Journal of Classification 35:394-421 (2018)
DOI: 10.1007/s00357-018-9271-0

Nonlinear Time Series Clustering Based on Kolmogorov-Smirnov 2D Statistic

Beibei Zhang

Capital University of Economics and Business, Beijing, China

Rong Chen

Rutgers University, New Jersey, USA

Abstract: Time series clustering is to assign a set of time series into groups that share certain similarity. It has become an attractive analytic tool as many applications require such classifications. Clustering may also result in more accurate parameter estimates when a group of time series are assumed to share common models and parameters, especially for short panel time series. Many existing time series clustering methods are based on the assumption that the time series are linear. However, linearity assumptions often fail to hold. In this paper we consider the problem of clustering nonlinear time series. We propose the use of a two dimensional Kolmogorov-Smirnov statistic as a distance measure of two time series by measuring the affinity of nonlinear serial dependence structures. It is nonparametric in nature hence no model assumption are needed. The approach is illustrated with simulation studies as well as real data examples.

Keywords: Cross validation; Dissimilarity measure; Hierarchical clustering; Generalized Ward's linkage.

B. Zhang's research is sponsored in part by National Bureau of Statistics of China scientific research planned project 2013LZ22 and National Natural Science Foundation of China 11301351, 11601349. R. Chen's research is partially supported under NSF grants DMS 1209085, DMS 1513409 and DMS 1737857.

Corresponding Author's Address: R. Chen, Department of Statistics and Biostatistics, Rutgers University, NJ 08854, USA, email: rongchen@stat.rutgers.edu.

Published online: 9 October 2018

1. Introduction

Clustering is an unsupervised learning method aimed at classifying objects under study into homogeneous groups so that the within-group distance is minimized and between-group distance is maximized. Objects determined to be similar can then be further analyzed to understand the underlying common structure and gain insights.

Due to the recent advances of data collection tools and the need of various applications, large panel of time series are observed more and more frequently. Often, a common model is used for all the time series, especially in the case of short time series when pooling is important to improve the accuracy of model estimation and prediction. But such an approach fails if the data-generating mechanism differs among the time series. For the purpose of improving estimation and forecast performance, within a panel of time series only those with similar dynamic properties should be pooled. Clustering hence becomes a crucial step in the analysis. It has become an important area of research, with applications in a wide variety of fields, including economics, marketing, business, finance, medicine, biology, physics, psychology, zoology, and many others. For example, Zhang (2013) developed a consistent clustering method for time series data based on trend parallelism, and applied it analyzing daily cellphone download activities to identify homogeneous subgroups for efficient advertising and marketing management. Ma and Zhong (2008) proposed a clustering method for large-scale functional data with multiple covariates, motivated by the need of clustering temporal expression data. Frühwirth-Schnatter and Kaufmann (2008) applied a model-based clustering method on regional income series and industrial production growth rates of some countries. An interesting overview on time series clustering methods and applications can be found in Liao (2005).

In any clustering problem, the starting point is to define a distance or dissimilarity measure and form a proximity matrix of the observed objects. Then various clustering algorithms can be applied to form homogeneous subgroups. However, clustering complex data type, such as a group of time series, generally requires specially designed dissimilarity measures, instead of using standard Euclidean distances of numerical vectors. The choice of the distance measure should of course take into account of the objective of the clustering procedure. Thus, the distance should be able to capture the particular discrepancies between observations that are relevant to the final objective. In this study, we focus on grouping time series according to their underlying nonlinear serial dependence structure.

Time series clustering has received more and more attention. Despite its fast advancement, most of the existing clustering methods are originally designed for linear time series and dissimilarity measures are con-

structed assuming that the underlying generating processes are linear and Gaussian. For example, Piccolo (1990) and Corduas and Piccolo (2008) developed a dissimilarity measure for autoregressive integrated moving average (ARIMA) model based on an AR representation and applied this measure to the identification of similarities between industrial production series. Maharaj (2000) proposed a classification method using the p-value of a hypothesis test of comparison of AR parameters of two stationary linear time series as a measure of similarity. Kalpakis, Gada, and Puttagunta (2001) utilized Euclidean distance between the Linear Predictive Coding (LPC) cepstra of two time series as the dissimilarity measure for clustering ARIMA time series. Xiong and Yeung (2004) proposed a model-based approach using mixtures of ARMA models. Recently, a hypothesis test was proposed by Liu and Maharaj (2013) for classifying stationary time series based on a bias-adjusted estimator of the fitted AR models. Most of these model-based measures are based on the linear assumption of the underlying process, and can not be generalized to nonlinear time series directly. However, a linear model is just a first and easy step in modeling an unknown dynamic relationship. The world is mostly nonlinear.

Since nonlinear relationship is often difficult to specify without strong prior knowledge, clustering methods based on nonparametric distance becomes an attractive option. Bohte, Cepar, and Kosmelj (1980) introduced several descriptive dissimilarity measures based on the comparison of autocorrelation functions (ACF). Galeano and Peña (2000) considered the Mahalanobis distance between ACFs. Caiado, Crato, and Peña (2006) applied the partial autocorrelation function (PACF) and inverse autocorrelation functions (IACF) to classify whether time series is stationary or non-stationary. D'Urso and Maharaj (2009) proposed a fuzzy clustering approach based on ACF. Although ACF related statistics have been widely used as one of the primary tools for exploring and testing times series, they only measure the Pearson correlation between X_t and X_{t+h}, reflecting the strength of linear dependence. Granger, Maasoumi, and Racine (2004) found that the usual correlation function measures are inadequate in recognizing nonlinear dependence. Some nonlinear time series have ACF and PACF behaving like white noise. In frequency domain, Caiado et al. (2006) proposed three different dissimilarity measures based on periodograms. Díaz and Vilar (2010) proposed several nonparametric distance measures based on the spectral densities of time series. Since ACF and spectral density are related by transformation, they carry the same information with respect to the dependence of a process. Hence most of the nonparametric dissimilarities mentioned above are all Pearson correlation related quantities, which can only measure linear dependence. They are not expected to perform well on nonlinear time series clustering.

There are very few studies of nonlinear time series clustering in the literature. Vilar, Alonso, and Vilar (2010) focused on grouping nonlinear time series that have similar or equal predictions, assuming the underlying generating process follows a nonparametric autoregressive model. Harvill, Ravishanker, and Ray (2013) proposed a procedure to distinguish among various nonlinear time series based on a distance measure computed from the square modulus of the estimated normalized bispectra. Lafuente-Rego and Vilar (2016) proposed a metric based on quantile autocovariances, which is shown to be able to cluster nonlinear and conditional heteroscedastic processes.

Nonlinear time series models have been used extensively to model complex dynamics not adequately represented by linear models (Tong, 1990; Fan, 2003). Since nonlinear time series have infinite number of possible forms, it is not easy to identify a proper time series dependence structure in practice. In this paper we avoid the use of model-based clustering method, but turn to a more flexible nonparametric approach. The intuition behind our proposed dissimilarity measure is that, if two series follow the same underlying model, their stationary marginal distribution and the joint distributions of lag variables should be the same. A classic nonparametric statistics, the Kolmogorov-Smirnov (KS) statistics, is proposed as the dissimilarity measure. Specifically, for two time series X_t and Y_t, we use a two dimensional KS statistics to measure the difference between the joint distribution of (X_t, X_{t+h}) and (Y_t, Y_{t+h}). By summing up the statistics for $h = 1, \ldots, K$, we obtain a measure of discrepancy of the serial dependence structure.

The main aim of this paper is to introduce a new distance measure for clustering nonlinear time series. It tries to overcome the limitation of the existing distance measures designed for linear processes. The proposed distance is based on a generalization of the KS statistic to two-dimensional distributions. Simulation studies show that, compared to other dependence-based distance, the proposed metric has good performance, particularly in its ability of classifying nonlinear models and discriminating between nonstationary and stationary models. Although the proposed distance measure is designed for nonlinear time series, empirical study shows that it also works for clustering linear processes. This is because the distance is defined in terms of discrepancy of the serial dependence structure, which includes linear structures.

The remaining of the paper is organized as follows. In Section 2, we briefly review the two dimensional KS statistic. In Section 3, we introduce the proposed clustering method with a discussion of some relevant statistical properties. Some simulation and real data results are shown in Section 4 and Section 5.

2. Two Dimensional Kolmogorov-Smirnov Test

The KS test is a nonparametric test of the equality of two probability distributions. One sample KS test compares the underlying distribution of a random sample with a reference probability distribution, while two sample KS test compares two distributions using two random samples.

The one-dimensional and one-sample KS test uses the largest absolute difference between the cumulative distributions of the reference population and the empirical distribution of a sample, denoted as D_{KS}. When two distributions are the same, the asymptotic distribution of the test statistic $Z_n = \sqrt{n}D_{KS}$, as the same sample n goes to infinity, follows

$$P(Z_n \leq x) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}exp(-2k^2x^2).$$

Two-sample KS test is defined similarly (Conover, 1999, Ch. 6).

We here discuss a useful generalization of the KS test to 2-dimensional distributions. This generalization is due to Fasano and Franceschini (1987), a variant on an earlier idea due to Peacock (1983). In a 2-dimensional distribution, each data point is characterized by a (x, y) pair of values. However, cumulative probability distribution is not well-defined. Fasano and Franceschini (1987) made use of the total number of points in each of the four quadrants anchored at a given point (x_i, y_i), namely, the fraction of data points in the regions $(x < x_i, y < y_i), (x < x_i, y > y_i), (x > x_i, y < y_i), (x > x_i, y > y_i)$. Within each quadrant, the difference between the observed and expected proportion of points is determined. The two dimensional KS (2DKS) statistics D_{2DKS} is defined as the maximum difference between data fractions in any two matching quadrants of the sample and of the parent population, ranging over all data points.

Specifically, given a point (c, d) we denote the probability of two random variables (X, Y) in the four natural quadrants around the point (c, d) as $p_{X,Y}(c+, d+)$, $p_{X,Y}(c-, d+)$, $p_{X,Y}(c+, d-)$ and $p_{X,Y}(c-, d-)$. With a set of samples of (X, Y), we use the fraction of data points in each quadrant as the estimate of the corresponding probabilities.

Let $S_1 = \{(x_{1i}, y_{1i}), i = 1, \ldots, n_1\}$ be the sample of (X_1, Y_1) and $S_2 = \{(x_{2j}, y_{2j}), j = 1, \ldots, n_2\}$ be that of (X_2, Y_2). Let D_{2DKS} be the maximum difference of the corresponding quadrant probabilities of (X_1, Y_1) and (X_2, Y_2) on the unit set of S_1 and S_2, i.e.

$$\begin{aligned} D_{2DKS} = \max_{(c,d)\in S_1\cup S_2} \{&|\hat{p}_{X_1,Y_1}(c+, d+) - \hat{p}_{X_2,Y_2}(c+, d+)|, \\ &|\hat{p}_{X_1,Y_1}(c-, d+) - \hat{p}_{X_2,Y_2}(c-, d+)|, \\ &|\hat{p}_{X_1,Y_1}(c+, d-) - \hat{p}_{X_2,Y_2}(c+, d-)|, \\ &|\hat{p}_{X_1,Y_1}(c-, d-) - \hat{p}_{X_2,Y_2}(c-, d-)|\}. \end{aligned} \tag{2.1}$$

Similar to the one dimensional case, $Z \equiv D_{2DKS}\sqrt{\frac{n_1 n_2}{n_1+n_2}}$ can be used as the test statistic. One would reject the null hypothesis of H_0 : $F_{X_1,Y_1}(\cdot,\cdot) = F_{X_2,Y_2}(\cdot,\cdot)$, if Z is significantly large. Fasano and Franceschini (1987) tabulated Monte Carlo result for the distribution of D_{2DKS} and Z under various setting. Lopes, Reid, and Hobson (2007) and Lopes, Hobson, and Reid (2008) surveyed and evaluated several algorithms version of the 2DKS test. Note that these results are obtained under *i.i.d* observations. For depended data, the testing problem is much more complicated. In this paper, we directly use D_{2DKS} as a distance measure, hence testing at a certain significant level is not our focus.

3. Time Series Clustering Method Based on 2DKS Distance Measure

Time series is mostly characterized by its serial dependency. For linear Gaussian process, the autocorrelation function completely determines the process. To cluster nonlinear time series, we are also interested in the similarity between their serial dependence structure. The existing literature on testing and measuring nonlinear dependence (Diks, 2009; Granger, Maasoumi, and Racine, 2004; Dufour, Lepage, and Zeidan, 1982) are mainly based on measuring the degree of deviation from independence, and do not discern nonlinear dynamic structures.

In this paper, we focus on serial dependence by examining the joint distribution of X_t and X_{t+h}. The lagged joint distribution provide abundant information on the dynamic of the underlying process. When two time series follow different models, they often have very different lagged joint distribution. For example Figure 1 displays two lagged scatterplots (X_{t-1}, X_t), where the series are simulated from a Bilinear model and an exponential autoregressive (EXPAR) model respectively. Numerically, 2DKS statistics can be used to test the difference between two 2-dimensional distributions.

Let $\{x_t, t = 1, 2, \ldots, n_1\}$ and $\{y_t, t = 1, 2, \ldots, n_2\}$ be the observed series of two stationary process $\{X_t\}$ and $\{Y_t\}$ respectively. We define $d_h(x_t, y_t)$ the dissimilarity between $\{x_t\}$ and $\{y_t\}$ at lag h, as the 2DKS statistics of (X_t, X_{t+h}) and (Y_t, Y_{t+h}) defined in (2.1). This dissimilarity measure has the usual properties of a metric distance as follows:

(1) (non-negativity) $0 \leq d_h(x_t, y_t) \leq 1$. If $\{x_t\} = \{y_t\}$ then $d_h(x_t, y_t) = 0$.
(2) (symmetry) $d_h(x_t, y_t) = d_h(y_t, x_t)$.
(3) (triangle inequality) $d_h(x_t, y_t) \leq d_h(x_t, z_t) + d_h(y_t, z_t)$, where x_t, y_t, z_t are three stationary processes.

It is straightforward to show that the distance $d_h(x_t, y_t)$ satisfies the non-negativity and symmetry conditions. To prove the triangle inequality, we define three two-dimensional sets: $S_x = \{(x_t, x_{t+h}), t = 1, 2, \ldots, n_1 -$

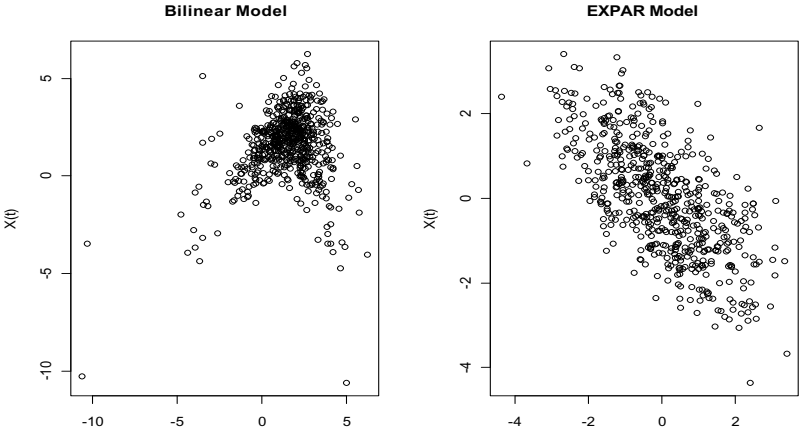


Figure 1. The lagged scatterplot (X_{t-1}, X_t) of time series

$h\}$, $S_y = \{(y_t, y_{t+h}), t = 1, 2, \ldots, n_2 - h\}$, and $S_z = \{(z_t, z_{t+h}), t = 1, 2, \ldots, n_3 - h\}$. Let (c, d) be any point in $S_x \cup S_y \cup S_z$. Since

$$\begin{aligned} |\hat{p}_{X_t,X_{t+h}}(c+, d+) &- \hat{p}_{Y_t,Y_{t+h}}(c+, d+)| \\ &\leq |\hat{p}_{X_t,X_{t+h}}(c+, d+) - \hat{p}_{Z_t,Z_{t+h}}(c+, d+)| \\ &+ |\hat{p}_{Z_t,Z_{t+h}}(c+, d+) - \hat{p}_{Y_t,Y_{t+h}}(c+, d+)|, \end{aligned}$$

we have

$$\begin{aligned} &\max_{(c,d)\in S_x\cup S_y\cup S_z} \{|\hat{p}_{X_t,X_{t+h}}(c+, d+) - \hat{p}_{Y_t,Y_{t+h}}(c+, d+)|\} \\ &\leq \max_{(c,d)\in S_x\cup S_y\cup S_z} \{|\hat{p}_{X_t,X_{t+h}}(c+, d+) - \hat{p}_{Z_t,Z_{t+h}}(c+, d+)|\} \\ &+ \max_{(c,d)\in S_x\cup S_y\cup S_z} \{|\hat{p}_{Z_t,Z_{t+h}}(c+, d+) - \hat{p}_{Y_t,Y_{t+h}}(c+, d+)|\}. \end{aligned} \tag{3.1}$$

Note that the empirical estimator $\hat{p}_{X_t,X_{t+h}}(\cdot, \cdot)$ only depends on the set S_x, and $\hat{p}_{Y_t,Y_{t+h}}(\cdot, \cdot)$ only on the set S_y. Hence, the maximum ranging over $S_x \cup S_y \cup S_z$ on the left hand side of (3.1) is equivalent to ranging over $S_x \cup S_y$. The two terms on the right hand side also have the similar property. Hence

$$\begin{aligned} &\max_{(c,d)\in S_x\cup S_y} \{|\hat{p}_{X_t,X_{t+h}}(c+, d+) - \hat{p}_{Y_t,Y_{t+h}}(c+, d+)|\} \\ &\leq \max_{(c,d)\in S_x\cup S_z} \{|\hat{p}_{X_t,X_{t+h}}(c+, d+) - \hat{p}_{Z_t,Z_{t+h}}(c+, d+)|\} \\ &+ \max_{(c,d)\in \cup S_y\cup S_z} \{|\hat{p}_{Z_t,Z_{t+h}}(c+, d+) - \hat{p}_{Y_t,Y_{t+h}}(c+, d+)|\}. \end{aligned}$$

Combine with similar inequalities for the other three quadrants, we have the triangle inequality

$$d_h(x_t, y_t) \leq d_h(x_t, z_t) + d_h(y_t, z_t).$$

To combine the impact of different lag value h, we define the distance between the two series x_t and y_t by combining $d_h(x_t, y_t)$ with different h ($h = 1, 2, \ldots, K$),

$$d_{2DKS}(x_t, y_t) = \sum_{h=1}^{K} w_h d_h(x_t, y_t), \tag{3.2}$$

where K is a suitable maximum lag order. The weight $w_h \geq 0$ can be used to assign different importance of different lag h. The linear combination remains to be a valid distance metric. Caiado et al. (2006) used the similar combining way for their ACF based clustering method. In our empirical studies we use equal weight w_h.

To cluster L time series, we obtain the pairwise dissimilarity matrix using the distance measure d_{2DKS} in (3.2), then based on this dissimilarity matrix, we do clustering using the commonly used clustering algorithms, such as the hierarchical clustering algorithm (Hastie, Tibshirani, and Friedman, 2009, p. 520) and Partitioning Around Medoids (PAM) algorithm (Kaufman and Rousseeuw, 2009, p. 68). Note that the clustering algorithms based on Euclidean distances such as k-mean algorithm and the standard Ward's linkage algorithm can not be used here as the proposed distance measure is not a Euclidean distance.

Remark 1: The distance measure d_{2DKS} in (3.2) depends on the parameter K. The most effective K depends on the unknown underlying serial dependence structure of the time series. Intuitively, a larger K enables the dissimilarity measure to detect a broader range of dependence structures, but it also tends to include more noises. Our simulation results (e.g. Table 3 in Section 4) show that in certain cases, a larger sample size seems to benefit more from using a large K, an indirect indication of the trade-off between signal and noise. Since different K yields different distance measures, comparing different K falls into the broader problem of how to evaluate and compare different dissimilarity measures in clustering, a challenging problem. Many criteria can be used, including the average Silhouette coefficient. In this paper, we choose to use the leave-one-out cross-validation criterion. In the time series setting, it is often the case that the strongest serial dependency occurs in small lags hence the 'optimal' K is often small and the result is often not too sensitive.

Remark 2: From our simulation studies shown in Section 4, it seems that the agglomerative hierarchical clustering algorithm using the Generalized Ward's linkage of Batagelj (1988) is particularly effective for the

proposed distance measure. The Generalized Ward's linkage extends the classical Ward's linkage, with a generalized notion of 'cluster center'. It has been shown that the Generalized Ward's linkage method satisfies the Lance-Williams formula (Lance and Williams, 1967) and shares the same parameters in the formula with the original Ward's linkage method. The use of the Generalized Ward's linkage with user-defined time series distance has been used in practice (e.g. Harvill et al., 2013; An, 2008; Kosmelj and Batagelj, 1990).

Remark 3: Computationally, the proposed clustering algorithm is efficient. For each pair of time series, the computational complexity of their 2DKS distance in (3.2) is of the order of $O(KTlog(T))$ (Xiao, 2017; Lopes et al., 2007), where K is the number of lag terms used in the measure. In comparison, d_{ACF}, d_{PACF}, and d_{PIC} all have complexity of $O(KT)$ (roughly), where K is the number of maximum lag considered and $K << T$. Vilar (2014) showed that d_M has computing times $O(T)$. In frequency domain, the periodogram-based distances require $O(Tlog(T))$ operations. The spectral dissimilarity measures involves integration of the differences between the spectral densities, with complexity $O(NTlog(T))$ where N is the number of subintervals for numerical integration. Hence, overall the proposed clustering algorithm has a similar computational complexity as most of the existing methods. All these distances included here will be described briefly in Section 4.

Remark 4: Since we are mainly interested in the difference in dynamic structure, we often normalize the time series first to remove location and scalar differences.

Remark 5: Our experiments show that d_{2DKS} is more sensitive to model differences than the differences on parameter values of the same model. It is in line with our objective to cluster the series according to their dynamic difference.

4. Simulation Studies

Díaz and Vilar (2010) reported the results of an extensive simulation study of comparing several parametric and nonparametric dissimilarity measures in different time series clustering setup. They considered three clustering settings: (i) to cluster non-linear time series models, (ii) to distinguish between stationary and non-stationary time series, and (iii) to classify different linear ARMA models. The setups are quite general and have a broad scope in many applications. For comparison, we adopt the same settings to demonstrate the finite sample behavior of the proposed 2DKS distance.

In each selection, the series were clustered via the hierarchical clustering algorithm using the Generalized Ward's linkage.

For comparison, in each simulation scenario we also performed clustering using some representative distance measures proposed in the existing literature, including ARIMA model-based measures, d_{PIC} of Piccolo (1990) and d_M of Maharaj (1996). We also make comparisons with model-free dissimilarity measures. In time domain, distance d_{ACF} and d_{PACF} are defined by Euclidean distance between estimated ACF and PACF, using a number of significant lags, while d_{ACFG} and d_{PACFG} include geometric weights decaying with the lag, $w_i = p(1-p)^i$, $0 < p < 1$. In our study, we considered different lag, and $p = 0.05$ are used. In the frequency domain, the dissimilarity measures are aimed to assess the discrepancy between the corresponding spectral densities. Caiado et al. (2006) proposed the distance measures based on the periodogram, specifically 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 frequency domain, $d_{W(DLS)}$, $d_{W(LK)}$, d_{GLK} and d_{ISD} are proposed by Díaz and Vilar (2010). They are different version of nonparametric spectral dissimilarity measures. The differences among them are that they applied different estimation method of spectral density and different discrepancy function. More details about these methods can be found in Vilar (2014).

In the simulation experiments, the ground truth is known in advance. Two clustering evaluation criteria based on known ground truth are used. One evaluates the clustering methods by comparing the cluster solutions with the true cluster partition, using the cluster similarity index of Gavrilov et al., (2000), defined as

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

where $G = \{G_1, G_2, \cdots, G_M\}$ are the set of the M true clusters, $A = \{A_1, A_2, \cdots, A_M\}$ is the cluster solution by a clustering method under evaluation, and

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

where $|\cdot|$ denotes the cardinality of the elements in each set. The similarity index has values ranging from 0 to 1, with 1 corresponding to the case when G and A are identical.

The other evaluation method uses a one-nearest-neighbour (1-NN) classifier evaluated by leave-one-out cross-validation (loo1NN). The 1-NN

classifier assigns each series to one cluster of G containing the nearest series according to the dissimilarity matrix, and then the proportion of correctly assigned series are calculated to evaluate the performance of the distance measure. The merit of loo1NN is that it is an intuitive and parameter-free evaluation procedure, which directly measures the efficiency of the distance regardless of the considered clustering algorithm (Manso and Vilar, 2013). For both of the criteria, the closer the index to 1, the better the cluster quality of the proposed distance.

4.1 Clustering of Non-linear Time Series

The first experiment is to access the performance of different measures in non-linear process clustering. These models were used in Tong and Yeung (1991) for linearity test and in the simulation study in Díaz and Vilar (2010). Specifically, each trial involved four different underlying models, with four series of length T=200 generated from each model. The models were: (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 $X_t = (0.3 - 10exp\{-X_{t-1}^2\})X_{t-1} + \varepsilon_t$; (3) Linear moving average (MA) model $X_t = \varepsilon_t - 0.4\varepsilon_{t-1}$; and (4) Non-linear moving average (NLMA) model $X_t = \varepsilon_t - 0.5\varepsilon_{t-1} + 0.8\varepsilon_{t-1}^2$.

In all cases, the error process $\varepsilon_t \sim N(0, 1)$ and white. Model (3) is linear and models (1), (2) and (4) are nonlinear in conditional mean. The experiment was repeated 100 times for each distance metric. Table 1 provides the clustering results under the two evaluation criteria.

We note that the clustering results based on 2DKS distance are generally better than the other distance measures. In particular, d_{2DKS} with $K = 1$ can distinguish almost perfectly among the nonlinear process, because all the models considered are order 1 autoregressive models. Figure 2 shows the scatterplots of (X_t, X_{t+1}) of a typical series of these models. Figure 3 shows the details of one clustering result using 2DKS distance and the Generalized Ward's linkage. If the dendrogram is divided into two clusters, TAR and EXPAR series are clustered in one group while NLMA and MA series are put into the second group. This result makes sense since the series from TAR and EXPAR models have an autoregressive structure while MA and NLMA process are composed of the error process. If we divide the dendrogram at the true number of the clusters (4 groups), the perfect cluster structure are formed. Even though the other value of K (up to 15) produce acceptable results, the clustering performance decreases as K increases.

Figure 4 is the multidimensional scaling (MDS) plot (Borg and Groenen, 2005) for visualizing, in two dimensions, the similarity among the time series based on the 2DKS distance. With the dissimilarity measures obtained

Table 1. Clustering of nonlinear processes. Series length $T = 200$. Number of trials $N = 100$. Generalized Ward's linkage.

Measure	Numbers of lags	Similarity Index	loo1NN	Measure	Similarity Index	loo1NN
	$L = 10$	0.723	0.601	(Model-based)		
d_{ACFG}	$L = 25$	0.679	0.565	d_{PIC}	0.793	0.741
	$L = 50$	0.692	0.574	d_M	0.755	0.702
	$L = 10$	0.722	0.645	(Periodograms)		
d_{PACF}	$L = 25$	0.697	0.535	d_P	0.545	0.389
	$L = 50$	0.664	0.489	d_{LP}	0.786	0.678
	$L = 10$	0.766	0.645	d_{NP}	0.571	0.332
d_{PACFG}	$L = 25$	0.736	0.616	d_{LNP}	0.577	0.348
	$L = 50$	0.742	0.614	(Non-parametric)		
	$K = 1$	**0.988**	**0.990**	$d_{W(DLS)}$	0.939	0.933
	$K = 2$	**0.977**	**0.974**	$d_{W(LK)}$	0.936	0.913
	$K = 3$	**0.936**	**0.952**	d_{GLK}	0.910	0.888
d_{2DKS}	$K = 4$	**0.915**	**0.923**	d_{ISD}	0.933	0.920
	$K = 5$	**0.885**	**0.900**			
	$K = 10$	**0.791**	**0.823**			
	$K = 15$	**0.725**	**0.752**			

L is the number of lags used to compute ACF and PACF distance measure.
K is the number of lags used for 2DKS distance.

from data, MDS plot seeks a lower-dimensional representation of the data that preserves the pairwise distances as closely as possible. Figure 4 shows a clear separation of the clusters, which can also be seen in the dendrogram in Figure 3.

We also use other existing distance measures to cluster the same set of time series. Both d_{PIC} and d_M are typical linear model-based method. In the nonlinear setting we consider here, the results are affected by the misspecification of the data generating process. Likewise d_{ACF} and d_{PACF} do not perform well. It suggests that ACF and PACF based distance measures are not appropriate for nonlinear process. For example, the first order autocorrelation coefficient of the TAR and EXPAR processes may not be significant even though the series have significant autoregressive dependency of order one. Therefore Pearson's correlation related quantities are not suitable to detect various nonlinear serial dependence structure. In the frequency domain, Euclidean distance between the periodograms, d_P, d_{LP}, d_{NP} and d_{LNP} yielded the worst result in this setting. Similar to autocorrelation function, periodogram mainly deals with linear correlation, so it has difficulties in distinguishing nonlinear structures. The distances based

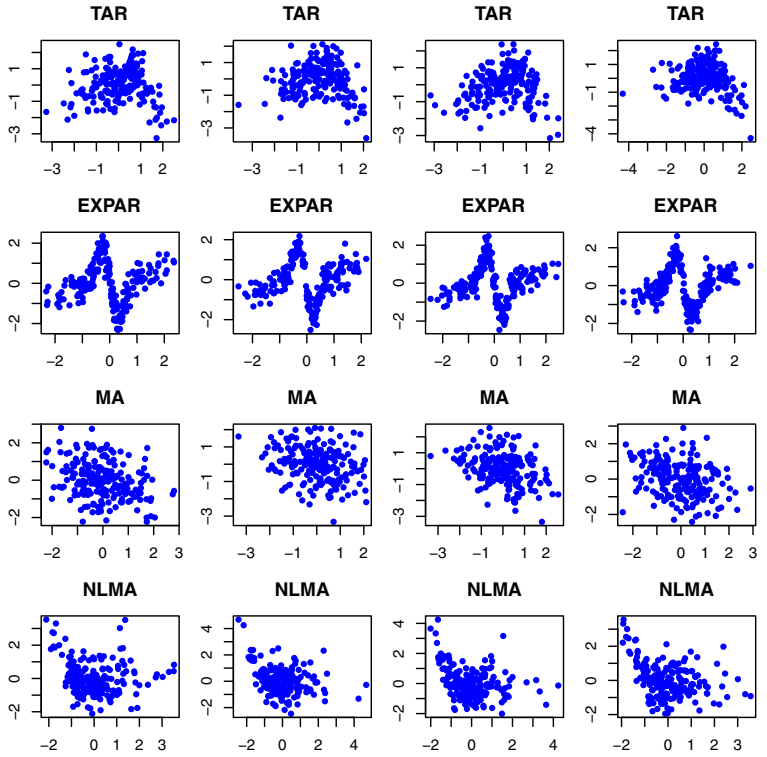


Figure 2. Lagged scatterplot (X_t, X_{t+1}) of nonlinear time series

on spectral density perform well in this scenario and led to similarity index around 0.9, comparable to that using 2DKS distance.

We also considered the performance of 2DKS distance using different clustering algorithms including PAM algorithm, single linkage algorithm (Gower and Ross, 1969), complete linkage algorithm (Defays, 1977), group average algorithm (Murtagh, 1984), and the agglomerative hierarchical clustering algorithm using the Generalized Ward's linkage. Figure 5 shows the results. It seems that the Generalized Ward's linkage and PAM method outperformed the other algorithm significantly. In general, the Generalized Ward's linkage always performs the best with equal cluster size. All of the five clustering algorithms attain their own best clustering result using $K = 1$, the true autoregressive order. The clustering similarity index decreases gradually as K increases.

To gain further insights into the 2DKS distance, a more difficult clustering task is conducted by increasing the differences between series within the same group. Specifically we generate 30 series from each model with

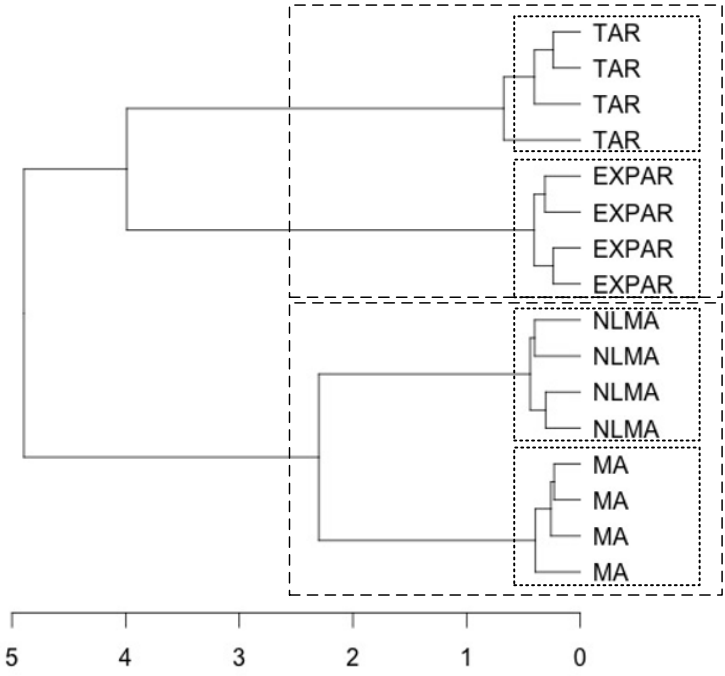


Figure 3. Generalized Ward's linkage dendrogram based on 2DKS distance

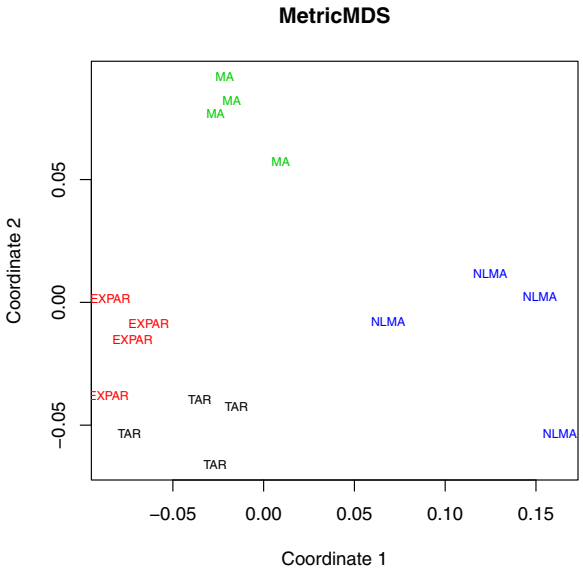


Figure 4. Multidimensional scaling plot for nonlinear time series based on 2DKS distance

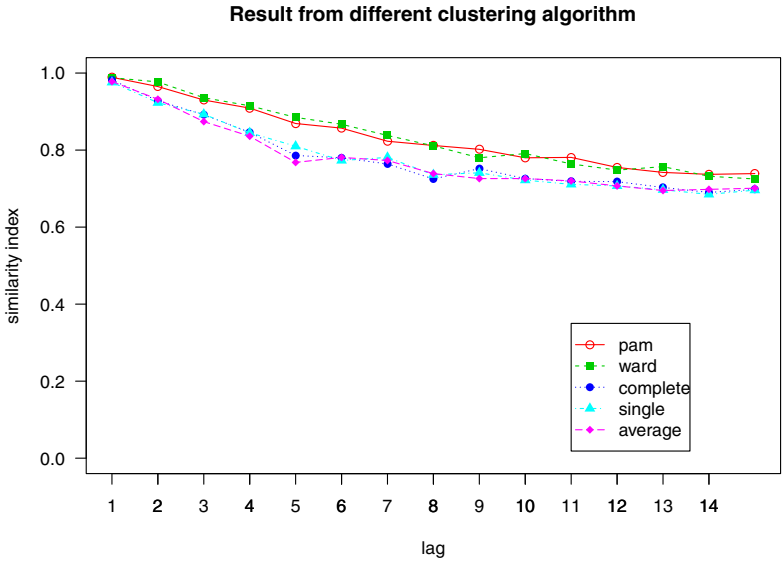


Figure 5. Clustering algorithm comparison when the maximum lag K varies from 1 to 15

varying coefficients, drawn from a given uniform distribution. The generating models are:
(1) Threshold autoregressive (TAR) model $X_t = a_1 X_{t-1} I(X_{t-1} \leq 0) - b_1 X_{t-1} I(X_{t-1} > 0) + \varepsilon_t, a_1 \sim U(0.3, 0.7), b_1 \sim U(1, 3)$;
(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) Linear moving average (MA) model $X_t = \varepsilon_t - a_3 \varepsilon_{t-1}, a_3 \sim U(0.2, 0.8)$; and
(4) Non-linear 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 Table 2. We found that the proposed 2DKS distance continues to produce the best performance among all distance measures.

In most cases, with a proper maximum order K, 2DKS distance is able to identify the genuine cluster structure. This simulation shows that 2DKS distance is robust to model coefficients, but more sensitive to the dynamic structure of the process, which allows us to group the same models in one cluster.

4.2 Classification of Time Series as Stationary or Non-Stationary

Although 2DKS distance measure was designed for nonlinear time series, it can be used to classify stationary and nonstationary time series

Table 2. Clustering of nonlinear processes with varying coefficients. 30 series in each group. Series length $T = 200$. Number of trials $N = 100$. Generalized Ward's linkage algorithm.

Measure	Numbers of lags	Similarity Index	loo1NN	Measure	Similarity Index	loo1NN
	$L = 10$	0.710	0.763	(Model-based)		
d_{ACFG}	$L = 25$	0.693	0.724	d_{PIC}	0.833	0.841
	$L = 50$	0.694	0.719	d_M	0.842	0.844
	$L = 10$	0.797	0.800	(Periodograms)		
d_{PACF}	$L = 25$	0.745	0.724	d_P	0.573	0.569
	$L = 50$	0.694	0.669	d_{LP}	0.790	0.758
	$L = 10$	0.818	0.814	d_{NP}	0.576	0.507
d_{PACFG}	$L = 25$	0.806	0.777	d_{LNP}	0.584	0.529
	$L = 50$	0.799	0.768	(Non-parametric)		
	$K = 1$	**0.992**	**0.992**	$d_{W(DLS)}$	0.929	0.941
	$K = 2$	**0.983**	**0.986**	$d_{W(LK)}$	0.926	0.928
	$K = 3$	**0.973**	**0.980**	d_{GLK}	0.900	0.873
d_{2DKS}	$K = 4$	**0.959**	**0.972**	d_{ISD}	0.930	0.932
	$K = 5$	**0.936**	**0.966**			
	$K = 10$	**0.827**	**0.933**			
	$K = 15$	**0.743**	**0.894**			

L is the number of lags used to compute ACF and PACF distance measure.
K is the number of lags used for 2DKS distance.

as well. This experiment is conducted by Caiado et al. (2006) and then extended by Díaz and Vilar (2010). It is done here to test the performance of 2DKS measure in distinguishing stationary or non-stationary series.

Consider a general ARIMA(*p,d,q*) process defined by

$$\phi(B)(1-B)^d x_t = \theta(B)\omega_t, t = 0, \pm 1, \ldots,$$

where *B* is the back-shift operator such that $B^r x_t = x_{t-r}$, $\phi(B) = 1 - \phi_1(B) - \ldots - \phi_p(B^p)$ is the *p*-order autoregressive operator, $\theta(B) = 1 - \theta_1(B) - \ldots - \theta_q(B^q)$ is the *q*-order moving average operator, *d* is the order of differentiating (so that $d = 0$ for a stationary process, $d \geq 1$ for non-stationary processes) and w_t is white noise. All differenced processes $y_t = (1-B)^d x_t$ are causal and invertible. As in Caiado et al. (2006), one realization from each of the following 12 ARIMA models (6 stationary models and 6 non-stationary models), were generated.

(i) AR(1) $\phi_1 = 0.9$
(ii) AR(2) $\phi_1 = 0.95, \phi_2 = -0.1$
(iii) ARMA(1,1) $\phi_1 = 0.95, \theta_1 = 0.1$

(iv) ARMA(1,1) $\phi_1 = -0.1, \theta_1 = -0.95$
(v) MA(1) $\theta_1 = -0.9$
(vi) MA(2) $\theta_1 = -0.95, \theta_2 = -0.1$
(vii) ARIMA(1,1,0) $\phi_1 = -0.1$
(viii) ARIMA(0,1,0)
(ix) ARIMA(0,1,1) $\theta_1 = 0.1$
(x) ARIMA(0,1,1) $\theta_1 = -0.1$
(xi) ARIMA(1,1,1) $\phi_1 = 0.1, \theta_1 = -0.1$
(xii) ARIMA(1,1,1) $\phi_1 = 0.05, \theta_1 = -0.05$

In all cases the error was Gaussian white noise with zero mean and unit variance. The experiment was performed with different lengths $T = 50, 200, 500$. As our objective here is the discrimination between stationary and nonstationary time series, we only focus on the last step of the agglomerative hierarchical clustering algorithm, where only two groups are formed. The clustering evaluation criterion consisted in computing the percentage of successes in the classification, namely, the percentage of time series that were classified in the correct cluster in accordance with their stationary or nonstationary behavior. The procedure was repeated $N = 300$ times and the percentage of success was averaged. Table 3 shows the result.

It can be seen that most of the time the 2DKS distance performs the best given an appropriate order. The result of 2DKS distance is comparable to Augmented Dickey-Fuller test (ADF test) and slightly inferior to Phillips-Perron test (PP test) of Perron (1987), both of which are designed specifically to test unit root type of nonstationary time series. However, 2DKS distance performs better than ARIMA model-based methods, most of the correlation based methods, periodograms based distances and nonparametric distances. In this example it seems that the choice of maximum lag K for 2DKS distance is important. A larger sample size seems to benefit more from a larger K.

The performance of 2DKS distance can be seen from Figure 6, which shows the MDS plot of one typical data set using 2DKS distance with $N = 500$ and $K = 10$. Figure 7 shows the lagged scatterplot of (X_t, X_{t+1}) of the 12 series, the difference of which can be recognized by 2DKS distance.

4.3 Clustering of ARMA Time Series

In this section, we will explore the performance of 2DKS distance in clustering stationary ARMA time series. The following generating models were used:

(i) AR(1): $X_t = 0.5X_{t-1} + \varepsilon_t$
(ii) MA(1): $X_t = \varepsilon_t + 0.7\varepsilon_{t-1}$
(iii) AR(2): $X_t = 0.6X_{t-1} + 0.2X_{t-2} + \varepsilon_t$

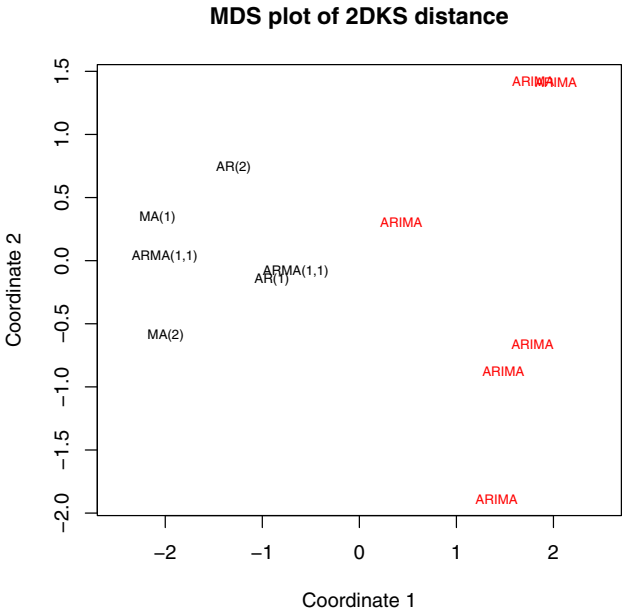


Figure 6. Multidimensional scaling plot based on 2DKS distance from 6 stationary and 6 non-stationary series , when $T = 500$ and $K = 10$

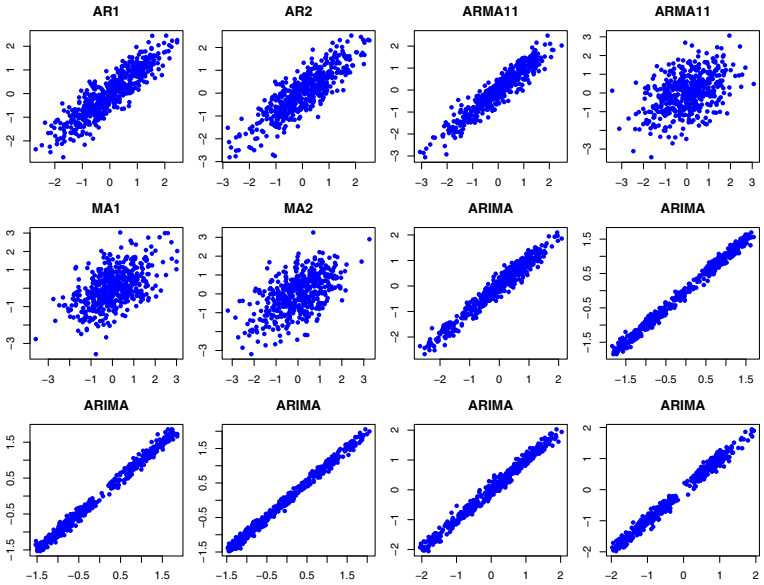


Figure 7. Lagged scatterplot (X_t, X_{t+1}) from 6 stationary and 6 non-stationary time series when T=500

Table 3. Percentage of success in the classification of 6 stationary and 6 non-stationary series with $N = 300$ iterations. Generalized Ward's linkage algorithm.

Measure	Numbers of lags	percentage of success $T = 50$	$T = 200$	$T = 500$	Measure	percentage of success $T = 50$	$T = 200$	$T = 500$
	$L = 10$	0.767	0.805	0.812	(Model-based)			
d_{ACFG}	$L = 25$	0.762	0.867	0.943	d_{PIC}	0.702	0.747	0.750
	$L = 50$	—	0.874	0.952	d_M	0.746	0.750	0.750
	$L = 10$	0.748	0.750	0.750	(Periodograms)			
d_{PACF}	$L = 25$	0.741	0.750	0.750	d_P	0.646	0.731	0.748
	$L = 50$	—	0.750	0.750	d_{LP}	0.688	0.752	0.750
	$L = 10$	0.743	0.750	0.750	d_{NP}	0.627	0.650	0.716
d_{PACFG}	$L = 25$	0.750	0.750	0.750	d_{LNP}	0.758	0.884	0.973
	$L = 50$	—	0.750	0.750	(Non-parametric)			
	$K = 1$	0.708	0.796	0.820	$d_{W(DLS)}$	0.732	0.750	0.750
	$K = 2$	**0.713**	0.778	0.780	$d_{W(LK)}$	0.746	0.750	0.750
d_{2DKS}	$K = 5$	0.705	0.802	0.848	d_{GLK}	0.740	0.750	0.750
	$K = 10$	0.683	**0.850**	0.917	d_{ISD}	0.740	0.750	0.750
	$K = 15$	0.669	0.846	**0.935**	(Unit-root tests)			
	$K = 50$	—	0.752	0.903	ADF	0.642	0.834	0.949
					PP	0.726	0.894	0.966

L is the number of lags used to compute ACF and PACF distance measure.
K is the number of lags used for 2DKS distance.
T is the length of the series.

(iv) MA(2): $X_t = \varepsilon_t + 0.8\varepsilon_{t-1} - 0.6\varepsilon_{t-2}$
(v) ARMA(1,1): $X_t = 0.8X_{t-1} + \varepsilon_t + 0.2\varepsilon_{t-1}$

This set of models was used by Maharaj (1996) to investigate the performance of d_M measure in clustering ARMA processes. Four series of length $T = 200$ were generated from each process. Then the agglomerative hierarchical clustering algorithm with the Generalized Ward's linkage was run on the collection of the 20 series with various distance measures. One hundred trials ($N = 100$) of this scheme were carried out, and Table 4 provides the results under the average cluster similarity indexes and average loo1NN.

Even though 2DKS distance is a kind of general dependence structure distance, 2DKS distance based clustering method performed worst except some periodograms based methods. It is especially difficulty in distinguishing AR(2) and ARMA(1,1) models under specific parameter values. Due

Table 4. Clustering of ARMA process. Series length $T = 200$. Number of trials $N = 100$. Generalized Ward's linkage.

Measure	Numbers of lags	Similarity Index	loo1NN	Measure	Similarity Index	loo1NN
	$L = 10$	0.750	0.724	(Model-based)		
d_{ACFG}	$L = 25$	0.736	0.674	d_{PIC}	0.855	0.916
	$L = 50$	0.721	0.655	d_M	0.965	0.961
	$L = 10$	0.966	0.940	(Periodograms)		
d_{PACF}	$L = 25$	0.911	0.856	d_P	0.524	0.444
	$L = 50$	0.884	0.784	d_{LP}	0.746	0.616
	$L = 10$	0.975	0.951	d_{NP}	0.586	0.437
d_{PACFG}	$L = 25$	0.960	0.932	d_{LNP}	0.752	0.637
	$L = 50$	0.960	0.930	(Non-parametric)		
	$K = 1$	0.692	0.574	$d_{W(DLS)}$	0.919	0.940
	$K = 2$	0.719	0.610	$d_{W(LK)}$	0.965	0.942
d_{2DKS}	$K = 3$	0.686	0.643	d_{GLK}	0.916	0.884
	$K = 4$	0.664	0.634	d_{ISD}	0.952	0.944
	$K = 5$	0.644	0.593			

L is the number of lags used to compute ACF and PACF distance measure.
K is the number of lags used for 2DKS distance.

to the fact that the two processes have very similar dependence structure (Figure 8), their joint Gaussian distributions of (X_t, X_{t+h}) can not be discriminated by 2DKS distance. When we change the parameters in AR(2) model to $X_t = -0.6X_{t-1} + 0.2X_{t-2} + \varepsilon_t$, the performance of 2DKS distance was improved significantly. The average similarity index can reach 0.940 for most maximum lag K. In this case the lag 1 autocorrelation coefficients are of the opposite sign for the AR(2) and ARMA(1,1) models used here. Therefore the two types of models differ more significantly in the joint distribution of (X_t, X_{t-1}).

5. Applications

We further illustrate the use of 2DKS distance for time series clustering with two real examples. The first application considers the annual population series of 20 US states and is aimed at identifying similarities among the population growth trend. The second example concerns the comparison of GDP growth in order to identify similar pattern of economic development of developed countries.

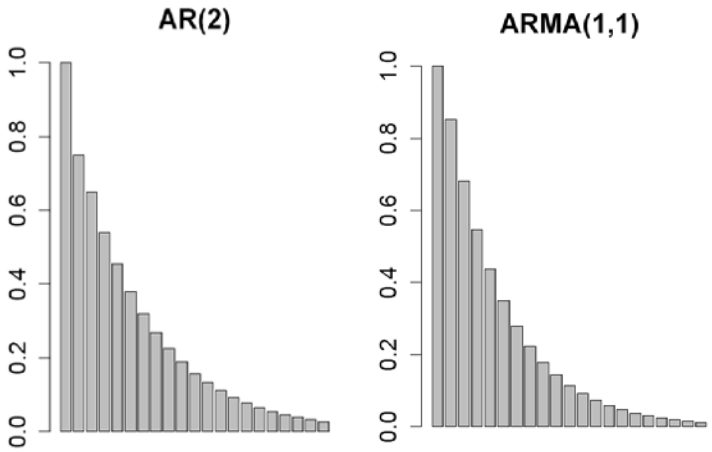


Figure 8. ACF for AR(2) and ARMA(1,1) Model

5.1 Case 1: Population Growth

This example is used in Kalpakis et al. (2001). The data is updated to 2010 here. It is a collection of time series of the population estimates from 1991 to 2010 in 20 states of the US. This data is obtained from the U.S Census Bureau, Population Distribution Division. In Kalpakis et al. (2001), 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. The performance of the clustering methods is evaluated with both the similarity index and loo1NN. We plotted two representative series from each group in Figure 9. The trend difference between the two groups are obvious.

In the following analysis, we use the growth series by taking log difference of the original series. Table 5 summarizes the results of clustering the time series with the agglomerative hierarchical clustering algorithm using the Generalized Ward's linkage. We observe that the 2DKS distance gives the most accurate clustering among all these dependence-based clustering method. The CEP method from Kalpakis et al. (2001) and ARMA mixture method from Xiong and Yeung (2004), have the similarity index 0.84 and 0.9 respectively. Apparently 2DKS method performs better than both of them. It is worth mentioning that 2DKS distance seems to have a significant advantage when loo1NN index is considered. This is because

Table 5. Clustering of population dataset. Number of cluster: 2. Generalized Ward's linkage clustering algorithm.

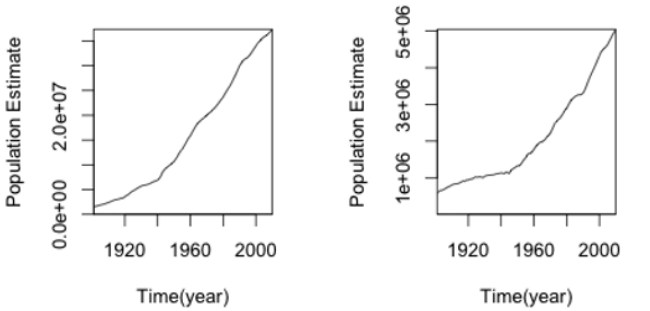
Measure	Numbers of lags	Similarity Index	loo1NN	Measure	Similarity Index	loo1NN
	$L = 10$	0.542	0.75	(Model-based)		
d_{ACFG}	$L = 25$	0.581	0.6	d_{PIC}	0.649	0.6
	$L = 50$	0.583	0.7	d_M	0.642	0.7
	$L = 10$	0.627	0.55	(Periodograms)		
d_{PACF}	$L = 25$	0.584	0.6	d_P	0.596	0.9
	$L = 50$	0.688	0.75	d_{LP}	0.596	0.7
	$L = 10$	0.749	0.6	d_{NP}	0.613	0.55
d_{PACFG}	$L = 25$	0.7	0.6	d_{LNP}	0.688	0.7
	$L = 50$	0.749	0.8	(Non-parametric)		
	$K = 1$	1	1	$d_{W(DLS)}$	0.596	0.85
	$K = 2$	0.688	1	$d_{W(LK)}$	0.581	0.55
d_{2DKS}	$K = 3$	1	1	d_{GLK}	0.581	0.55
	$K = 4$	1	1	d_{ISD}	0.581	0.55
	$K = 5$	0.688	1			

L is the number of lags used to compute ACF and PACF distance measure.
K is the number of lags used for 2DKS distance.

this criterion directly evaluates the efficiency of the dissimilarity distance, and pays less attention to the clustering algorithm. Comparing with similarity index, clustering of series is more robust to the choice of K under the criterion of loo1NN index. On the other hand, to learn the impact of the linkage algorithm, we also cluster the series with other linkage algorithms including single, complete and average linkage methods. The three methods show similar results, with similarity index all less than 0.7. It seems that the Generalized Ward's linkage algorithm works well in this case.

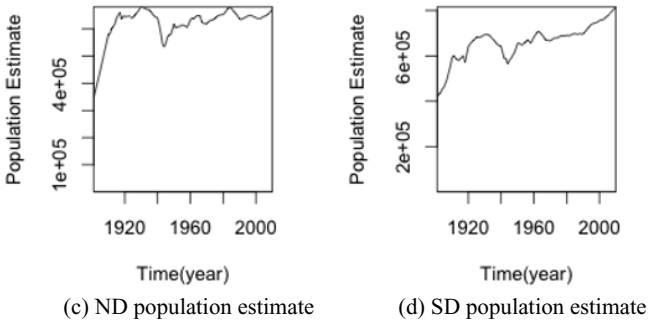
The dendrogram clustered by 2DKS distance with the Generalized Ward's linkage method are shown in Figure 10. By sectioning the dendrogram at the highest level we can obtain two groups, which are exactly consistent with the ones Kalpakis et al. (2001) identified. The main discriminating feature is the nonstationary trend of the time series for this dataset. Even though we remove the nonstationarity before clustering, the difference between exponentially increasing trend and stabilizing trend can still be captured by 2DKS distance.

In practice, one does not know how many groups generate these time series. In general, the optimal number of clusters could be chosen according to some objective criterion, such as the average Silhouette coefficient



(a) CA population estimate (from group 1)

(b) CO population estimate (from group 1)



(c) ND population estimate (from group 2)

(d) SD population estimate (from group 2)

Figure 9. Original time series from group 1 and group 2

(Kaufman and Rousseeuw, 2009, p. 120). In this application, the optimal number of clusters is 3, by maximizing the average Silhouette coefficient. The three clusters consist of the stabilizing trend group and two exponential increasing trend group. From Figure 10, the 3-group solution divides the group of exponentially increasing trend into two subgroups: C1={WA, CO, TX, FL, CA}, and C2={VA, MD, TN, SC, NC, GA}. The state members in each subgroup are quite homogeneous. Specifically, the majorities of subgroup C1 are located in the western area except FL, while the states in the subgroup C2 are all located in the eastern area and geographically adjacent to each other. It is interesting that this grouping has a clear geographical component.

5.2 Case 2: Annual Real GDP dataset

Here we consider the total real Gross Domestic Product (GDP) data obtained from http://www.conference-board.org/data/economydatabase/index.cfm?id=2346. It contains the annual real GDP of the 23 most devel-

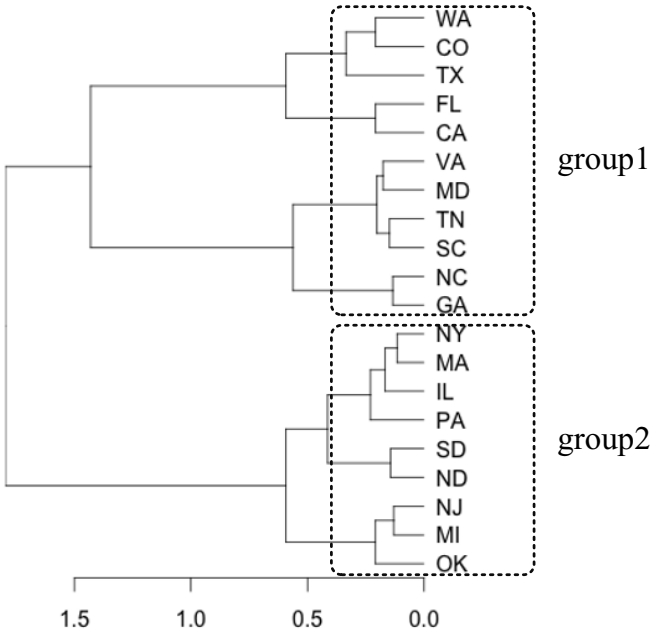


Figure 10. Dendrogram based on 2DKS distance with $K = 1$. The Generalized Ward's linkage algorithm.

oped countries from 1950 to 2011: Austria, Belgium, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Italy, Luxembourg, Netherland, Norway, Portugal, Spain, Sweden, Switzerland, United Kingdom, Canada, United States, Australia, New Zealand, and Japan. We consider the data normalized by EKS method (Atkinson and Bourguignon, 2000, p. 949).

We use the annual GDP growth rate $log(GDP_t) - log(GDP_{t-1})$ in the clustering procedures rather than the actual annual GDP. These series are clustered via the Generalized Ward's linkage based on 2DKS distance. The result is relatively insensitive to the maximum lag K used, with similar clustering result from K ranging from 2 to 9. The dendrogram is shown in Figure 11. The average Silhouette coefficients were examined for different number of clusters, and it seems that three clusters yields a compact solution (Figure 12). Figure 13 shows the grouping of three clusters.

It is interesting to note that the countries are mostly grouped by their geographical locations. The first cluster is formed by 9 countries, including Switzerland, Denmark, United Kingdom, Sweden, Belgium, France, Germany, Italy, and New Zealand. In this group, most of them are core EU members, including the original EU member states joined in 1979, except Switzerland, Sweden, and New Zealand. The second one includes Portu-

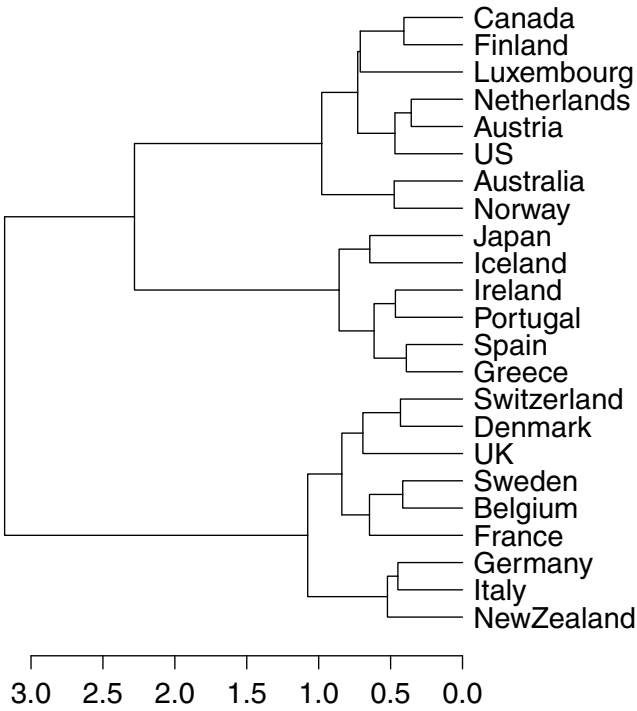


Figure 11. GDP clustering dendrogram based on 2DKS distance with $K = 3$. Generalized Ward's linkage algorithm.

gal, Spain, Greece, Ireland, Island, and Japan. This group does not change with different K. They are all located in the outmost reaches of Europe except Japan. Portugal, Spain, Greece are at the most southern part of Europe, while Ireland and Island are at the westernmost peninsula of Eurasia. Portugal and Spain have similar history and politic background, therefore their economics behavior are similar, both less developed than the other countries in Europe. The third cluster, includes most of the non-Europe developed large countries, such as United States, Canada and Australia. The economies of Canada and United States would behavior similarly to each other because of their geographical locations and trade agreement.

References

AN, L. (2008), "Dynamic Clustering of Time Series Gene Expression", Thesis, Purdue University, ProQuest Dissertations Publishing.

ATKINSON, A.B., and BOURGUIGNON, F. (2000), *Handbook of Income Distribution*, Elvesier.

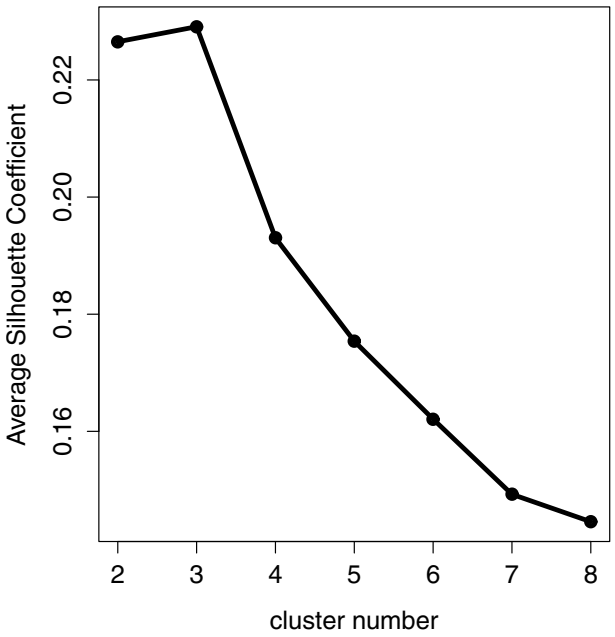


Figure 12. Plot of average Silhouette coefficient ($K = 3$). Generalized Ward's linkage algorithm.

Figure 13. Three groups: 2DKS distance ($K = 3$).

BATAGELJ, V. (1988), "Generalized Ward and Related Clustering Problems", in *Classification and Related Methods of Data Analysis*, ed. H.H. Bock, pp 67–74.

BOHTE, Z., CEPAR, D., and KOSMELJ, K. (1980), "Clustering of Time Series", in *Compstat* (Vol. 80), pp 587–593.

BORG, I., and GROENEN, P.J. (2005), *Modern Multidimensional Scaling: Theory and Applications,* Springer Science and Business Media.

CAIADO, J., CRATO, N., and PEÑA, D. (2006), "A Periodogram-Based Metric for Time Series Classification", *Computational Statistics and Data Analysis, 50(10)*, 2668–2684.

CONOVER, W. (1999), *Practical Nonparametric Statistics*, New York: John Wiley and Sons.

CORDUAS, M., and PICCOLO, D. (2008), "Time Series Clustering and Classification by the Autoregressive Metric", *Computational Statistics and Data Analysis, 52(4)*, 1860–1872.

DEFAYS, D. (1977), "An Efficient Algorithm for a Complete Link Method", *Computer Journal, 20(4)*, 364–366.

DÍAZ, S.P., and VILAR, J.A. (2010), "Comparing Several Parametric and Nonparametric Approaches to Time Series Clustering: A Simulation Study", *Journal of Classification, 27(3)*, 333–362.

DIKS, C. (2009), "Nonparametric Tests for Independence", in *Encyclopedia of Complexity and Systems Science*, Springer, pp 6252–6271.

DUFOUR, J.M., LEPAGE, Y., and ZEIDAN, H. (1982), "Nonparametric Testing for Time Series: A Bibliography", *Canadian Journal of Statistics, 10(1)*, 1–38.

D'URSO, P., and MAHARAJ, E.A. (2009), "Autocorrelation-Based Fuzzy Clustering of Time Series", *Fuzzy Sets and Systems, 160(24)*, 3565–3589.

FAN, J. (2003), *Nonlinear Time Series: Nonparametric and Parametric Methods*, Springer.

FASANO, G., and FRANCESCHINI, A. (1987), "A Multidimensional Version of the Kolmogorov-Smirnov Test", *Monthly Notices of the Royal Astronomical Society, 225*, 155–170.

FRÜHWIRTH-SCHNATTER, S., and KAUFMANN, S. (2008), "Model-Based Clustering of Multiple Time Series", *Journal of Business and Economic Statistics, 26(1)*, 78–89.

GALEANO, P., and PEÑA, D.P. (2000), "Multivariate Analysis in Vector Time Series", *Resenhas, 4*, 383–404.

GAVRILOV, M., ANGUELOV, D., INDYK, P., and MOTWANI, R. (2000), "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, pp 487–496.

GOWER, J.C., and ROSS, G.J.S. (1969), "Minimum Spanning Trees and Single Linkage Cluster Analysis", *Journal of the Royal Statistical Society, 18(1)*, 54–64.

GRANGER, C., MAASOUMI, E., and RACINE, J. (2004), "A Dependence Metric for Possibly Nonlinear Processes", *Journal of Time Series Analysis, 25(5)*, 649–669.

HARVILL, J.L., RAVISHANKER, N., and RAY, B.K. (2013), "Bispectral-Based Methods for Clustering Time Series", *Computational Statistics and Data Analysis, 64(C)*, 113–131.

HASTIE, T., TIBSHIRANI, R., and FRIEDMAN, J. (2009), *The Elements of Statistical Learning* (2nd ed.), New York: Springer.

KALPAKIS, K., GADA, D., and PUTTAGUNTA, V. (2001), "Distance Measures for Effective Clustering of ARIMA Time-Series", in *Proceedings IEEE International Conference on Data Mining, 2001. ICDM 2001*, pp. 273–280.

KAUFMAN, L., and ROUSSEEUW, P.J. (2009), *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley and Sons.

KOSMELJ, K., and BATAGELJ, V. (1990), "Cross-Sectional Approach for Clustering Time Varying Data", *Journal of Classification, 7(1)*, 99–109.

LAFUENTE-REGO, B., and VILAR, J. (2016), "Clustering of Time Series Using Quantile Autocovariances", *Advances in Data Analysis and Classification, 10(3)*, 391–415.

LANCE, G.N., and WILLIAMS, W.T. (1967), "A General Theory of Classificatory Sorting Strategies. Hierarchical Systems", *The Computer Journal, 9(4)*, 373–380.

LIAO, T.W. (2005), "Clustering of Time Series Data: A Survey", *Pattern Recognition, 38(11)*, 1857–1874.

LIU, S., and MAHARAJ, E.A. (2013), "A Hypothesis Test Using Bias-Adjusted ar Estimators for Classifying Time Series in Small Samples", *Computational Statistics and Data Analysis, 60*, 32–49.

LOPES, R.H., REID, I., and HOBSON, P.R. (2007), "The Two-Dimensional Kolmogorov-Smirnov Test", in *XI International Workshop on Advanced Computing and Analysis Techniques in Physics Research*, Nikhef, Amsterdam, The Netherlands.

LOPES, R.H., HOBSON, P.R., and REID, I.D. (2008), "Computationally Efficient Algorithms for the Two-Dimensional Kolmogorov-Smirnov Test", in: *Journal of Physics: Conference Series* (Vol. 119), IOP Publishing, pp. 2438–2571.

MA, P., and ZHONG, W. (2008), "Penalized Clustering of Large-Scale Functional Data with Multiple Covariates", *Journal of the American Statistical Association, 103(482)*, 625–636.

MAHARAJ, E.A. (1996), "A Significance Test for Classifying ARMA Models", *Journal of Statistical Computation and Simulation, 54(4)*, 305–331.

MAHARAJ, E.A. (2000), "Cluster of Time Series", *Journal of Classification, 17(2)*, 297–314.

MANSO, P.M., and VILAR, J. (2013), "TSclust: Time Series Clustering Utilities", http://CRAN.R-project.org/package=TSclust, R package version 1.1.

MURTAGH, F. (1984), "Complexities of Hierarchic Clustering Algorithms: State of the Art", *Computational Statistics Quarterly, 1(2)*, 1041–1080.

PEACOCK, J. (1983), "Two-Dimensional Goodness-of-Fit Testing in Astronomy", *Monthly Notices of the Royal Astronomical Society, 202*, 615–627.

PERRON, P. (1987), "Testing for a Unit Root in Time Series Regression", *Biometrika, 75(2)*, 335–346.

PICCOLO, D. (1990), "A Distance Measure for Classifying ARIMA Models", *Journal of Time Series Analysis, 11(2)*, 153–164.

TONG, H. (1990), *Non-Linear Time Series: A Dynamical System Approach*, Oxford University Press.

TONG, H., and YEUNG, I. (1991)," On Tests for Self-Exciting Threshold Autoregressive-Type Nonlinearity in Partially Observed Time-Series", *Applied Statistics-Journal of the Royal Statistical Society Series C, 40(1)*, 43–62.

VILAR, J. (2014), "Tsclust: An R Package for Time Series Clustering", *Journal of Statistical Software, 62(1)*, 1–43.

VILAR, J.A., ALONSO, A.M., and VILAR, J M. (2010), "Non-Linear Time Series Clustering Based on Non-Parametric Forecast Densities", *Computational Statistics and Data Analysis, 54(11)*, 2850–2865.

XIAO, Y. (2017), "A Fast Algorithm for Two-Dimensional Kolmogorov-Smirnov Two Sample Tests", *Computational Statistics and Data Analysis, 105(C)*, 53–58.

XIONG, Y., and YEUNG, D.Y. (2004), "Time Series Clustering with ARMA Mixtures", *Pattern Recognition, 37(8)*, 1675–1689.

ZHANG, T. (2013), "Clustering High-Dimensional Time Series Based on Parallelism", *Journal of the American Statistical Association, 108(502)*, 577–588.