Università degli Studi di Salerno

DIPARTIMENTO DI SCIENZE ECONOMICHE E STATISTICHE

Corso di Dottorato di ricerca in
Economia e Politiche dei Mercati e delle Imprese
Ciclo: XXXIII
Curriculum: Metodi Statistici

High-Dimensional Time Series Clustering: Nonparametric Trend Estimation

Candidato:
Giuseppe Feo
Matricola 8801000030

Tutor:
Ch.mo Prof.
Francesco Giordano

Coordinatore:
Ch.ma Prof.ssa
Alessandra Amendola

La borsa di dottorato è stata cofinanziata con risorse del
Programma Operativo Nazionale Ricerca e Innovazione 2014-2020 (CCI 2014IT16M2OP005),
Fondo Sociale Europeo, Azione I.1 "Dottorati Innovativi con caratterizzazione Industriale"

UNIONE EUROPEA
Fondo Sociale Europeo

Ministero dell'Università e della Ricerca

PON RICERCA E INNOVAZIONE 2014 - 2020

Anno Accademico 2020 - 2021

Contents

Introduction

The era of big data has produced extensive methodologies for extracting features/patterns from complex time series data. From a data science perspective these methodologies have emerged from multiple disciplines, including statistics, signal processing/engineering, and computer science. Clustering is a solution for classifying enormous data when there is not any previous knowledge about classes obtaining numerosity reduction for instance.

Considering time series as discrete objects, conventional clustering procedures can be used to cluster a set of individual time series with respect to their similarity such that similar time series are grouped into the same cluster. From this perspective time series clustering techniques have been developed, most of them critically depend on the choice of distance (i.e., similarity) measure. In general, the literature defines three different approaches to cluster time series: (i) *Shape-based clustering*, clustering is performed based on the shape similarity, where shapes of two time series are matched using a nonlinear stretching and contracting of the time axes; (ii) *Feature-based clustering*, raw time series are transformed into the feature vector of lower dimension where, for each time series, a fixed-length and an equal-length feature vector is created (usually a set of statistical characteristics); (iii) *Model-based clustering* assumes a mathematical model for each cluster and attempts to fit the data into the assumed model.

Choosing an appropriate representation method can be considered as the key component which effects the efficiency and accuracy of the clustering solution. High-dimensionality and noise are characteristics of the most time series data, consequently, dimensionality reduction methods are used in time series clustering in order to address this issues and promote the performance. Time series trend composition is a very important topic in data analysis, especially in the more recent literature of clustering High-dimensional time series. Checking trend composition is the first step for a further statistical

analysis conducted on a time series. In fact, many of the clustering procedures proposed in the literature are based on the assumption that all the time series considered follow the same trend structure. The latter can be absent, linear or nonlinear. Actually, the true structure of the trend is unknown, therefore a procedure that allows this distinction is necessary before any clustering analysis. With this in mind, the proposed thesis aims to fill this gap.

In particular, the proposal discussed in this thesis regards an embryonic analysis for carrying out a correct further clustering analysis on time series. Precisely, it regards the classification of nonstationary time series, where the nonstationarity is given by the presence of a deterministic trend, by looking at the first derivative of the trend in a context of high-dimensionality and without requiring a pre specified form for the trend. This is achieved by means of a nonparametric estimator which has a very simple form. The idea is to classify the time series by checking the trend first derivative. If the trend is constant, then its first derivative is zero, if the trend is linear, then its first derivative is constant. If none of the previous happens, then the trend is of course nonlinear and then its first derivative will be not constant. In this way the time series can be divided into three groups. This approach can be included in the category of "clustering of time series based on features", since the trend composition can be considered as a feature of the time series. Once the time series are classified it will be possible to apply the most appropriate clustering technique.

The chapters of the thesis will be organized as follows. The first Chapter gives a multidisciplinary overview on the literature of clustering time series with particular emphasis on the statistical point of view; the second Chapter introduces the general setting for the idea behind the proposed procedure together with the statistical tools that will be used in order to present the statistics based on the proposed first derivative estimator; the third Chapter focuses on the theoretical results of the proposed estimator and its statistics in the context of high-dimensionality; the fourth Chapter illustrates simulation studies which evaluate the performances of the proposed procedure for classifying high-dimensional time series; the fifth Chapter presents an application of the proposed procedure on electrical consumption data; finally some concluding remarks and ideas for future works are presented in order to conclude the thesis.

Chapter 1

Time series Clustering

The era of big data has produced extensive methodologies for extracting features/patterns from complex time series data. From a data science perspective these methodologies have emerged from multiple disciplines, including statistics, signal processing/engineering, and computer science. Clustering is a solution for classifying enormous data when there is not any previous knowledge about classes obtaining numerosity reduction for instance.

The goal of clustering is to identify structure in an unlabelled data set by organizing data into homogeneous groups where the within-group dissimilarity is minimized and the between-group dissimilarity is maximized. Clustering is necessary when no labelled data are available regardless of whether the data are binary, categorical, numerical, interval, ordinal, relational, textual, spatial, temporal, spatio-temporal, image, multimedia, or mixtures of the above data types. Data are called static if all their feature values do not change with time, or the change negligible. The most of clustering analyses has been performed on static data. Clustering methods developed for handing this type of data can be classified into five major categories: partitioning methods, hierarchical methods, density based methods, grid-based methods, and model-based methods. Furthermore, the most relevant key points in the literature are: selecting a suitable clustering criterion, computational issues (identifying a sensible search strategy for the latent allocations, choosing sensible starting values) and selecting the number of clusters among others.

Just like static data clustering, time series clustering requires a clustering algorithm or procedure to form clusters given a set of unlabelled data objects and the choice of clustering algorithm depends both on the type of data available and on the particular purpose and application. As far as

time series data are concerned, distinctions can be made as to whether the data are discrete-valued or real-valued, uniformly or non-uniformly sampled, univariate or multivariate, and whether data series are of equal or unequal length.

From a mere multidisciplinary view point to statistics and statistical learning literature, a further distinctions can be enlightened. The linear, stationary paradigm which yields a mathematically elegant and powerful framework for analyzing and interpreting time series data, may not be always suitable for modeling more complicated time series. While methods have been developed for nonlinear and/or nonstationary time series, the literature on classification and clustering nonlinear, nonstationary series is relatively more recent. Furthermore, while statistic tests may be seen as similarity or distance metrics, a key point in the classical clustering literature, the distinguishing characteristic of a time series may be seen as the definition of the clustering problem itself. For instance the notion of parallelism may be a substitution for the membership criterion. Clustering techniques have been proposed for both time and frequency domain. The main literature concentrates on multivariate approach and does not often refer to the High-Dimensional frame.

1.1 General settings

Cluster Analysis and the time series represent the main ingredients composing the aim of this thesis. A multidisciplinary and general overview of this key concepts will be presented.

1.1.1 Cluster and Clustering

Cluster analysis or Clustering is an important tool in a variety of scientific areas including pattern recognition, information retrieval, micro-arrays and data mining. In general, a family of exploratory data analysis methods can be used to discover structures in data. These methods aim to obtain a reduced representation of the initial data and, like principal components analysis, factor analysis or multidimensional scaling, are one form of data reduction. The aim of cluster analysis is the organization of the set into homogenous classes or natural classes, in a way which ensures that objects within a class are similar to one another.

Sometimes some confusion about the use of the terms classification and Clustering occurs. Classification, whose task is to assign objects to classes or groups on the basis of measurements of the objects, is more general and can be divided into Supervised, Semisupervised and Unsupervised.

Supervised classification (or discrimination) seeks to create a classifier for the classification of future observations, starting from a set of labeled objects (a training or learning set). More precisely, data are composed of n individuals $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ belonging to a space $\mathcal{X}$ of dimension p, and also of an associated partition in K groups $G_1, \ldots, G_K$. This partition is denoted by $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$, where $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})'$ is a vector of $\{0,1\}^K$ such that $z_{ik} = 1$ if individual $\mathbf{x}_i$ belongs to the kth group G_k, and $z_{ik} = 0$ otherwise $(i = 1, \ldots, n; k = 1, \ldots, K)$. The data set is thus composed of all pairs $D = (\mathbf{x}, \mathbf{z}) = ((\mathbf{x}_1, \mathbf{z}_1), \ldots, (\mathbf{x}_n, \mathbf{z}_n))$, generally denoted as the learning data set. The aim is to estimate the group $\mathbf{z}_{n+1}$ of any new individual $\mathbf{x}_{n+1}$ in $\mathcal{X}$ for which the group would be unknown. This aim can be reformulated as the estimation of an allocation rule r from D and defined as follows:

$$\begin{aligned} r : &\mathcal{X} \to \{1, \ldots, K\} \\ &\mathbf{x}_{n+1} \mapsto r(\mathbf{x}_{n+1}) \end{aligned}$$

In *Semisupervised classification*, the aim is the same as in supervised classification but the data set is composed of n^l individuals $(0 \leq n^l \leq n)$ $\mathbf{x}^l = (\mathbf{x}_1, \ldots, \mathbf{x}_{n^l})$ for which group memberships $\mathbf{z}^l = (\mathbf{z}_1, \ldots, \mathbf{z}_{n^l})$ are known, whereas the $n^u = n - n^l$ remaining individuals $\mathbf{x}^u = (\mathbf{x}_{n^l+1}, \ldots, \mathbf{x}_n)$ have unknown labels $\mathbf{z}^u = (\mathbf{z}_{n^l+1}, \ldots, \mathbf{z}_n)$. Then, $D = (D_l, D_u)$ with $D_l = (\mathbf{x}^l, \mathbf{z}^l)$ and $D_u = \mathbf{x}^u$. The main idea is thus that the unlabelled individuals may be useful to learn an allocation rule. Usually, unlabelled individuals are expected to be more numerous than the labelled ones since the latter are clearly cheaper to obtain.

Finally, in *Unsupervised classification*, or Clustering, only individuals $\mathbf{x}$ are known and thus observed data are restricted to $D = \mathbf{x}$. The aim is focused to estimating the partition $\mathbf{z}$ related to $\mathbf{x}$ and not to estimate a partition of all the space $\mathcal{X}$. However, in some cases a partition of all the space $\mathcal{X}$ can be given as a simple by-product. In its more general, but also more difficult, version, the number of groups K is unknown as the number of individuals in a group, or Cluster.

In order to organize the objects of into homogenous clusters, the definition of homogenity is needed. Often similarity or dissimilarity measures

can be used to such task. Many clustering methods require the data to be presented as a set of proximities. This notion of proximity, which is a quantitative measure of closeness, is a general term for similarity, dissimilarity and distance: two objects are close when their dissimilarity or distance is small or their similarity large. Formally, a dissimilarity on the set Ω can be defined as a function d from $\Omega \times \Omega$ to $\mathbb{R}$ such that:

1. $d(x, y) > 0$ for all $x \neq y$ belonging to Ω
2. $d(x, x) = 0$ for all x belonging to Ω
3. $d(x, y) = d(y, x)$ for all x, y belonging to Ω.

A dissimilarity satisfying the triangle inequality

$$d(x, z) \leq d(x, y) + d(y, z) \quad \forall x, y, z \in \Omega$$

is a distance. Sometimes these proximities are the form in which the data naturally occur. In most clustering problems, however, each of the objects under investigation will be described by a set of variables or attributes. The first step in clustering, possibly the most important, is to define these proximities. Different kinds of definitions depending on the type of variables (continuous, binary, categorical or ordinal) are to be found in the literature. For instance, in the absence of information allowing the appropriate distance to be employed, the Euclidean distance between two vectors $\mathbf{x} = (x_1, \ldots, x_p)$ and $\mathbf{y} = (y_1, \ldots, y_p)$ in $\mathbb{R}^p$, defined by

$$d(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}|| = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$

is the most frequently used distance for continuous data. Moreover, before computing these proximities, it is often necessary to consider scaling or transforming the variables, since variables with large variances tend to impact the resulting clusters more than those with small variances. Other transformations can be used according to the nature of data.

1.1.2 Time series

A time series is a set of observations x_t, each one being recorded at a specified time t. A discrete-time series, x_t, is one which the set T_0 of times at which

observations are made is a discrete set. Continuous-time series, $x(t)$, is one which $T_0 = [0, 1]$ (Brockwell et al., 1991). Any observed data representing a physical phenomenon can be broadly classified as being either deterministic or nondeterministic (Bendat and Piersol, 2011). *Deterministic* data are those that can be described by an explicit mathematical relationship. A classical example consider a rigid body that is suspended from a fixed foundation by a linear spring, it defines the exact location of the body at any instant of time in the future. However, there are many other physical phenomena that produce data that are *nondeterministic*. For example, the height of waves in a confused sea, the acoustic pressures generated by air rushing through a pipe, and the electrical output of a noise generator represent data that cannot be described by explicit mathematical relationships. There is no way to predict an exact value at a future instant of time. These data are random in character and must be described in terms of probability statements and statistical averages rather than by explicit equations.

While Deterministic data can be further divided into Periodic and Nonperiodic, Nondeterministic (or stochastic) data can be classified into Stationary and Nonstationary (Brockwell et al., 1991). In general, for all the aforementioned categories, time series may present different characteristics.

- Seasonality is a periodical pattern observed for a time series. It is the effects of seasons such as months or fiscal year on the volatility and the volume traded within a period of time.

- Cycle is a dynamic pattern observed over a period of time (e.g., year). For instance, it is expected to observe some cyclic behaviour during harvesting time (e.g., cotton harvesting time).

- Trend is a long-term movement in a given time series without considering time or some other external influential factors.

- Unpredictable components, often calculated or retrieved after trend-cycle and seasonal components are removed from the time series. The remaining parts are unpredictable, since it only represents non-cyclic and the characteristics that are unique to the underlying time series.

Moreover, regardless the categories, due to the collection of several information, time series are characterized by High-Dimensionality. Given a p-dimensional time series $\{\mathbf{x}_t,\ t \in T_0\}$, p grows as the sample size $t \to \infty$. In

particular the order at which p grows (as function of t) characterizes the distinction between High-Dimensional, $p = O(t^c)$, or Ultra-High-Dimensional time series $\log(p) = O(t^c)$, where $c \in (0, 1)$. The serial or temporal dependence gives additional complications for analysis purposes based on this type of time series. In particular, Clustering High-dimensional Time series has shown increasing interest in recent years as a means to undertake further analysis.

1.2 The Time series Clustering Problem

Time series clustering has been shown effective in extracting useful information from time series data in various application domains expecially from a data mining point of view (Fu, 2011). Aghabozorgi et al. (2015) give a wide multidisciplinary overview of time series clustering procedures.

Considering time series as discrete objects, conventional clustering procedures can be used to cluster a set of individual time series with respect to their similarity such that similar time series are grouped into the same cluster. From this perspective time series clustering techniques have been developed, most of them critically depend on the choice of distance (i.e., similarity) measure. In general, there are three different approaches to cluster time series (Liao, 2005).

1. In *Shape-based approach*, clustering is performed based on the shape similarity, where shapes of two time series are matched using a nonlinear stretching and contracting of the time axes. An appropriate distance measure, specifically adapted for time series, and a convectional clustering methods are used in the shape-based clustering. An example is that proposed by Paparrizos and Gravano (2015) that present a novel algorithm for time series clustering called k-Shape. k-Shape relies on a scalable iterative refinement procedure, which creates homogeneous and well-separated clusters. The algorithm uses a normalized version of the cross-correlation measure in order to consider the shapes of time series while comparing them. Based on the properties of that distance measure, a method to compute cluster centroids is developed, which are used in every iteration to update the assignment of time series to clusters.

2. In *Feature-based clustering*, raw time series are transformed into the

feature vector of lower dimension (i.e., for each time series a fixed-length and an equal-length feature vector is created, usually a set of statistical characteristics). Then, a standard measure and a conventional clustering algorithm are applied on the lower dimension feature vector. The extracted features are usually application dependent which implies that one set of features that are useful for one application might not be relevant and useful for another one. Guijo-Rubio et al. (2020) propose a novel technique of time series clustering based on two clustering stages. In a first step, a least squares polynomial segmentation procedure is applied to each time series, which is based on a growing window technique that returns different-length segments. Then, all the segments are projected into same dimensional space, based on the coefficients of the model that approximates the segment and a set of statistical features. After mapping, a first hierarchical clustering phase is applied to all mapped segments, returning groups of segments for each time series. These clusters are used to represent all time series in the same dimensional space, after defining another specific mapping process. In a second and final clustering stage, all the time series objects are grouped.

3. *Model-based clustering* assumes a model for each cluster and attempts to fit the data into the assumed model. Then, each raw time series data is transformed into either model parameters (one model for each time series) or into a mixture of underlying probability distributions. An example is proposed by McDowell et al. (2018) which address the problem of clustering gene expression time series data uncertainty with infinite mixture models using a Dirichlet process (DP) prior. This Bayesian nonparametric approach is used in the Infinite Gaussian Mixture Model (GIMM). One of the major problems of the model-based approaches is the scalability problems, and its performance deteriorated when the clusters are very similar (Mitsa, 2010).

There are also methods which combine the aforementioned approaches. Asadi et al. (2016) employed Hidden Markov models (HMMs) for modeling and analysis of sequence data. Besides, ensemble methods, which employ multiple models to obtain the target model, revealed good performances in experiments. All these facts are a high level of motivation to employ HMM ensembles in the task of classification and clustering of time series data. The

hybrid approach derives benefits from both similarity-based and model-based methods.

Essentially, depending on the application, each of the aforementioned time series clustering approaches may use some or all of five major components: 1) time series representation, 2) distance measurement (or similarity), 3) the clustering algorithm, 4) prototype definition and 5) clusters' evaluation. The reason of having each component is as follows: the time series representation is usually used to fit data in memory. Afterwards, a clustering algorithm is performed on the data using a similarity (distance) measure, and as a result a prototype is created which shows a summarization of the time series. Finally, the created clusters are evaluated using different criteria.

1.2.1 Time series Representation Methods

The first major component of the time series clustering is time series representation. The new representation transforms the time series to another space such that if two time series are similar in the original space, their representations are also similar too. Choosing an appropriate time series representation method plays a significant role in the efficiency and accuracy of the clustering. Ding et al. (2008) have provided a comprehensive study on eight different representation methods which are performed on 38 datasets.

In particular, Model-based approaches give a special kind of time series representation methods that are used to represent a time series in a stochastic way such as HMM, statistical models and Auto-Regressive Moving Average (ARMA). In literature, a common distinction between models is that of Parametric and Nonparametric, more specification about the argument will be given in Section 1.3.

1.2.2 Time series Similarity/Distance Measures

Time series clustering are highly dependent on the choice of similarity and distance metric. An appropriate choice for similarity/distance extremely relies on representation methods, the length, the characteristic and the objective of clustering of the time series. In general, there are three *objectives*:

1. Finding Similar time series in Time, in this approach, similar time series are discovered on each time step. Euclidean distance and correlation based distances are appropriate distance measures for this method.

However, these distance measures are calculated using raw time series which is extremely expensive. Hence, the calculation is performed on transformed time series such as Piece-wise Aggregate Approximation (PAA), wavelets, or Fourier transformation.

2. Finding Similar time series in Shape, similar time series are identified according to similar shape features regardless of time points. To do so, similar trends occurring at different time or similar pattern of changes in data are captured. Elastic methods, such as Dynamic Time Warping (DTW), are used to measure distance for this approach. Note that, similarity in time is a special case of similarity in shape.

3. Finding Similar time series in Change, also known as structural similarity, in this approach the time series data is first modelled using modeling methods such as Hidden Markov Models or ARMA process. Then, similarity metric is measured based on global feature extracted from the obtained models. This is an appropriate approach for long time series, and usually may not be effective for short or modest time series.

Then, depending on the objective and on the length of time series, the distance measure can be roughly classified into 4 categories: Shape, Feature, Complexity and Model based measure. For each category, a further distinction may be extended to Elastic or Non-elastic measure if comparison between different length time series is allowed. Montero et al. (2014) provide a detailed summary on the most used ones.

Shape-based similarity measure

Shape-based similarity measure is usually used to find the similar time series in time and shape. It is a group of measures which are proper for short time series and do not take into account the stochastic properties of the time series. The most used is the Minkowski distance, a non-elastic measure with the Euclidean one as particular case. Another example is the Frechet distance (Fréchet, 1906) which measures the proximity between continuous curves. Unlike the Minkowski distance, it does not just treat the series as two points sets, but it takes into account the ordering of the observations, then can be computed on series of different length.

More recent Shape-based measures is the Minimal Variance Matching (MVM) (Latecki et al., 2005), which represents a distance (more appropriately an algorithm) for elastic matching of two time series of different lengths. It outperforms both Dynamic Time Warping distance (DTW) (Sakoe and Chiba, 1978) and Longest Common Sub-Sequence (LCSS) (Das et al., 1997) in the sense that it computes the distance value between two time series directly based on the distances of corresponding elements, just as DTW does, and it allows the query sequence to match to only a subsequence of the target sequence, just as LCSS does. Considering each time series as a piecewise linear function, the Short time series distance (STS), (Möller-Levet et al., 2003) can be defined as the sum of the squared differences of the slopes in two time series being compared. To remove the effect of scale, the standardization of the series is recommended.

The Adaptive dissimilarity index, (Chouakria and Nagabhushan, 2007), introduces a dissimilarity measure addressed to cover both conventional measures for the proximity on observations and temporal correlation for the behaviour proximity estimation. The proximity between the dynamic behaviours of the series is evaluated by means of the first order temporal correlation coefficient, while an adaptive tuning function automatically modulates a conventional raw-data distance. The modulating function should work increasing (decreasing) the weight of the dissimilarity between observations as the temporal correlation decreases from 0 to -1 (increases from 0 to $+1$).

Feature-based similarity measure

Feature-based similarity measures are proper for long time series and usually are represented by time series' statistics or coefficients that come from a previous time series' transformation. Though most feature extraction methods are generic in nature, the extracted features are usually application dependent. That is, one set of features that work well on one application might not be relevant to another. Some studies even take another feature selection step to further reduce the number of feature dimensions after feature extraction.

An example is represented by the Pearson's correlation coefficient (cc) and related distances, such as the two cross-correlation-based distances used

by Golay et al. (1998) in the fuzzy c-means algorithm

$$d_{cc}^1 = \left(\frac{1-cc}{1+cc}\right)^{\beta} \tag{1.1}$$

$$d_{cc}^2 = 2(1-cc) \tag{1.2}$$

where $\beta > 0$ allows to regularize the decreasing of d_{cc}^1 when $cc = -1$.

Autocorrelation-based distances, instead, allow to take into acoount the dependence over time. Let $\hat{\rho}_{X_t} = (\hat{\rho}_{1,X_t}, \ldots, \hat{\rho}_{L,X_t})'$ and $\hat{\rho}_{Y_t} = (\hat{\rho}_{1,Y_t}, \ldots, \hat{\rho}_{L,Y_t})'$ be the estimated autocorrelation vectors of X_t and Y_t respectively, for some L such that $\hat{\rho}_{i,X_t} \approx 0$ and $\hat{\rho}_{i,Y_t} \approx 0$ for $i > L$. Galeano and Peña (2001) define a distance between $\mathbf{X}_t$ and $\mathbf{Y}_t$ as follows.

$$d_{ACF}(\mathbf{X}_t, \mathbf{X}_t) = \sqrt{(\hat{\rho}_{X_t} - \hat{\rho}_{Y_t})'\Omega(\hat{\rho}_{X_t} - \hat{\rho}_{Y_t})} \tag{1.3}$$

where Ω is a matrix of weights.

Periodogram-based distances. Let $I_{X_T}(\lambda_k) = T^{-1}|\sum_{t=1}^{T} X_t e^{-i\lambda_k t}|^2$ and $I_{Y_T}(\lambda_k) = T^{-1}|\sum_{t=1}^{T} Y_t e^{-i\lambda_k t}|^2$ be the periodograms of X_t and Y_t, respectively, at frequencies $\lambda_k = 2\pi k/T$, $k = 1, \ldots, n$, with $n = [(T-1)/2]$ (in this case $[x]$ denotes the integer part of x). Three dissimilarity measures based on periodograms were analyzed by Caiado et al. (2006). More precisely, all of them use the Euclidean distance directly, on the normalized version or on the log normalized version of the periodogram respectively. De Lucas (2010) considers a distance measure based on the cumulative versions of the periodograms, i.e., the integrated periodograms. The author argues that the approaches based on the integrated periodogram present several advantages over the ones based on periodograms. In particular, the periodogram is an asymptotically unbiased but inconsistent estimator of the spectral density while the integrated periodogram is a consistent estimator of the spectral distribution; the spectral distribution always exists, but the spectral density exists only under absolutely continuous distributions.

Nonparametric Spectral estimators distances. Kakizawa et al. (1998) proposed a general spectral disparity measure between two series given by

$$d_W(X_T, Y_T) = \frac{1}{4\pi}\int_{-\pi}^{\pi} W\left(\frac{f_{X_T}(\lambda)}{f_{Y_T}(\lambda)}\right) d\lambda, \tag{1.4}$$

where f_{X_T} and f_{Y_T} denote the spectral densities of X_T and Y_T, respectively, and $W(\cdot)$ is a divergence function satisfying appropriate regular conditions to ensure that d_W has the quasi-distance property. If, for example,

$W(x) = \log(\alpha x + (1-\alpha)) - \alpha log x$, with $0 < \alpha < 1$, then d_W corresponds to the limiting spectral approximation of the Chernoff information in the time domain (Shumway and Unger (1974)). d_W is not a distance, it is not symmetric and does not satisfy the triangle inequality. In order to obtain the distance property, the divergence function can be modified by setting $\tilde{W}(x) = W(x) + W(x^{-1})$.

In practice, the spectra f_{X_T} and f_{Y_T} are unknown and must be previously estimated. Vilar and Pértega (2004) studied the asymptotic properties of d_W when f_{X_T} and f_{Y_T} are replaced by nonparametric estimators constructed via local linear regression. These approximations can be done in three different ways (Fan and Kreutzberger (1998)), thus resulting three different versions of the d_W dissimilarity measure. Specifically,

- $d_{W(DLS)}$, when the spectra are replaced by local lineal smoothers of the periodograms, obtained via least squares.
- $d_{W(LS)}$, when the spectra are replaced by the exponential transformation of local linear smoothers of the log-periodograms, obtained via least squares.
- $d_{W(LK)}$, when the spectra are replaced by the exponential transformation of local linear smoothers of the log-periodograms, here obtained by using the maximum local likelihood criterion.

In particular, the default value of the bandwidth is an automatic plug-in selector specifically designed for local linear Gaussian kernel regression (see Ruppert et al. (1995))

Two alternative nonparametric spectral dissimilarity measures introduced by Díaz and Vilar (2010). The first alternative comes from the generalized likelihood ratio test approach introduced by Fan and Zhang (2004) to check whether the density of an observed time series belongs to a parametric family while the second one evaluates the integrated squared differences between nonparametric estimators of the log-spectra.

The Discrete wavelet transform (DWT) is a useful feature extraction technique often used to measure dissimilarity between time series. DWT performs a scale-wise decomposing of the time series in such a way that most of the energy of the time series can be represented by only a few coefficients. The basic idea is to replace the original series by their wavelet approximation coefficients in an appropriate scale, and then to measure the dissimilarity between the wavelet approximations. A detailed description of wavelet methods

for time series analysis can be seen in Percival and Walden (2000). There is indeed a key point when using the DWT technique for clustering: the choice of an appropriate scale to obtain an accurate clustering. The algorithm was proposed by Zhang et al. (2006) and it is aimed to select the scale by levering two conflicting requirements: an efficient reduction of the dimensionality and preserving as much information from the original data as possible.

Complexity-based similarity measure

Complexity-based similarity, suitable for short and long time series, represents a group of dissimilarity measures based on comparing levels of complexity of time series. The similarity of two time series does not rely on specific serial features or the knowledge of underlying models, but on measuring the level of shared information by both time series. The mutual information between two series can be formally established using the Kolmogorov complexity concept, a measure of randomness of strings based on their information content, proposed by Kolmogorov in 1965 to quantify the randomness of strings and other objects in an objective and absolute manner. The Kolmogorov complexity $K(\mathbf{x})$ of a string $\mathbf{x}$ is defined as the length of the shortest program capable of producing $\mathbf{x}$ on a universal computer (i.e., a Turing machine). Intuitively, $K(\mathbf{x})$ is the minimal quantity of information required to generate $\mathbf{x}$ by an algorithm. In practice measure cannot be computed and must be approximated. The most common Complexity-measure is the Compression-based Dissimilarity Measures (CDM) Keogh et al. (2007).

Under many dissimilarity measures, pairs of time series with high levels of complexity frequently tend to be further apart than pairs of simple series. This way, complex series are incorrectly assigned to classes with less complexity. In order to mitigate this effect, Batista et al. (2011) propose to use information about complexity difference between two series as a correction factor for existing dissimilarity measures, such as a Complexity-invariant dissimilarity measure.

Permutation distribution clustering (PDC) represents an alternative complexity-based approach to clustering time series. Dissimilarity between series is described in terms of divergence between permutation distributions of order patterns in m-embedding of the original series. Specifically, given X_t , an m-dimensional embedding is constructed by considering

$$\mathcal{X}'_m = \{X'_m = (X_t, \ldots, X_{t+m}), \quad t = 1, \ldots, T - m\}$$

Then, for each $X'_m \in \mathcal{X}'_m$, permutation $\Pi(X'_m)$ obtained by sorting X'_m in ascending order (so-called *codeword* of X'_m) is recorded, and the distribution of these permutations on $\mathcal{X}'_m$, $P(X_t)$ (so-called *codebook* of X_t), is used to characterize the complexity of X_t. Furthermore, dissimilarity between two time series X_t and Y_t is measured in terms of the dissimilarity between their codebooks $P(X_t)$ and $P(Y_t)$, respectively. Brandmaier (2011) establishes this dissimilarity as the α-divergence between codebooks. The α-divergence concept (Amari, 2007) generalizes the Kullback-Leibler divergence and the parameter α can be chosen to obtain a symmetric divergence.

Model-based similarity measure

Model-based similarity measure, for long time series. Model-based dissimilarity measures assume that the underlying models are generated from specific model or mixture of distribution structures. The main approach in the literature is to assume that the generating processes of $\mathbf{X}_t$ and $\mathbf{Y}_t$ follow invertible ARIMA models. In such a case, the idea is fitting an ARIMA model to each series and then measuring the dissimilarity between the fitted models. First step requires estimating the structure and the parameters of ARIMA models. The structure is either assumed to be given or automatically estimated using, for example, the Akaike's information criterion (AIC) or the Schawartz's Bayesian information criterion (BIC). The parameter values are commonly fitted using generalized least squares estimators. Some of the most relevant dissimilarity measures derived will be briefly described.

The Piccolo-Distance (Piccolo (1990)) defines a dissimilarity measure in the class of invertible ARIMA processes as the Euclidean distance between the AR(∞) operators approximating the corresponding ARIMA structures. If the series are non-stationary, differencing is carried out to make them stationary, and if the series possess seasonality, then it should be removed before further analysis. Then, a definite criterion such as AIC or BIC is used to truncated AR(∞) models of orders k_1 and k_2 that approximate the generating processes of $\mathbf{X}_t$ and $\mathbf{Y}_t$, respectively. This approach allows to overcome the problem of obtaining ad-hoc ARMA approximations for each of the series subjected to clustering. If $\hat{\Pi}_{X_t} = (\hat{\pi}_{1,X_t}, \ldots, \hat{\pi}_{k_1,X_t})'$ and $\hat{\Pi}_{Y_t} = (\hat{\pi}_{1,Y_t}, \ldots, \hat{\pi}_{k_2,Y_t})'$ denote the vectors of AR(k_1) and AR(k_2) parameter

estimations for $\mathbf{X}_t$ and $\mathbf{Y}_t$, respectively, then the distance takes the form

$$d_{PIC}(\mathbf{X}_t, \mathbf{Y}_t) = \sqrt{\sum_{j=1}^{k} \left(\hat{\pi}'_{j,X_t} - \hat{\pi}'_{j,Y_t}\right)^2}, \tag{1.5}$$

where $k = \max(k_1, k_2)$, $\hat{\pi}'_{j,X_t} = \hat{\pi}_{j,X_t}$ if $j \leq k_1$, and $\hat{\pi}'_{j,X_t} = 0$ otherwise, and analogously $\hat{\pi}'_{j,Y_t} = \hat{\pi}_{j,Y_t}$ if $j \leq k_2$, and $\hat{\pi}'_{j,Y_t} = 0$ otherwise. Besides satisfying the properties of a distance (non-negativity, symmetry and triangularity), d_{PIC} always exists for any invertible ARIMA process since $\sum \pi_j$, $\sum ||\pi_j||$ and $\sum \pi_j^2$ are well defined quantities.

The Piccolo-distance does not take into account the variance of the white noise processes associated with the observed series, while the Maharaj-Distance (Maharaj, 1996) involves these variances in its definition. For the class of invertible and stationary ARMA processes, two discrepancy measures based on hypotheses testing to determine whether or not two time series have significantly different generating processes are defined. The first of these metrics is given by the test statistic

$$d_{MAH}(\mathbf{X}_t, \mathbf{Y}_t) = \left(\hat{\Pi}'_{X_t} - \hat{\Pi}'_{Y_t}\right)^T \hat{\mathbf{V}}^{-1} \left(\hat{\Pi}'_{X_t} - \hat{\Pi}'_{Y_t}\right) \tag{1.6}$$

where are the AR(k) parameter estimations of $\mathbf{X}_t$ and $\mathbf{Y}_t$, respectively, with k selected as in the Piccolo-distance, and $\hat{\mathbf{V}}$ is an estimator of $\mathbf{V} = 1/T(\sigma^2_{X_t}\mathbf{R}^{-1}_{X_t} + \sigma^2_{Y_t}\mathbf{R}^{-1}_{Y_t})$, with $\sigma^2_{X_t}$ and $\sigma^2_{Y_t}$ denoting the variances of the white noise processes associated with $\mathbf{X}_t$ and $\mathbf{Y}_t$, and $\mathbf{R}_{X_t}$ and $\mathbf{R}_{Y_t}$ the sample $k \times k$ covariance matrices of both series. d_{MAH} is asymptotically χ^2 distributed under the null hypothesis of equality of generating processes, i.e., by assuming that $\hat{\Pi}'_{X_t} = \hat{\Pi}'_{Y_t}$. Therefore, the dissimilarity between $\hat{\Pi}'_{X_t}$ and $\hat{\Pi}'_{Y_t}$ can also be measured through the associated p value, i.e., by considering

$$d_{MAHp}(\mathbf{X}_t, \mathbf{Y}_t) = P(\chi^2_k > d_{MAH}(\mathbf{X}_t, \mathbf{Y}_t)) \tag{1.7}$$

Both the test statistic d_{MAH} and the associated p value d_{MAHp} satisfy the properties of non-negativity and symmetry so that any of them can be used as dissimilarity measure between $\mathbf{X}_t$ and $\mathbf{Y}_t$.

Measures d_{MAH} and d_{MAHp} come from a hypothesis testing procedure designed to compare two independent time series. To overcome this limitation, Maharaj (2000) introduced a new testing procedure that can be applied to time series that are not necessarily independent. In this case, a pooled model

including collectively the models fitted to $\mathbf{X}_t$ and $\mathbf{Y}_t$ is considered, and the combined vector of $2k$ AR parameters $\Pi = (\Pi_{X_t}, \Pi_{Y_t})'$ is estimated by using generalized least squares. Assuming that the two models are correlated at the same points in time but uncorrelated across observations, the proposed test statistic (say d_{MAHext}) is also asymptotically distributed as χ^2 with k degrees of freedom. As before, a dissimilarity measure (say $d_{MAHextp}$) based on the p values associated with this new test can be constructed.

Kalpakis et al. (2001) propose the Cepstral-based distance which uses the linear predictive coding (LPC) cepstrum for clustering $ARIMA$ time series. The cepstrum is defined as the inverse Fourier transform of the short-time logarithmic amplitude spectrum. The cepstrum constructed by using the autoregression coefficients from linear model of the signal is referred to as the LPC Cepstrum, since it is derived from the linear predictive coding of the signal. Only a few LPC cepstral coefficients retains high amount of information on the underlying $ARIMA$ model. Consider a time series X_t following an $AR(p)$ structure, the LPC cepstral coefficients can be derived from the autoregressive coefficients ϕ_r as follows:

$$\psi_h = \begin{cases} \phi_1 & \text{if } h = 1 \\ \phi_h + \sum_{m=1}^{h-1}(\phi_m - \psi_{h-m}) & \text{if } 1 < h \leq p \\ \sum_{m=1}^{p}(1 - \frac{m}{h})\phi_m \psi_{h-m} & \text{if } p < h \end{cases}$$

In order to measure the distance between two time series X_t and Y_t , the Euclidean distance between their corresponding estimated LPC cepstral coefficients is considered

$$d_{LPC,Cep}(X_t, Y_t) = \sqrt{\sum_{i=1}^{T} \left(\psi_{X_t,i} - \psi_{Y_t,i}\right)^2} \tag{1.8}$$

Originally developed by Kumar et al. (2002) in their study of clustering seasonality patterns. They defined the similarity/distance between two seasonalities, A_i and A_j , as the probability of accepting/rejecting the null hypothesis $H_0 : A_i \sim A_j$. Assuming A_i and A_j , each comprised T independent samples drawn from Gaussian distributions with means x_{it} and x_{jt} and standard deviations σ_{it} and σ_{jt} , respectively, the statistic

$$\sum_{t=1}^{T} \frac{(x_{it} - x_{jt})^2}{\sigma_{it}^2 + \sigma_{jt}^2} \sim \chi^2_{T-1},$$

consequently,

$$d_{ij}(A_i, A_j) = \chi^2_{T-1} \left(\sum_{t=1}^{T} \frac{(x_{it} - x_{jt})^2}{\sigma_{it}^2 + \sigma_{jt}^2} \right) \tag{1.9}$$

The null hypothesis $A_i \sim A_j$ denotes $\mu_{it} = \mu_{jt}$ (i.e., the true seasonalities) for $t = 1, \ldots, T$.

The Kullback-Liebler distance, Kullback (1997). Let P_1 and P_2 be matrices of transition probabilities of two Markov chains (MCs) with s probabilities each and $p_{1_{ij}}$ and $p_{2_{ij}}$ be the $i \to j$ transition probability in P_1 and P_2. The asymmetric Kullback–Liebler distance of two probabilities is

$$d_{KL}(p_{1_i}, p_{2_i}) = \sum_{j=1}^{s} p_{1_{ij}} \log \left(\frac{p_{1_{ij}}}{p_{2_{ij}}} \right).$$

The symmetric version of Kullback–Liebler distance of two probabilities is

$$d_{KL_s}(p_{1_i}, p_{2_i}) = \frac{1}{2} \left[d_{KL}(p_{1_i}, p_{2_i}) + d_{KL}(p_{2_i}, p_{1_i}) \right] \tag{1.10}$$

The average distance between P_1 and P_2 is then

$$d_{KL_m}(P_1, P_2) = \frac{1}{s} \sum_{i=1}^{s} d_{KL_s}(p_{1_i}, p_{2_i}).$$

1.2.3 Time series Clustering Algorithms

According to the algorithm used, time series clustering can be classified into six categories (Halkidi et al., 2001)

1. Hierarchical time series Clustering. In this approach, a hierarchy of clusters is generated using either agglomerative (or bottom-up) or divisive (or top-down) approaches. In agglomerative methods, each item is considered as a cluster then appropriate clusters are merged together; whereas, in divisive approach all the items are included in one cluster, then the cluster is split into multiple clusters. Once the hierarchy is generated, it cannot adjust with any further changes. There fore, the quality of hierarchical clustering is weak and other clustering approaches are leveraged to remedy this issue.

2. Partitioning time series Clustering. In this approach, k groups of clusters are generated. One of the most common algorithms of partitioning clustering is called k-mean clustering, where k clusters are generated and the mean value of all the elements inside a cluster is considered as a cluster prototype.

3. Density-based time series Clustering. In this approach, a cluster is defined as a subspace of dense objects. One of the most common algorithms of density-based clustering is called DBSCAN (Ester et al., 1996), where a cluster is extended if its neighbors are dense.

4. Grid-based time series Clustering. In the grid-based clustering, the space is divided into a finite number of cells which are called grids, then clustering is done on the grids. STING (Wang et al., 1997) and WaveCluster (Sheikholeslami et al., 1998) are two common grid-based clustering algorithms.

5. Model-based time series Clustering. In this approach, a model is used for each cluster, then the best fit of data for the model is discovered. In model-based clustering approaches, either statistical approaches or neural network methods can be used. One example is Self-Organizing Maps (SOM) (Kohonen, 1990) which is a model-based clustering approach based on neural networks. While, another one is using Gaussian Mixture Models (GMMs) (see Biernacki (2017), Malsiner-Walli et al. (2016), Grün (2018)). The central assumption in model-based clustering is that the N time series from H hidden classes. Within each class, say h, all time series can be characterized by the same data generating mechanism (also called a clustering kernel), which is defined in terms of a probability distribution for the entire time series, depending on an unknown class-specific parameter $\boldsymbol{\vartheta}_h$. To address serial dependence among the observations for each subject, model-based clustering of time series data is often based on dynamic clustering kernels derived from first order Markov processes. The clustering kernel is formulated for the truncated time series $\mathbf{y}_i = \{y_{i,1}, \ldots, y_{i,T_i}\}$ conditional on the first observation y_{i0}, i.e.:

$$p(\mathbf{y}_i|\boldsymbol{\vartheta}_h) = \prod_{t=1}^{T_i} p(y_{it}|y_{i,t-1}, \mathbf{x}_{it}; \boldsymbol{\vartheta}_h), \tag{1.11}$$

where $\mathbf{x}_{it}$ are unknown values conditioning y_{it}. Other approaches dealing with model-based clustering ignore serial dependence and assume that the T_i observations are independent, given $\boldsymbol{\vartheta}_h$. In this case, the clustering kernel formulated for the entire time series $\mathbf{y}_i = \{y_{i0}, \ldots, y_{i,Ti}\}$ is called a locally independent clustering kernel. A class assignment index S_i taking a value in the set $\{1, \ldots, H\}$ is introduced for each time series $\mathbf{y}_i$ to indicate which class the time series belongs to

$$p(\mathbf{y}_i|S_i, \boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_H) = p(\mathbf{y}_i|\boldsymbol{\vartheta}_{S_i}) = \begin{cases} p(\mathbf{y}_i|\boldsymbol{\vartheta}_1), & \text{if } S_i = 1, \\ \vdots & \vdots \\ p(\mathbf{y}_i|\boldsymbol{\vartheta}_H), & \text{if } S_i = H. \end{cases} \tag{1.12}$$

In model-based clustering it is assumed that the class assignment indices $S_1, \ldots, S_N$ are random and independently distributed apriori, with prior class assignment distribution $Pr(S_i = h|\boldsymbol{\phi}) = p(h|\boldsymbol{\phi})$, $h = 1, \ldots, H$, where $\boldsymbol{\phi}$ is a model parameter. Then $Pr(S_1 = h_1, \ldots, S_N = h_N|\boldsymbol{\phi}) = p(h_1|\boldsymbol{\phi}) \ldots p(h_N|\boldsymbol{\phi})$. This leads to a representation of the marginal distribution $p(\mathbf{y}_i|\boldsymbol{\theta}_H)$, given $\boldsymbol{\theta}_H = (\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_H, \boldsymbol{\phi})$, in terms of a finite mixture distribution with H components:

$$p(\mathbf{y}_i|\boldsymbol{\theta}_H) = \sum_{h=1}^{H} p(h|\boldsymbol{\phi}) p(\mathbf{y}_i|\boldsymbol{\vartheta}_h). \tag{1.13}$$

It is assumed that the time series are independent for a given parameter value $\boldsymbol{\theta}_H$, meaning that the data generating mechanism is formulated by defining the data generating mechanism independently for each $\mathbf{y}_i$. Hence, to set up model-based clustering, two modeling assumptions have to be made, choosing the clustering kernel $p(\mathbf{y}_i|\boldsymbol{\vartheta}_h)$ and choosing the prior class assignment distribution $Pr(S_i = h|\boldsymbol{\phi})$, $h = 1, \ldots, H$. A example of Clustering kernels for real-valued time series observations y_{it} are typically based on dynamic regression models

$$y_{it} = \zeta_h + \delta_h y_{i,t-1} + \mathbf{x}_{it}\boldsymbol{\beta}_h + \sigma_h \varepsilon_{it}, \tag{1.14}$$

where ε_{it} is a random noise having zero expectation and variance equal to one. All parameters of the clustering kernel (1.14) are class-specific, however, it makes sense for specific applications to assume that certain parameters are the same across the classes.

6. Multi-step time series clustering. Multi-step time series clustering refers to a combination of methods (also called a hybrid method), which is used to improve the quality of cluster representation.

1.2.4 Time series Cluster Prototypes

One of the most significant and desired component in time series clustering is cluster prototype or cluster representative. Cluster prototype refers to the summarization of time series and is obtained using different methods. The quality of clustering is highly dependent on the quality of cluster prototypes. In literature there are three main methods to obtain the cluster prototype: Medoids, Averaging and Local search.

Medoids

Using Medoid as Prototype. Medoid is defined as a member of cluster such that its dissimilarity to all other members in the cluster is minimum. The concept of medoid is similar to that of centroids (which is used in K-mean clustering) and means. However, medoids are members of cluster; whereas, centroids and means are not. Medoids are useful when centroids or means cannot be defined as graphs.

Averaging

Using Averaging Prototype. In averaging prototype methods, mean of time series at each point is calculated. Averaging prototype is used when the time series have equal length and distance metric is a non-elastic metric (e.g., Euclidean distance). Sometimes computing average of time series is not trivial. For example, when the similarity between time series is based on the shape, then finding the average shape is challenging so in this case averaging prototype is evaded. In general, if the similarity of time series is based on elastic approaches (such as DTW or LCSS), averaging prototype is not trivial and is evaded. Two averaging methods using DTW and LCSS are:

- Shape averaging using DTW, in this approach, one method to define the prototype of a cluster is by combination of pairs of time series hierarchically or sequentially. For example, shape averaging using Dynamic Time Warping, until only one time series is left. The drawback of this

method is its dependency on the ordering of choosing pairs which results in different final prototypes. Another method is the approach mentioned by Abdulla et al. (2003), where the medoid is find as the initial guess, then all sequences are aligned by DTW to the medoid, and then the average time- series is computed. The resulting time series has the same length as the medoid, but the method is invariant to the order of processing sequences. In another study, the authors present a global averaging method for defining the prototypes (Petitjean et al., 2011). They use an averaging approach where the distance method for clustering or classification is DTW. However, its accuracy is dependent on the length of the initial averages equence and value of its coordinates.

- Shape averaging using LCSS, the longest common subsequence generally permits to make a summary of a set of sequences. This approach supports the elastic distances and unequal size time series. Usually, a fuzzy clustering approach for time series clustering is used, and the averaging method by LCSS as prototype is performed.

Local search

Using Local Search Prototype. In local search prototype, the medoid of cluster is computed, then warping paths techniques are used to calculate averaging prototype. Finally, for the obtained averaged prototype new warping paths are calculated.

1.2.5 Time series Clustering Evaluation Measures

Evaluating the extracted clusters is not a trivial task and has been extensively researched. Different clustering algorithms obtain different clusters and different clustering structures, thus evaluating clustering results is quite important, in order to evaluate clustering structures objectively and quantitatively. There are two different testing criteria: external criteria and internal criteria. External criteria uses class labels (also known as ground truth) for evaluating the assigned labels. Note that the ground truth is not used during the clustering algorithm. On the other hand, internal criteria evaluates the goodness of a clustering structure without respect to external information.

Internal metrics

Among the different internal criteria, the most important ones differ if the cluster representative is needed or not.

The Sum of Squared Error index (SSE) measures the compactness of a given clustering, independently of the distance to other clusters. Better clusterings have lower values of SSE.

$$SSE = \frac{1}{T} \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} d_E(\mathbf{x}, \bar{C}_i)^2$$

where $\bar{C}_i$ is the representative of the ith cluster C_i, k is the number of clusters and $d_E(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{T}(x_i - y_i)^2}$ is the Euclidean distance between the time series $\mathbf{x}$ and $\mathbf{y}$. While the Normalized SSE index (NSSE) looks for well-separated groups, maximizing the distance intra-clusters.

$$NSSE = \frac{SSE}{(T-1)! \sum_{i=1}^{k} \sum_{j=i+1}^{k} d_E(\bar{C}_i, \bar{C}_j)^2}$$

where $\bar{C}_i$ and $\bar{C}_j$ are the representatives of the ith and jth cluster respectively.

The Silhouette index (SI) combines both cohesion and separation, so it is based on the intra-cluster ($a(\mathbf{x}; C_i)$) and inter-cluster ($b(\mathbf{x}; C_i)$) distances respectively. This distances are given as follows:

$$a(\mathbf{x}; C_i) = \frac{1}{T_{C_i}} \sum_{\mathbf{y} \in C_i} d(\mathbf{x}, \mathbf{y})$$

$$b(\mathbf{x}; C_i) = \min_{C_l, l \neq i} \{a(\mathbf{x}; C_l)\}$$

where T_{C_i} is the cardinality of the cluster C_i and $d(\mathbf{x}, \mathbf{y})$ is a generic distance between $\mathbf{x}$ and $\mathbf{y}$ time series, as defined before. Finally, SI index is defined as:

$$SI = \frac{1}{T} \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} s(\mathbf{x}, C_i)$$

where $s(\mathbf{x}, C_i) = \frac{b(\mathbf{x}, C_i) - a(\mathbf{x}, C_i)}{\max\{b(\mathbf{x}, C_i), a(\mathbf{x}, C_i)\}}$.

For an extensive comparison of the other indices such as Calinski and Harabasz (CH), Davies-Bouldin (DB), Dunn index (DU) and COP index

(COP) see Arbelaitz et al. (2013). In particular, CH, SI, COP, DU and variants have to be maximised. Conversely, DB, SSE and NSSE have to be minimised. The most common measures in the literature are CH, DU and SSE. Furthermore, Information Criteria such as AIC (Akaike, 1974), BIC (Schwarz et al., 1978) and ICL (Biernacki et al., 2000) can be used as internal metrics if a Mixture of Gaussian distribution approach (Model-based) is used.

External metrics

There is not a compromise and universally accepted technique to evaluate clustering approaches, though there are many candidates which can be discounted for a variety of reasons. External indices measure the similarity between the cluster assignment and the ground truth, which has to be given as a form of evaluation but should not be used during the clustering. There are many metrics in the literature (Amigó et al., 2009).

One of the ways to measure the quality of a clustering solution is *cluster purity*. Purity is a simple and transparent evaluation measure. Considering $G = \{G_1, \ldots, G_M\}$ as ground truth clusters, and $C = \{C_1, \ldots, C_M\}$ as the clusters made by a clustering algorithm under evaluations, in order to compute the purity of cluster C with respect to G, each cluster is assigned to the class which is most frequent in the cluster, and then the accuracy of this assignment is measured by counting the number of correctly assigned objects and dividing by number of objects in the cluster. A bad clustering has purity value close to 0, and a perfect clustering has a purity of 1. However, high purity is easy to achieve when the number of clusters is large, in particular, purity is 1 if each objects gets its own cluster. Thus, one cannot only rely on purity as the quality measure.

Assuming a one to one mapping between clusters C and categories G, and relying on precision and recall concepts, other measures can be obtained. The most popular measures for cluster evaluation are the afore mentioned Purity, Inverse Purity and their harmonic mean such as the F-measure. This one is a well-established measure for assessing the quality of any given clustering solution with respect to ground truth. F-measure compares how closely each cluster matches a set of categories of ground truth.

$$F = \sum_i \frac{|G_i|}{N} \max_j \{F(G_i, C_j)\}$$

where

$$\begin{aligned} F(G_i, C_j) =& \frac{2 \times \text{Recall}(G_i, C_j) \times \text{Precision}(G_i, C_j)}{\text{Recall}(G_i, C_j) + \text{Precision}(G_i, C_j)} \\ \text{Recall}(G, C) =& \text{Precision}(C, G) \\ \text{Precision}(C_i, G_j) =& \frac{|C_i \cap G_j|}{|C_i|} \end{aligned}$$

Another approach to define evaluation metrics for clustering is considering statistics over *pairs of time series*. Let SS be the number of pairs, each belongs to one cluster in G and are clustered together in C; DS be the number of pairs that belong to one cluster in G, but are not clustered together in C; SD be the number of pairs that are clustered together in C, but are not belong to one cluster in G; DD be the number of pairs, each neither clustered together in C, nor belongs to the same cluster in G. The most used is the Rand index (RI), which measures the agreement between two partitions and shows how much clustering results are close to the ground truth.

$$\text{RI} = \frac{(SS + DD)}{SS + SD + DS + DD} \tag{1.15}$$

A drowback of RI is that it does not take a constant value (such as zero) for two random clustering. Hence, Hubert and Arabie (1985) suggest a corrected-for-chance version of the RI, such as the Adjusted Rand Index (ARI).

Other External metrics are Folkes and Mallow index (FM), (usually used for time series clustering in multimedia domain)and the Jaccard Score (J).

The Entropy of a cluster shows, instead, how dispersed classes are with a cluster (this should be low). Entropy is a function of the distribution of classes in the resulting clusters.

$$\text{Entropy} = -\sum_j \frac{n_j}{n} \sum_i p(i, j) \times \log_2 p(i, j)$$

being $p(i, j)$ the probability of finding an element from the category i in the cluster j, n_j is the number of time series in cluster j and n is the total number of time series in the distribution. Other metrics based on entropy have also been defined, for instance, the Normalized Mutual Information (NMI) (Studholme et al., 1999). High purity in the large number of clusters is a drawback of purity measure. In order to make trade-off between the quality

of the clustering against the number of clusters, NMI is utilized as quality measure. Moreover, NMI can be used to compare clustering approaches with different numbers of clusters, because this measure is normalized.

One of the most popular approaches for quality evaluation of clusters is external indices to find how good the finding cluster results are (Halkidi et al., 2001). However,it is not directly applicable in real-life unsupervised tasks, because the ground truth is not available in many cases. Therefore, in the case that ground truth is not available, internal index is used.

1.3 The Model-based Representation

As mentioned in Section 1.2.1, choosing an appropriate representation method can be considered as the key component which effects the efficiency and accuracy of the clustering solution. High-dimensionality and noise are characteristics of the most time series data, consequently, dimensionality reduction methods are used in time series clustering in order to address this issues and promote the performance.

From a statistical prospective, it is natural to model time series as stochastic processes and a further classification lies in the specification of the model. A stochastic process is a family of random variables $\{X_t,\ t \in T\}$ defined on a probability space $(\Omega, \mathfrak{F}, P)$ while its realizations (or sample-paths) are the functions $\{X_\cdot(\omega),\ \omega \in \Omega\}$ on T. Due the unpredictable nature of future observations, it is natural to suppose that each observation x_t, is a realized value of a certain random variable X_t. The time series $\{x_t,\ t \in T_o\}$ is then a realization of the family of random variables $\{X_t,\ t \in T_o\}$. These considerations suggest modeling the data as a realization (or part of a realization) of a stochastic process $\{X_t,\ t \in T\}$ where $T \supseteq T_o$ (Brockwell et al., 1991).

In literature it is common to distinguish the class of Parametric stochastic model from the class of Nonparametric one. There are infinitely many stochastic processes that can generate the same observed data, as the number of observations is always finite. However, some of these processes are more plausible and admit better interpretation than others. Without further constraints on the underlying process, it is impossible to identify the process from a finite number of observations. A popular approach is to confine the probability law to a specified family and then to select a member in that family that is most plausible. The former is called modeling and the latter is called estimation, or more generally statistical inference. When the form of

the probability laws in a family is specified except for some finite-dimensional defining parameters, such a model is referred to as a parametric model. When the defining parameters lie in a subset of an infinite dimensional space or the form of probability laws is not completely specified, such a model is often called a nonparametric model (Fan and Yao, 2008).

1.3.1 Parametric model specification

The class of Parametric models have been widely used to deal with both linear and nonlinear time series. While examples of nonlinear parametric models include, among others, the $ARCH$-modeling of fluctuating volatility of financial data and the threshold modeling of biological and economic data, the most popular linear time series models are the autoregressive moving average ($ARMA$) models. $ARMA$ models are frequently used to model linear dynamic structures, to depict linear relationships among lagged variables, and to serve as vehicles for linear forecasting (Fan and Yao, 2008). The process $\{X_t, t = 0, \pm 1, \pm 2, \dots\}$ is said to be an $ARMA(p, q)$ process if for every t,

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \tag{1.16}$$

where $Z_t \sim WN(0, \sigma^2)$. Moreover X_t is an $ARMA(p, q)$ process with mean μ if $X_t - \mu$ is an $ARMA(p, q)$ process (Brockwell et al., 1991). ARMA process can be easily extended to multivariate case. Let random vectors $\mathbf{X}_1, \dots, \mathbf{X}_T$ be drawn from a stationary process $\{\mathbf{X}_t\}_{t=-\infty}^{\infty}$, and $\mathbf{X} = [\mathbf{X}_1, \dots, \mathbf{X}_t, \dots, \mathbf{X}_T]' \in \mathbb{R}^{T \times d}$, where $\mathbf{X}_t = (x_1, \dots, x_d)' \in \mathbb{R}^d$ is a d-dimensional vector and each column of $\mathbf{X}$ is a one-dimensional time series with T samples. In particular, each $\mathbf{X}_t$ can be modeled by a vector ARMA model of order p and q ($VARMA(p, q)$)

$$\mathbf{X}_t - \mathbf{A}_1 \mathbf{X}_{t-1} - \cdots - \mathbf{A}_p \mathbf{X}_{t-p} = \mathbf{Z}_t + \mathbf{B}_1 \mathbf{Z}_{t-1} + \cdots + \mathbf{B}_q \mathbf{Z}_{t-q},$$

where $\mathbf{A}_1, \dots, \mathbf{A}_p, \mathbf{B}_1, \dots, \mathbf{B}_q$ are real $d \times d$ matrices and $\mathbf{Z}_t \sim WN(0, \mathbf{\Psi})$. Assuming $p = 1$ and $q = 0$, the latter equation can be rewritten as a first-order vector autoregressive model $VAR(1)$

$$\mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1} + \mathbf{Z}_t.$$

To secure the above process to be stationary, the transition matrix $\mathbf{A}$ must have bounded spectral norm, i.e., $||\mathbf{A}||_2 < 1$.

1.3.2 Nonparametric model specification

Many data in applications exhibit nonlinear features such as nonnormality, asymmetric cycles, bimodality, nonlinearity between lagged variables, and heteroscedasticity. They require nonlinear models to describe the law that generates the data. However, beyond the linear time series models, there are infinitely many nonlinear forms that can be explored. Then the number of parametric models that should be considered increases. A natural alternative is to use nonparametric methods. The most flexible model is the saturated (full) nonparametric model, which does not impose any particular form on autoregression functions

$$X_t = f\left(X_{t-1}, \ldots, X_{t-p}\right) + \sigma\left(X_{t-1}, \ldots, X_{t-p}\right)\varepsilon_t.$$

Where $f(\cdot)$ and $\sigma(\cdot)$ are unknown functions, and $\{\varepsilon_t\} \sim IID(0,1)$. Instead of imposing concrete forms on functions f and σ, some qualitative assumptions can be made, such as that the functions f and σ are smooth (Fan and Yao, 2008). A further generalization is the "classical decomposition" model with no periodic component,

$$Y_t = \mu_t + X_t \tag{1.17}$$

which allows to represent $\{Y_t\}$ as the sum of a slowly varying trend component μ_t and a zero-mean stationary stochastic component X_t (Fan and Yao, 2008).

1.4 Clustering High-dimensional time series

As mentioned so far, time series are characterized by High-Dimensionality if, given a p-dimensional time series $\{\mathbf{x}_t,\ t \in T_0\}$, p grows as the sample size $t \to \infty$. In particular the order at which p grows (as function of t) characterizes the distinction between High-Dimensional, $p = O(t^c)$, or Ultra-High-Dimensional time series $\log(p) = O(t^c)$, where $c \in (0,1)$. Furthermore, from a statistical prospective, it is natural to model time series as stochastic processes i.e. a family of random variables $\{X_t,\ t \in T\}$ defined on a probability space $(\Omega, \mathfrak{F}, P)$ and the common distinction is given by the class of Parametric stocastic model and the class of Nonparametric one. So, from now on, the focus will be on the clustering procedure for high-dimensional time series (HDTS) seen as realization of stochastic processes.

In this light, distance measurement, clustering algorithm, prototype definition and clusters' evaluation became the four HDTS clustering components. Each one may characterize a deeper difference among various work introduced so far. Examples can be formulated by means of the last two works considered.

In the context of stationary d-dimensional time series $\mathbf{X}_t$ where d is allowed to grow exponentially with t but under $\log d = o(t)$ and $\mathbf{X}_t$ follows a vector autoregressive model of order 1, Hong et al. (2017) propose a new *pairwise similarity measure* for *high-dimensional time series* called *Cross-Predictability* (CP) which represents the degree to which a future value in each time series is predicted by past values of the others. The setting is that of Ultra-high-dimensionality with time series represented as parametric autoregressive model. Under the further assumptions that $\mathbf{Z}_t \sim N(0, \mathbf{\Psi})$ is i.i.d. additive noise independent of $\mathbf{X}_t$, and $\mathbf{X}_t$ has zero mean and a covariance matrix $\mathbf{\Sigma}$, i.e., $\mathbf{X}_t \sim N(0, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = E[\mathbf{X}_{t-1}\mathbf{X}'_{t-1}]$ is the autocovariance matrix, the lag-1 autocovariance matrix $\mathbf{\Sigma}_1 = E[\mathbf{X}_{t-1}\mathbf{X}'_t]$ can be rewritten as

$$\mathbf{\Sigma}\mathbf{A}' = \mathbf{\Sigma}_1. \tag{1.18}$$

Furthermore, since $\{\mathbf{X}_t\}_{t=-\infty}^{\infty}$ is stationary, the covariance matrix $\mathbf{\Sigma}$ depends on $\mathbf{A}$ and $\mathbf{\Psi}$, i.e., $\mathbf{\Sigma} = \mathbf{A}'\mathbf{\Sigma}\mathbf{A} + \mathbf{\Psi}$. A non-zero entry $\mathbf{A}_{ij}$ implies that the jth time series is predictive for the ith time series, with the magnitude $|\mathbf{A}_{ij}|$ indicating how much the predictive power is. This is a measure of crosspredictive relationship between time series. Given the numbero of clusters, the clustering algorithm first estimates the cross-predictability among the time series, and then identifies the clustering structure based on the estimated relationship. Inspired by the relationship in (1.18), the main idea is to estimate $\mathbf{A}$ based on the relationship between $\mathbf{A}$ and the autocovariance and lag-1 autocovariance matrices. This motivates the following Dantzig selector type estimator (see Candes et al. (2007) and Han et al. (2015)),

$$\hat{\mathbf{A}} = arg\min_{\mathbf{A}} ||\mathbf{A}||_1 \quad \text{s.t. } ||\hat{\mathbf{\Sigma}}\mathbf{A}' - \hat{\mathbf{\Sigma}}_1||_{\infty,\infty} \leq \mu \tag{1.19}$$

where $||\mathbf{M}||_1 = \max_{||v||_1=1} ||\mathbf{M}v||_1$ and $||\mathbf{M}||_{\infty,\infty} = \max_{ij} |M_{ij}|$ are the matrix norms l_1 and l_∞ of the matrix $\mathbf{M}$ respectively, $\mu > 0$ is a tuning parameter, $\hat{\mathbf{\Sigma}} = \mathbf{X}'_{\mathcal{S}}\mathbf{X}_{\mathcal{S}}/(T-1)$, $\hat{\mathbf{\Sigma}}_1 = \mathbf{X}'_{\mathcal{S}}\mathbf{X}_{\mathcal{T}}/(T-1)$, with $\mathbf{X}_{\mathcal{S}} = [\mathbf{X}_1, \ldots, \mathbf{X}_{T-1}]' \in \mathbb{R}^{(T-1)\times d}$, $\mathbf{X}_{\mathcal{T}} = [\mathbf{X}_2, \ldots, \mathbf{X}_T]' \in \mathbb{R}^{(T-1)\times d}$. The above optimization problem can be decomposed into d independent sub-problems and solved individually

as follows:

$$\hat{\beta}_i = arg\min_{\beta_i} ||\beta_i||_1 \quad \text{s.t. } ||\hat{\mathbf{\Sigma}}\beta_i' - \hat{\gamma}_i||_{\infty,\infty} \leq \mu \tag{1.20}$$

where $||v||_1 = \sum_i |v_i|$ and $||v||_{\infty,\infty} = \max_i |v_i|$ are the vector norms l_1 and l_∞ of the vector v respectively, $\hat{\gamma}_i = (\hat{\mathbf{\Sigma}}_1)_{*i} = \mathbf{X}_{\mathcal{S}}'(\mathbf{X}_{\mathcal{T}})_{*i}/(T-1)$, i.e., $\hat{\gamma}_i$ is the ith column of $\hat{\mathbf{\Sigma}}_1$, and $\hat{\mathbf{A}} = [\hat{\beta}_1, \ldots, \hat{\beta}_d]' \in \mathbb{R}^{d\times d}$ with each $\hat{\beta}_i \in \mathbb{R}^d$. Therefore, the $\hat{\beta}_i$ in (1.20) is an estimation of the ith row of the transition matrix $\mathbf{A}$. Now, for each $\mu > 0$, there always exists a $\lambda > 0$ such that (1.20) is equivalent to the following regularized Dantzig selector type estimator:

$$\hat{\beta}_i = arg\min_{\beta_i} \lambda||\hat{\mathbf{\Sigma}}\beta_i' - \hat{\gamma}_i||_{\infty,\infty} + ||\beta_i||_1 \tag{1.21}$$

where λ is a regularization parameter to determine the sparsity of the estimation. After solving the problem in the last equation, an affinity matrix $\mathbf{W}$ based on $\hat{\mathbf{A}}$ is constructed by symmetrization, and compute the corresponding Laplacian to perform standard spectral clustering (Ng et al., 2002) to recover the clusters in the input time series. *Sparsity assumption* is then defined as only time series in the same cluster share significant CP (i.e. sparsity of the cross-predictability matrix assumption). The *cluster recovery* of the algorithm is proved under the assumptions, 1) the individual time series can be modelled by an Autoregressive model (AR), 2) the transition matrix for the Vector Autoregressive model (VAR) is block-diagonal.

On the other hand, Zhang (2013) considers the problem of clustering *high-dimensional time series* based on trend parallelism. The underling p-dimensional time series $\mathbf{Y}_t$ where $p = O(T^l)$, with $l < 1/4$, is modeled as a *nonparametric trend function with local stationary errors*, i.e. in (1.17) each component of $\mathbf{Y}_t$ is assumed to be a *nearly stationary processes* (Draghicescu et al., 2009). The setting is then that of high-dimensionality with time series represented nonparametrically. For each group where the parallelism holds, its representative trend function is estimated semiparametrically. Instead of pairwise dissimilarity measure, an in-group one is considered using an information criterion which takes into account for the number of clusters and the estimated common trend functions can be seen as cluster prototypes. More precisely, following Draghicescu et al. (2009), suppose to observe p (which can grow to infinity) time series $\{y_{k,i}\}_{i=1}^n$, $k = 1, \ldots, p$, according to the model

$$y_{k,i} = \mu_k(i/n) + e_{k,i}, \quad i = 1, \ldots, n, \tag{1.22}$$

where $\mu_k : [0,1] \to \mathbb{R}$ are unknown smooth trend functions, and $\{e_{k,i}\}_{i=1}^n$ are locally stationary zero mean error processes and such that

$$e_{k,i} = G(i/n, \mathcal{F}_{k,i}), \quad \mathcal{F}_{k,i} = \{\ldots, \epsilon_{k,i-1}, \epsilon_{k,i}\},$$

where $\epsilon_{l,j}, l, j \in \mathbb{Z}$, are independent and identically distributed (iid) random variables, and G is a measurable function. The scaling device i/n in (1.22) is useful in characterizing the smoothness and it is necessary for providing asymptotic justification for any nonparametric smoothing estimators. The objective in Zhang (2013) is to find a minimal number of nonoverlapping subgroups $\mathcal{S}_1 \cup \cdots \cup \mathcal{S}_Q = \{1, \ldots, p\}$ such that for each $q = 1, \ldots, Q$ and $k \in \mathcal{S}_q$,

$$\mu_k(t) = \mu_{\mathcal{S}_q}(t) + c_k, \quad t \in [0,1], \tag{1.23}$$

for some common trend function $\mu_{\mathcal{S}_q} : [0,1] \to \mathbb{R}$ and individual shifts $c_k \in \mathbb{R}$. Furthermore, to ensure the identifiability

$$\sum_{k \in \mathcal{S}_q} c_k = 0.$$

For each individual time series $\{y_{k,i}\}_{i=1}^n$, its trend function (along with its derivative) can be estimated nonparametrically by the local linear estimate (Fan and Gijbels, 1996). Using Epanechnikov kernel $K(\cdot)$, the closed form solution

$$\hat{\mu}_k(t) = \sum_{i=1}^n y_{k,i} w_i(t) \tag{1.24}$$

is obtained, where $w_i(t) = K\{(i/n - t)/b_n\}\{S_2(t) - (t - i/n)S_1(t)\}/\{S_2(t)S_0(t) - S_1^2(t)\}$ are the local linear weights, $S_j(t) = \sum_{i=1}^n (t - i/n)^j K\{(i/n - t)/b_n\}$ and b_n is the bandwidth selected using the generalized cross-validation (GCV) selector. Suppose that the parallelism assumption holds for the subgroup $\mathcal{S}_q$. Then its common trend function $\mu_{\mathcal{S}_q}(\cdot)$ as in (1.23) can be estimated by

$$\hat{\mu}_{\mathcal{S}_q}(t) = |\mathcal{S}_q|^{-1} \sum_{k \in \mathcal{S}_q} \hat{\mu}_k(t), \quad t \in [0,1], \tag{1.25}$$

where $|\mathcal{S}_q|$ is the cardinality of $\mathcal{S}_q$. By (1.23), the individual shifts c_k, $k \in \mathcal{S}_q$, can be estimated by

$$\hat{c}_k = n^{-1} \sum_{i=1}^n \left[\hat{\mu}_k(i/n) - \hat{\mu}_{\mathcal{S}_q}(i/n)\right]$$

Instead of pairwise similarity measure, an in-group one is defined by

$$RSS(\mathcal{S}_q) = \sum_{k \in \mathcal{S}_q} \sum_{i=1}^{n} \left[y_{k,i} - \hat{\mu}_{\mathcal{S}_q}(i/n) - \hat{c}_k \right]^2$$

Since $\hat{\mu}_{\mathcal{S}_q}(i/n) + \hat{c}_k$ is the semiparametric estimate of $E(y_{k,i})$, $k \in \mathcal{S}_q$, $i = 1, \ldots, n$, the above statistic provides the residual sum of squares for the subgroup $\mathcal{S}_q$ under the parallelism (1.23). Let $\mathcal{P} = \{\mathcal{S}_1, \ldots, \mathcal{S}_Q\}$ be the implied partition with cardinality $|\mathcal{P}| = Q$, and $RSS(\mathcal{P}) = \sum_{q=1}^{Q} RSS(\mathcal{S}_q)$ be the residual sum of squares across all subgroups. The Extended Bayesian Information criterion (EBIC) (Chen and Chen, 2008) is considered for clustering high-dimensional time series data, which takes the form

$$EBIC(\mathcal{P}) = (np) \log \left[RSS(\mathcal{P})/(np) \right] + \tau_n |\mathcal{P}|. \tag{1.26}$$

The criterion depends on a tuning parameter τ_n. If $\tau_n = \log n$, then (1.26) becomes the traditional BIC. Larger τ_n leads to stronger penalization on the number of clusters, and vice versa. The true partition $\mathcal{P}_0$ is estimated by minimizing (1.26). This work not only generates prototypes but provides also a special algorithm which make feasible the computation of the EBIC. The Adjusted Rand Index remains the reference for simulation studies for both works.

Chapter 2

The proposal

In the Chapter 1, a large part of the literature on the topic of clustering on time series was presented. Some of the aforementioned approaches can be used in the presence of stationary time series, while others consider time series whose nonstationarity is linked to the presence of a trend.

Time series trend composition is a very important topic in data analysis. Checking trend composition is the first step for a further statistical analysis conducted on a time series. A very common question about the trend concerns its existence and if it has a linear or a nonlinear composition. *The proposal that will be discussed in this thesis regards the classification of nonstationary time series, where the nonstationarity is given by the presence of a deterministic trend, before undertaking a cluster analysis. This is accomplished by looking at the first derivative of the trend in a context of high-dimensionality and without requiring a pre specified form for the trend.*

The idea is to classify the time series by checking the trend first derivative. If the trend is constant, then its first derivative is zero, if the trend is linear, then its first derivative is constant. If none of the previous happen, then the trend is of course nonlinear and then its first derivative will be not constant. In this way the time series can be divided into tree groups.

In the following, the first section gives a brief introduction to the most recent literature on clustering time series based on the use of a statistical test in the nonparametric trend estimation context; the second section introduces the general setting for the idea behind the proposed procedure together with the statistical instruments that will be used; finally, the third section concentrates on the explanation of the proposed procedure. More precisely, the procedure can be included in the category of "clustering of time series based

on features", since the trend composition can be considered as a feature of the time series and can therefore be used before conducting a further and more in-depth cluster analysis which can concern both stationary and non-stationary time series (where the nonstationarity is linked to the presence of the trend).

2.1 Recent works on testing for trends

Lyubchich and Gel (2016) propose a nonparametric test for synchronism of trends exhibited by multiple linear time series where the number of time series p can be large but fixed (i.e. $p < \infty$). The core idea of the approach is based on employing the local regression test statistic, which allows to detect possibly non-monotonic nonlinear trends. The finite sample performance of the new synchronism test statistic is enhanced by a nonparametric hybrid bootstrap approach.

They argue that Degras et al. (2011) and Zhang (2013) extend the Integated Squared Error (ISE) based approach of Vilar-Fernández et al. (2007) to a case of multiple time series with weakly dependent nonstationary errors and then their methods involve the selection of multiple nuisance parameters, such as bandwidth, level of smoothness, and window size for a longrun variance function. Those peculiarities often lead to inadequate performance, especially in samples of moderate size.

The previous criticism leads to the core idea of the approach of Lyubchich and Gel (2016) which is based on *generalizing the nonparametric local factor (lf) test statistic* (originally developed for detecting a trend in a single process), which allows to assess whether p weakly dependent time series exhibit a *joint nonmonotonic nonlinear trend that belongs to a prespecified parametric family of functions*. The test procedure employs: an artificial balanced one-way analysis of variance, where each distinct time point is viewed as a category, and an associated cell which includes all observations within a surrounding local window. Each cell is choosen using the data-driven heuristic m-out-of-n bootstrap selection algorithm of Bickel et al. (1997). The new trend synchronism statistic is asymptotically normally distributed. However, as convergence to the asymptotic distribution might be slow, the finite sample properties of the test statistic is enhanced by a hybrid bootstrap procedure. The proposed test yields noticeably more accurate estimates of the size of the test, compared with the test of Degras et al. (2011), especially when there

is either a small number of observed time series or when the autocorrelation structure of an individual time series is allowed to include negative terms, which is a frequent situation for economic and environmental studies.

Let observe p time series processes

$$Y_{it} = m_i(t/T) + \varepsilon_{it} \quad (i = 1, \ldots, p; t = 1, \ldots, T), \tag{2.1}$$

where $m_i(u)$ $(0 < u \leq 1)$ are unknown smooth trend functions, and the noise ε_{it} satisfies the following assumptions:

(A) The noise is a finite order autoregressive process

$$\varepsilon_{it} = AR(d_i) = \sum_{k=1}^{d_i} \phi_{ik}\varepsilon_{i,t-k} + e_{it} \quad (i = 1, \ldots, p; t = 1, \ldots, T),$$

where conditions on e_{it} are specified in assumption (B), and the polynomial $\phi_i(\lambda)$ has all its roots outside the closed unit disk.

(A') The noise is an infinite order autoregressive process

$$\varepsilon_{it} = AR(\infty) = \sum_{k=1}^{\infty} \phi_{ik}\varepsilon_{i,t-k} + e_{it} \quad (i = 1, \ldots, p; t = 1, \ldots, T),$$

where conditions on e_{it} are specified in assumption (B), and the $AR(\infty)$ does not degenerate to a finite dimensional autoregressive representation of order d.

(B) In assumptions (A) and (A'), e_{it} $(i = 1, \ldots, p)$ are independent and identically distributed random variables $E(e_{it}) = 0$, $E(e_{it}^2) = \sigma_i^2$, $E(e_{it}^4) < \infty$, and $\{e_{it}\}_{t=1}^T$ and $\{e_{jt}\}_{t=1}^T$ are independent if $i \neq j$.

Lyubchich and Gel (2016) are interested in testing whether p observed time series have the same trend of some pre-specified smooth parametric form $f(\theta, u)$:

$$H_0 : m_i(u) = c_i + g(\theta, u) \quad (i = 1, \ldots, p),$$
$$H_1 : \text{there exists i, such that } m_i(u) \neq c_i + g(\theta, u),$$

where the reference curve $g(\cdot, u) : (0, 1] \to \mathbb{R}$ belongs to a known family of smooth parametric functions $S = \{g(\theta, \cdot), \theta \in \Theta\}$, and Θ is a set of possible parameter values and also is a subset of Euclidean space. For identifiability,

$\sum_{i=1}^{p} c_i = 0$ is assumed. Notice that the hypothesis include (but are not limited to) three special cases: $g(\theta, u) \equiv 0$, i.e. testing for no trend; $g(\theta, u) = \theta_0 + \theta_1 u$, i.e. testing for a common linear trend; $g(\theta, u) = \theta_0 + \theta_1 u + \theta_2 u^2$, i.e. testing for a common quadratic trend.

The Algorithm for testing H_0 consists on the following steps:

Step 1 Estimate the joint hypothetical trend $g(\theta, \cdot)$ using the aggregated sample $\{\bar{Y}_{\cdot t} = 1/T \sum_{i=1}^{p} Y_{it}\}_{t=1}^{T}$ with a $\sqrt{T}$-consistent estimator (e.g. the nonlinear least squares method).

Step 2 Apply the local factor test statistic to each de-trended and filtered series of residuals $\hat{e}_{it}$, which under H_0 behave asymptotically like independent and identically distributed e_{it}:

$$\begin{aligned}\hat{e}_{it} &= \hat{\varepsilon}_{it} - \sum_{j=1}^{d_i(T)} \hat{\phi}_{ij} \hat{\varepsilon}_{i,t-j} \\ &= \{Y_{it} - \sum_{j=1}^{d_i(T)} \hat{\phi}_{ij} Y_{i,t-j}\} - \{g(\hat{\theta}, u_t) - \sum_{j=1}^{d_i(T)} \hat{\phi}_{ij} g(\hat{\theta}, u_{t-j})\}\end{aligned}$$

with $\hat{\phi}_{ij}$ a $\sqrt{T}$-consistent estimator of ϕ_{ij} $(j = 1, \ldots, d_i(T))$ obtained from $\{\hat{\varepsilon}_{it}\}_{t=1}^{T}$ for the ith time series. Hence, the individual local factor test statistic for each observed process takes the form of the classical F-statistic as the ratio of mean square for treatments (mst) and mean square for errors (mse) as in Wang et al. (2008) whose initials give the statistic its name:

$$\begin{aligned}wavk_i(k_{iT}) = F_T &= \frac{mst}{mse} \\ &= \frac{k_{iT}}{T-1} \sum_{t=1}^{T} (\bar{V}_{t\cdot} - \bar{V}_{\cdot\cdot})^2 / \frac{1}{T(k_{iT}-1)} \sum_{t=1}^{T} \sum_{j=1}^{k_{it}} (V_{tj} - \bar{V}_{t\cdot})^2\end{aligned}$$

where, k_{iT} is the number of the nearest values of $\hat{e}_{it}$ used to construct a local window W_{it} around each t (this for each observed time series), $\{V_{t1}, \ldots, V_{tk_{iT}}\} = \{\hat{e}_{ij} : j \in W_{it}\}$, $\bar{V}_{t\cdot}$ and $\bar{V}_{\cdot\cdot}$ are the mean of the tth group and the grand mean, respectively.

Step 3 Construct a sequence of p statistics $\{wavk_1(k_{1T}), \ldots, wavk_p(k_{pT})\}$. Then,

the new synchronism test statistic is

$$S_T = \sum_{i=1}^{p} k_{iT}^{-1/2} wav k_i(k_{iT}).$$

Under assumptions (A)(or (A')), (B)-(D), if d_i is finite but unknown, or if ε_{it} follows (A'), then $d_i(T) = O(\ln(T))$. Then, under H_0 for fixed d, as $T \to \infty$, $k_{iT} \to \infty$, $k_{iT}^{3/2}/T \to 0$ and $k_{iT}/(\ln(T))^m \to \infty$, $m > 1$,

$$T^{1/2} S_T \Rightarrow N\left(0, \frac{4}{3}\sum_{i=1}^{p} \sigma_i^4\right).$$

As with other local factor methods, convergence of the test statistic S_T to its asymptotic distribution might be slow for small sample sizes. To enhance the finite sample performance, they propose to employ a sieve bootstrap procedure (Bühlmann et al., 1997) which provides an asymptotically correct size of the test (i.e., α-level under H_0) even when the linear noise does not degenerate to a finite-dimensional representation.

Chen and Wu (2019) propose statistical inference for trends of high-dimensional time series. Based on a modified $\mathcal{L}^2$ distance between parametric and nonparametric trend estimators, they propose a de-diagonalized quadratic form test statistic (which takes into account both temporal and spatial dependences) for testing patterns on trends, such as linear, quadratic, or parallel forms. They develop an asymptotic theory for the test statistic. A Gaussian multiplier bootstrap testing procedure is proposed for an improved finite sample performance and a faster convergence rate. Additionally, they consider estimation of long-run covariance matrices and propose normalized Frobenius norm consistency.

Suppose to observe p-dimensional time series $X_t = (X_{t1}, \ldots, X_{tp})'$, $t = 1, \ldots, T$, $p \geq 1$, based on the following model:

$$X_t = m(t/T) + \varepsilon_t, \tag{2.2}$$

where $\varepsilon_t \in \mathbb{R}^p$ is a zero-mean p-dimensional stationary process and $m(\cdot) = (m_1(\cdot), \ldots, m_p(\cdot))' : [0,1] \to \mathbb{R}^p$, is an unknown trend function. They are interested in testing the null hypothesis that the trend function belongs to some given parametric family

$$H_0 : m_j(u) = g_j(\theta_j, u), \quad j = 1, \ldots, p, \tag{2.3}$$

where θ_j is an unknown parameter vector and function $g_j(\cdot,\cdot) : \mathbb{R}^{d+1} \to \mathbb{R}$ has a known pre-specified form. They argue that in Degras et al. (2011) (and then Zhang (2013)) and Lyubchich and Gel (2016) the process ε_t have independent components. To incorporate both *temporal and spatial dependencies*, they consider the widely used moving average (MA) process

$$\varepsilon_t = \sum_{i \geq 0} A_i \eta_{t-i} \tag{2.4}$$

where $\eta_t = (\eta_{t1}, \ldots, \eta_{tl})'$ and η_{tj} (with $t, j \in \mathbb{Z}$), are independent and identically distributed (iid) random variables with zero mean and unit variance, and A_i, $i \geq 0$, are matrices in $\mathbb{R}^{p \times l}$ such that ε_t is a proper random vector. If $A_i = 0$ for all $i \geq 1$, then the noise sequences are temporally independent; if $l = p$ and matrices A_i are diagonal, then the sequences are spatially independent. In the latter case, $\{\varepsilon_{tj}\}_{t=1}^T$ becomes a MA sequence independently distributed with respect to different j.

In order to test (2.3), they define a modified ISE based test statistic

$$\hat{I}_{p,T}^M = \sum_{|i-j| \geq M} a_{i,j} \hat{\varepsilon}_i' \hat{\varepsilon}_j, \quad \hat{\varepsilon}_t = X_t - g(\hat{\theta}, t/T), \tag{2.5}$$

where $a_{i,j} = \int_0^1 \omega_b(i/T, u)\omega_b(j/T, u)du$ (with $\omega_b(u, v)$ the local linear weights as in Fan and Gijbels (1996)), $\hat{\epsilon}_t$ are estimates of ϵ_t and $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)'$ is an estimate of $\theta = (\theta_1, \ldots, \theta_p)'$.

Under assumptions that the kernel function is Lipschitz continuous on $\mathbb{R}$ with compact support $[-2, 2]$ and mild restrictions on the dependence of the noise sequence ε_t, they show the asymptotic normality of $\hat{I}_{p,T}^M$. To ensure this result in the high-dimensional setting, they consider the restriction $p = o(h^{(\beta+1/2)/\gamma} T^{\beta/\gamma})$, where $0 < \gamma \leq 1/2$, h is the kernel bandwidth such that $h^2 T \to \infty$ and $\beta > 1$ such that for all $k \geq 0$, $\sum_{|i| \geq k} tr|(\Gamma_i)|/tr(\Sigma) \leq c(k \vee 1)^{-\beta}$, with $\Gamma_k := E\varepsilon_0 \varepsilon_k'$ and $\Sigma := \sum_{t=-\infty}^{+\infty} E\varepsilon_0 \varepsilon_t'$. This restriction on p is a generalization of the case $1 \leq \varrho \leq \sqrt{p}$ where $\varrho = tr(\Sigma)/|\Sigma|_F$. This last equality come from the sufficient condition for boundedness of the noise coefficients which, under $p \to \infty$ case, are supposed to be diagonal. Namely, if $\varrho \asymp 1$ then no restriction on p is needed to ensure boundedness, on the other hand, if $\varrho \asymp p^\gamma$ then a restriction needs to be imposed on the rate of p. They show also that is possible to extend the test statistic to the case of nonlinear time series.

2.1.1 Common considerations

Note that in the cited works the central idea is to test if the time series belong to the same predefined, hence known, family of parametric trend functions. This implies that those tests are quite restrictive in the sense that one needs to know in advance the parametric family before undertaking the test. It is possible to achieve a more general result considering a procedure which does not need this knowledge and that is able to distinguish such a characteristic. This is the aim of this thesis, how to achieve this will be shown from here on out.

2.2 Checking for Trends in High-Dimensional Time Series

Suppose to observe p time series of the form

$$Y_{it} = m_i(t/T) + \varepsilon_{it}, \quad i = 1, \dots, p; t = 1, \dots, T \tag{2.6}$$

where $m_i : [0,1] \to \mathbb{R}$ are unknown trend functions and $\{\varepsilon_{it}\}_{t=1}^{T}$ are zero mean error processes. Suppose now to be interested in testing the following hypothesis

$$H_0 : m_i(u) = g(\theta_i, u), \quad i = 1, \dots, p, \tag{2.7}$$

where the function $g(\cdot, \cdot)$ has a known form and θ_i is a parameter vector of g being identically zero, constant, and special cases $g \equiv 0$, $g \equiv a$ constant and $g(\theta_i, u) = \theta_{i0} + \theta_{i1}u$ for some $\theta_i = (\theta_{i0}, \theta_{i1}) \in \mathbb{R}^2$ correspond to testing whether a signal exists, is time-varying and nonlinear, respectively.

In the context of grouping time series based on feature and using the previous setting, of particular interest is to group the time series according to whether their trend is constant, linear or nonlinear. In order to make this partition, one can test the following

$$H_0 : m_i^{(1)}(u) \equiv \theta_{i1} \tag{2.8}$$

where $m_i^{(1)}(u)$ is the trend first derivative at u of the ith time series and $\theta_i^* \in \mathbb{R}$ represents the angular coefficient if $m_i(u) = \theta_{i0} + \theta_{i1}u$. Note also that if the trend is a constant (i.e. $m_i(u) = \theta_{i0}$), then $\theta_{i1} = 0$ and

$$H_0 : m_i^{(1)}(u) \equiv 0. \tag{2.9}$$

On the other hand, if the trend is not a linear function of u, one can always define

$$H_1 : m_i^{(1)}(u) \not\equiv \theta_{i1}. \tag{2.10}$$

The use of the first derivative instead of the main function presents multiple advantages: (i) on the mathematical point of view is quite intuitive the use of the first derivative to highlight the linearity of a function; (ii) one can assert if a trend is linear or not without imposing a predefined model for trend but only looking if the angle coefficient is constant; (iii) this type of test makes a partition of the set of the given time series which may be used in a further analysis as starting point (i.e. it gives a useful previous knowledge on the trend composition); (iv) it does not impose restrictions on the trend composition such as those which are imposed when the presence of parallelism is tested (Zhang, 2013).

A natural way to estimate the first derivative of an unknown function is to use a nonparametric estimator which is able to estimate the underlying function with very few assumptions. In particular, the Local Polynomial estimator has the appealing characteristic to include in its definition the first, say $d + 1$, derivatives. Once the estimate of the function first derivative is obtained, one needs to test it on the support of each trend function. A common choice is to use the Integrated Square Error (ISE) based test statistic. With this setting in mind, one can test the trend first derivative of multiple time series together (see Degras et al. (2011)) even in the context of high-dimensionality (see Chen and Wu (2019) also), namely when p goes to infinity as function of T.

In the following subsections the Local Polynomial estimator for α-mixing processes will be presented together with the testing procedure based on the Integrated Squared Error proposed by the literature. The assumption on the error to be α-mixing is due to the fact that one want to estimate the first derivative of the trend by using a nonparametric estimator under one of the least restrictive dependence conditions for the error term.

2.2.1 Local polynomial estimator for Mixing processes

Following the monographs of Wand and Jones (1994) and Fan and Gijbels (1996), nonparametric regression is studied in both fixed design and random design contexts. In the univariate fixed design case the design consists of $x_1, \ldots, x_n$ which are ordered non-random numbers. An equally spaced fixed

design is one for which $x_{i+1} - x_i$ is constant for all i. For the fixed design case the response variables are assumed to satisfy

$$Y_i = m(x_i) + v(x_i)\varepsilon_i, \quad i = 1, \ldots, n \tag{2.11}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent random variables with mean zero and variance σ_ε^2. m is the mean regression function, or simply the regression function, since $E(Y_i) = m(x_i)$, while v is called the variance function since $Var(Y_i) = v^2(x_i)\sigma_\varepsilon^2$. The random design regression model arises when the bivariate sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of random pairs is observed, in which case the model can be written as

$$Y_i = m(X_i) + v(X_i)\varepsilon_i, \quad i = 1, \ldots, n \tag{2.12}$$

where, conditional on $X_1, \ldots, X_n$ the ε_i are independent random variables with zero mean and finite variance. However, in the random design context

$$m(x) = E(Y|X = x) \text{ and } \sigma^2(x) = \sigma_\varepsilon^2 v^2(x) = Var(Y|X = x) \tag{2.13}$$

are, respectively, the conditional mean and variance of Y given $X = x$.

Now, suppose one is interested in estimate the regression function $m(x) = E(Y|X = x)$ and its derivatives $m^{(1)}(x), m^{(2)}(x), \ldots, m^{(d)}(x)$. Suppose also that the $(d+1)th$ derivative of $m(x)$ at the point x_0 exists. One can then approximate the unknown regression function $m(x)$ locally by a polynomial of order d. A Taylor expansion gives, for x in a neighborhood of x_0,

$$m(x) \approx \sum_{j=0}^{d} \frac{m^{(j)}(x_0)}{j!}(x - x_0)^j. \tag{2.14}$$

This polynomial is fitted locally by a weighted least squares regression problem

$$\min_{\beta_j, j=0,\ldots,d} \sum_{i=1}^{n} \left[Y_i - \sum_{j=0}^{d} \beta_j (X_i - x_0)^j \right]^2 K_h(X_i - x_0), \tag{2.15}$$

where $K_h(u) = \frac{1}{h} K\left(\frac{u}{h}\right)$ with h is a bandwidth controlling the size of the neighborhood at x_0 and K is a kernel function assigning weights to each point. Given $\{\hat{\beta}_j, j = 0, \ldots, d\}$ the solution of (2.15), then the estimator of $m^{(\nu)}(x_0), \nu = 0, \ldots, d$, is $\hat{m}^{(\nu)}(x_0) = \nu!\hat{\beta}_\nu$. In order to estimate the entire function $m^{(\nu)}(\cdot)$, one needs to solve (2.15) for all x_0 in the domain of interest.

Local Polynomial fitting is an attractive method both from theoretical and practical point of view. Other commonly used kernel estimators, such as the Nadaraya-Watson estimator and the Gasser-Miiller estimator suffer from some drawbacks. The Nadaraya-Watson estimator leads to an undesirable form of the bias, while the Gasser-Miiller estimator has to pay a price in variance when dealing with a random design model. Local polynomial fitting also has other advantages. The method adapts to various types of designs such as random and fixed designs, highly clustered and nearly uniform designs. Furthermore, there is an absence of boundary effects: the bias at the boundary stays automatically of the same order as in the interior, without use of specific boundary kernels. This is remarkably different from the other methods. Another attractive characteristic is that, since the polynomial is fitted locally, one does not need to know whether $Var(Y|X = x)$ remains constant or not, because it is approximately the same in a local neighborhood.

Even if Local Polynomial fitting has quite useful characteristics, there are several important issues which have to be discussed.

- The choice of the bandwidth parameter h, which plays a rather crucial role. A too large bandwidth under-parametrizes the regression function, causing a large modelling bias, while a too small bandwidth over-parametrizes the unknown function and results in noisy estimates. However this theoretical choice is not directly practically usable since it depends on unknown quantities.

- Another issue in local polynomial fitting is the choice of the order d of the local polynomial. Since the modelling bias is primarily controlled by the bandwidth, this issue is less crucial however. For a given bandwidth h, a large value of d would expectedly reduce the modelling bias, but would cause a large variance and a considerable computational cost. There is a general pattern of increasing variability: for estimating $m^{(\nu)}(x_0)$, there is no increase in variability when passing from an even $d = \nu + 2q$ order fit to an odd $d = \nu + 2q + 1$ order fit for any $q \in \mathbb{N}$, but when passing from an odd $d = \nu + 2q + 1$ order fit to the consecutive even $d = v + 2q + 2$ order fit there is a price to be paid in terms of increased variability. Therefore, even order fits $d = \nu + 2q$ are not recommended, but odd order fit of $d = \nu + 2q + 1$ are.

- Another question concerns the choice of the kernel function K. Since the estimate is based on the local regression (2.15) no negative weight

K should be used. Theorem 3.4 in Fan and Gijbels (1996) shows that for all choices of d and ν the optimal weight function is $K(z) = \frac{3}{4}(1 - z^2)_+$, the Epanechnikov kernel, which minimizes the Asymptotic Mean Squared Error (AMSE) of the resulting local polynomial estimators.

In order to derive theoretical results for Local Polynomial estimator, rewrite (2.15) in matrix form is more convenient

$$\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{W} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tag{2.16}$$

where,

$$\mathbf{X} = \begin{pmatrix} 1 & (X_1 - x_0) & \cdots & (X_1 - x_0)^d \\ \vdots & \vdots & & \vdots \\ 1 & (X_n - x_0) & \cdots & (X_n - x_0)^d \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_d \end{pmatrix} \tag{2.17}$$

and $\mathbf{W} = diag\left[K_h(X_i - x_0)\right]$. The solution vector is then given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}. \tag{2.18}$$

In particular the matrix $(\mathbf{X}'\mathbf{W}\mathbf{X})$ is positive definite as long as there are at least $d + 1$ local effective design points. This assumption is granted with probability tending to one assuming $nh \to \infty$. The (2.18) allows for a very useful representation of the local estimate of the νth derivative

$$\begin{aligned} \hat{m}^{(\nu)}(x_0) &= \nu! \mathbf{e}'_{\nu+1} \hat{\boldsymbol{\beta}} \\ &= \nu! \hat{\beta}_\nu, \end{aligned} \tag{2.19}$$

where $\mathbf{e}_j$ is the $(d + 1)$ length vector with 1 in the jth position and zeros elsewhere. The last equation highlights that the bias and variance of the local polynomial derivatives estimator depend on those of $\hat{\boldsymbol{\beta}}$. More precisely, under the assumption of i.i.d. $(X_1, Y_1), \ldots, (X_n, Y_n)$, and $\mathbb{X} = \{X_1, \ldots, X_n\}$

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}|\mathbb{X}) &= (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{m} \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{r} \\ Var(\hat{\boldsymbol{\beta}}|\mathbb{X}) &= (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X})(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}, \end{aligned} \tag{2.20}$$

where $\mathbf{m} = (m(X_1), \ldots, m(X_n))'$, $\mathbf{r} = \mathbf{m} - \mathbf{X}\boldsymbol{\beta}$ and $\boldsymbol{\Sigma} = diag[K_h^2(X_i - x_0)\sigma^2(X_i)]$. Since $\mathbf{r}$ and $\boldsymbol{\Sigma}$ are unknown quantities, there is a need for approximating bias and variance. Let $f(\cdot)$ be the probability density function of the generic X_j. Assuming $f(x_0) > 0$, and that $f(\cdot)$, $m^{(d+1)}(\cdot)$ and $\sigma^2(\cdot)$ are continuous in a neighborhood of x_0, while $h \to 0$ and $nh \to \infty$. By Theorem 3.1 in Fan and Gijbels (1996), the asymptotic conditional variance of $\hat{m}^{(\nu)}(x_0)$ is given by

$$Var(\hat{m}^{(\nu)}(x_0)|\mathbb{X}) = \mathbf{e}'_{\nu+1} S^{-1} S^* S^{-1} \mathbf{e}_{\nu+1} \frac{\nu!^2 \sigma^2(x_0)}{f(x_0) n h^{1+2\nu}} + o_P\left(\frac{1}{nh^{1+2\nu}}\right), \tag{2.21}$$

where $S = (\mu_{j+l})_{0 \le j,l \le d}$, $S^* = (\nu_{j+l})_{0 \le j,l \le d}$, $\mu_j = \int u^j K(u) du$ and $\nu_j = \int u^j K^2(u) du$. On the other hand, its asymptotic conditional bias takes two forms. If $d - \nu$ is odd, then

$$\begin{aligned} Bias(\hat{m}^{(\nu)}(x_0)|\mathbb{X}) = \mathbf{e}'_{\nu+1} S^{-1} c_d \frac{\nu!}{(d+1)!} m^{(d+1)}(x_0) h^{d+1-\nu} \\ + o_P\left(h^{d+1-\nu}\right), \end{aligned} \tag{2.22}$$

if $d - \nu$ is even, then

$$\begin{aligned} Bias(\hat{m}^{(\nu)}(x_0)|\mathbb{X}) = \mathbf{e}'_{\nu+1} S^{-1} \tilde{c}_d \frac{\nu!}{(d+2)!} \Bigg\{ m^{(d+2)}(x_0) \\ + (d+2) m^{(d+1)}(x_0) \frac{f^{(1)}(x_0)}{f(x_0)} \Bigg\} h^{d+2-\nu} \\ + o_P\left(h^{d+2-\nu}\right), \end{aligned} \tag{2.23}$$

where $c_d = (\mu_{d+1}, \ldots, \mu_{2d+1})'$, $\tilde{c}_d = (\mu_{d+2}, \ldots, \mu_{2d+2})'$ and provided that $f^{(1)}(\cdot)$ and $m^{(d+2)}(\cdot)$ are continuous in a neighborhood of x_0 and $nh^3 \to \infty$. From these two last equations, it is clear that there is a theoretical difference between the even and odd case. In the odd $d - \nu$ case, the bias has a simpler form which does not depend on $f^{(1)}(x_0)$. A more general result is that polynomial fit with $d - \nu$ odd outperform those with $d - \nu$ even in terms of increase of variability (see Section 3.3 in Fan and Gijbels (1996)). From (2.22) and (2.23) one can see that, fixing ν, for higher order approximation the bias reduces. Looking at (2.21) seems that d does not affect the variance. This is not true if one explores the behaviour of the constant term: in moving

from an even order to its consecutive odd order there is no loss in variability. Noteworthy is the fact that the variance is non decreasing in d, which means that a lower d will be preferred.

Masry and Fan (1997) highlight that local polynomial fitting can be applied to nonlinear time series modeling also. Namely, under assumption of dependence, without requiring a Markovian structure, they established joint asymptotic normality for derivative estimation when the processes are strongly mixing (said also α-mixing). More precisely, let $\mathcal{F}_i^k$ be the σ-algebra of events generated by the random variables $\{X_j, Y_j, i \leq j \leq k\}$. The stationary processes $\{X_j, Y_j\}$ are called strongly mixing if

$$\sup_{A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_k^\infty} |P(AB) - P(A)P(B)| = \alpha(k) \to 0, \quad \text{as } k \to \infty. \tag{2.24}$$

The condition indicates the maximum dependence between two events k steps apart. Local polynomial fitting techniques continue to apply under the weak dependence in medium or long term, namely, when k is large. The short term dependence does not have much effect on the local smoothing method. The reason is that for any two given random variables X_i and X_j and a point x, the random variables $K_h(X_i - x)$ and $K_h(X_j - x)$ are nearly uncorrelated as $h \to 0$. This property is, however, not shared by parametric estimators. Note that the local polynomial fit has a quite general setup which allows to estimate functions of the form

$$m_\psi(x) = E(\psi(Y)|X = x). \tag{2.25}$$

Examples are $\psi(Y) = I_{\{Y \leq y\}}$, which corresponds to the problem of estimating the conditional distribution $m_\psi(x) = P(Y \leq y|X = x)$, and, in particular, $\psi(Y) = Y^2$ which corresponds to estimating the conditional second moment. By Theorem 6 in Masry and Fan (1997) for (2.25), under conditions

(i) for some $\delta > 2$ and $a > 1 - 2\delta$, the kernel function K is a bounded density satisfying $u^{2\delta d+2}K(u) \to 0$ as $|u| \to \infty$ and with a compact support $[-1, 1]$;

(ii) the joint density of (X_0, X_l), $f(u, v; l) \leq M_1 < \infty$ and $E(\psi(Y_1)^2 + \psi(Y_l)^2|X_1 = u, X_l = v) \leq M_2 < \infty$, $\forall l \geq 1$ and for u, v in a neighborhood of x_0;

(iii) $\sum l^a[\alpha(l)]^{1-2\delta} < \infty$, $E(|\psi(Y_0)|^\delta|X = x) \leq M_3 < \infty$ for x in a neighborhood of x_0;

(iv) for $h \to 0$ and $Th \to \infty$, there exists a sequence of positives integers satisfying $s_T \to \infty$ and $s_T = o((Th)^{1/2})$ such that $(T/h)^{1/2}\alpha(s_T) \to 0$, as $T \to \infty$;

(v) the conditional distribution of $Y|X = x$ is continuous at point $x = x_0$;

if $h = O(T^{1/(2d+3)})$, then as $T \to \infty$

$$Bias(\hat{m}_\psi^{(\nu)}(x_0)) = \frac{m_\psi^{(d+1)}(x_0)\nu! B_\nu}{(d+1)!} h^{d+1-\nu}, \tag{2.26}$$

$$Var(\hat{m}_\psi^{(\nu)}(x_0)) = \frac{(\nu!)^2 V_\nu \sigma_\psi^2(x_0)}{Th^{2\nu+1} f(x_0)}, \tag{2.27}$$

where B_ν and V_ν are the νth element of $S^{-1}c_d$ and the νth diagonal element of $S^{-1}S^*S^{-1}$ respectively and $\sigma_\psi^2 = Var(\psi(Y)|X = x_0)$.

Note that, for the second point listed in the issues of Local Polynomial estimator, the bandwidth depends on the order of the derivative ν. Hence, also the Bias is function of the true derivative of order $\nu + 2$ as one can see from (2.26).

Following Masry and Fan (1997) and under mild assumptions on the covariace structure, Francisco-Fernández and Vilar-Fernández (2001) show very similar results for the case of fixed design

$$Bias(\hat{m}_\psi^{(\nu)}(x_0)) = \frac{m_\psi^{(d+1)}(x_0)\nu! B_\nu}{(d+1)!} h^{d+1-\nu}, \tag{2.28}$$

$$Var(\hat{m}_\psi^{(\nu)}(x_0)) = \frac{(\nu!)^2 V_\nu c(\varepsilon)}{Th^{2\nu+1}}(1 + o(1)), \tag{2.29}$$

where $c(\varepsilon)$ represents the sum of all covariances.

2.2.1.1 The choice of h

In the i.i.d. setting of $(X_1, Y_1), \ldots, (X_n, Y_n)$, a theoretical optimal local bandwidth for estimating $m^{(\nu)}(x_0)$ is obtained by minimizing the conditional Mean Squared Error (MSE) given by

$$[Bias(\hat{m}^{(\nu)}(x_0)|\mathbb{X})]^2 + Var(\hat{m}^{(\nu)}(x_0)|\mathbb{X}).$$

For the same reasons for (2.20), the MSE can be approximated by the asymptotic MSE (AMSE) which gives the asymptotic optimal local bandwidth via

$$h_{\nu,opt}(x_0) = \frac{1}{n^{1/(2d+3)}} C_{\nu,d}(K) \left[\frac{\sigma^2(x_0)}{[m^{(d+1)}(x_0)]^2 f(x_0)} \right]^{1/(2d+3)} \tag{2.30}$$

where

$$C_{\nu,d}(K) = \left[\frac{(d+1)!^2(2\nu+1)\int K_\nu^{*2}(u)du}{2(d+1-\nu)(\int u^{d+1}K_\nu^*(u)du)^2} \right]^{1/(2d+3)}$$

and K_ν^* is the equivalent kernel (Fan and Gijbels, 1996). Note that the MSE is a measure of local loss, then if one is not interested in local measures, others need to be taken into account.

A commonly used, simple measure of global loss is the weighted Mean Integrated Squared Error (MISE). Minimization of the conditional weighted MISE

$$\int \left([Bias(\hat{m}^{(\nu)}(x)|\mathbb{X})]^2 + Var(\hat{m}^{(\nu)}(x)|\mathbb{X}) \right) \omega(x)dx.$$

with $\omega > 0$ some weight function, leads to a theoretical optimal constant bandwidth

$$h_{\nu,opt} = \frac{1}{n^{1/(2d+3)}} C_{\nu,d}(K) \left[\frac{\int \sigma^2(x)\omega(x)/f(x)dx}{\int [m_\psi^{(d+1)}(x)]^2 \omega(x)dx} \right]. \tag{2.31}$$

In the dependence setting of $\alpha-$mixing processes, by (2.26) and (2.27), the optimal local bandwidth is

$$h_{\nu,opt}(x_0) = \frac{1}{T^{1/(2d+3)}} \left[\frac{[(d+1)!]^2 V_\nu \sigma_\psi^2(x_0)/f(x_0)}{2(d+1-\nu)[m_\psi^{(d+1)}(x_0)]^2 B_\nu^2} \right]^{1/(2d+3)} \tag{2.32}$$

which, in the fixed design case (Francisco-Fernández and Vilar-Fernández, 2001) can be restated as

$$h_{\nu,opt}(x_0) = \frac{1}{T^{1/(2d+3)}} C_{\nu,d}(K) \left[\frac{c(\varepsilon)}{[m_\psi^{(d+1)}(x_0)]^2} \right]^{1/(2d+3)}, \tag{2.33}$$

where $c(\varepsilon)$ as in (2.29). The associated global optimal constant bandwidth is then given by

$$h_{\nu,opt} = \frac{1}{T^{1/(2d+3)}} C_{\nu,d}(K) \left[\frac{c(\varepsilon)}{\int [m_{\psi}^{(d+1)}(x)]^2 dx} \right]^{1/(2d+3)}. \tag{2.34}$$

These asymptotically optimal bandwidths depend on unknown quantities such as the design density $f(\cdot)$, the conditional variance $\sigma_{\psi}^2(\cdot)$ and the derivative function $m_{\psi}^{(d+1)}(\cdot)$, and hence further work is needed for achieving practical bandwidth selection procedures. For an exaustive review on the different methods for bandwidths selection, see Fan and Gijbels (1996) and Wand and Jones (1994). In particular various techniques have been proposed: boostrap techniques, modified versions of cross-validation, plug-in approaches and procedures based on correlation among others. A very popular method for bandwidth selection is the Leave-one-out Cross-validation (see Chapter 5 of Härdle (1990)).

2.2.2 Testing trend first derivative

A natural way to check a nonparametric regression function defined on the interval $[0, 1]$, is to rely on the $\mathcal{L}^2$ distance between the regression function $m(\cdot)$ and its linear estimate $\hat{m}(\cdot) = \sum_{t=1}^{T} \omega_t(\cdot) Y_t$, where $\omega_t(\cdot)$, $t = 1, \ldots, T$ are weight functions which depend on the fixed design points $x_1, \ldots, x_T$. Then, one can define the quadratic form statistic

$$I = \int_0^1 [\hat{m}^{(1)}(u) - m^{(1)}(u)]^2 du. \tag{2.35}$$

Since one can define $x_t = t/T$, the statistic can be used as a mean to test the first derivative of the trend function estimated by Local Polynomial estimator. Generalizations of (2.35) has been studied under different settings, see for example Hall et al. (1984), Ioannides (1992) and Manteiga and Fernandez (1995).

2.3 The proposed procedure

The proposal discussed in this thesis regards the classification of time series by looking at the first derivative of the deterministic trend in a context of

high-dimensionality by means of a nonparametric estimator. If the trend is constant, then its first derivative is zero, if the trend is linear, then its first derivative is constant. If none of the previous happen, then the trend is of course nonlinear and then its first derivative will be not constant. In this way the time series can be divided into three groups. This approach can be included in the category of "clustering of time series based on features", since the trend composition can be considered as a feature of the time series.

Suppose to observe p (which may goes to infinity as function of the time horizon) independent time series of the form

$$Y_{it} = m_i(t/T) + \varepsilon_{it}, \quad i = 1, \dots, p; t = 1, \dots, T \tag{2.36}$$

where $m_i : [0,1] \to \mathbb{R}$ are unknown trend functions and $\{\varepsilon_{it}\}_{t=1}^{T}$ are zero mean, strongly mixing error processes. In order to partitioning those time series according to their trend composition (constant, linear or nonlinear), one can estimate the first derivative of the trend by using a nonparametric estimator under one of the least restrictive dependence conditions of the error term. The proposed nonparametric estimator for the trend first derivative, at point $x \in [0,1]$, has the form

$$\hat{\beta}(x) = \frac{1}{Th^2} \sum_{t=1}^{T} K_h(t/T - x)(t/T - x)Y_t, \tag{2.37}$$

where $K_h(u) = \frac{1}{h}K\left(\frac{u}{h}\right)$ with $K(\cdot)$ is a symmetric Lipshitz continuous kernel function with bounded support, $h = h_T > 0$ is the bandwidth such that $Th^4 \to \infty$ as $T \to \infty$. The proposed estimator, based on the guiding line of Local Polynomial estimator with fixed design, has the appealing characteristic that it is not only asymptotically normal distributed, as shown in Proposition 2, but also its expected value is proportional to the true first derivative by a known quantity as $T \to \infty$.

Under the reasonable assumption that the number of time series with nonlinear trend is finite, the proposed partition procedure consists on two stages. In the first one, the proposed estimator $\hat{\beta}(x)$ is tested to be zero or not, by the following statistic

$$\hat{I}_\beta = \frac{n^{4/7}}{\mu_2^* c(\varepsilon)} \sum_{j=1}^{k_T} \hat{\beta}(x_j)^2, \tag{2.38}$$

where $\mu_2^* = \int_{-1}^{1} u^2 K(u)^2 du$, $c(\varepsilon) = \gamma_\varepsilon(0) + 2\sum_{k=1}^{\infty} \gamma_\varepsilon(k)$ and $k_T = O(T)$. Once $c(\varepsilon)$ is substituted by the spectral density nonparametric consistent estimator valued at frequency zero $\hat{c}(\varepsilon)$, $\hat{I}_\beta$ allows to distinguish the time series with constant trend. Under the hypothesis that the time series has a constant trend, it follows a chi-squared distribution. In the second one, the difference between the estimator at different points is used in a screening approach to make the further linear/non linear partition of the remaining time series from the previous stage. More precisely, defining

$$\hat{D}(x_1, x_2) = \hat{\beta}(x_1) - \hat{\beta}(x_2), \tag{2.39}$$

where $x_1, x_2 \in (h, 1-h)$, the statistic

$$\hat{I}_D(x) = \frac{1}{k_T} \sum_{j=1}^{k_T} D(x, x_j)^2, \tag{2.40}$$

is used to rank the remaining time series. This ensures, with probability tending to 1, that one can estimate the set which contains the true set of time series with nonlinear trend under the sparsity assumption that the latter has a finite number of elements. Furthermore, the consistency results for both stages are guaranteed by Theorems 2 and 3 in the High-dimensional setting. In particular, as it will be discussed in Chapter 3, Theorem 3 implies the Sure Screening property (Fan and Lv, 2008).

In other words, the first stage is used to select the time series with constant trend by using a testing procedure while the second is a screening procedure which gives the set which contains, with probability tending to 1, the true set of time series with nonlinear trend. The Algorithm below gives the details of the various steps. Setting the first type error α and the bandwidth for each time series, the test statistic $\hat{I}_\beta$ is calculated for each time series and compared against a chi-squared random variable. If the test holds, then the time series is labelled to belong to the set of time series with constant trend C_1. If the set of the remaining time series is not empty, then the statistic $\hat{I}_D$ is used to rank them and the first say s are labelled to belong to the set C_3, which contains with probability tending to one the time series with nonlinear trend. The procedure ends by assigning all the others to the set C_2 of the time series with linear trend.

The use of the mentioned approach presents multiple advantages: (i) on the mathematical point of view is quite intuitive the use of the first derivative

Algorithm Classify HD Time Series by Trend

Set $U := \{1, \ldots, p\}$, $C_1 = C_2 = C_3 = \emptyset$
Set the parameters α, s and h_i, $i \in U$
for $i \in U$ **do**
    Perform the "Trend/NoTrend Test Statistic" $\hat{I}_{\beta,i}$
    if $\hat{I}_{\beta,i} < \chi^2_{(1-\alpha/p,k_T)}$ **then**
        Set $C_1 := C_1 \cup \{i\}$
Set $U := U \backslash C_1$
if $U = \emptyset$ **then**
    return C_1, C_2, C_3
else Perform the "Lin/NoLin Statistic" $\hat{I}_{D,i}$, $i \in U$, and sort them as $\hat{I}_{D,\sigma(1)} \geq \cdots \geq \hat{I}_{D,\sigma(p_2)}$
    Set $C_3 := \{\sigma(1), \ldots, \sigma(s)\}$ and $C_2 := U \backslash C_3$
    return C_1, C_2, C_3

to highlight the linearity of a function; (ii) one can assert if a trend is linear or not without imposing a predefined mathematical model; (iii) this type of procedure makes a partition of the set of the given time series which may be used in a further analysis as starting point (i.e. it gives a useful previous knowledge on the trend composition for a deeper clustering analysis); (iv) it does not impose restrictions on the trend composition such as those which are imposed when the presence of parallelism is tested; (v) it gives mathematical guarantees in the high-dimensional setting since it is consistent in the case of $p = o\left(T^{1/2}/\log T\right)$.

Chapter 3

Theoretical results

For the sake of the reader, the setting will be restated. Suppose to observe p independent time series of the form

$$Y_{it} = m_i(t/T) + \varepsilon_{it}, \quad i = 1, \dots, p; t = 1, \dots, T \tag{3.1}$$

where $m_i : [0,1] \to \mathbb{R}$ are unknown trend functions and $\{\varepsilon_{it}\}_{t=1}^{T}$ are zero mean, strongly mixing error processes. In order to partitioning those time series according to their trend composition (constant, linear or nonlinear), one can use the proposed nonparametric estimator for the first derivative $m^{(1)}(\cdot)$ of the signal $m(\cdot)$

$$\hat{\beta}(x) = \frac{1}{Th^2} \sum_{t=1}^{T} K_h(t/T - x)(t/T - x)Y_t, \tag{3.2}$$

together with the following statistics:

- $$\hat{I}_\beta = \frac{n^{4/7}}{\mu_2^* c(\varepsilon)} \sum_{j=1}^{k_T} \hat{\beta}(x_j)^2, \tag{3.3}$$

  where $\mu_2^* = \int_{-1}^{1} u^2 K(u)^2 du$, $c(\varepsilon) = \gamma_\varepsilon(0) + 2\sum_{k=1}^{\infty} \gamma_\varepsilon(k)$ and $k_T = O(T)$ which represents the number of equally spaced points by which the subinterval of the support has been divided, to evaluate the presence of the trend;

- $$\hat{I}_D(x) = \frac{1}{k_T} \sum_{j=1}^{k_T} D(x, x_j)^2, \tag{3.4}$$

  with $\hat{D}(x, x_j) = \hat{\beta}(x) - \hat{\beta}(x_j)$, to evaluate the linearity of the trend.

In Section 3.2 and 3.3 the statistical properties of (3.2)-(3.4) will be given under the assumptions described in Section 3.1. Section 3.4 will describe consistency results obtained by the proposed procedure. Section 3.5 concludes with theoretical extensions for the procedure.

3.1 Assumptions

In order to achieve the results in the following sections, a list of assumptions will be presented.

(A1) $K(\cdot)$ is a symmetric Lipshitz continuous kernel function with bounded support.

(A2) The sequence of bandwidths $\{h_T\}$, satisfies $h = h_T > 0$, $h \to 0$, $Th^4 \to \infty$ as $T \to \infty$.

(A3) $Cov(\varepsilon_t, \varepsilon_{t+k}) = \sigma^2 c(k)$, $k = 0, \pm 1, \ldots$, such that $\sum_{k=1}^{\infty} k|c(k)| < \infty$.

(A4) $E|\varepsilon_t|^{2+\delta} < \infty$ for some $\delta > 0$.

(A5) $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is a strictly stationary α-mixing process with mixing coefficients $\alpha(k)$ such that $\sum_{k=1}^{\infty} \alpha(k)^{\delta/(2+\delta)} < \infty$. Furthermore, there exists a sequence of positive integers $\{s_T\}$, $s_T \to \infty$ as $T \to \infty$ with $s_T = o\left((Th^3)^{1/2}\right)$ and such that $(Th^{-1})^{1/2} \sum_{k=s_T}^{\infty} \alpha(k)^{1-\gamma} < \infty$, with $\gamma = 2/(2+\delta)$.

(A6) $h = O\left(T^{-1/7}\right)$.

(A7) $|NL| = O(1)$.

In particular, while (A1) and (A2) are common asumptions on the kernel functon and on its bandwidth, (A3) assumes that the strength of correlation between error terms is independent of the sample size, i.e. it depends only on

the lag. (A5) requires that the random component is strong mixing which is one of the least restrictive dependence conditions satisfied by many processes, for example $ARMA$ and $GARCH$ processes. For more insights see Doukhan (1994) and Zhengyan and Chuanrong (1997). (A7) assumes that the set of time series with nonlinear trend (NL) has a finite number of elements. Since $p \to \infty$, (A7) also represents a sparsity assumption for NL. The last one is a key assumption for checking the Sure Screening property of $\hat{I}_D$.

3.2 The Beta Estimator and its properties

In this section will be showed, under (A1)-(A6), the theoretical results for the first derivative trend estimator $\hat{\beta}(x)$ and for the $\hat{D}(x_1, x_2)$. The latter constitute the building blocks for the theoretical results which will be shown in the next section.

Starting with $\hat{\beta}(x)$, Proposition 1 is related to its bias and variance, while Proposition 2 shows that this estimator is distributed asymptotically as a multivariate normal random variable. In particular the mean vector is proportional to the true trend first derivative $m^{(1)}(x)$ by a quantity $\mu_2 = \int_{-1}^{1} K(z) z^2 dz$ (i.e. the second moment of the kernel) which is known. This feature is the one with the greatest attraction since links, in an easy way, the behaviour of the simple proposed estimator $\hat{\beta}(x)$ to that of $m^{(1)}(x)$. Another surprising feature is that the proposed estimator, evaluated in different points of the support, is asymptotically independent. The latter allows for a simplified treatment of this estimator in terms of theoretical results.

Proposition 1. *Under (A1)-(A3), for every* $x \in (h, 1-h)$,

$$\begin{aligned} Bias[\hat{\beta}(x)] &= E[\hat{\beta}(x) - m^{(1)}(x)] \\ &= m^{(1)}(x)\,(\mu_2 - 1) + \frac{m^{(3)}(x) h^2}{6}\mu_4 + O(h^4) + O\left(\frac{1}{Th^2}\right), \end{aligned} \tag{3.5}$$

$$Var[\hat{\beta}(x)] = \frac{c(\varepsilon)}{Th^3}\mu_2^* + o\left(\frac{1}{Th^3}\right), \tag{3.6}$$

where $\mu_j = \int_{-1}^{1} K(z) z^j dz$, $\mu_2^* = \int_{-1}^{1} K^2(z)\, z^2 dz$ *and* $c(\varepsilon) = \sigma^2 \left(c(0) + 2\sum_{k=1}^{\infty} c(k)\right)$.

The proof of the proposition is referred to in the appendix.

Proposition 2. *Let* $X = \left[\hat{\beta}(x_j)\right]_{j=1,\ldots,k_T}$, *under (A1)-(A6)*

$$\sqrt{Th^3}\left(X - \boldsymbol{\mu}\right) \Rightarrow N\left(\mathbf{0}, c(\varepsilon)\mu_2^* \boldsymbol{I}_{k_T}\right), \tag{3.7}$$

where $\boldsymbol{\mu}$ *is the* k_T*-dimensional vector of expected values and in particular*

$$\sqrt{Th^3}\left[\hat{\beta}(x) - m^{(1)}(x)\mu_2 - \frac{m^{(3)}(x)h^2}{6}\mu_4\right] \Rightarrow N\left(0, \mu_2^* c(\varepsilon)\right), \tag{3.8}$$

where μ_j, *with* $j = 2, 4$, μ_2^* *and* $c(\varepsilon)$ *as in Proposition 1.*

The proof of the proposition is referred to in the appendix.

Proposition 3 presents the expected value and variance of $\hat{D}(x_1, x_2)$. These results are very useful as they constitute the milestones for the proof of Theorem 3. Looking more carefully, it can be seen that it reflects the same characteristics observed thanks to Proposition 1 for the estimator $\hat{\beta}(x)$. Its expected value is proportional to the difference $\Delta^{(1)}(x_1, x_2) = m^{(1)}(x_1) - m^{(1)}(x_2)$ by the same known quantity μ_2. Intuitively, if a time series has a nonlinear trend then this difference is not zero, while if a time series has a linear trend $\Delta^{(1)}(x_1, x_2) \equiv 0$. Again, a very simple statistic which has a very desirable characteristic. Note that its variance is independent from the point of the support. This feature is shared also by $\hat{\beta}(x)$ and is very helpful in the proof of the next section results.

Proposition 3. *Under (A1)-(A3), for every* $x_1, x_2 \in (h, 1-h)$

$$\begin{aligned} E[\hat{D}(x_1, x_2)] =& E\left[\hat{\beta}(x_1) - \hat{\beta}(x_2)\right] \\ =& \mu_2 \Delta^{(1)}(x_1, x_2) + \frac{\mu_4 h^2}{6}\Delta^{(3)}(x_1, x_2) + O(h^4) + O\left(\frac{1}{Th^2}\right), \end{aligned} \tag{3.9}$$

$$\begin{aligned} Var[\hat{D}(x_1, x_2)] =& Var\left[\hat{\beta}(x_1) - \hat{\beta}(x_2)\right] \\ =& 2\frac{c(\varepsilon)}{Th^3}\mu_2^* + o\left(\frac{1}{Th^3}\right), \end{aligned} \tag{3.10}$$

where $\mu_j = \int_{-1}^{1} K(z) z^j dz$, $\Delta^{(j)}(x_1, x_2) = m^{(j)}(x_1) - m^{(j)}(x_2)$, μ_2^* *and* $c(\varepsilon)$ *as in Proposition 1. Furthermore, for* $x_1 \neq x_2$

$$Cov(\hat{\beta}(x_1), \hat{\beta}(x_2)) = o\left(\frac{1}{Th^3}\right). \tag{3.11}$$

The proof of the proposition is referred to in the appendix.

3.3 Theoretical properties of the Test and Screening Statistics

In this section the theoretical results will be given for the statistics $\hat{I}_\beta$ and $\hat{I}_D(x)$ based on the proposed first derivative trend estimator $\hat{\beta}(x)$.

Regarding the following theorems, $c(\varepsilon)$ is assumed to be known. How its estimate affects the theoretical results will be shown in Section 3.5.

Starting with Theorem 1, it highlights that $\hat{I}_\beta$ is proportional to the following quadratic form

$$Q = \sum_{j=1}^{k_T} \hat{\beta}(x_j)^2, \tag{3.12}$$

where $k_T = O(T)$ and represents the number of equally spaced points by which the subinterval of the support has been divided. In this way it is easy to prove that it follows a noncentral chi-squared distribution. This result remembers the ISE statistic discussed in section 2.2.2, since if one is interested in testing if a trend has the same value over the support, one can evaluate the squared error of the proposed estimator from the zero value over the support. In this case, the Theorem 1 shows that $\hat{I}_\beta$ follows a noncentral chi-squared distribution whit noncentrality parameter $\mu^* = 0$ under the hypothesis that the trend is constant (no trend).

Theorem 1. *Let $k_T = k$, under (A1)-(A6)*

$$\frac{T^{4/7}}{\mu_2^* c(\varepsilon)} Q \Rightarrow \chi_k^2(\mu^*) \tag{3.13}$$

where $\mu^* = \mu_2^2 \sum_{j=1}^{k} m^{(1)}(x_j)^2$, μ_2^*, μ_2 and $c(\varepsilon)$ as in Proposition 1.

The proof of the theorem is referred to in the appendix.

On the other hand, Theorem 2 shows that the proposed standardized statistic $\hat{I}_\beta$ is consistent in the sense that it is able to distinguish the time series which has a constant trend from the others as the time horizon goes to infinity. In this way one is able to select the time series with constant trend under mathematical guarantee. Note that the standardized $\hat{I}_\beta$ is considered in order to take into account for $k_T = O(T)$.

Theorem 2. *Let* $\hat{I}_\beta = \frac{T^{4/7}}{\mu_2^* c(\varepsilon)} Q$ *with* $k_T = O(T)$*, under (A1)-(A6), where (A4) holds for* $\delta > 2$ *and*

$$\alpha(T) \leq c_1 e^{-c_2 T}, \quad c_1, c_2 > 0,$$

then

$$P\left(\frac{\hat{I}_\beta - E\hat{I}_\beta}{\sqrt{Var\hat{I}_\beta}} > \sqrt{2\log T}\right) = O\left(T^{-1/2}\log T\right) \quad \textit{if the time series has constant trend,}$$

$$P\left(\frac{\hat{I}_\beta - E\hat{I}_\beta}{\sqrt{Var\hat{I}_\beta}} < \sqrt{2\log T}\right) = O\left(T^{-4/7}\right) \quad \textit{otherwise.}$$

The proof of the theorem is referred to in the appendix.

Finally, Theorem 3 gives the guarantee that the proposed statistic $\hat{I}_D(x)$ can be used for the screening procedure. Namely, it postulates the existence of a threshold η which is able to reduce the probability to commit the error of separating the time series between linear and nonlinear trend by $\hat{I}_D(x)$.

Theorem 3. *Under (A1)-(A3) and (A6), there exists* $\eta > 0$ *such that*

$$P(\hat{I}_D > \eta) = O\left(T^{-4/7}\right) \quad \textit{if the time series has a linear trend}$$
$$P(\hat{I}_D < \eta) = o(1) \quad \textit{if the time series has a nonlinear trend.}$$

The proof of the theorem is referred to in the appendix.

3.4 Consistency of the Proposed Procedure in High-Dimensionality

In this section, the consistency of the proposed procedure in the case of high-dimensionality will be showed by means of the results in Theorems 2 and 3 in conjunction with (A7). Precisely, (A7) assumes that the set of time series with nonlinear trend (NL) has a finite number of elements, i.e. $|NL| = O(1)$.

Let

$$p = o\left(\frac{T^{1/2}}{\log T}\right). \tag{3.14}$$

The proposed procedure is composed by two stages, the first is used to test if a time series has constant trend while the second ranks in the first positions a time series with nonlinear trend. In order to prove the consistency of the whole procedure, it is necessary to prove this property for both parts.

Starting with the first stage, considering Theorem 2 and by Boole's inequality

$$\begin{aligned} pP(E_{1,1}) + pP(E_{1,2}) =& pO\left(T^{-1/2}\log T\right) + pO\left(T^{-4/7}\right) \\ =& o(1), \end{aligned} \tag{3.15}$$

where:

- $E_{1,1}$ is the event $\left\{\frac{\hat{I}_\beta - E\hat{I}_\beta}{\sqrt{Var\hat{I}_\beta}} > \sqrt{2\log T}\right\}$ which represents the error of not selecting a time series with constant trend given that it has;
- $E_{1,2}$ is the event $\left\{\frac{\hat{I}_\beta - E\hat{I}_\beta}{\sqrt{Var\hat{I}_\beta}} < \sqrt{2\log T}\right\}$ represents the error of not selecting a time series with no constant trend given that it has not a constant trend.

The equation (3.15) shows the consistency of the first stage of the proposed procedure if one considers as dimensionality (3.14).

For the second stage, considering Theorem 3, (A7) and again by Boole's inequality

$$\begin{aligned} pP(E_{2,1}) + |NL|P(E_{2,2}) =& pO\left(T^{-4/7}\right) + |NL|o(1) \\ =& o(1), \end{aligned} \tag{3.16}$$

where:

- $E_{2,1}$ is the event $\left\{\hat{I}_D > \eta\right\}$ which represents the error when a time series which has a linear trend is considered;
- $E_{2,2}$ is the event $\left\{\hat{I}_D < \eta\right\}$ which represents the error when a time series whit a nonlinear trend is considered.

Equation (3.16) suggests that also the second stage of the proposed procedure is consistent if one considers the dimensionality (3.14). Furthermore,

Theorem 3, (3.14) and (A7) imply that the proposed procedure has the *Sure Screening property* (Fan and Lv, 2008)

$$P(NL \subset \widehat{NL}) \to 1, \tag{3.17}$$

where NL is the true set of time series with non linear trend and $\widehat{NL}$ is the estimated one. Namely, it is able to detect the set which contains the true set of time series with nonlinear trend with probability tending to one as the time horizon goes to infinity.

Remark 1. *Note that, when a screening approach is used, the threshold does not need to be estimated, see for example Fan and Lv (2008).*

Remark 2. *Both results, (3.15) and (3.16), highlight that the dimensionality which ensures the consistency of the proposed procedure is* $p = o\left(T^{1/2}/\log T\right)$. *This result is due by the use of Berry-Essen theorem for mixing processes (see Proof of Theorem 2 in the Appendix) which gives a not so finer approximation result. As a future development a better approximation method could be considered in order to increase the achievable dimensionality.*

3.5 Extensions

In the proofs of Theorems 2 and 3 the crucial points are the rates at which some quantities go to zero. More precisely, in all the proofs $c(\varepsilon)$ is assumed to be known, but in practice it needs to be estimated by the use of the nonparametric Spectral Window estimator for each one of the p time series considered. Furthermore, $c(\varepsilon)$ is function of an unobservable quantity, ε. The latter constitutes an added complication for the derivation of the theoretical results.

Assuming for now that ε is observable so that one can focus on the rate at which $\hat{c}(\varepsilon)$ converges to the true $c(\varepsilon)$. As in Priestley (1981) or Brockwell et al. (1991), this rate depends on the window parameter

$$M = O(T^{1/5}), \tag{3.18}$$

when a Daniell window is used for example. The last implies that

$$Var[\hat{c}(\varepsilon)] = O\left(T^{-4/5}\right) \quad \text{and} \quad Bias[\hat{c}(\varepsilon)] = O\left(T^{-2/5}\right) \tag{3.19}$$

which gives for every $\eta > 0$

$$P\left(|\hat{c}(\varepsilon) - c(\varepsilon)| > \eta\right) = O\left(T^{-4/5}\right). \tag{3.20}$$

The equations (3.19) and (3.20) suggest that one can use the spectral window estimator not afflicting the whole proposed procedure consistency.

On the other hand, assuming now that ε is unobservable, one may estimates it by the use of Local Polynomial estimator

$$\hat{\varepsilon}_t = Y_t - \hat{m}(t/T) \tag{3.21}$$

with $h = O\left(T^{-1/5}\right)$ (see Section 2.2). Now,

$$[d_2(\hat{F}, F)]^2 \leq \frac{1}{T}\sum_{t=1}^{T}\left(\hat{\varepsilon}_t - \varepsilon_t\right)^2 = O_P\left(T^{-4/5}\right), \tag{3.22}$$

where $d_2(\hat{F}, F)$ is the Mallows distance of order 2 (Bickel and Freedman, 1981), $\hat{F}$ and F represents the empirical distributions of $\hat{\varepsilon}_t$ and ε_t, respectively. The latter implies that

$$d_2(\hat{F}, F) = O_P\left(T^{-2/5}\right). \tag{3.23}$$

The last equation suggests that, by replacing ε_t with $\hat{\varepsilon}_t$ for each time series in the whole procedure, the consistency still holds.

Remark 3. *In order to evaluate the consistency one has to consider this rate in conjunction with that one obtained when $\hat{c}(\varepsilon)$ is used. The impact must therefore be assessed for the whole procedure. The last constitutes one of the theoretical developments that this procedure will still have to evaluate.*

Chapter 4

Simulation studies

In the following sections the proposed procedure for classifying high-dimensional time series by trend will be studied by means of Monte Carlo simulations. In general, since the procedure is composed by two stages, they will be checked separately. Precisely, Section 4.1 will show how the quantities h_i and $c_i(\varepsilon)$, $i = 1, \dots, p$, will be estimated, since they are fundamental components of the whole procedure. Section 4.2 will give the general setting for the data generation. Sections 4.3 and 4.4 present the performance results for the testing procedure and the screening stage, respectively. Finally, Section 4.5 reports the results in case the whole procedure is employed.

4.1 Procedure implementation

In order to implement the testing procedure, one needs to choose the cutoff value, estimate $c_i(\varepsilon)$ and select the bandwidth h_i, $i = 1, \dots, p$.

Regarding the choice of the cutting value, Theorem 1 in Chapter 3 highlights that $\hat{I}_\beta$ follows a chi-squared distribution under the null (i.e. the ith time series has constant trend) with k degrees of freedom. Then the cutoff value (i.e. the quantile of the chi-squared random variable) depends on the definition of k. In the simulations the recommended formula is $k_T = 0.3T$, more precisely it is used the $[0.3T, 0.6T]$ interval for the computation of each statistic. It gives less computational burden respect to wider intervals, guaranteeing at the same time completely similar results.

Since $\hat{I}_\beta$ depends on $c(\varepsilon) = \gamma_\varepsilon(0) + 2\sum_{k=1}^{\infty} \gamma_\varepsilon(k)$, in practice it needs to be estimated. The choice to use a nonparametric consistent estimator narrows

the focus on the use of a spectral density estimator valued at zero frequency. In particular, a Spectral Windows estimator is used with a Daniell window (for more details see Chapter 6 of Priestley (1981)) which in practice gives good results.

The last critical point is to estimate the bandwidth for each time series. Considering the results in Sections 4.3 and 4.4, since the estimation of the bandwidth is not the aim of this thesis, it has been calculated by using the (2.34) in Chapter 2 which gives the global optimal estimate. For a more detailed description on the procedure see Francisco-Fernández and Vilar-Fernández (2001). Considering instead the results of Section 4.5 a FeedForward Neural Network (FFNN) estimator is used in order to obtain a plug-in estimator for the optimal bandwidths (Giordano and Parrella, 2019) in order to evaluate the performance of the whole proposed procedure.

4.2 The general setting

In this section the general setting used to generate the independent p time series is presented. Consider the representation

$$Y_{i,t} = m(t/T) + \varepsilon_{i,t}. \tag{4.1}$$

Since it includes the hypothesis of α-mixing errors, two classic cases are considered for the generation of the errors $\{\varepsilon_{i,t}\}$:

- an ARMA(1,1) process

$$\varepsilon_{i,t} = \phi\varepsilon_{i,t-1} + \beta a_{i,t-1} + a_{i,t}, \tag{4.2}$$

  with $\phi = 0.5$ and $\beta = 0.2$;

- a GARCH(1,1) process

$$\begin{aligned} \varepsilon_{i,t} =& v_t a_{i,t} \\ v_t =& \sqrt{\omega + \alpha\varepsilon_{i,t-1}^2 + \beta v_{t-1}^2}, \end{aligned} \tag{4.3}$$

  with $\omega = 0.5$, $\alpha = 0.4$ and $\beta = 0.3$.

For both the compositions of the error process, the $a_{i,t}$ follow an i.i.d. $N(0, \sigma_a^2)$ with $\sigma_a = 0.5$. Note that $E\varepsilon_{i,t} = 0$ in both cases, the long-run variance is

$\gamma_{ARMA} = \frac{1+2\phi\beta+\beta^2}{1-\phi^2}\sigma_a^2$ for the ARMA(1,1) case and $\gamma_{GARCH} = \frac{\omega\sigma_a^2}{1-(\alpha\sigma_a^2+\beta)}$ for the GARCH(1,1) case. For both ARMA and GARCH cases, the following structures of the signal are considered:

- $m(t/T) = 0$, (NoTrend);
- $m(t/T) = at/T$, (Lin);
- $m(t/T) = a\sin(6\pi t/T)$, (NoLin);

with $a \in \mathbb{R}$ fixed in order to obtain a Noise Proportion $\left(\text{NP} = \frac{Var(\varepsilon_i)}{Var(y_i)}\right)$ of: 10%, 20% e 30%. In this way it will be possible to evaluate with more precision the behavior of the proposed procedure considering different incidences of the error respect to the signal.

Finally, in all the simulations conducted, the Epanechnikov kernel $K(u) = \frac{3}{4}\max(0, 1-u^2)$ is used.

Remark 4. *The fact that a sinusoidal signal has been used is due to the need to use a signal structure which is intentionally more complicated than a quadratic one, for example. The sinusoidal signal tends to have greater persistence subjected to differentiation. Furthermore, the example could be seen as a case linked to the deterministic part of the Wold's decomposition, in particular. This underlines the variety of application of the proposed approach even if the case in question may not be included as an example of a nonstationary time series.*

4.3 The testing stage performances

In this section the performances of the proposed $\hat{I}_\beta$ test statistic will be shown. For each structure of the signal presented in the previous section (NoTrend, Lin, NoLin), 1000 realizations are generated with different combinations of T (200, 500 and 1000) and error type ($ARMA(1,1)$ and $GARCH(1,1)$). Moreover, for each generated time series, the $\hat{I}_\beta$ test statistic is performed given $\alpha = 5\%$ and for various levels of h. The following tables show the percentage for which the null hypothesis is accepted (i.e. H_0 : the ith time series has no trend). Note that for those simulations the Bonferroni correction (see line 5 in the Algorithm) is not used since the aim is not to evaluate a family of null hypothesis.

NP	h	$T = 200$	$T = 500$	$T = 1000$
		$ARMA(1,1)$		
100%	0.40	1.000	0.888	0.944
	0.30	0.415	0.551	0.687
	0.20	0.045	0.087	0.119
		$GARCH(1,1)$		
100%	0.40	0.942	0.945	0.992
	0.30	0.782	0.800	0.937
	0.20	0.227	0.235	0.432

Table 4.1: Acceptance probability of 95% for the setting of $m(u) = 0$ and noise as $ARMA(1,1)$ (upper part) and as $GARCH(1,1)$ (lower part). 1000 realizations of the previous setting are generated with different combinations of T and h.

In general, the *size* of a test is the probability of incorrectly rejecting the null hypothesis if it is true. The *power* of a test is the probability of correctly rejecting the null hypothesis if it is false. For a given hypothesis and test statistic, one constrains the size of the test to be small and attempts to make the power of the test as large as possible. With this reminder, Table 4.1 displays that the test statistic $\hat{I}_\beta$ has a size which increases as h shrinks especially for errors with an autoregressive linear structure. This effect is reduced with the increasing of the length of the time series. On the other hand, Tables 4.2 and 4.3 display that the power of the proposed $\hat{I}_\beta$ is very high for reasonable proportions of noise once the bandwidth moves in a neighbourhood of the optimal global one. This last result is highlighted by Table 4.3 in which a too wide deviation from the optimal bandwidth gives a significant drop in power whatever the length of the time series.

The aim of this thesis is not to give a new type of estimator or a new procedure to estimate the bandwidth. With this in mind, the performances of the proposed test statistic $\hat{I}_\beta$ are proved to be quite good and in line to the theoretical results obtained in the previous chapter. It is then able to select the time series with no trend.

NP	h	$T = 200$	$T = 500$	$T = 1000$
		$ARMA(1,1)$		
10%	0.40	0.000	0.000	0.000
	0.30	0.000	0.000	0.000
	0.20	0.000	0.000	0.000
20%	0.40	0.000	0.000	0.000
	0.30	0.000	0.000	0.000
	0.20	0.000	0.000	0.000
30%	0.40	0.000	0.000	0.000
	0.30	0.001	0.000	0.000
	0.20	0.001	0.000	0.000
		$GARCH(1,1)$		
10%	0.40	0.000	0.000	0.000
	0.30	0.000	0.000	0.000
	0.20	0.000	0.000	0.000
20%	0.40	0.000	0.000	0.000
	0.30	0.000	0.000	0.000
	0.20	0.000	0.000	0.000
30%	0.40	0.000	0.000	0.000
	0.30	0.000	0.000	0.000
	0.20	0.000	0.000	0.000

Table 4.2: Acceptance probability of 95% for the setting of $m(u) = au$ and noise as $ARMA(1,1)$ (upper part) and as $GARCH(1,1)$ (lower part). 1000 realizations of the previous setting are generated with different combinations of T and h.

NP	h	$T = 200$	$T = 500$	$T = 1000$
		$ARMA(1,1)$		
10%	0.30	0.853	0.993	0.998
	0.20	0.000	0.000	0.000
	0.10	0.000	0.000	0.000
20%	0.30	0.830	0.972	1.000
	0.20	0.000	0.000	0.000
	0.10	0.000	0.000	0.000
30%	0.30	0.762	0.974	1.000
	0.20	0.000	0.000	0.000
	0.10	0.000	0.000	0.000
		$GARCH(1,1)$		
10%	0.30	1.000	1.000	1.000
	0.20	0.000	0.000	0.000
	0.10	0.000	0.000	0.000
20%	0.30	1.000	1.000	1.000
	0.20	0.000	0.000	0.000
	0.10	0.000	0.000	0.000
30%	0.30	1.000	1.000	1.000
	0.20	0.000	0.000	0.000
	0.10	0.000	0.000	0.000

Table 4.3: Acceptance probability of 95% for the setting of $m(u) = a\sin(6\pi u)$ and noise as $ARMA(1,1)$ (upper part) and as $GARCH(1,1)$ (lower part). 1000 realizations of the previous setting are generated with different combinations of T and h. The last is considered in a neighbourhood of the global optimal one $h_{opt} \approx 0.10$.

4.4 The screening stage performances

In this section the screening stage performances will be analysed. In doing so, a typical screening performance measure, the Minimum Model Size (MMS), is used. This measure consists, for the case in use, of the highest position at which a time series with a nonlinear trend has been assigned by the ranking induced by the proposed statistic $\hat{I}_D$. Note that the results of this section are obtained without following the whole procedure, this means to obtain the number of time series with nonlinear trend those time series have not been tested as time series with no trend from the previous stage. To avoid ambiguities, for the time series with no nonlinear trend the bandwidth value is fixed at 0.40. This is coherent not only with the results obtained in the Table 4.1 and 4.2, but also to keep the focus on the second stage of the proposed procedure.

Table 4.4 shows the median of the MMS, with standard deviation in parenthesis, for 200 repetitions of the whole procedure using the various composition of the signal together, divided for type of error. In other words, in the p time series which constitute each repetition, some have no trend, others have a linear trend and still others have a nonlinear trend. The repetitions are performed for three different couples of T and p. Furthermore, various settings of NP and h are considered. The number of time series with nonlinear trend are 5, 8 and 10 for $p = 20, 30, 50$ respectively. The best achievable result is to have all the time series with nonlinear trend in the first top positions. In the neighbourhood of the global optimal bandwidth the statistic $\hat{I}_D$ is able to rank in the top positions the exact number of time series with nonlinear trend. These results show the effectiveness of the proposed procedure in the screening phase.

NP	h	$T = 200$ $p = 20$	$T = 500$ $p = 30$	$T = 1000$ $p = 50$
		$ARMA(1,1)$		
10%	0.30	0 (0)	0 (0)	0 (0)
	0.20	5 (0)	8 (0)	10 (0)
	0.10	5 (0)	8 (0)	10 (0)
20%	0.30	0 (0)	0 (0)	0 (0)
	0.20	5 (0)	8 (0)	10 (0)
	0.10	5 (0)	8 (0)	10 (0)
30%	0.30	0 (0.35)	0 (0)	0 (0)
	0.20	5 (0)	8 (0)	10 (0)
	0.10	5 (0)	8 (0)	10 (0)
		$GARCH(1,1)$		
10%	0.30	5 (2.41)	8 (0.94)	10 (0)
	0.20	5 (0)	8 (0)	10 (0)
	0.10	5 (0)	8 (0)	10 (0)
20%	0.30	8 (3.85)	9 (3.68)	10 (3.30)
	0.20	5 (0)	8 (0)	10 (0)
	0.10	5 (0)	8 (0)	10 (0)
30%	0.30	10 (4.05)	12 (5.24)	11 (5.90)
	0.20	5 (0)	8 (0)	10 (0)
	0.10	5 (0)	8 (0)	10 (0)

Table 4.4: Median of MMS with standard deviation in parenthesis for 200 iterations of the whole procedure over realizations which contains all the three types of signal for the two type of errors and for various h. The last is considered in a neighbourhood of the global optimal one $h_{opt} \approx 0.10$. The number of time series with nonlinear trend is respectively 5, 8 and 10 for the three combination of T and p.

4.5 The whole procedure performances

For the sake of the reader the Algorithm, which shows the whole procedure, is restated. To improve the reading and understanding of what will be covered in this section, the sets listed in the first line of the Algorithm will be reintroduced.

C_1 represents the set in which the procedure assigns the time series with no trend at the first stage;

C_3 represents the set in which the procedure assigns the time series with nonlinear trend at the second stage;

C_2 represents the set in which the procedure assigns all the remaining time series not listed in C_1 and C_3.

Algorithm Classify HD Time Series by Trend

Set $U := \{1, \ldots, p\}$, $C_1 = C_2 = C_3 = \emptyset$
Set the parameters α, s and h_i, $i \in U$
for $i \in U$ **do**
    Perform the "Trend/NoTrend Test Statistic" $\hat{I}_{\beta,i}$
    if $\hat{I}_{\beta,i} < \chi^2_{(1-\alpha/p,k_T)}$ **then**
        Set $C_1 := C_1 \cup \{i\}$
Set $U := U \backslash C_1$
if $U = \emptyset$ **then**
    return C_1, C_2, C_3
else Perform the "Lin/NoLin Statistic" $\hat{I}_{D,i}$, $i \in U$, and sort them as $\hat{I}_{D,\sigma(1)} \geq \cdots \geq \hat{I}_{D,\sigma(p_2)}$
    Set $C_3 := \{\sigma(1), \ldots, \sigma(s)\}$ and $C_2 := U \backslash C_3$
    return C_1, C_2, C_3

In this section the whole procedure is tested. Namely, given the true number of time series with no trend P and its complement $N = p - P$, the False Positive and False Negative Rates

$$FPR = \frac{FP}{N}, \quad FNR = \frac{FN}{P},$$

are calculated for the "Trend/NoTrend" stage, where FP and FN are the number of time series incorrectly labelled to have no trend and to have a trend, respectively. Subsequently, the MMS Ratio is computed for the "Lin/NoLin" stage. In the latter, the MMS (computed in the same way as Section 4.4) is divided by the number of true positives time series with nonlinear trend which is corrected to take into account that the procedure has two stages in which the outcome of the second depends on the first. More precisely, the true Positives of the second stage P_2 are all those time series which have a nonlinear trend that are not kept in C_1 in the first stage. In formulas

$$\frac{MMS}{P_2} \geq 1 \quad \text{where} \quad P_2 = |NL \cap \overline{C_1}|.$$

The simulated scenarios are very similar to those presented in Section 4.4. In particular, 100 repetitions of the whole procedure using the various composition of the signal together, divided for type of error are created. In the p time series which constitute each repetition, some have no trend, others have a linear trend and still others have a nonlinear trend. The repetitions are performed for three different couples of T and p. Various settings of NP are considered. The number of time series with no trend are 8, 12 and 20 while those with nonlinear trend are 5, 8 and 10 for $p = 20, 30, 50$ cases respectively. For all 100 repetitions, the previous described Ratios are calculated. The best achievable result is to have all the FPR and FNR as close as possible to zero while the MMS Ratio as close as possible to 1. For this last rate, a value greater than 1 means that the $\hat{I}_D$ statistic ranks the remaining true positives P_2 time series not in the first P_2 positions. Consequently the smallest set which contains all the P_2 time series with nonlinear trend has cardinality (i.e. the MMS) greater than P_2.

Table 4.5 reports the mean for the FPR and FNR calculated for the various scenarios, Table 4.6 reports the median for the MMS Ratio. Both tables gives the value of standard deviation in parenthesis.

The results in Table 4.5 suggest a very good performance of the proposed procedure in terms of FPR. In all the scenarios considered the error made by selecting a time series with a nonzero trend at the first stage is always zero. On the other hand, the performance in terms of FNR improves when the number of observations for each time series, increases. The best performances are achieved when the error is generated by a GARCH process.

Table 4.6 highlights that the screening procedure at the second stage

keeps all the time series with nonlinear trend in the first positions of the ranking. Considering also the results in the previous table, since the only error occurs when a time series without a trend is considered as one that has a trend, this does not compromise the results obtained in the second phase.

Note that, in order to compute $\hat{I}_\beta$ for each time series, two optimal bandwidths need to be estimated as showed by (2.34) in section 2.2.1. Namely, one for $\hat{c}(\varepsilon)$ since it requires the estimation of the trend function and the other one for $\hat{I}_\beta$. Both are function of latent quantities: the integrated second and third derivative of the trend function, respectively. To overcome this issues the same approach used in Giordano and Parrella (2019) is used. More precisely, an FFNN estimator is used in order to obtain a plug-in estimator for the optimal bandwidths.

First stage: "Trend/NoTrend"

	NP	$T = 200$ $p = 20$	$T = 500$ $p = 30$	$T = 1000$ $p = 50$
		$ARMA(1,1)$		
FPR	30%	0.00 (0)	0.00 (0)	0.00 (0)
	20%	0.00 (0)	0.00 (0)	0.00 (0)
	10%	0.00 (0)	0.00 (0)	0.00 (0)
FNR	30%	0.17 (0.16)	0.14 (0.11)	0.09 (0.06)
	20%	0.20 (0.16)	0.13 (0.11)	0.08 (0.06)
	10%	0.22 (0.16)	0.14 (0.10)	0.08 (0.06)
		$GARCH(1,1)$		
FPR	30%	0.00 (0)	0.00 (0)	0.00 (0)
	20%	0.00 (0)	0.00 (0)	0.00 (0)
	10%	0.00 (0)	0.00 (0)	0.00 (0)
FNR	30%	0.06 (0.08)	0.04 (0.05)	0.02 (0.03)
	20%	0.06 (0.08)	0.04 (0.06)	0.02 (0.03)
	10%	0.06 (0.08)	0.04 (0.06)	0.02 (0.03)

Table 4.5: Mean of False Positive and False Negative Rates for the "Trend/NoTrend" part with standard deviation in parenthesis. The results are obtained by running the procedure over 100 simulated scenarios and considering $\alpha = 0.05$.

Second stage: "Lin/NoLin"			
NP	$T = 200$	$T = 500$	$T = 1000$
	$p = 20$	$p = 30$	$p = 50$
$ARMA(1,1)$			
30%	1.00 (0)	1.00 (0)	1.00 (0)
20%	1.00 (0)	1.00 (0)	1.00 (0)
10%	1.00 (0)	1.00 (0)	1.00 (0)
$GARCH(1,1)$			
30%	1.00 (0)	1.00 (0)	1.00 (0)
20%	1.00 (0)	1.00 (0)	1.00 (0)
10%	1.00 (0)	1.00 (0)	1.00 (0)

Table 4.6: Median of MMS Ratio for the "Lin/NoLin" part with standard deviation in parenthesis. The results are obtained by running the procedure over 100 simulated scenarios.

Chapter 5

Real data Application

In this chapter an example to illustrate the application of the proposed procedure to classify high-dimensional time series by trend on real data is presented. Before proceeding with the actual application, a summary of the context in which the proposed procedure is applied will be introduced.

5.1 Data description

The problem of interest here is to classify the energy consumption, in kWh, of some London Householders. The problem was proposed on Kaggle platform (https://www.kaggle.com), an online community of data scientists and machine learning practitioners which allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.

Among the many datasets, the one called "Smart meter data from London" was chosen. It is a reorganization of an existing dataset (https://old.datahub.io/dataset/smartmeter-energy-use-data-in-london-households), which contains the energy consumption readings for a sample of 5,567 London Households that took part in the UK Power Networks led Low Carbon London project between November 2011 and February 2014. The data contains also infomations on the ACORN (a segmentation tool which categorises the UK's population into demographic types) classification details that can be found in the website of CACI (https://acorn.caci.co.uk/downloads/Acorn-User-guide.pdf).

The descriptive report gives the following information:

"To better follow the energy consumption, the British government wants energy suppliers to install smart meters in every home in England, Wales and Scotland. There are more than 26 million homes for the energy suppliers to get to, with the goal of every home having a smart meter by 2020. This roll out of meter is lead by the European Union who asked all member governments to look at smart meters as part of measures to upgrade the energy supply and tackle climate change. After an initial study, the British government decided to adopt smart meters as part of their plan to update the ageing energy system. The daily dataset contains 112 block files with the daily information like the number of measures, minimum, maximum, mean, median, sum and standard deviation of the dataset which contains the corresponding block files with the half-hourly smart meter measurement."

Figure 5.1 gives an example of the mentioned mean daily time series belonging to block 107. Each time series refers to an id code reported on the top-left which distinguishes the householders. This means that each time series can be uniquely identified through this code. For example the code "MAC000313" identify the first time series of the block 107. In particular, the time series MAC000641 shows a non linear trend which looks like a sinusoid. On the other hand, the time series MAC000635 seems to have no trend or at most a slight linear downward trend.

5.2 Classification of energy consumption

Now, setting the error of the first type $\alpha = 0.05$ for the first stage of selection, the threshold $s = 20$, which is the cardinality of the set of time series with non linear trend, and by using a FFNN estimator in order to obtain the optimal bandwidths, the procedure is applied to all the 112 blocks by discarding in advance those time series with a length less than 100.

Figure 5.2 gives a sketch of the output of the procedure applied to the first 5 time series of block 107. As previously stated, the time series MAC000635 is classified to have no trend, the MAC000313 seems to have a linear trend while the remaining a non linear one. For completeness, Table 5.1 reports the results obtained for all the time series of the block 107.

Considering now the results obtained for all the blocks, they are summarized in Figure 5.2 which contains the cardinalities of the three classes. In

particular, the first and last blocks are characterized, for example, by having not only the absence of time series without a linear trend but also with a cardinality of the third group (of the time series with a non linear trend) lower than that previously set ($s = 20$). This last peculiarity is due to the nature of the proposed procedure, the screening phase is applied to the residual set of the first selection phase (Trend/NoTrend). It is therefore logical to expect that, in some blocks, the estimated set of time series with a non linear trend could not only have cardinality lower than the predetermined threshold but that it may also contain time series with a linear trend. While this may seem like a disadvantage of the proposed procedure, it undoubtedly gives rise to further improvements such as turning screening into selection.

In order to conclude the analysis, Figure 5.2 shows the distributions of the cardinality of the sets which suggest that the time series with no trend constitute the main part of each block and that the median cardinalities among blocks are 23 for the No Trend, 7 for the Lin Trend and 20 for the No Lin Trend. On the other hand, the total result is of 2526 time series with no trend, 810 with a linear trend and 2219 with a non linear trend.

These results greatly confirm the importance of the proposed procedure. It is now possible to apply one of the many clustering techniques on time series seen in Chapter 1. For example, once the respective trends have been removed from the Lin and No Lin sets, one could conduct a cluster analysis based on the Piccolo-Distance applied to a K-means algorithm for the three sets separately. Such an analysis would be somewhat misleading if conducted without first using the proposed procedure. It could happen that time series with and without trends are included in the same group. In terms of electricity consumption, it would mean including users with quite different behaviours in the same cluster.

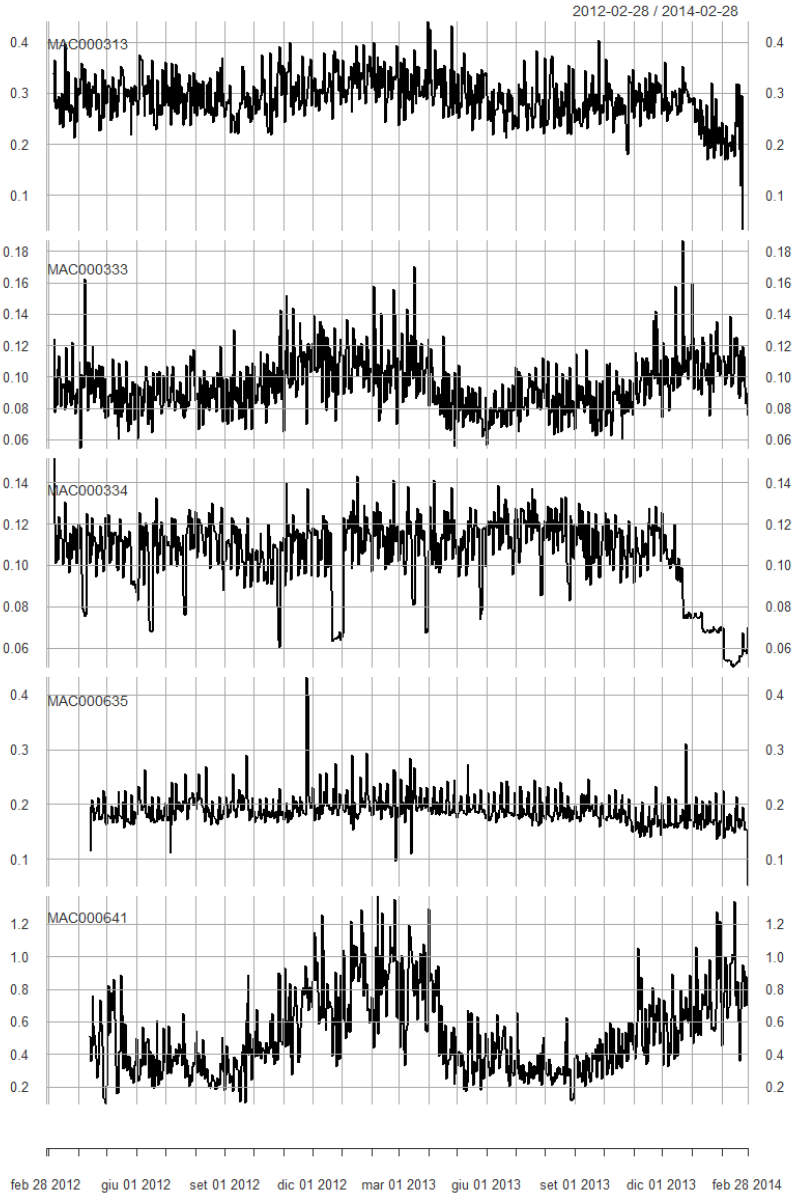

Figure 5.1: First 5 time series belonging to block 107.

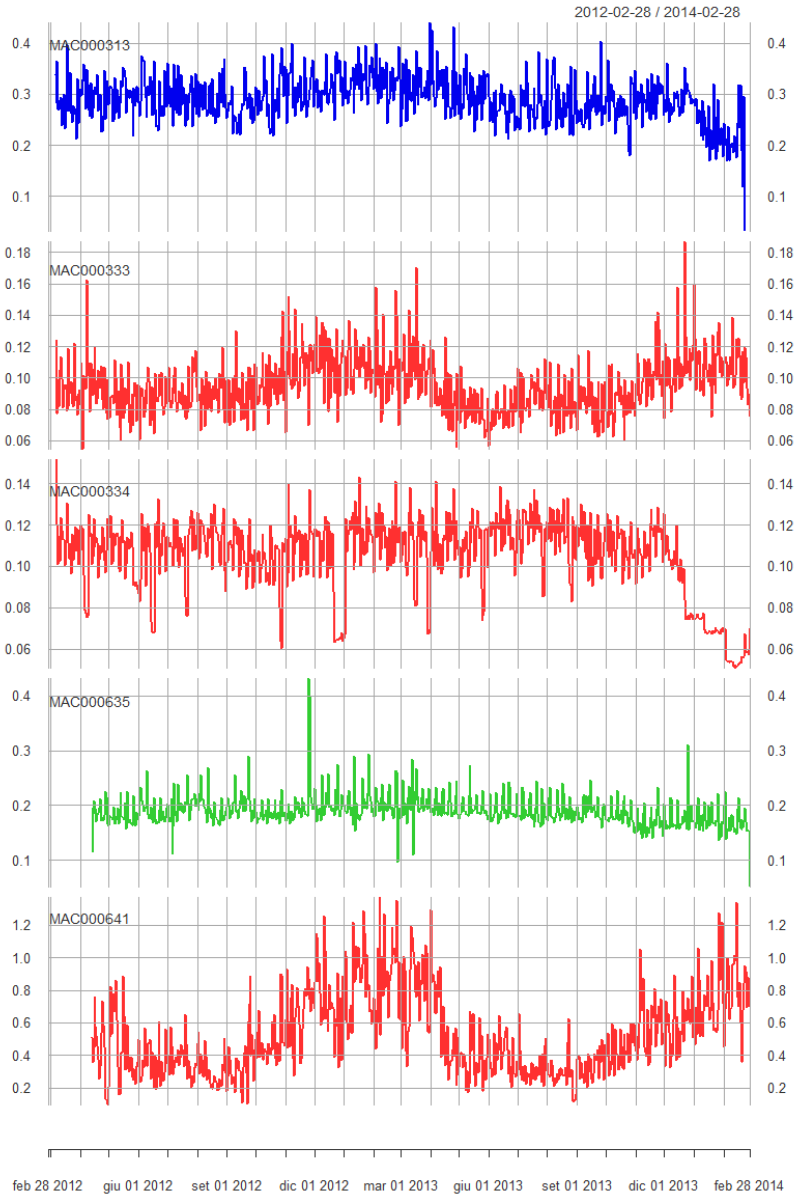

Figure 5.2: First 5 time series belonging to block 107. Each time series is colored according to the output of the proposed procedure: in "green" those with "No Trend", in "blue" those with "Lin Trend" finally in "red" those with "No Lin Trend".

Block	No Trend	Lin Trend	No Lin Trend
107	MAC000635	MAC000313	MAC001811
	MAC001449	MAC002438	MAC001777
	MAC001554	MAC001791	MAC001663
	MAC001648	MAC002427	MAC001715
	MAC001657	MAC002117	MAC000641
	MAC001672	MAC001640	MAC001829
	MAC001673	MAC001509	MAC000334
	MAC001692		MAC002501
	MAC001703		MAC002483
	MAC001714		MAC001744
	MAC001738		MAC001684
	MAC001747		MAC005458
	MAC001773		MAC002475
	MAC001785		MAC005435
	MAC001798		MAC001755
	MAC001824		MAC001817
	MAC001840		MAC001751
	MAC001904		MAC000333
	MAC002419		MAC002426
	MAC002424		MAC001748
	MAC002511		
	MAC005414		
	MAC005452		
♯	23	7	20

Table 5.1: Results of the proposed procedure for classify high-dimensional time series by trend on block 107.

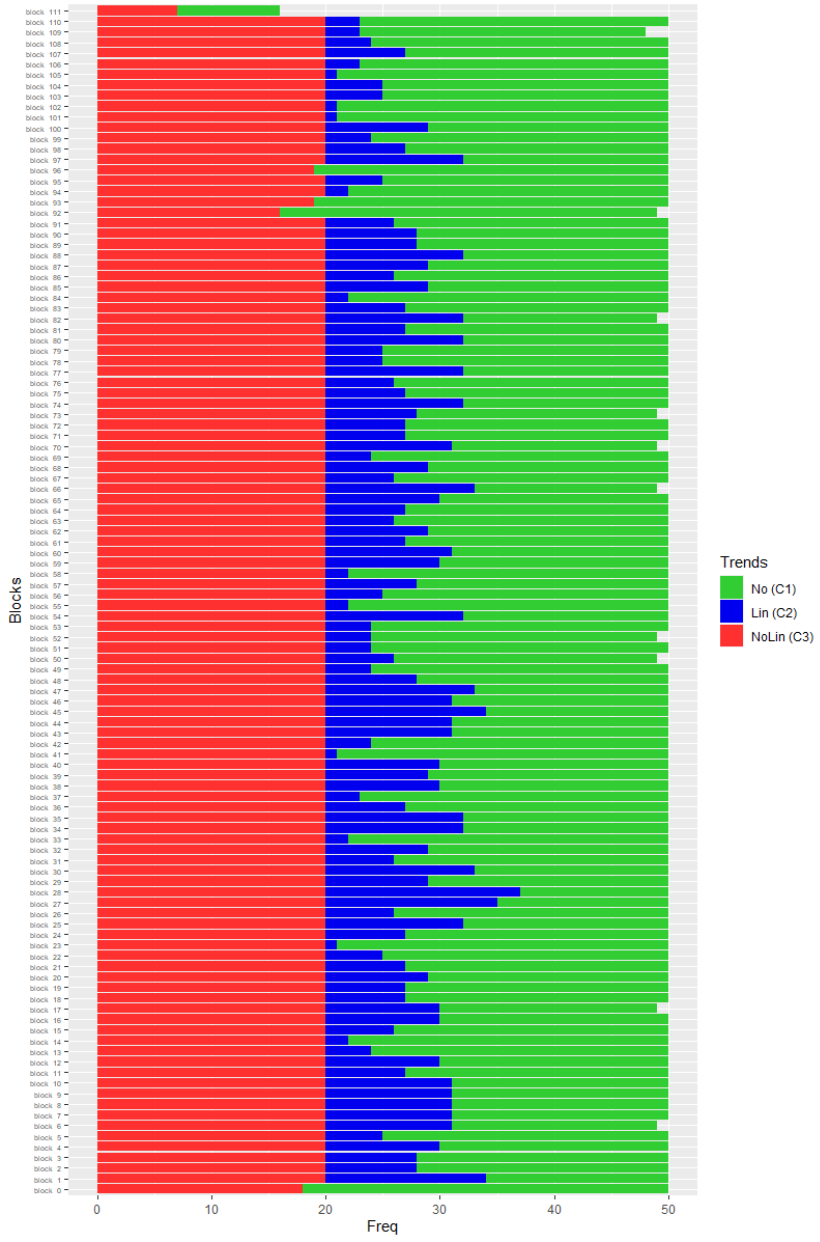

Figure 5.3: Cardinality of the sets obtained by applying the proposed procedure for classify high-dimensional time series by trend to all the blocks.

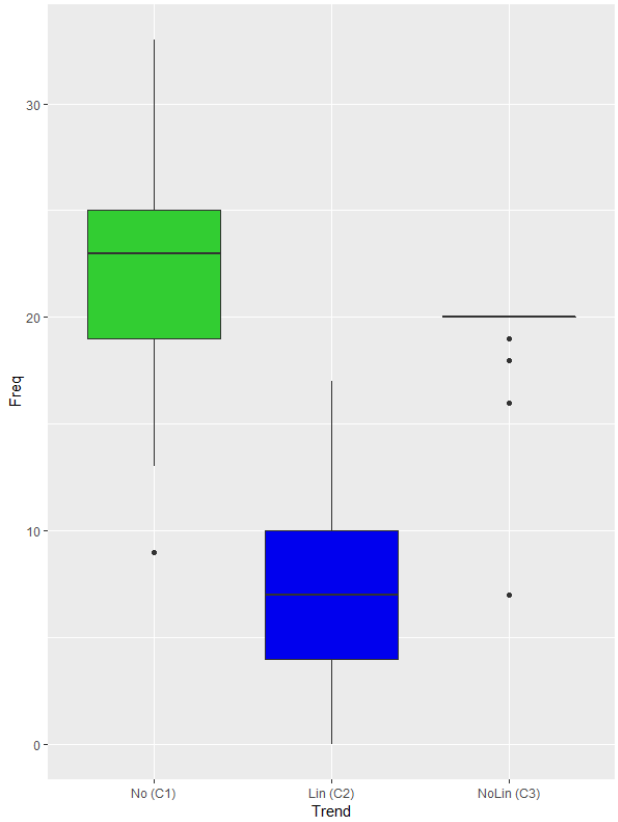

Figure 5.4: Distributions of the cardinality of the sets obtained by applying the proposed procedure for classify high-dimensional time series by trend to all the blocks.

Conclusions

In this thesis a new procedure is presented as an embryonic analysis for carrying out a correct further clustering analysis on time series. It regards the classification of nonstationary time series, where the nonstationarity is given by the presence of a deterministic trend, by looking at the first derivative of the trend in a context of high-dimensionality and without requiring a pre specified form for the trend. This is achieved by using the proposed first derivative trend estimator $\hat{\beta}(x)$ which is based on the Local Polynomial estimator for fixed design and also presents the desirable characteristic of a simple form.

Under the reasonable assumption that the number of time series with nonlinear trend is finite, the proposed partition procedure consists in two stages. In the first one, the proposed estimator is tested to be zero or not, which allows to distinguish the time series with constant trend (no trend). In the second one, the difference between the estimator at different points is used in a screening approach to make the further linear/nonlinear partition of the remaining time series from the previous stage. In other words, the first stage is used to select the time series with constant trend by using a testing procedure while the second one is a screening procedure which gives the set which contains, with probability tending to 1, the true set of time series with nonlinear trend, i.e. it has the Sure Screening property. Furthermore, an Algorithm is given in order to show the easy implementation of the whole procedure.

The use of the mentioned approach presents multiple advantages: (i) on the mathematical point of view, it is quite intuitive the use of the first derivative to highlight the linearity of a function; (ii) one can assert if a trend is linear or not without imposing a predefined mathematical model; (iii) this type of procedure makes a partition of the set of the given time series which may be used in a further analysis as starting point (i.e. it gives

a useful previous knowledge on the trend composition for a deeper clustering analysis); (iv) it does not impose restrictions on the trend composition such as those which are imposed when the presence of parallelism is tested; (v) it gives mathematical guarantees in the high-dimensional setting since it is consistent in the case of $p = o\left(T^{1/2}/\log T\right)$, where p is the number of time series.

The performances of the proposed procedure are studied by an extensive use of Monte Carlo simulations with different scenarios. The performances are checked not only for each part of the procedure but also for the whole procedure. The results obtained highlight that the proposed procedure confirms the theoretical results. An example of application on real data has been proposed to show the actual goodness and necessity of the procedure before applying a cluster analysis on time series.

Future developments regarding the proposed procedure concern the transformation of the second stage into a selection procedure which allows to identify with greater precision the true set of time series with nonlinear trend and the increase of the achievable dimensionality reached by the procedure.

Appendices

Proof of Propositions 1, 2 and 3

Proof of Proposition 1. Sarting with the bias of $\hat{\beta}(x)$, for the linearity of the expected value, it is sufficient to develop the following expected value using: the Riemann sum, the change of variable $z = \frac{u-x}{h}$, Taylor's expansion up to the fourth power, (A1) and finally (A2).

$$\begin{aligned}
E[\hat{\beta}(x)] =& \frac{1}{Th^2} \sum_{t=1}^{T} K_h(t/T - x)\ (t/T - x)\ m(t/T) \\
=& \frac{1}{h^3} \int_0^1 K\left(\frac{u-x}{h}\right) (u-x)\ m(u)\ du + O\left(\frac{1}{Th^2}\right) \\
=& \frac{1}{h} \int_{-1}^1 K\left(z\right)\ z\ m(x+hz)\ dz + O\left(\frac{1}{Th^2}\right) \\
=& m^{(1)}(x) \int_{-1}^1 K\left(z\right)\ z^2\ dz + \frac{m^{(3)}(x)h^2}{6} \int_{-1}^1 K\left(z\right)\ z^4\ dz + O(h^4) + O\left(\frac{1}{Th^2}\right)
\end{aligned}$$

In order to derive the variance it is sufficient to use the Proposition 2 in Francisco-Fernández and Vilar-Fernández (2001) noting that $\hat{\beta}(x) - E[\hat{\beta}(x)] = \frac{1}{h} h^{-j} t_j^*$ with $j = 1$. □

Proof of Proposition 2. The proof follows the approach of Francisco-Fernández and Vilar-Fernández (2001) and Masry and Fan (1997) used to prove the asymptotic normality. Let $X_0 = X - E[X]$, then

$$\begin{aligned}
\hat{\beta}_0(x) =& \frac{1}{Th^2} \sum_{t=1}^{T} K_h(t/T - x)(t/T - x)(Y_t - m(t/T)) \\
=& \frac{1}{Th^2} \sum_{t=1}^{T} K_h(t/T - x)(t/T - x)\varepsilon_t.
\end{aligned} \tag{1}$$

Let Q_T be an arbitrary linear combination of $\hat{\beta}_0(x)$

$$Q_T = \sum_{j=1}^{k_T} a_j \hat{\beta}_0(x_j) \tag{2}$$

where $a_j \in \mathbb{R}$, $j = 1, \ldots, k_T$. Now it remains to prove the asimptotic normality of $\sqrt{Th^3}Q_T$ and subsequently use the Cramer-Wold device. In order to obtain the result it is sufficient to use the small-blocks large-blocks method applying the same steps as in the proof of Proposition 3 of Francisco-Fernández and Vilar-Fernández (2001) noting that, by (1), $E[Q_T] = 0$, while

$$\lim_{T\to\infty} Th^3 VarQ_T = c(\varepsilon) \int C^2(u)du = \sigma_Q^2, \tag{3}$$

and

$$\sqrt{Th^3}Q_T = \frac{h}{\sqrt{T}} \sum_{t=1}^{T} Z_t = \frac{h}{\sqrt{T}} S_T, \tag{4}$$

where

$$Z_t = \sqrt{h} C_h(t/T)\varepsilon_t,$$

with

$$C(u) = \sum_{j=1}^{k_T} a_j (u - x_j) K\left(\frac{u - x_j}{h}\right) \quad \text{and} \quad C_h(u) = \frac{1}{h} C(u).$$

□

Proof of Proposition 3. To prove the first part of the statement it is sufficient to apply the first part of the Proposition 1 for $\hat{\beta}(x_1)$ and $\hat{\beta}(x_2)$, with $x_1 \neq x_2$.

To prove, instead, the second part of the statement, it is useful to proceed first with the proof of the (3.11).

Let $X_0 = X - E[X]$, then it can be defined the following

$$\begin{aligned}\hat{\beta}_0(x) =& \frac{1}{Th^2} \sum_{t=1}^{T} K_h(t/T - x)(t/T - x)(Y_t - m(t/T)) \\ =& \frac{1}{Th^2} \sum_{t=1}^{T} K_h(t/T - x)(t/T - x)\varepsilon_t.\end{aligned}$$

In this way $Cov(\hat{\beta}(x_1), \hat{\beta}(x_2)) = E[\hat{\beta}_0(x_1)\hat{\beta}_0(x_2)]$ which gives the easiest form

$$E[\hat{\beta}_0(x_1)\hat{\beta}_0(x_2)] = \frac{1}{T^2h^4}\sum_{i,t=1}^{T} K_h(i/T - x_1)(i/T - x_1)K_h(t/T - x_2)(t/T - x_2)E\varepsilon_i\varepsilon_t. \tag{5}$$

From the latter, by (A1)-(A3), it can be obtained, using the same arguments as the proof of Proposition 2 in Francisco-Fernández and Vilar-Fernández (2001) the following decomposition

$$\begin{aligned}
E[\hat{\beta}_0(x_1)\hat{\beta}_0(x_2)] =& \frac{\sigma^2}{T^2h^4}\sum_{i,t=1}^{T} K_h(i/T - x_1)(i/T - x_1)K_h(t/T - x_2)(t/T - x_2)c(|i-t|) \\
=& \frac{\sigma^2}{T^2h^4}\sum_{i=1}^{T} K_h(i/T - x_1)(i/T - x_1)K_h(i/T - x_2)(i/T - x_2)\sum_{t=1}^{T} c(|i-t|) \\
&+ \frac{\sigma^2}{T^3h^5}\sum_{i=1}^{T} K_h(i/T - x_1)(i/T - x_1)K_h(i/T - x_2)\sum_{t=1}^{T}(t-i)c(|i-t|) \\
&+ \frac{\sigma^2}{T^3h^5}\sum_{i=1}^{T} K_h(i/T - x_1)(i/T - x_1)(i/T - x_2)\sum_{t=1}^{T}(t-i)K^*(i,t)c(|i-t|) \\
&+ \frac{\sigma^2}{T^3h^6}\sum_{i=1}^{T} K_h(i/T - x_1)(i/T - x_1)\sum_{t=1}^{T}\frac{(t-i)^2}{T}K^*(i,t)c(|i-t|) \\
=& \Delta_1 + \Delta_2 + \Delta_3 + \Delta_4,
\end{aligned} \tag{6}$$

where $K^*(i,t) = \int_0^1 K^{(1)}\left(\frac{i/T - x_2 + y(t/T - i/T)}{h}\right)dy$. Using the Riemann sum, the chnge of variable $z = \frac{u - x_1}{h}$ and remembering (A1) and (A3),

$$|\Delta_1| \leq \frac{\sigma^2 c_1}{Th^3}\int_{-1}^{1} K(z)|z|K\left(\frac{x_1 - x_2}{h} + z\right)\left|\frac{x_1 - x_2}{h} + z\right|dz + O\left(\frac{1}{T^2h^5}\right),$$

where $c_1 = 2\sum_{k=1}^{\infty}|c(k)|$. Noting that, given $z \in [1-, 1]$,

$$K\left(\frac{x_1 - x_2}{h} + z\right)\left|\frac{x_1 - x_2}{h} + z\right| = o(1)$$

and by using the same approach for Δ_2, Δ_3 and Δ_4, the proof is completed.

In order to prove the (3.10) it is sufficient to apply the second part of the Proposition 1 for $\hat{\beta}(x)$ noting that it does not depend on a precise point of the support and by using the (3.11). □

Proof of Theorems 1, 2 and 3

Proof of Theorem 1. The proof is conducted noting that, by Proposition 2, it is possible to restate Q as a quadratic form

$$Q = X'X, \qquad \text{where } X = \left[\hat{\beta}(x_j)\right]_{j=1,\dots,k}.$$

Applying the Continuous Mapping theorem to the quadratic form defined above and by using the results in Section 3.5 of Serfling (2009), the proof is concluded. □

Proof of Theorem 2. In order to prove the theorem it is sufficient to consider $Z_T^2 = \frac{T^{4/7}}{\mu_2^* c(\varepsilon)}\hat{\beta}(x)^2$ and show that

$$P\left(Z_T^2 > 2\log T\right) = O\left(T^{-1/2}\log T\right) \qquad \text{if the time series has constant trend,}$$
$$P\left(Z_T^2 < 2\log T\right) = O\left(T^{-4/7}\right) \qquad \text{otherwise.}$$

The first part can be proved by using the Berry-Essen theorem for strong-mixing processes (see Chapter 7 of Zhengyan and Chuanrong (1997)). Noting that

$$P\left(|Z_T| > \sqrt{2\log T}\right) \leq 2P\left(Z_T > \sqrt{2\log T}\right), \tag{7}$$

and that, under the constant trend assumption, (i) $E[Z_T] = 0$ by (3.2) and (A1). Furthermore, (ii) by (A4) there exists a $\delta > 2$ such that $E|\varepsilon_t|^{2+\delta} < \infty$, (iii) by assumption the mixing coefficients decrease exponentially. With this in mind and by Remark 7.1.1 in Zhengyan and Chuanrong (1997),

$$\Delta_T = |F_T(x) - \Phi(x)| = O\left(T^{-1/2}\log T\right), \tag{8}$$

where $F_T(x) = P\left(Z_T < x\right)$. Now,

$$\begin{aligned} P\left(Z_T > x\right) &= \left|P\left(Z_T > x\right) - \left(1 - \Phi(x)\right) + \left(1 - \Phi(x)\right)\right| \\ &\leq \Delta_T + |1 - \Phi(x)| \\ &= O\left(T^{-1/2}\log T\right), \end{aligned}$$

since $1-\Phi(x)=o\left(T^{-1}\right)$.

For the second part, note that, under the non constant trend assumption, Proposition 1 and (A6),

$$E[Z_T]=T^{2/7}c, \quad \text{with } c\neq 0.$$

Since,

$$P\left(Z_T^2<2\log T\right)=P\left(-\sqrt{2\log T}-E[Z_T]<Z_T-E[Z_T]<\sqrt{2\log T}-E[Z_T]\right),$$

it is sufficient to prove that

$$P\left(|Z_T-EZ_T|>|\sqrt{2\log T}-E[Z_T]|\right)=O\left(T^{-4/7}\right)$$

to complete the proof. Now, using Chebyshev's inequality, Proposition 1 and (A6),

$$P\left(|Z_T-EZ_T|>|\sqrt{2\log T}-E[Z_T]|\right)\leq\frac{1}{c^2T^{4/7}}T^{4/7}Var[\hat{\beta}(x)],$$

the latter is obtained.

□

Proof of Theorem 3. Using Markov's inequality, Proposition 3 and (A6), under the linear trend assumption

$$\begin{aligned}P\left(\hat{I}_D(x)>\eta\right)=&P\left(\frac{1}{k_T}\sum_{j=1}^{k_T}\hat{D}(x,x_j)^2>\eta\right)\\ \leq&\frac{1}{\eta k_T}\sum_{j=1}^{k_T}E[\hat{D}(x,x_j)^2]=O\left(T^{-4/7}\right),\end{aligned}$$

the first part is proved.

For the second part, note that, under the nonlinear trend assumption and by Proposition 3,

$$\hat{I}_D\xrightarrow{P}c^*>0, \quad \text{where } c^*=\mu_2^2\int_0^1\Delta^{(1)}(x,y)^2dy.$$

Now, choosing $0<\eta<c^*$, it follows that

$$P\left(\hat{I}_D<\eta\right)=o(1)$$

which completes the proof. □

Bibliography

Abdulla, W. H., Chow, D., and Sin, G. (2003). Cross-words reference template for dtw-based speech recognition systems. In *TENCON 2003. Conference on convergent technologies for Asia-Pacific region*, volume 4, pages 1576–1579. IEEE.

Aghabozorgi, S., Shirkhorshidi, A. S., and Wah, T. Y. (2015). Time-series clustering–a decade review. *Information Systems*, 53:16–38.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Amari, S.-i. (2007). Integration of stochastic models by minimizing α-divergence. *Neural computation*, 19(10):2780–2796.

Amigó, E., Gonzalo, J., Artiles, J., and Verdejo, F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information retrieval*, 12(4):461–486.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., PéRez, J. M., and Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46(1):243–256.

Asadi, N., Mirzaei, A., and Haghshenas, E. (2016). Creating discriminative models for time series classification and clustering by hmm ensembles. *IEEE transactions on cybernetics*, 46(12):2899–2910.

Batista, G. E., Wang, X., and Keogh, E. J. (2011). A complexity-invariant distance measure for time series. In *Proceedings of the 2011 SIAM international conference on data mining*, pages 699–710. SIAM.

Bendat, J. S. and Piersol, A. G. (2011). *Random data: analysis and measurement procedures*, volume 729. John Wiley & Sons.

Bickel, P., Götze, F., and vanZwet, W. (1997). Resampling fewer than n observations: Gains, losses, and remedies for losses. *STATISTICA SINICA*, 7(1).

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *The annals of statistics*, pages 1196–1217.

Biernacki, C. (2017). Mixture models. In Droesbeke, J.-J., Saporta, G., and Thomas-Agnan, C., editors, *Choix de modèles et agrégation*. Technip.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, 22(7):719–725.

Brandmaier, A. M. (2011). Permutation distribution clustering and structural equation model trees.

Brockwell, P. J., Davis, R. A., and Fienberg, S. E. (1991). *Time series: theory and methods: theory and methods*. Springer Science & Business Media.

Bühlmann, P. et al. (1997). Sieve bootstrap for time series. *Bernoulli*, 3(2):123–148.

Caiado, J., Crato, N., and Peña, D. (2006). A periodogram-based metric for time series classification. *Computational Statistics & Data Analysis*, 50(10):2668–2684.

Candes, E., Tao, T., et al. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The annals of Statistics*, 35(6):2313–2351.

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.

Chen, L. and Wu, W. B. (2019). Testing for trends in high-dimensional time series. *Journal of the American Statistical Association*, 114(526):869–881.

Chouakria, A. D. and Nagabhushan, P. N. (2007). Adaptive dissimilarity index for measuring time series proximity. *Advances in Data Analysis and Classification*, 1(1):5–21.

Das, G., Gunopulos, D., and Mannila, H. (1997). Finding similar time series. In *European Symposium on Principles of Data Mining and Knowledge Discovery*, pages 88–100. Springer.

De Lucas, D. C. (2010). *Classification techniques for time series and functional data*. PhD thesis, Universidad Carlos III de Madrid.

Degras, D., Xu, Z., Zhang, T., and Wu, W. B. (2011). Testing for parallelism among trends in multiple time series. *IEEE Transactions on Signal Processing*, 60(3):1087–1097.

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.

Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., and Keogh, E. (2008). Querying and mining of time series data: experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment*, 1(2):1542–1552.

Doukhan, P. (1994). *Mixing: Properties and Examples*. Lecture Notes in Statistics 85. Springer-Verlag New York, 1 edition.

Draghicescu, D., Guillas, S., and Wu, W. B. (2009). Quantile curve estimation and visualization for nonstationary time series. *Journal of Computational and Graphical Statistics*, 18(1):1–20.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, volume 66. CRC Press.

Fan, J. and Kreutzberger, E. (1998). Automatic local smoothing for spectral density estimation. *Scandinavian Journal of Statistics*, 25(2):359–369.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Fan, J. and Yao, Q. (2008). *Nonlinear time series: nonparametric and parametric methods*. Springer Science & Business Media.

Fan, J. and Zhang, W. (2004). Generalised likelihood ratio tests for spectral density. *Biometrika*, 91(1):195–209.

Francisco-Fernández, M. and Vilar-Fernández, J. M. (2001). Local polynomial regression estimation with correlated errors. *Communications in Statistics-Theory and Methods*, 30(7):1271–1293.

Fréchet, M. M. (1906). Sur quelques points du calcul fonctionnel. *Rendiconti del Circolo Matematico di Palermo (1884-1940)*, 22(1):1–72.

Fu, T.-c. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164–181.

Galeano, P. and Peña, D. (2001). Multivariate analysis in vector time series.

Giordano, F. and Parrella, M. L. (2019). Efficient nonparametric estimation and inference for the volatility function. *Statistics*, 53(4):770–791.

Golay, X., Kollias, S., Stoll, G., Meier, D., Valavanis, A., and Boesiger, P. (1998). A new correlation-based fuzzy logic clustering algorithm for fmri. *Magnetic resonance in medicine*, 40(2):249–260.

Grün, B. (2018). Model-based clustering. *Handbook of mixture analysis*, pages 163–198.

Guijo-Rubio, D., Durán-Rosal, A. M., Gutiérrez, P. A., Troncoso, A., and Hervás-Martínez, C. (2020). Time-series clustering based on the characterization of segment typologies. *IEEE Transactions on Cybernetics*.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of intelligent information systems*, 17(2-3):107–145.

Hall, P. et al. (1984). Integrated square error properties of kernel estimators of regression functions. *The Annals of Statistics*, 12(1):241–260.

Han, F., Lu, H., and Liu, H. (2015). A direct estimation of high dimensional stationary vector autoregressions. *The Journal of Machine Learning Research*, 16(1):3115–3150.

Härdle, W. (1990). *Applied nonparametric regression*. Number 19. Cambridge university press.

Hong, D., Gu, Q., and Whitehouse, K. (2017). High-dimensional time series clustering via cross-predictability. In *Artificial Intelligence and Statistics*, pages 642–651.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):193–218.

Ioannides, D. A. (1992). Integrated square error of nonparametric estimators of regression function: The fixed design case. *Statistics & probability letters*, 15(2):85–94.

Kakizawa, Y., Shumway, R. H., and Taniguchi, M. (1998). Discrimination and clustering for multivariate time series. *Journal of the American Statistical Association*, 93(441):328–340.

Kalpakis, K., Gada, D., and Puttagunta, V. (2001). Distance measures for effective clustering of arima time-series. In *Proceedings 2001 IEEE international conference on data mining*, pages 273–280. IEEE.

Keogh, E., Lonardi, S., Ratanamahatana, C. A., Wei, L., Lee, S.-H., and Handley, J. (2007). Compression-based data mining of sequential data. *Data Mining and Knowledge Discovery*, 14(1):99–129.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480.

Kullback, S. (1997). *Information theory and statistics*. Courier Corporation.

Kumar, M., Patel, N. R., and Woo, J. (2002). Clustering seasonality patterns in the presence of errors. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 557–563.

Latecki, L. J., Megalooikonomou, V., Wang, Q., Lakaemper, R., Ratanamahatana, C. A., and Keogh, E. (2005). Elastic partial matching of time series. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 577–584. Springer.

Liao, T. W. (2005). Clustering of time series data—a survey. *Pattern recognition*, 38(11):1857–1874.

Lyubchich, V. and Gel, Y. R. (2016). A local factor nonparametric test for trend synchronism in multiple time series. *Journal of Multivariate Analysis*, 150:91–104.

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).

Malsiner-Walli, G., Frühwirth-Schnatter, S., and Grün, B. (2016). Model-based clustering based on sparse finite gaussian mixtures. *Statistics and computing*, 26(1-2):303–324.

Manteiga, W. G. and Fernandez, J. V. (1995). Testing linear regression models using non-parametric regression estimators when errors are non-independent. *Computational Statistics & Data Analysis*, 20(5):521–541.

Masry, E. and Fan, J. (1997). Local polynomial estimation of regression functions for mixing processes. *Scandinavian Journal of Statistics*, 24(2):165–179.

McDowell, I. C., Manandhar, D., Vockley, C. M., Schmid, A. K., Reddy, T. E., and Engelhardt, B. E. (2018). Clustering gene expression time series data using an infinite gaussian process mixture model. *PLoS computational biology*, 14(1):e1005896.

Mitsa, T. (2010). *Temporal data mining*. CRC Press.

Möller-Levet, C. S., Klawonn, F., Cho, K.-H., and Wolkenhauer, O. (2003). Fuzzy clustering of short time-series and unevenly distributed sampling points. In *International symposium on intelligent data analysis*, pages 330–340. Springer.

Montero, P., Vilar, J. A., et al. (2014). Tsclust: An r package for time series clustering. *Journal of Statistical Software*, 62(1):1–43.

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856.

Paparrizos, J. and Gravano, L. (2015). k-shape: Efficient and accurate clustering of time series. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 1855–1870.

Percival, D. B. and Walden, A. T. (2000). *Wavelet methods for time series analysis*, volume 4. Cambridge university press.

Petitjean, F., Ketterlin, A., and Gançarski, P. (2011). A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, 44(3):678–693.

Piccolo, D. (1990). A distance measure for classifying arima models. *Journal of Time Series Analysis*, 11(2):153–164.

Priestley, M. (1981). *Spectral Analysis and Time Series*. Number v. 1-2 in Probability and mathematical statistics : a series of monographs and texbooks. Academic Press.

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, 90(432):1257–1270.

Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Serfling, R. J. (2009). *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons.

Sheikholeslami, G., Chatterjee, S., and Zhang, A. (1998). Wavecluster: A multi-resolution clustering approach for very large spatial databases. In *VLDB*, volume 98, pages 428–439.

Shumway, R. and Unger, A. (1974). Linear discriminant functions for stationary time series. *Journal of the American Statistical Association*, 69(348):948–956.

Studholme, C., Hill, D. L., and Hawkes, D. J. (1999). An overlap invariant entropy measure of 3d medical image alignment. *Pattern recognition*, 32(1):71–86.

Vilar, J. A. and Pértega, S. (2004). Discriminant and cluster analysis for gaussian stationary processes: Local linear fitting approach. *Journal of Nonparametric Statistics*, 16(3-4):443–462.

Vilar-Fernández, J. M., Vilar-Fernández, J. A., and González-Manteiga, W. (2007). Bootstrap tests for nonparametric comparison of regression curves with dependent errors. *Test*, 16(1):123–144.

Wand, M. P. and Jones, M. C. (1994). *Kernel smoothing*. Crc Press.

Wang, L., Akritas, M. G., and Van Keilegom, I. (2008). An anova-type nonparametric diagnostic test for heteroscedastic regression models. *Journal of Nonparametric Statistics*, 20(5):365–382.

Wang, W., Yang, J., Muntz, R., et al. (1997). Sting: A statistical information grid approach to spatial data mining. In *VLDB*, volume 97, pages 186–195.

Zhang, H., Ho, T. B., Zhang, Y., and Lin, M.-S. (2006). Unsupervised feature extraction for time series clustering using orthogonal wavelet transform. *Informatica*, 30(3).

Zhang, T. (2013). Clustering high-dimensional time series based on parallelism. *Journal of the American Statistical Association*, 108(502):577–588.

Zhengyan, L. and Chuanrong, L. (1997). *Limit theory for mixing dependent random variables*, volume 378. Springer Science & Business Media.