)/2. \tag{9}$$

5 Simulation results and discussion

To test the proposed approach, a dataset was selected from http://data.worldbank.org/indicator. This dataset contains the World Bank indicators of human development. It contains about 20 indicators for 235 countries in the world. It is a clean dataset with 2355 data values. A mind map containing attributes of the World Bank development indicators is given in Fig. 5. This dataset is a compilation of relevant and high-quality information about the quality of people's lives. The data are internationally comparable statistics about development. The dataset has been used by many researchers (World Bank, 2012; Devarajan, 2013; Susantitaphong et al., 2013). For the purpose of this simulation, we selected random pairs of attributes to test GRCC and ρ on the predictive metric criteria. A data model was created using the GRCC algorithm. This model used the traditional and GRCC algorithms on defined data values. The aim of this simulation was to compare the existing and the proposed correlation metrics to see the difference between them. The comparative analysis is given in Table 2. The symbols used are given in Table 3.

The value g lies between -1 and 1; hence, it is

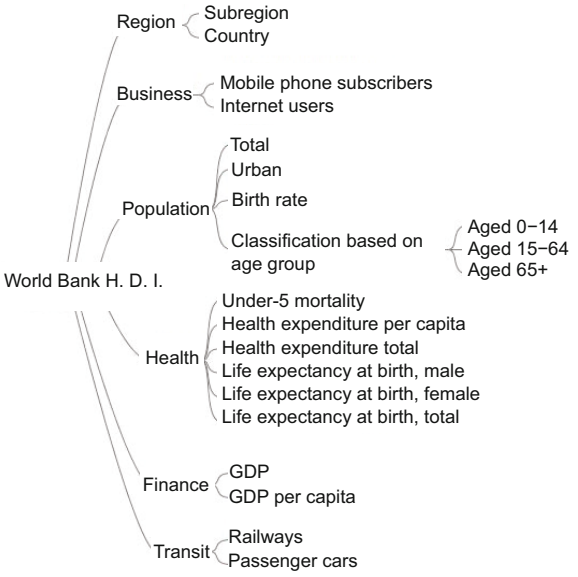

Fig. 5 A mind map representing attributes of World Bank human development indicator dataset

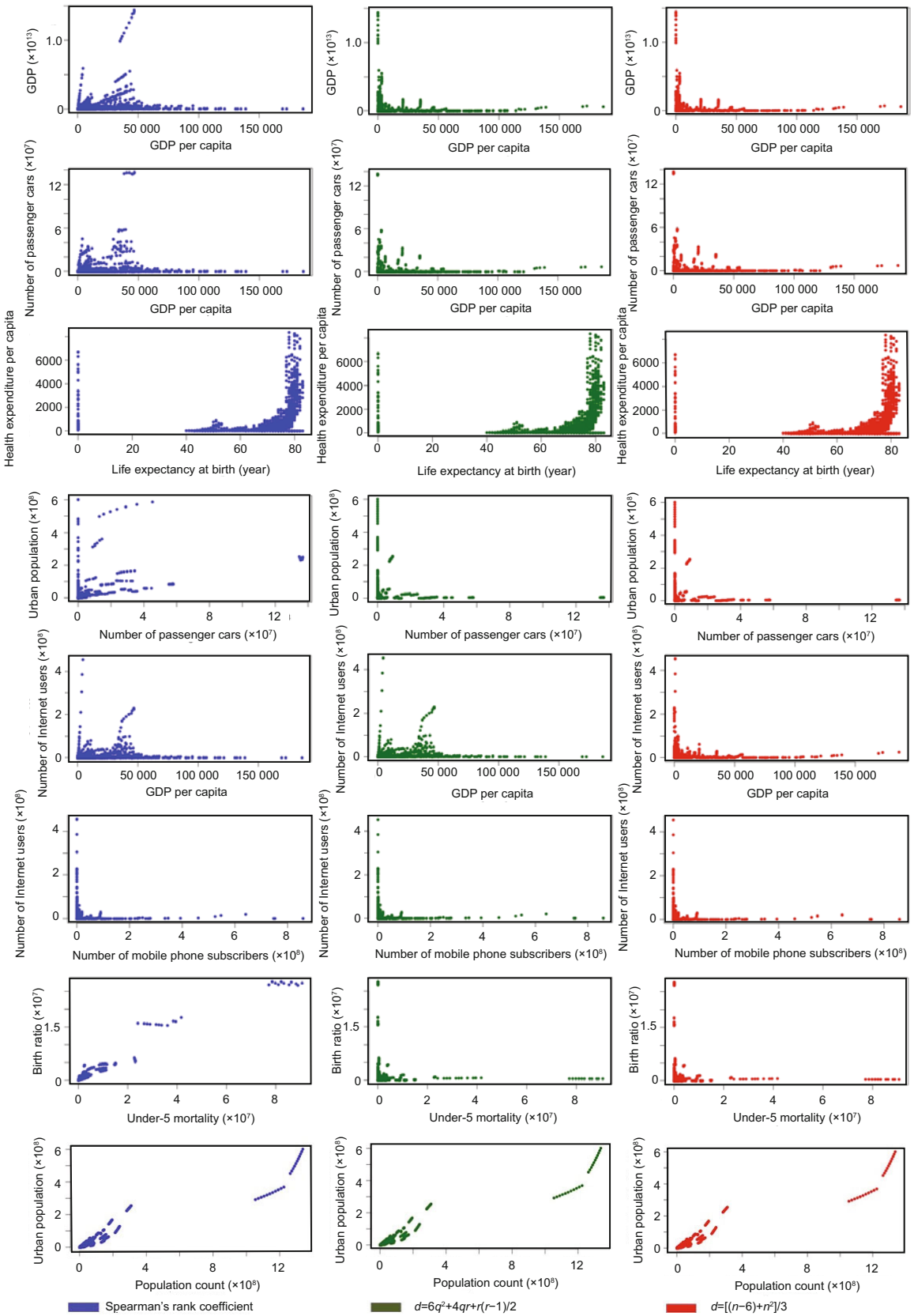

Fig. 6 Scatterplots depicting correlations between variables A and B by applying the two correlation coefficients. References to color refer to the online version of this figure

bounded. It has a bimodal distribution with a small dip near 0. This means that near 0, no patterns would be found. It can be used for the detection of outliers because it uses rank distances instead of squared distances. The squared distances have outliers heavily weighing on them. These results prove that the general correlation coefficient fulfills all the conditions of an effective predictive metric. The comparison of ρ and GRCC based on predictive metric criteria is given in Table 4.

Fig. 6 shows scatterplots depicting correlations between variables A and B by applying the two correlation coefficients. If we study the values of the correlations obtained, it becomes clear that the general correlation coefficient is more intuitive than the traditional Spearman's rank correlation coefficient. The proposed metric has higher correlations between the variables that should have a high correlation, according to the human understanding of the world. For instance, there is a very high correlation between 'car ownership and the GDP per capita', which makes sense, and the relatively low value of Spearman's rank correlation is not indicative of that. Similarly, the high correlation values among attributes such as 'life expectancy at birth (year) and health expenditure per capita', 'passenger cars and urban population', and 'GDP per capita and Internet users' indicate the same. Also, certain attributes demonstrate low values of correlation as compared with Spearman's rank correlation coefficient, like 'under-5 mortality and birth rate' and 'total population and urban population'. Intuitively, the correlation between these parameters should be low. In some cases, the two coefficients have almost equal values, like

Table 3 Symbol table

Symbol	Description
A, B	Attributes in which correlations are to be found
p value	Statistical significance parameter
S value	Statistical significance parameter
ρ	Spearman's rank correlation
g	General rank-based correlation coefficient

Table 2 Comparison of Spearman's rank correlation coefficient and GRCC

Data attribute		Number of data points	Spearman's rank correlation coefficient			GRCC g			
						$d=d_1$		$d=d_2$	
A	B		p value	S value	ρ	$c=1$	$c=2$	$c=1$	$c=2$
GDP per capita	GDP	2327	$<2.2\times10^{-16}$	900 579 181	0.4600	0.4302	0.3216	0.4130	0.3092
GDP per capita	Passenger cars	2354	$<2.2\times10^{-16}$	1 451 827 312	0.3321	1.0000	1.0000	1.0000	1.0000
Life expectancy	Health expenditure per capita	2327	$<2.2\times10^{-16}$	858 399 753	0.5380	0.9999	0.9984	0.9997	0.9842
Passenger cars	Urban population	2354	$<2.2\times10^{-16}$	1 365 948 542	0.3717	0.7480	0.9890	0.7030	0.9420
GDP per capita	Internet users	2354	$<2.2\times10^{-16}$	1 101 037 602	0.4935	0.7551	0.8821	0.8155	0.8122
Mobile subscribers	Internet users	2354	$<2.2\times10^{-16}$	224 906 329	0.8965	0.7551	0.6128	0.4130	0.8551
Under-5 mortality	Birth rate	2354	$<2.2\times10^{-16}$	120 426 906	0.9440	0.5110	0.4626	0.4780	0.4400
Total population	Urban population	2354	$<2.2\times10^{-16}$	104 317 436	0.9520	0.4303	0.3416	0.3193	0.3253

$d_1=6q^2+4qr+r(r-1)/2, d_2=[(n-6)+n^2]/3$

Table 4 Comparison of ρ and GRCC based on predictive metric criteria

Data attribute		Fulfills rank correlation ρ?					Fulfills GRCC g?				
A	B	SS	V	C	S	I	SS	V	C	S	I
GDP per capita	GDP	×	√	×	√	√	√	√	√	√	√
GDP per capita	Passenger cars	×	√	×	×	×	√	√	√	√	√
Life expectancy	Health expenditure per capita	×	√	×	×	×	√	√	√	√	√
Passenger cars	Urban population	×	√	×	×	×	√	√	√	√	√
GDP per capita	Internet users	×	√	×	×	×	√	√	√	√	√
Mobile subscribers	Internet users	×	√	×	√	√	√	√	√	√	√
Under-5 mortality	Birth rate	×	√	×	×	×	√	√	√	√	√
Total population	Urban population	×	√	×	√	×	√	√	√	√	√

SS: sample size; V: value; C: complexity; S: sensitivity; I: intuitiveness

'GDP per capita and GDP' and 'mobile phone subscribers and Internet users', showing that the two values do tend to converge for some cases. These cases are no exception to the rule. They just show that there might be some instances when these results are possible, and the results of the proposed and the existing metric would coincide.

Another simulation was performed using synthetic datasets. We generated data using random number distributions to test GRCC at n=15 000, 45 000, 75 000, and 100 000 data points. The aim of this simulation was to test the behavior of GRCC at different values of n and compare it to ρ. The random number distributions are binomial distribution, exponential distribution, normal distribution, Poisson distribution, and uniform distribution. The results of the simulation are given in Table 5 and the corresponding graphs are represented in Figs. 7–11. It can also be observed in all the distributions that GRCC is independent of sample size because it continues to give consistent results irrespective of the sample size. As the sample size increases, GRCC for $c = 1$ performs better than ρ, making it more accurate for large datasets.

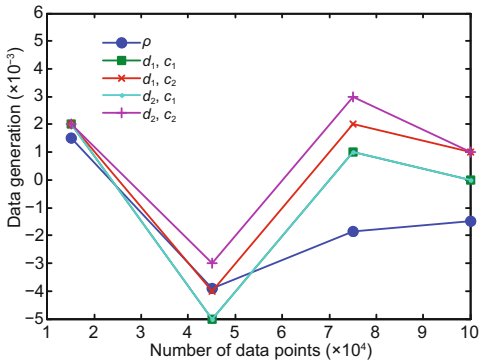

Fig. 7 Comparison of GRCC with ρ for the binomial distribution

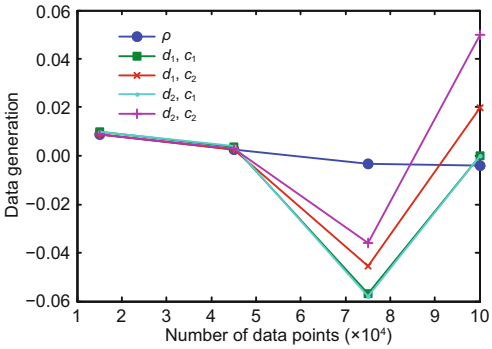

Fig. 8 Comparison of GRCC with ρ for the exponential distribution

Table 5 Comparison of ρ and GRCC at c=1 and c=2

Simulation method	Data generation		ρ	GRCC g			
	Parameter	Number of data points		$d = d_1$		$d = d_2$	
				$c = 1$	$c = 2$	$c = 1$	$c = 2$
Uniform random variables		15 000	−0.0059	−0.0061	−0.0059	−0.0060	−0.0059
		45 000	−0.0020	−0.0025	−0.0020	−0.0015	−0.0020
		75 000	−0.0030	0	−0.0030	0	−0.0030
		100 000	0.0030	0	0.0010	0	0.0010
Exponential random variables	Mean: 0.2	15 000	0.0088	0.0099	0.0089	0.0099	0.0088
		45 000	0.0027	0.0035	0.0027	0.0040	0.0028
		75 000	−0.0033	−0.0568	−0.0454	−0.0580	−0.0358
		100 000	−0.0042	0	0.0200	0	0.0500
Normal random variables	Mean: 15; standard deviation: 2	15 000	−0.0047	−0.0050	−0.0035	−0.0054	−0.0047
		45 000	−0.0053	−0.0070	−0.0060	−0.0080	−0.0055
		75 000	0.0025	0	0.0030	0	0.0050
		100 000	0.0011	0	0	0	0
Poisson random variables	Mean: 10	15 000	0.0017	0.0020	0.0015	0.0030	0.0019
		45 000	0.0074	0.0090	0.0070	0.0090	0.0060
		75 000	0.0048	0	0.0040	0	0.0050
		100 000	−0.0058	0	0.0050	0	0.0030
Binomial random variables	Mean: 10; probability of success: 0.3	15 000	0.0015	0.0020	0.0020	0.0020	0.0020
		45 000	−0.0039	−0.0050	−0.0040	−0.0050	−0.0030
		75 000	−0.0019	0.0010	0.0020	0.0010	0.0030
		100 000	−0.0015	0	0.0010	0	0.0010

$d_1 = 6q^2 + 4qr + r(r-1)/2, d_2 = [(n-6) + n^2]/3$

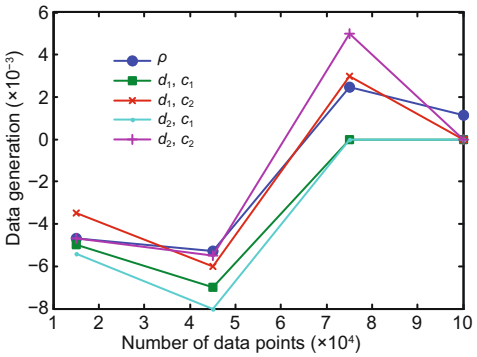

Fig. 9 Comparison of GRCC with ρ for the normal distribution

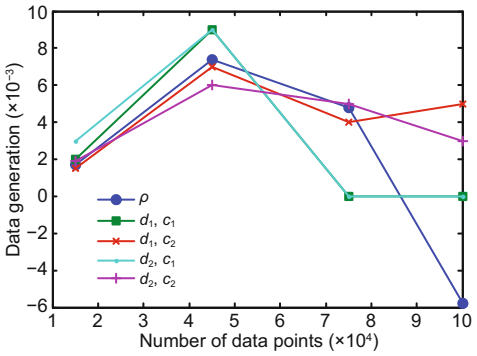

Fig. 10 Comparison of GRCC with ρ for the Poisson distribution

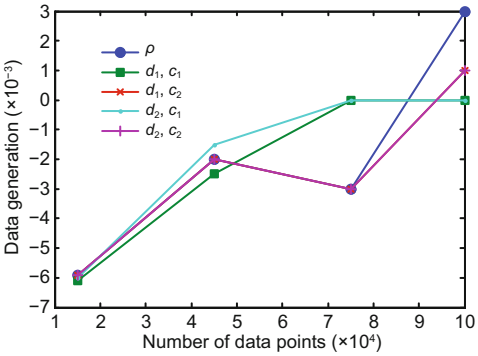

Fig. 11 Comparison of GRCC with ρ for the uniform distribution

6 Conclusions and future scope

We have discussed an algorithm to improve a traditional correlation metric. It defines a new general rank-based correlation coefficient that gives a more intuitive correlation between two variables. It is more accurate when the number of observations is large. The purpose of this metric is to quantify the meaning of correlation analysis in the world of big data. This metric is an improvement over the existing Spearman's rank correlation in terms of predictive powers. It is also independent of the size of the sample, not sensitive to outliers, works on measurement of the degree of monotonicity instead of linearity, and is more intuitive than the existing metric.

The mathematical proofs for the required equations and simulations performed on the proposed scheme validated it. A lot of work can be done to refine this metric and make it more concrete, like finding its statistical significance. The most important application of this technique will be in the field of prediction analysis. It can be used in many applications, such as online predictive systems, health care information systems, political prediction systems, and financial market recommendation systems.

References

Chaudhuri B, Bhattacharya A, 2001. On correlation between two fuzzy sets. *Fuzzy Sets Syst*, 118(3):447-456. https://doi.org/10.1016/S0165-0114(98)00347-9

Chen H, Chiang RHL, Storey VC, 2012. Business intelligence and analytics: from big data to big impact. *MIS Q*, 36(4):1165-1188.

Chen N, Xu Z, Xia M, 2013. Correlation coefficients of hesitant fuzzy sets and their applications to clustering analysis. *Appl Math Model*, 37(4):2197-2211. https://doi.org/10.1016/j.apm.2012.04.031

Davenport T, Barth P, Bean R, 2013. How 'Big Data' is Different. https://sloanreview.mit.edu/article/how-big-data-is-different/

Deufemia V, Giordano M, Polese G, et al., 2014. A visual language-based system for extraction-transformation-loading development. *Softw Pract Exper*, 44(12):1417-1440. https://doi.org/10.1002/spe.2201

Devarajan S, 2013. Africa's statistical tragedy. *Rev Income Wealth*, 59(S1):9-15. https://doi.org/10.1111/roiw.12013

Didelez V, Pigeot I, 2001. Judea Pearl: causality: models, reasoning, and inference. *PVS*, 42(2):313-315. https://doi.org/10.1007/s11615-001-0048-3

Ginsberg J, Mohebbi MH, Patel RS, et al., 2009. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012-1014. https://doi.org/10.1038/nature07634

Granville V, 2014. Developing analytic talent: becoming a data scientist. John Wiley & Sons, Inc., Indianapolis, USA.

Gratton G, Kolotilin A, 2015. Euclidean fairness and efficiency. *Econ Inq*, 53(3):1689-1690. https://doi.org/10.1111/ecin.12193

Hauke J, Kossowski T, 2011. Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. *Quaest Geograph*, 30(2):87-93. https://doi.org/10.2478/v10117-011-0021-1

Hong DH, 2006. Fuzzy measures for a correlation coefficient of fuzzy numbers under T_W (the weakest t-norm)-based fuzzy arithmetic operations. *Inform Sci*, 176(2):150-160. https://doi.org/10.1016/j.ins.2004.11.005

Hung WL, 2001. Using statistical viewpoint in developing correlation of intuitionistic fuzzy sets. *Int J Uncert Fuzz Knowl Based Syst*, 9(4):509-516. https://doi.org/10.1142/S0218488501000910

Huo X, Székely GJ, 2016. Fast computing for distance covariance. *Technometrics*, 58(4):435-447. https://doi.org/10.1080/00401706.2015.1054435

Kitano H, 2002. Systems biology: a brief overview. *Science*, 295(5560):1662-1664. https://doi.org/10.1126/science.1069492

Kong J, Klein BEK, Klein R, et al., 2012. Using distance correlation and SS-ANOVA to assess associations of familial relationships, lifestyle factors, diseases, and mortality. *PNAS*, 109(50):20352-20357. https://doi.org/10.1073%2Fpnas.1217269109

Li R, Zhong W, Zhu L, 2012. Feature screening via distance correlation learning. *J Am Stat Assoc*, 107(499):1129-1139. https://doi.org/10.1080/01621459.2012.695654

Liao H, Xu Z, Zeng X, et al., 2015a. Qualitative decision making with correlation coefficients of hesitant fuzzy linguistic term sets. *Knowl Based Syst*, 76:127-138. https://doi.org/10.1016/j.knosys.2014.12.009

Liao H, Xu Z, Zeng X, 2015b. Novel correlation coefficients between hesitant fuzzy sets and their application in decision making. *Knowl Based Syst*, 82:115-127. https://doi.org/10.1016/j.knosys.2015.02.020

Linden G, Smith B, York J, 2003. Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Intern Comput*, 7(1):76-80. https://doi.org/10.1109/MIC.2003.1167344

Liu S, Kao C, 2002. Fuzzy measures for correlation coefficient of fuzzy numbers. *Fuzzy Sets Syst*, 128(2):267-275. https://doi.org/10.1016/S0165-0114(01)00199-3

Lyons R, 2013. Distance covariance in metric spaces. *Ann Probab*, 41(5):3284-3305. https://doi.org/10.1214/12-AOP803

McGregor C, 2013. Big data in neonatal intensive care. *Computer*, 46(6):54-59. https://doi.org/10.1109/MC.2013.157

Mitchell HB, 2004. A correlation coefficient for intuitionistic fuzzy sets. *Int J Intell Syst*, 19(5):483-490. https://doi.org/10.1002/int.20004

Murthy CA, Pal SK, Majumder DD, 1985. Correlation between two fuzzy membership functions. *Fuzzy Sets Syst*, 17(1):23-38. https://doi.org/10.1016/0165-0114(85)90004-1

Reshef DN, Reshef YA, Finucane HK, et al., 2011. Detecting novel associations in large data sets. *Science*, 334(6062):1518-1524. https://doi.org/10.1126/science.1205438

Ritala P, Golnam A, Wegmann A, 2014. Coopetition-based business models: the case of Amazon.com. *Ind Mark Manag*, 43(2):236-249. https://doi.org/10.1016/j.indmarman.2013.11.005

Sen A, Dacin PA, Pattichis C, 2006. Current trends in web data analysis. *Commun ACM*, 49(11):85-91. https://doi.org/10.1145/1167838.1167842

Susantitaphong P, Cruz DN, Cerda J, et al., 2013. World incidence of AKI: a meta-analysis. *Clin J Am Soc Nephrol*, 8(9):1482-1493. https://doi.org/10.2215/CJN.00710113

Székely GJ, Rizzo ML, 2012. On the uniqueness of distance covariance. *Stat Probab Lett*, 82(12):2278-2282. https://doi.org/10.1016/j.spl.2012.08.007

Volpone SD, Tonidandel S, Avery DR, et al., 2015. Exploring the use of credit scores in selection processes: beware of adverse impact. *J Bus Psychol*, 30(2):357-372. https://doi.org/10.1007/s10869-014-9366-5

World Bank, 2012. World Development Indicators 2012. World Development Indicators, Washington DC, USA. https://openknowledge.worldbank.org/handle/10986/6014

Xiao C, Ye J, Esteves R, et al., 2015. Using Spearman's correlation coefficients for exploratory data analysis on big dataset. *Concurr Comput Pract Exp*, 28(14):3866-3878. https://doi.org/10.1002/cpe.3745