Mathematical Biosciences and Engineering

MBE, 21(1): 1356–1393.
DOI: 10.3934/mbe.2024059
Received: 16 July 2023
Revised: 30 October 2023
Accepted: 23 November 2023
Published: 27 December 2023

http://www.aimspress.com/journal/MBE

Research article

New research for detecting complex associations between variables with randomness

Yuwen Du, Bin Nie*, Jianqiang Du, Xuepeng Zheng, Haike Jin and Yuchao Zhang

School of Computer, Jiangxi University of Chinese Medicine, Nanchang 330004, China

* **Correspondence:** Email: ncunb@163.com.

Abstract: Many correlation analysis methods can capture a wide range of functional types of variables. However, the influence of uncertainty and distribution status in data is not considered, which leads to the neglect of the regularity information between variables, so that the correlation of variables that contain functional relationship but subject to specific distributions cannot be well identified. Therefore, a novel correlation analysis framework for detecting associations between variables with randomness (RVCR-CA) is proposed. The new method calculates the normalized RMSE to evaluate the degree of functional relationship between variables, calculates entropy difference to measure the degree of uncertainty in variables and constructs the copula function to evaluate the degree of dependence on random variables with distributions. Then, the weighted sum method is performed to the above three indicators to obtain the final correlation coefficient R. In the study, which considers the degree of functional relationship between variables, the uncertainty in variables and the degree of dependence on the variables containing distributions, cannot only measure the correlation of functional relationship variables with specific distributions, but also can better evaluate the correlation of variables without clear functional relationships. In experiments on the data with functional relationship between variables that contain specific distributions, UCI data and synthetic data, the results show that the proposed method has more comprehensive evaluation ability and better evaluation effect than the traditional method of correlation analysis.

Keywords: correlation analysis; information entropy; cubic B-spline; copula function; analytic hierarchy process

1. Introduction

Correlation analysis is particularly important in various fields. The important relationships that may be implied between different variables can be found by measuring the closeness between variables [1]. Correlation analysis is widely used in many fields such as finance, medicine, industry and biology. For example, Zhou et al. [2] investigated the correlation between the nutritional status and prognosis of COVID-19 patients by using multivariate logistic regression analysis. Xu et al. [3] revealed the correlations between the factors and the system status through statistical properties of data, and explored the related factors affecting the operating status of power systems. Liu et al. [4] investigated the relationship between intestinal flora content and hypertension, the results showed that the content of intestinal flora has a significant correlation with hypertension.

At present, the correlation analysis methods of variables can be roughly divided into three categories, namely, the correlation measure methods based on statistics, the correlation measure methods based on information theory and the measure methods based on similarity [5]. The correlation measure methods based on statistics include Pearson correlation coefficient, Kendell's coefficient and Spearman's coefficient etc. [6]. The measure methods based on information theory [7] include mutual information, maximum information coefficient (MIC) [8] and information gain etc. [9]. The similarity measure methods include distance correlation [10], Jaccard correlation and cosine similarity, etc. [11]. However, the traditional methods of correlation analysis do not comprehensively consider the impact of data uncertainty and distribution [12,13], which leads to the neglect of strong regularity information in the variables, so that the correlation coefficient value between variables with strong regularity is too small, and it is believed that there is no correlation or weak correlation between variables. Since the Pearson coefficient can reflect only the degree of linear correlation between variables, and it requires that variables conform to normal distribution. The MIC can capture a variety of functional relationship types, but it is not sensitive to the functional relationships containing specific distributions.

There are many studies on resolving the distribution consistency, uncertainty and randomness in data currently. For example, Gabriela et al. [14] use the paraconsistent logic, which can provide a compelling quantitative analysis approach in classification algorithms because it deals directly with inaccurate, inconsistent and incomplete data. Xin et al. [15] apply the chance theory to deal with the analysis of indeterminacy, including both uncertainty and randomness, to study two types of linear quadratic (LQ) optimal control models for multistage uncertain random systems. The first model is an LQ model with additive noises, while the second model is an LQ model with both multiplicative noises and additive noises. Yang et al. [16] provide a flexible framework for characterizing uncertainty in the outputs of physical systems due to randomness in their inputs or noise in their observations. Ayensa et al. [17] consider that data are never uncertainty-free and a suitable approach is needed to face data measurement errors and their intrinsic randomness in problems with well-established physical constraints. Additionally, Villiers et al. [18] proposes that if uncertainties in the modeling process are not accounted for, fusion processes may provide under- or overconfident results, or in some cases incorrect results. The authors establish four abstract processes to verify the situation in which uncertainty affects the modeling process. The above literatures all proved that the noise, uncertainty and randomness in data have certain influence on data analysis.

However, for the correlation analysis of data, most of the methods apply some preprocessing methods to alleviate the impact of noise and randomness on the correlation evaluation. For example, Niven et al. [19] believe that traditional means of calculating correlation coefficients are known to be

adversely affected by outlier data, thus a new method for calculating a robust correlation coefficient is proposed based on a weighted average correlation calculated from different combinations or subsets of the original data, which is more robust than Pearson's or Spearman's correlation coefficients, but the uncertainty and randomness in the data are not analyzed and measured. Johnson et al. [20] removed the outliers in the data first and then calculated the correlation from the remaining data, but the outlier data may be reliable data in some case. In addition, the copula function is considered that it contains all the dependent relationships between random variables. Moreover, the copula theory believes that it is unreasonable to analyze the correlation between variables if they conform to different distribution statuses [21]. Therefore, in the copula theory, the variables are transformed into same distribution first, and then the correlation between variables is described by constructing the joint distribution of random variables. The copula theory has strong practicability, flexibility and robustness in analyzing the correlation structure of random variables [22]. Ma et al. [23] study the statistical relationship between random variables from data with association measures, and copula entropy is used to measure the degree of independence between joint distribution and edge distribution for random variables, so as to quantify the dependence of multivariate random variables.

As mentioned above, there are many studies on data noise, uncertainty and randomness indeed, there are few researches that link factors such as data noise, uncertainty and randomness with correlation between variables. Most correlation analysis methods solve the problems of noise and randomness through some pre-processing methods. Furthermore, we find that most correlation analysis methods take the degree that variables conform to certain functional relationship as the only size indicator for correlation. Suppose that the closer the distribution of data points for variables is to the image distribution of a quadratic function, the larger the correlation coefficient of the variable; however, if some data points are added to make the image distribution of data points close to the quadratic function gradually wider, the correlation coefficient will gradually decrease, but the correlation between the variables is strong, so the problem is that the correlation between variables is not weak, but the value of correlation coefficient become smaller. Furthermore, there exist strong correlation and regularity between functional relationship variables with specific distributions according to the scatter plots, so it is necessary to comprehensively evaluate the correlation between variables from multiple perspectives.

Furthermore, in many traditional methods of correlation analysis, it is considered that there exists complete functional relationship between variables when the coefficient value is equal to 1, that is, there exist accurate expression between variables. When the coefficient value is between 0 and 1, it is indicated that there exists dependent relationship between variables. The variables are independent of each other when the correlation value is equal to 0. For correlation relationships, there are no one-to-one functional mapping between the variables, so the tendency of a scatter plot for data can be observed. If there exist regularity between variables, the correlation between variables can be analyzed by regression analysis or correlation analysis methods. Additionally, the correlation between variables needs to be evaluated from multiple perspectives.

The rest of the paper is organized as follows. In Section 2, the related theories applied in this paper are described. In Section 3, the novel correlation analysis framework proposed in this paper is introduced. In Section 4, the proposed correlation analysis framework is applied to various datasets and the correlation of variables are evaluated. In Section 5, we discuss the correlation analysis accuracy compared traditional correlation analysis methods. we provide a summary of this paper in Section 6.

2. Related theories

2.1. The approximate fitting based on cubic B-spline

2.1.1. The cubic B-spline

A spline is a smooth curve through nodes, which is defined by the constraints of control points and nodes. The spline fitting can be regarded as a piecewise fitting, that is, the specific data is divided into multiple segments, and each segment can be fitted to obtain a polynomial. When the basis function of the spline curve is a cubic polynomial, the curve obtained by fitting it is a cubic B-spline [24] curve. The definition of the basis function of cubic B-splines is as follows:

$$N_{i,k}(t) = \frac{1}{k!}\sum_{m=0}^{k-i}(-1)^m \binom{m}{k+1}\left(t+k-m-j\right)^k \tag{1}$$

t represents the node, $t \in [t_0, t_1, t_2, ..., t_m]$, $k = 3$, j represents the number of control points and m represents the number of nodes.

Then, the B-spline curve equation obtained by the basis function of the cubic B-spline is:

$$C(t) = P_j \bullet N_{i,k}(t) \tag{2}$$

where P_j represents the control point, and $\bullet$ represents the point multiplication. The process of spline fitting is the process of inversely finding the control points by enumerating the curve equations.

2.1.2. The approximate based on cubic B-spline

Spline fitting can divide into interpolation fitting and approximate fitting. The difference between the two is whether all data points are passed through by the curve during fitting. The interpolation curve may pass through all data points instead of closely following the data polygon. In order to overcome this problem, approximation techniques are introduced, which relax the strict requirement that the curve must contain all data points.

In global approximation, the curve need not contain every data point except the first and last ones. To measure how well a curve "approximates" to a given data polygon, the concept of error distance is used. The error distance is the distance from the data point to the "corresponding" point on the curve. Therefore, if the sum of these error distances is minimum, the curve should closely follow the shape of the data polygon. Curves obtained in this way are called approximate curves.

2.2. Information entropy

Information entropy is the measure of reduction degree in the uncertainty of an event, which can be used to measure the uncertainty of information [25,26].

According to the information entropy formula proposed by Shannon, for any random variable X, suppose the variable X is a discrete variable with n number of values, where the number of unequal values is s, then its information entropy is defined as follows:

$$H(X) = -\sum_{i=1}^{s} p(x_i) \log p(x_i) \tag{3}$$

where $x \in X$, x_i represents the *i*-th unequal values that belongs to variable *X*, and $p(x)$ is the probability of something happening, and the unit of entropy is bit.

2.3. Copula function

Copula function can connect the joint distribution of multidimensional random variables $x_i, x_2, ..., x_n$ with their respective marginal distributions $F_1(x_1), F_2(x_2), ..., F_n(x_n)$ Sklar's [27] theorem has demonstrated that in the general case there exists a multivariate real function $C(u_1, u_2, ...u_n)$, making

$$F(x_1, x_2, ..., x_n) = C(F_1(x_1), F_2(x_2), ..., F_n(x_n)) \tag{4}$$

This multivariate function *C* is the copula.

Copula is considered to contain all dependencies between random variables, which can reflect linear relationships and describe nonlinear relationships. Therefore, choosing which copula function to construct the correlation structure of variables is an important subject [28]. Copula models can be mainly divided into two categories, which include elliptic copula and Archimedes copula. Archimedes Copula, including three types of functions: Clayton copula, Gumbel copula and Frank copula, which are often used to analyze the correlation structure of binary random variables. We can evaluate the effect of each copula model by calculating the errors of empirical copula and theoretical copula, as well as Akaike information criterion (AIC) and Bayesian Information Criterion (BIC). The calculation formulas of AIC and BIC are shown in Eq (5):

$$\begin{aligned} AIC &= m \ln MSE + 2k \\ BIC &= m \ln MSE + 2k \ln m \end{aligned} \tag{5}$$

where m represents the number of samples, k represents the number of parameters of the model and the calculation formula of MSE is as follows:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (F(x_i) - F_i(x_i))^2 \tag{6}$$

2.4. Spectral clustering

Spectral clustering [29] is a graph-based machine learning algorithm. The graph-based algorithm regards the sample data as the vertices of the graph, constructs edges according to the distances between the data points to form a graph with weights and completes the functions required by the algorithm through processing the graph. For the clustering problem, it is realized by graph cutting, which is to divide the graph into multiple subgraphs. These subgraphs are the clusters. The spectral clustering algorithm constructs an adjacency graph (also known as a similarity graph) of the sample set and obtains the Laplacian matrix of the graph. Next, the matrix is decomposed into eigenvalues, and clusters are constructed by processing the eigenvectors.

2.5. Analytic hierarchy process

Analytic Hierarchy Process (AHP) [30,31] is a systematic and hierarchical analysis method combining qualitative and quantitative, which can assign weights to multiple different indicators. In this paper, we use the largest eigenvalue method in AHP to determine the weight coefficient. The first step is to construct the relative importance matrix of the factors, then to calculate the eigenvector corresponding to the maximum eigenvalue of the matrix, and the eigenvector is normalized to obtain the final weight matrix; The second step is to perform consistency check on the obtained weight matrix, and calculate the consistency index CR. Equation (7) is shown below,

$$CI = \frac{\lambda_{\max} - n}{n-1}, CR = \frac{CI}{RI} \tag{7}$$

where RI represents the average random consistency, which can be acquired by checking out the table; CR < 0.1, it can be judged that the matrix has consistency, that is, the matrix meets consistency requirements.

3. The proposed correlation analysis framework

In order to consider the degree of functional relationship between variables, the uncertainty in variables and the degree of dependence on the variables containing distributions, a new correlation analysis framework is proposed in this paper.

The relationships between variables are divided into the functional relationships with exact expression and the correlation relationships without explicit function expression, so that we will use the proposed analysis framework to analyze the correlation of the two relationship types. The proposed correlation analysis framework is shown in Figure 1, and the correlation analysis framework is specific to continuous bivariate in this paper.

As shown in Figure 1, the correlation analysis framework proposed in this paper is divided into three modules, namely, the correlation analysis module of functional variables with specific distribution, the correlation analysis module of variables without explicit functional relationship, and the calculation module for weight coefficients. For each correlation analysis module, which is also divided into three stages:

1) Measure the degree of functional relationship between variables. Calculating the RMSE of random distribution to evaluate the degree of functional relationship between variables.

2) Measure the degree of uncertainty in the distribution of variables. Calculating the uncertainty information caused by random distribution between variables first, and then the uncertainty degree of random distribution between variables is evaluated according to the information entropy difference between the functional relationship variables and the functional relationship variables containing specific distribution.

3) Measure the degree of dependence on the random distribution of the variables. First, the copula model for random distribution of variables is constructed to represent the correlation structure of variables. Then the optimal copula function is selected according to the error between empirical copula and theoretical copula, AIC and BIC. Finally, the relevant parameter corresponding to the optimal copula is selected to evaluate the degree of dependence on the random distribution of variables.

The stages of calculation module for weight coefficient include construct the factor importance matrix, consistency test and maximum eigenvalue method to generate final weighted coefficients.

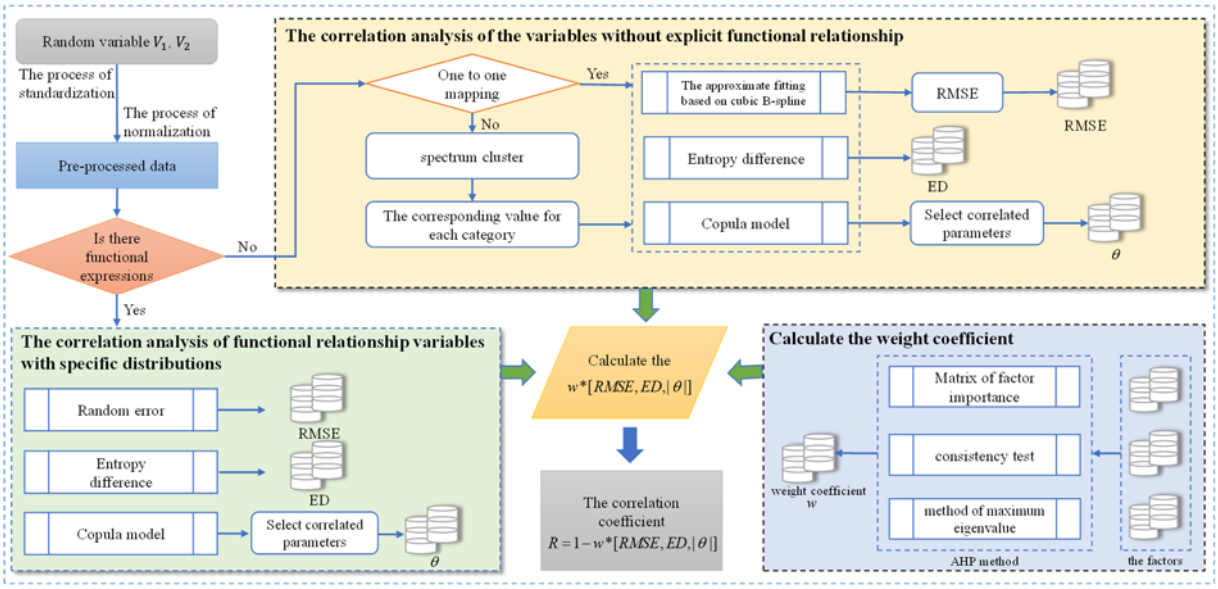

Figure 1. the proposed correlation analysis framework in this paper.

3.1. The description of the correlation analysis framework proposed in this paper

In this section, the correlation analysis framework for the functional relationship variables with distribution and the variables without explicit functional relationship will be introduced, respectively, and the framework is described by flowchart and construction process in detail. The flowchart of correlation analysis framework for functional relationship variables with specific distributions is shown in Figure 2, and the flowchart of correlation analysis framework for variables without explicit function relationship is shown in Figure 3.

3.1.1 The correlation analysis framework for functional relationship variables with specific distributions

The construction process of the correlation analysis framework for functional relationship variables with specific distributions is as follows:

Assuming random variables X and Y, and there is a clear functional relationship between X and Y. The function is denoted by $Y = CX + \sigma$, σ represents the values that satisfy specific distribution.

Step 1: Standardize and normalize the random variables. The standardized and normalized variables E_0, E_1 and E_2 corresponding to X, Y and Y_1 is obtained. The functional relationship between variables Y_1 and X is denoted by $Y_1 = CX$.

Step 2: Calculate the root mean square error (RMSE) between E_1 and E_2, then the normalized RMSE is obtained.

Step 3: Calculate the normalized information entropy of variables X and Y_1. First, we use MIC to obtain the optimal partition interval for the variables X and Y_1, and enumerating all the partition results. Then, the information entropy of variables X and Y_1 is calculated after the intervals is divided, and the information entropy is normalized. Finally, the normalized information entropy

corresponding to the optimal partition interval is obtained.

Step 4: Construct the copula function. First, the random variables are transformed into the probability distribution function which conforms to the normal distribution. Then, the copula function is constructed on the transformed data.

Step 5: Obtain the correlation parameter θ. Calculate the error between empirical copula and theoretical copula, and obtain the AIC and BIC of the copula model. Then, the correlation parameter corresponding to the best copula function is selected.

Step 6: Obtain the weight w. The factor importance matrix is constructed, and the AHP method is applied to obtain the weight matrix w. The consistency test is also used to determine whether the factor importance matrix is reasonable.

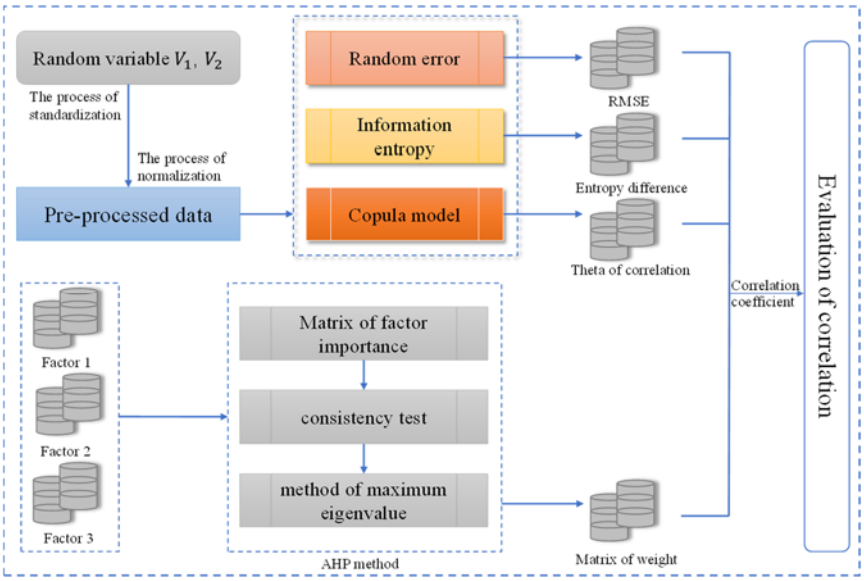

Figure 2. The flowchart of correlation analysis framework for functional relationship variables with specific distributions.

3.1.2 The correlation analysis framework for variables without explicit functional relationship

The construction process of the correlation analysis framework for variables without an explicit functional relationship is as follows:

Assuming random variables X and Y, and there does not exist specific functional relationship between X and Y. The value obtained by the spline fitting is denoted by Y_1.

Step 1: Standardize and normalize the random variables. The standardized and normalized variables E_0, E_1 and E_2 corresponding to X, Y and Y_1 is obtained. The functional relationship between variables Y_1 and X is denoted by $Y_1 = P_i \sum_{k=1}^{m} N_{i,p}(t_k)$.

Step 2: Perform the spectrum cluster. If there doesn't exist one to one mapping between variables, the spectrum cluster is performed to obtain multiple category data.

Step 3: Perform the approximate fitting based on cubic B-spline. The approximation fitting based on cubic B-spline is used to each category data, and the curve expression after fitting is obtained

Step 4: Calculate the root mean square error (RMSE). The final spline RMSE is the sum of the errors of each category.

Step 5: Calculate the normalized information entropy of variables X and Y_1. First, using MIC to obtain the optimal partition interval for the variables X and Y_1, and enumerating all the partition results. Then, the information entropy of variables X and Y_1 is calculated after the intervals is divided, and the information entropy is normalized. Finally, the normalized information entropy corresponding to the optimal partition interval is obtained.

Step 6: Construct the copula function. First, the random variables are transformed into the probability distribution function which conforms to the normal distribution. Then, the copula function is constructed on the transformed data.

Step 7: Obtain the correlation parameter θ. Calculating the error between empirical copula and theoretical copula, the AIC and BIC of the copula model. Then, the correlation parameter corresponding to the best copula function is selected.

Step 8: Obtain the weight w. The factor importance matrix is constructed, and the AHP method is applied to obtain the weight matrix w. The consistency test is also used to determine whether the factor importance matrix is reasonable.

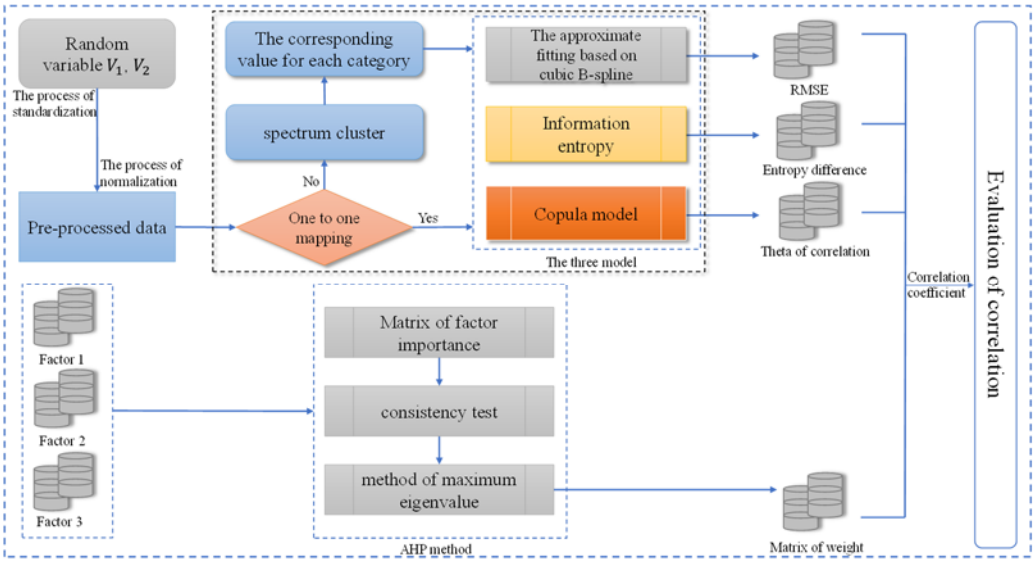

Figure 3. The flowchart of a correlation analysis framework for variables without an explicit functional relationship.

3.2. The degree of functional relationship between variables

There may exist functional relationships between correlated variables but subject to some error distributions and randomness. In statistics, regression analysis is often used to explore the relationship between variables. Through regression analysis, the regression equation between variables can be obtained, which can approximately reflect the closeness of the association and the general regulation of changes between variables. Therefore, we can measure the degree of functional relationship between variables by calculating the error between the original values and the fitted values after regression. If the error is equal to 0, it means that there exists perfect functional relationship between variables; the smaller the error, the closer the association between variables, and it indicates that the functional relationship between variables is stronger.

In this paper, the relationship of variables is divided into two cases, which include the explicit functional relationship containing specific distribution and the correlated relationship without explicit function. Algorithm 1 shows the pseudo-code for calculating the RMSE of variables with defined function or undefined function relationship. In order to remove the dimension of data and make the results obtained from the model comparable, so the data are standardized and normalized first, then the preprocessed data are between 0 and 1. The formula of standardization and normalization is shown in Eq (8).

For functional relationship variables with specific distributions, the degree of functional relationship between variables is evaluated by calculating the root mean square error (RMSE) in the distribution of variables. The calculation formula of RMSE is shown in Eq (9).

$$E = \frac{x - \overline{x}}{\sigma}, e = \frac{E - E_{\max}}{E_{\max} - E_{\min}} \tag{8}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(e_i - \hat{e}_i)^2} \tag{9}$$

For the variables without an explicit functional relationship, the function expression between variables was obtained by approximate fitting based on cubic B-spline [32,33]. Then, the degree of the functional relationship between variables is evaluated by calculating the RMSE between the original data and the fitted data. First, we divide data into segments, then the approximate fitting based on cubic B-spline is performed on each segment, and the function expression corresponding to each segment is obtained; finally, the function expression function based on spline curve is obtained by summing all the segmented expressions [34]. Since the interpolation fitting requires the curve to pass each data point, and the curve obtained by interpolation fitting is complex and irregular, which cannot reflect the shape trend of the original data well. Therefore, the idea of the least square is introduced in an approximate technique [35], and the condition of minimizing the error distance between the original value and the functional value is added into the constraints of cubic B-spline fitting. The approximate fitting does not strictly require the curve to pass all data points, and the curve obtained by the fitting follows the shape of the data smoothly and closely, so the method of approximate fitting based on cubic B-spline can better reflect the functional correlation of the original variables. The specific process is as follows:

$$C(t) = \sum_{i=0}^{h} N_{i,p}(t) P_i \tag{10}$$

where $P_0, P_1, ..., P_h$ represent $h+1$ numbers of unknown control points. Since the spline will pass the first and last data points, we have $D_0 = C(0) = P_0$ and $D_n = C(1) = P_h$, so there are only $h-1$ numbers of unknown control points. Taking this into consideration, the curve equation becomes the following:

$$C(t) = N_{0,p}(t) P_i D_0 + (\sum_{i=1}^{h-1} N_{i,p}(t) P_i) + N_{h,p}(t) D_n \tag{11}$$

Introduce the approximation technique, then the sum of all squared error distances between

original value and the spline fitting value is as follows:

$$f(P_1,P_2,...,P_{h-1})=\sum_{k=1}^{n-1}|D_k-C(t_k)|^2 \tag{12}$$

Our goal is to find those control points $P_1,P_2,...,P_{h-1}$ such that the function $f(.)$ is minimized.

Algorithm 1: the RMSE of variables with defined function or undefined function relationship

Input: continue random variables X and Y
$Y=CX+\sigma$ (the σ represents the random error with specific distribution)
$Y_1=CX$
Output: the normalized RMSE
if there is a clear functional relationship between X and Y:
$n=len(X)$
standardize and normalize the random variables X, Y and Y_1
obtain the variables E_0, E_1 and E_2
for i in range n do:

$$normal\_RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(E_1-E_2)^2}$$

Else do:
Applying the approximate fitting based on cubic B-spline
The function expression is obtained:

$$C(t_k)=\sum_{k=1}^{n-1}N_{i,p}(t_k)P_i$$

Normalized $C(t_k)$ and obtain E_3
for i in range n do:

$$normal\_RMSE^*=\sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}(E_1-E_3)^2}$$

3.3. The degree of uncertainty for the distribution of variables

The more cluttered the data distribution, the higher the uncertainty of the information. Therefore, there exist uncertainty difference between the functional relationship variables after regression fitting and the original variables, the uncertainty of variables can adversely affect correlation results, and the degree of uncertainty of variables has an important impact on the correlation analysis of variables. The information entropy can be applied to measure the uncertainty of information, and the degree of uncertainty in the distribution of variables is evaluated by calculating the difference of information entropy between the functional relationship variables and the original variables. The information entropy is calculated by the Eq (3), and the difference entropy (ED) between functional relationship variables and original variables is calculated by

$$ED=H_0-H_1 \tag{13}$$

Obviously, the unit of ED is bit, the lower the degree of uncertainty comes the less ED.

The correlation analysis framework is specific to continuous bivariate, and the information entropy is mainly for discrete data. Therefore, the partition idea of MIC is referenced, the partition way corresponding to the maximum mutual information is used as the way of discretization process of the original variables, the probability of data points between variables after discretization is expressed as p, and the probability of each part is calculated as the number of data points of the corresponding part divided by the total number of data points. The calculation formula is given in Eq (14), and Figure 4 shows an example of dividing data points between variables, the number of red lines represents the number of dividing intervals. Then, the information entropy of each division is calculated, and in order to facilitate the comparison of the results, the information entropy is normalized so that the value of each information entropy is between 0 and 1. Finally, the normalized information entropy corresponding to the best division can be obtained.

$$p_i(x_1, x_2) = \frac{m_{(part)}}{m_{total}} \tag{14}$$

where m_{part} represents the number of data points in a partition, and m_{total} represents the total number of data points in the variable.

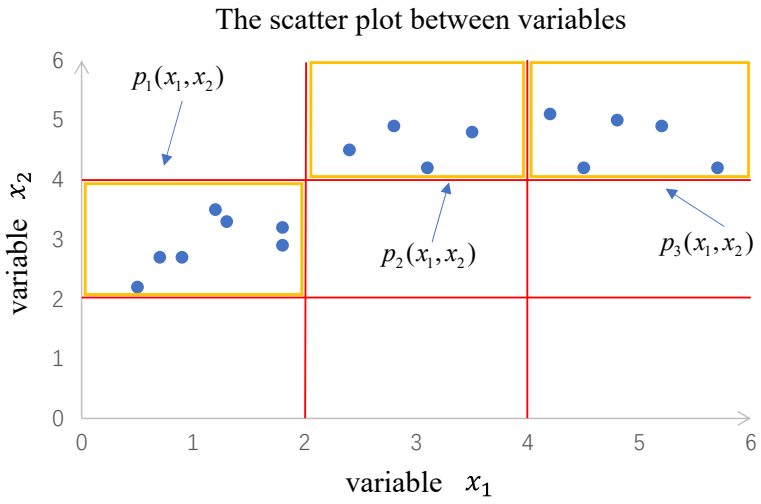

Figure 4. The probability of each part corresponding to the best partition between variables.

3.4. The degree of dependence on the random distribution of variables

The values of variables studied in this paper are affected by random errors and distributions [36], so that there exist uncertainty and randomness in variables, but the probability that these values fall into a certain range is definite. When studying the correlation of variables that are subject to distributions, it is necessary to consider the distribution state of the variables. The copula theory is applied to analyze the correlation of variables affected by distributions.

According to Copula theory, if the distribution status of variables is inconsistent, it is unreasonable to analyze the correlation between variables. Therefore, the random variables are first transformed into probability distribution functions that conform to the same distribution before

constructing the correlation structure of random variables, so that the values of random variables are between 0 and 1. Then, the function construction is performed on the transformed data by the copula function, which represents the correlation structure of the random variables. Additionally, all the dependency relationships are contained after constructing the functional relationship, which is the theoretical copula of random variables. The empirical copula is the bridge between the theoretical copula and the actual data. The RMSE between the empirical copula and the theoretical copula, AIC and BIC are used as indicators to evaluate the copula model. Finally, the correlation parameters corresponding to the optimal copula function are selected to measure the dependence on random variables with distributions.

Correlated parameter estimation includes maximum likelihood estimation method and non-parametric estimation method [37]. In this paper, the parameters of the function are estimated by non-parametric estimation method, i.e., the parameters are estimated by the relationship between the correlated parameters and the Kendell's coefficients, and the range of parameters corresponding to the Archimedean copula is shown in Table 1.

Table 1. Introduction to the Archimedean copula function.

Function name	Function expressions	The range of parameters	The relationship between τ and θ
Clayton copula	$C(u,v)=(u^{-\theta}+v^{-\theta}-1)^{-1/\theta}$	$\theta\in[-1,\infty),\theta\neq 0$	$\tau=\frac{\theta}{2+\theta}$
Gumbel copula	$C(u,v)=\exp(-[(-\log u)^{1/\theta}+(-\log v)^{1/\theta}]^{\theta})$	$\theta\in(0,1]$	$\tau=1-\theta$
Frank copula	$C(u,v)=-\frac{1}{\theta}\log\left(1+\frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right)$	$\theta\in(-\infty,\infty),\theta\neq 0$	$\tau=1+\frac{4}{\theta}[\frac{1}{\theta}\int_0^1\frac{t}{\exp(t)-1}dt$

3.5. The measure of correlation between variables

The relationship between variables is divided into deterministic relationships and non-deterministic relationships. Deterministic relationships refer to the existence of clear functional relationship between variables, that is, there exists accurate function expression, and the non-deterministic relationship refers to the correlation relationship without clear functional relationship. In many measure methods of correlation, when the correlation value is equal to 1, it is considered that there exists complete functional relationship between variables; when the correlation value is between 0 and 1, there exist correlated relationship between variables; and when the correlation value is equal to 0, the variables are independent. However, many traditional methods do not fully consider the impact of data uncertainty and distribution, which leads to ignoring the strong regularity information in the variables, so that the correlation coefficient value obtained by the variables with strong regularity is too small. Then, it is considered that there exist no correlation or weak correlation between the variables.

Based on the above, in this section, the correlation of the functional relationship variables containing specific distributions and the variables without explicit functional relationships will be analyzed, respectively. By measuring the degree of functional relationship between variables, the

degree of uncertainty in the distribution of variables and the dependence on the random distribution of variables, the correlation between the two relationship types is comprehensively analyzed. Additionally, the process of correlation analysis and its rationality are described in detail.

3.5.1. The measure of correlation for the functional relationship variables with specific distributions

1) The degree of functional relationships between variables

For the functional relationship variables with specific distribution, there exist functional relationships between variables that are subject to some error distributions. The degree of functional relationship between variables is evaluated by calculating the RMSE of random distribution of the functional relationship variables with specific distribution. The smaller the error, the closer the association between variables, and it indicates that the functional relationship between variables is stronger. The larger the error, the sparser the degree of functional relationship between the variables.

2) The degree of uncertainty for the distribution of variables

Algorithm 2: the entropy difference of the variables X and Y

Input: continue random variables X and Y
$Y = CX + \sigma$ (the σ represents the random error)
$Y_1 = CX$
Output: the entropy difference ED
For i in range (2, len(X)):
  For j in range (2, len(Y)):
    If $i * j \le \lfloor len(X) **0.6 \rfloor$:
      output the number of divided intervals (i, j)
      $a = cal\_mic(i, j)$
      $b = cal\_entropy(i, j)$
      mics.append(a)
      entropy.append(b)
      if $a \ge \max(mics)$:
        output the optimal divided interval (i, j)
normalize the matrix. b .
output the normalized entropy H_0 corresponding to the optimal divided interval (i, j)
similarly, calculate the normalized entropy H_1 corresponding to the optimal divided interval between X and Y_1 $ED = | H_0 - H_1 |$

For the functional relationship variables containing specific distribution, the uncertainty of variables is brought by the distribution of variables. The uncertainty of information is measured by calculating the information entropy in this paper. First, the best partition corresponding to the MIC is calculated for the variables with functional relationships, then enumerating all partition results and calculating the information entropy corresponding to each partition. Finally, the information entropy is normalized to obtain the normalized information entropy corresponding to the best partition. Similarly, the above process was repeated for the functional relationship data containing specific distribution, and the difference of information entropy between the functional relationship variables

and the functional relationship variables with distributions is obtained. The uncertainty degree of the functional relationship variable containing specific distribution is evaluated by the difference of normalized information entropy. The greater the difference in the information entropy of the variables, the higher the degree of uncertainty of the variables. Algorithm 2 shows the pseudo-code for computing the entropy difference of variables.

3) The degree of dependence on the random distribution of variables

Algorithm 3: measure the correlation of variables X and Y with distribution and uncertainty

Input: continue random variables X and Y

$Y = CX + \sigma$ (the σ represents the random error)

$Y_1 = CX$

Output: the correlation parameter θ

$u = stats.norm(E_0)$, $v = stats.norm(E_1)$

calculate the Kendell coefficient τ

for i in range (len(u)):

$\theta = 1/(1-\tau)$

$gumbel\_func\_copula = \exp(-[(-\log u)^{1/\theta} + (-\log v)^{1/\theta}])^{\theta}$

Gumbel_copula.append(Gumbel_func_copula)

for i in range (len(u)):

$\theta = 2\tau/(1-\tau)$

$clayton\_func\_copula = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}$

Clayton_copula.append(Clayton_func_copula)

for i in range (len(u)):

$$\tau = 1 + \frac{4}{\theta}[\frac{1}{\theta}\int_0^1 \frac{t}{\exp(t)-1}dt - 1]$$

$$frank\_func\_copula = -\frac{1}{\theta}\log\left(1 + \frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right)$$

Frank_copula.append(Frank_copula)

for i in range (len(u)):

for j in range(len(v)):

if $x[j] \le x[i]$ and $y[j] \le y[i]$:

count=count+1

empirical_copula.append(count/len(x)):

for i in range (len(theory_copula)):

error+=(theory_copula-emirical_copula)**2

error=np.sqrt(error/len(theory_copula))

error1=error**2

$AIC = m\ln MSE + 2k$

$BIC = m\ln MSE + 2k\ln m$

The function type of Archimedean copula and the construction process of copula function are introduced in Section 3.4. The correlation parameters corresponding to the optimal copula function are used to measure the degree of dependence on variables that subject to distributions. The function

expressions of bivariate copulas and the range of parameters corresponding to copula function are shown in Table 1. For the Clayton copula, the range of parameter is $(0,\infty)$, the larger the absolute value of parameter, the greater the correlation of variables containing distributions. Then, when the parameter $|\theta|<1$, the degree of dependence on variables containing distributions is weak. For Gumbel copula, the range of the parameter is $(0,1]$, and the closer the correlation parameter is to 0, the more correlated the variables. For Frank copula, the range of the parameter is $(-\infty,\infty)$, and the larger the absolute value of parameter, the greater the correlation of variables containing distributions. Then, when the parameter $|\theta|<1$, the degree of dependence on variables containing distributions is weak. In order to facilitate the comparison, the reciprocal of the correlation parameters of Clayton copula and Frank copula functions is taken, so that the directions of change of parameter value size and the strength of the correlation is consistent with the Gumbel copula. In addition, in order to compare the correlation values obtained by other methods, and measure the correlation strength of variables more intuitively through the correlation value, and if the absolute value of the parameter that is obtained by constructing the Clayton copula and Frank copula functions of variables is less than 1, it is considered that the correlation between variables is extremely weak. Thus, the inverse of parameter $|\theta|<1$ is the value of infinity, so the correlation coefficient value R obtained by $1-w*[RMSE,ED,|\theta|]$ is negative. The parameter $|\theta|\geq 1$ corresponding to correlation coefficient values obtained by the new method are between 0 and 1. The closer the correlation value is to 1, the stronger the correlation between variables. Algorithm 3 shows the pseudo-code for measure the correlation of variables with distribution and uncertainty.

3.5.2. The measure of correlation for the variables without explicit functional relationships

1) The degree of functional relationship between variables

Algorithm 4: the approximate fitting based on cubic B-spline

Input: $n+1$ data points $D_0,D_1,...,D_n$, degree=3,
The number of control points $h+1$
Output: A B-spline curve of degree 3
Obtain a set of parameters $t_0,t_1,...,t_n$ and a knot vector U
Let $P_0=D_0$ and $P_h=D_n$;
For $k=1$ to $n-1$ do:
  Compute Q_k : $Q_k=D_k-N_{0,p}(t_k)D_0-N_{h,p}(T_k)D_0$
For $i=1$ to $h-1$ do:
  Compute the following and save it to the $i-th$ row of matrix Q
  $\sum_{k=1}^{n-1}N_{i,p}(t_k)Q_k$
For $k=1$ to $n-1$ do:
  For $i=1$ to $h-1$ do:
   Compute $N_{i,p}(t_k)$ and save to row k and column of i of N;
Compute $M=N^TN$;
Solving for P from $M\bullet P=Q$;
Row i of P is control point P_i;
Control points $P_0,P_1,...,P_h$ knot vector U and degree determines an approximation B-spline curve;

In the actual variables, the functional relationship between variables is not clear, and it is necessary to apply the regression model to obtain the function expression, which can approximately reflect the closeness of the association and the general regulation of changes between variables. The error between the original data and the fitted data will be used to measure the degree of functional relationship between the original variables. In this paper, the approximation fitting based on cubic B-spline is applied to obtain functional relationship expressions between the variables, and then the spline error between the original data and the fitted data is calculated. Due to the idea of minimizing the error distance was adopted by the cubic B-spline. Then, when there exist one-to-many situations in the data or many outliers, the spline function obtained by the spline fitting is complex and does not conform to the shape distribution of the original data. Therefore, before the spline fitting, the data need to be standardized and normalized, and the spectral clustering is performed on the data in the case of one-to-many or many-to-many first. Then, the spline fitting is performed on each category data. Finally, the spline error is the sum of the RMSE after fitting each category data. Algorithm 4 shows the pseudo-code for approximate fitting based on the cubic B-spline.

2) The degree of uncertainty for the distribution of variables

The degree of uncertainty in variables is evaluated by the entropy difference, and the calculated method is as follows: First, the functional expression between variables is obtained by the B-spline approximate fitting, and the MIC between the variables is calculated to obtain the best partition of the data; then the information entropy corresponding to different partitions is calculated, and the information entropy is normalized; finally, the normalized information entropy of the variables after spline fitting corresponding to the best partition is obtained, and the normalized information entropy of the original variables is also derived. The degree of uncertainty for the original data is measured by the difference between the normalized information entropy of the fitting variables and the original variables. The larger the difference, the higher the degree of uncertainty. Similarly, the smaller the difference, the smaller the degree of uncertainty.

3) The degree of dependence on the random distribution of variables

Mentioned in Section 3.4, the correlation of variables with distributions and uncertainty is evaluated by estimating the correlation parameters of the copula function. First, the function of Clayton copula, Gumbel copula and Frank copula in the Archimedes copulas for variables are constructed, and the correlation parameters of the three functions are estimated. Then, the optimal copula function is selected by calculating the fitting error between the empirical copula and theoretical copula, AIC and BIC to represent all dependence structures between variables. Finally, the correlation parameters corresponding to the optimal copula are selected as the measure indicator of correlation for variables containing distributions.

3.6. The weighted coefficient

Section 3.5 described how to measure the correlation of variables by evaluating the degree of functional relationship, the degree of uncertainty in variables and the degree of dependence on the variables containing distributions. Then, the above three indicators will be integrated into a single

correlation value R to evaluate the correlation between variables more intuitively. The weight coefficients of the three indicators will be determined by the maximum eigenvalue method in the AHP method [38], and the relative importance matrices for measure indicators are constructed as shown in Table 2. In addition, the constructed importance matrices all pass the consistency test, and the final weight matrix is [0.346, 0.200, 0.454]. The smaller the RMSE, entropy difference and correlated parameter, the closer the association between variables, and the values of three indictors are between 0 and 1, so that the weighted results after integration are transformed to $R = 1 - w * [RMSE, ED, |\theta|]$, the larger the correlation value R, the more correlated the variables. Where w represents the weight matrix, the $RMSE$ is the root mean square error between the original variable and the variable after constructing a functional relationship, ED represents the information entropy of variables, and $|\theta|$ represents the absolute values of correlated parameters corresponding to the optimal copula function. Algorithm 5 shows the pseudo-code for AHP method.

Table 2. The relative importance matrix for evaluation indicators.

A_i / A_j	A_1(RMSE)	A_2(Entropy difference)	A_3(The correlated parameter)
A_1(RMSE)	1	2	1/1.5
A_2((Entropy difference)	1/2	1	1/2
A_3(The correlated parameter)	1.5	2	1

Algorithm 5: AHP method

Input: the RMSE, the entropy difference ED , and the correlation parameter θ

RI =[0, 0, 0.52, 0.89, 1.12, 1.26, 1.36, 1.41,

1.46, 1.49, 1.52, 1.54, 1.56, 1.58, 1.59]

Output: the value of the correlation coefficient *R*

Construct the importance matrix M of the indicators

eig_val, eig_vector = np.linalg.eig(array *M*)

max_eig_val = np.max(eig_val)

max_eig_vector = eig_vector [:, np.argmax(self.eig_val)]

max_eig_val = $\lambda_{\max}$

max_eig_vector = w

$$CI = \frac{\lambda_{\max} - n}{n-1}, CR = \frac{CI}{RI}$$

If $CI < 0.1$: the matrix M pass the conformance test

Else: the matrix M failed the consistency test

$R = 1 - w * [RMSE, ED, |\theta|]$

4. The experiment

4.1. The data description

In order to verify the effectiveness of the method, two types of data were selected for experiments. The first one is the functional relationship variable that contains a specific distribution. Table 3 lists 12 function expressions that contain a specific distribution, where represents the random error that conform to uniform distribution. The expression types include general expressions and parameter expressions. The variable range and distribution data value range are shown in Table 3. The second is to use 4 UCI datasets and 4 artificially synthesized datasets with no explicit functional relationship. The UCI datasets include Iris Data Set, seeds Data Set, Glass Identification Data Set and Wine Data Set. The synthetic datasets include Two_cluster, Twomoons, Five_cluster and Roll. The detailed information of datasets is shown in Table 4.

4.1.1. Experimental data types that possess specific distribution and functional relationship variables simultaneously

Table 3. The functional relationship expression that implies specific distribution.

Function expressions	Range of independent variables	Range of values for random distribution
$y = \lvert x \rvert + \delta$	(-1,1)	(-1,1)
$y = 4(x^2 - 0.5)^2 + \delta$	(-1,1)	(-1,1)
$y = 2x^2 + \delta$	(-1,1)	(-1,1)
$y = 2(x^3 - \delta)^4$	(-1,1)	(-1,1)
$y = \delta x^5$	(-1,1)	(-1,1)
$y = \pm(x^2 + \delta)$	(-1,1)	(0,0.5)
$\begin{cases} x = \sin \pi t + \delta \\ y = \cos \pi t + \delta \end{cases}$	(-1,1)	(0,1/8)
$\begin{cases} x = 2\sin t - \sin 2t + \delta \\ y = 2\cos t - \cos 2t + \delta \end{cases}$	(-5,5)	(0,0.5)
$\begin{cases} x = t\sin(\pi t) + \delta \\ y = t\cos(\pi t) + \delta \end{cases}$	(-5,5)	(0,0.5)
$\begin{cases} x = 2\sin(5t)\cos t + \delta \\ y = 2\sin(5t)\sin t + \delta \end{cases}$	(-5,5)	(0,0.5)
$\begin{cases} x = 2(\cos(360t))^3 + \delta \\ y = 2(\sin(360t))^3 + \delta \end{cases}$	(-5,5)	(0,0.5)
$\begin{cases} x = -9\sin(2t) - 5\sin(3t) + \delta \\ y = 9\cos(2t) - 5\cos(3t) + \delta \end{cases}$	$(0,2\pi)$	(0,2)

4.1.2. Datasets with no explicit functional relationship variables

Table 4. Information description of UCI datasets and synthetic datasets.

Index	Datasets	Number of attributes	Number of samples
1	iris	4	150
2	seeds	7	210
3	glass	10	214
4	wine	13	178
5	Two_cluster	3	400
6	Twomoons	3	1502
7	Five_cluster	3	2000
8	Roll	3	2000

4.2. Correlation analysis

4.2.1. Correlation analysis of functional relationship variables containing specific distribution

The first step to carry out the correlation analysis of variables is to draw the scatter plots of the variables to observe the trend of the variable data. Then, the degree of functional relationship between variables is evaluated by the random distribution error of the variables, the entropy difference of the variables with random distribution is calculated to measure the degree of uncertainty of the variables and the copula function models are constructed to measure the degree of dependence on the variables containing distributions. Finally, the strength of correlation of the variables is evaluated integrally.

1) The scatter plots of the variables

For the functional relationship variables containing specific distributions, the scatter plots of the variables are shown in Figure 5, in which the MIC and Pearson coefficient values between the variables are also displayed. It can be seen that all scatter plots have strong regularity from the figure, but the MIC value is relatively small and the Pearson coefficient value is almost equal to 0, which indicates that MIC can measure only the correlation with a specific type of functional relationship but cannot identify the type of variable with distribution, and the Pearson coefficient can evaluate only the type of linear function relationship between variables but the rest of the relationship types cannot be well measured.

2) The degree of functional relationships of the variables

For variables with explicit functional relationship involving specific distributions, the degree of functional relationships that is evaluated by the error brought by the specific distribution. Their distribution errors are expressed by the root mean square error (RMSE) of random distribution of the functional relationship variables with specific distribution. The smaller the error, the higher the degree of functional relationship of the variables. The larger the error, the smaller the degree of functional relationship of the variables. The specific results are shown in the second column of Table 5. Owing to there exists different degree of random distribution between variables, the scatter plots and the results of normalized RMSE for variables are also different, and the greater the degree of random distribution

between variables, the larger the value of normalized RMSE for variables, which indicates that the random distribution between variables will affect the degree of functional relationships between variables.

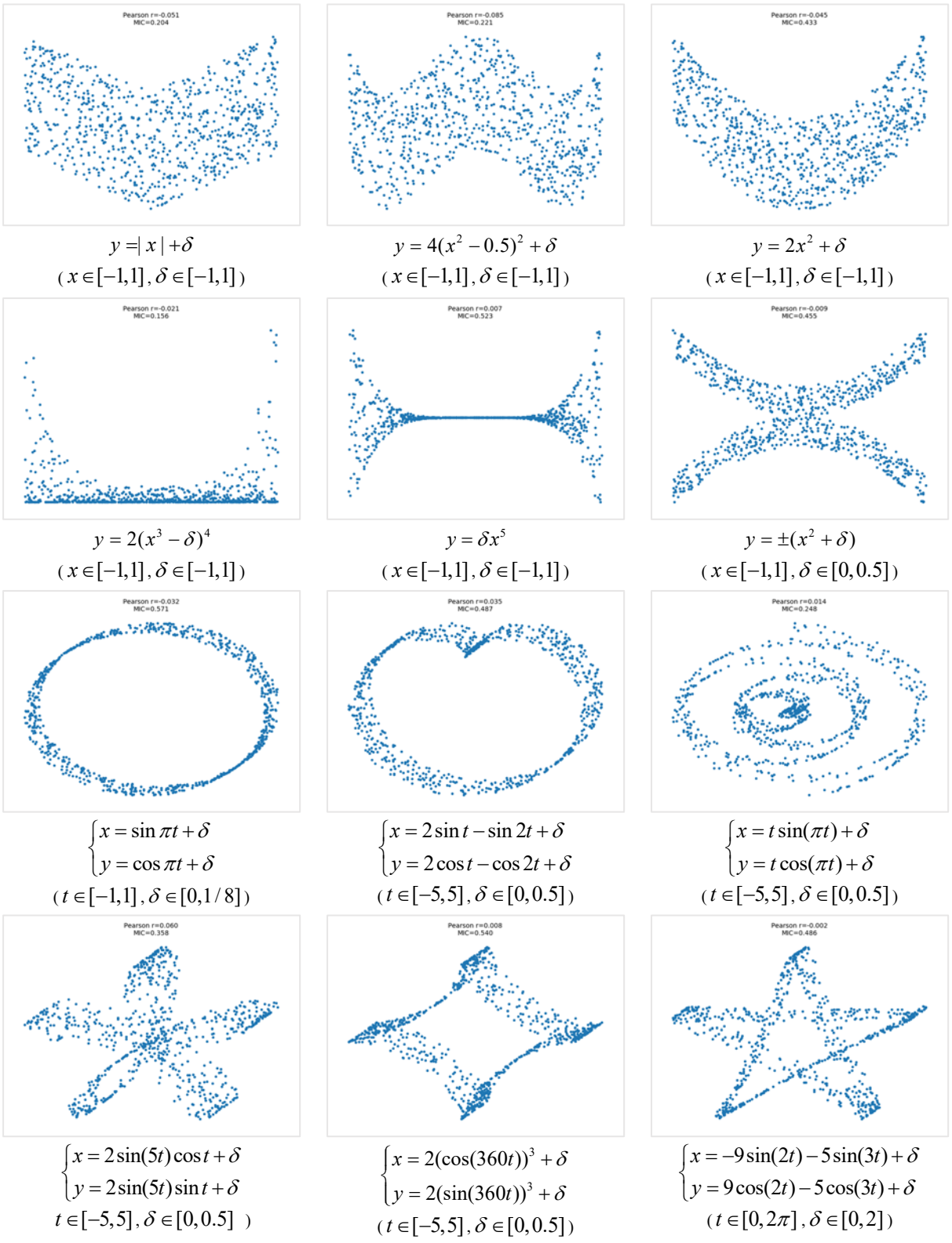

Figure 5. The scatter plots of functional relationship variables that contain specific distribution.

Table 5. The experimental results of the normal RMSE, normal entropy difference (ED), absolute value of correlated parameters, unweighted correlation coefficient R and weighted correlation coefficient R for variables with specific distribution and explicit functional relationship.

Function expressions	Normal_RMSE	Normal ED	Absolute value of correlated parameters	Unweighted R	Weighted R
$y = \lvert x \rvert + \delta$	0.272	0.255	0.967	0.502	0.417
$y = 4(x^2 - 0.5)^2 + \delta$	0.306	0.201	0.944	0.516	0.425
$y = 2x^2 + \delta$	0.227	0.137	0.980	0.552	0.449
$y = 2(x^3 - \delta)^4$	0.177	0.044	1.008	0.590	0.454
$y = \delta x^5$	0.197	0.020	0.997	0.595	0.481
$y = \pm(x^2 + \delta)$	0.306	0.139	1.006	0.506	0.397
$\begin{cases} x = \sin \pi t + \delta \\ y = \cos \pi t + \delta \end{cases}$	0.232	0.050	0.484	0.745	0.690
$\begin{cases} x = 2\sin t - \sin 2t + \delta \\ y = 2\cos t - \cos 2t + \delta \end{cases}$	0.042	0.015	1.038	0.635	0.389
$\begin{cases} x = t\sin(\pi t) + \delta \\ y = t\cos(\pi t) + \delta \end{cases}$	0.019	0.018	1.026	0.646	0.524
$\begin{cases} x = 2\sin(5t)\cos t + \delta \\ y = 2\sin(5t)\sin t + \delta \end{cases}$	0.045	0.041	1.067	0.616	0.492
$\begin{cases} x = 2(\cos(360t))^3 + \delta \\ y = 2(\sin(360t))^3 + \delta \end{cases}$	0.044	0.025	1.043	0.629	0.506
$\begin{cases} x = -9\sin(2t) - 5\sin(3t) + \delta \\ y = 9\cos(2t) - 5\cos(3t) + \delta \end{cases}$	0.260	0.011	1.007	0.574	0.440

3) Uncertainty measures for distributions between variables

The uncertainty of functional relationship variables containing specific distributions is measured by comparing the information entropy difference brought by variable distribution. First, calculate the MIC value of the specific functional relationship type between the variables, then exhaustively list all the divisions and calculate the information entropy under each division and normalize it. Finally, the normalized information entropy corresponding to the best division is acquired. Furthermore, calculating the normalized information entropy of the optimal division of variables with specific distribution and the difference between the two information entropies is obtained. The experimental results are shown in Table 6, where Normal ED represents the difference of information entropy between functional relationship variables and the variables with specific distributions, Normal Ent represent the normal information entropy of variables. The results show that the distribution of the data will bring a certain degree of information entropy difference.

Table 6. The experimental results of the optimal partition for the data points of variables (Optimal partition), the normal entropy of variables without distribution, the normal entropy (Normal Ent) and the normal entropy difference (Normal ED) between variables with specific distribution and explicit functional relationship.

Function expressions	Optimal partition	Normal Ent of variables without distribution	Normal Ent	Normal ED
$y = \| x \| + \delta$	(24,2)	0.700	0.955	0.255
$y = 4(x^2 - 0.5)^2 + \delta$	(26,2)	0.785	0.986	0.201
$y = 2x^2 + \delta$	(21,2)	0.720	0.852	0.137
$y = 2(x^3 - \delta)^4$	(3,18)	0.413	0.457	0.044
$y = \delta x^5$	(19,2)	0.844	0.864	0.020
$y = \pm(x^2 + \delta)$	(14,3)	0.623	0.762	0.139
$\begin{cases} x = \sin \pi t + \delta \\ y = \cos \pi t + \delta \end{cases}$	(10,5)	0.680	0.730	0.050
$\begin{cases} x = 2\sin t - \sin 2t + \delta \\ y = 2\cos t - \cos 2t + \delta \end{cases}$	(12,4)	0.750	0.765	0.015
$\begin{cases} x = t\sin(\pi t) + \delta \\ y = t\cos(\pi t) + \delta \end{cases}$	(17,3)	0.835	0.853	0.018
$\begin{cases} x = 2\sin(5t)\cos t + \delta \\ y = 2\sin(5t)\sin t + \delta \end{cases}$	(12,4)	0.728	0.769	0.041
$\begin{cases} x = 2(\cos(360t))^3 + \delta \\ y = 2(\sin(360t))^3 + \delta \end{cases}$	(3,16)	0.671	0.696	0.025
$\begin{cases} x = -9\sin(2t) - 5\sin(3t) + \delta \\ y = 9\cos(2t) - 5\cos(3t) + \delta \end{cases}$	(16,3)	0.808	0.819	0.011

4) Dependency measures for random distribution of variables

In the experiment of measuring the dependency degree of random distribution of variables, the optimal copula function is selected to construct the correlation structure of variables, and the relevant parameters of the optimal copula are used as the correlation evaluation index of random distribution. The binary copula functions include Clayton copula, Gumbel copula and Frank copula. The fitting errors, AIC and BIC of empirical and theoretical copulas obtained after constructing the function, are used as the index for selecting the optimal copula function. Appendix to Table A1 shows the results of the best copula function, and the best results are shown in bold.

The final results of correlation analysis are shown in Table 5. The correlation of variables is comprehensively evaluated through the degree of functional relationship, degree of uncertainty and degree of random distribution dependence of the variables. Then, the weight coefficients of the three indicators are calculated and summed up. Since the smaller the three indicators, the stronger the correlation, and the value range is between [0, 1], so the weighted sum is subtracted from 1. Finally, the weighted correlation coefficient value is obtained. The larger the value, the stronger the correlation. Likewise, unweighted correlation results can be drawn.

4.2.2. Correlation analysis of variables without explicit functional relationship

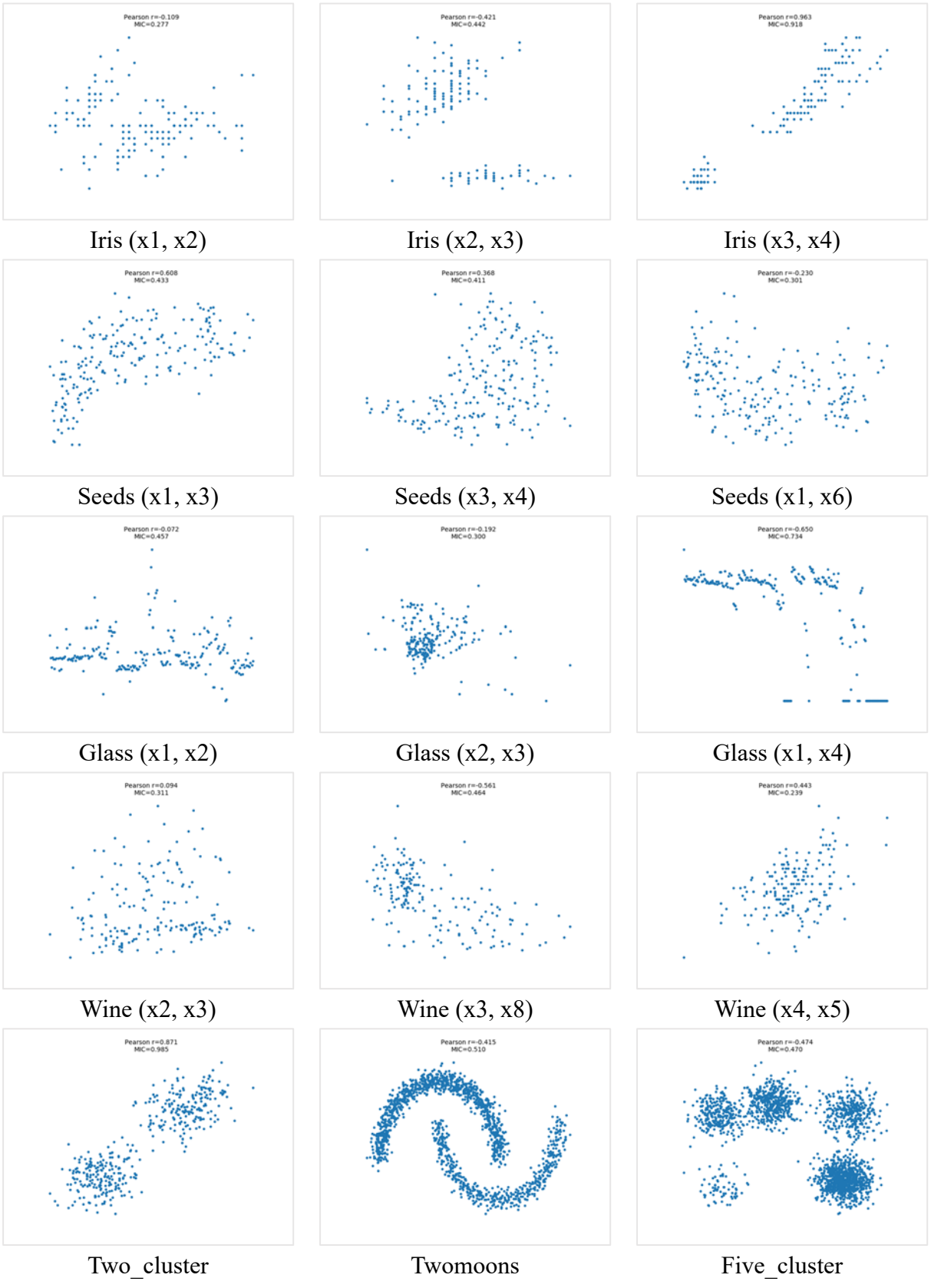

Figure 6. Scatter plot of variables in UCI datasets and synthetic datasets.

To carry out the correlation analysis of variables, the first step is to draw the scatter plots of the variables to observe the approximate distribution and regularity of the data. Then, the degree of functional relationships between variables is evaluated by the spline error, the entropy difference of the variables is calculated to measure the degree of uncertainty of the data, the copula function model is constructed to measure the degree of dependence of the random distribution of the variables, and the correlation size of the variables is evaluated integrally.

1) The scatter plots of the variables

The scatter plots of variables without explicit functional relationships in the UCI datasets and the artificially synthetic datasets are shown in Figure 6, which also shows the MIC and Pearson coefficient values between partial variables in each dataset. Combining the scatter plots of variables, MIC and Pearson coefficient values, it can be analyzed that Pearson coefficient can well evaluate their linear correlation when there is an obvious monotonic linear correlation shown in the scatter plots, and MIC can do better when the scatter plots show a perfect function image of specific function. However, for the functional relationship variables containing specific distributions, neither can perform well, and the obtained MIC and Pearson coefficient values are small.

2) The degree of functional relationship of the variables

Table 7. The results of the optimal partition for the data points of variables (Optimal partition), the normalized RMSE of variable after approximate fitting based on cubic B-spline (Normal_RMSE of spline fitting) and the normalized entropy of variables after approximate fitting based on cubic B-spline (Normal Ent of spline functional variables) for variables with no explicit functional relationship.

Datasets	Optimal partition	Normal_RMSE of spline fitting	Normal Ent of spline functional variables
Iris(x1,x2)	(9,2)	0.120	0.775
Iris(x2,x3)	(8,2)	0.385	1.000
Iris(x3,x4)	(3,4)	0.255	0.427
Seeds(x1,x3)	(8,2)	0.175	0.707
Seeds(x3,x4)	(10,2)	0.245	0.719
Seeds(x1,x6)	(10,2)	0.187	0.712
Glass(x1,x2)	(7,2)	0.135	0.517
Glass(x2,x3)	(8,2)	0.159	0.571
Glass(x1,x4)	(2,10)	0.221	0.684
Wine(x2,x3)	(7,2)	0.218	0.509
Wine(x3,x8)	(9,2)	0.190	0.619
Wine(x4,x5)	(3,7)	0.176	0.267
Two_cluster	(2,18)	0.203	0.854
Twomoons	(20,4)	0.085	0.805
Five_cluster	(2,44)	0.369	0.689
Roll	(32,3)	0.275	0.707

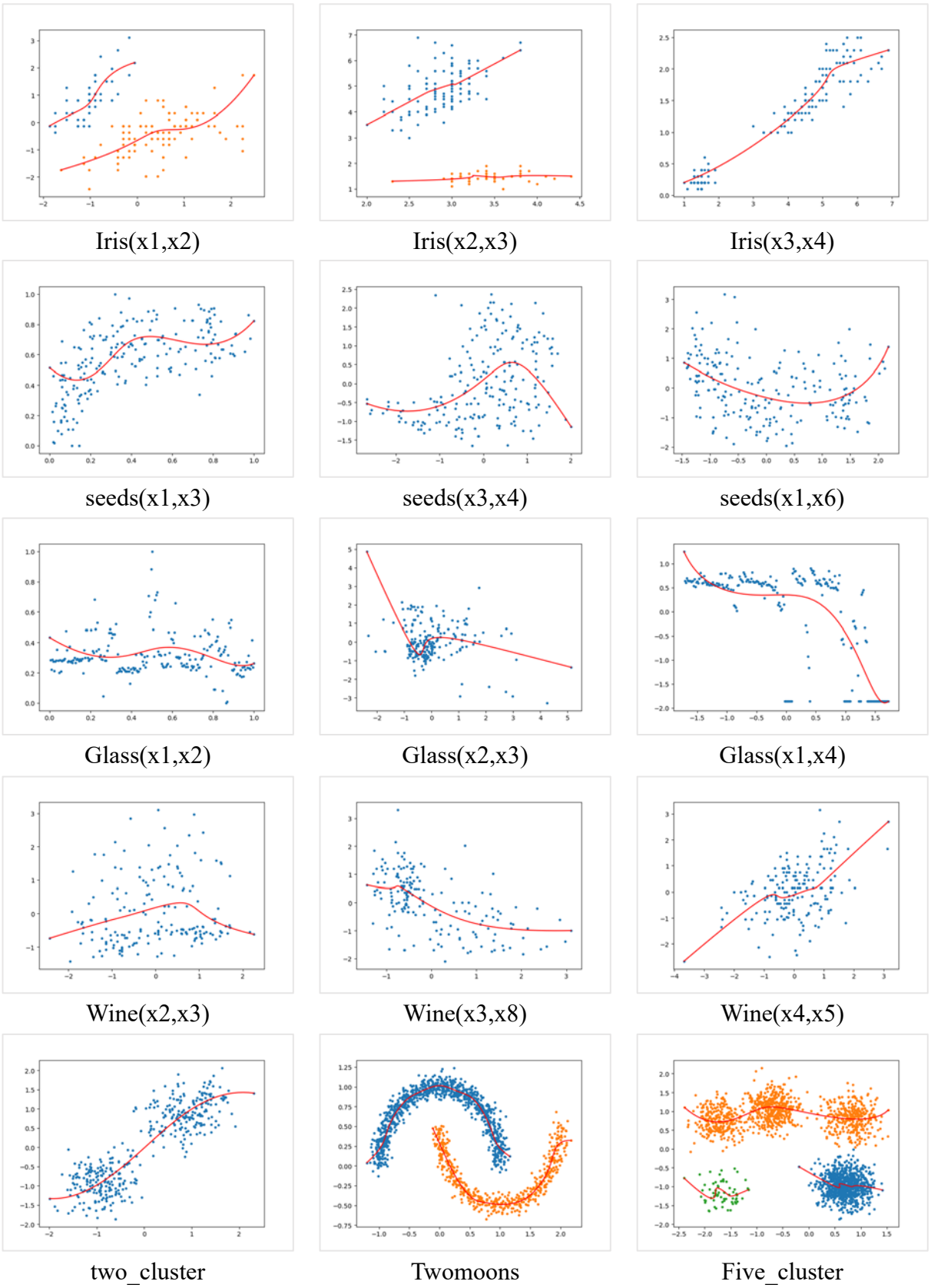

Figure 7. The results of the approximate fitting based on cubic B-spline and spectral clustering.

For variables without a clear functional relationship, the degree of functional relationship between the variables was evaluated by the error between the true value and the spline fitting value. The cubic B-spline approximate fitting was performed on the value corresponding to each category. The number of cubic B-spline control points used in this experiment was 4 and the number of nodes was 3. The mean squared errors of the true and fitted values in each category were calculated and summed up. The spline fitting results are shown in Figure 7, and the error are shown in Table 7. Combining the figures and the results in the table, it indicates that the approximate fitting based on cubic B-spline can follow the trend of data points well, and for variables with stronger distribution regularity of data points, the better the fitting effect.

3) The degree of uncertainty for the distribution of variables

For the measurement of uncertainty in variables without explicit functional relationships, the MIC values among variables with spline functional relationships are calculated first, then all divisions are enumerated to derive the best division of variables, the information entropy under each division is calculated and normalized, and finally the normalized information entropy corresponding to the best division is obtained. Furthermore, the normalized information entropy of the original variables corresponding to the best division is also calculated. The degree of uncertainty of the variables without explicit functional relationships was measured by comparing the difference in information entropy between the original data and the spline fitted data, and the results are shown in Table 8. According to the results in the table, the stronger the randomness of distribution for data points of the variables in the datasets, the greater the information entropy obtained and the larger the entropy difference calculated.

Table 8. The results of maximum information coefficient (MIC), normal entropy (Normal Ent) and normal entropy difference (Normal ED) for variables without explicit functional relationships.

Datasets	MIC	Normal Ent	Normal ED
Iris(x1,x2)	0.277	0.901	0.126
Iris(x2,x3)	0.442	0.869	0.131
Iris(x3,x4)	0.918	0.523	0.096
Seeds(x1,x3)	0.433	0.758	0.051
Seeds(x3,x4)	0.411	0.876	0.157
Seeds(x1,x6)	0.301	0.894	0.182
Glass(x1,x2)	0.457	0.697	0.180
Glass(x2,x3)	0.300	0.800	0.229
Glass(x1,x4)	0.734	0.549	0.135
Wine(x2,x3)	0.311	0.730	0.221
Wine(x3,x8)	0.464	0.850	0.231
Wine(x4,x5)	0.239	0.737	0.470
Two_cluster	0.985	0.971	0.117
Twomoons	0.510	0.840	0.035
Five_cluster	0.470	0.969	0.280
Roll	0.396	0.829	0.122

4) The degree of dependence on the random distribution of variables

For variables without explicit functional relationships, the correlated parameter corresponding to the optimal copula function is selected to evaluate the correlation degree between variables subject to random distributions. The bivariate copula functions include Clayton copula, Gumbel copula and Frank copula, and the fitting errors of empirical copula and theoretical copula, AIC and BIC are derived after constructing the copula models of variables, and the three indicators are used as evaluation index for selecting the optimal copula function, and the correlation parameter corresponding to the optimal copula is used as an evaluation method for correlation evaluation of random variables. Appendix to Table A2 shows the results of the selection of optimal copula function, and the best results are shown in bold.

Table 9. The experimental results of the normal RMSE of variables after approximate fitting based on cubic B-spline (Normal_RMSE of spline fitting), normal entropy difference (ED), absolute value of correlated parameters, unweighted correlation coefficient R and weighted correlation coefficient R for variables without explicit functional relationship.

Datasets	Normal_ RMSE of spline fitting	Normal ED	Absolute value of correlated parameters	Unweight-ed R	Weighted R
Iris(x1,x2)	0.120	0.126	0.932	0.607	0.510
Iris(x2,x3)	0.385	0.131	0.593	0.631	0.572
Iris(x3,x4)	0.255	0.096	0.123	0.842	0.837
Seeds(x1,x3)	0.175	0.051	0.610	0.722	0.652
Seeds(x3,x4)	0.245	0.157	1.531	0.356	0.189
Seeds(x1,x6)	0.187	0.182	0.846	0.595	0.515
Glass(x1,x2)	0.135	0.180	0.949	0.579	0.487
Glass(x2,x3)	0.159	0.229	1.033	0.526	0.431
Glass(x1,x4)	0.221	0.135	0.223	0.807	0.796
Wine(x2,x3)	0.218	0.221	1.103	0.487	0.374
Wine(x3,x8)	0.190	0.231	0.505	0.692	0.659
Wine(x4,x5)	0.176	0.470	0.406	0.650	0.322
Two_cluster	0.203	0.117	0.136	0.849	0.845
Twomoons	0.085	0.035	0.419	0.821	0.774
Five_cluster	0.369	0.280	0.436	0.639	0.619
Roll	0.275	0.122	1.071	0.536	0.409

The correlation analysis results for the functional relationship variables without clear function are shown in Table 9. The correlation of variables without clear functional relationships is evaluated by the degree of functional relationship between variables, the degree of uncertainty of variables and the degree of dependence on the variables containing distributions. Then, the weight coefficient of the three indicators is calculated, and the weighted sum of the three indicators is performed. Owing to the smaller the three indicators, the stronger the correlation, and the value range is between 0 and 1, so the weighted correlation value is obtained by subtracting the value after weighted summation from 1. The

higher the value, the stronger the correlation. Similarly, the unweighted correlation results can be derived. According to the results, the stronger the regularity of the variable, the greater the value of the correlation coefficient.

In addition, in order to compare the performance of correlation analysis methods for the two types of functional relationship variables with specific distribution and the variables without explicit functional relationships, the experimental designs are conducted as follows: First, the correlation analysis performance of the two methods is compared under the same functional expression with specific distribution; second, the performance of the two correlation analysis methods is compared for the same manifold dataset without functional relationships. Combined with the above experiments, the results are analyzed in detail and the reasonable conclusions are drawn below.

Table 10. The comparison results of correlation analysis methods for the variables with specific distributions and variables without explicit functional relationships.

Datasets	The methods for variables with specific distributions		The methods for variables without explicit functional relationships	
	Unweighted R	Weighted R	Unweighted R	Weighted R
$y = 4(x^2 - 0.5)^2 + \delta$	0.516	0.425	**0.548**	0.422
$\begin{cases} x = t\sin(\pi t) + \delta \\ y = t\cos(\pi t) + \delta \end{cases}$	**0.646**	0.524	0.575	0.467
two_cluster	0.766	0.794	**0.849**	0.845
Roll	0.511	0.394	**0.536**	0.409

According to the experimental results, the performance of the two correlation analysis methods is similar in the case of the same functional relationship expression with specific distribution, and the correlation analysis method aiming at the functional relation variable with specific distribution is slightly better. For manifold datasets without explicit functional relationships, the correlation analysis method without functional relation variables has better performance. However, it is also valid for functional relation variables with specific distributions. According to the experimental results, the performance of the two correlation analysis methods is similar in the case of the same functional relation expression with specific distribution, and the correlation analysis method aiming at the functional relation variable with specific distribution is slightly better. For manifold data sets without explicit functional relations, the correlation analysis method without functional relation variables has better performance. However, it is also valid for functional relation variables with a specific distribution. Moreover, for variables without clear functional relationships, it considers that the correlation analysis method for such variables will fit the function expression between variables first, and then conduct comprehensive analysis to the variables. Therefore, the relationship between the two correlation analysis methods is only for different variable types, and the evaluation methods are actually consistent. When the function expression between variables is wrong, the evaluation method for variables without clear functional relationships can be used to fit the function relationship between variables to get the correct function expression, and then correlation analysis will carry out to get the final correlation coefficient.

5. Discussion

Table 11. The comparison results of a new method with the maximum information coefficient (MIC), Pearson coefficient and mutual information (MI).

Datasets	Unweighted R	Weighted R	MIC	Pearson	MI
$y=\|x\|+\delta$	**0.502**	0.417*	0.204	-0.051	0.203
$y=4(x^2-0.5)^2+\delta$	**0.516**	0.425*	0.221	0.000	0.219
$y=2x^2+\delta$	**0.552**	0.449*	0.433	-0.045	0.356
$y=2(x^3-\delta)^4$	**0.590**	0.454*	0.156	-0.021	0.081
$y=\delta x^5$	**0.595**	0.481	0.523*	0.007	0.517
$y=\pm(x^2+\delta)$	0.506	0.397	0.455*	-0.009	**0.549**
$\begin{cases}x=\sin\pi t+\delta\\y=\cos\pi t+\delta\end{cases}$	**0.745**	0.690*	0.571	-0.032	0.792
$\begin{cases}x=2\sin t-\sin 2t+\delta\\y=2\cos t-\cos 2t+\delta\end{cases}$	0.635*	0.389	0.487	0.035	**0.807**
$\begin{cases}x=t\sin(\pi t)+\delta\\y=t\cos(\pi t)+\delta\end{cases}$	**0.646**	0.524*	0.248	0.014	0.298
$\begin{cases}x=2\sin(5t)\cos t+\delta\\y=2\sin(5t)\sin t+\delta\end{cases}$	0.616*	0.492	0.358	0.060	**0.643**
$\begin{cases}x=2(\cos(360t))^3+\delta\\y=2(\sin(360t))^3+\delta\end{cases}$	0.629*	0.506	0.540	0.008	**0.871**
$\begin{cases}x=-9\sin(2t)-5\sin(3t)+\delta\\y=9\cos(2t)-5\cos(3t)+\delta\end{cases}$	**0.574**	0.440	0.486	-0.002	0.522*
Iris(x1,x2)	**0.607**	0.510*	0.277	-0.109	0.265
Iris(x2,x3)	**0.631**	0.572*	0.442	-0.421	0.328
Iris(x3,x4)	0.842	0.837	0.918*	**0.963**	0.855
Seeds(x1,x3)	**0.722**	0.652*	0.443	0.608	0.348
Seeds(x3,x4)	0.356	0.189	0.411	0.368*	0.213
Seeds(x1,x6)	**0.595**	0.515*	0.301	0.230	0.170
Glass(x1,x2)	**0.579**	0.487*	0.457	-0.072	0.232
Glass(x2,x3)	**0.526**	0.431*	0.300	-0.192	0.208
Glass(x1,x4)	**0.807**	0.796*	0.734	0.650	0.535
Wine(x2,x3)	**0.487**	0.374*	0.311	0.094	0.115
Wine(x3,x8)	**0.692**	0.659*	0.464	-0.561	0.331
Wine(x4,x5)	**0.650**	0.322	0.239	0.443*	0.215
Two_cluster	0.849	0.845	**0.985**	0.871	0.876*
Twomoons	**0.821**	0.774*	0.510	-0.415	0.771
Five_cluster	**0.639**	0.619*	0.470	-0.474	0.396
Roll	**0.536**	0.409	0.386	0.122	7.601

Note: the best results are shown in bold.

According to the algorithm theory in this paper, the value range of the correlation value calculated by the method in this paper is between 0 and 1, and the larger the value, the stronger the correlation. In order to verify the effectiveness of the proposed method, the weighted and unweighted correlation coefficients R obtained by the new method are compared with MIC, Pearson coefficient and Mutual Information (MI). The results of the comparison are shown in Table 11, with the best results are shown in bold, and the suboptimal results are marked with the * sign. In addition, for the new method, MIC and Pearson coefficients, the correlation coefficient values of them is equal to 0 when the variables are uncorrelated, while the correlation coefficient value is equal to 1 when the variables containing functional relationship with each other. For MI, the value of MI is close to 0 when the variables are uncorrelated with each other, while the value of MI is close to positive infinity when the variables with functional relationship.

1) According to the results in Table 10, in general, most of the correlation coefficients obtained by the new method are better than MIC and Pearson coefficients, and all are better than mutual information. For the functional relationship variables containing distributions, the MI value is close to 0, and the Pearson coefficient value is almost equal to 0, but there are clear regularities in the scatter plot of the variables, and the correlation coefficient value obtained by the new method and MIC is not equal to 0, which indicates that there exists correlation between the functional relationship variables containing specific distributions, but the MI and Pearson coefficient cannot measure their correlation. Therefore, according to the experimental results, Pearson coefficient can well evaluate the correlation of variables with linear relationship, MIC can better measure the correlation between variables with specific function types, and the new method can measure the functional relationship variables well. In addition, for variables without explicit functional relationship, the values obtained by the MIC and Pearson coefficient are small. However, there exist strong linear or nonlinear correlations between variables in the iris datasets, seeds datasets and the three synthetic datasets according the scatter plots which are shown in Figure 6. This indicates that MIC and the Pearson coefficient cannot evaluate the correlation of variables that subject to distributions and randomness well, but the new method can measure their correlations well

2) For the functional relationship variables containing specific distributions, according to the analysis results of the functional relationship variables that contain specific distribution, the new method has better evaluation ability for variables with obvious regularity in the scatter plots. According to the scatter plots shown in Figure 5, influenced by random error and distribution, the values of variables are non-deterministic, and there exist uncertainty and randomness in variables, but the probability that these values fall into a certain range is definite, and there exist strong regularity between variables. As shown in Table 11, there are 12 numbers of functional relationship variables that contain specific distributions in total, the unweighted results obtained by the new method are all better than MIC and Pearson coefficients, and the weighted correlation coefficient results are better than Pearson coefficients, with 8 numbers of weighted results that are better than MIC, indicating that the correlation analysis framework proposed in this paper can better evaluate the correlation of functional relationship variables that contain distributions.

3) For the experimental results of the variables without explicit functional relationships, there are 16 numbers of variable groups of datasets in total, there are 13 numbers of unweighted results obtained by the new method are better than MIC and Pearson coefficient, and for the weighted results, there are 13 numbers of results are better than MIC and 12 numbers of results are better than Pearson coefficient, respectively. Furthermore, there exist strong linear or nonlinear correlations between variables in the

iris dataset, seeds dataset, wine dataset and three synthetic datasets according to the scatter plots, which are shown in Figure 6. For the variables of the iris dataset, seeds dataset, wine dataset, Two_cluster, Twomoons and Five_cluster, the correlation coefficient values of variables obtained in the proposed method are almost larger than MIC, Pearson coefficient and MI, which indicates that the new correlation analysis framework can evaluate not only correlations of variables subject to distributions and randomness, but also can evaluate linear and nonlinear correlations between variables without explicit functional relationships well.

4) The comprehensive analysis shows that MIC can measure only the correlation with specific types of functional relationships but cannot identify the type of variable with distribution, and the Pearson coefficient can evaluate only the type of linear functional relationship between variables but the rest of the relationship types cannot be well measured. However, the proposed correlation analysis framework in this paper cannot only measure the linear or nonlinear correlation of variables containing any functional relationship types, but can also evaluate the relationship of variables with functional relationships involving specific distributions. In addition, the correlation of variables with obvious regularity in the scatter plot and the correlation of variables without clear functional relationship can also be evaluated well. The proposed method comprehensively considers the degree of functional relationship between variables, the degree uncertainty of variables and the degree of dependence on the variables containing distributions, which has a more comprehensive evaluation ability and better evaluation effect.

6. Conclusions

Most correlation analysis methods do not comprehensively consider the impact of uncertainty and distribution in variables when analyzing the correlation between variables, which leads to the neglect of regularity information in the variables, especially the strong regularity information in variables, so that the correlation coefficient values between variables with regularity calculated by traditional methods are too small, particularly the correlation coefficient values between variables with strong regularity. Moreover, it is believed that there is no correlation or weak correlation between variables. Based on the above problems, a novel correlation analysis framework of variables (RVCR-CA) is proposed in this paper, which considers the degree of functional relationship, the degree of uncertainty for the distribution of variables and the degree of dependence on the variables containing distributions. The correlation analysis framework cannot only evaluates the correlation of variables with obvious regularity in the scatter plots, but can also measure the correlation of variables without regularity in the scatter plots. In addition, for the variables without explicit functional relationships, the framework can analyze the correlation between different variables more comprehensively. In this paper, the experimental design is carried out from the perspective of the functional relationship variables containing specific distribution and the variables without clear functional relationships. Comparing the evaluation methods of correlation, such as MIC, Pearson coefficient and MI, the proposed framework cannot only measure the correlation of variables with any specific functional relationship, but can also evaluate the functional relationship variables containing specific distribution at the same time. For the variables without clear functional relationships, the new method can also better measure the correlation of variables. The proposed correlation analysis framework has a more comprehensive evaluation ability and better evaluation effect.

Use of AI tools declaration

The authors declare that they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant No 82260849); the National Natural Science Foundation of China (Grant No 61562045); and Jiangxi University of Chinese Medicine Science and Technology Innovation Team Development Program (Grant No CXTD22015).

Conflict of interest

The authors declare that there are no conflicts of interest.

References

1. N. J. Gogtay, U. M. Thatte, Principles of correlation analysis, *J. Assoc. Physicians India*, **65** (2017), 78–81.
2. J. Zhou, Y. Ma, Y. Liu, Y. Xiang, C. Cao, H. Yu, et al., A correlation analysis between the nutritional status and prognosis of COVID-19 patients, *J. Nutr. Health Aging*, **25** (2020), 1–10. https://doi.org/10.1007/s12603-020-1457-6
3. X. Xu, X. He, A. Qian, R. C. Qiu, A correlation analysis method for power systems based on random matrix theory, *IEEE Trans. Smart Grid*, **8** (2015), 1811–1820. https://doi.org/10.1109/tsg.2015.2508506
4. J. Liu, N. An, C. Ma, X. F. Li, J. Zhang, W. Zhu, et al., Correlation analysis of intestinal flora with hypertension, *Exp. Ther. Med.*, **16** (2018), 2325–2330. https://doi.org/10.3892/etm.2018.6500
5. C. E. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.*, **27** (1948), 584093. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
6. K. Pearson, Determination of the coefficient of correlation, *Science*, **30** (1909), 23–25. https://doi.org/10.1126/science.30.757.23
7. V. Aguiar, I. Guedes, Shannon entropy, Fisher information and uncertainty relations for log-periodic oscillators, *Physica A*, **423** (2015), 72–79. https://doi.org/10.1016/j.physa.2014.12.031
8. D. N. Reshef, Y. A. Reshef, H. K. Finucane, R. F. Grossman, G. McVean, P. J. Turnbaugh, et al., Detecting novel associations in large datasets, *Science*, **334** (2011), 1518–1524. https://doi.org/10.1126/science.1205438
9. H. Xiong, P. Shang, Weighted multifractal cross-correlation analysis based on Shannon entropy, *Commun. Nonlinear Sci. Numer. Simul.*, **30** (2016), 268–283. https://doi.org/10.1016/j.cnsns.2015.06.029
10. G. J. Székely, M. L. Rizzo, N. K. Bakirov, Measuring and testing dependence by correlation of distances, *Ann. Stat.*, **35** (2007), 2769–2794. https://doi.org/10.1214/009053607000000505
11. O. H. Diserud, F. Odegaard, A multiple-site similarity measure, *Biol. Lett.*, **3** (2007), 20–22. https://doi.org/10.1098/rsbl.2006.0553

12. R. H. Hariri, E. M. Fredericks, K. M. Bowers, Uncertainty in big data analytics: survey, opportunities, and challenges, *J. Big Data*, **6** (2019), 1–16. https://doi.org/10.1186/s40537-019-0206-3
13. C. S. Lai, Y. Tao, F. Xu, W. W. Y. Ng, Y. W. Jia, H. L. Yuan, et al., A robust correlation analysis framework for imbalanced and dichotomous data with uncertainty, *Inf. Sci.*, **470** (2019), 58–77. https://doi.org/10.1016/j.ins.2018.08.017
14. W. G. Favieiro, A. Balbinot, Paraconsistent random forest: an alternative approach for dealing with uncertain data, *IEEE Access*, **7** (2019), 149714–147927. https://doi.org/10.1109/access.2019.2946256
15. X. Chen, Y. Zhu, Uncertain random linear quadratic control with multiplicative and additive noises, *Asian J. Control*, **23** (2020), 2849–2864. https://doi.org/10.1002/asjc.2460
16. Y. Yang, P. Perdikaris, Adversarial uncertainty quantification in physics-informed neural networks, *J. Comput. Phys.*, **394** (2019), 136–152. https://doi.org/10.1016/j.jcp.2019.05.027
17. J. Ayensa-Jiménez, H. M. Doweidar, A. J. Sanz-Herrera, M. Doblaré, A new reliability-based data-driven approach for noisy experimental data with physical constraints, *Comput. Methods Appl. Mech. Eng.*, **328** (2018), 752–774. https://doi.org/10.1016/j.cma.2017.08.027
18. J. P. de Villiers, K. Laskey, A. L. Jousselme, E. Blasch, A. Waal, G. Pavlin, et al., Uncertainty representation, quantification and evaluation for data and information fusion, in *2015 18th International Conference on Information Fusion (Fusion)*, (2015), 50–57.
19. B. E. Niven, V. C. Deutsch, Calculating a robust correlation coefficient and quantifying its uncertainty, *Comput. Geosci*., **40** (2012), 1–9. https://doi.org/10.1016/j.cageo.2011.06.021
20. R. A. Johnson, D. W. Wichern, *Applied Multivariate Statistical Analysis*, Springer Berlin, Heidelberg, (2015), 517. https://doi.org/10.1007/978-3-662-45171-7
21. R. B. Nelsen, Concordance and copulas: a survey, in *Distributions with Given Marginals and Statistical Modelling*, Springer Netherlands, (2002), 167–177. https://doi.org/10.1007/978-94-017-0061-0_18
22. S. Kotz, Encyclopedia of statistical sciences, *J. Am. Stat. Assoc.*, **93** (1998), 281–317. https://doi.org/10.2307/2669895
23. M. Jian, Discovering association with copula Entropy, preprint, arXiv:1907.12268.
24. I. J. Schoenberg, Contributions to the problem of approximation of equidistant data by analytic functions, Part A: On the problem of smoothing or graduation, a first class of analytic approximation formulas, *Quart. Appl. Math.*, **4** (1946), 112–141.
25. D. H. Zhang, Entropy—A measure of uncertainty of random variable, *Syst. Eng. Electron.*, **11** (1997), 3–7.
26. I. Farrance, R. Frenkel, Measurement uncertainty and the importance of correlation, *Clin. Chem. Lab. Med.*, **59** (2021), 7–9. https://doi.org/10.1515/cclm-2020-1205
27. A. Sklar, Fonctions de répartition à N dimensions et leurs marges, in *Annales de l'ISUP*, **8** (1959), 229–231.
28. S. Ly, K. H. Pho, S. Ly, W. K. Wong, Determining distribution for the product of random variables by using copulas, *Risks*, **7** (2019), 23.
29. W. E. Donath, A. J. Hoffman, Lower bounds for the partitioning of graphs, *IBM J. Res. Dev.*, **17** (1973), 420–425. https://doi.org/10.1147/rd.175.0420
30. T. L. Saaty, K. P. Kearns, The analytic hierarchy process, *Anal. Plann.*, **1985** (1985), 19–62. https://doi.org/10.1016/b978-0-08-032599-6.50008-8
31. M. Brunelli, Introduction to the analytic hierarchy process, *Springer*, (2014), 33–44. https://doi.org/10.1016/B978-0-12-416727-8.00003-5

32. C. D. Boor, Least squares cubic spline approximation I-fixed knots and II-fixed knots, *Purdue Univ. Rep.*, (1968), 10014776809.
33. D. Mukherjee, An error reduced and uniform parameter approximation in fitting of B-spline curves to data points, preprint, arXiv:2005.08468.
34. O. Nave, Modification of semi-analytical method applied system of ODE, *Mod. Appl. Sci.*, **14** (2020), 75. https://doi.org/10.5539/mas.v14n6p75
35. B. Nie, Y. W. Du, J. Q. Du, Y. Rao, Y. C. Chao, X. P. Zheng, et al., A novel regression method: Partial least distance square regression methodology, *Chemom. Intell. Lab. Syst.*, **237** (2023), 104827. https://doi.org/10.1016/J.CHEMOLAB.2023.104827
36. A. Shemyakin, A. Kniazev, Random variables and distributions, in *Introduction to Bayesian Estimation and Copula Models of Dependence*, John Wiley & Sons, Inc., (2017), 103–140. https://doi.org/10.1002/9781118959046.ch1
37. C. Genest, L. P. Rivest, Statistical inference procedures for bivariate Archimedean copulas, *J. Am. Stat. Assoc.*, **88** (1993), 1034–1043. https://doi.org/10.1080/01621459.1993.10476372
38. S. Lipovetsky, Understanding the analytic hierarchy process, *Technometrics*, **2** (2021), 278–279. https://doi.org/10.1007/978-3-319-33861-3_2

Appendix

Table A1. The results of Copula function selection for functional relationship variables with a specific distribution.

Function expressions	Indicator	Clayton copula	Gumbel copula	Frank copula
$y = \lvert x \rvert + \delta$	theta	-0.065	**0.968**	-0.301
	Kendell's coefficient	-0.033	-0.033	-0.033
	AIC	-7648.202	**-8069.438**	-7660.186
	BIC	-7636.832	**-8058.069**	-7648.817
	fitting error	0.008	**0.006**	0.008
$y = 4(x^2 - 0.5)^2 + \delta$	theta	-0.111	**0.944**	-0.533
	Kendell's coefficient	-0.059	-0.059	-0.059
	AIC	-6548.274	**-6900.723**	-6558.867
	BIC	-6536.905	**-6889.353**	-6547.498
	fitting error	0.017	**0.013**	0.017
$y = 2x^2 + \delta$	theta	-0.040	**0.980**	-0.183
	Kendell's coefficient	-0.020	-0.020	-0.020
	AIC	-7597.857	**-8427.888**	-7603.843
	BIC	-7586.488	**-8416.518**	-7592.474
	fitting error	0.009	**0.005**	0.009
$y = 2(x^3 - \delta)^4$	theta	0.017	**1.008**	0.074
	Kendell's coefficient	0.008	0.008	0.008
	AIC	-9005.123	**-9105.323**	-9086.352
	BIC	-8993.754	**-9093.953**	-9074.983
	fitting error	0.004	**0.003**	**0.003**

Continued on next page

Function expressions	Indicator	Clayton copula	Gumbel copula	Frank copula
$y = \delta x^5$	theta	-0.005	**0.998**	-0.022
	Kendell's coefficient	-0.002	-0.002	-0.002
	AIC	-6770.312	**-6770.915**	-6767.358
	BIC	-6758.943	**-6759.545**	-6755.989
	fitting error	**0.015**	**0.015**	**0.015**
$y = \pm(x^2 + \delta)$	theta	0.028	**1.006**	-0.749
	Kendell's coefficient	0.061	0.061	0.061
	AIC	-6216.165	**-6918.485**	-6903.422
	BIC	-6204.796	**-6907.116**	-6892.053
	fitting error	0.020	**0.013**	**0.013**
$\begin{cases} x = \sin \pi t + \delta \\ y = \cos \pi t + \delta \end{cases}$	theta	0.566	1.283	**2.067**
	Kendell's coefficient	0.221	0.221	0.221
	AIC	-6694.823	-6975.023	**-7390.698**
	BIC	-6683.454	-6963.654	**-7379.3298**
	fitting error	0.015	0.013	**0.010**
$\begin{cases} x = 2\sin t - \sin 2t + \delta \\ y = 2\cos t - \cos 2t + \delta \end{cases}$	theta	0.076	**1.038**	0.331
	Kendell's coefficient	0.037	0.037	0.037
	AIC	-6164.096	**-6191.257**	-6158.928
	BIC	-6152.727	**-6179.888**	-6147.559
	fitting error	**0.021**	**0.021**	**0.021**
$\begin{cases} x = t\sin(\pi t) + \delta \\ y = t\cos(\pi t) + \delta \end{cases}$	theta	0.053	**1.026**	0.230
	Kendell's coefficient	0.026	0.026	0.026
	AIC	-6205.095	**-6223.546**	-6201.534
	BIC	-6193.726	**-6212.176**	-6190.165
	fitting error	0.021	**0.020**	0.021
$\begin{cases} x = 2\sin(5t)\cos t + \delta \\ y = 2\sin(5t)\sin t + \delta \end{cases}$	theta	0.136	**1.067**	0.575
	Kendell's coefficient	0.064	0.064	0.064
	AIC	-6118.383	**-6164.713**	-6109.877
	BIC	-6107.013	**-6153.343**	-6098.508
	fitting error	0.022	**0.021**	0.022
$\begin{cases} x = 2(\cos(360t))^3 + \delta \\ y = 2(\sin(360t))^3 + \delta \end{cases}$	theta	0.085	**1.043**	0.368
	Kendell's coefficient	0.041	0.041	0.041
	AIC	-6172.419	**--6203.614**	-6166.508
	BIC	-6161.050	**-6192.245**	-6155.139
	fitting error	0.021	**0.021**	**0.021**
$\begin{cases} x = -9\sin(2t) - 5\sin(3t) + \delta \\ y = 9\cos(2t) - 5\cos(3t) + \delta \end{cases}$	theta	0.014	**1.007**	0.062
	Kendell's coefficient	0.007	0.007	0.007
	AIC	-6696.385	**-6702.803**	-6695.131
	BIC	-6685.015	**-6691.434**	-6683.761
	fitting error	0.015	**0.015**	0.015

Note: the best results are shown in bold.

Table A2. The results of the selection of the optimal copula function for variables.

Datasets	Indicator	Clayton copula	Gumbel copula	Frank copula
Iris(x1,x2)	theta	-0.135	**0.933**	-0.652
	Kendell's coefficient	-0.072	-0.072	-0.072
	AIC	-1029.535	**-1037.073**	-1032.074
	BIC	-1021.514	**-1029.052**	-1024.053
	fitting error	0.032	**0.031**	0.032
Iris(x2,x3)	theta	-0.309	0.846	**-1.687**
	Kendell's coefficient	-0.182	-0.182	-0.182
	AIC	-1176.229	-1304.415	**-1309.328**
	BIC	-1168.208	-1296.393	**-1301.307**
	fitting error	0.020	**0.013**	**0.013**
Iris(x3,x4)	theta	**8.153**	5.077	18.501
	Kendell's coefficient	0.803	0.803	0.803
	AIC	**-1418.980**	-1374.472	-1395.611
	BIC	**-1410.959**	-1366.450	-1387.590
	fitting error	**0.009**	0.010	**0.009**
Seeds(x1,x3)	theta	**1.638**	1.819	4.899
	Kendell's coefficient	0.450	0.450	0.450
	AIC	**-1407.815**	-1375.620	-1391.465
	BIC	**-1399.121**	-1366.926	-1382.771
	fitting error	**0.035**	0.038	0.036
Seeds(x3,x4)	theta	**0.654**	1.327	2.333
	Kendell's coefficient	0.246	0.246	0.246
	AIC	**-1534.568**	-1516.419	-1522.695
	BIC	**-1525.874**	-1507.725	-1514.001
	fitting error	**0.026**	0.027	0.027
Seeds(x1,x6)	theta	-0.307	**0.847**	-1.675
	Kendell's coefficient	-0.181	-0.181	-0.181
	AIC	-1464.723	**-1489.371**	1472.437
	BIC	-1456.029	**-1480.677**	-1463.742
	fitting error	0.030	**0.029**	0.030
Glass(x1,x2)	theta	-0.100	**0.950**	-0.475
	Kendell's coefficient	-0.053	-0.053	-0.053
	AIC	-1850.694	**-1861.513**	-1856.081
	BIC	-1841.962	**-1852.781**	-1847.349
	fitting error	**0.013**	**0.013**	**0.013**
Glass(x2,x3)	theta	0.066	**1.033**	0.288
	Kendell's coefficient	0.032	0.032	0.032
	AIC	-1749.192	**-1765.237**	-1746.311
	BIC	-1740.460	**-1756.505**	-1737.579
	fitting error	0.017	**0.016**	0.017

Continued on next page

Datasets	Indicator	Clayton copula	Gumbel copula	Frank copula
Glass(x1,x4)	theta	-0.595	0.703	**-4.487**
	Kendell's coefficient	-0.423	-0.423	-0.423
	AIC	-1786.484	-1260.540	**-1825.382**
	BIC	-1777.752	-1252.177	**-1816.650**
	fitting error	0.015	0.029	**0.014**
Wine(x2,x3)	theta	0.207	**1.104**	0.851
	Kendell's coefficient	0.094	0.094	0.094
	AIC	-1286.481	**-1294.058**	-1290.871
	BIC	-1278.117	**-1285.694**	-1282.508
	fitting error	0.027	**0.026**	**0.026**
Wine(x3,x8)	theta	-0.350	0.825	**-1.980**
	Kendell's coefficient	-0.212	-0.212	-0.212
	AIC	-1190.365	1155.657	**-1198.454**
	BIC	-1182.001	-1147.294	**-1190.090**
	fitting error	0.035	0.039	**0.034**
Wine(x4,x5)	theta	0.697	1.348	**2.460**
	Kendell's coefficient	0.258	0.258	0.258
	AIC	-1405.072	-1332.711	**-1415.150**
	BIC	-1396.709	-1324.347	**-1406.787**
	fitting error	**0.019**	0.024	**0.019**
Two_cluster	theta	2.739	2.369	**7.369**
	Kendell's coefficient	0.578	0.578	0.578
	AIC	-3717.193	-3550.687	**-3622.591**
	BIC	-3707.210	-3540.704	**-3612.608**
	fitting error	0.010	0.012	**0.006**
Twomoons	theta	-0.402	0.799	**-2.385**
	Kendell's coefficient	-0.251	-0.251	-0.251
	AIC	-17,417.483	-17,881.203	**-18,373.622**
	BIC	-17,404.854	-17,868.574	**-18,360.993**
	fitting error	0.003	**0.002**	**0.002**
Five_cluster	theta	-0.390	0.805	**-2.291**
	Kendell's coefficient	-0.242	-0.242	-0.242
	AIC	-17,246.295	-18,408.549	**-19,268.310**
	BIC	-17,233.093	-18,395.347	**-19,255.108**
	fitting error	0.013	0.010	**0.008**
Roll	theta	0.200	**1.071**	0.647
	Kendell's coefficient	0.067	**0.067**	0.067
	AIC	-21,451.205	**-23,255.264**	-21,448.762
	BIC	-21,438.003	**-23,242.062**	-21,435.560
	fitting error	0.005	**0.003**	0.005

Note: the best results are shown in bold.