

RESEARCH ARTICLE

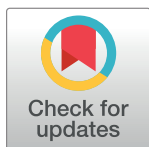
Clustering time series based on dependence structure

Beibei Zhang¹, Baiguo An¹✉*

School of Statistics, Capital University of Economics and Business, Beijing, China

✉ These authors contributed equally to this work.

* anbg200@163.com



Abstract

The clustering of time series has attracted growing research interest in recent years. The most popular clustering methods assume that the time series are only linearly dependent but this assumption usually fails in practice. To overcome this limitation, in this paper, we study clustering methods applicable to time series with a general and dependent structure. We propose a copula-based distance to measure dissimilarity among time series and consider an estimator for it, where the strong consistency of the estimator is guaranteed. Once the pairwise distance matrix for time series has been obtained, we apply a hierarchical clustering algorithm to cluster the time series and ensure its consistency. Numerical studies, including a large number of simulations and analysis of practical data, show that our method performs well.

OPEN ACCESS

Citation: Zhang B, An B (2018) Clustering time series based on dependence structure. PLoS ONE 13(11): e0206753. <https://doi.org/10.1371/journal.pone.0206753>

Editor: Cathy W.S. Chen, Feng Chia University, TAIWAN

Received: May 25, 2018

Accepted: October 18, 2018

Published: November 12, 2018

Copyright: © 2018 Zhang, An. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The data underlying this study are third party. The first GDP dataset is available from the following link: <https://www.conference-board.org/retrievefile.cfm?filename=Output-Labor-and-Labor-Productivity-1950-20111.xls&type=subtitle>. The second population growth dataset can be obtained from the following link: <https://www.census.gov/programs-surveys/popest/data/data-sets.2007.html>. In this website, the intercensal estimates of population are saved separately every ten years. One can download them separately and then put them together to obtain the final dataset used in our paper, i.e., the population estimates from 1991 to 2010 in 20

Introduction

Massive amounts of data relating to time series are frequently collected in fields ranging from science, engineering, and business to economics, healthcare, and government. It is often desirable to find groups of similar time series in a set or panel of such series using clustering techniques. In the case of panel time series, especially for short time series, it is beneficial for model estimation and forecasting performance to pool time series with similar data-generating mechanisms [1]. Consequently, as one of the most frequently used exploratory techniques in data mining, clustering is a crucial step in time series analysis.

A popular approach to clustering time series is model-based clustering. This approach assumes that each time series is generated by a specific underlying model or a mixture of underlying probability distributions. The time series are considered similar when the models characterizing them are similar. The most commonly considered models include the ARIMA process [2], the GARCH model [3], and the dynamic regression model [1]. Researchers in machine learning have also used Markov chains [4] and hidden Markov models [5].

Unlike model-based clustering methods, distance-based methods cluster time series in a simple and efficient way, where the choice of a proper distance or dissimilarity measure is a critical step. Once the dissimilarity measure is determined, an initial pairwise dissimilarity matrix can be obtained and many conventional clustering algorithms can then be used to form

states of the US. For example, we can find 2000–2010 intercensal estimates of population from the website: <https://www2.census.gov/programs-surveys/popest/tables/2000-2010/intercensal/state/st-est00int-01.xls>. The authors did not have special access privileges and interested researchers will be able to access the data in the same manner as the authors.

Funding: The research is supported by National Natural Science Foundation of China (No. 11601349), scientific research planned project of National Bureau of Statistics of China (No. 2017LZ15). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

groups of objects. Different distances are pursued according to the aim of time series clustering. The selected distance should be able to capture particular discrepancies between series that are relevant to the final objective of the clustering. The R package TSclust [6] provides a brief overview of well-established peer-reviewed time series dissimilarity measures, including measures based on raw data, extracted features, underlying parametric models, levels of complexity, and forecast behaviors. An interesting overview of time series clustering methods and their applications can be found in [7].

A central issue in the analysis of time series data is to consider the structure of the temporal dependence. It is often helpful for model estimation to cluster time series into several groups according to their underlying dependency structures. In most research on the issue, it is assumed that the temporal dependences of time series are only linear. The ARIMA model is the most commonly used linear model (see, for example, [2, 8–12]). However, the assumption of linearity often fails to hold in practice. When time series are nonlinearly dependent, linear methods suffer from a severe model mismatch problem. Scant attention has been paid in the literature thus far to the clustering of nonlinear time series. Nonparametric model-free methods are usually employed to deal with nonlinear problems. Dissimilarity in nonparametric distance-based clustering methods is measured by comparing serial features extracted from the original series that aim to represent the dynamic structure of each series, such as autocorrelation [13, 14], partial autocorrelation [15], cross-correlation [16], and spectral features [17]. Even though nonparametric methods do not make any model assumptions, most aforementioned dissimilarities are quantities related to Pearson's correlation, which can only measure linear dependence. It is thus natural that these features are inadequate at recognizing more general, temporal, and nonlinear dependence structures. They are not expected to perform well at clustering more general time series, which was also shown in our simulation experiments in this study.

We ignore model-based clustering methods and focus on the distance-based nonlinear time series clustering approach due to its popularity and simplicity. In view of the aforementioned considerations, a distance measure of global dependence is required. Specifically, we need to construct a distance measure that can capture linear and nonlinear dependencies without requiring the specification of any kind of model. However, to the best of our knowledge, no prevalent measure of association or dependence can satisfy this requirement. Numerous diagnostics or tests can be used to only examine departures from independence, and include mutual information, entropy measure, and the Hellinger distance. They cannot distinguish between dependence structures or gauge dissimilarity between them.

In this paper, we propose a distance measure based on a copula function to measure the dissimilarity among the general serial dependence of two time series. The advantages of this distance measure can be summarized as follows: First, it overcomes the limitations of prevalent time series clustering methods designed for linear processes. Our simulations show that the proposed measure performs well, particularly in terms of classifying nonlinear models. As the proposed distance measure is designed in terms of a discrepancy in the serial structure of global dependence, which includes linear structures, it can also be used for linear processes. Second, it is nonparametric. To obtain the distance measure, we rely on an empirical estimator of the copula function. The superiority of the proposed estimation approach resides precisely in its ability to account for the divergence of global dependencies with no need to specify an exact model. Third, this is a rank-invariant approach, as copulas possess an invariance property with respect to a monotonically increasing transformation of the variables [18]. Fourth, we can theoretically guarantee the consistency of the distance estimator. Fifth, this distance measure takes a close form that can be efficiently computed.

The remainder of the paper is organized as follows: In Section 2, we described the proposed clustering method with relevant statistical properties. Simulations and analysis of data from a practical scenario are provided in Sections 3 and 4, respectively. A short conclusion is provided in Section 5. The proofs of all theorems are in the Appendix.

Methodology

Notations and copula-based distance

Let $X_i = (X_{i1}, \dots, X_{iT_i})^T$ be the i -th time series ($1 \leq i \leq n$), where T_i is its length and n is the number of time series. We assume that these time series are strictly stationary drawn from J_0 clusters, and the time series in each cluster share a common dynamic pattern. The purpose is to identify these J_0 clusters. We propose a method based on the copula function to represent the dynamic pattern of the time series. Specifically, for a fixed positive integer h , $(X_{ij}, X_{i(j+h)})$ and $(X_{i'j}, X_{i'(j+h)})$ have the same copula function if time series X_i and $X_{i'}$ belong to a common cluster.

For random variables X and Y with continuous marginal cumulative distribution functions $F_X(x)$ and $F_Y(y)$, respectively, we denote the joint cumulative distribution functions by $F(x, y)$. Sklar's theorem [19] claims that a unique copula function $C(u, v)$ exists connecting $F(x, y)$ to $F_X(x)$ and $F_Y(y)$ via $F(x, y) = C(F_X(x), F_Y(y))$, which is equivalent to $C(u, v) = F(F_X^{-1}(u), F_Y^{-1}(v))$. This means that the copula function $C(u, v)$ can capture the structure of dependence between random variables X and Y . We use the copula to capture the dynamic pattern of the time series.

For each $1 \leq i \leq n$, we denote by $C_{i,h}(u, v)$ the copula function of $(X_{ij}, X_{i(j+h)})$. For arbitrary $i \neq i'$, we define the following copula-based Cramér-von Mises distance to measure the dissimilarity between time series X_i and $X_{i'}$:

$$D_h(i, i') = \sqrt{\int \int_{[0,1]^2} (C_{i,h}(u, v) - C_{i',h}(u, v))^2 du dv}. \quad (1)$$

The copula-based distance $D_h(i, i')$ satisfies the following three classical properties:

1. (Nonnegativity) $D_h(i, i') \geq 0$. Moreover, $D_h(i, i') = 0$ if and only if $(X_{ij}, X_{i(j+h)})$ and $(X_{i'j}, X_{i'(j+h)})$ share a common copula function.
2. (Symmetry) $D_h(i, i') = D_h(i', i)$.
3. (Triangle inequality property) $D_h(i, i') \leq D_h(i, k) + D_h(k, i')$.

Nonnegativity and symmetry are apparent in the foregoing, and the triangle inequality property can be obtained from the Cauchy–Schwarz inequality. Traditionally, the dependence structure of time series is often captured using an autocovariance-based linear correlation, which fails in nonlinear dependency structures. Nonlinear dependence may be caused by various nonlinear structures. The copula function does not make any assumptions about the model, and is a flexible method that can capture them. Moreover, copula-based distance is not affected by strict monotonic transformations (i.e., logarithmic transform and exponential transform) due to the property of mapping invariance of the copula function [18].

If time series $X_i (1 \leq i \leq n)$ are dependent in the order of one, we let $h = 1$ and directly use $D_1(i, i')$ as the dissimilarity measure of X_i and $X_{i'}$. If the time series are dependent in a higher

order, the dissimilarity measure of $X_i, X_{i'}$ can be defined as the following weighted version:

$$D(i, i') = \sum_{h=1}^K \omega_h D_h(i, i'),$$

where K is the highest order of dependence of time series X_i and $X_{i'}$ considered in the dissimilarity measure definition, and ω_h is the weight of $D_h(i, i')$. We can allow ω_h to decrease as h becomes larger.

The selection of K depends on the unknown underlying model. We may entertain several possible values of K and use model selection criteria such as AIC and BIC to select the optimal value of K . On this point, [20] claimed that the aim is not the goodness-of-fit to the underlying models but clustering them properly. We thus do not make any significant effort on this issue. In practice, the value of K should be chosen with specific knowledge of the application. It is often the case that the strongest serial dependency occurs in small lags, and a larger value of K means greater computational time and redundant information. Moreover, we can always obtain reasonably satisfactory results using small values of K in our simulations and applications. A large number of studies focusing on application have shown that it suffices for a large number of time series to use a lag of one to obtain goodness of fit. Therefore, in practice $K = 1$ is highly recommended.

Estimation of copula-based distance

In practice, the copula function is usually unknown, and thus the copula-based distance (1) cannot be used directly. We propose a nonparametric estimation of the copula function that can be plugged into (1) to obtain the estimated distance.

Specifically, we denote the empirical distribution functions of $(X_{it}, t = 1, \dots, T_i)$ by $\hat{F}_i(x) = T_i^{-1} \sum_{t=1}^{T_i} I(X_{it} \leq x)$, where $I(\cdot)$ is an indicator function. For $1 \leq t \leq (T_i - h)$, we further define

$$U_{it} = \frac{T_i}{T_i + 1} \hat{F}_i(X_{it}), \quad V_{it} = \frac{T_i}{T_i + 1} \hat{F}_i(X_{i(t+h)}).$$

Then, the nonparametric estimator for the copula function of $(X_{it}, X_{i(t+h)})$ (i.e., $C_{i,h}(u, v)$) is defined as

$$\hat{C}_{i,h}(u, v) = \frac{1}{T_i - h} \sum_{t=1}^{T_i-h} I(U_{it} \leq u) I(V_{it} \leq v).$$

We replace $C_{i,h}(u, v)$ in (1) by $\hat{C}_{i,h}(u, v)$ and obtain the corresponding copula-based distance estimator:

$$\hat{D}_h(i, i') = \sqrt{\int \int_{[0,1]^2} (\hat{C}_{i,h}(u, v) - \hat{C}_{i',h}(u, v))^2 du dv}. \quad (2)$$

If we further assume that the time series are α mixing processes, $\hat{C}_{i,h}(u, v), \hat{D}_h(i, i')$ are strong consistency estimators for $C_{i,h}(u, v)$ and $D_h(i, i')$, respectively. We summarize the results as the following theorem:

Theorem 1 Assume that time series $X_i = (X_{it}, 1 \leq t \leq T_i) (1 \leq i \leq n)$ are strictly stationary α mixing processes. Then, $\hat{C}_{i,h}(u, v) \xrightarrow{a.s.} C_{i,h}(u, v)$ as $T_i \rightarrow \infty$ and $\hat{D}_h(i, i') \xrightarrow{a.s.} D_h(i, i')$, as $T_i, T_{i'} \rightarrow \infty$, where $\xrightarrow{a.s.}$ denotes convergence with probability 1.

The proof of Theorem 1 is provided in Appendix A. In practice, we need to calculate the value of $\hat{D}_h(i, i')$, which is easy. For the details of the calculation of $\hat{D}_h(i, i')$, see the following proposition:

Proposition 1 For $i, i' = 1, \dots, n$, define

$$L_{i,i'} = \frac{\sum_{t=1}^{T_i-h} \sum_{t'=1}^{T_{i'}-h} (1 - \max(U_{it}, U_{i't'}))(1 - \max(V_{it}, V_{i't'}))}{(T_i - h)(T_{i'} - h)}.$$

Then, for the copula-based distance estimator $\hat{D}_h(i, i')$,

$$\hat{D}_h(i, i') = \sqrt{L_{i,i} - 2L_{i,i'} + L_{i',i'}}. \quad (3)$$

The proof of Proposition 1 is provided in Appendix B.

Clustering algorithms based on copula distance

Once the pairwise distance between the time series has been obtained, we can apply any clustering method that uses a distance matrix as input, such as partition around medoids (PAM) [21], spectral clustering [22], and hierarchical clustering [23]. In PAM methods, we need to find a centroid time series in each cluster to represent the group. PAM methods require that the user specify the number of clusters. Through a hierarchical clustering algorithm, we can obtain a tree dendrogram built starting from the leaves and combining clusters to the trunk. The hierarchical clustering algorithm applied in this paper can be summarized as follows:

Clustering Algorithm

Step 1. For $i, i' = 1, \dots, n$ and $i \neq i'$, compute the estimator $\hat{D}(i, i')$ for the copula-based distance between time series pair X_i and $X_{i'}$. Then, we can obtain the distance matrix. Treat each time series as its own cluster.

Step 2. For $J = n, n - 1, \dots, 2$:

- a. Compare all pairwise dissimilarity measures among the J clusters and identify the pair of clusters most similar. Merge them into as one.
- b. Compute the pairwise dissimilarities of the new $J - 1$ clusters.

It is worth mentioning that dissimilarities should be defined among clusters based on distances between time series prior to applying the clustering algorithm. Single linkage, complete linkage, average linkage, centroid linkage, and Ward's linkage are the most common methods to extend dissimilarities among observations to dissimilarities among clusters [24].

In practice, The computations of clustering can be performed in a more efficient way. [25] introduced a recurrence formula which can be used to compute the updated inter-cluster distance values efficiently. If a clustering procedure does not satisfy such a recurrence relation, the initial data should be retained throughout the entire process when updating cluster distances. All these aforementioned linkage methods fit into the Lance-Williams formalism, and can therefore easily be implemented with user-defined time series distance. In this paper, the efficient method of implementing clustering is to store the matrix of copula distances and update inter-cluster distance using Lance-Williams recursive formula.

Specifically, the traditional Ward's linkage method minimizes the increase in total within-cluster sum of squared error. This increase is proportional to the squared Euclidean distance between cluster centers. Here, we extends the classical Ward's linkage method that relies on Euclidean distance to copula distance, which still shares the same parameters in the update

formula with original Ward's linkage method [26]. With the Lance-Williams formula, we do not need to assign the cluster center to compute the new inter-cluster distance.

Recall that we assume these time series are drawn from J_0 clusters. We denote these J_0 clusters by $\{\mathcal{M}_1, \dots, \mathcal{M}_{J_0}\}$. Let $C_{j_0,h}(u, v)$ be the common copula shared by the time series in the j -th cluster. We further define

$$D_{0,h}(j, j') = \sqrt{\int \int_{[0,1]^2} (C_{j_0,h}(u, v) - C_{j'_0,h}(u, v))^2 du dv},$$

which is the copula-based distance between the j -th and the j' -th clusters. Let $\epsilon = \min_{j \neq j'} D_{0,h}(j, j')$, which is the minimal distance between the clusters. Then, for the foregoing clustering algorithm, the following theorem holds:

Theorem 2 Assume that time series $X_i = (X_{it}, 1 \leq t \leq T_i) (1 \leq i \leq n)$ are strictly stationary α mixing processes and $\epsilon > 0$. For $J > J_0$ in Step 2 of Clustering Algorithm, the obtained $J - 1$ clusters are denoted by $\{\widehat{\mathcal{M}}_1, \dots, \widehat{\mathcal{M}}_{J-1}\}$. Then, as $\min_i \{T_i\} \rightarrow \infty$, with probability 1 we have that for every $\widehat{\mathcal{M}}_j$ with $1 \leq j \leq J - 1$, there exists a j' such that $\widehat{\mathcal{M}}_j \subset \mathcal{M}_{j'}$.

We call the results of Theorem 2 the consistency of clustering. The proof of Theorem 2 is provided in Appendix C. If J_0 is known, we can directly cluster the time series into J_0 groups. However, the number of clusters J_0 is usually unknown, and we should detect the optimal number of clusters in practice. A useful approach to determine this optimal number is the silhouette method [27]. For each time series X_i with $i = 1, 2, \dots, n$, its silhouette width $s(i)$ is defined as

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))},$$

where $a(i)$ is the average copula distance between series X_i and all other time series of the cluster to which X_i belongs, and $b(i)$ is the average copula distance between series X_i and time series in neighboring cluster, i.e., the nearest cluster to which it does not belong. Let the set of time series be partitioned into J clusters. The corresponding average silhouette width is defined as

$$Sil(J) = \frac{\sum_{i=1}^n s(i)}{n}.$$

The average silhouette of the clusters is calculated according to the number of clusters. A high average silhouette width indicates good clustering. Thus, the optimal number of clusters J is that which maximizes the average silhouette over a range of possible values for J . We choose the number of clusters as J^* , which yields the maximum value of $Sil(J)$.

Simulations

We used three examples to assess the performance of nonlinear time series clustering based on our proposed distance measure. For the sake of comparison, in each simulation scenario, we performed clustering using some representative dissimilarity measures proposed in the literature, including model-based measures d_{PIC} of [8] and d_M of [28], which are based on ARIMA models. We also made comparisons with model-free dissimilarity measures. In the temporal domain, distances d_{ACF} and d_{PACF} are defined as Euclidean distances between the estimated ACF and PACF using a number of significant lags. In the frequency domain, the dissimilarity measures were designed to assess the discrepancy between corresponding spectral densities. [15] proposed distance measures based on the periodogram, Euclidean distances between

periodograms (d_P), log-periodograms (d_{LP}), normalized periodograms (d_{NP}), and log-normalized periodograms (d_{LNP}). The other nonparametric dissimilarity measures in the frequency domain, $d_{W(DLS)}$, $d_{W(LK)}$, d_{GLK} , and d_{ISD} proposed in [17], were also considered. They are different versions of nonparametric spectral dissimilarity measures. The differences among them were in terms of estimation methods of spectral density and discrepancy functions. For more details concerning these methods, the interested reader can refer to [6].

In the simulation experiments, the ground truth was known in advance. We assessed the clustering methods by using the cluster similarity index proposed in [29], which is defined as

$$Sim(G, A) = \frac{1}{J_0} \sum_{i=1}^{J_0} \max_{j \leq J_0} Sim(G_i, A_j),$$

where $G = \{G_1, G_2, \dots, G_{J_0}\}$ are the set of J_0 true clusters, $A = \{A_1, A_2, \dots, A_{J_0}\}$ is the solution to the clusters by the clustering method evaluated, and

$$Sim(G_i, A_j) = \frac{2|G_i \cap A_j|}{|G_i| + |A_j|},$$

where $|\cdot|$ denotes the cardinality of an arbitrary set. The similarity index has values ranging from zero to one, with one corresponding to the case when G and A are identical.

Example 1: Nonlinear time series clustering

In this example, we considered the following four models:

1. Threshold autoregressive (TAR) model

$$X_t = 0.5X_{t-1}I(X_{t-1} \leq 0) - 2X_{t-1}I(X_{t-1} > 0) + \varepsilon_t,$$

2. Exponential autoregressive (EXPAR) model

$$X_t = (0.3 - 10\exp\{-X_{t-1}^2\})X_{t-1} + \varepsilon_t,$$

3. Linear moving average (MA) model

$$X_t = \varepsilon_t - 0.4\varepsilon_{t-1},$$

4. Nonlinear moving average (NLMA) model

$$X_t = \varepsilon_t - 0.5\varepsilon_{t-1} + 0.8\varepsilon_{t-1}^2,$$

where the error process ε_t independently followed $N(0, 1)$. These models were used in [30] for linearity tests and [17] to study time series clustering. Except for model (3), the others were conditional mean nonlinear models. The stationarity of these models can be guaranteed. We here only take the TAR model as an instance to illustrate its stationarity. For a TAR model

$$X_t = (\gamma + \alpha X_{t-1})I(X_{t-1} \leq r) + (\delta + \beta X_{t-1})I(X_{t-1} > r) + \varepsilon_t,$$

where $\{\varepsilon_t\}$ are independent and identically distributed with zero mean and a finite variance.

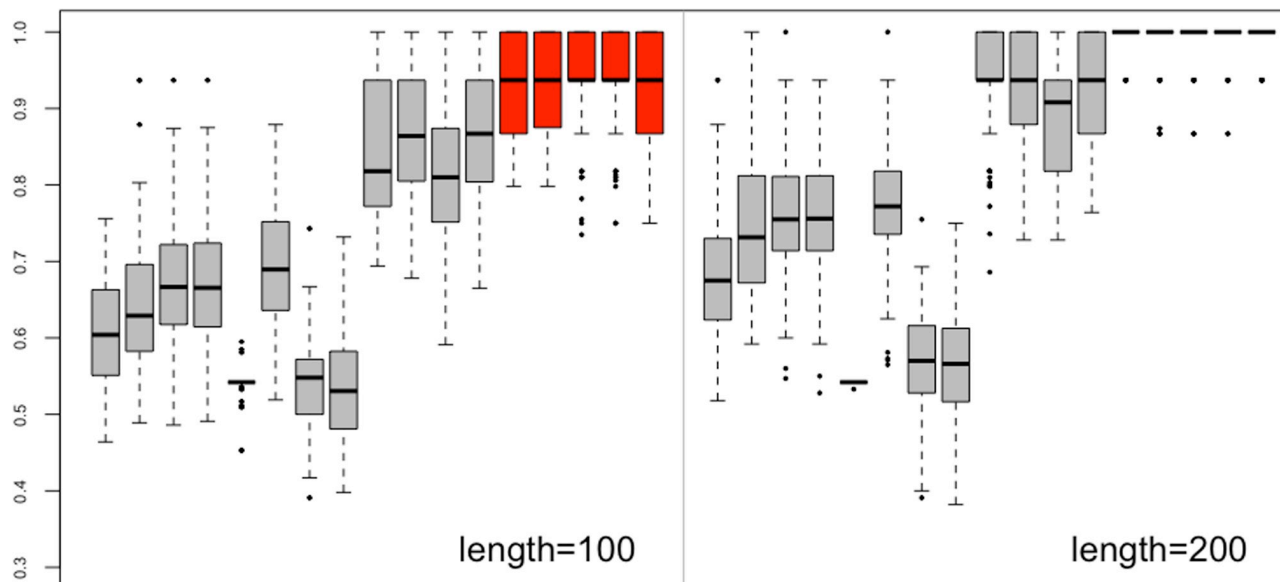


Fig 1. Example 1. Boxplot of clustering similarity indices. The distances from left to right: d_{ACF} , d_{PACF} , d_{PIC} , d_M , d_P , d_{LP} , d_{NP} , d_{LNP} , d_{DLS} , d_{LK} , d_{GLK} , d_{ISD} , $Copula(K=1)$, $Copula(K=2)$, $Copula(K=3)$, $Copula(K=4)$, and $Copula(K=5)$.

<https://doi.org/10.1371/journal.pone.0206753.g001>

Then the necessary and sufficient condition for the strictly stationarity to above TAR model when $\gamma = \delta = 0$ is $\alpha < 1$, $\beta < 1$ and $\alpha\beta < 1$ [31–34]. In model (1), we set $\gamma = \delta = r = 0$, and $\alpha = 0.5$, $\beta = -2$, hence the necessary and sufficient condition holds, and further the stationarity of the TAR model (1) is guaranteed.

We generated four time series from each model, and thus the sample size was 16. We set the lengths of all series as a common parameter T , and we considered two values of T (i.e. $T = 100, 200$). For sake of simplicity, in all of the experiments we used uniform weight for copula distance, that is $w_h = 1$ for $1 \leq h \leq K$. The experiment was repeated 100 times using all the aforementioned distance measures. The clustering similarity indices were calculated and summarized by the boxplot in Fig 1. The distance based on the copula function always yielded the best performance. For each of the distances, a larger series size seems to benefit more. When the series length $T = 200$, the similarity indices of the copula distance were almost equal to one, which meant that the copula distance could cluster the series into the true group from which they were generated. The first eight distances ($d_{ACF} \sim d_{LNP}$) needed the assumption of linearity of the time series. Therefore, their results were significantly worse than those of the other nonparametric distances ($d_{DLS} \sim d_{ISD}$) in this example.

Fig 2 shows the multidimensional scaling (MDS) plot [35] used to visualize observations in two dimensions ($T = 200$), where the dissimilarity among the time series was based on our proposed copula distance. With the dissimilarity measures obtained from the data, the MDS plot sought a lower-dimensional representation of the data that preserved pairwise distances as closely as possible. Fig 2 shows a clear separation of the four clusters and the capability of copula distance to discriminate among them. Furthermore, it is evident that the series from the MA models and NLMA models are closer to one another because both models expressed time series as a function of white noise.

To gain further insights into copula distance, based on Example 1, we designed two more challenging clustering tasks. The first involved increasing heterogeneity within each cluster and the second explored the performance of copula distance when the strength of nonlinear dependence was changing.

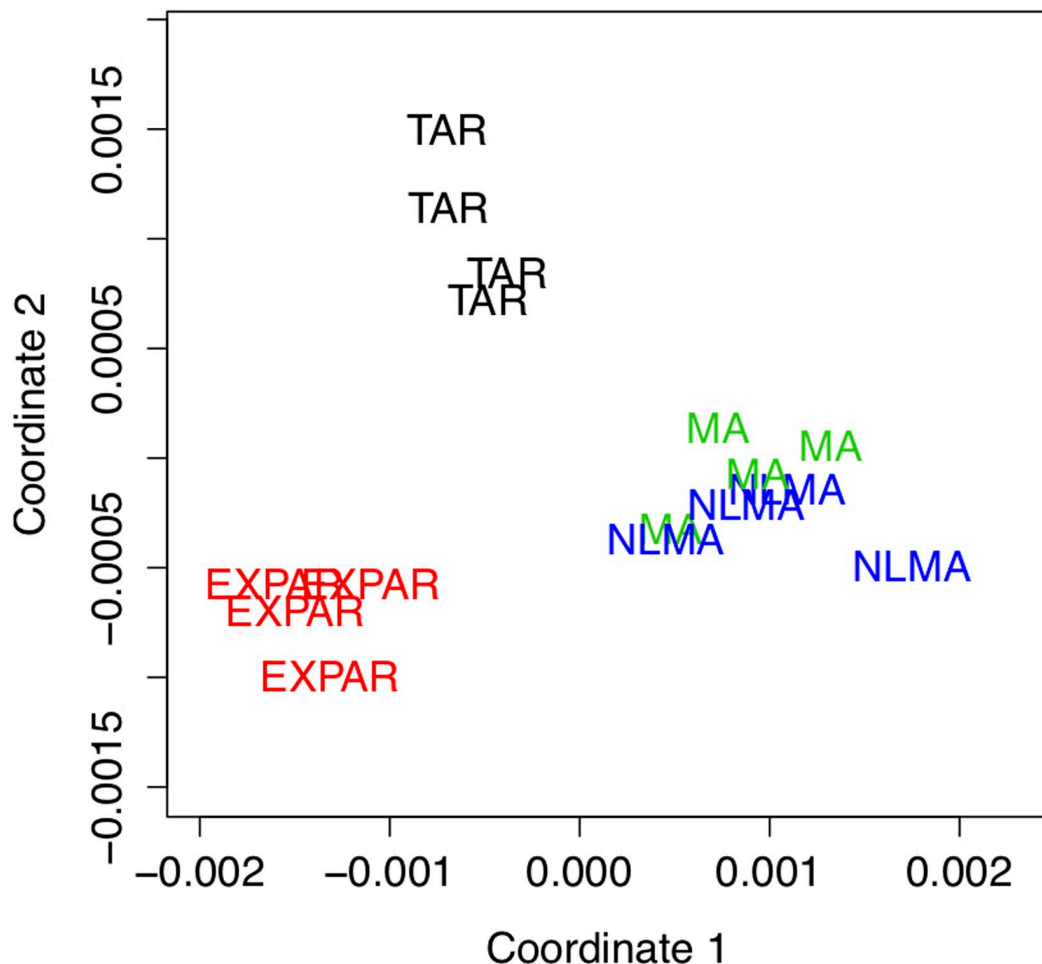


Fig 2. Example 1. Multidimensional scaling plot.

<https://doi.org/10.1371/journal.pone.0206753.g002>

Example 2: Nonlinear time series clustering with increased intra-cluster heterogeneity

In this example, we considered models similar to those in Example 1. The length of each time series was set to $T = 200$ here. To enhance the difficulty of nonlinear time series clustering, we made some changes to each model. For model (1), TAR model is generalized to the Smooth Transition Autoregressive (STAR) model, which allows for higher degree of flexibility in model parameters. Similarly, for model (3), linear MA model was replaced by the Smooth Transition Moving Average (STMA) model. For models (2) and (4), instead of using a fixed constant in each model, we used varying coefficients generated randomly from some given uniform distributions. Specifically, the models considered were:

1. Smooth transition autoregressive (STAR) model

$$X_t = 0.5X_{t-1} - 2.5X_{t-1}F(X_{t-1}) + \varepsilon_t,$$

where $F(X_{t-1}) = (1 + \exp(-X_{t-1}))^{-1}$ is the smooth transition function.

2. Exponential autoregressive (EXPAR) model

$$X_t = (a_2 - b_2 \exp\{-X_{t-1}^2\})X_{t-1} + \varepsilon_t, a_2 \sim U(0.2, 0.6), b_2 \sim U(6, 12);$$

3. Smooth transition moving average (STMA) model

$$X_t = \varepsilon_t - 0.4\varepsilon_{t-1} + 0.8\varepsilon_{t-1}F(X_{t-1}) + \varepsilon_t,$$

where the smooth transition function is specified as

$$F(X_{t-1}) = 1 - \exp(-X_{t-1}^2).$$

4. Nonlinear moving average (NLMA) model

$$X_t = \varepsilon_t - a_4\varepsilon_{t-1} + b_4\varepsilon_{t-1}^2, a_4 \sim U(0.3, 0.7), b_4 \sim U(0.2, 0.9).$$

The results are shown in Fig 3. As the heterogeneity in each group increased, almost all of the distance measured led to different levels of performance degradation except copula distance. The copula distance still generated the best performance among all measures, and its performance has hardly any degradation.

Example 3: Nonlinear time series clustering by adjusting nonlinear strength

In this example, we studied the sensitivity of copula distance to the nonlinear strength of the time series with the length $T = 200$. We wanted to determine the clustering performance of copula distance when the nonlinear strength of the time series varied. We controlled the strength of nonlinearity by adjusting the coefficients of the models considered in Example 1 as follows:

1. Threshold autoregressive (TAR) model

$$X_t = 0.5X_{t-1}I(X_{t-1} \leq 0) - b_1X_{t-1}I(X_{t-1} > 0) + \varepsilon_t;$$

2. Exponential autoregressive (EXPAR) model

$$X_t = (0.3 - b_2 \exp\{-X_{t-1}^2\})X_{t-1} + \varepsilon_t;$$

3. Linear moving average (MA) model

$$X_t = \varepsilon_t - 0.4\varepsilon_{t-1};$$

4. Nonlinear moving average (NLMA) model

$$X_t = \varepsilon_t - 0.5\varepsilon_{t-1} + b_4\varepsilon_{t-1}^2.$$

We can see that model (3) remains identical to that in Example 1 and, if (b_1, b_2, b_4) was equal to $(2, 10, 0.8)$, the remaining three models were identical to those in Example 1. On the

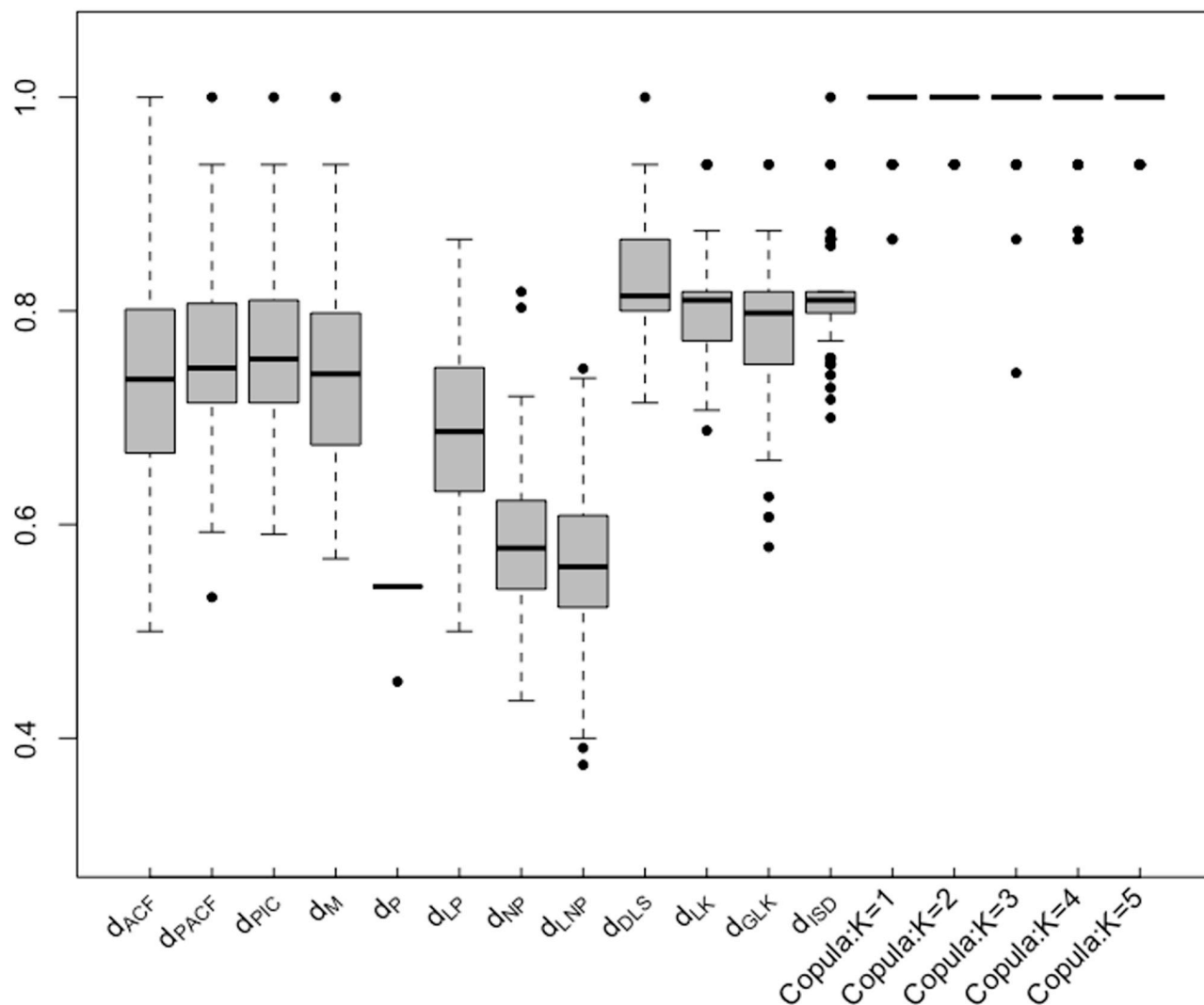


Fig 3. Example 2. Boxplot of clustering similarity indices: heterogeneity enlarged.

<https://doi.org/10.1371/journal.pone.0206753.g003>

contrary, if we set (b_1, b_2, b_4) equal to $(-0.5, 0, 0)$, the models degenerated to the following linear models:

1. $X_t = 0.5X_{t-1} + \varepsilon_t$;
2. $X_t = 0.3X_{t-1} + \varepsilon_t$;
3. $X_t = \varepsilon_t - 0.4\varepsilon_{t-1}$;
4. $X_t = \varepsilon_t - 0.5\varepsilon_{t-1}$.

We can see that the strength of nonlinear dependency decreased when (b_1, b_2, b_4) changed from $(2, 10, 0.8)$ to $(-0.5, 0, 0)$. Therefore, if we assigned $(b_1, b_2, b_4) = (2.5\alpha - 0.5, 10\alpha, 0.8\alpha)$, $0 \leq \alpha \leq 1$, the size of α represented the strength of nonlinear dependence. The larger the value of α , the stronger the nonlinear dependence. When $\alpha = 1$, this strength was the highest, which is the situation in Example 1. On the contrary, when $\alpha = 0$ the strength of nonlinear dependence was the weakest, which was linear. We provided six uniformly spaced values in $[0, 1]$ to α (i.e., $\alpha = 0, 0.2, 0.4, 0.6, 0.8, 1$), and the results of clustering are shown in Fig 4 below.

From Fig 4, we see that when the time series were simulated from linear models, i.e., $\alpha = 0$, the distance measures based on the assumption of model linearity yielded the best performance, such as d_{PACF} , d_{PIC} , and d_M . The performance of copula distance was not inferior to that of the other methods. As nonlinear strength increased (i.e., α increased), the advantage of nonparametric distance measures become ever more apparent while the clustering performance of methods based on linearity degenerated. In most cases, copula distance yielded far better results than the competition.

Real data analysis

We further illustrate the use of copula distance for time series clustering with two practical examples.

Case 1: Annual real GDP data analysis

We considered data concerning the gross domestic product (GDP) obtained from <https://www.conference-board.org/retrievefile.cfm?filename=Output-Labor-and-Labor-Productivity-1950-2011.xls&type=subsite>. It contained the annual real GDP of the 23 most developed

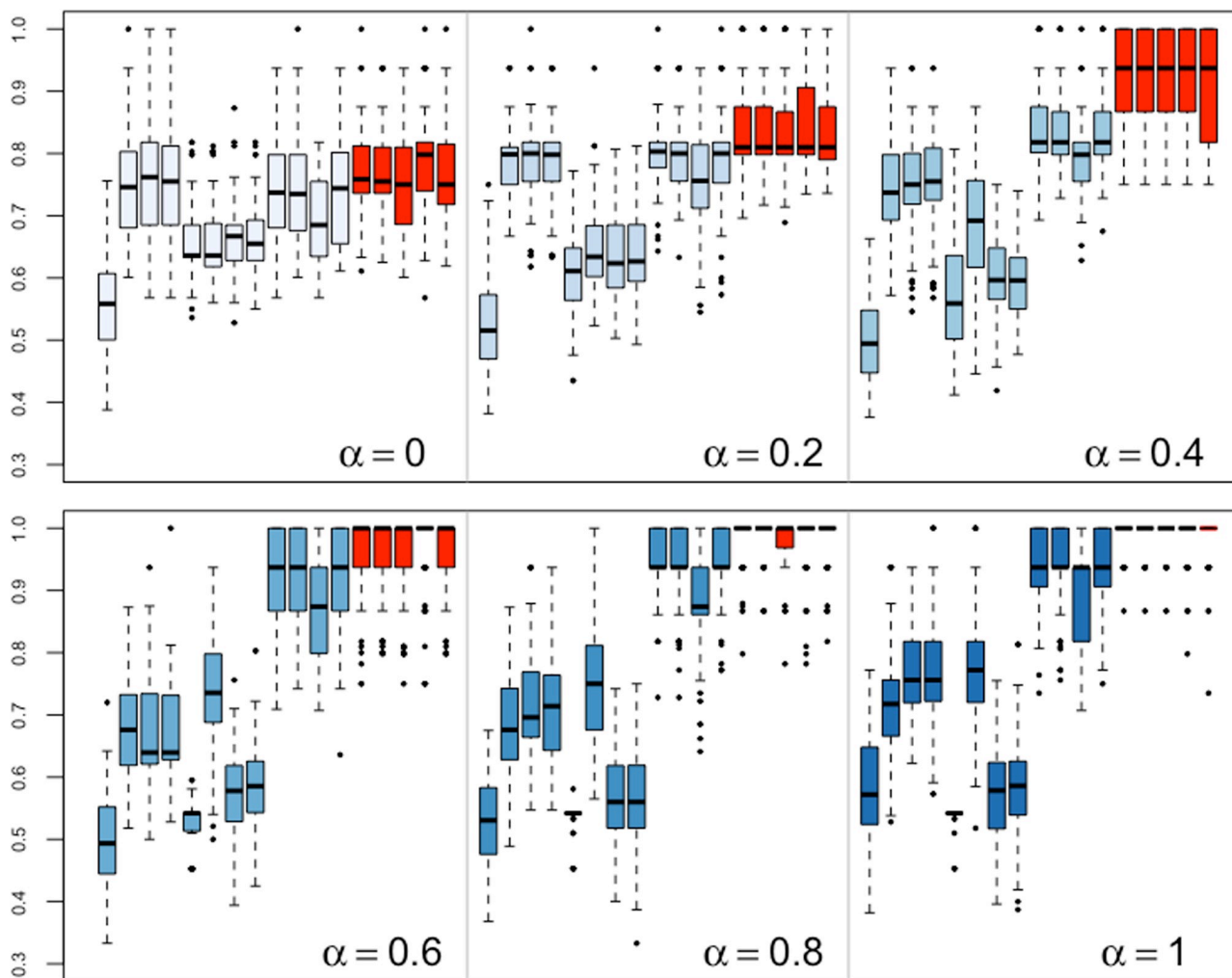


Fig 4. Example 3. Boxplots of clustering similarity indices: adjusting nonlinear strength. The distances from left to right: d_{ACF} , d_{PACF} , d_{PIC} , d_M , d_P , d_{LP} , d_{NP} , d_{LNP} , d_{DLS} , d_{LK} , d_{GLK} , d_{ISD} , $Copula(K=1)$, $Copula(K=2)$, $Copula(K=3)$, $Copula(K=4)$, and $Copula(K=5)$.

<https://doi.org/10.1371/journal.pone.0206753.g004>

countries in the world from 1950 to 2011: Austria, Belgium, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Italy, Luxembourg, the Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, United Kingdom, Canada, the United States, Australia, New Zealand, and Japan. We considered data normalized by the EKS method [36]. We used annual GDP growth rate $\log(GDP_t) - \log(GDP_{t-1})$ in the clustering procedures rather than annual GDP. These series were clustered via Ward's linkage method based on copula distance. When $K = 2$, the dendrogram is shown in Fig 5. The result was relatively insensitive to the maximum lag K used, with similar clustering results obtained when K ranged from two to nine.

In practice, we do not know how many distinct populations generate n time series. In general, as in any cluster analysis, the optimal number of clusters can be chosen according to some objective criterion, such as the average silhouette criterion. The average silhouette of the data is a useful criterion for assessing the natural number of clusters, which can be determined by maximizing the coefficient. It is a measure of how tightly grouped all data in the cluster are. The average silhouette coefficients were examined for different numbers of clusters, and two clusters appeared to yield a compact solution (Fig 6).

Fig 7 shows the grouping of the two clusters. In this figure, the two colors represent two groups. To display the clustering effect of copula distance, the area occupied by Europe is magnified many times. It is interesting to note that the countries were grouped primarily by geographical location. The group in blue contains five northern European countries and nearly all developed non-European countries except Japan. The northern European countries have a

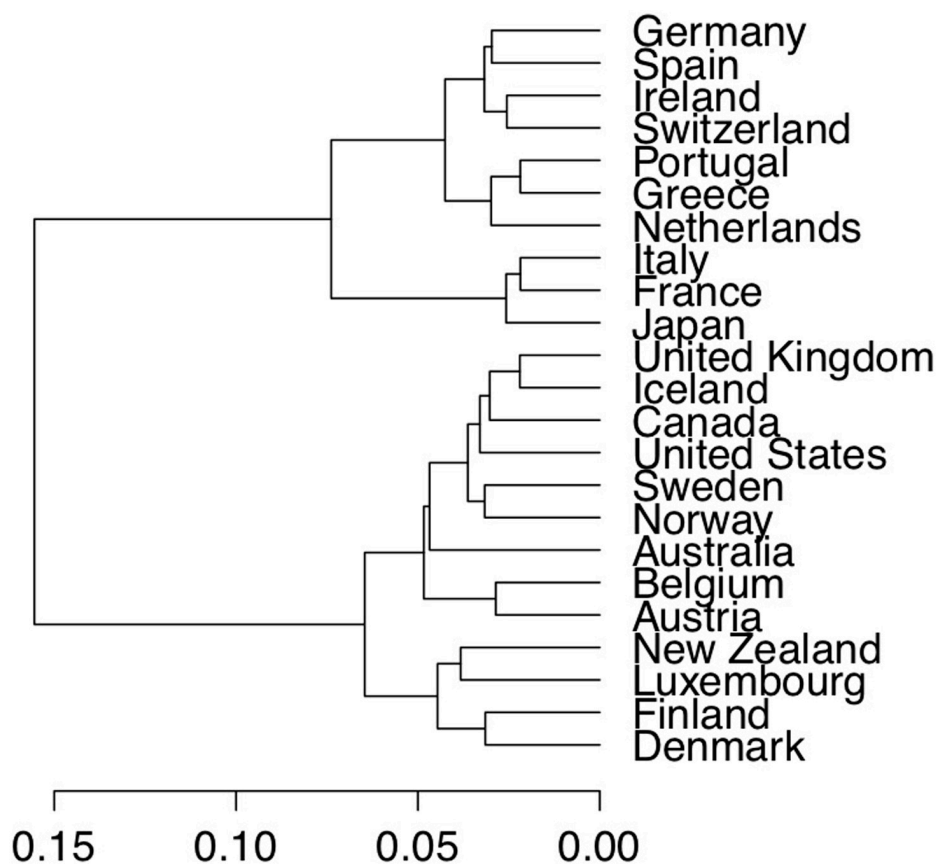


Fig 5. Annual real GDP data analysis. GDP clustering dendrogram based on copula distance with $K = 2$.

<https://doi.org/10.1371/journal.pone.0206753.g005>

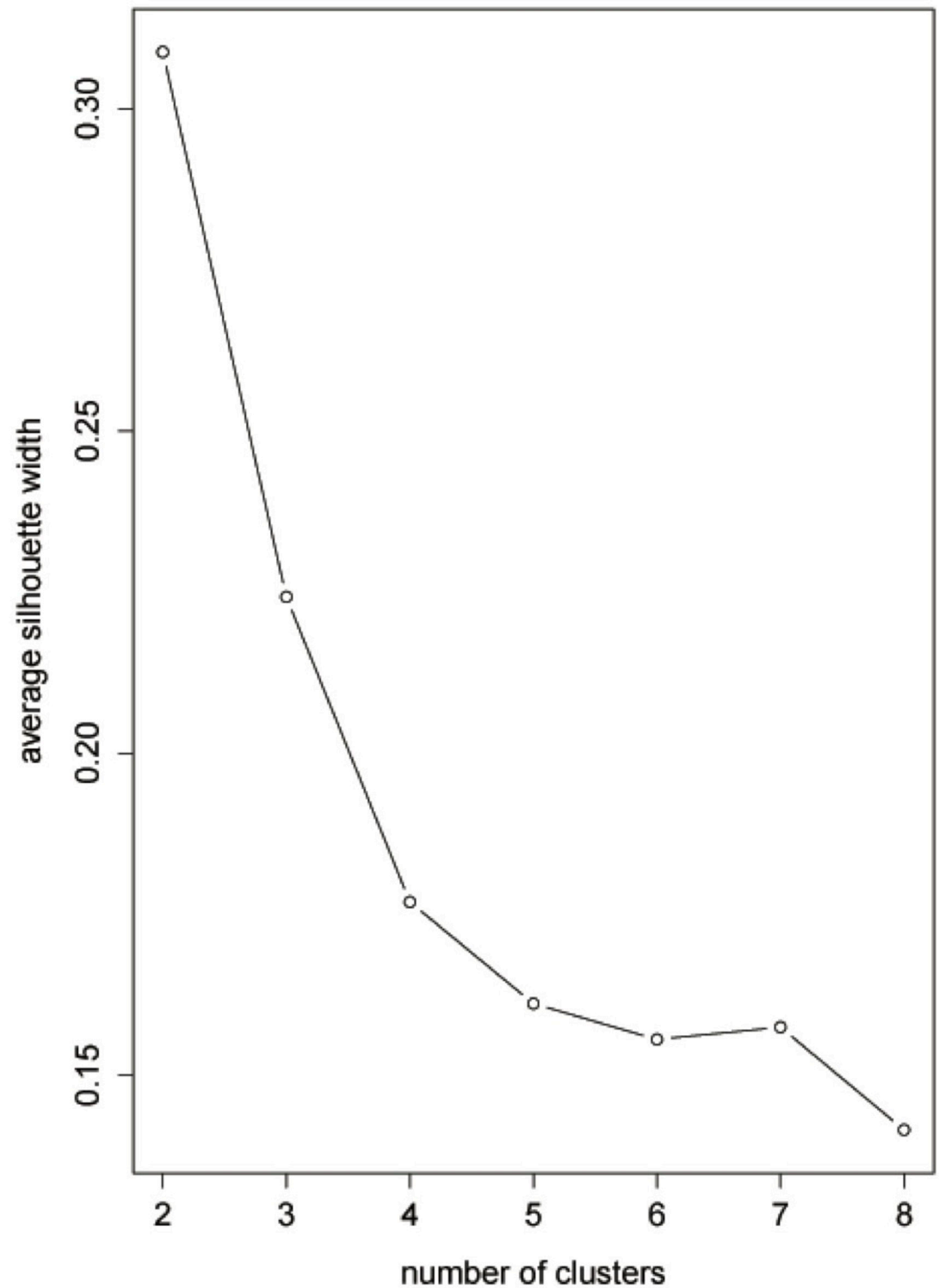


Fig 6. Annual real GDP data analysis. Plot of average silhouette coefficient ($K = 2$).

<https://doi.org/10.1371/journal.pone.0206753.g006>

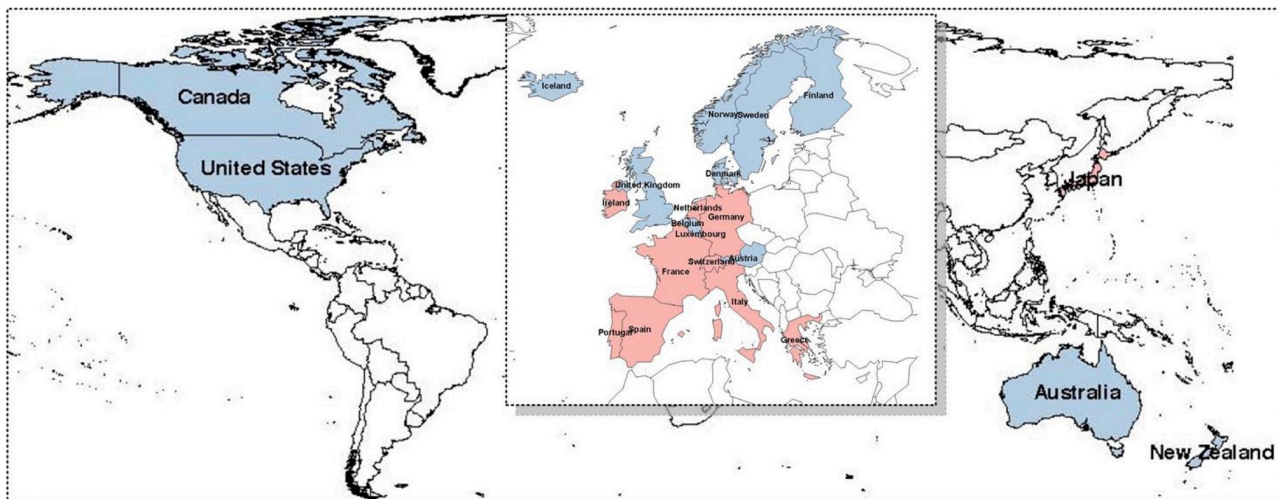


Fig 7. Annual real GDP data analysis. Two group of GDP data in map based on copula distance ($K = 2$).

<https://doi.org/10.1371/journal.pone.0206753.g007>

long history of cooperation and so have much in common. The second group is in red, and includes central European countries, southern European countries, and western European countries to the south.

Case 2: Population growth data analysis

This application considers the annual population series of 20 US states, and is aimed at identifying similarities among the population growth trend. This data is obtained from the U.S Census Bureau, Population Distribution Division (<https://www.census.gov/programs-surveys/popest/data/data-sets.2007.html>). It is a collection of time series of the population estimates from 1991 to 2010 in 20 states of the US, and has been used in [10]. In [10], two different groups of time series in the dataset were identified. Group 1 consisting of CA, CO, FL, GA, MD, NC, SC, TN, TX, VA, and WA had an exponentially growth trend, while Group 2 consisting of IL, MA, MI, NJ, NY, OK, PA, ND, and SD had a stable trend. In this case, we assume the above finding is the ground truth. In the following analysis, we use the growth series by taking log difference of the original time series. We cluster the data with hierarchical clustering algorithm using Ward's linkage method. The dendrogram clustered by copula distance with Ward's linkage method are shown in Fig 8. As we know, in this application the optimal number of clusters is 2. We can get the exact same optimal cluster number by maximizing the average silhouette width (see Fig 9). The copula distance method has the similarity index 0.8. If we section the dendrogram at the highest level we can obtain two groups, and all of the states are correctly classified except three states, CA, FL and MD.

Conclusion

The ability to successfully cluster sets of time series is a popular area of research in many fields. In this paper, we proposed a clustering method that is not limited to linear processes. Because of the diversity of the structures of dependence of time series, we avoided model-based clustering and used the distance-based method instead. We proposed a distance measure based on the copula function to measure the dissimilarity between the general serial dependence of time series. This distance can be calculated easily by the empirical estimator of the copula function. We theoretically guaranteed the consistency of the distance estimator as well. This method

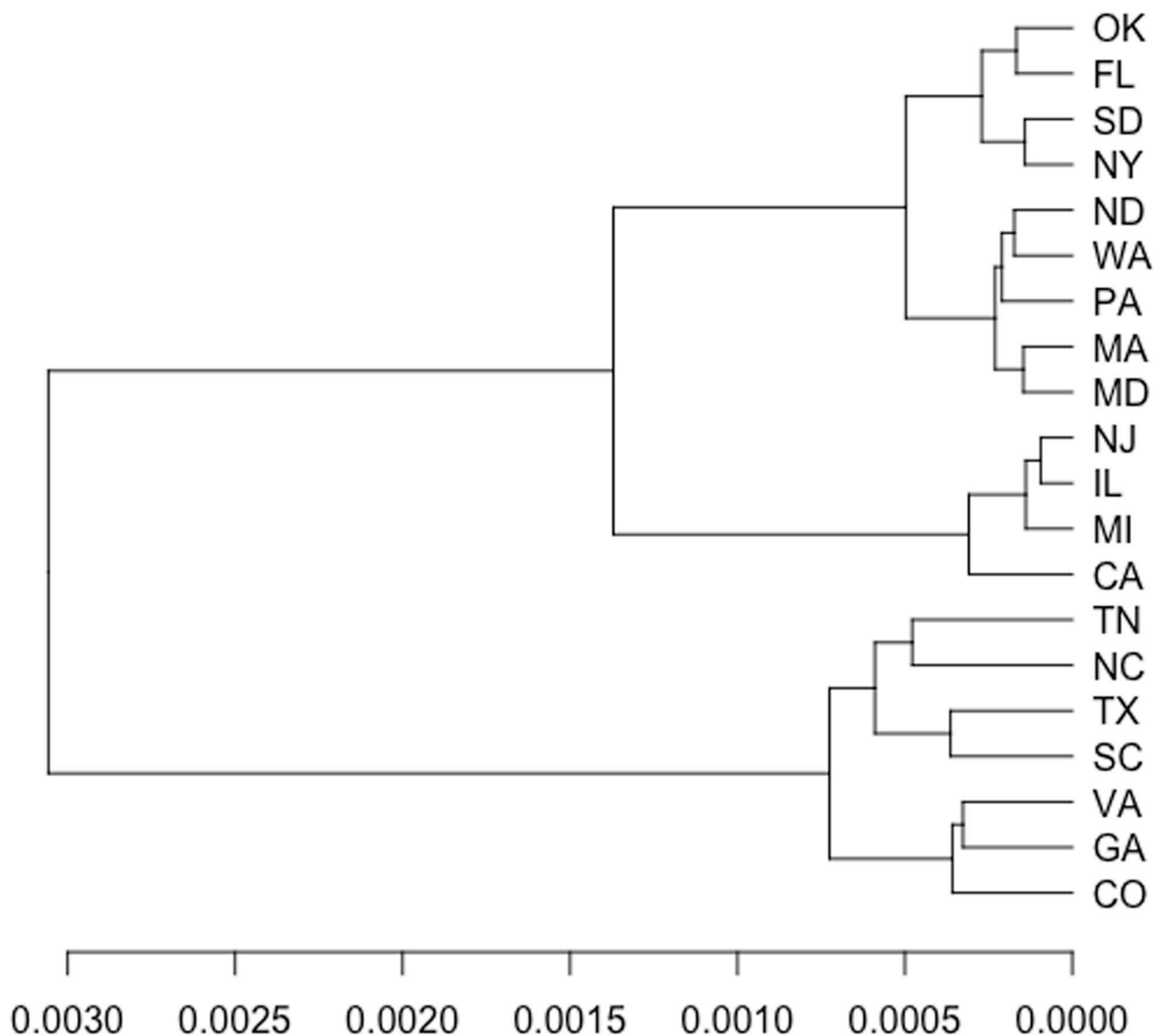


Fig 8. Population growth data analysis. Population Growth clustering dendrogram based on copula distance with $K = 2$.

<https://doi.org/10.1371/journal.pone.0206753.g008>

takes advantage of the ability of the copula function to measure the global dependence of the time series and fills gaps in research on time series clustering. Three simulation examples and the examination of a practical scenario illustrated the usefulness of our method. However, our proposed method has its limitation. Beyond the linear time series, there are infinitely many nonlinear forms to be explored. In this paper we only explore nonlinear relationship between lagged variables, which is only one general kind of nonlinear time series.

Appendixes

Appendix A: Proof of Theorem 1

For every $1 \leq i \leq n$, denote by $F_i(x)$ the marginal distribution function of X_{it} . Then we will prove the theorem by the following four steps.

Step 1. For arbitrary $u, v \in [0, 1]$, define

$$\tilde{C}_{i,h}(u, v) = (T_i - h)^{-1} \sum_{t=1}^{T_i-h} I(F_i(X_{it}) \leq u, F_i(X_{i(t+h)}) \leq v).$$

In this step, we will prove that as $T_i \rightarrow \infty$, $\tilde{C}_{i,h}(u, v) \xrightarrow{a.s.} C_{i,h}(u, v)$.

Specifically, for $1 \leq i \leq n$ and $1 \leq t \leq T_i - h$, let $Y_{it} = I(F_i(X_{it}) \leq u, F_i(X_{i(t+h)}) \leq v)$. Because $X_i = (X_{it}, 1 \leq t \leq T_i)$ is a strictly stationary α mixing process, one can see that $Y_i = \{Y_{it}, 1 \leq t \leq T_i\}$ is also a strictly stationary α mixing process. Then by Proposition 2.8 of [37], one can see that as $T_i \rightarrow \infty$,

$$\tilde{C}_{i,h}(u, v) \xrightarrow{a.s.} P(F_i(X_{it}) \leq u, F_i(X_{i(t+h)}) \leq v) = C_{i,h}(u, v).$$

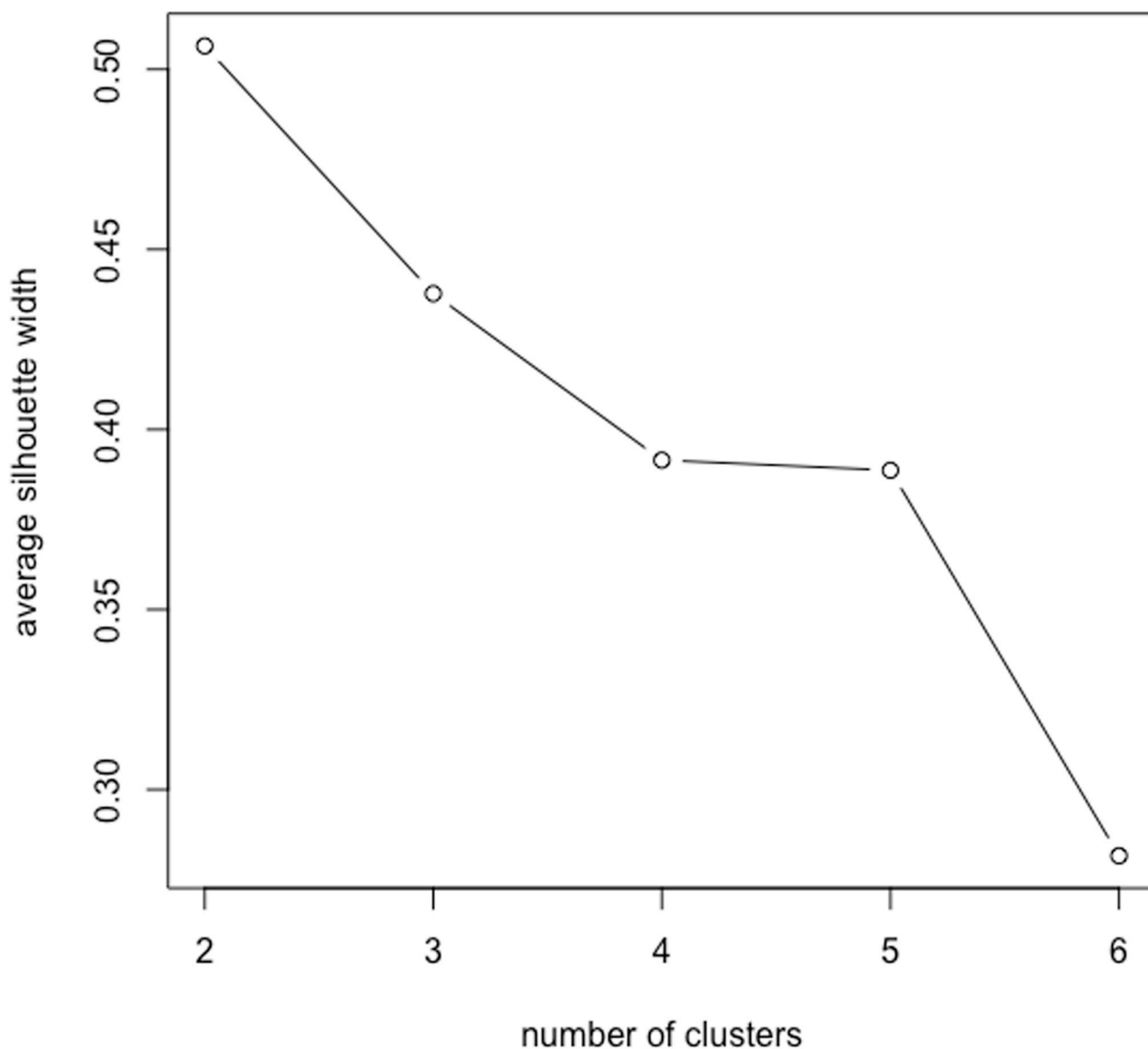


Fig 9. Population growth data analysis. Plot of average silhouette coefficient ($K = 2$).

<https://doi.org/10.1371/journal.pone.0206753.g009>

Step 2. In this step, we will prove that as $T_i \rightarrow \infty$,

$$\sup_{-\infty < x < \infty} |\hat{F}_i(x) - F_i(x)| \xrightarrow{a.s.} 0.$$

Because $X_i = (X_{it}, 1 \leq t \leq T_i)$ is a strictly stationary α mixing process, X_i is ergodic. Then by the theorem of [38], the Glivenko-Cantelli theorem holds for time series X_i , hence we have that as $T_i \rightarrow \infty$,

$$\sup_{-\infty < x < \infty} |\hat{F}_i(x) - F_i(x)| \xrightarrow{a.s.} 0.$$

Step 3. In this step, we will prove that $\hat{C}_{i,h}(u, v) \xrightarrow{a.s.} C_{i,h}(u, v)$. Define

$$A_i = \left\{ \lim_{T_i \rightarrow \infty} \sup_{-\infty < x < \infty} |\hat{F}_i(x) - F_i(x)| = 0 \right\}.$$

By the result of Step 2, we know that $P(A_i) = 1$. Consequently, it suffices to prove

$\hat{C}_{i,h}(u, v) \xrightarrow{a.s.} C_{i,h}(u, v)$ on the event A_i .

Specifically,

$$\begin{aligned} \hat{C}_{i,h}(u, v) &= \frac{1}{T_i - h} \sum_{t=1}^{T_i-h} I(U_{it} \leq u) I(V_{it} \leq v) \\ &= \frac{1}{T_i - h} \sum_{t=1}^{T_i-h} I\left(\frac{T_i}{T_i + 1} \hat{F}_i(X_{it}) \leq u, \frac{T_i}{T_i + 1} \hat{F}_i(X_{i(t+h)}) \leq v\right) \\ &= \frac{1}{T_i - h} \sum_{t=1}^{T_i-h} I\left(F_i(X_{it}) \leq \frac{T_i + 1}{T_i} u - (\hat{F}_i(X_{it}) - F_i(X_{it}))\right) \\ &\quad I\left(F_i(X_{i(t+h)}) \leq \frac{T_i + 1}{T_i} v - (\hat{F}_i(X_{i(t+h)}) - F_i(X_{i(t+h)}))\right) \end{aligned}$$

On the event A_i , for an arbitrary ϵ , there exists a T such that $|\hat{F}_i(x) - F_i(x)| < \epsilon$ for $\forall x \in (-\infty, \infty)$. Hence we have that $I\left(F_i(X_{it}) \leq \frac{T_i + 1}{T_i} u - \epsilon\right) \leq$

$$I\left(F_i(X_{it}) \leq \frac{T_i + 1}{T_i} u - (\hat{F}_i(X_{it}) - F_i(X_{it}))\right) \leq I\left(F_i(X_{it}) \leq \frac{T_i + 1}{T_i} u + \epsilon\right) \text{ and}$$

$$I\left(F_i(X_{i(t+h)}) \leq \frac{T_i + 1}{T_i} v - \epsilon\right) \leq I\left(F_i(X_{i(t+h)}) \leq \frac{T_i + 1}{T_i} v - (\hat{F}_i(X_{i(t+h)}) - F_i(X_{i(t+h)}))\right) \leq$$

$$I\left(F_i(X_{i(t+h)}) \leq \frac{T_i + 1}{T_i} v + \epsilon\right). \text{ Consequently,}$$

$$\frac{1}{T_i - h} \sum_{t=1}^{T_i-h} I\left(F_i(X_{it}) \leq \frac{T_i + 1}{T_i} u - \epsilon, F_i(X_{i(t+h)}) \leq \frac{T_i + 1}{T_i} v - \epsilon\right) \leq \hat{C}_{i,h}(u, v),$$

and

$$\frac{1}{T_i - h} \sum_{t=1}^{T_i-h} I\left(F_i(X_{it}) \leq \frac{T_i + 1}{T_i} u + \epsilon, F_i(X_{i(t+h)}) \leq \frac{T_i + 1}{T_i} v + \epsilon\right) \geq \hat{C}_{i,h}(u, v).$$

This means that $\tilde{C}_{i,h}\left(\frac{T_i + 1}{T_i} u - \epsilon, \frac{T_i + 1}{T_i} v - \epsilon\right) \leq \hat{C}_{i,h}(u, v) \leq \tilde{C}_{i,h}\left(\frac{T_i + 1}{T_i} u + \epsilon, \frac{T_i + 1}{T_i} v + \epsilon\right)$. Then let $T_i \rightarrow \infty$, by the result of Step 1, with probability 1 we have

$$C_{i,h}(u - \epsilon, v - \epsilon) \leq \lim_{T_i \rightarrow \infty} \hat{C}_{i,h}(u, v) \leq \overline{\lim}_{T_i \rightarrow \infty} \hat{C}_{i,h}(u, v) \leq C_{i,h}(u + \epsilon, v + \epsilon).$$

Finally, by the arbitrary property of ϵ , one can obtain that as $T_i \rightarrow \infty$,

$$\widehat{C}_{i,h}(u, v) \xrightarrow{a.s.} C_{i,h}(u, v).$$

Step 4. In this step, we will prove that $\widehat{D}_h(i, i') \xrightarrow{a.s.} D_h(i, i')$ as $T_i, T_{i'} \rightarrow \infty$. By the results of Step 3, we have as $T_i, T_{i'} \rightarrow \infty$, $\widehat{C}_{i,h}(u, v) \xrightarrow{a.s.} C_{i,h}(u, v)$ and $\widehat{C}_{i',h}(u, v) \xrightarrow{a.s.} C_{i',h}(u, v)$, which means that $(\widehat{C}_{i,h}(u, v) - \widehat{C}_{i',h}(u, v))^2 \xrightarrow{a.s.} (C_{i,h}(u, v) - C_{i',h}(u, v))^2$. Moreover, $(\widehat{C}_{i,h}(u, v) - \widehat{C}_{i',h}(u, v))^2 \leq 1$, then by the dominated convergence theorem, one can obtain that $\widehat{D}_h(i, i') \xrightarrow{a.s.} D_h(i, i')$ as $T_i, T_{i'} \rightarrow \infty$.

This completes the whole proof of Theorem 1.

Appendix B: Proof of Proposition 1

By the Eq (2), we have

$$\widehat{D}_h(i, i') = \sqrt{\int \int_{[0,1]^2} (\widehat{C}_{i,h}(u, v)^2 - 2\widehat{C}_{i,h}(u, v)\widehat{C}_{i',h}(u, v) + \widehat{C}_{i',h}(u, v)^2) dudv}.$$

By the definition of $\widehat{C}_{i,h}(u, v)$, one can see that

$$\begin{aligned} & \int \int_{[0,1]^2} \widehat{C}_{i,h}(u, v)^2 dudv \\ &= \frac{1}{(T_i - h)^2} \int \int_{[0,1]^2} \sum_{t=1}^{T_i-h} \sum_{t'=1}^{T_i-h} I(U_{it} \leq u) I(V_{it} \leq v) I(U_{it'} \leq u) I(V_{it'} \leq v) dudv \\ &= \frac{1}{(T_i - h)^2} \sum_{t=1}^{T_i-h} \sum_{t'=1}^{T_i-h} \int_0^1 I(U_{it} \leq u) I(U_{it'} \leq u) du \cdot \int_0^1 I(V_{it} \leq v) I(V_{it'} \leq v) dv \\ &= \frac{1}{(T_i - h)^2} \sum_{t=1}^{T_i-h} \sum_{t'=1}^{T_i-h} (1 - \max(U_{it}, U_{it'})) (1 - \max(V_{it}, V_{it'})) \\ &= L_{i,i}. \end{aligned}$$

Similarly, one can also verify that $\int \int_{[0,1]^2} \widehat{C}_{i,h}(u, v) \widehat{C}_{i',h}(u, v) dudv = L_{i,i'}$. Hence we have that

$$\widehat{D}_h(i, i') = \sqrt{L_{i,i} - 2L_{i,i'} + L_{i',i'}}.$$

This completes the proof of Proposition 1.

Appendix C: Proof of Theorem 2

By the results of Theorem 1, we have that as $\min_i \{T_i\} \rightarrow \infty$, $\widehat{D}_h(i, i') \xrightarrow{a.s.} D_h(i, i')$ for all $i, i' = 1, \dots, n$. Hence one can prove the Theorem on the event

$\{\widehat{D}_h(i, i') \rightarrow D_h(i, i') \text{ for all } i, i' = 1, \dots, n\}$, which is denoted by \mathcal{E} .

Assume that X_i and $X_{i'}$ belong to the j -th cluster and the j' -th cluster respectively, then we have that $D_h(i, i') = D_{0,h}(j, j')$. One can see that $D_{0,h}(j, j') = 0$ for $j = j'$, and $D_{0,h}(j, j') \geq \epsilon$ for $j \neq j'$. Consequently, on the event \mathcal{E} , as long as $\min_i \{T_i\}$ is large enough we have $\widehat{D}_h(i, i') < \epsilon/2$ if $X_i, X_{i'}$ belong to a common cluster, and $\widehat{D}_h(i, i') > \epsilon/2$ if $X_i, X_{i'}$ belong to different clusters. Let $(i_1, i'_1) = \arg \min_{(i, i')} \widehat{D}_h(i, i')$, then $X_{i_1}, X_{i'_1}$ must belong to the common cluster, and for $J = n$ in the algorithm, $X_{i_1}, X_{i'_1}$ are merged together as a cluster. This means that for $J = n$, the theorem holds.

Next assume the theorem holds for $J = J_1 + 1 > J_0 + 1$, we will prove the theorem still holds for $J = J_1$. For $J = J_1 + 1$, we denote the J_1 clusters as $\widehat{\mathcal{M}}_1, \dots, \widehat{\mathcal{M}}_{J_1}$. Here one should define the dissimilarities among clusters. Without loss of generality, we only consider the single linkage method here and the proof for other dissimilarities among clusters is similar. For single linkage, the distance between $\widehat{\mathcal{M}}_j, \widehat{\mathcal{M}}_{j'}$ is defined as $\widehat{D}(\widehat{\mathcal{M}}_j, \widehat{\mathcal{M}}_{j'}) = \min_{X_i \in \widehat{\mathcal{M}}_j, X_{i'} \in \widehat{\mathcal{M}}_{j'}} \widehat{D}(i, i')$. Based on the assumption that the theorem holds for $J = J_1 + 1 > J_0$, we know that time series in each $\widehat{\mathcal{M}}_j$ share a common copula function. Then we have that $\widehat{D}(\widehat{\mathcal{M}}_j, \widehat{\mathcal{M}}_{j'}) < \epsilon/2$ if $\widehat{\mathcal{M}}_j, \widehat{\mathcal{M}}_{j'}$ belong a common cluster, and $\widehat{D}(\widehat{\mathcal{M}}_j, \widehat{\mathcal{M}}_{j'}) > \epsilon/2$ otherwise. Furthermore, due to $J_1 > J_0$ there exist at least two $\widehat{\mathcal{M}}_j, \widehat{\mathcal{M}}_{j'}$, time series in which also share a common copula function. Let $(\widehat{j}, \widehat{j}') = \arg \min_{j \neq j'} \widehat{D}(\widehat{\mathcal{M}}_j, \widehat{\mathcal{M}}_{j'})$. For $J = J_1$, $\widehat{\mathcal{M}}_{\widehat{j}}$ and $\widehat{\mathcal{M}}_{\widehat{j}'}$ will be merged together as a new cluster. Moreover, by the above analysis, we know $\widehat{\mathcal{M}}_{\widehat{j}}$ and $\widehat{\mathcal{M}}_{\widehat{j}'}$ share a common copula function, and they are truly belong to a common cluster. This means that the claims of Theorem 2 is true for $J = J_1$.

This completes the whole proof of Theorem 2.

Acknowledgments

We thank the Editor, the AE and two referees for their helpful comments and valuable suggestions, which greatly improved the article. The research of Baiguo An and Beibei Zhang is supported by National Natural Science Foundation of China (No.11601349), scientific research planned project of National Bureau of Statistics of China (No. 2017LZ15).

Author Contributions

Formal analysis: Beibei Zhang.

Funding acquisition: Baiguo An.

Investigation: Baiguo An.

Methodology: Baiguo An.

Software: Beibei Zhang.

Validation: Beibei Zhang, Baiguo An.

Writing – original draft: Beibei Zhang, Baiguo An.

References

1. Frühwirth-Schnatter S, Kaufmann S. Model-based clustering of multiple time series. *Journal of Business & Economic Statistics*. 2008; 26(1):78–89. <https://doi.org/10.1198/073500107000000106>
2. Xiong Y, Yeung DY. Time series clustering with ARMA mixtures. *Pattern Recognition*. 2004; 37(8):1675–1689. <https://doi.org/10.1016/j.patcog.2003.12.018>
3. Otranto E. Clustering heteroskedastic time series by model-based procedures. *Computational Statistics & Data Analysis*. 2008; 52(10):4685–4698. <https://doi.org/10.1016/j.csda.2008.03.020>
4. Ramoni M, Sebastiani P, Cohen P. Bayesian Clustering by Dynamics. *Machine Learning*. 2002; 47(1):91–121. <https://doi.org/10.1023/A:1013635829250>
5. Oates T, Firoiu L, Cohen PR. Clustering Time Series with Hidden Markov Models and Dynamic Time Warping. In *Proceedings of the IJCAI-99 Workshop on Neural, Symbolic and Reinforcement Learning Methods for Sequence Learning*; 1999.
6. Montero P, Vilar JA. TSclust: An R package for time series clustering. *Journal of Statistical Software*. 2014; 62(1):1–43. <https://doi.org/10.18637/jss.v062.i01>

7. Liao TW. Clustering of time series data: a survey. *Pattern Recognition*. 2005; 38(11):1857–1874. <https://doi.org/10.1016/j.patcog.2005.01.025>
8. Piccolo D. A distance measure for classifying ARIMA models. *Journal of Time Series Analysis*. 1990; 11(2):153–164. <https://doi.org/10.1111/j.1467-9892.1990.tb00048.x>
9. Maharaj EA. Cluster of time series. *Journal of Classification*. 2000; 17(2):297–314. <https://doi.org/10.1007/s003570000023>
10. Kalpakis K, Gada D, Puttagunta V. Distance measures for effective clustering of ARIMA time-series. In: *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*. IEEE; 2001. p. 273–280.
11. Corduas M, Piccolo D. Time series clustering and classification by the autoregressive metric. *Computational Statistics & Data Analysis*. 2008; 52(4):1860–1872. <https://doi.org/10.1016/j.csda.2007.06.001>
12. Liu S, Maharaj EA. A hypothesis test using bias-adjusted AR estimators for classifying time series in small samples. *Computational Statistics & Data Analysis*. 2013; 60:32–49. <https://doi.org/10.1016/j.csda.2012.11.014>
13. Galeano P, Peña DP. Multivariate Analysis in Vector Time Series. *Resenhas*. 2000; 4:383–404.
14. D'Urso P, Maharaj EA. Autocorrelation-based fuzzy clustering of time series. *Fuzzy Sets and Systems*. 2009; 160(24):3565–3589. <https://doi.org/10.1016/j.fss.2009.04.013>
15. Caiado J, Crato N, Peña D. A periodogram-based metric for time series classification. *Computational Statistics & Data Analysis*. 2006; 50(10):2668–2684. <https://doi.org/10.1016/j.csda.2005.04.012>
16. Bohte Z, Cepar D, Kosmelj K. Clustering of time series. In: *Compstat*. vol. 80; 1980. p. 587–593.
17. Díaz SP, Vilar JA. Comparing several parametric and nonparametric approaches to time series clustering: A simulation study. *Journal of classification*. 2010; 27(3):333–362. <https://doi.org/10.1007/s00357-010-9064-6>
18. Rémillard B, Scaillet O. Testing for equality between two copulas. *Journal of Multivariate Analysis*. 2009; 100(3):377–386. <https://doi.org/10.1016/j.jmva.2008.05.004>
19. Sklar M. Fonctions de répartition à n dimensions et leurs marges. *Publ inst statist univ Paris*. 1959; 8:229–231.
20. Lafuente-Rego B, Vilar J. Clustering of time series using quantile autocovariances. *Advances in Data Analysis & Classification*. 2016; 10(3):391–415. <https://doi.org/10.1007/s11634-015-0208-8>
21. Kaufman L, Rousseeuw PJ. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons; 2009.
22. Luxburg U. A tutorial on spectral clustering. *Statistics & Computing*. 2007; 17(4):395–416. <https://doi.org/10.1007/s11222-007-9033-z>
23. Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning* 2nd edition. New York: Springer; 2009.
24. James G, Witten D, Hastie T, Tibshirani R. *An introduction to statistical learning with Applications in R*. Springer; 2013.
25. Lance GN, Williams WT. A General Theory of Classificatory Sorting Strategies. Hierarchical systems. *The Computer Journal*. 1967; 9(4):373–380. <https://doi.org/10.1093/comjnl/9.4.373>
26. Batagelj V. Generalized Ward and Related Clustering Problems. In: *Hh Bock, Classification & Related Methods of Data Analysis*; 1988. p. 67–74.
27. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*. 1987; 20:53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
28. Maharaj EA. A significance test for classifying ARMA models. *Journal of Statistical Computation and Simulation*. 1996; 54(4):305–331. <https://doi.org/10.1080/00949659608811737>
29. Gavrilov M, Anguelov D, Indyk P, Motwani R. Mining the stock market (extended abstract): which measure is best? In: *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM; 2000. p. 487–496.
30. Tong H, Yeung I. On Tests for Self-Exciting Threshold Autoregressive-type Nonlinearity in Partially Observed Time-Series. *Applied Statistics-Journal of the Royal Statistical Society Series C*. 1991; 40(1):43–62.
31. Petrucci JD, Woolford S. A Threshold AR(1) Model. *Journal of Applied Probability*. 1984; 21(02):270–286. <https://doi.org/10.1017/S0021900200024670>
32. Chan KS, Petrucci JD, Tong H, Woolford S. A Multiple-Threshold AR(1) Model. *Journal of Applied Probability*. 1985; 22(02):267–279. <https://doi.org/10.2307/3213771>

33. Chan KS, Tsay RS. Limiting properties of the least squares estimator of a continuous threshold autoregressive model. *Biometrika*. 1998; 85(2):413–426. <https://doi.org/10.1093/biomet/85.2.413>
34. Liu W, Ling S, Shao Q. On non-stationary threshold autoregressive models. *Bernoulli*. 2011; 17(3):969–986. <https://doi.org/10.3150/10-BEJ306>
35. Borg I, Groenen PJ. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media; 2005.
36. Atkinson AB, Bourguignon F. *Handbook of Income Distribution*. Elsevier; 2000.
37. Fan J. *Nonlinear time series: nonparametric and parametric methods*. Springer; 2003.
38. Tucker HG. A Generalization of the Glivenko-Cantelli Theorem. *Annals of Mathematical Statistics*. 1959; 30(3):828–830. <https://doi.org/10.1214/aoms/1177706212>