JOURNAL OF TIME SERIES ANALYSIS
J. Time Ser. Anal. **44**: 667–685 (2023)
Published online 17 June 2023 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1111/jtsa.12706

SPECIAL ISSUE ARTICLE

A TESTING APPROACH TO CLUSTERING SCALAR TIME SERIES

DANIEL PEÑA[a,b] AND RUEY S. TSAY[a,b]

[a]*Department of Statistics, Universidad Carlos III de Madrid, Madrid, Spain*
[b]*Booth School of Business, University of Chicago, Chicago, IL, USA*

This article considers clustering stationary scalar time series using their marginal properties and a hierarchical method. Two major issues involved are to detect the existence of clusters and to determine their number. We propose a new test statistic for detecting whether a data set consists of multiple clusters and a new procedure to determine the number of clusters. The proposed method is based on the jumps, that is, the increments, in the heights of the dendrogram when a hierarchical clustering is applied to the data. We use autoregressive sieve bootstrap to obtain a reference distribution of the test statistics and propose an iterative procedure to find the number of clusters. The clusters found are internally homogeneous according to the test statistics used in the analysis. The performance of the proposed procedure in finite samples is investigated by Monte Carlo simulations and illustrated by some empirical examples. Comparisons with some existing methods for selecting the number of clusters are also investigated.

Received 28 August 2022; Accepted 28 May 2023

Keywords: Autogressive sieve bootstrap; dendrogram; distance; gap statistic; hierarchical clustering; jump; Silhouette statistic; similarity.

MOS subject classification: 62F40; 62H30; 62M10.

1. INTRODUCTION

Selecting clusters (or groups) of time series with similar characteristics is a common problem with many applications in diverse scientific areas, ranging from economics to medicine. Indeed, similarity between series plays a key role in all classification problems with dynamic dependence data. See Pértega and Vilar (2010), Aghabozorgi *et al.* (2015), and Maharaj *et al.* (2019) for surveys on the topic.

We assume that the available information for clustering is included in a data matrix $\mathbf{X}_{T\times k}$ with k stationary scalar time series observed in T time periods, even though some methods discussed remain applicable when series have different sample sizes. Specifically, we assume that each time series x_{it} of $\mathbf{X}_{T\times k}$, for $i = 1, \ldots, k$, is strictly stationary and has the following linear AR(∞) representation

$$x_{it} - \mu_i = \sum_{j=1}^{\infty} \pi_{i,j}(x_{i,t-j} - \mu_i) + a_{it}, \tag{1}$$

It is equivalent to MA(1). i.e. It is stationary

ait is White Noise

where $\mu_i = E(x_{it})$ and $\{a_{it}\}$ is a sequence of independent and identically distributed random variables with mean zero, variance $E(a_{it}^2) = \sigma_i^2 > 0$, and $E(a_{it}^4) < \infty$, $\sum_{j=1}^{\infty}|\pi_{ij}| < \infty$, and $\pi_i(z) = 1 - \sum_{j=1}^{\infty}\pi_{ij}z^j \neq 0$ for $|z| \leq 1$. In this article, we focus mainly on the dynamic structure of the time series x_{it} so that we can assume $\mu_i = 0$ for all i. Also, for simplicity, we assume that $\{a_{it}\}$ have the same distribution for all i. If the level of x_{it} and the distribution of a_{it} are also of interest, then one can include μ_i and some characteristics of a_{it} in the clustering analysis.

Consider the sequence $\{\pi_{ij}\}_{j=1}^{\infty} \equiv \mathbf{\Pi}_i$, for $i = 1, \ldots, k$. If $\pi_{ij} = \pi_j$ for all i and j, then $\mathbf{\Pi}_i = \mathbf{\Pi}_v$ for $1 \leq i, v \leq k$. In this case, all time series share the same AR coefficients and there is only a single cluster in the data. On the

* Correspondence to: Daniel Peña, Departamento de Estadistica, Universidad Carlos III de Madrid, calle Madrid 126, Getafe 28903, Madrid, Spain. Email: daniel.pena@uc3m.es

other hand, if $\{\mathbf{\Pi}_i\}_{i=1}^k$ consists of G distinct sequences, then there are G clusters in the data set $\mathbf{X}_{T\times k}$. The goal of this article is to test for the homogeneity of the data, that is, to check whether the series in $\boldsymbol{X}_{T\times k}$ belong to a single cluster. Furthermore, if the null hypothesis of a single cluster is rejected, we apply an iterative testing procedure to determine the number of clusters.

Roughly speaking, there are three main approaches for finding clusters in the data matrix $\mathbf{X}_{T\times k}$. The first one is to split the k time series into clusters based on their marginal features. This approach has been extensively studied in the literature. See, for instance, Maharaj *et al.* (2019). The second one is to group series by their dynamic dependency using their cross correlations; see, for instance, Alonso and Peña (2019). A third approach is to split the time period into segments such that the series exhibit homogeneous joint dynamic behavior within a segment, but show some differences between different segments. This third approach has not been fully explored in the statistical literature.

In this article we focus on the first approach, clustering stationary scalar time series using their marginal properties via a hierarchical method. Usually the methods consider a distance or dissimilarity measure applied to some selected dynamic features of individual series. The choice of features often depends on the objective of clustering. For a given set of selected features, we first propose a test statistic for detecting the existence of multiple clusters in $\boldsymbol{X}_{T\times k}$. An autoregressive sieve (AR-sieve) bootstrap procedure is used to generate a reference distribution for the test statistic. See, for instance, Bühlmann (2002), Kreiss *et al.* (2011), and the references therein for AR-sieve bootstrap. If the null hypothesis of a single cluster is rejected, we propose a new approach to select the number of clusters. The proposed method makes use of increments in the heights (or jumps) of the dendrogram obtained by a hierarchical clustering. Testing for the existence of clusters in $\mathbf{X}_{T\times k}$ is less studied in the literature, because their existence is often assumed in applications. From a statistical point of view, it would be useful to test for the homogeneity in the data instead of assuming the existence of multiple clusters *a priori*.

The contributions of this article are as follows: (i) a new test statistic for testing the homogeneity of a time series data set is proposed; (ii) it demonstrates clearly the usefulness of jumps in the dendrogram in selecting the number of clusters; (iii) an iterative procedure is proposed to find the number of clusters with the property that the clusters found are internally homogeneous; and (iv) the proposed procedure is shown to work well in practice.

The article is organized as follows. We briefly review some existing clustering procedures for time series in Section 2 and some commonly used criteria for selecting the number of clusters in Section 3. Section 4 presents our proposed method and Section 5 shows some simulation results, including comparisons with some commonly used methods for selecting the number of clusters in the literature. Section 6 presents two empirical applications and Section 7 concludes.

2. CLUSTERING PROCEDURES FOR TIME SERIES

Clustering time series has a long history in the literature based on the basic idea that the observed series have been generated by a mixture of different distributions. In a non-parametric approach these distributions are assumed to be different, but not specified, and some distance between vectors of selected features of the series is used to perform clustering, usually using a hierarchical cluster method. These features can also be used in a partitioning algorithm, such as the k-means. In the parametric, or model-based, approach the clusters are found by estimating a mixture of different distributions or time series models. The estimation is carried out either by the EM algorithm or by Markov chain Monte Carlo (MCMC), or other Bayesian estimation methods. See, for instance, Bouveyron *et al.* (2019) for clustering independent data. Clustering has also been considered in the machine learning literature using neural networks.

2.1. Dissimilarities for Hierarchical Clustering

The most commonly used non-parametric approach to clustering is to summarize each time series in a vector of selected features or characteristics and use distances or dissimilarities between these vectors and a hierarchical

Journal of Time Series Analysis published by John Wiley & Sons Ltd. DOI: 10.1111/jtsa.12706

clustering algorithm to perform clustering. We focus in this section on the measures of distance used in hierarchical clustering.

For the stationary linear processes of (1), several ways to summarize the series have been proposed. The first one is to use the autocorrelation coefficients (ACF) of the series. Let $\widehat{\boldsymbol{\rho}}_x = (\widehat{\rho}_x(1), \ldots, \widehat{\rho}_x(h))'$ be the vector of the first h lags of sample ACF of x_t. The Euclidean distance between the selected sample ACFs of series x_t and y_t is given by

Euclidean distance doesn't take into account the correlation structure of the stationary data because Euclidean distance is invariant to transformations.

$$d(x_t, y_t, h) = \sqrt{\sum_{j=1}^{h} [\widehat{\rho}_x(j) - \widehat{\rho}_y(j)]^2}, \tag{2}$$

and is referred to as the *ACF distance* (Galeano and Peña, 2000). In (2), all ACFs have the same weight. This can be modified by introducing a weight matrix with weights being a decreasing function of the lag. For instance, the weights may decrease linearly. For sample ACFs, the weights may depend on their asymptotic variances so that high-order coefficients have lower weights, because they tend to have higher variability. A third and more complex alternative is to take into account that sample ACFs are correlated and employ a Mahalanobis distance between the two vectors of sample ACFs. Thus, a general distance between the ACFs is

$$d_W(x_t, y_t, h) = \sqrt{(\widehat{\boldsymbol{\rho}}_x - \widehat{\boldsymbol{\rho}}_y)' \mathbf{W} (\widehat{\boldsymbol{\rho}}_x - \widehat{\boldsymbol{\rho}}_y)},$$

where the matrix $\boldsymbol{W}$ is the identity matrix if all the coefficients have the same weight, as in (2), and is a diagonal matrix if the weights decrease with the lag, and is a full-rank matrix if $\boldsymbol{W}$ is the inverse covariance matrix of the sample ACFs used.

An alternative way to distinguish time series is by their predictability, defined as the ratio of the explained (or predicted) variance and the total variance. For instance, two linear processes may have different parameters and autocorrelations, but the same predictability. Equation (1) has a moving-average representation,

$$x_{it} - \mu_i = \sum_{j=0}^{\infty} \psi_{i,j} a_{i,t-j}, \tag{3}$$

where $\psi_{i,0} = 1$ and $\psi_{i,j}$ can be obtained by $\pi_i(z)\psi_i(z) = 1$, where $\psi_i(z) = 1 + \sum_{j=1}^{\infty} \psi_{i,j} z^j$ and $\pi_i(z) = 1 - \sum_{j=1}^{\infty} \pi_{i,j} z^j$, and $x_{it} = \widehat{x}_{it|t-1} + a_t$. Then, the predictability of x_{it} can be defined as $P_i = \sigma^2_{\widehat{x}_{t|t-1}} / \sigma^2_x = 1 - (\sum_{i=0}^{\infty} \psi_i^2)^{-1}$, where σ_y^2 denotes the variance of a stationary series y_t. It is easily seen that the vector of ACFs $\widehat{\boldsymbol{\rho}}_x$ defined earlier is not a good measure of predictability. For instance, the two MA processes, $x_t = a_t - 0.5a_{t-1}$ and $y_t = a_t - 0.5a_{t-2}$, have different autocorrelations and parameters, but the same predictability. If we are interested in clustering by predictability the sample ACFs can be included in the determinant of autocorrelation matrix of order h, given by

$$\widehat{\mathbf{R}}_h = \begin{vmatrix} 1 & \widehat{\rho}(1) & \ldots & \widehat{\rho}(h) \\ \widehat{\rho}(1) & 1 & \ldots & \widehat{\rho}(h-1) \\ \ldots & \ldots & \ddots & \ldots \\ \widehat{\rho}(h) & \widehat{\rho}(h-1) & \ldots & 1 \end{vmatrix}.$$

This determinant verifies $0 \leq |\mathbf{R}_h| \leq 1$. Peña and Rodríguez (2002) proposed to standardize this determinant and define the total correlation statistic $\widehat{D}_h$,

Portmanteau Test Statistic ->

$$\widehat{D}_h = T\left[1 - |\widehat{\mathbf{R}}_h|^{1/h}\right], \tag{4}$$

DOI: 10.1111/jtsa.12706 *Journal of Time Series Analysis* published by John Wiley & Sons Ltd.

to quantify the dynamic dependence or memory of a scalar time series. It can be shown that this statistic can be used as a global correlation coefficient or predictability measure. It is also a weighted function of the first h lags of partial autocorrelation coefficients (PACFs). The total correlation is robust to the choice of h when most of the time series have no serial dependence after certain lag. Peña and Tsay (2021) propose to use the following statistic as a similarity measure of predictability between series x_t and y_t:

$$S(x_t, y_t) = \left| |\hat{\mathbf{R}}_{x,h}|^{1/h} - |\hat{\mathbf{R}}_{y,h}|^{1/h} \right|,$$

where the subscripts x and y are added to signify the series used.

The linear dependency of a time series can also be summarized according to Wold' theorem (see, for instance, Box *et al.*, 2015) by the coefficients of an AR fit to the series. Piccolo (1990), see also Corduas and Piccolo (2008), employs the vector of the first h lags of estimated AR coefficients $\hat{\boldsymbol{\phi}}_x = [\hat{\phi}_x(1), \ldots, \hat{\phi}_x(h)]$ and computes the Euclidean distance between the coefficient vectors by

$$d_{\phi}(x_t, y_t, h) = \sqrt{\sum_{j=1}^{h} [\hat{\phi}_x(j) - \hat{\phi}_y(j)]^2}. \tag{5}$$

In practice, the lag or order h used must be sufficiently large to describe properly the linear dynamic dependence of a series, especially when the series involved has certain seasonality. This approach can be extended to include the intercept and residual variance in $\hat{\boldsymbol{\phi}}_x$, especially when the level of x_t or its variability is of interest.

A different way to summarize the dynamic dependence of a time series is to use its periodogram. A distance between the logarithms of normalized periodograms was proposed by Caiado *et al.* (2006) as

$$d_{NP}(x_t, y) = \sqrt{\sum_{j=1}^{T/2} [\log NI_x(2\pi j/T) - \log NI_y(2\pi j/T)]^2}. \tag{6}$$

where $NI(2\pi j/T)$ is the normalized periodogram at frequency j/T.

Other features of scalar time series have also been proposed for clustering analysis. Wang *et al.* (2006) proposed to use a vector of structural characteristics consisting of trend, seasonality, periodicity, serial correlations, marginal skewness and kurtosis, and some nonlinear features. Some authors have proposed distance measures based on nonlinear characteristics (see Montero and Vilar, 2014), or with the objective to compare the shape of two time series with possible different sample sizes. For instance, in speech recognition we want to identify words that have been recorded with different speaking speeds. In this case, we search for an optimal alignments into the two series and this is the objective of the *dynamic time warping* measure which minimizes some distance between values of the series taken in sequence, but these values may be for different time points, so that a given shape may not happen at the same time (see Petitjean *et al.*, 2011).

For integrated non-stationary time series one can take the proper difference and compares properties of the resulting stationary series. A mixture of stationary and non-stationary features can also be used (Wang *et al.*, 2006). One can also use the similarity between forecasting densities of the series for a given horizon to perform clustering, as proposed by Alonso *et al.* (2006). Other forecasting based measures include the distances between vectors of out-of-sample forecasts and the total correlation introduced before. Finally, as the data may contain outliers, it is important to clean the series or use some robust methods to compute the selected features. For instance, instead of sample ACFs, Spearman's rank correlation coefficients can be used; see, for instance, Tsay (2020).

Instead of the marginal dependency some authors have proposed to use the joint dependency to cluster time series. Zhang and An (2018) used a copula-based distance to measure dissimilarity among time series and employed a non-parametric estimator. Alonso and Peña (2019) introduced the generalized cross correlation

DOI: 10.1111/jtsa.12706

coefficient to summarize the linear joint dependency between two time series and used a hierarchical clustering method to find groups.

2.2. Other Approaches for Clustering Scalar Time Series

Xiong and Yeung (2004) consider a set of independent time series and proposed a mixture of ARMA models and a model-based clustering procedure, where the models are estimated by the EM algorithm. A similar setup is adopted by Fruhwirth-Schnatter and Kaufmann (2008) from a Bayesian point of view with MCMC estimation methods. They assume G clusters each with a unique AR(p) model for its member series and introduce classification variables that follow a multinomial distribution. Given the parameters, the classification variables, and conjugate priors for the parameter vector, the authors assume that the individual series is normally distributed. Wang and Tsay (2019) generalize this approach by allowing for structural breaks. Once breaks and group memberships are found, it is possible to pool information across all series in the same group for estimation and prediction.

Clustering methods have been also proposed for time series with non-normal distributions as, for instance, financial data which usually follow heavy-tailed distributions. Discrimination and clustering of non-Gaussian time series have been studied by Zhang and Taniguchi (1994), Sakiyama and Taniguchi (2004), Watanabe *et al.* (2010), Durante *et al.* (2014), Dias *et al.* (2015), and Liu *et al.* (2017), among many others. Model-based procedures for the joint dependency of high dimensional time series often employ a factor model framework under which the clusters of the series are generated by specific dynamic factors. See, for instance, Ando and Bai (2017), Alonso *et al.* (2020) and Oh and Patton (2023).

Some time series clustering procedures in the machine learning and artificial intelligence literature are based on deep neural networks that can be considered as semi-parametric approaches. See Sezer *et al.* (2020) for a survey of these methods in financial applications, Bandara *et al.* (2020) for the use of recurrent deep networks, and Alqahtani *et al.* (2021) for a review of deep time series clustering.

2.3. Clustering Multivariate Time Series

multivariate-multiple features

Clustering analysis has been extended to the case of multivariate time series in which a h-dimensional time series is treated as an item and the data set consists of k such items. Kakizawa *et al.* (1998) in a pioneering work analyzed discriminant and cluster approaches for such a data set. Specifically, they consider k data matrices $\{\boldsymbol{X}_i\}_{i=1}^{k}$, where $\boldsymbol{X}_i$ is a $T_i \times h$ data matrix with T_i being the sample size of the ith item. For clustering, they postulate that the vector time series has been generated by different probability distributions and their goal is to cluster these k h-dimensional time series into different groups. They assume that the probability distribution of each vector time series is multivariate normal with zero mean and propose different similarity measures between these probability distributions. For simplicity, we assume that $T_i = T$, for $i = 1, \ldots, k$. In this case, their symmetric J divergence is given by

$$J(\mathbf{f}_T; \mathbf{g}_T) = \frac{1}{2T} \sum_s [\mathrm{tr}(\mathbf{f}_T \mathbf{g}_T^{-1}) + \mathrm{tr}(\mathbf{g}_T \mathbf{f}_T^{-1}) - 2h]. \tag{7}$$

where $\mathbf{f}_T$ and $\mathbf{g}_T$ are $h \times h$ spectral density matrices of two vector time series with T observations and the summation is summing over frequencies $\lambda_s = 2\pi s/T$, for $s = 1, \ldots, T$. (Note that in Equations (7) and (18) of Kakizawa *et al.* (1998) there is a typo and p should be m, the dimension of their vector time series, that is h here.) An alternative similarity measure they considered is the Chernoff information given by

$$B_\alpha(\mathbf{f}_T; \mathbf{g}_T) = \frac{1}{2T} \sum_s \left[\log \left(\frac{|\alpha \mathbf{f}_T + (1-\alpha)\mathbf{g}_T|}{|\mathbf{g}_T|} \right| + \log \left(\frac{|\alpha \mathbf{g}_T + (1-\alpha)\mathbf{f}_T|}{|\mathbf{f}_T|} \right] . \tag{8}$$

DOI: 10.1111/jtsa.12706 *Journal of Time Series Analysis* published by John Wiley & Sons Ltd.

These authors apply the two dissimilarity measures to a set of 17 bivariate time series of seismic events and use hierarchical clustering to show two main groups, namely earthquakes and mining explosions. This work was the first in considering joint dependencies of a multivariate time series for clustering, see also Taniguchi and Kakizawa (2012). Finally, Huang *et al.* (2016) introduced a smooth weighting function at each time point and an algorithm similar to that of k-means to cluster vector time series and Li (2019) proposed clustering several vector time series with possibly different time lengths but the same dimension h by using common principal components, computed from the average covariance matrix of all the vectors and a k-means type algorithm.

Categories of Clustering:

1. Measure of Homogenity
2. Estimate no. of components in a mixture of distributions

3. SELECTING THE NUMBER OF CLUSTERS

Several procedures have been developed in the literature for selecting the number of clusters in a data set. See Gordon (1999) for an overview of classical approaches. Those procedures were developed for independent data, but they can also be used in clustering time series. One can roughly classify those methods into three main categories. Methods in the first category use a criterion to measure the homogeneity within the clusters and maximize it. These methods assume the existence of multiple clusters, so that they cannot be used to decide whether the data need to be split into clusters. For clustering with variables a commonly used method, proposed by Calinski and Harabasz (1974), is to maximize the ratio of the average between clusters sum of squares and the average within-cluster sum of squares. For hierarchical clustering a popular criterion is the Silhouette statistic of Rousseeuw (1987), which is discussed in the next subsection.

The second category of methods consists of those that estimate the number of components in a mixture of distributions. In model-based clustering this is usually carried out by an information criterion; see Fraley and Raftery (2002). In a Bayesian approach, it is by the posterior probability of each mixture (Fruhwirth-Schnatter and Kaufmann, 2008). Non-parametric approaches search for the number of modes in the data; see, for instance, Cuevas *et al.* (2000) and Peña *et al.* (2012). modes=connected components of estimator of {f>c}

The third category includes methods that select the number of clusters by comparing a homogeneity measure of the classification obtained from the data to the same measure obtained in random samples generated by some reference distribution under the assumption of a single cluster. This category was initiated by the Gap criterion of Tibshirani *et al.* (2001), that is based on comparing the observed distances between elements to be clustered to a reference distribution assuming homogeneity. For independent data, Sebastiani and Perls (2016) compute the distribution of the $k-1$ distances or heights in the dendrogram, denoted by $H_o = \{h_{1o}, \ldots, h_{k-1,o}\}$, at which two subjects are merged in a hierarchical agglomerative clustering and compare the distribution of H_o to those obtained by applying B random permutations to the columns of the data matrix. In this way, B new sets of dendrogram heights are obtained. They are denoted by $H_i = \{h_{1i}, \ldots, h_{k-1,i}\}$, for $i = 1, \ldots, B$, and the average reference heights are, $H_R = \{h_{1R}, \ldots, h_{k-1,R}\}$, where $h_{jR} = \sum_{i=1}^{B} h_{ji}/B$. They proposed to standardized the heights by dividing them by $\sqrt{k}$, where k is the number of variables, and build a QQ-plot of the standardized observed heights, SH_o, and the expected standardized heights under homogeneity, SH_R. A large deviation from a straight line of the QQ-plot is indicative of the existence of clusters, although a formal test is not proposed. The number of clusters is found by cutting the dendrogram at some large percentile of the reference distribution of the heights.

The procedure we propose belongs to this third category and will be presented in the next section. It has the advantage of being coherent, that is, if the clusters found are considered as a new sample and tested for homogeneity, then they will be found as homogeneous. These property is not shared by two of the most commonly used criteria for selecting the number of hierarchical clusters, the Silhouette and the Gap statistics, that are briefly review next.

It measures how well an object matches the clustering.

3.1. Silhouette Statistic

A graphical display partioning technique based on comparison of tightness and separation

The Silhouette statistic of Rousseeuw (1987) computes a measure of the quality of the clustering and selects the number of clusters to maximize this measure. Suppose we have made a classification of the time series in g clusters with $g > 1$ by using some distance or dissimilarity measure between the series i and i', say $d(i, i')$. Let $a(i)$ be

DOI: 10.1111/jtsa.12706

the average distance of the ith time series to the others in its group and $c(i)$ to those of the closet group, that is the minimum average distance to the series in the other groups. The *Silhouette measure* defines the quality of the classification of the ith series as the relative difference between the average distance of this series to those in the same group and this average distance with series of the closest group. Then,

$$s_g(i) = \frac{c(i) - a(i)}{\max(a(i), c(i))},$$

where g is the number of clusters. If $s_g(i) > 0$ the ith time series is well classified and the maximum possible value of the Silhouette measure is $s_g(i) = 1$, which occurs when $a(i) = 0$. A global measure of the quality of the classification is the average value of $s_g(i)$ over all time series. This is the Silhouette statistic, is given by

$$S(g) = \frac{1}{k}\sum_{i=1}^{k} s_g(i). \tag{9}$$

The larger $S(g)$ the better the global classification, and the number of clusters is selected by maximizing $S(g)$ in (9) over the possible set $g \in \{2, \ldots, G\}$ with G being a prespecified maximum number of clusters. Note that the Silhouette statistic is not appropriate to decide whether we should form clusters or not, because neither the Silhouette measure nor the Silhouette statistic are defined for a single cluster.

The Silhouette criterion works well when the clusters are of similar sizes, but may fail to find clusters with a small number of time series. The reason is that a few series with low negative values in (9) would have a small effect on the average. Also, the presence of outliers may affect the statistic and a robust Silhouette statistics has been proposed in Alonso *et al.* (2020).

3.2. Gap Statistic

An alternative method to select the number of clusters is the Gap statistics, introduced by Tibshirani *et al.* (2001), that compares a measure of the average distance in the sample within the clusters with the expected value of this measure under the hypothesis of homogeneity, that is, no clusters or a single cluster.

Consider a classification of observations in $g > 1$ clusters. Let G_j denote the indices of observations in the jth cluster. An indicator of the homogeneity of the observations in the cluster G_j, for $1 \le j \le g$, is the average value of the squared distances between the series in this cluster G_j, given by

$$D_j = \frac{1}{2k_j}\sum_{i,i' \in G_j} d^2(i, i'), \tag{10}$$

where k_j is the number of series in G_j, $\sum_{j=1}^{g} k_j = k$, and in the definition of D_j as each distance is included twice, the sum is divided by $2k_j$. A quality measure of the clustering made with g clusters is the sum of the average squared distances:

$$W_g = \sum_{j=1}^{g} D_j. \tag{11}$$

The gap statistic is the difference, in logarithm, between the average value of B bootstrap samples, $W_{b,g}$, for $b = 1, \ldots, B$, computed using samples of data generated by the same reference distribution, and the observed W_g value. Specifically, for $g = 1, \ldots, G$, where G is the maximum number of clusters, the following value is computed,

$$Gap(g) = \sum_{b=1}^{B} \log(W_{b,g})/B - \log(W_g), \tag{12}$$

and the larger the Gap the stronger the discrepancy between what is expected under homogeneity and what is observed in the data. Therefore the selection of the number of cluster is made by maximizing this difference, subject to the sample variability.

To estimate the gap statistics a reference distribution to represent homogeneity (no clusters) must be chosen. The authors proved that for a single feature, among all uni-modal distributions, the uniform distribution is the most likely to produce spurious clusters. When clustering with several features they proposed two alternatives to generate the reference values from each feature. The first one is from an uniform distribution over the range of values for this feature. The second one from a uniform distribution over a box aligned with the principal components of the data (see Tibshirani *et al.*, 2001 for details). Then, the number of clusters is selected by the smallest value of g that verifies,

$$Gap(g) \geq Gap(g+1) - sd_g\sqrt{1+1/B}, \tag{13}$$

where sd_g is the standard deviation of the values $\log(W_{b,g})$ in the B bootstrap samples.

4. THE PROPOSED JUMP DETECTION PROCEDURE

Consider the collection of time series $\mathbf{X}_{T\times k}$ in (1). Suppose that there are exactly G clusters in the data. Then, for the ith time series $\boldsymbol{X}_i = (x_{i1}, \ldots, x_{iT})'$, where $1 \leq i \leq k$, we can write the mixture distribution of $\boldsymbol{X}_i$ as

$$f(\boldsymbol{X}_i) = \sum_{g=1}^{G} \alpha_g f_g(\boldsymbol{X}_i), \tag{14}$$

where $\sum_{g=1}^{G}\alpha_g = 1, \alpha_g > 0$, and the α_g and the $f_g(\boldsymbol{X})$ distribution are unknown. Selecting clusters in this case is amount to grouping the series that have been generated by the same distribution $f_g(\boldsymbol{X})$, or by the same sequence of coefficients $\{\pi_{ij}\}_{j=1}^{\infty} \equiv \boldsymbol{\Pi}_i$ defined in (1). Assume also that we have defined a distance or dissimilarity measure between two time series, say $D(\boldsymbol{X}_i, \boldsymbol{X}_j) = D_{ij} \geq 0$ that is consistent, that is, $\lim_{T\to\infty} D_{ij} = 0$ if $f_i(\boldsymbol{X}) = f_j(\boldsymbol{X})$ and $\lim_{T\to\infty} D_{ij} > 0$, when $f_i(\boldsymbol{X}) \neq f_j(\boldsymbol{X})$. This is not a stringent assumption, for instance, the distance between sample autocorrelations will be consistent, as the sample ACFs converge to their population counterparts under rather weak conditions. Using this measure the clusters in the population are well defined.

Consider a realization of k time series with T observations, denoted by $\boldsymbol{X}_{T\times k}$. Each column $\boldsymbol{X}_i$ is a time series that has been generated from the population mixture distribution in (14). If we use the dissimilarity measures D_{ij} between two series, these values will tend to zero as T increases for series of the same cluster and they will be positive for series of different clusters. Thus, we expect that, for sufficiently large sample size T,

$$\max_{1\leq g\leq G} \max_{i,j\in C_g} D_{ij} < \min_{1\leq g\leq G} \min_{i\in C_g, j\notin C_g} D_{ij}. \tag{15}$$

where C_g denotes the gth cluster.

Next, we extend the dissimilarity measures to two groups of series by choosing a given rule (e.g. single linkage, complete linkage, etc.), and apply a hierarchical clustering to the data. The output of the clustering will be a sequence of dendrogram heights, $\mathbf{h}^o = \{h_1^o \leq h_2^o \leq \cdots \leq h_{k-1}^o\}$ that correspond to the distances or dissimilarities between groups of series in each of the $k-1$ steps of merging the k series. In the dendrogram, the merging of groups could be made under two different scenarios: (i) the series merged are from the same cluster; (ii) the series merged belong to two different clusters. Let $E(h_i^s)$ and $E(h_i^d)$ be the expected values of the random variable h_i^o under the two scenarios, where the superscript 's' and 'd' denote that the series merged are from a single or different clusters, respectively. By the condition (15), we have that $E(h_i^d) > E(h_i^s)$ and denote their difference as

$$\Delta_i = E(h_i^d) - E(h_i^s) > 0. \tag{16}$$

DOI: 10.1111/jtsa.12706

The observed sequence of heights then assumes the form

$$h_i^o = E(h_i^s) + I_i\Delta_i + \varepsilon_i \tag{17}$$

where I_i is a dummy variable that takes the value 0 if the ith merge is combining series from the same cluster and assumes the value 1, otherwise, that is, merging series from two different clusters. The noise ε_i is the difference between the ith observed height and its expected value under one of the two possible scenarios and, conditioning on the scenario, will be a random variable with zero mean.

Define the ith jump as $J_i = h_{i+1}^o - h_i^o$, for $i = 1, \ldots, k-2$, and denote the collection of all jumps as $\mathbf{J}_k^o = \{J_1^o, J_2^o, \ldots, J_{k-2}^o\}$. Note that the jumps must be non-negative, but they are not necessarily monotonously increasing. To see this, we have

$$J_i^o = h_{i+1}^o - h_i^o = E(h_{i+1}^s) - E(h_i^s) + I_{i+1}\Delta_{i+1} - I_i\Delta_i + u_i, \tag{18}$$

where $u_i = \varepsilon_{i+1} - \varepsilon_i$ is, given the ith and $(i+1)$th scenarios, a zero mean random noise variable and $E(h_{i+1}^s) > E(h_i^s)$. If $I_i = 0$, $I_{i+1} = 1$, and $I_{i+2} = 0$, then the $(i+1)$th merge in the dendrogram combines two different groups while the ith and $(i+2)$th merges are combining series from the same group, and we have

1. Do multiple clusters exist
2. If exist, how many?

$$J_i^o = E(h_{i+1}^s) - E(h_i^s) + \Delta_{i+1} + u_i \tag{19}$$

$$J_{i+1}^o = E(h_{i+2}^s) - E(h_{i+1}^s) - \Delta_{i+1} + u_{i+1}. \tag{20}$$

If Δ_{i+1} is large, it is likely that $J_i^o > J_{i+1}^o$. On the other hand, if all the series are homogeneous and there is a single cluster, then $I_{i+1} = I_i = 0$, and, taking expectation in Equation (18), the jumps are expected to vary slowly following the sequence $E(h_{i+1}^s) - E(h_i^s) > 0$. However, when multiple clusters exist, we would expect some large jumps in $\mathbf{J}_k^{(o)}$ caused by merging different groups of series together. Furthermore, the number of large jumps in $\mathbf{J}_k^{(o)}$ should be informative in selecting the number of clusters.

Consider the asymptotic case and suppose that we are using a consistent criterion so that $\lim_{T\to\infty} D_{ij}^s = 0$, when both series belong to the same cluster and $\lim_{T\to\infty} D_{ij}^d > 0$, when they belong to different clusters. Then, $\lim_{T\to\infty} E(h_i^s) = 0$ and the jumps will have a positive value when series from different groups are merged. Therefore, when $T \to \infty$ the number of positive jumps plus one will indicate the number of groups.

In finite samples we face two difficulties. The first one is how to decide when a jump in height is statistically significant. This is equivalent to detecting the existence of multiple clusters in the data set $\boldsymbol{X}_{T\times k}$. The second difficulty is to determine the number of clusters, if they exist. To overcome these two difficulties we need a reference distribution for the observed jumps $\boldsymbol{J}_k^o$ under the assumption of homogeneity, or a single cluster, that can be used to perform a statistical test for the existence of multiple clusters. Moreover, when the test statistic rejects the null hypothesis of a single cluster, we can apply again the test statistic to individual cluster to check for the existence of sub-clusters. When all clusters cannot be further divided, the number of clusters is identified.

Run the test again for sub-clusters to determine the existence of single cluster or not.

4.1. Reference Distribution of the Jumps Under Homogeneity

Consider the jumps of the dendrogram of the data $\boldsymbol{X}_{T\times k}$, namely $\boldsymbol{J}_k^o = \{J_1^o, J_2^o, \ldots, J_{k-2}^o\}$. We want to compare these jumps with those of a reference distribution generated under the null hypothesis of a single cluster, that is, all the series are homogeneous. This reference distribution is generated by selecting a 'representative' series from $\boldsymbol{X}_{T\times k}$ and then generating B new data sets of k time series, denoted $\boldsymbol{X}^{(b)}$, for $b = 1, \ldots, B$, where all the series in a given new data have been generated following the same distribution as the selected series. The same clustering method is then applied to each of the B bootstrap data set to obtaining the jumps

$\boldsymbol{J}_k^{(b)} = \{J_1^{(b)}, J_2^{(b)}, \ldots, J_{k-2}^{(b)}\}$. Note that $J_i^{(b)}$ are non-negative values but they do not necessarily form an increasing sequence for $1 \le i \le k-2$.

The representative series can be selected in many ways. For instance, it can be the median empirical dynamic quantile of $\boldsymbol{X}_{T\times k}$ (see Peña *et al.*, 2019). The median dynamic quantile is the time series in $\boldsymbol{X}_{T\times k}$ that minimizes the L_1 distance between it and other series in $\boldsymbol{X}_{T\times k}$. The selection of this series has the advantage of being, in global terms, in the middle of the data set. A simpler alternative is to select a series randomly from $\boldsymbol{X}_{T\times k}$. This approach has computational advantage when k is large, but the selected series may not be a good representative for the data. In practice, one can repeat the proposed method multiple times if this simple approach is used. A third approach is to select an extreme series from $\boldsymbol{X}_{T\times k}$ based on some criteria of extremeness.

The problem of generating random samples $\boldsymbol{X}^{(b)}$ from a given representative series has been extensively studied in the literature. See Bühlmann (2002) and Kreiss and Lahiri (2012) for some reviews. Indeed, several methods have been proposed and each requires different conditions for its validity. The first method is the block bootstrap procedure, which is non-parametric; see Hall (1985) and Künsch's (1989). The second method, proposed by Bühlmann (1998), is the AR sieve bootstrap based on resampling (centered) residuals from an autoregressive approximation of the given process. This procedure can also be regarded as model-free within the class of linear processes. A smoothed sieve bootstrap procedure has been proposed by Alonso *et al.* (2002). Finally, a parametric bootstrap method has also been used by sampling from the residuals of the fitted ARMA model; see, for instance, Tsay (1992).

Kreiss *et al.* (2011) showed that the AR-sieve bootstrap is valid for stationary processes of (1). Therefore, we use the AR sieve bootstrap approach to generate the bootstrap samples in this article and employ the following procedure to generate a reference distribution for $\boldsymbol{J}_k^o$:

1. Select the number of bootstrap samples B. We set $B = 100$ as the default.
2. Find the median EDQ (empirical dynamic quantile) of $\boldsymbol{X}_{T\times k}$. Denote the selected series by x_{ut} and fit an AR(p) model to it with the order p being selected by an information criterion, such as AIC. Compute the residuals of the fitted AR(p), center the residuals by removing their sample mean, and denote the centered residuals $\{\widehat{a}_{ut}\}$.
3. For $b = 1, \ldots, B$, do
   (i) Sample with replacements T values from $\{\widehat{a}_{ut}\}$ and use the fitted AR coefficients to generate a time series. This process is repeated k times and denote the resulting bootstrap samples by $\boldsymbol{X}^{(b)}$, for $b = 1, \ldots, B$.
   (ii) Use the same hierarchical agglomerative procedure to cluster $\boldsymbol{X}^{(b)}$ and let $\boldsymbol{J}_k^{(b)} = \{J_1^{(b)}, J_2^{(b)}, \ldots, J_{k-2}^{(b)}\}$ be the sequence of jumps of the dendrogram for $\boldsymbol{X}^{(b)}$. Note that $J_i^{(b)}$ are non-negative values but they do not necessarily form an increasing sequence for $1 \le i \le k-2$.
4. Let $\boldsymbol{J}$ be the $(k-2)\times B$ matrix such that its bth column is $\boldsymbol{J}_k^{(b)}$, for $b = 1, \ldots, B$. We use $\boldsymbol{J}$ to make statistical inference.

4.2. A Jump Test for the Existence of Clusters

Given the reference bootstrap distribution $\boldsymbol{J}$, we consider the hypothesis testing H_0: A single cluster versus H_a: More than one cluster. That is, $H_0 : G = 1$ vs $H_a : G > 1$, under the mixture model in (14). As discussed in the previous section, large jumps in $\mathbf{J}_k^{(o)}$ should occur at the upper tail so that, for a given small upper tail probability α, one can employ the $(1-\alpha)$ quantile to perform the test. Specifically, let $J_{1-\alpha}^o$ be the $1-\alpha$ quantile of $\boldsymbol{J}_k^o$ and $\{J_{1-\alpha}^{(i)}\}$ be the $1-\alpha$ quantile of $\boldsymbol{J}_k^{(i)}$, for $i = 1, \ldots, B$. Here $\{J_{1-\alpha}^{(i)}\}_{i=1}^B$ form a reference distribution of $J_{1-\alpha}^o$. Consequently, the proposed test statistic is

$$T(\alpha) = J_{1-\alpha}^o. \tag{21}$$

DOI: 10.1111/jtsa.12706

For a given type-I error α_1, one can choose the $1-\alpha_1$ quantile of $\{J^{(i)}_{1-\alpha}\}_{i=1}^{B}$ as the critical value for the test statistic in (21). Let $C(\alpha_1)$ be the $1-\alpha_1$ quantile of $\{J^{(i)}_{1-\alpha}\}_{i=1}^{B}$. The null hypothesis of a single cluster is rejected if $T(\alpha) > C(\alpha_1)$.

In practice, it remains to choose proper values for the small upper tail probability α and type-I error α_1 to perform the proposed test. A good choice would depend on many factors, including the sample size T, the number of series k, the selected features to compute the distance, and the underlying dynamic dependence of the series. To simplify the choices, we propose to employ $\alpha = \alpha_1$, but instead of using a single α, we employ three test statistics $\{T(0.01), T(0.025), T(0.05)\}$ to perform the hypothesis test. The decision rule is that the null hypothesis H_0 is rejected if it is rejected by any one of the three test statistics. Our recommendation is based on several considerations. First, if the data under study consist of multiple clusters of different sizes, then a single α is likely to encounter some difficulties as the number of series in a cluster plays an important role in determining the reference distribution of the proposed test statistic. It is then reasonable to use multiple values of α. Second, we have conducted some simulation studies with both small and large cluster sizes and found that the use of three values, that is, $\alpha \in \{0.01, 0.025, \text{ and } 0.05\}$, works reasonably well. These three α's seem to be a good compromise among different cluster sizes. Indeed, our simulation results and empirical data analysis of the next two sections show that the recommended testing procedure with $\{T(0.01), T(0.025), T(0.05)\}$ works reasonably well in finite samples.

In our implementation of the proposed test procedure, we always use 100 AR-sieve bootstrap to obtain the reference distribution. Also, if the number series k is not large, say $k \leq 100$, we use the median empirical dynamic quantile as the representative series in AR-sieve bootstraps. For large k, we may use simple random sample to select a representative series in AR-sieve bootstraps to reduce the computing time.

4.3. Selecting the Number of Clusters

If the test statistic of (21) rejects the null hypothesis of a single cluster, the data set $\boldsymbol{X}_{T\times k}$ has multiple clusters. The next question is how to determine the number of clusters. To this end, we take advantages of the powerful test statistic in (21). Specifically, let $G_{\max}$ be the maximum number of clusters allowed for the observed data. If the test statistic suggests the existence of multiple clusters, we consider the following procedure to select the number of clusters.

1. Start with $\hat{g} = 2$.
2. For a given $\hat{g}$, let $S(\hat{g}) = \{C_1, \ldots, C_{\hat{g}}\}$ be the clusters obtained assumed that there were $\hat{g}$ clusters. This is done using the hierarchical clustering of $\boldsymbol{X}_{T\times k}$.
3. For $j = 1, \ldots, \hat{g}$, perform the proposed test using the data set C_j. If the null hypothesis is rejected, then let $\hat{g} = \hat{g} + 1$ and go the step 2. If the test statistic fails to reject the null hypothesis of a single cluster for all data sets in $S(\hat{g})$, then $G = \hat{g}$ is identified and stop of the search process.
4. The process also stops when $\hat{g}$ reaches $G_{\max}$.

In other words, we start with two clusters and test the null hypothesis that each cluster is indeed homogeneous based on the test statistic (21). If any of the cluster rejects the null hypothesis of a single cluster, we increase the number of clusters by one and repeat the testing procedure. In this way, the number of clusters is selected when all clusters cannot be divided further based on the proposed test statistic.

5. SIMULATION AND COMPARISON

We conduct simulation studies to check the performance of the proposed clustering method in finite samples and to compare the proposed method of selecting the number of clusters with the Silhouette and Gap statistics of Sections 3.1 and 3.2. The last two methods are available in the R package **SLBDD**.

We consider several data generating processes (DGP) to examine the finite sample performance of the proposed method and to compare different methods in selecting the number of clusters. Unless specified otherwise, the innovation series $\{a_{it}\}$ are sequences of independent and identically distributed N(0, 1) random variates. We use the

first five lags of sample ACF as the selected features and employ Euclidean distance to quantify the dissimilarity between time series. Also, the complete linkage is used in the hierarchical clustering. For each time series in a given data set $\boldsymbol{X}_{T\times k}$, we start with zero initial values, but discard the first 20 observations to mitigate the impact of starting values. For each simulation study, 500 iterations are used. The AR order of the selected representative series of the proposed method is chosen by the Akaike information criterion of the **ar** command in R.

The five DGP used in the simulation are given below:

1. Simulation 1: A single cluster with $k = 25$. The model used in $x_{it} = 1.3x_{i,t-1} - 0.4x_{i,t-2} + a_{it}$, for $i = 1, \ldots, 25$.
2. Simulation 2: Two clusters with cluster size (20, 10).
   - Cluster 1: $x_{it} = 0.8x_{i,t-1} + a_{it}$, for $i = 1, \ldots, 20$.
   - Cluster 2: $x_{it} = a_{it} + 0.6a_{i,t-2}$, for $i = 21, \ldots, 30$.
3. Simulation 3: Three clusters with cluster size (20, 15, 10).
   - Cluster 1: $x_{it} = 0.8x_{i,t-1} + a_{it} - 0.4a_{i,t-1}$, for $i = 1, \ldots, 20$.
   - Cluster 2: $x_{it} = a_{it} + 0.5a_{i,t-1}$, for $i = 21, \ldots, 35$.
   - Cluster 3: $x_{it} = 1.4x_{i,t-1} - 0.45x_{i,t-2} + a_{it}$, for $i = 36, \ldots, 45$.
4. Simulation 4: Three clusters with cluster size (10, 15, 8) and Student t_{10} innovations.
   - Cluster 1: $x_{it} = 1.3x_{i,t-1} - 0.4x_{i,t-2} + a_{it}$, for $i = 1, \ldots, 10$.
   - Cluster 2: $x_{it} = 0.8x_{i,t-1} + a_{it}$, for $i = 11, \ldots, 25$.
   - Cluster 3: $x_{it} = a_{it} - 0.1a_{i,t-1} - 0.42a_{i,t-2}$, for $i = 26, \ldots, 33$.
5. Simulation 5: Four clusters with cluster size (10, 15, 15, 8).
   - Cluster 1: $x_{it} = 1.4x_{i,t-2} - 0.48x_{i,t-2} + a_{it}$, for $i = 1, \ldots, 10$.
   - Cluster 2: $x_{it} = 0.8x_{i,t-1} + a_{it}$, for $i = 11, \ldots, 25$.
   - Cluster 3: $x_{it} = a_{it} + 0.1a_{i,t-1} - 0.42a_{i,t-2}$, for $i = 26, \ldots, 40$.
   - Cluster 4: $x_{it} = a_{it} + 0.6a_{i,t-1}$, for $i = 41, \ldots, 48$.

Table I reports the percentages of correct selection of the number of clusters for the five DGP listed above when the total number of series k is small, say $k < 50$. In all cases, the sample size used is $T = 300$ and the number of iterations is 500. From the table, we make the following observations. First, no selection method consistently dominates the others in the five simulation studies. Second, the Silhouette statistic fails to select the correct number of clusters in three of the five simulations. Except for the single cluster case of Simulation 1, this finding is rather surprising. But the Silhouette statistic works well in the simple case of two clusters. Third, the Gap statistic works reasonably well except for the difficult case of Simulation 5. As a matter of the fact, it fails to select the correct number of clusters in this case. Fourth, the proposed testing procedure works in all five simulations. As expected, it fares well when there is a single cluster. More importantly, it is the only method that fares reasonably well in the difficult case of Simulation 5. Finally, both the proposed testing procedure and the Gap statistic are less sensitive to the heavy tailed t-distribution.

Table II shows the results of correct specification of the number of clusters for three methods entertained when the total number of series k is large. From the table, we see that, as expected, it is harder for any clustering method to correctly specify the number of clusters when k is large. The Gap statistic is particularly sensitive to the increase in the number of series. For the case of two clusters the proportion of times that the Gap statistic finds the true number of clusters decreases from 0.954 with 30 series to 0.068 with 200 series. For a single cluster its successful rate drops from 0.932 to 0.298 when k increases from 25 to 200. However, it is less affected in the case of three clusters. On the other hand, the proposed testing procedure is still capable of specifying the correct number of clusters, albeit at a lower successful rate. The Silhouette method works well in the simple case of two clusters, but its performance deteriorates markedly when the data consist of more than two clusters. Finally, we increased the sample size from 300 to 500 for the case of four clusters and found that all three methods improve in specifying

Journal of Time Series Analysis published by John Wiley & Sons Ltd. DOI: 10.1111/jtsa.12706

Table I. Empirical probabilities of selecting the correct number of clusters by four methods

DGP simulation	Cluster sizes	Noise a_{it}	Methods Jump	Sil	Gap
Single AR(1)	(25)	N(0, 1)	0.930	0.000	0.932
AR(1), AR(2)	(20,10)	N(0, 1)	0.784	1.000	0.954
ARMA(1,1), AR(1), AR(2)	(20,15,10)	N(0, 1)	0.566	0.468	0.950
AR(2), AR(1), ARMA(1,1)	(10,15,8)	t_{10}	0.660	0.000	0.728
AR(2), AR(1), ARMA(1,1), MA(1)	(10,15,15,8)	N(0, 1)	0.774	0.000	0.014

Note: In the table, DGP stands for data generating process, and Jump, Sil, and Gap denote the proposed testing method, the Silhouette criterion, and Gap statistics, respectively. The results are for sample size $T = 300$ and are based on 500 repetitions.

Table II. Empirical probabilities of selecting the correct number of clusters by six methods

DGP simulation	Cluster sizes	Sample size	Methods Jump	Sil	Gap
Single AR(1)	(200)	300	0.910	0.000	0.298
AR(1), AR(2)	(120,80)	300	0.654	1.000	0.068
ARMA(1,1), AR(1), AR(2)	(80,60,40)	300	0.488	0.424	0.742
AR(2), AR(1), ARMA(1,1), MA(1)	(40,70,60,40)	300	0.328	0.000	0.232
AR(2), AR(1), ARMA(1,1), MA(1)	(40,70,60,40)	500	0.568	0.332	0.476

Note: In the table, DGP stands for data generating process and Jump, Sil, and Gap denote the proposed testing procedure, Silhouette and Gap statistics, respectively. In all simulations, the noises are independent N(0, 1) random variates. The results are based 500 simulation repetitions.

the correct number of clusters. This result indicates the importance of obtaining accurate estimates of the selected features in time series clustering.

We note that the proposed testing procedure requires intensive computation when the total number of series k is large. On the other hand, both Silhouette and Gap statistics can easily be computed. It is important to seek procedures to reduce the computing intensity of the proposed testing procedure.

6. EMPIRICAL ANALYSIS

We consider two empirical analyses in this section. The two data sets used are available from the **SLBDD** package on R CRAN. The first data set, called UMEdata20002018, consists of quarterly gross domestic product (GDP) at market prices, Household and NPISH final consumption expenditure (CON), and Gross Fixed Capital Formation (INV) of 19 Euro Area member countries from 2000 to 2018. The original data were extracted on 7 August 2019 from Eurostat. The second data set consists of hourly $PM_{2.5}$ measurements from 508 different locations in Taiwan in March 2017. It is called TaiwanAirBox032017.

Example 1. (UME data) The three time series used (GDP, CON, and INV) represent the status of the 19 Euro Area economies. We analyze the growth rates of the data, for example, the first difference of log series, with $T = 75$ observations and $k = 57$ series. Figure 1 shows the time plots of the 57 growth series. From the plots, most of the growth rates are small as expected, but a few series seem to have high variabilities after 2015. For clustering analysis, we choose the first five partial autocorrelation coefficients of each series as their selected features, employ Euclidean distances between the features as similarity measures, and use the complete linkage. In this particular instance, both the Silhouette and Gap statistics select two clusters, but our proposed testing approach specifies 3 clusters using the three recommended test statistics ($T(0.01), T(0.025), T(0.05)$). As stated in the proposed procedure, we use the median empirical dynamic quantile as the representative series in AR-sieve bootstrap with 100 bootstrap samples. Figure 2 shows the resulting dendrogram of the data set. From the plot,

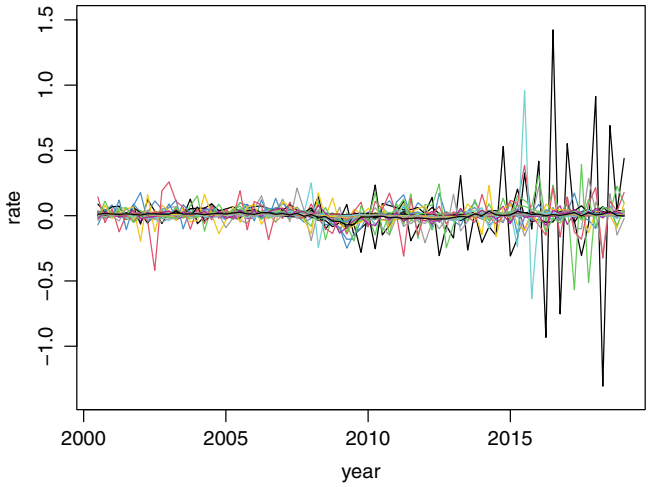

Figure 1. Time plots of quarterly growth rates of gross domestic product, consumption, and investment of 19 Euro Area countries from 2000 to 2018

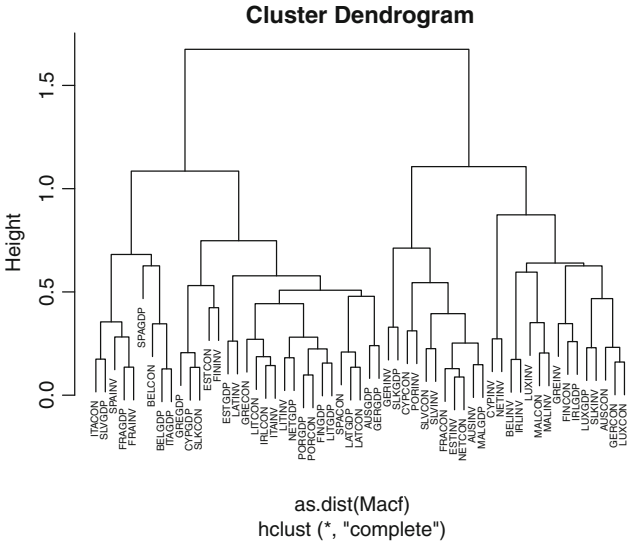

Figure 2. Dendrogram of UME data based on the first five partial autocorrelation coefficients, Euclidean distance, and complete linkage

it seems that the existence of multiple clusters is justified and the number of clusters could range from 2 to 4, depending on the height used to form the clusters.

With three clusters, the numbers of time series in the three clusters are 31, 15, and 11, respectively. Figure 3 shows the time plots of the growth series of the three clusters. Examining the scale of the y-axis, one can easily see that (i) Cluster 2 consists mainly of those growth rate series that have high variabilities in the later part of the time period (i.e., after 2015), but were not seriously impacted by the 2008 financial crisis, (ii) Cluster 3 contains the growth series that are more homogeneous throughout the entire time period and have lower variabilities, and (iii) Cluster 1 consists of growth series that were seriously and negatively affected by the 2008 subprime financial crisis. Checking the series in Cluster 1, it is not surprising to see that 15 out of the 31 series are GDP growth rates, as the 2008 subprime financial crisis led to downturn in many economies worldwide. On the other hand, 8 out of

Journal of Time Series Analysis published by John Wiley & Sons Ltd. DOI: 10.1111/jtsa.12706

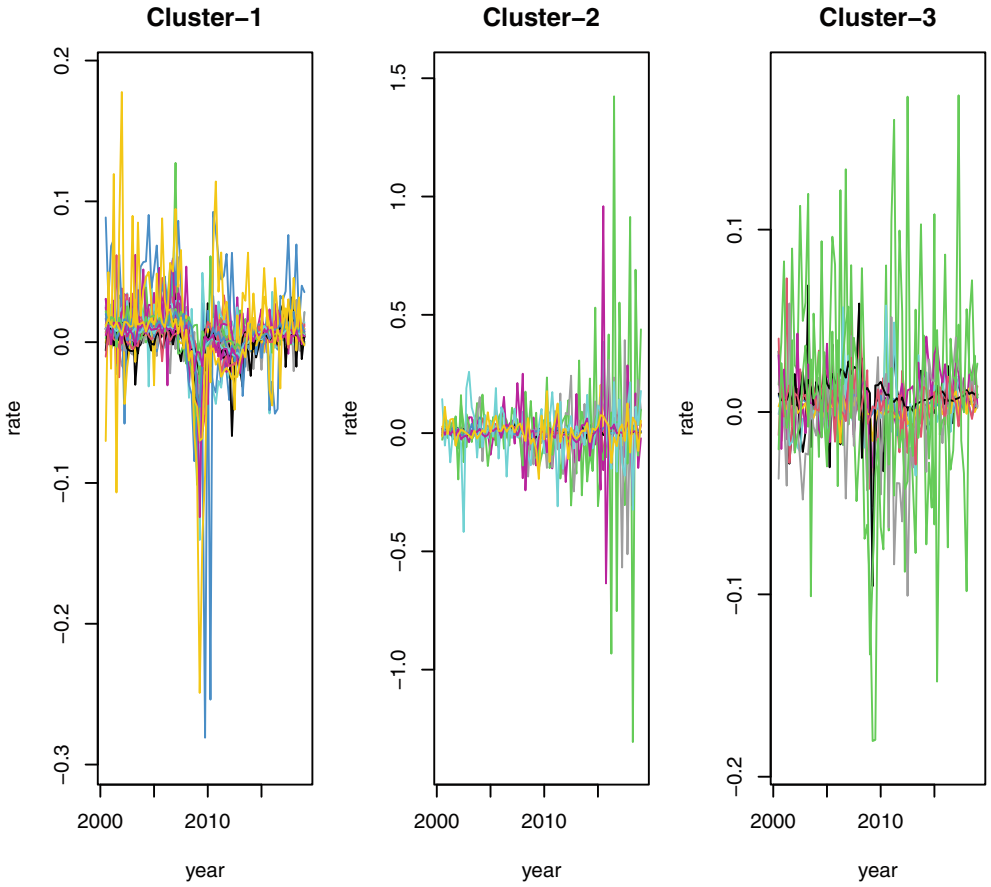

Figure 3. Time plots of the UME growth rate series via clusters

15 series in Cluster 2 are growth rates of gross fixed capital formation, which were less sensitive to the subprime financial crisis. The only two GDP series in Cluster 2 are Ireland and Luxembourg. These two countries were known to fare better during the 2008 financial crisis.

If two clusters are used, the numbers of time series in the clusters are 31 and 26, respectively. Thus, in this particular application, the proposed jump testing approach went one-step farther by dividing the second cluster of Silhouette and Gap methods into two clusters. It is interesting to see that, from the plots in Figure 3, such division is justified and provides more details about the structures of the growth series, especially their behavior during the 2008 financial crisis.

Example 2. (AirBox data) The Taiwan AirBox data include hourly measurements of $PM_{2.5}$ made in Taiwan by Air boxes in March 2017. The data set consists of 516 series each with 744 observations taken at different locations throughout the island. Based on some initial outlier analysis, eight series, (1, 29, 35, 46, 70, 118, 155 and 157), were found as outliers and the analysis we present here employs the remaining 508 series.

Figure 4 shows a time plot of the original data including three dynamic quantiles defined in Peña *et al.* (2019). As some of the series appear to be non-stationary (e.g. contain potential time trends), we apply the first difference to all series to achieve stationarity. As the results obtained using the selected features of sample autocorrelation functions (ACF), partial autocorrelation functions (PACF), and the total correlation are similar, we only present here the results using the Euclidean distance of the first six sample ACF between the series.

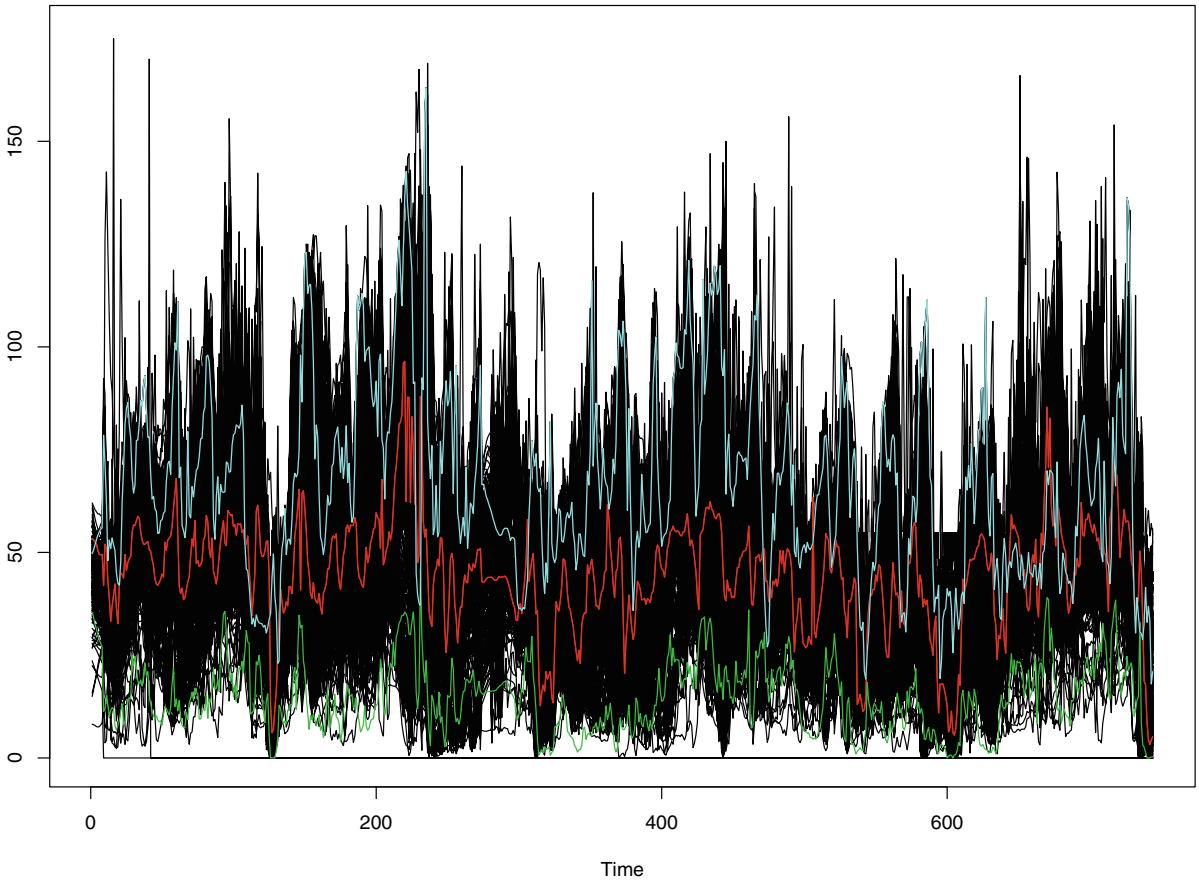

Figure 4. The AirBox Taiwan Data and the empirical dynamic quantiles with probabilities 0.05, 0.5 and 0.95

The dendrogram obtained is shown in Figure 5. From the plot, the existence of multiple clusters seems possible. In this particular application, the Silhouette and Gap methods select 2 and 4 clusters, respectively. On the other hand, the proposed testing approach selects 7 clusters. Again, the median empirical dynamic quantiles are used in the AR-sieve bootstrap as a representative series of a cluster and 100 bootstrap iterations are used.

If two clusters are employed, then the cluster sizes are (384, 124). If four clusters are used, the cluster sizes become (27, 357, 94, 30). It seems that the Gap statistic further divides each cluster identified by the Silhouette method into two sub-clusters. If seven clusters are employed, then the cluster sizes become (8, 19, 180, 177, 94, 26, 4). We see that the proposed testing method divides Clusters 2 obtained by the Gap statistic into two sub-clusters of sizes (180, 177), Cluster 1 of Gap statistic also into two sub-clusters of sizes (8, 19), and Cluster 4 of Gap statistics into two sub-clusters of sizes (26, 4).

Table III provides the cluster size and average sample autocorrelation coefficients (lag-1 to lag-6) of the seven clusters specified by the proposed testing procedure. From the table, we make the following observations. First, Cluster 1 seems to contain the MA(1) type of series whereas Cluster 4 seems to consist of white noise processes. Second, Cluster 6 also contains the MA(1) type of series, but, in contrast to those in Cluster 1, the series have negative lag-1 ACF. Third, Cluster 2 consists of MA(4) type of series with serial dependence focusing on lags 1 and 4. Fourth, Cluster 3 consists of MA(3) type of time series. On the other hand, Cluster 5 contains MA(2) type of series. Finally, Cluster 7 contains only four time series that have negative serial correlations at lags 1 and 3. These observations provide empirical justifications for having seven clusters.

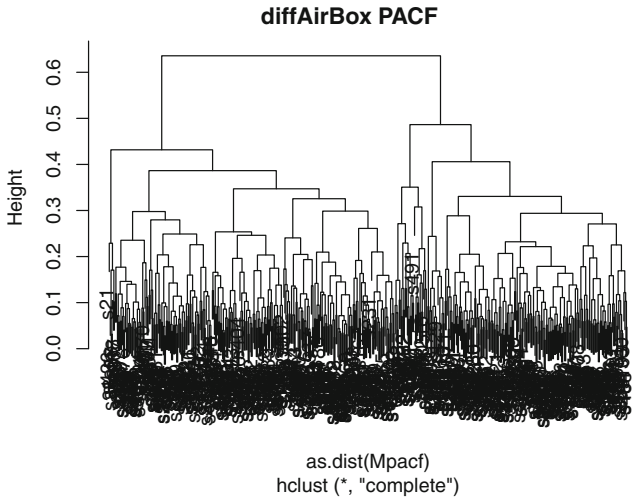

Figure 5. Dendrogram obtained with complete linkage between the first six sample autocorrelation coefficients for the first difference of Taiwan AirBox Data

Table III. Average sample autocorrelations ($\bar{r}_i$) of seven clusters found by the proposed procedure for the AirBox data

Cluster	1st	2nd	3rd	4th	5th	6th	7th
$\bar{r}_1$	0.229	0.207	0.109	0.059	−0.019	−0.137	−0.184
$\bar{r}_2$	0.057	−0.024	−0.104	−0.045	−0.112	−0.035	0.077
$\bar{r}_3$	0.021	−0.061	−0.105	−0.031	−0.056	−0.035	−0.112
$\bar{r}_4$	−0.035	−0.126	−0.041	−0.040	−0.025	−0.032	−0.023
$\bar{r}_5$	−0.046	−0.098	−0.012	−0.046	−0.018	0.012	−0.058
$\bar{r}_6$	−0.066	−0.049	−0.028	−0.039	−0.020	−0.047	−0.018
Size	8	19	180	177	94	26	4

Note: The cluster size is also given.

7. CONCLUSIONS

We have presented a testing approach to determine the number of clusters in a given set of univariate time series based on some selected features of individual series. The main tool used by the proposed procedure is a test on a high upper quantile of the observed jumps (or height changes) in the dendrogram. This upper quantile is compared to its reference distribution which is generated by the AR-sieve bootstrap method assuming that there is only a single cluster.

The proposed jump procedure seems to work well for clustering stationary linear time series, but it has some limitations. First, it may not work well for clustering nonlinear time series. Second, it assumes that the data set consists of independent time series. These limitations can be investigated in future work. For example, block bootstrap can be used to generate a reference distribution for nonlinear time series, especially if the serial dependence is not strong. Also, the ideas presented in this article can be extended to obtain a new criterion for clustering time series by their cross-dynamic dependency. This extension is non-trivial because the generation of bootstrap samples to form a data set in which all series share a similar cross-correlation structure is not straightforward. Finally, we have not considered clustering segments of time index so that the joint distribution of the series is similar in a segment. This extension would be useful in applications if stationarity of the time series involved is questionable. It can also be used to detect global structural breaks among a set of univariate time series.

DOI: 10.1111/jtsa.12706 *Journal of Time Series Analysis* published by John Wiley & Sons Ltd.

ACKNOWLEDGMENTS

This work is dedicated to Masanobu Taniguchi, who made pioneering work on clustering time series that we admire very much. We are very thankful to the Editors of this issue for inviting us to participate and to the two referees for their helpful comments that have stimulated us to improve our procedure. The research of D. Peña was partially supported by the Spanish Agencia Nacional de Evaluación under grant PID2019-109196GB-I00 and the research of R. Tsay is supported in part by the Booth School of Business, University of Chicago. The authors wish to thank Mr. S.C. Huang for assistance in running the large scale simulation studies.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in the R package SLBDD available at CRAN (https://cran.r-project.org/package=SLBDD).

REFERENCES

Aghabozorgi S, Shirkhorshidi AS, Wah TY. 2015. Time-series clustering: a decade review. *Information Systems* **53**:16–38.

Alonso AM, Berrendero JR, Hernández A, Justel A. 2006. Time series clustering based on forecast densities. *Computational Statistics and Data Analysis* **51**:762–766.

Alonso AM, Peña D. 2019. Clustering time series by linear dependency. *Statistics and Computing* **29**:655–676.

Alonso AM, Galeano P, Peña D. 2020. A robust procedure to build dynamic factor models with cluster structure. *Journal of Econometrics* **216**(1):35–52.

Alonso AM, Peña D, Romo J. 2002. Forecasting time series with sieve bootstrap. *Journal of Statistical Planning and Inference* **100**(1):1–11.

Alqahtani A, Ali M, Xie X, Jones MW. 2021. Deep time-series clustering: a review. *Electronics* **10**(23):3001.

Ando T, Bai J. 2017. Clustering huge number of financial time series: a panel data approach with high-dimensional predictors and factor structures. *Journal of the American Statistical Association* **112**(519):1182–1198.

Bandara K, Bergmeir C, Smyl S. 2020. Forecasting across time series databases using recurrent neural networks on groups of similar series: a clustering approach. *Expert Systems with Applications* **140**:112896.

Bouveyron C, Celeux G, Murphy TB, Raftery AE. 2019. *Model-Based Clustering and Classification for Data Science: With Applications in R* Cambridge University Press, Cambridge, UK.

Box GE, Jenkins GM, Reinsel GC, Ljung GM. 2015. *Time Series Analysis: Forecasting and Control* John Wiley and Sons, Hoboken, NJ.

Bühlmann P. 2002. Bootstraps for time series. *Statistical Science* **17**:52–72.

Bühlmann P. 1998. Sieve bootstrap for smoothing in nonstationary time series. *The Annals of Statistics* **26**(1):48–83.

Caiado J, Crato N, Peña D. 2006. A periodogram-based metric for time series classification. *Computational Statistics and Data Analysis* **50**:2668–2684.

Calinski T, Harabasz J. 1974. A dendrite method for cluster analysis. *Communications in Statistics – Theory and Methods* **3**:1–27.

Corduas M, Piccolo D. 2008. Time series clustering and classification by the autoregressive metric. *Computational Statistics and Data Analysis* **52**:1860–1872.

Cuevas A, Febrero M, Fraiman R. 2000. Estimating the number of clusters. *Canadian Journal of Statistics* **28**(2):367–382.

Dias JG, Vermunt JK, Ramos S. 2015. Clustering financial time series: new insights from an extended hidden Markov model. *European Journal of Operational Research* **243**(3):852–864.

Durante F, Pappada R, Torelli N. 2014. Clustering of financial time series in risky scenarios. *Advances Data Analysis and Classification* **8**:359–376.

Fraley C, Raftery AE. 2002. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association* **97**(458):611–631.

Fruhwirth-Schnatter S, Kaufmann S. 2008. Model-based clustering of multiple time series. *Journal of Business and Economic Statistics* **26**:78–89.

Galeano P, Peña D. 2000. Multivariate analysis in vector time series. *Resenhas IMB-USP* **4**(4):383–403.

Gordon AD. 1999. *Classification* CRC Press, London, UK.

Huang X, Ye Y, Xiong L, Lau RY, Jiang N, Wang S. 2016. Time series *k*-means: a new *k*-means type smooth subspace clustering for time series data. *Information Sciences* **367**:1–13.

Li H. 2019. Multivariate time series clustering based on common principal component analysis. *Neurocomputing* **349**:239–247.

Liu Y, Nagahata H, Uchiyama H, Taniguchi M. 2017. Discriminant and cluster analysis of possibly high-dimensional time series data by a class of disparities. *Communications in Statistics – Simulation and Computation* **46**(10):8014–8027.

Maharaj EA, D'Urso P, Caiado J. 2019. *Time Series Clustering and Classification* Chapman and Hall/CRC, London, UK.
Montero P, Vilar JA. 2014. TSclust: an R package for time series clustering. *Journal of Statistical Software* **62**:1–43.
Kakizawa Y, Shumway RH, Taniguchi M. 1998. Discrimination and clustering for multivariate time series. *Journal of the American Statistical Association* **93**(441):328–340.
Kreiss JP, Paparoditis E, Politis DN. 2011. On the range of validity of the autoregressive sieve bootstrap. *The Annals of Statistics* **39**(4):2103–2130.
Kreiss JP, Lahiri SN. 2012. *Bootstrap methods for time series*. In *Handbook of Statistics*, Vol. 30, Elsevier, Amsterdam, Netherlands; 3–26.
Oh DH, Patton AJ. 2023. Dynamic factor copula models with estimated cluster assignments. *Journal of Econometrics* in press.
Peña D, Rodríguez J. 2002. A powerful portmanteau test of lack of fit for time series. *Journal of the American Statistical Association* **97**(458):601–610.
Peña D, Tsay RS. 2021. *Statistical Learning for Big Dependent Data* John Wiley and Sons, Hoboken, NJ.
Peña D, Tsay RS, Zamar R. 2019. Empirical dynamic quantiles for visualization of high-dimensional time series. *Technometrics* **61**:429–444.
Peña D, Viladomat J, Zamar R. 2012. Neares-neighbors medians clustering. *Statistical Analysis and Data Mining* **5**(4):349–362.
Petitjean F, Ketterlin A, Ganarski P. 2011. A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition* **44**(3):678–693.
Pértega S, Vilar JA. 2010. Comparing several parametric and nonparametric approaches to time series clustering: a simulation study. *Journal of Classification* **27**:333–362.
Piccolo D. 1990. A distance measure for classifying ARMA models. *Journal of Time Series Analysis* **2**:153–163.
Rousseeuw PJ. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20**:53–65.
Sebastiani P, Perls TT. 2016. Detection of significant groups in hierarchical clustering by resampling. *Frontiers in Genetics* **7**:144.
Sakiyama K, Taniguchi M. 2004. Discriminant analysis for locally stationary processes. *Journal of Multivariate Analysis* **90**:282–300.
Sezer OB, Gudelek MU, Ozbayoglu AM. 2020. Financial time series forecasting with deep learning: a systematic literature review: 2005–2019. *Applied Soft Computing* **90**:106181.
Taniguchi M, Kakizawa Y. 2012. *Asymptotic Theory of Statistical Inference for Time Series* Springer Science and Business Media, Berlin, Germany.
Tibshirani R, Walther G, Hastie T. 2001. Estimating the number of clusters in a data set via the Gap statistic. *Journal of the Royal Statistical Society, Series B* **63**:411–423.
Tsay RS. 1992. Model checking via parametric bootstraps in time series analysis. *Applied Statistics* **41**:1–15.
Tsay RS. 2020. Testing serial correlations in high-dimensional time series via extreme value theory. *Journal of Econometrics* **216**(1):106–117.
Xiong Y, Yeung DY. 2004. Time series clustering with ARMA mixtures. *Pattern Recognition* **37**(8):1675–1689.
Wang X, Smith K, Hyndman R. 2006. Characteristic-based clustering for time series data. *Data Mining and Knowledge Discovery* **13**(3):335–364.
Wang Y, Tsay RS. 2019. Clustering multiple time series with structural breaks. *Journal of Time Series Analysis* **40**:182–202.
Watanabe T, Shiraishi H, Taniguchi M. 2010. Cluster analysis for stable processes. *Communications in Statistics Theory and Methods* **39**:8–9.
Zhang B, An B. 2018. Clustering time series based on dependence structure. *PLoS ONE* **13**(11):e0206753.
Zhang G, Taniguchi M. 1994. Discriminant analysis for stationary vector series. *Journal of Time Series Analysis* **5**:117–126.

DOI: 10.1111/jtsa.12706 *Journal of Time Series Analysis* published by John Wiley & Sons Ltd.