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
Cilco: XXXIII
Curriculum: Metodi Statistici

Summary of the thesis
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

Anno Accademico 2020–2021

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. 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. 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.

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 non-linear 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.

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

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

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 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

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

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 to be proportional to the real first derivative since its bias depends only on 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 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. 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 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 and shows the easy implementation of the whole procedure.

Simulations studies with different settings for the error term, T and h are conducted in order to check not only each part of the procedure individually but also the whole procedure. The results obtained confirm what has been theoretically proved for the two-stage procedure.

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

Finally, an example of application on real data ("Smart meter data from London" available at `https://www.kaggle.com`) has been proposed to show the actual goodness and necessity of the procedure before applying a cluster analysis on time series.

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 in which the number of time series tends to infinity, that is $p = o\left(T^{1/2}/\log T\right)$.

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.