Journal of Intelligent Manufacturing (2022) 33:1699–1713
https://doi.org/10.1007/s10845-021-01751-w

Nonparametric-copula-entropy and network deconvolution method for causal discovery in complex manufacturing systems

Yanning Sun[1] · Wei Qin[1] · Zilong Zhuang[1]

Received: 27 December 2019 / Accepted: 20 February 2021 / Published online: 11 March 2021

Abstract

To clarify the causality among process parameters is a core issue of data-driven production performance analysis and product quality optimization. The difficulty lies in accurately measuring and distinguishing direct and indirect associations of complex manufacturing systems. In this work, the nonparametric-copula-entropy and network deconvolution method is proposed for causal discovery in complex manufacturing systems. Firstly, based on copula theory and kernel density estimation method, the nonparametric-copula-entropy is introduced to improve the accuracy of association measurement between parameters, and its superiority is verified by comparing with the results of different association measurement methods. Then, the global association matrix is constructed by the nonparametric-copula-entropy, and network deconvolution method is employed to extract the direct information from the global association matrix. The proposed method is tested by using an open gene expression dataset. Finally, as an experimental application, the causal analysis for a diesel engine production line is carried out by the proposed method. The results show that the proposed method can reveal causal relationship between process parameters and quality parameters in the diesel engine production line well, which provide theoretical guidance and implementation approach for the optimal control of complex manufacturing system.

Keywords Causal discovery · Complex manufacturing system · Data-driven · Nonparametric-copula-entropy · Network deconvolution

List of symbols

Symbols

$a_1(x_i)$, $a_2(x_i)$	Parameters of multivariate kernel function $K(\cdot)$
B	Base-number of logarithmic functions
$c(u_1, u_2, \ldots, u_N)$	Probability density function corresponding to copula function
$C(u_1, u_2, \ldots, u_N)$	Copula function
$f(x_1, x_2, \ldots, x_N)$	Joint probability density function of N random variables X_1, $X_2, \ldots, X_N$
$f_1(x_1), f_2(x_2) \ldots, f_N(x_N)$	Marginal probability density function of N random variables $X_1, X_2, \ldots, X_N$
$F(x_1, x_2, \ldots, x_N)$	Joint distribution function of N random variables $X_1, X_2, \ldots, X_N$
$F_1(x_1)$, $F_2(x_2) \ldots, F_N(x_N)$	Marginal distribution function of N random variables X_1, X_2, $\ldots, X_N$
$G(\cdot)$	Distribution function of univariate kernel function
G	Global association matrix
G_{dir}	Direct association matrix
h	Bandwidth of univariate kernel density estimation
h_i	Bandwidth of multivariate kernel density estimation for the ith random variable
I	Identity matrix
$k(\cdot)$	Univariate kernel function
$K(\cdot)$	Multivariate kernel function
$\mathrm{MI}(X_1, X_2, \ldots, X_N)$	Mutual information of multi-dimension random variables X_1, $X_2, \ldots, X_N$
N	Number of random variables
n	Rotational speed

✉ Wei Qin
wqin@sjtu.edu.cn

[1] School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai, China

$p(x_i)$	Subfunctions of multivariate kernel function $K(\cdot)$
P	Power
T	Torque
u_i	Independent variable of copula function, $u_i = F_i(x_i)$
U_i	Random variable i subject to uniform distribution
$\mathbf{U}$	Vector form of $[u_1, u_2, \ldots, u_N]$
x_i	Observation values of the *i*th random variable
X_i	The *i*th random variable

Abbreviations

AMV	Association measurement values
AUC	Area under ROC curve
CE	Copula entropy
FN	The numbers of false negatives
FP	The numbers of false positives
FPR	False positive rate
KDE	Kernel density estimation
MES	Manufacturing execution system
NCE	Nonparametric copula entropy
ND	Network deconvolution
OLE	Object linking and embedding
PDF	Probability density function
QAS	Quality assurance system
RMSE	Root mean square error
ROC	Receiver operating characteristic
TN	The numbers of true negatives
TP	The numbers of true positives
TPR	True positive rate
WS	Work station

Introduction

With the trend of information and intelligence in manufacturing enterprises, mass manufacturing process data are collected and stored. These data contain the inherent laws of complex manufacturing process, so they have huge practical values and numerous opportunities for scientific discovery. Mining meaningful data relationships, especially causality in complex manufacturing systems is a promising theoretical and engineering problem. Association and directionality are important topics in data-driven causal discovery. Given a observation dataset that manufacturing process parameters take in different conditions, a topological structure can be inferred by computing the pairwise correlation, such as mutual information or other association metrics (Feizi et al. 2013). For example, a true topological path $X_i \rightarrow X_k \rightarrow X_j$ can result in an observable response between X_i and X_j, falsely suggesting the existence of a direct association between them (Barzel and Barabási, 2013). The association between X_i and X_k, and X_k and X_j can be called direct associations, which represent causality. And the association between X_i and X_j, can be called indirect association, which means non-causality. For a complex manufacturing system, the implementation of each process may have a direct or indirect effect on the final product quality. There are also direct and indirect associations among process parameters (Qin et al. 2018). Therefore, how to accurately measure and distinguish direct and indirect associations poses a challenge for causal discovery in complex manufacturing systems.

To meet the above research challenge, a two-stage causal discovery method for complex manufacturing systems is proposed in this paper. Stage 1: According to copula theory and kernel density estimation (KDE) method, the nonparametric copula entropy (NCE) is used to quantify the association relationship between parameters, and construct the global association matrix of the system. Stage 2: The network deconvolution (ND) algorithm is designed to extract the direct association information from the global association matrix.

The rest of this paper is structured as follows: Related studies are reviewed in "Literature review" section. In "NCE-ND method for causal analysis" section, the NCE-ND causal analysis method is proposed and validated by an open gene expression dataset. In "Experimental application" section, the experimental application and analysis discussion are carried out for the historical data from the real diesel engine manufacturing system. The section of "Conclusions and future works" summarizes the main conclusions of this paper, and puts forward some issues that may need further research in the future.

Literature review

In this section, studies related to the research topic are reviewed, furthermore some contributions and gaps to the extant literature are also summarized.

Traditionally, Pearson correlation coefficient is widely used to measure the linear relationship between the observed variables. Frigieri et al. (2019) conducted correlation analysis between noise density and machining parameters in hardened steel turning by Pearson correlation coefficient. In addition, the Kendall and Spearman rank correlation coefficients can be used to measure the consistency of the changes between the two random variables. They belong to the nonparametric statistical method, which does not require the edge distributions of random variables, thus ensuring a wider range of applications (Croux & Dehon, 2010).

However, these three methods can only measure the linear correlation between random variables.

Mutual information (MI) is a way of information measurement in information theory (Cover & Thomas, 2012), which can be understood as the information of one random variable contained in another random variable. If two random variables are independent, the corresponding MI is zero, and if they have a certain correlation, the MI is a positive value. MI can describe nonlinear correlation, and it is not susceptible to noise and data transformation in the process of initialization, so it has attracted wide attention. For example, MI is used to construct gene regulatory networks for gene expression data in the field of gene regulation (Altay & Emmert-Streib, 2010). But it is also true that estimating MI is not always easy. In practical applications, the estimation accuracy of MI has an directly effect on the identification of dependency between random variables (Han & Ren, 2015). Based on entropy estimates by *k*-nearest neighbor distance, Kraskov et al. (2004) presented two classes of improved MI estimators from random point samples based on joint probability density distribution. Using the Kraskov's methods, Rossi et al. (2006) suggested the use of the MI to estimate the association of spectral variables in a prediction problem and presented a effective method to select the spectral variables based on their MI with the output. More recently, based on a similar approach, Fang et al. (2015) improved feature selection method for multidimensional time series to achieve dimension reduction. However, these *k*-nearest neighbors-based approaches have been confirmed to be inferior to the kernel density estimators (Khan et al., 2007), which is a nonparametric representation of the probability density function (PDF) of random variables (Silverman, 1986). Thomas et al. (2014) used KDE together with the tuning of the kernel's covariance matrix using data attributes to calculate average MI, and achieved good results in MATLAB software.

In addition, combining copula function (Sklar, 1959) with Shannon entropy theory (Shannon, 1948), Ma and Sun (2008) proposed the definiation of copula entropy (CE). Their research pointed out that MI is equivalent to the opposite number of CE. Therefore, the correlation between variables can be measured by the CE. As a result, the MI estimation method based on CE is widely used in many fields, especially in hydrology and water resources (Chen, 2013; Zachariah & Reddy, 2013; Chen & Guo, 2019), finance (Hu, 2006; Patton, 2002; Xu, 2005; Zhao & Lin, 2011), and industry (Gu et al., 2019; Jeon et al., 2019; Wei et al., 2019). The common parametric copula functions have the elliptic copula functions, such as n copula and t copula, and the Archimedes copula functions, such as Clayton copula, Frank copula and Gumbel copula (Embrechts et al., 2001). These parameterized forms are based on the ideal hypothesis model, which limits the application of copula theory. At present, some scholars have done some researches on nonparametric copula function. Nicoloutsopoulos (2005) studied parametric and nonparametric methods of copula estimation, especially for Archimedean class of copulas. In order to describe the correlation structure between variables, Huard et al (2006) used Bayesian theory to select copula function. However, further research is rare in the literature.

In addition to the difficulty of accurate estimation, MI cannot detect the direct associations in the network. As an extension of MI, conditional mutual information, point mutual information and partial correlation are widely used to infer network structure or detect direct associations in many areas, including biology, logistics, engineering and social research (Shi et al., 2019). Zhang et al. (2012, 2015, 2016) employed conditional mutual information, conditional mutual inclusion information and point mutual information, respectively, to infer the direct associations in gene regulatory network through gene expression data. It is pointed out that conditional mutual information can quantify the nonlinear relationship between observed data variables, which is superior to linear measures, but there is a serious underestimation problem. In terms of accurately measuring the relationship in the network, point mutual information and condition mutual inclusion information have certain advantages. However, using the methods mentioned above, it is generally assumed that the observed data are normal distribution in order to simplify the amount of calculation, otherwise the calculation cost is unacceptable. The parameters in the production process often do not obey the normal distribution, which makes it difficult for the traditional methods to quantitatively analyze the relationship between process quality parameters and product quality in the manufacturing process. Feizi et al. (2013) proposed ND method to remove chained noise in association networks. It has been verified and achieved good accuracy in protein amino acid association network, gene regulation network and social network. This method has the advantages of simple calculation, high accuracy and wide range of application, which has been applied to detect the correlation relationship of manufacturing system (Qin et al., 2018). Qin et al. (2018) combined normalized MI and ND to analyze causal variables in diesel engine production, but the effects of calculation accuracy of MI was not studied.

In conclusion, although MI can be used for measurement of associations, its complex calculation formula makes it difficult to obtain accurate values. The value of MI can be calculated indirectly by CE, which is fast and accurate, but further research is rare about nonparametric copula function. Moreover, there is a lack of literature on the comparison of various association measurement methods and their practical application in the field of industrial manufacturing.

NCE-ND method for causal analysis

The definition and basic properties of NCE

Based on Sklar's theorem (Sklar, 1959), if $F(x_1, x_2, \ldots, x_N)$ is a multivariate joint distribution function of N random variables $X_1, X_2, \ldots, X_N$ with respective marginal distributions $F_1(x_1)$, $F_2(x_2)$, …, $F_N(x_N)$, then it is possible to establish a functional relationship between N-dimensional joint distribution function and univariate margins $F_1(x_1)$, $F_2(x_2)$, …, $F_N(x_N)$, as follows:

$$\begin{aligned} F(x_1, x_2, \ldots, x_N) &= C[F_1(x_1), F_2(x_2), \ldots, F_N(x_N)] \\ &= C(u_1, u_2, \ldots, u_N) \end{aligned} \tag{1}$$

where $F_i(x_i) = u_i$ for $i = 1, 2, \ldots, N$, with $U_i \sim U(0, 1)$, and C is a copula function (Chen et al. 2014). And the corresponding PDF is $c\ (u_1, u_2, \ldots, u_N)$, which can be denoted as:

$$\begin{aligned} c(u_1, u_2, \ldots, u_N) &= \frac{\partial^N C(u_1, u_2, \ldots, u_N)}{\partial u_1 \partial u_2 \ldots \partial u_N} \\ &= \frac{\partial^N C(u_1, u_2, \ldots, u_N)}{\partial F_1(x_1) \partial F_2(x_2) \ldots \partial F_N(x_N)} \\ &= \frac{\partial^N F(x_1, x_2, \ldots, x_N)}{f_1(x_1) f_2(x_2) \ldots f_N(x_N) \partial x_1 \partial x_2 \ldots \partial x_N} \\ &= \frac{f(x_1, x_2, \ldots, x_N)}{f_1(x_1) f_2(x_2) \ldots f_N(x_N)} \end{aligned} \tag{2}$$

where f denotes PDF. It can be seen that the copula function is a PDF of random variables $u_1, u_2, \ldots, u_N$, and it has the range of [0,1]. In order to avoid making model hypothesis, KDE method is introduced to obtain $c\ (u_1, u_2, \ldots, u_N)$. The KDE of marginal PDF $f_i(x_i)$ is firstly given by:

$$f_i(x_i) = \frac{1}{nh} \sum_{t=1}^{n} k\left(\frac{x_i - x_{it}}{h}\right) \tag{3}$$

where $x_{i1}, x_{i2}, \ldots, x_{in}$ are n random samples from an unknown distribution X_i, $k(\cdot)$ is the kernel function, and h is the bandwidth. Gaussian kernel function (Bowman & Azzalini, 1997) is employed to build KDE and Scott's rule (Scott, 2015) is used to calculate the optimal bandwidth in this paper.

The KDE of marginal distribution function $F_i(x_i)$ can be written as:

$$u_i = F_i(x_i) = \int_{-\infty}^{x_i} f_i(t_i) dt_i = \frac{1}{n} \sum_{t=1}^{n} G\left(\frac{x_i - x_{it}}{h}\right) \tag{4}$$

$$G(x) = \int_{-\infty}^{x} k(t) dt \tag{5}$$

Further, the nonparametric estimation of copula function is expressed as follows:

$$\begin{aligned} & c(u_1, u_2, \ldots, u_N) \\ &= \frac{1}{n h_1 h_2 \ldots h_N} \sum_{t=1}^{n} K\left(\frac{u_1 - u_{1t}}{h_1}, \frac{u_2 - u_{2t}}{h_2}, \ldots, \frac{u_N - u_{Nt}}{h_N}\right) \end{aligned} \tag{6}$$

$$K(x_1, x_2, \ldots, x_N) = p(x_1) p(x_2) \ldots p(x_N) \tag{7}$$

where u_{it} denotes the tth random samples from a uniform distribution U_i, $K(\cdot)$ is the multivariate kernel function considering the support set of copula function as [0, 1], and its subfunctions $p(x_i)$ are given by (Jones, 1993):

$$p(x_i) = \frac{(a_2(x_i) - a_1(x_i) x_i) k(x_i)}{a_0(x_i) a_2(x_i) - a_1^2(x_i)} \tag{8}$$

$$a_1(x_i) = \int_{(x_i - 1)/h_i}^{x_i/h_i} u_i k(u_i) du_i \tag{9}$$

$$a_2(x_i) = \int_{(x_i - 1)/h_i}^{x_i/h_i} u_i^2 k(u_i) du_i \tag{10}$$

By substituting Eq. (6) into to the formula of Shannon entropy (Shannon, 1948), NCE can be expressed as:

$$\begin{aligned} \mathrm{NCE} = - \int_0^1 \ldots \int_0^1 c(u_1, u_2, \ldots, u_N) \\ \log_b c(u_1, u_2, \ldots, u_N) \mathrm{d}u_1 \mathrm{d}u_2 \ldots \mathrm{d}u_N \end{aligned} \tag{11}$$

where the unit of NCE depends on base-number b. The unit is bit when $b = 2$, the unit is nat when $b = \mathrm{e}$, and the unit is Hart when $b = 10$. And $b = \mathrm{e}$ in this paper. Avoiding model assumption, the calculation process of NCE completely depends on the actual data. It is a model-free method with strong practical application prospect.

It is confirmed in literatures (Chen et al., 2014; Zhao & Lin, 2011) that MI (see Eq. (12)) is equal to the negative NCE. Therefore, both of them are important ways of association measurement.

$$\begin{aligned} & \mathrm{MI}(X_1, X_2, \ldots, X_N) \\ &= \int \ldots \int f(x_1, x_2, \ldots, x_N) log_b \frac{f(x_1, x_2, \ldots, x_N)}{f_1(x_1) f_2(x_2) \ldots f_N(x_N)} \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_N \end{aligned} \tag{12}$$

However, compared with MI, NCE has good properties in computational complexity and efficiency:

(1) For the formula of MI, KDEs of marginal PDF $f_1(x_1)$, $f_2(x_2)$, …, $f_N(x_N)$ and joint PDF $f\ (x_1, x_2, \ldots, x_N)$ are needed. And then there are two multiplication operations, one summation operation and one integration operation. It should be noted that the integral range is all real numbers in theory.

(2) For NCE, KDEs of marginal distribution $u_1, u_2, \ldots, u_N$ and joint PDF $c\ (u_1, u_2, \ldots, u_N)$ are needed. And then there are one multiplication operations, one summation operation and one integration operation with the integral range [0, 1]. This reduced integration range will greatly improve calculation accuracy and efficiency.

At the end of this subsection, to demonstrate the validity of NCE for association measurment, three linear correlation coefficients (Spearson, Pearson and Kendall), five common parametric copula functions (t copula, n copula, Clayton copula, Frank copula and Gumbel copula) are employed to compare with accurate MI (calculated by Eq. (12), as a benchmark) and NCE. The 10,000 × 2 Gaussian distribution data is randomly generated. The mathematical expectation values are [20, 50], the variance is [1, 1], and the correlation coefficient ranges from 0 to 1.

We defined the association measurement value (AMV) as the strength of correlation between variables. The larger the AMV, the stronger the correlation between variables. From the perspective of information theory, the larger the AMV, the greater the amount of common information among variables. Furthermore, AMV obtained from the mentioned methods are plotted in Fig. 1, which shows that benchmark MI, parametric copula-based method and NCE are all better than linear correlation coefficients on the measurement of nonlinear association, and NCE is in good agreement with benchmark MI. Further, root mean square error (RMSE) of different association measurement methods relative to benchmark MI are listed in Table 1. It shows that NCE has high calculation accuracy and its RMSE relative to benchmark MI is only 0.0236. Kendall correlation coefficient has the best performance in three linear correlation coefficients and Clayton copula has the best performance in the parametric copula-based method.

Causal inference based on NCE-ND

After the association measurement is carried out, the global association matrix of manufacturing parameters network can be written as:

$$G=\begin{bmatrix} 1 & -\mathrm{NCE}_{12} & \cdots & -\mathrm{NCE}_{1N} \\ -\mathrm{NCE}_{21} & 1 & \cdots & -\mathrm{NCE}_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ -\mathrm{NCE}_{N1} & -\mathrm{NCE}_{N2} & \cdots & 1 \end{bmatrix} \tag{13}$$

where NCE_{ij} is the nonparametric copula entropy of variables i and j ($i = 1, 2, \ldots, N$; $j = 1, 2, \ldots, N$; and $i \neq j$). G represents an information network. Because of the association transfer among variables, this information network contains the effects of direct and indirect associations. Therefore, in order to extract direct information from the information network, the ND method is employed to construct a direct association matrix from the global association matrix. The method assumes that the indirect associations can be approximated as the product of direct associations and they are summable. Based on the assumption, the global association matrix can be written by the convolution formula:

$$G = G_{\mathrm{dir}} + G_{\mathrm{dir}}^2 + G_{\mathrm{dir}}^3 + \cdots = G_{\mathrm{dir}}(I - G_{\mathrm{dir}})^{-1} \tag{14}$$

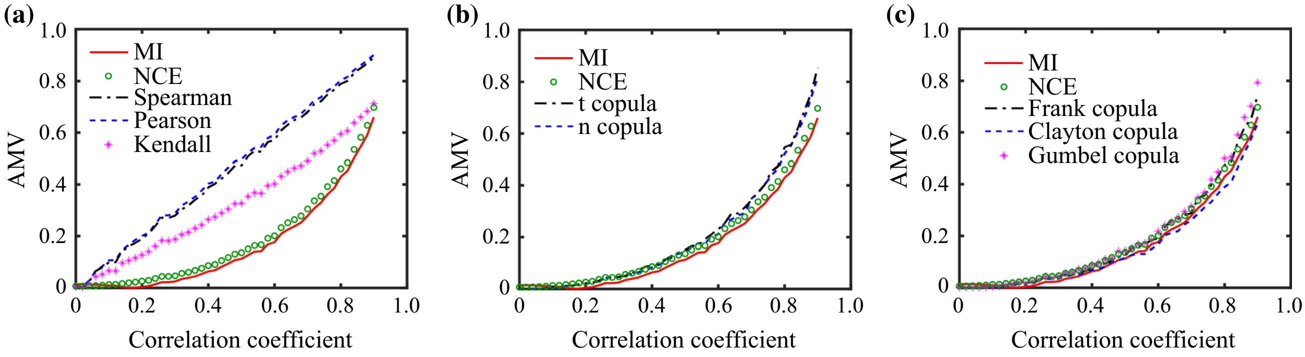


Fig. 1 Association measurement values for Gaussian distribution data obtained from different methods

Table 1 Comparison of association measurement RMSE for Gaussian distribution data

AMV	MI	NCE	Spearman	Pearson	Kendall	t copula	n copula
RMSE	–	0.0236	0.3030	0.3158	0.1690	0.0639	0.0551
AMV	Frank copula	Clayton copula	Gumbel copula				
RMSE	0.0290	0.0167	0.0438				

where G_{dir} is direct association matrix and I is identity matrix. Equation (14) can be transformed into the deconvolution formula:

$$G_{\mathrm{dir}}{=}G(I+G)^{-1} \tag{15}$$

which can be calculated by eigen decomposition and the more detailed algorithmic process about ND can be found in (Feizi et al., 2013). By the ND method, more accurate association network is reconstructed.

It is obvious find that all the values in the direct association matrix G_{dir} is over zero, that is to say, the defined network structure is a fully connected. What follows is to select an appropriate threshold t. We defined the adjacent matrix G_{adj} as,

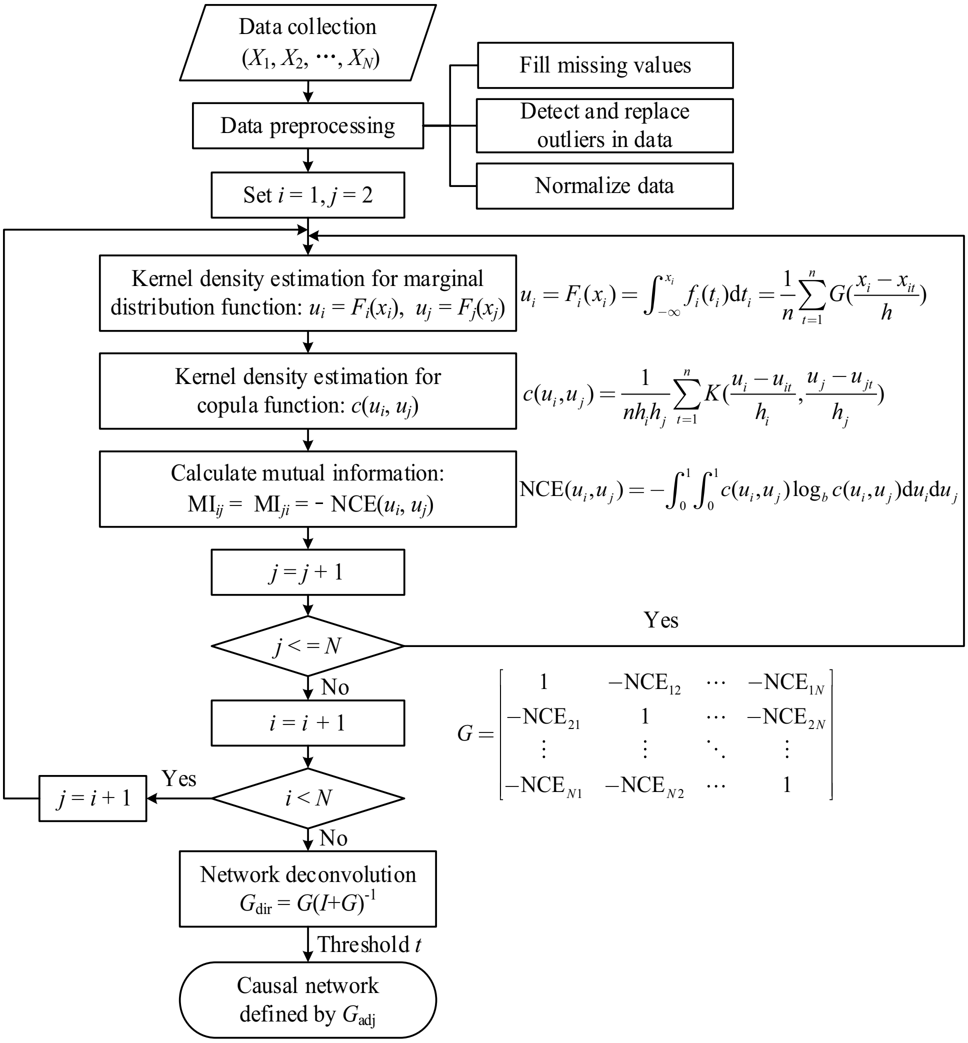
$$G_{\mathrm{adj}}{=}\begin{bmatrix} 0 & b_{12} & \dots & b_{1N} \\ b_{21} & 0 & \dots & b_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ b_{N1} & b_{N2} & \dots & 0 \end{bmatrix} \tag{16}$$

where b_{ij} is the element of adjacency matrix G_{adj}. If the b_{ij} is greater than t, it is set as 1 and the node i is connected with the node j. Otherwise, $b_{ij}=0$ and the edge between nodes i and j does not exist. Nodes with connected edges in the network have causality. From this point of view, we can see that different t will produce different causality network structure. How to find an optimal t is still open problem. In this paper, an appropriate threshold t is chosen by the experimental method and to avoid too dense the network, the 50%-60% edges with the highest values in G_{dir} are retained (Sun et al., 2017). The procedure of the causal inference is shown in Fig. 2.

Performance evaluation

In this subsection, the proposed method is validated by a synthetic nonlinear expression dataset from the dialogue for reverse engineering assessment and methods (DREAM) challenge. The dataset was a widely used benchmark network in the field of data-driven associations measurement, in

Fig. 2 Flow chart of causal network inference


Data collection
$(X_1, X_2, \cdots, X_N)$
Fill missing values
Data preprocessing
Detect and replace outliers in data
Set $i = 1, j = 2$
Normalize data
Kernel density estimation for marginal distribution function: $u_i = F_i(x_i)$, $u_j = F_j(x_j)$
$u_i = F_i(x_i) = \int_{-\infty}^{x_i} f_i(t_i)\mathrm{d}t_i = \frac{1}{n}\sum_{t=1}^{n} G(\frac{x_i - x_{it}}{h})$
Kernel density estimation for copula function: $c(u_i, u_j)$
$c(u_i,u_j) = \frac{1}{nh_ih_j}\sum_{t=1}^{n} K(\frac{u_i - u_{it}}{h_i}, \frac{u_j - u_{jt}}{h_j})$
Calculate mutual information: $\mathrm{MI}_{ij} = \mathrm{MI}_{ji} = -\mathrm{NCE}(u_i, u_j)$
$\mathrm{NCE}(u_i,u_j) = -\int_0^1\int_0^1 c(u_i,u_j)\log_b c(u_i,u_j)\mathrm{d}u_i\mathrm{d}u_j$
$j = j + 1$
Yes
$j <= N$
No
$i = i + 1$
$G = \begin{bmatrix} 1 & -\mathrm{NCE}_{12} & \cdots & -\mathrm{NCE}_{1N} \\ -\mathrm{NCE}_{21} & 1 & \cdots & -\mathrm{NCE}_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ -\mathrm{NCE}_{N1} & -\mathrm{NCE}_{N2} & \cdots & 1 \end{bmatrix}$
Yes
$j = i + 1$
$i < N$
No
Network deconvolution
$G_{\mathrm{dir}} = G(I{+}G)^{-1}$
Threshold t
Causal network defined by G_{adj}

which the expressed data were generated by a system of nonlinear ordinary differential equation, and the network structures were determined with the detailed dynamics of both transcriptional and translational processes. More detailed information about the dataset is available in the reference (Marbach et al., 2010). And the gene expression data with network size 100 and sample number 101 was adopted to test the performance of proposed method. In order to facilitate the readers' intuitive understanding of the selected data, the benchmark network structure is shown in Fig. 3a, and the heat map of the adjacency matrix for the top 20 genes is plotted in Fig. 3b.

Based on the proposed NCE-ND method, firstly, the global association matrix is generated by the NCE and the heat map of global association matrix for the top 20 genes is shown in Fig. 4a. From the figure, it can be seen that almost all genes are not weakly associated with other genes because of the existence of indirect associations. Then, to filter out these indirect associations, the ND method is used to extract direct association matrix, and the heat map of direct association matrix for the top 20 genes is shown in Fig. 4b. Lastly, the network structure can be inferred by setting a certain threshold.

Compared with Figs. 3b and 4b, it can be seen that the proposed method has a certain misjudgment rate, which is inevitable for all methods. The receiver operating characteristic (ROC) curve is used to evaluate the performance of different association measurement methods mentioned by "The definition and basic properties of NCE" subsection. With the false positive rate (FPR) as the transverse coordinate and the true positive rate (TPR) as the longitudinal coordinate, the ROC curve is drawn by setting different threshold values, as shown in Fig. 5. The formulas for calculating TPR and FPR are represented as:

$$\mathrm{TPR} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN}) \tag{17}$$

$$\mathrm{FPR} = \mathrm{FP}/(\mathrm{FP}+\mathrm{TN}) \tag{18}$$

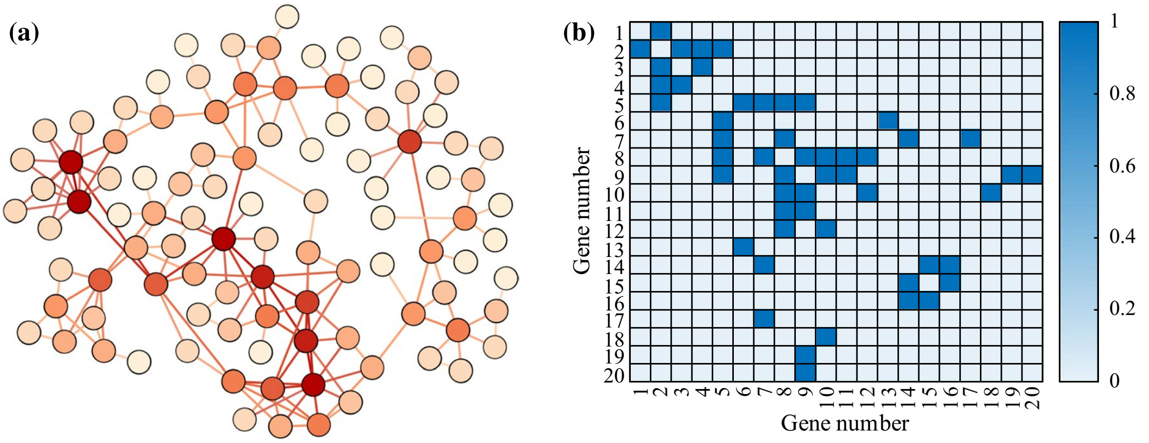


Fig. 3 Selected data schematic: **a** the network structure **b** the heat map of adjacency matrix for the top 20 genes

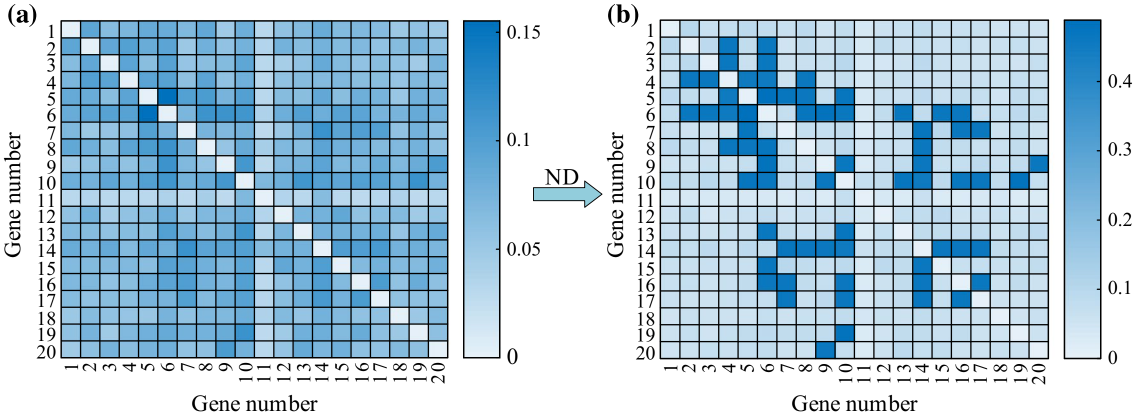


Fig. 4 Heat map of association matrix for the top 20 genes: **a** global association matrix **b** direct association matrix

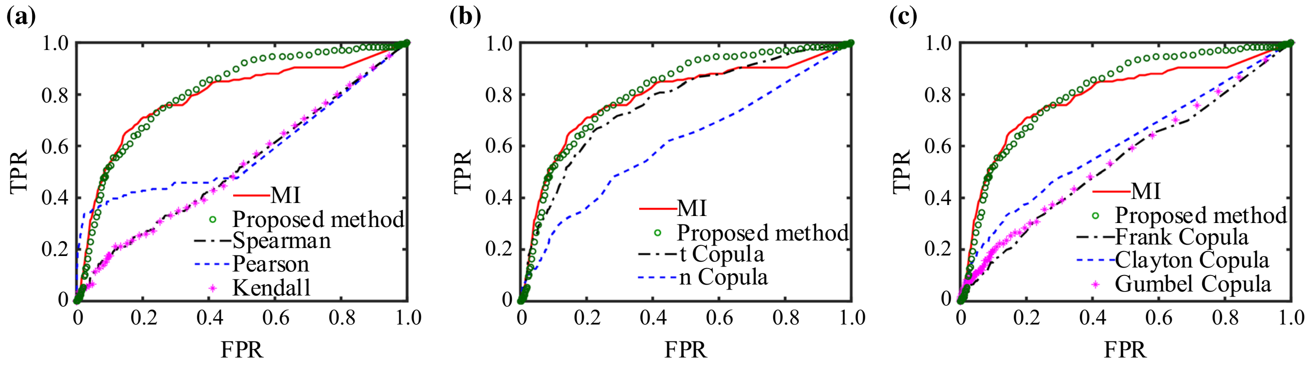


Fig. 5 ROC curves of different methods for the selected dataset

where TP, FP, TN and FN are the numbers of true positives, false positives, true negatives and false negatives respectively.

The larger the area under ROC curve (AUC), the better the method is. Comparison of AUC values of different methods for the selected dataset is listed in Table 2. It can be seen from the result that the method combining linear correlation coefficients and ND cannot detect the nonlinear relationships, thus the poor results is obtained (the AUCs are less than 0.6); The AUC of t couple-ND is 0.7666, which is better than that of n copula (only 0.6087). And the parametric copula function is not suitable for the data in this section because of its parametric form, so the results are poor (AUC maximum is only 0.6061). In other words, if the form of copula function is not suitable, it will cause a large loss of accuracy, which is also a reason why this paper proposes the NCE for associations measurement. Besides, the performance of MI-ND is also inferior to proposed method due to the accumulation of errors caused by multiple integrals in whole real number field. Thus, the proposed method got the best performance among all mentioned association measurement methods (the AUC is 0.8077).

Experimental application

In this section, the causal network of a diesel engine production line (as a typical complex manufacturing system, see Fig. 6) is inferred based on the proposed NCE-ND method, and some results are discussed.

Dataset description

The production line of diesel engine consists of four parts: main assembly line, sub-assembly lines 1–5, performance test line and package line. The diesel engine block and cooler, crankshaft, oil pump and camshaft, piston connecting rod unit, and cylinder are assembled by the sub assembly lines 1–5 respectively, and delivered to main assembly line. And the main assembly line carries on the overall assembly of the diesel engine. The performance indexes of diesel engine are tested in performance test line, such as power, fuel consumption rate, etc. As is shown in Fig. 6, the layout of the production line adopts serial and parallel connection, thus achieving the improvement of space utilization, the realization of cross-station operation, the rational utilization of manufacturing resources, and the reduction of cost. The manufacturing process data of diesel engine contains assembly process parameters and performance test parameters, which are collected through industrial automation sensors. Furthermore, those data are transmitted to manufacturing execution system (MES) and quality assurance system (QAS) through object linking and embedding (OLE), and process control is carried out accordingly.

There are more than 100 work stations (WS) on the diesel engine assembly line. A total of 172 assembly process parameters are tested. Diesel engine enters the stage of performance test after assembly. For demonstration purposes, as is shown in Table 3, 16 assembly process parameters (numbers 1–16) and 24 performance test parameters (numbers 17–40) are selected to construct the dataset of this section.

Table 2 Comparison of AUC values of different methods for the selected dataset

Methods	MI-ND	Proposed method	Spearman-ND	Pearson-ND	Kendall-ND	t copula-ND	n copula-ND
AUC	0.7902	0.8077	0.5260	0.5837	0.5255	0.7666	0.6087
Methods	Frank copula-ND	Clayton copula-ND	Gumbel copula-ND				
AUC	0.5372	0.6061	0.5584				

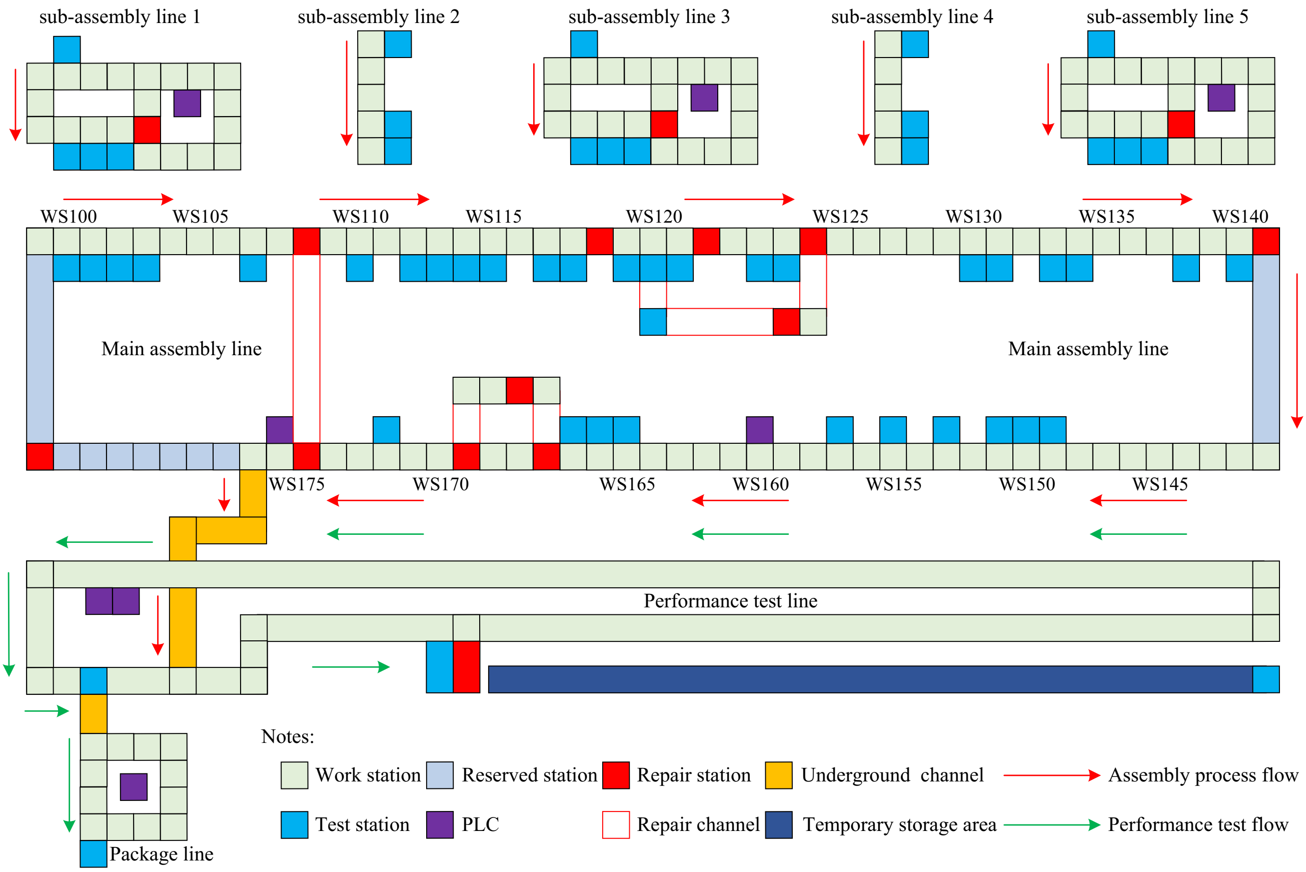


Fig. 6 Diagram of diesel engine manufacturing system

The marginal density functions obtained from the kernel estimation method is shown in Fig. 7. The symbols x_1, x_7 and x_8 denote cylinder liner protrusion height 1, starting torque and running torque, which are typical assembly process parameters. And the symbols x_{20}, x_{29} and x_{40} denote piston leakage, fuel consumption and power respectively, which are typical performance test parameters. This figure shows that the distribution of sample data of the diesel engine system is complex, not all of them obey Gaussian distribution, and KDE method retains the true characteristics of the sample data very well. Take the starting and running torques for example, the nonparametric estimation result of copula function is shown in Fig. 8. Copula function (distribution function, see Fig. 8b) and the corresponding PDF (see Fig. 8a) can be evaluated flexibly following actual data by the method of the previous section.

Results analysis and discussions

AMV between power (number 40) and the other 39 parameters of the diesel engine is shown in Fig. 9. The results show that the AMV between power and torque (number 26), rotational speed (number 39) is maximal. Besides, when the correlation is large, the range of AMV based on parameterized copula may exceed 1, that is to say, They are not a normalization method. NCE, MI, Spearman, Pearson and Kendall can ensure that AMV is in the range of 0 to 1. The comparison of association measurement RMSE is listed in Table 4. It can be seen that NCE has the best performance in all methods and the RMSE of the parametric copula-based method increases obviously. Furthermore, as we all know, $P = 9559 \cdot T \cdot n$, where P is power, T is torque and n is rotational speed. Based on this expert knowledge, there is a strong correlation between power and torque, speed. This is in line with our results.

The causal network of this diesel engine production line is inferred based on NCE-ND method and plotted in Fig. 10 by Gephi software (Bastian et al. 2009). As you can see from Fig. 10a, the number of edges connected to nodes ST and RT is the largest, that is, ST and RT are the two most important parameters for the whole system. Figure 10b shows the parameters associated with power: T, FCR, RS, Th, ST and RT, where the first two are assembly process parameters and the last three are performance test parameters. Therefore, the power of diesel engine can be controlled by regulating assembly process parameters ST and RT. Moreover,

Table 3 Selected manufacturing process parameters

Number	Symbol	Description	Number	Symbol	Description
1	CLPH1	Cylinder liner protrusion height 1	21	OT	Oil temperature
2	CLPH2	Cylinder liner protrusion height 2	22	OP	Oil pressure
3	CLPH3	Cylinder liner protrusion height 3	23	IAT	Intake air temperature
4	CLPH4	Cylinder liner protrusion height 4	24	IWT	Inlet water temperature
5	CLPH5	Cylinder liner protrusion height 5	25	CFC	Cumulative fuel (gas) consumption
6	CLPH6	Cylinder liner protrusion height 6	26	T	Torque
7	ST	Starting torque	27	ET	Exhaust temperature
8	RT	Running torque	28	FT	Fuel temperature
9	AC	Axial clearance	29	FC	Fuel consumption
10	CTM	Crankshaft turning moment	30	FCR	Fuel consumption rate
11	PPH1	Piston protrusion height 1	31	WG	Water gate
12	PPH2	Piston protrusion height 2	32	SI	Smoke Intensity
13	PPH3	Piston protrusion height 3	33	Th	Throttle
14	PPH4	Piston protrusion height 4	34	Rt	Running time
15	PPH5	Piston protrusion height 5	35	IOT	Intercooler outlet temperature
16	PPH6	Piston protrusion height 6	36	IOP	Intercooler outlet pressure
17	OWT	Outlet water temperature	37	IIT	Intercooler inlet temperature
18	AH	Ambient humidity	38	IIP	Intercooler inlet pressure
19	AP	Atmospheric pressure	39	RS	Rotational speed
20	PL	Piston leakage	40	P	Power

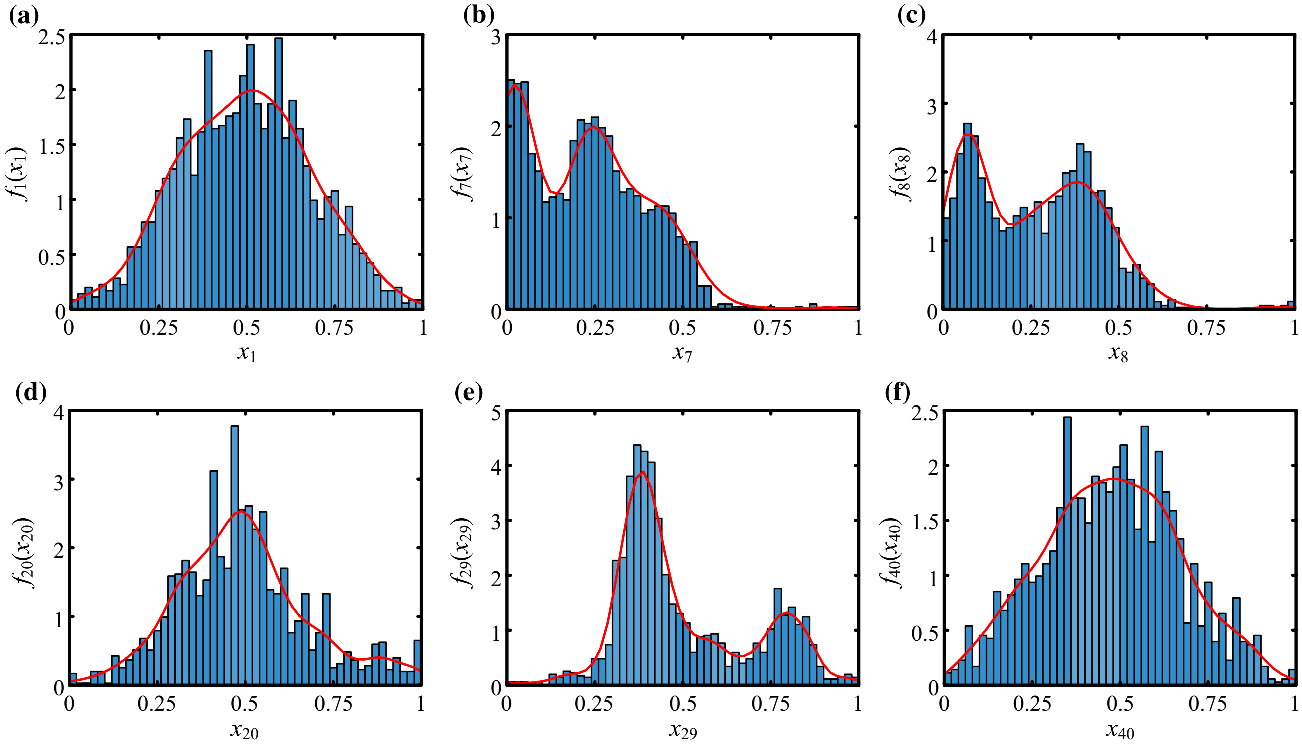


Fig. 7 Marginal density functions obtained from the kernel estimation method: **a** cylinder liner protrusion height 1, **b** crankshaft turning moment, **c** outlet water temperature, **d** piston leakage, **e** fuel consumption, **f** power

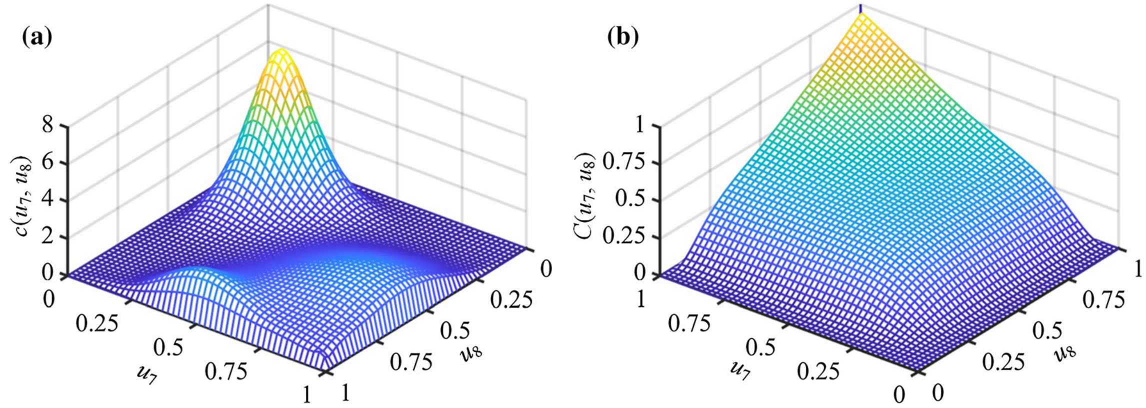


Fig. 8 Copula function obtained from the multivariate kernel estimation method: **a** density function, **b** distribution function

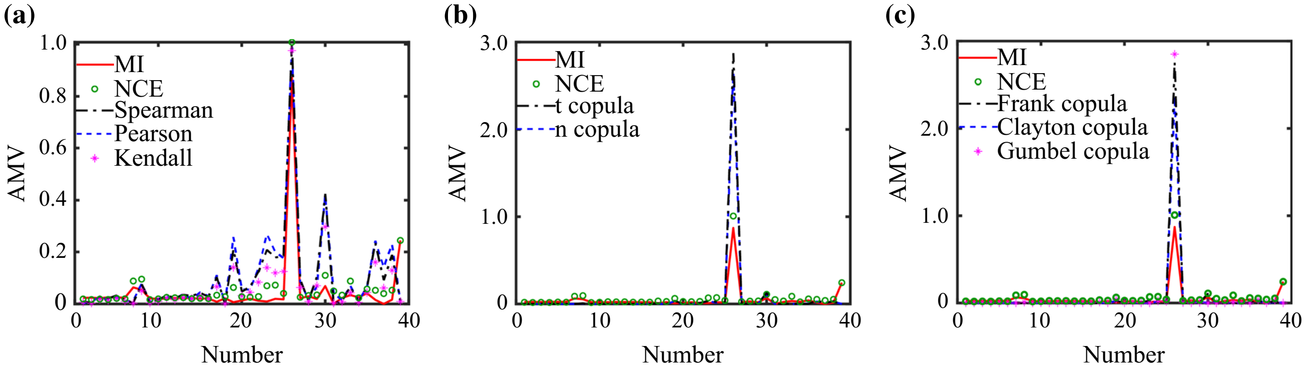


Fig. 9 Association measurement values for diesel engine manufacturing data obtained from different methods

Table 4 Comparison of association measurement RMSE for diesel engine manufacturing

AMV	MI	NCE	Spearman	Pearson	Kendall	t copula	n copula
RMSE	–	0.0339	0.1023	0.1119	0.0676	0.3212	0.2850
AMV	Frank copula	Clayton copula	Gumbel copula				
RMSE	0.3027	0.2414	0.3178				

the scatter plot of power versus its associated parameters is shown in Fig. 11. It can be seen from the figure that the first three parameters are approximately linear with power. However, the latter three parameters have complex nonlinear association with power, so it is difficult to find the obvious change trend in the visualization map.

The rated power of produced diesel engines is 254 kW. According to the deviation between the actual power and the rated power, it is stipulated that if the actual power falls within the range of 254 ± 3% kW (246.38–261.62 kW), the diesel engines are qualified products, and if the power deviation is more than ± 3%, the diesel engines are considered to be unqualified products. The original control limit of ST is [0, 40] and the original control limit of RT is [0, 8]. That is to say, ST has a wider range of control than RT. In practice, the regulation of ST will get more benefits. So it is an effective way to reduce unqualified product rate by adjusting the upper limit of ST. The trend of unqualified product rate with the upper limit of ST is shown in Fig. 12, which reflects clearly that unqualified product rate is the lowest (7.46%) when the upper limit of ST is 10 N·m.

Furthermore, based on the proposed method, association rules among assembly process parameters can be mined to realize the overall regulation of the system. Different from performance test parameters, assembly process parameters imply time sequence, so we can combine the causal network (see Fig. 13a) and temporal sequence to determine the impact mechanism of these parameters. As is shown in Fig. 13b, parameters above the diagonal of the heat map will affect the parameters below because The parameters at

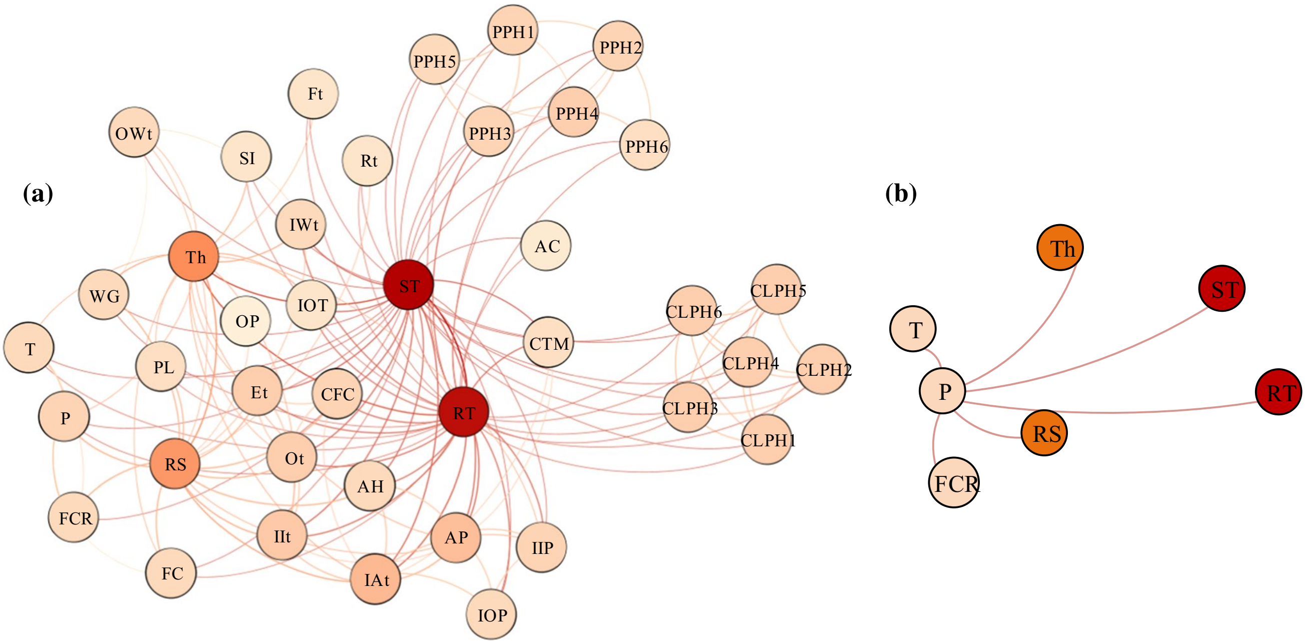


Fig. 10 Parameter association network of the diesel engine manufacturing system: **a** all parameters, **b** parameters associated with power

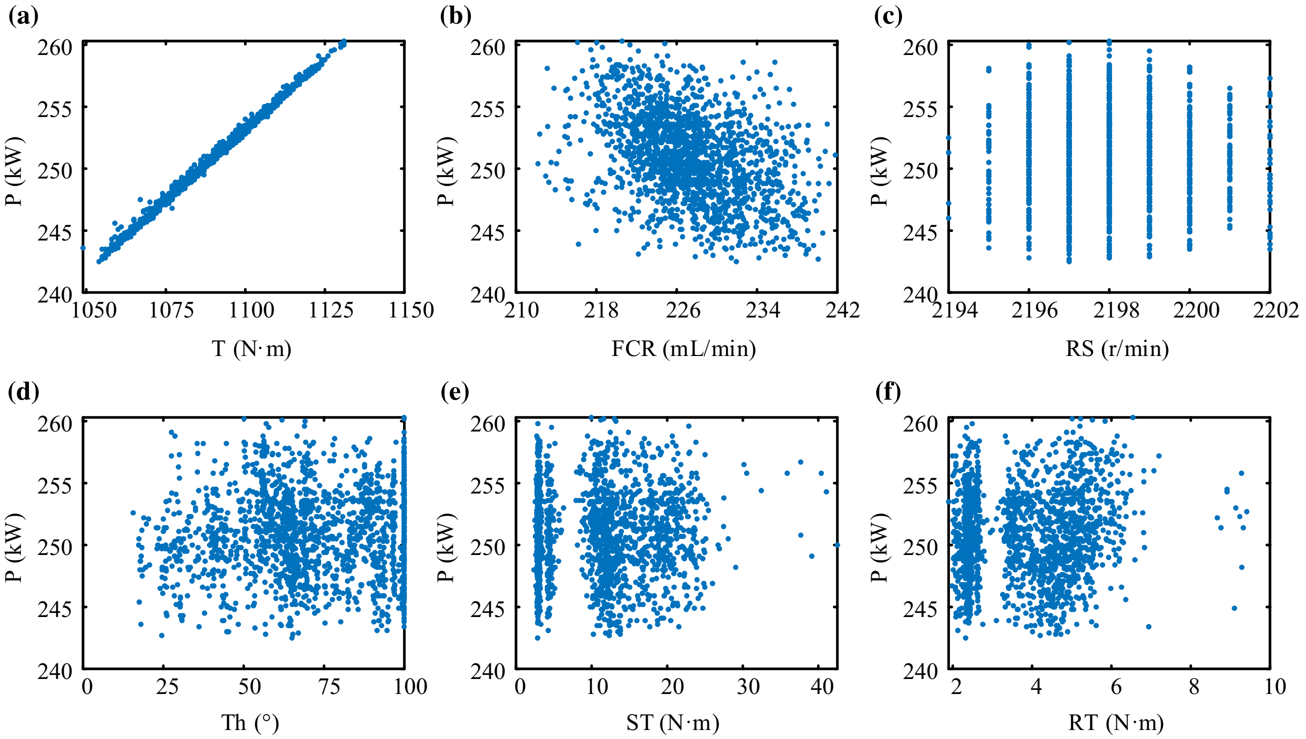


Fig. 11 Scatter plot of power versus its associated parameters

the top are generated before the parameters at the bottom. For example, the parameters associated with parameter 8 are 7, 10, and 13; parameter 7 is above the diagonal, and parameters 10 and 13 are below the diagonal, so parameter 8 is affected by 7 and affects 10 and 13. According to this method, the influence mechanism among arbitrary parameters with time series relation in a complex manufacturing system can be identified.

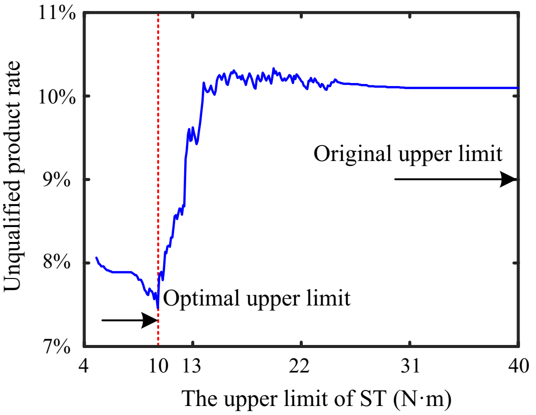


Fig. 12 The trend of unqualified product rate with the upper limit of ST

Conclusions and future works

With a data-driven approach, an effective causal discovery method combining nonparametric copula entropy and network deconvolution is developed to analyze a complex manufacturing system in this paper. Based on the copula theory and kernel density estimation, the nonparametric copula entropy is introduced to quantify association relationship and verified by comparing with the results of mutual information formula, 3 linear correlation coefficients and 5 parametric copula functions. By the network deconvolution method, a more accurate association network (i.e. causal network) is reconstructed. Furthermore, the proposed method is validated by using a gene expression dataset, and the application of the method in diesel engine manufacturing system is analyzed and discussed. Some detailed conclusions are summarized as follows:

(1) As an extension of mutual information theory, nonparametric copula entropy achieve high and reliable accuracy. Besides, Kendall correlation coefficient has the best performance in three linear correlation coefficients and Clayton copula has the best performance in the parametric copula-based method.
(2) The proposed method can be used to infer network structure and has a good performance. For the gene expression dataset, the linear correlation coefficient cannot detect the nonlinear relationships. Previous copula function form is not suitable, resulting in a greater loss of accuracy, and due to the accumulation of error caused by multiple integrals, the performance of mutual information is inferior to that of nonparametric copula entropy.
(3) For the diesel engine manufacturing system in this paper, starting and running torques are the two most important parameters for the whole system. The parameters associated with power are starting and running torques, throttle, rotational speed, fuel consumption rate and torque. Therefore, the power of diesel engine can be controlled by regulating these parameters.
(4) The causal network and temporal sequence relationship can be used to determine the impact mechanism of these parameters. Since the parameters at the top are generated before the parameters at the bottom, the parameters above the diagonal of the heatmap chart of association matrix will affect the parameters below. According to this principle, the influence mechanism among arbitrary parameters with time series relation in the complex manufacturing system can be identified.

As a preliminary study to discover the causality in complex manufacturing systems, many issues need to be further studied. Firstly, when the direct association matrix is transformed into a causal network structure, a more scientific and reasonable threshold setting method should be further

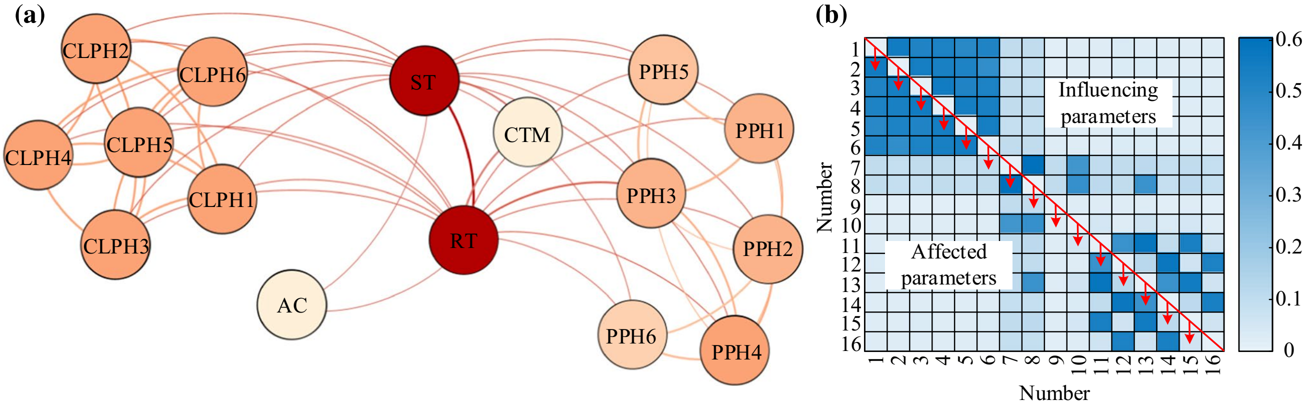


Fig. 13 Association relationship of assembly process parameters: **a** association network, **b** heatmap chart of association matrix

proposed. The complex network statistical characteristics-based method may be an effective way. Then, on the basis of this study, we will further propose an accurate causal directionality learning method. Structural equation method and transfer entropy are potential research directions. Lastly, more effective performance indicators need to be further developed for the evaluation of reliability and robustness of causal discovery in complex manufacturing systems. It is a good research idea to take the prediction effect of machine learning as the performance indicator.

Acknowledgements This project is supported by Shanghai Aerospace Science and Technology Innovation Fund (No. SAST2016048), and the National Natural Science Foundation (Grant Nos. 51435009 and 51775348).

References

Altay, G., & Emmert-Streib, F. (2010). Revealing differences in gene network inference algorithms on the network level by ensemble methods. *Bioinformatics, 26*(14), 1738–1744.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. In *Third International AAAI Conference on Weblogs and Social Media*.

Bowman, A. W., & Azzalini, A. (1997). *Applied smoothing techniques for data analysis*. Oxford University Press Inc.

Chen, L., & Guo, S. (2019). *Copulas and its application in hydrology and water resources*. Springer.

Chen, L., Ye, L., Singh, V., Zhou, J., & Guo, S. (2014). Determination of input for artificial neural networks for flood forecasting using the copula entropy method. *Journal of Hydrologic Engineering, 19*(11), 04014021–04014031.

Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory*. Wiley.

Croux, C., & Dehon, C. (2010). Influence functions of the Spearman and Kendall correlation measures. *Statistical Methods & Applications, 19*(4), 497–515.

Embrechts, P., Lindskog, F., & McNeil, A. (2001). Modelling dependence with copulas, rapport technique, *Dép. de Math. Inst. Féd. de Technol. de Zurich, Zurich.*

Fang, L., Zhao, H., Wang, P., Yu, M., Yan, J., Cheng, W., & Chen, P. (2015). Feature selection method based on mutual information and class separability for dimension reduction in multidimensional time series for clinical data. *Biomedical Signal Processing and Control, 21,* 82–89.

Feizi, S., Marbach, D., Médard, M., & Kellis, M. (2013). Network deconvolution as a general method to distinguish direct dependencies in networks. *Nature Biotechnology, 31*(8), 726.

Frigieri, E. P., Ynoguti, C. A., & Paiva, A. P. (2019). Correlation analysis among audible sound emissions and machining parameters in hardened steel turning. *Journal of Intelligent Manufacturing, 30*(4), 1753–1764.

Gu, Y. K., Fan, C. J., Liang, L. Q., & Zhang, J. (2019). Reliability calculation method based on the Copula *function* for mechanical systems with dependent failure. *Annals of Operations Research*. https://doi.org/10.1007/s10479-019-03202-5

Han, M., & Ren, W. (2015). Global mutual information-based feature selection approach using single-objective and multi-objective optimization. *Neurocomputing, 168,* 47–54.

Hu, L. (2006). Dependence patterns across financial markets: a mixed copula approach. *Applied Financial Economics, 16*(10), 717–729.

Huard, D., Évin, G., & Favre, A. C. (2006). Bayesian copula selection. *Computational Statistics & Data Analysis, 51*(2), 809–822.

Jeon, H. W., Lee, S., & Wang, C. (2019). Estimating manufacturing electricity costs by simulating *dependence* between production parameters. *Robotics and Computer-Integrated Manufacturing, 55,* 129–140.

Jones, M. C. (1993). Simple boundary correction for kernel density estimation. *Statistics and Computing., 3*(3), 135–146.

Khan, S., Bandyopadhyay, S., Ganguly, A. R., Saigal, S., Erickson, D. J., et al. (2007). Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. *Physical Review E, 76,* 1–15.

Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical Review E, 69*(6), 066138.

Marbach, D., Prill, R. J., Schaffter, T., Mattiussi, C., Floreano, D., & Stolovitzky, G. (2010). Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the National Academy of Sciences, 107*(14), 6286–6291.

Nicoloutsopoulos, D. (2005). Parametric and Bayesian non-parametric estimation of copulas. *Doctoral dissertation, University of London*.

Patton, A. J. (2002). Applications of Copula Theory in Financial Econometrics, *Ph.D. Dissertation, University of California*.

Qin, W., Zha, D., & Zhang, J. (2018). An effective approach for causal variables analysis in diesel engine production by using mutual information and network deconvolution. *Journal of Intelligent Manufacturing*, 1–11.

Rossi, F., Lendasse, A., François, D., Wertz, V., & Verleysen, M. (2006). Mutual information for the selection of relevant variables in spectrometric nonlinear modelling. *Chemometrics and intelligent laboratory systems, 80*(2), 215–226.

Scott, D. W. (2015). *Multivariate density estimation: theory, practice, and visualization*. Wiley.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27,* 379–423.

Shi, J., Zhao, J., Li, T., & Chen, L. (2019). Detecting direct associations in a network by information theoretic approaches. *Science China Mathematics, 62*(5), 823–838.

Silverman, B. W. (1986). Density estimation for statistics and data analysis. *Monographs on Statistics and Applied Probability*, 26.

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris, 8,* 229–231.

Sun, J. C., et al. (2017). Complex network construction of multivariate time series using information geometry. *IEEE Transactions on Systems, Man, and Cybernetics: Systems., 49*(1), 107–122.

Thomas, R. D., Moses, N. C., Semple, E. A., & Strang, A. J. (2014). An efficient algorithm for the computation of average mutual information: Validation and implementation in Matlab. *Journal of Mathematical Psychology, 61,* 45–59.

Wei, J., Pan, Z., Lin, X., Qin, D., Zhang, A., & Shi, L. (2019). Copula-function-based analysis model and dynamic reliability of a gear transmission system considering failure correlations. *Fatigue & Fracture of Engineering Materials & Structures, 42*(1), 114–128.

Xu, Y. (2005). Applications of Copula-based Models in Portfolio Optimization, *Ph.D. Dissertation, University of Miami*.

Zachariah, M., & Reddy, M. J. (2013). Development of an entropy-copula-based stochastic simulation model for generation of monthly inflows into the Hirakud Dam. *ISH Journal of Hydraulic Engineering, 19*(3), 267–275.

Zhang, X., Zhao, X. M., He, K., Lu, L., Cao, Y., Liu, J., et al. (2012). Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics, 28*(1), 98.

Zhang, X., Zhao, J., Hao, J. K., Zhao, X. M., & Chen, L. (2015). Conditional mutual inclusive information enables accurate quantification of associations in gene regulatory networks. *Nucleic Acids Research, 43*(5), e31–e31.

Zhao, N., & Lin, W. T. (2011). A copula entropy approach to correlation measurement at the country level. *Applied Mathematics and Computation, 218*(2), 628–642.