Hindawi
Mathematical Problems in Engineering
Volume 2018, Article ID 5840296, 10 pages
https://doi.org/10.1155/2018/5840296

Research Article

Semiparametric Estimation and Panel Data Clustering Analysis Based on D-Vine and C-Vine

Hong Li,[1] **Yuantao Xie**,[2] **Juan Yang**,[3] **and Di Wang**[1]

[1] *School of Economics, Peking University, China*
[2] *School of Insurance and Economics, University of International Business and Economics, China*
[3] *Institute of Comprehensive Development, Chinese Academy of Science and Technology for Development, China*

Correspondence should be addressed to Di Wang; surpassyourself@vip.163.com

Received 8 February 2018; Revised 13 June 2018; Accepted 26 July 2018; Published 18 September 2018

Academic Editor: Mahdi Jalili

This paper proposed a panel data clustering model based on D-vine and C-vine and supported a semiparametric estimation for parameters. These models include a two-step inference function for margins, two-step semiparameter estimation, and stepwise semiparametric estimation. In similarity measurement, similarity coefficients are constructed by a multivariate Hierarchical Nested Archimedean Copula (HNAC) model and compound PCC models, which are HNAC and D-vine compound model and HNAC and C-vine compound model. Estimation solutions and models evaluation are given for these models. In the case study, the clustering results of HNAC and D-vine compound model and HNAC and C-vine compound model are given, and the effect of different copula families on clustering results is also discussed. The result shows the models are effective and useful.

1. Introduction

When we use panel data to cluster, it is difficult to use the spatial correlation and cross-correlation of temporal dimension simultaneously. There is very little literature in the field. This paper would discuss how to build a multivariate copula model for panel data analysis, which would reflect the hierarchical structure relationship between time and index for realizing comprehensive evaluation and dynamic clustering.

1.1. Literature Review. In the traditional literature of data mining, the clustering method is divided into five methods: the method based on classification, on the hierarchy, on density, on the grid, and on the model. Frey et al. [1] also proposed a neighbor propagation algorithm based on clustering center.

Currently, hierarchy clustering methods are used for panel data analysis: Ren et al. [2], Juarez [3], Nie [4], Zhu et al. [5], Zheng et al. [6], LI et al. [7], Yang et al. [8] Zheng et al. [9], and Xie et al. [10].

There is also some authors' research based on models method. For example, Bartolucci [11] analyzed the binary panel data model; Zheng et al. [9] applied the traditional clustering method to panel data analysis by constructing a panel data matrix. Ren et al. [2] proposed a clustering method based on multi-index panel data by reestablishing the ward function based on the extended Frobenius criterion. Juárez [12] proposed a non-Gaussian panel data clustering method model based on the skew T distribution, which divided the cluster for the object by the dynamic behavior of the potential, equilibrium level and covariance effect of the regression and non-Gaussian model. The Nie [4] proposed a new clustering measurement, which can be used to calculate the weight flexibly. If the user paid more attention to the latest data, the more weight would be allocated.

A special case of vertical data would be proposed by panel data analysis, which also has been discussed by some authors. For example, Chiou et al. [13] proposed a clustering analysis of vertical data by establishing a nonparametric stochastic effect model based on Karhunen-Loeve expansion and nonparametric iterative mean-variance model.

De la Cruz-Mesia [14] proposed a model clustering method based on the measurement of cluster individuals over a period. With reference to the changes of data as the main feature in considering nonlinear layered model to the mixed layer model, the MCMC sampling method to explore the posterior distribution of the target which was formulated by the EM algorithm discussed the maximum likelihood estimation of the model.

Nielsen et al. [15] assumed that the count data follows the nonhomogeneous Poisson process whose intensity is a time-homogeneous and adaptive spline function. The spline functioned with smoothness and covariance of the change by time, which is also used to control the shape of the curve. And then an adaptive semiparametric longitudinal counting data clustering method is proposed. Jackknife, bootstrap, and pseudo-quasi methods were used as the parameter estimation method.

Shaikh et al. [16] applied the clustering method based on the model to cluster the missing vertical data. They used the mixed normal distribution and improved the Cholesky method for decomposing the covariance structure and then used the EM algorithm to estimate the parameters of the method model.

Yang et al. [17] proposed a panel data clustering analysis based on density. Xie et al. [18] proposed panel data clustering with affinity propagation and gave an agriculture risk regionalization analysis case. Guan et al. [19] gave MRI data analysis of affinity propagation clustering based on similarity matrix reduction.

The above vertical data clustering literature was based on the model clustering method in which assumption data and mixed models were applied to the vertical data clustering analysis. The merits of this method are its frequent user of all information and features of vertical data and the method with rigorous assumptions and statistical inferences. And its demerit is that, because of the data types of the vertical data with specific assumptions, such clustering methods are not widely applicable.

The above literature uses some extracting features of panel data to calculate similarity coefficient or distance, but not using overall characteristics of the panel data to consider how to calculate the similarity measure of panel data. Therefore, clustering of panel data could be studied on how to comprehensively measure its characteristics, proposed effective, and feasible method. In this paper, a classification method would be proposed to consider the multiple indexes of panel data and present a similarity coefficient, a clustering method based on density and near-neighbor propagation which were not used in the above literature. Based on the model clustering method, this paper would also present a composite PCC clustering method for the flexibility of the copula method (Genest et al. [20], Joe [21], and Bedford et al. [22, 23]).

1.2. The Innovation and Structure of the Study. According to the literature of panel data clustering, the purpose of panel data clustering mainly includes the following: to classify individual by clusters or classes; to classify indicator by clusters or classes; to find outliers or noise; and to classify individual by their shape characteristics, numerical characteristics, surface features, and another clustering purpose. Therefore, different clustering purposes are needed to propose different clustering methods. For example, some clustering purpose is to embody the overall characteristics of the data, or to reflect the hierarchical structure of the indicators, or to reflect the dynamic development characteristics of the data. Related state-of-the-art methods use some indicators to model the correlations in the data, of which its metrics are too single to extract complex dependency structures hidden in panel data, Therefore, it is necessary to propose the clustering method that can use comprehensive information and adapt to different clustering purposes. Copula method is a good approach.

In this paper, Pair-Copula Construction (PCC) would be discussed, which includes three types of models: HNAC model, the composite model of D-vine and HNAC, and the composite model of C-vine and HNAC. The third part of this paper discusses the statistical inference of composite PCC panel data clustering.

A general expression of composite PCC method would be presented first in this paper, and the parameters of the composite PCC method estimates would also be discussed, including the maximum likelihood method, marginal inference function of the two-stage estimate, semiparametric two stages, and semiparametric step-by-step; then the degree of fit and test would be discussed. The fourth part of this paper would also discuss the application of compound PCC method in panel data clustering and analyze panel data by clustering. And the final part would be the summary of the paper.

2. Model Building

2.1. The Construction of Multiple Copula. The construction of multiple copula mainly used EAC (the Exchangeable Archimedean copula) and NAC (the Nest Archimedean copula) (Joe [21]).

Joe [21] first proposed the structure of PCC, which is a new method and manifests as a waterfall structure similar to EAC and NAC. The difference of PCC, EAC, and NAC is that the PCC will deconstruct complex multivariate joint probability density into relatively simple multiple two-dimensional copula and marginal probability density; among them, the two-dimensional copula is not limited to Archimedes copula, also using any copula class, even mixed with a variety of copula classes. Bedford and Cooke [22, 23] provide a likelihood estimation method of PCC. Kurowicka et al. [24] simulated linear dependence by using partial correlation coefficient and determining the correlation coefficient matrix. Aas et al. [25] proposed a maximum pseudo-likelihood estimation method.

Berg and Aas [25] compared the EAC, HNAC, and PCC from max number of the copula, parameter constraint, and the choice of copula class. Conclusions are as follows: the EAC and NAC can only choose Archimedes copula class; parameters satisfy the constraint conditions; the scope of its application is restricted. The main advantage of PCC is that PCC is more flexible than NAC; there is no more $d(d-1)/2$ copula in PCC; the degrees of freedom of HNAC will decrease by the Mosaic layer increasing; therefore, the fitting effect of real data by the PCC method is better than by HNAC. Berg et al. [26] also compared the computational efficiency of HNAC and PCC. The results showed that the efficiency of PCC was higher than that of HNAC. See Table 1.

Aas et al. [25] pointed out the obvious deficiency of the PCC method of not having a unique structure. So Bedford and Cooke [22, 23] proposed a solution, the graph model

TABLE 1: The comparison of three ways of copula construction.

copula	Number of copulas, self-defined	Parameter constraint	Copula family
EAC	1	No	Archimedes copula
HNAC	d-1	$\theta_{ij} \geq \theta_{kj}\ i < k < d$	Archimedes copula
PCC	d(d-1)/2	No	Any copula

Here HNAC is a representative of NAC.
Here d represents the number of dimensions of the copula.

called Regular Vine (R-vine). But the structure of R-vine is still complex. Kurowicka et al. [24] presented two special forms of R-vine: Canonical Vines (C-vine) and Drawable Vines (D-vine). Aas et al. [25] First presented the application of vine copula in financial data, and the development of vine copula theory in the literature of Kurowicka et al. [24], Haff et al. [27], and Czado [28].

The R-vine method is used to solve the problem of multiple structures of the PCC; however, R-vine also has a variety of structures. C-vine and D-vine are two special forms of R-vine. Aas et al. [25] pointed out that, in the d dimension, C-vine has $d!/2$ kinds of structures and D-vine also has $d!/2$ structures, and finally made a conclusion of four variables, C-vine and D-vine structure and the joint probability density function. Bedford et al. [23] proposed an R-vine structure of five variables. Yang et al. [8] proposed the general formula of probability density function in the case of d dimension.

Brechmann et al. [29] analyzed the method of pruning at the top of the tree and discussed how to choose the independent copula. Smith et al. [30] selected independent copula and a given copula function in MCMC and model indexes. Yang et al. [8] proposed the sequence tree wise method to select the copula function.

2.2. The Dependency Structure of the Composite PCC Metric Panel Data. Pair-copula is suitable for the cross section data, Joe [21], Smith et al. [30], and Sun et al. [31]. Pair-copula was used in time series but rarely used in panel data and vertical data.

In order to reflect the dependency structure between indexes of panel data, this paper would build a composite method of PCC; the basic idea is that the upper (outer) uses HNAC structure, and PCC structure is nested in the lower (inner), including D-vine and HNAC composite model, and the composite model of C-vine and HNAC.

2.2.1. Notation. In this paper, the CDF of copula can be noted as C, while the pdf of copula can be noted as c. F is the CDF for united distributions or marginal distributions, while f is the pdf. Panel data is represented by $\mathbf{X}_{g_{i,n}}$, g for the unit (for example, Beijing for $g = 1$), the total number of unit for G (for example, for the country $G = 31$) contains the index number for $d_g = 4$, each unit for the g (i.e., 1 = health index, 2 = education index, 3 = life index, and 4 = social index), so C_g is g a set of common $g_1, \ldots, g_{d_g}$ indicators d_g (dimensions) of the dependency structure. The first variable g_i contains an observation n_g. For example, Beijing's education index $\mathbf{X}_{1_2}$ contains a total of observations $n_1 = 9$ from 2004 to 2012, specifically, $X_{1_{2,1}}$ (Beijing education 2004) to $X_{1_{2,9}}$ (Beijing education 2012).

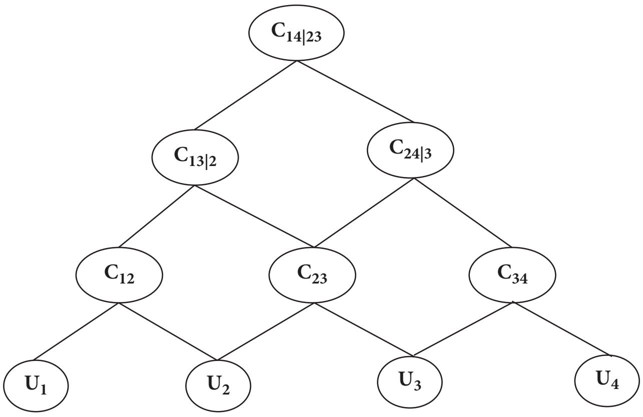


FIGURE 1: D-vine and HNAC construction.

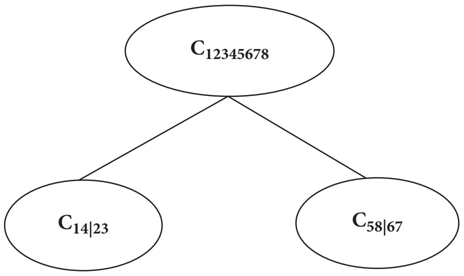


FIGURE 2: D-vine and HNAC structure.

2.2.2. Using D-Vine and HNAC to Measure Dependency Structure. The D-vine structure is used in the D-vine form with HNAC nesting. The D-vine structure is adopted in the subblock structure, as shown in Figure 1.

The HNAC structure is used in the superstructure of D-vine as shown in Figure 2.

2.2.3. Using C-Vine and HNAC to Measure Dependency Structure. The C-vine form with HNAC is used in the C-vine structure, and the subblock structure is adopted in C-vine structure, as shown in Figure 3.

The HNAC structure is used in the superstructure of C-vine, which is structured by $C_{14|23}$and $C_{58|67}$as shown in Figure 4.

3. Statistical Inference

3.1. Semiparametric Gradual Estimation. When the number of parameters of PCC is too much, the previously discussed estimation methods are less efficient ways. The more efficient estimation method would be discussed in the next.

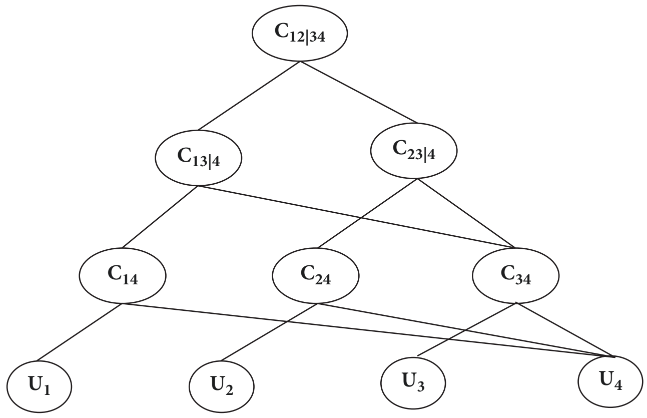


Figure 3: C-vine and HNAC construction.

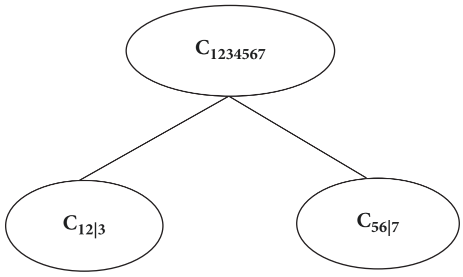


Figure 4: C-vine and HNAC structure.

Assuming that $\boldsymbol{\alpha}$ represents the parameter of the marginal distribution; $\boldsymbol{\theta}$ represents the parameter of the PCC copula; $\boldsymbol{\theta}^H$ are the dependent parameters of HNAC. The joint distribution can be expressed as

$$\begin{aligned} & f_{1_1,\ldots,1_{d_1},\ldots,G_1,\ldots,G_{d_G}}\left(x_{1_1},\ldots,x_{1_{d_1}},\ldots,x_{G_1},\ldots,x_{G_{d_G}};\boldsymbol{\alpha},\boldsymbol{\theta},\boldsymbol{\theta}^H\right) \\ & \quad = c_{1_1,\ldots,1_{d_1},\ldots,G_1,\ldots,G_{d_G}}\left(C_1,\ldots,C_G\right)\prod_{g=1}^{G} c_g \end{aligned} \tag{1}$$

where f is the pdf for united distributions as mentioned before. The probability density function of g is satisfied:

$$\begin{aligned} c_g &= f_{g_1,\ldots,g_{d_g}}\left(x_{g_1},\ldots,x_{g_{d_g}};\boldsymbol{\alpha},\boldsymbol{\theta}\right) \\ &= c_{g_1,\ldots,g_{d_g}}\left(F_{g_1}\left(x_{g_1};\boldsymbol{\alpha}_{g_1}\right),\ldots,F_{g_{d_g}}\left(x_{g_{d_g}};\boldsymbol{\alpha}_{g_{d_g}}\right);\boldsymbol{\theta}\right) \\ &\quad \cdot \prod_{i=1}^{d_g} f_{g_i}\left(x_{g_i};\boldsymbol{\alpha}_{g_i}\right) \end{aligned} \tag{2}$$

Defining $\boldsymbol{\theta}_{g_{i\longrightarrow i+j}} = \{\boldsymbol{\theta}_{g_{s,s+k|v_{sk}}} : (s, s+k) \in w_{ij}\}$, $\boldsymbol{\theta}_{g_{i\longrightarrow i}} = \varnothing$, $\boldsymbol{\theta}_{g_j} = \{\boldsymbol{\theta}_{g_{s,s+k|v_{sk}}} : |v_{sk}| = j-1\}$. $|\cdot|$ is the cardinality. And $\boldsymbol{\theta}_{g_j}$ is the data set of all parameters from g to i

For D-vine, the joint probability density function can be expressed as

$$\begin{aligned} c_g &= f_{g_1,\ldots,g_{d_g}}\left(x_{g_1},\ldots,x_{g_{d_g}};\boldsymbol{\alpha},\boldsymbol{\theta}\right) \\ &= \prod_{j=1}^{d_g-1}\prod_{i=1}^{d_g-j} c_{g_{i,i+j|v_{ij}}}\left(F_{g_{i|v_{ij}}}\left(x_{g_i} \mid \mathbf{x}_{g_{v_{ij}}};\boldsymbol{\alpha}_{g_{w_{i,j-1}}},\boldsymbol{\theta}_{g_{i\longrightarrow i+j-1}}\right),\right. \\ &\quad \left. F_{g_{i,i+j|v_{ij}}}\left(x_{g_{i+j}} \mid \mathbf{x}_{g_{v_{ij}}};\boldsymbol{\alpha}_{g_{w_{i+1,j-1}}},\boldsymbol{\theta}_{g_{i+1\longrightarrow i+j}}\right);\boldsymbol{\theta}_{g_{i,i+j|v_{ij}}}\right) \\ &\quad \times \prod_{i=1}^{d_g} f_{g_i}\left(x_{g_i};\boldsymbol{\alpha}_{g_i}\right) \end{aligned} \tag{3}$$

In the next analysis, the paper assumes that D-vine has similar formulas for C-vine and another vine. In the semiparametric two-stage estimation, the marginal distribution parameters of the logarithmic likelihood function are replaced with nonparametric. The semiparametric gradual estimation is similar to this idea, estimating the PCC parameters at a level by level. The log likelihood for a given sample can be expressed as

$$\begin{aligned} & \varphi_j\left(u_{1_1},\ldots,u_{G_{d_G}};\boldsymbol{\theta}_1\cdots,\boldsymbol{\theta}_j\right) \\ & \quad = \sum_{g=1}^{G}\sum_{i=1}^{d_g-j}\log\left\{c_{g_{i,i+j|v_{ij}}}\left[e_{g_{i,i+j}}\left(u_{g_i},\ldots,u_{g_{i+j-1}};\boldsymbol{\theta}_{g_{i\longrightarrow i+j-1}}\right),h_{g_{i,i+j}}\left(u_{g_{i+1}},\ldots,u_{g_{i+j}};\boldsymbol{\theta}_{g_{i+1\longrightarrow i+j}}\right);\boldsymbol{\theta}_{g_{i,i+j|v_{ij}}}\right]\right\} \end{aligned} \tag{4}$$

where $i = 1,\ldots,d_g - j$, $j = 1,\ldots,d_g - 1$.

The pseudo-likelihood function of the dependent part can be written as

$$\begin{aligned} & pl_c(\boldsymbol{\theta};\mathbf{x}) \\ & \quad = \sum_{k=1}^{n_g}\log\left[c_{1_1,\ldots,G_{d_G}}\left(F_{1_{1,n_g}}\left(x_{1_{1,k}}\right),\ldots,F_{G_{d_G,n_g}}\left(x_{G_{d_G,k}}\right);\boldsymbol{\theta}\right)\right] \\ & \quad = \sum_{k=1}^{n_g}\sum_{l=1}^{j}\varphi_l\left(F_{1_{1,n_g}}\left(x_{1_{1,k}}\right),\ldots,F_{G_{d_G,n_g}}\left(x_{G_{d_G,k}}\right);\boldsymbol{\theta}_1\cdots,\boldsymbol{\theta}_l\right) \end{aligned} \tag{5}$$

By gradually substituting parameters, normal functions can be constructed and solved $\widehat{\boldsymbol{\theta}}^{SSP}$. The normal equations are

$$\begin{aligned} & \sum_{k=1}^{n_g}\Delta_{(j-1)(d_g-j/2)+i}^{SSP}\left(F_{1_{1,n_g}}\left(x_{1_{1,k}}\right),\ldots,F_{G_{d_G,n_g}}\left(x_{G_{d_G,k}}\right);\right. \\ & \quad \left.\widehat{\boldsymbol{\theta}}^{SSP}\right) = \mathbf{0} \end{aligned} \tag{6}$$

where $i = 1, \ldots, d_g - j$, $j = 1, \ldots, d_g - 1$.

$$\begin{aligned}\Delta^{SSP}_{(j-1)(d_g-j/2)+i}&\left(u_{1_1}, \ldots, u_{G_{d_G}}; \boldsymbol{\theta}_1 \cdots, \boldsymbol{\theta}_j\right)\\ &= \frac{\partial\varphi_j\left(u_{1_1}, \ldots, u_{G_{d_G}}; \boldsymbol{\theta}_1 \cdots, \boldsymbol{\theta}_j\right)}{\partial\boldsymbol{\theta}_{g_{i,i+j|v_{ij}}}}\end{aligned} \tag{7}$$

$\widehat{\boldsymbol{\theta}}^{SSP}$ is plugged into the likelihood function, estimating $\widehat{\boldsymbol{\theta}}^{H\ SSP}$. To be specific, the steps of gradual estimation are as follows.

Step 1. Estimate $\widehat{\boldsymbol{\theta}}_1^{SSP}$, the first level of dependency parameters.

Step 2. Plug $\widehat{\boldsymbol{\theta}}_1^{SSP}$ into (5). Then obtain $\widehat{\boldsymbol{\theta}}_2^{SSP}$, the dependent parameters of the second level, by estimating the maximum likelihood method.

Step 3. Take parameters $\widehat{\boldsymbol{\theta}}_1^{SSP}$ and $\widehat{\boldsymbol{\theta}}_2^{SSP}$ in (5), and estimate the dependent parameters $\widehat{\boldsymbol{\theta}}_3^{SSP}$.

Step 4. Repeat the above steps, estimating $\widehat{\boldsymbol{\theta}}_j^{SSP}$ by $\widehat{\boldsymbol{\theta}}_1^{SSP}, \ldots, \widehat{\boldsymbol{\theta}}_{j-1}^{SSP}$.

Step 5. Plug $\widehat{\boldsymbol{\theta}}^{SSP}$ into the likelihood function, and estimate $\widehat{\boldsymbol{\theta}}^{H\ SSP}$ and all dependent parameters.

Haff [29] proved that the semiparametric gradually estimation has consistent, progressive, and normal and robustness characters. In this paper, the good properties of the estimations are as follows:

(1) $\begin{pmatrix}\widehat{\boldsymbol{\theta}}\\ \widehat{\boldsymbol{\theta}}^H\end{pmatrix}^{SSP}$ is the consistent estimation of$\begin{pmatrix}\boldsymbol{\theta}\\ \boldsymbol{\theta}^H\end{pmatrix}$;

(2) $\begin{pmatrix}\widehat{\boldsymbol{\theta}}\\ \widehat{\boldsymbol{\theta}}^H\end{pmatrix}^{SSP}$ has asymptotic normality:

$$\begin{aligned}&\sqrt{n}\left(\begin{pmatrix}\widehat{\boldsymbol{\theta}}\\ \widehat{\boldsymbol{\theta}}^H\end{pmatrix}^{SSP} - \begin{pmatrix}\boldsymbol{\theta}\\ \boldsymbol{\theta}^H\end{pmatrix}\right)\\ &\quad\xrightarrow{d} N\left(\mathbf{0}, \mathbf{J}_\Theta^{-1}\mathbf{K}_\Theta\left(\mathbf{J}_\Theta^{-1}\right)^T + \mathbf{J}_\Theta^{-1}\mathbf{L}_\Theta\left(\mathbf{J}_\Theta^{-1}\right)^T\right)\end{aligned} \tag{8}$$

where

$$\Theta = \begin{pmatrix}\boldsymbol{\theta}\\ \boldsymbol{\theta}^H\end{pmatrix} \tag{9}$$

$$\begin{aligned}\mathbf{K}_\Theta &= E\left[\Delta^{SSP}\left(\Delta^{SSP}\right)^T\right]\\ &= \begin{bmatrix}\mathbf{K}_{\Theta,1_1,1_1} & \cdots & \cdots & \mathbf{0}\\ \vdots & \ddots & \vdots & \vdots\\ \mathbf{0} & \cdots & \mathbf{K}_{\Theta,G_{d_G}-2,G_{d_G}-2} & \mathbf{0}\\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{K}_{\Theta,G_{d_G}-1,G_{d_G}-1}\end{bmatrix}\end{aligned} \tag{10}$$

$$\begin{aligned}\mathbf{J}_\Theta &= E\left[\Delta^{SSP}\left(\Delta^{SSP}\right)^T\right]\\ &= \begin{bmatrix}\mathbf{J}_{\Theta,1_1,1_1} & \cdots & \cdots & \mathbf{0}\\ \vdots & \ddots & \vdots & \vdots\\ \mathbf{0} & \cdots & \mathbf{J}_{\Theta,G_{d_G}-2,G_{d_G}-2} & \mathbf{0}\\ \mathbf{I}_{\Theta,G_{d_G}-1,1_1} & \cdots & \mathbf{I}_{\Theta,G_{d_G}-1,G_{d_g}-2} & \mathbf{I}_{\Theta,G_{d_G}-1,G_{d_G}-1}\end{bmatrix}\end{aligned} \tag{11}$$

$$\begin{aligned}\mathbf{B}_i\left(\mathbf{U}; \boldsymbol{\theta}, \boldsymbol{\theta}^H\right) &= \int \frac{\partial^2 \log c_{1_1,\ldots,G_{d_G}}\left(\mathbf{u}; \boldsymbol{\theta}, \boldsymbol{\theta}^H\right)}{\partial\Theta\partial u_i} I\left(U_i\right.\\ &\quad\left.\le u_i\right) dC_{1,\ldots,d}\left(\mathbf{u}; \boldsymbol{\theta}, \boldsymbol{\theta}^H\right)\end{aligned} \tag{12}$$

$$\begin{aligned}\mathbf{L}_\Theta &= \operatorname{var}\left[\sum_{l=1}^{G_{d_G}} \mathbf{B}_l\left(\mathbf{U}; \boldsymbol{\theta}, \boldsymbol{\theta}^H\right)\right]\\ &\quad+ \sum_{l=1}^{G_{d_G}} \operatorname{cov}\left[\frac{\partial \log c_{1_1,\ldots,G_{d_G}}\left(\mathbf{U}; \boldsymbol{\theta}, \boldsymbol{\theta}^H\right)}{\partial\Theta}, \mathbf{B}_l\left(\mathbf{U}; \boldsymbol{\theta}, \boldsymbol{\theta}^H\right)\right]\end{aligned} \tag{13}$$

(3) $\begin{pmatrix}\widehat{\boldsymbol{\theta}}\\ \widehat{\boldsymbol{\theta}}^H\end{pmatrix}^{SSP}$ are robustness estimates.

In practical analysis, this paper considers firstly implementing semi-ginseng gradual regression, then plugging the estimate of the parameter $\begin{pmatrix}\widehat{\boldsymbol{\theta}}\\ \widehat{\boldsymbol{\theta}}^H\end{pmatrix}^{SSP}$ as the initial value into the maximum likelihood estimation method, and finally getting the parameter estimator.

3.2. Model Evaluation. Front, parameter estimation method has been proposed for the same set of data; the difference of the results of using different marginal distribution or using a different copula is very large. The problem with this paper is how to compare fitting effects between different marginal distribution and different copula. Further discussions in the paper are model evaluation comparison and fitting optimization test. A Hit test would be discussed in this session, which does not only compare with the figure and numerical value but also gives a test; the traditional information criterion evaluation criteria would be used to intuitively compare the fitting effect of different combinations, and the fitting optimization test would be discussed later.

3.2.1. Hit Test. When using multiple marginal distributions or different copula, the fitting test will be different from the traditional fitting degree test. The original Hit test method is adopted in this paper. First, divide the interval into known regions, as shown in Figure 5. In contrast to the Hit inspection of Pattern [32], the paper adopts the overall Hit test for the shaded part of Figure 5; different axes represent different marginal distribution function. For two-dimensional case, the chessboard form is shown above. For the three-dimensional case, a cubic (cube) is obtained. The high dimension is corresponding to the hypercube (super cube).

The specific steps are as follows.

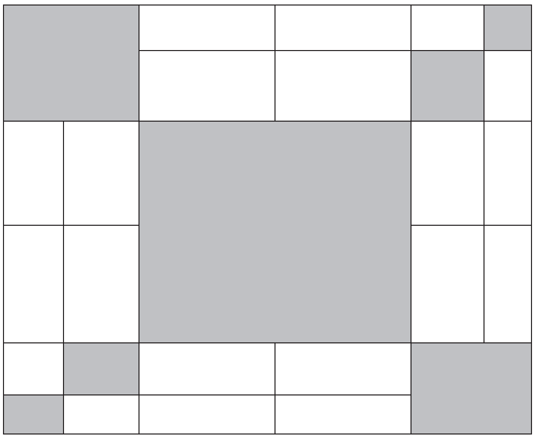


Figure 5: Sample area map.

Step 1. Divide experience data by the above critical value. Each data point is only in a subtable (a subcube or a hypercube). Calculate $N_{1_1,\ldots,G_{d_G}}$, the number of data points on each subtable (a subcube or a hypercube).

Step 2. With reference to the previous section method to estimate the marginal distribution parameters, and according to these parameters, calculate $n_{1_1,\ldots,G_{d_G}}$, the number of fitting data points falling in each subtable (a subcube or a hypercube).

Step 3. Construct the Hit test according to the number of empirical data points and the number of fitting data points.

$$\sum \frac{\left(N_{1_1,\ldots,G_{d_G}} - n_{1_1,\ldots,G_{d_G}}\right)^2}{n_{1_1,\ldots,G_{d_G}}} \sim \chi^2_\alpha (N-k-1) \tag{14}$$

where N is the number of the total sample observation points and k is the number of estimated parameters.

3.2.2. Model Evaluation. Traditional statistical analysis indicators include information statistics and entropy criterion. Only information criterion is discussed in this paper, and other criteria can be calculated by information criterion.

The correction of AIC criterion (AICC, abbreviated as AICC) can be used to balance the optimal degree of fitting and the number of parameters. The paper proposes the statistics of HQC (Hannan-Quinn information criterion, abbreviated as HQC). The maximum likelihood estimation is calculated as follows:

$$\begin{aligned} HQC^{ML} &= -2\log\left[c_{1_1,\ldots,1_{d_1},\ldots,G_1,\ldots,G_{d_G}}\left(C_1,\ldots,C_G\right)\right] \\ &\quad - 2\sum_{g=1}^{G}\sum_{k=1}^{n_g}\sum_{j=1}^{d_g-1}\sum_{i=1}^{d_g-j}\log\left\{c_{g_{i,i+j|v_{ij}}}\left(F_{g_{i|v_{ij}}}\left(x_{g_i} \mid \mathbf{x}_{g_{v_{ij}}}; \widehat{\boldsymbol{\alpha}}_{g_{w_{i,j-1}}}, \widehat{\boldsymbol{\theta}}_{g_{i\longrightarrow i+j-1}}\right), F_{g_{i,i+j|v_{ij}}}\left(x_{g_{i+j}} \mid \mathbf{x}_{g_{v_{ij}}}; \widehat{\boldsymbol{\alpha}}_{g_{w_{i+1,j-1}}}, \widehat{\boldsymbol{\theta}}_{g_{i+1\longrightarrow i+j}}\right); \widehat{\boldsymbol{\theta}}_{g_{i,i+j|v_{ij}}}\right)\right\} \\ &\quad - 2\sum_{g=1}^{G}\sum_{k=1}^{n_g}\sum_{i=1}^{d_g}\log\left[f_{g_i}\left(x_{g_{ik}}; \widehat{\boldsymbol{\alpha}}_{g_i}\right)\right] + 2\log\left(\log\left(N\right)\right)k \end{aligned} \tag{15}$$

For semiparametric gradual estimation, HQC is calculated as

$$\begin{aligned} HQC^{SSP} &= -2\log\left[c_{1_1,\ldots,1_{d_1},\ldots,G_1,\ldots,G_{d_G}}\left(C_1,\ldots,C_G\right)\right] - 2\sum_{g=1}^{G}\sum_{k=1}^{n_g}\sum_{l=1}^{j}\varphi_l\left(F_{1_{1,n_g}}\left(x_{1_{1,k}}\right),\ldots,F_{G_{d_G,n_g}}\left(x_{G_{d_G,k}}\right); \widehat{\boldsymbol{\theta}}^{SSP}_{g_1}\cdots,\widehat{\boldsymbol{\theta}}^{SSP}_{g_l}\right) \\ &\quad - 2\sum_{g=1}^{G}\sum_{k=1}^{n_g}\sum_{i=1}^{d_g}\log\left[f_{g_i}\left(x_{g_{ik}}\right)\right] + 2\log\left(\log\left(N\right)\right)k \end{aligned} \tag{16}$$

The constructed information criterion in the front is only a numerical value, which cannot be tested in its own way. So there are many disadvantages. In practice, a test of goodness of fit is

$$\begin{aligned} S &= n\int_{[0,1]^{\sum_{g=1}^{G} g}} \left\{\widehat{C}\left(\mathbf{u}\right) - C_{\widehat{\boldsymbol{\theta}}}\left(\mathbf{u}\right)\right\}^2 d\widehat{C}\left(\mathbf{u}\right) \\ &= \sum_{j=1}^{n}\left\{\widehat{C}\left(\mathbf{U}_j\right) - C_{\widehat{\boldsymbol{\theta}}}\left(\mathbf{U}_j\right)\right\}^2 \end{aligned} \tag{17}$$

more effective.

$$\begin{aligned} \widehat{C}\left(\mathbf{u}\right) &= \frac{1}{n+1}\sum_{j=1}^{n} 1\left(U_{1_1 j} \le u_{1_1},\ldots,U_{G_{d_G} j} \le u_{G_{d_G}}\right)\mathbf{u} \\ &= \left(u_1,\ldots,u_d\right) \in (0,1)^d \end{aligned} \tag{18}$$

Table 2: Models evaluation and goodness of fit with different Archimedes copula.

Copula	HNAC		PCC with D-vine and HNAC		PCC with C-vine and HNAC	
	HQC	S	HQC	S	HQC	S
Gumbel	-150.6136	85.3 (0.1802)	-151.1203	84.2 (0.1921)	-151.1039	80.7 (0.2025)
Clayton	-150.5281	86.1 (0.1632)	-151.0218	84.4 (0.1805)	-150.9812	84.6 (0.1791)

4. The Empirical Analysis

4.1. Introduction to the Panel Data Clustering. The China Development Index (RCDI) is compiled by the China Survey and data center of Renmin University of China. The index is composed of four indices of health, education, economy, and social environment, and 15 subindexes. The development of the states and the 31 regions since 2004 has been comprehensively measured. According to the similarity matrix of composite pair-copula, the proper clustering algorithm is selected and the rationality of the clustering algorithm is tested and compared.

According to the similarity matrix, the similarity coefficient is transformed into $(0, 1)$, and we select the corresponding clustering algorithm, such as the ward method commonly used in the multivariate statistics, hierarchical clustering method.

4.2. The Empirical Analysis. The following is an empirical analysis of China's development index panel data.

This paper gives the HNAC clustering results and composite method of PCC clustering results to contrast HNAC and composite method of PCC for the effect of panel data clustering; among them, the composite PCC method discussed C-vine and HNAC composite model, as well as D-vine and HNAC composite model.

In order to make the results of this paper comparable, the upper HNAC Archimedes copula function chose Gumbel copula. The PCC section also uses Gumbel copula. It is a way to unify and ensure the comparability of analysis.

4.2.1. Selection of Different Copula Functions. As mentioned, the Gaussian copula is applicable to data without tail dependencies; the Clayton copula is applicable to the data of the bottom tail; Gumbel copula applies to the data on the tail; Student copula is applied to data of the bottom tail and on the tail.

Since each copula uses a different range, the coefficient of fitting will also be different, and its structure will change, which may influence the clustering result. Therefore, we need to combine the Hit test and the fitting method to select the most suitable copula.

If different copula is selected, the fitting effect of the three methods will be different. The above section has discussed the model of evaluation and test of goodness of fit; this paper mainly chooses HQC index and S inspection; considering that the value of S depends on $\widehat{\theta}$, simulate 2000 times to give an approximate estimate of the adjoint probability in this paper by using Monte Carlo simulation.

As can be seen from Table 2, the Gumbel copula model works well, so Gumbel copula is selected for the following analysis.

4.2.2. The Clustering Results of HNAC. Taking Beijing and Shanghai as an example, the copula dependency relationship after data processing is correct in Figures 6 and 7.

The distribution $C_{1_{1,2,3,4}2_{1,2,3,4}}$ is shown in Figure 7.

The HNAC and composite PCC methods depend on parameter estimation.

The dependent parameters are processed by the unit, and then the clustering analysis; the results of the clustering are drawn in Figure 8.

As can be seen from Figure 8,s Beijing and Shanghai are classified together, which are very different from any other provinces. The eastern coastal areas (except Fujian province and Hebei province), including the northeast border area, are obviously a category; Hainan and Xinjiang are the third provinces, and some of the provinces in the west are the fourth provinces. The correlation between regions is reflected.

4.2.3. C-Vine and HNAC Composite PCC Clustering Results. The dependent parameters are processed by the unit, and then the clustering analysis; the results of the clustering are drawn as shown in Figure 9.

As can be seen from Figure 9. Beijing and Shanghai are one class, which is different from any other province. The eastern coastal areas (except Fujian province and Hebei province), including the northeast border area, are obviously a category; Hainan provinces and Xinjiang (excluding Guizhou) are the third provinces, and some provinces in the west are the fourth provinces. The correlation between regions is reflected.

Through graphics contrast, it can be seen that in front of the C-vine and HNAC composite model relative to the D-vine and HNAC composite model is more similar, the results of the two models can reflect the various provinces and regions section, and the different time sequence on inertia of economics, and the characteristic of the composite panel data dependency structure can be clearly reflected; for panel data clustering, the effect is very stable.

5. Conclusions

Among these models, the composite model of D-vine and HNAC and the compound model of C-vine and HNAC are collectively referred to as the composite PCC method. The composite PCC method is the new structure proposed in this paper, which can reflect the hierarchical structure of panel data indexes.

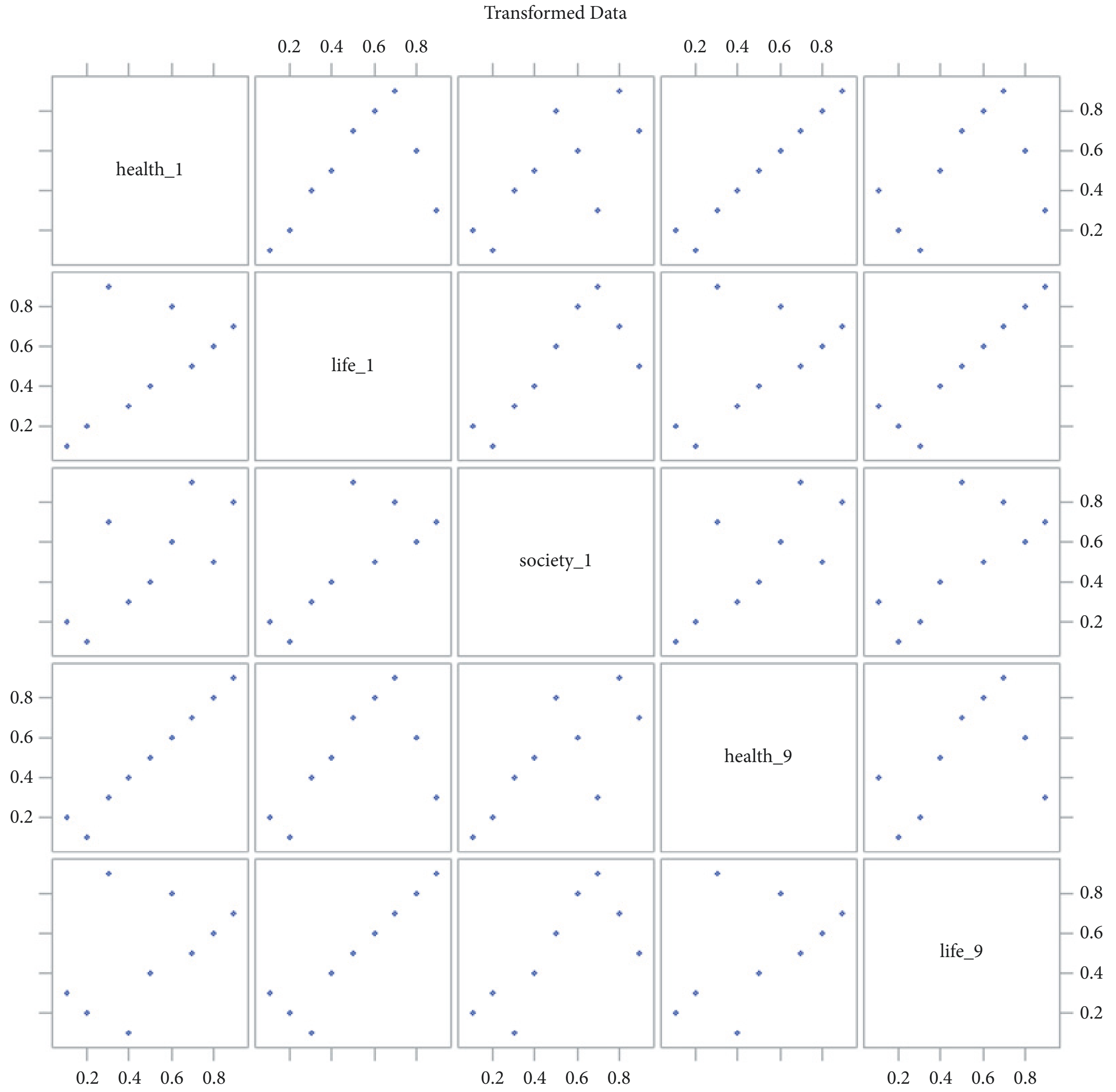


Figure 6: Sample dependency graph.

In PCC, the conditional probability functions used by D-vine, C-vine, and R-vine are different, so there is a huge difference in independence structure. D-vine is appropriate for variable equivalence (similar to the exchangeable order, see Barbe [33]), such as education, economic index, social index, and index of life, the four partial comparative equivalences (intuitively, the number of copula links to each variable is equal); but for the C-vine, the number of links of copulas connects different variables, four indexes were put in position which need to be well designed and argued, the number of copulas linked to each variable in R-vine is different, and there is a certain sequence. Refer to Rehman [34] for research on nonparametric estimating abundance.

In this paper, a panel data clustering method based on composite PCC is proposed to summarize this method according to the evaluation of the clustering algorithm.

(1) Based on the method of composite PCC, it has better scalability. When the estimated parameters are large and the data volume is too large, the operation speed drops.

(2) Some clustering algorithms are sensitive to parameters, and the parameters are more difficult to be determined for some data sets with a large number of the unit. Such clustering algorithm is not practical.

Based on the method of compound PCC, there are too many parameters in all marginal distribution, conditional distribution, and dependent structure, so it is difficult to bring parameter estimation and hypothesis testing.

(3) Some clustering algorithms are sensitive to noise data, and such clustering methods are not useful, but we

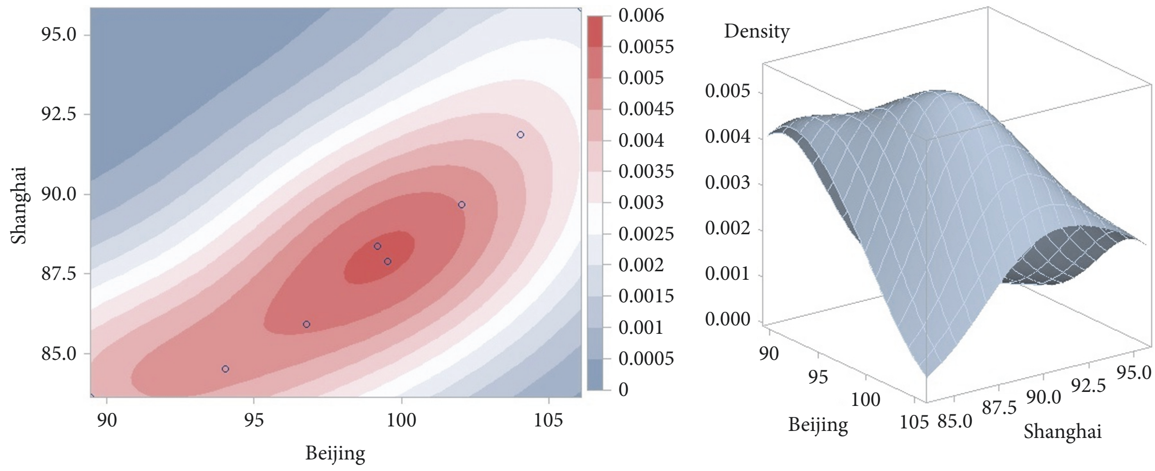


FIGURE 7: Distribution $C_{1_{1,2,3,4}2_{1,2,3,4}}$ and 3D density map.

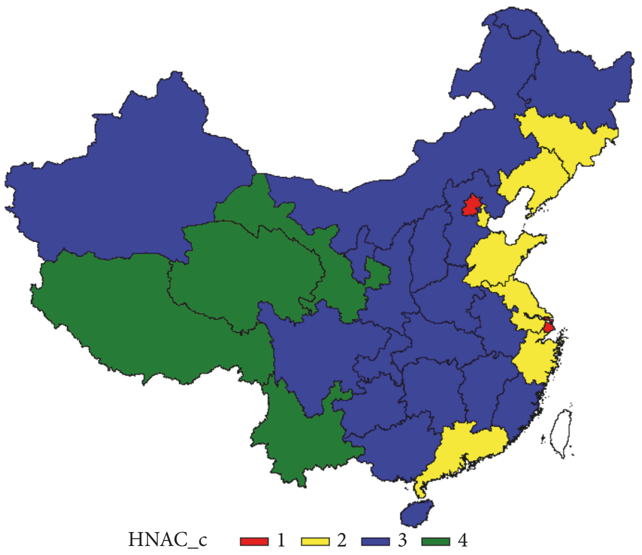


FIGURE 8: D-vine and HNAC composite PCC method clustering results.

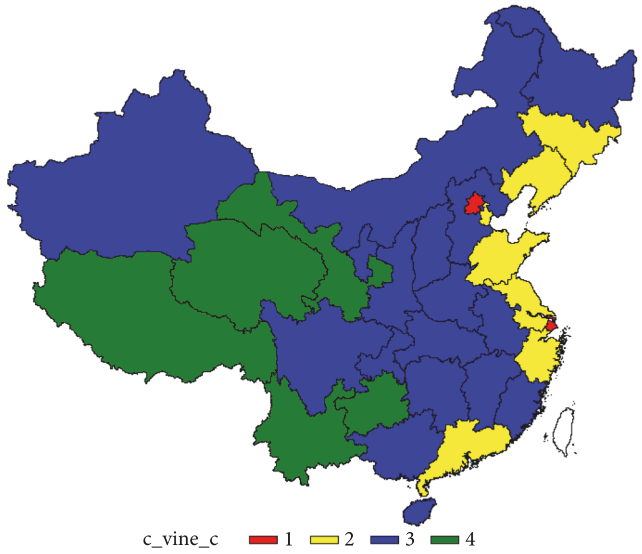


FIGURE 9: C-vine and HNAC composite PCC method clustering results.

sometimes need to be able to recognize the clustering algorithm of noise. The sensitivity to noise data for the method of compound PCC depends on the copulas connected and marginal distribution. The result can be sensitive to noise data and also cannot be sensitive, depending on the construction.

(4) From the input data order: some clustering algorithms are sensitive to the order of input data; such clustering algorithm is not practical. The above clustering method is not sensitive to the input data order. But the setting of the dependent structure itself is sequential, depending on the analyst's understanding and grasp of the problem.

(5) From processing high-dimensional data: in the field of genetics and biology, or in the data set of e-commerce, the number of observations is often far less than the number of indicators (variables or attributes). Based on the method of compound PCC, it is easy to set the complex structure with high computational complexity when dealing with high dimension. It is sensitive to an initial value and it is convergent to the local optimal solution or even does not converge.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors hereby declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The paper was financially supported by the National Social Science Fund of China "Research on the Redistribution Function of Social Security System: Research on the National Social Security Fund's Intervention in Pension Insurance Payment" (18BJY212) and "the Fundamental Research Funds for the Central Universities" in UIBE (CXTD9-04).

References

[1] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[2] J. Ren and S.-L. Shi, "Multivariable panel data ordinal clustering and its application in competitive strategy identification of appliance-wiring listed companies," in *Proceedings of the 2009 16th International Conference on Management Science and Engineering, ICMSE 2009*, pp. 253–258, Russia, September 2009.

[3] M. A. Juárez and M. F. Steel, "Model-based clustering of non-Gaussian panel data based on skew-t distributions," *of Business Economic Statistics*, vol. 28, no. 1, pp. 52–66, 2010.

[4] G. Nie, Y. Chen, L. Zhang, and Y. Guo, "Credit card customer analysis based on panel data clustering," *Computer Science*, vol. 1, no. 1, pp. 2489–2497, 2010.

[5] J. H. Niu, "The Cluster Analysis of Multivariable Panel Data and its Application," *Applied Mechanics and Materials*, vol. 220-223, pp. 2668–2671, 2012.

[6] B. Zheng, "The clustering analysis of multivariable panel data and its application," *Application of Statistics and Management*, vol. 27, no. 2, pp. 265–270, 2008.

[7] J. Ai, R. Zhang, Y. Li et al., "Circulating microRNA-1 as a potential novel biomarker for acute myocardial infarction," *Biochemical and Biophysical Research Communications*, vol. 391, no. 1, pp. 73–77, 2010.

[8] J. Yang, Y. T. Xie, and Y. B. Guo, "Panel Data Clustering Analysis based on Composite PCC: a Parametric Approach," *Cluster Computing*, vol. 2, pp. 1–11, 2018.

[9] T. Zheng, D. Zhu, X. Wang, and B. Yu, "Panel Data Clustering and its Application to Discount Rate," in *Proceedings of the of B Stock in China. Information and Computing Science, 2009*, vol. 1, pp. 163–166, 2009.

[10] Y. T. Xie, Z. X. Li, and R. Parsa, "Extension and Application of Credibility Models in Predicting Claim Frequency," *Mathematical Problems in Engineering*, vol. 2018, Article ID 6250686, 8 pages, 2018.

[11] F. Bartolucci and V. Nigro, "Maximum likelihood estimation of an extended latent Markov model for clustered binary panel data," *Computational Statistics & Data Analysis*, vol. 51, no. 7, pp. 3470–3483, 2007.

[12] M. A. Jußrez, M. F. J. Steel, and M. A. Juárez, "Non-Gaussian dynamic Bayesian modeling for panel data," *Mpra Paper*, vol. 25, no. 7, pp. 1128–1154, 2006.

[13] J. M. Chiou and P. L. Li, "Functional clustering and identifying substructures of longitudinal data," *Journal of the Royal Statistical Society B: Statistical Methodology*, vol. 69, no. 4, pp. 679–699, 2007.

[14] R. De la Cruz-Mesía, F. A. Quintana, and G. Marshall, "Model-based clustering for longitudinal data," *Computational Statistics & Data Analysis*, vol. 52, no. 3, pp. 1441–1457, 2008.

[15] J. D. Nielsen and C. B. Dean, "Adaptive functional mixed NHPP models for the analysis of recurrent event panel data," *Computational Statistics & Data Analysis*, vol. 52, no. 7, pp. 3670–3685, 2008.

[16] M. Shaikh, P. D. McNicholas, and A. F. Desmond, "A pseudo-EM algorithm for clustering incomplete longitudinal data," *The International Journal of Biostatistics*, vol. 6, no. 1, 2010.

[17] J. Yang and Y. T. Xie, "Panel Data Clustering Analysis Based On Density," *Statistics & Information Forum*, 2014.

[18] "Agriculture Risk Regionalization Analysis Based on Panel Data Clustering with Affinity Propagation," *Statistics & Information Forum*, 2017.

[19] X. Guan, W. Zeng, and N. Wang, "MRI Data Analysis of Affinity Propagation Clustering Based on Similarity Matrix Reduction," *Computer Engineering*, 2016.

[20] C. Genest and J. Mackay, "The joy of copulas: Bivariate distributions with uniform marginals," *The American Statistician*, vol. 40, no. 4, pp. 280–283, 1986.

[21] H. Joe, *Multivariate Models and Dependence Concepts*, vol. 73 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London, UK, 1997.

[22] T. Bedford and R. M. Cooke, "Probability density decomposition for conditionally dependent random variables modeled by vines," *Annals of Mathematics and Artificial Intelligence*, vol. 32, no. 1-4, pp. 245–268, 2001.

[23] T. Bedford and R. M. Cooke, "Vines - A new graphical model for dependent random variables," *Annals of Statistics*, vol. 30, no. 4, pp. 1031–1068, 2002.

[24] D. Kurowicka and H. Joe, *Dependence modeling: vine copula handbook*, D. Kurowicka and H. Joe, Eds., World Scientific, 2011.

[25] K. Aas, C. Czado, A. Frigessi, and H. Bakken, "Pair-copula constructions of multiple dependence," *Insurance: Mathematics & Economics*, vol. 44, no. 2, pp. 182–198, 2009.

[26] K. Aas and D. Berg, "Models for construction of multivariate dependence - a comparison study," *European Journal of Finance*, vol. 15, no. 7-8, pp. 639–659, 2009.

[27] I. H. Haff, "Parameter estimation for pair-copula constructions," *Bernoulli Society for Mathematical Statistics and Probability*, vol. 19, no. 2, pp. 462–491, 2013.

[28] C. Czado, U. Schepsmeier, and A. Min, "Maximum likelihood estimation of mixed C-vines with application to exchange rates," *Statistical Modelling*, vol. 12, no. 3, pp. 229–255, 2012.

[29] E. C. Brechmann, C. Czado, and K. Aas, "Truncated regular vines in high dimensions with application to financial data," *The Canadian Journal of Statistics*, vol. 40, no. 1, pp. 68–85, 2012.

[30] M. Smith, A. Min, C. Almeida, and C. Czado, "Modeling longitudinal data using a pair-copula decomposition of serial dependence," *Journal of the American Statistical Association*, vol. 105, no. 492, pp. 1467–1479, 2010.

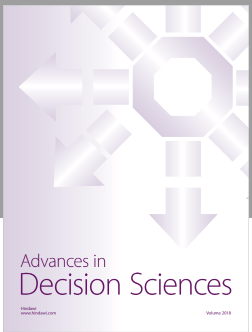
[31] J. Sun, E. W. Frees, and M. A. Rosenberg, "Heavy-tailed longitudinal data modeling using copulas," *Insurance: Mathematics and Economics*, vol. 42, no. 2, pp. 817–830, 2008.

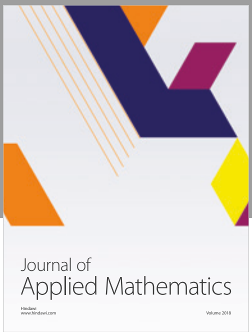
[32] A. J. Patton, "Estimation of Copula Models for Time Series of Possibly Different Lengths," *SSRN Electronic Journal*.

[33] P. Barbe, C. Genest, K. Ghoudi, and B. Rémillard, "On Kendall's process," *Journal of Multivariate Analysis*, vol. 58, no. 2, pp. 197–229, 1996.

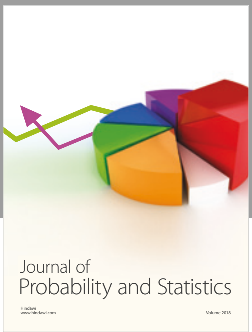
[34] Z. Rehman and Y. T. Xie, "Reply to Comment: Estimating abundance: a non parametric mark recapture approach for open and closed systems," *Environmental and Ecological Statistics*, vol. 24, no. 4, pp. 595–598, 2017.

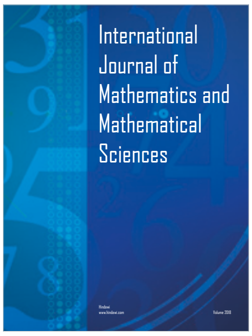
Advances in
Operations Research


Advances in
Decision Sciences

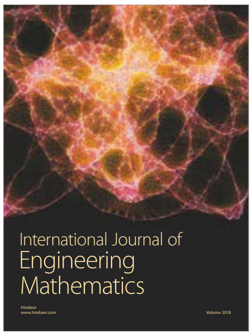

Journal of
Applied Mathematics

The Scientific
World Journal

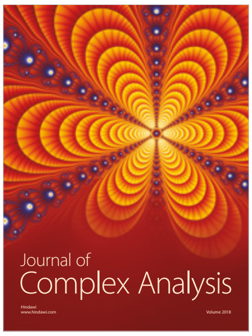

Journal of
Probability and Statistics

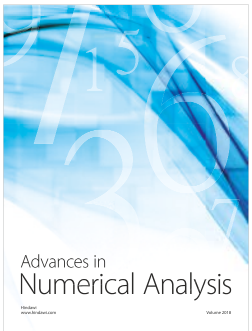

International
Journal of
Mathematics and
Mathematical
Sciences

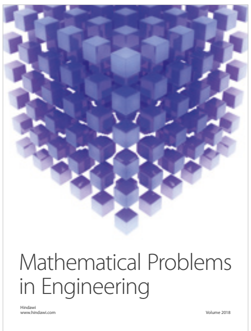
Journal of
Optimization

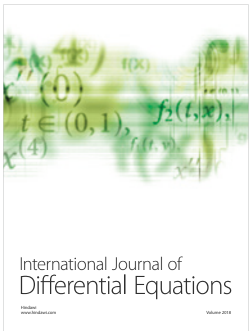

International Journal of
Engineering
Mathematics

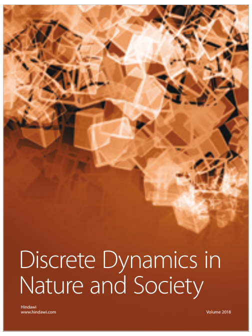
International Journal of
Analysis

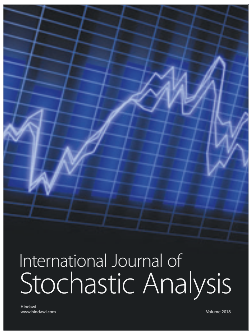

Journal of
Complex Analysis


Advances in
Numerical Analysis


Mathematical Problems
in Engineering

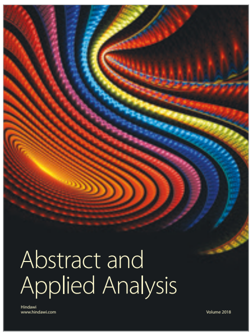

International Journal of
Differential Equations

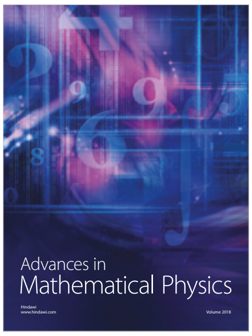

Discrete Dynamics in
Nature and Society


International Journal of
Stochastic Analysis

Journal of
Mathematics

Journal of
Function Spaces


Abstract and
Applied Analysis


Advances in
Mathematical Physics