New methodological contributions in time series clustering

Borja Raúl Lafuente Rego

Tese de doutoramento UDC/2017

New methodological contributions in time series clustering

Autor: Borja Raúl Lafuente Rego

Tese de doutoramento UDC/2017

Director:

José Antonio Vilar Fernández

Departamento de Matemáticas

UNIVERSIDADE DA CORUÑA

El abajo firmante hace constar que es el director de la Tesis Doctoral titulada "New methodological contributions in time series clustering", realizada por Borja Raul Lafuente Rego en la Universidade da Coruña (Departamento de Matemáticas) en el marco del programa interuniversitario (UDC, USC y UVigo) de doctorado en Estadística e Investigación Operativa, dando su consentimiento para que su autor proceda a su presentación y posterior defensa.

O abaixo asinante fai constar que é o director da Tese Doutoral titulada "New methodological contributions in time series clustering", desenvolta por Borja Raul Lafuente Rego na Universidade da Coruña no marco do programa interuniversitario (UDC, USC e UVigo) de doutoramento en Estatística e Investigación de Operacións, dando o seu consentimento para que o seu autor proceda a súa presentación e posterior defensa.

A Coruña, 24 de abril de 2017.

Director:

Dr. José Antonio Vilar

Doctorando:

Borja Raul Lafuente Rego

A mi hermana.

Agradecimientos

En primer lugar desearía expresar mi gratitud al director de esta tesis, el profesor José Antonio Vilar Fernández, por su apoyo y dedicación sin la cual este trabajo no habría salido adelante. Ha sido todo un privilegio poder contar con su guía y ayuda.

Agradecer a todos los profesores del departamento de Matemáticas por su ayuda y consejos durante todos estos años.

También quisiera agradecer a todos mis compañeros del laboratorio 2.1 muy especialmente a mi amigo Miguel, por su compañerismo y atención, y por hacer mucho más divertidas las jornadas de trabajo, y a Bea que, a pesar de llevar menos tiempo, me ha apoyado en todo momento. También dar las gracias a todos los compañeros que han pasado por esos largos cafés en la cafetería de la facultad.

A toda mi familia, especialmente a mis padres y mi hermana por su comprensión, paciencia y cariño.

Dar las gracias a Eloy, Nuria, Raquel y Sebas por su amistad a lo largo de todos estos años y por prestarme su apoyo, especialmente en los momentos de más agobio.

Quiero hacer constar mi agradecimiento a la Xunta de Galicia y al Ministerio de Ciencia e Innovación por su apoyo financiero a través de los proyectos 2012/130, MTM2011-22392 y MTM2014-52876-R.

Abstract

This thesis presents new procedures to address the analysis cluster of time series. First of all a two-stage procedure based on comparing frequencies and magnitudes of the absolute maxima of the spectral densities is proposed. Assuming that the clustering purpose is to group series according to the underlying dependence structures, a detailed study of the behavior in clustering of a dissimilarity based on comparing estimated quantile autocovariance functions (QAF) is also carried out. A prediction-based resampling algorithm proposed by Dudoit and Fridlyand is adjusted to select the optimal number of clusters. The asymptotic behavior of the sample quantile autocovariances is studied and an algorithm to determine optimal combinations of lags and pairs of quantile levels to perform clustering is introduced. The proposed metric is used to perform hard and soft partitioning-based clustering. First, a broad simulation study examines the behavior of the proposed metric in crisp clustering using hierarchical and PAM procedure. Then, a novel fuzzy C-medoids algorithm based on the QAF-dissimilarity is proposed. Three different robust versions of this fuzzy algorithm are also presented to deal with data containing outlier time series. Finally, other ways of soft clustering analysis are explored, namely probabilistic D-clustering and clustering based on mixture models.

Resumo

Esta tese presenta novos procedementos para abordar a análise cluster de series temporais. En primeiro lugar proponse un procedemento en duas etapas basado na comparación de frecuencias y magnitudes dos máximos absolutos das densidades espectrais. Supoñendo que o propósito é agrupar series de acordo coas estruturas de dependencia subxacentes, tamén se leva a cabo un estudo detallado do comportamento en clustering dunha disimilaridade baseada na comparación deas funcións estimadas das autocovariancias cuantil (QAF). Un algoritmo de remostraxe baseado na predición proposto por Dudoit e Fridlyand adáptase para seleccionar o número óptimo de clusters. Tamén se estuda o comportamento asintótico das autocovariancias cuantís e se introduce un algoritmo para determinar as combinacións óptimas de lags e pares de niveles de cuantís para levar a cabo a clasificación. A métrica proposta utilizase para realizar análise cluster baseado en particións "hard" e "soft". En primeiro lugar, un amplo estudo de simulación examina o comportamento da métrica proposta en clúster "hard" utilizando os procedementos xerárquico e PAM. A continuación, proponse un novo algoritmo "fuzzy" C-medoides baseado na disimilaridade QAF. Tamén se presentan tres versións robustas deste algoritmo "fuzzy" para tratar con datos que conteñan atípicos. Finalmente, se exploranse outras vías de análise cluster "soft", concretamente, D-clustering probabilístico e clustering baseado en modelos mixtos.

Resumen

Esta tesis presenta nuevos procedimientos para abordar el análisis cluster de series temporales. En primer lugar se propone un procedimiento en dos etapas basado en la comparación de frecuencias y magnitudes de los máximos absolutos de las densidades espectrales. Suponiendo que el propósito es agrupar series de acuerdo con las estructuras de dependencia subyacentes, también se lleva a cabo un estudio detallado del comportamiento en clustering de una disimilaridad basada en la comparación de las funciones estimadas de las autocovariancias cuantil (QAF). Un algoritmo de remuestreo basado en predicción propuesto por Dudoit y Fridlyand se adapta para seleccionar el número óptimo de clusters. También se estudia el comportamiento asintótico de las autocovariancias cuantiles y se introduce un algoritmo para determinar las combinaciones óptimas de lags y pares de niveles de cuantiles para llevar a cabo la clasificación. La métrica propuesta se utiliza para realizar análisis cluster basado en particiones "hard" y "soft". En primer lugar, un amplio estudio de simulación examina el comportamiento de la métrica propuesta en clúster "hard" utilizando los procedimientos jerárquico y PAM. A continuación, se propone un nuevo algoritmo "fuzzy" C-medoides basado en la disimilaridad QAF. También se presentan tres versiones robustas de este algoritmo "fuzzy" para tratar con datos que contengan atípicos. Finalmente, se exploran otras vías de análisis cluster "soft", concretamente, D-clustering probabilístico y clustering basado en modelos mixtos.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Contents

1.1 Clustering of time series: An introduction

Time series clustering is aimed to split a set of partial realizations of time series into different categories or clusters. Partition is performed in such a way that series in the same cluster are more similar to each other than series in different clusters. Time series clustering is a central problem in many application fields and it is nowadays an active research area in a vast range of fields such as finance and economics, medicine, engineering, physics, pattern recognition, among many others. These arguments account for the growing interest on this topic, which has resulted in a huge number of contributions. Some illustrative examples of these applications are: classification of industrial production series (Piccolo, 1990; Corduas and Piccolo, 2008), comparison of seismological data as in the classical case of distinguishing between earthquake and nuclear explosion waveforms (Kakizawa et al., 1998), clustering of ecological dynamics (Li et al., 2001), comparison of daily hydrological time series (Grimaldi, 2004), clustering of industrialized countries according to historical data of CO_2 emissions (Alonso et al., 2006), detection of similar immune response behaviors of CD4C cell number progression in patients affected by immunodeficiency virus (HIV) (Chouakria and Nagabhushan, 2007), identification of active genes during the cell division process (Douzal-

Chouakria et al., 2009), classification of chemometrical data (D'Urso and Giovanni, 2014), clustering based on daily nitrogen monoxide emissions (D'Urso et al., 2015a), and analysis of navigation patterns of users visiting news web sites (García-Magariños and Vilar, 2015), among others.

A crucial issue in time series clustering is determining a suitable measure to assess dissimilarity between two time series data. Unlike conventional clustering on static data objects, time series are inherently dynamic, with underlying autocorrelation structures, and therefore the similarity searching must be governed by the behavior of the series over their periods of observation. As an illustrative example, the Euclidean distance treats the observations as if they were independent so that, in particular, it is invariant to permutations over time, and hence it does not take into account the underlying correlation structure. This fact is highlighted in Figure 1.1 with a simple example. Figure 1.1 (a) shows the profiles of realizations simulated from AR(1) (black) and MA(1) (red) processes with parameters $\phi = 0.7$ and $\theta = 0.3$, respectively. An arbitrary permutation of each one of these realizations is depicted in Figure 1.1 (b). By definition, the Euclidean distance assumes the i-th observation in one sequence is aligned with the i-th observation in the other, and therefore the realizations in Figures 1.1 (a) and (b) are separated by the same Euclidean distance (18.65). Nevertheless, one expects that changes and distortions in the temporal behaviors lead to different levels of dissimilarity. This goal can be attained by using distances or dissimilarities regarding the underlying dynamic component. A simple way to tackle this issue is comparing sequences of estimated autocorrelations, which involves information on the lineal dependence structure. In fact, the Euclidean distance between estimated autocorrelations leads to the values 2.55 and 0.57 for the realizations in the left and right panels of Figure 1.1, respectively. In sum, the Euclidean distance between raw data cannot be considered a good measure of dissimilarity between time series data. Overall the choice of a proper dissimilarity measure between time series is a non trivial issue, and a large number of criteria have been proposed in the last two decades. This point is one of the challenges in the current dissertation and its importance motivates the short overview provided in Section 1.2 of this Introduction.

Although selecting a proper metric plays a key role, there are additional difficulties to be addressed in time series clustering. For instance, many clustering applications in real-life involve a huge number of very long series, i.e. one faces the high-dimensionality problem. In fact, the observed time series often contain thousands of data, which in cluster analysis translates into thousands of classification variables. Therefore, algorithms working directly on the raw time series could become inefficient, or simply unfeasible. To overcome the high-dimensionality problem, we will focus throughout the entire thesis on the feature-based approach, where the raw data are replaced by a lower dimension vector of extracted

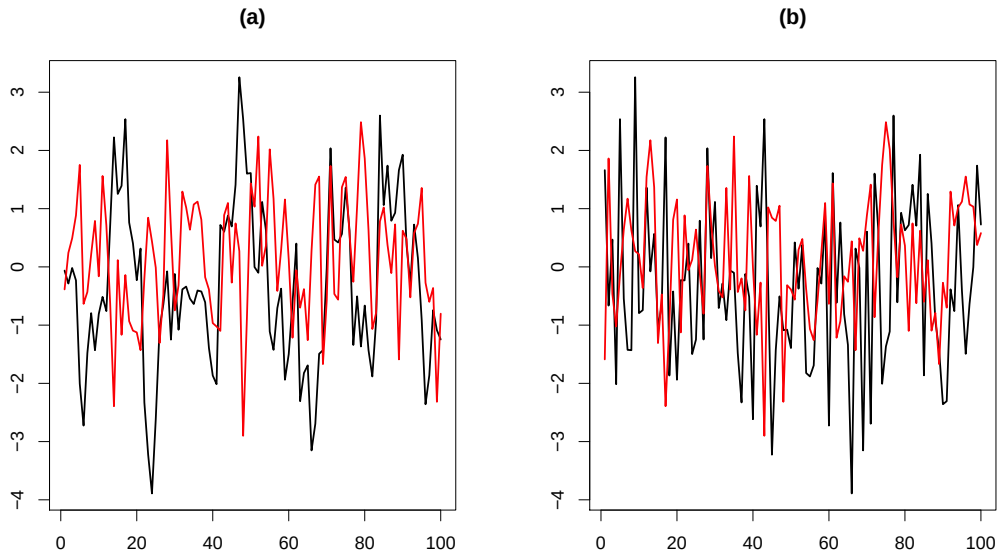

Figure 1.1: Simulated realizations from AR(1) (black) and MA(1) (red) processes (a), and arbitrary permutations of these realizations (b).

features that represent the dynamic structure of each series, thus allowing a dimensionality reduction and a meaningful saving in computation time. This way, dissimilarity between time series is measured in terms of discrepancy between these representations.

Also, when dealing with partitioning-based algorithms, the concept of centroid is particularly complex. As it is well known, the centroids are representative objects for the clusters and sometimes the target of the clustering process is to identify these prototypes rather than performing an accurate classification. In the time series setting, a centroid determines a specific temporal pattern and it is often important to get insight into these patterns in order to carry out predictions or establish differences between behaviors over time. Nevertheless, caution must be taken to properly define the centroid by dealing with time series. For example, the most popular partitioning-based algorithm is the C-means procedure, where the centroids are defined as the average objects within the clusters. Such an approach might generate inconsistencies whether a distance based on feature vectors is used because of the average of a set of features does not necessarily characterizes a time series model, and therefore it cannot be a representative object of the temporal behavior of the cluster. In other words, the centroids could be fictitious time series and thus failing in providing a suitable model of the cluster dynamics. Furthermore, it could be that the distance between a single time series object and the average of the group is not well-defined. A natural way to overcome these drawbacks is to perform a k-medoids-based algorithm where the candidates

have to be selected among the data points. In fact, k-medoids-based procedures will be adopted later in this thesis.

Other points to be considered in cluster analysis of time series are indeed related to the nature of the series in study, the final clustering purpose and the computational complexity of the employed procedures. Certainly, a suitable distance to deal with series generated from the linear models may be inappropriate to face non-linear models, and a cluster algorithm designed to discriminate between stationary processes will hardly be useful to group series showing similar trends. In the big data era, algorithms with a very high computational complexity will be unfeasible to perform clustering on databases including many and very long time series so that the computational efficiency and the capability to manage unbalanced time series are not minor properties.

In summary, the high level of complexity and particularities associated to time series clustering together with its enormous interest in a broad range of applications account for the great focus of attraction that this topic has led over the past decades in research, mainly into the fields of Statistics, Data Mining and Artificial Intelligence. Comprehensive surveys on time series clustering can be seen in Liao (2005), and more currently in Fu (2011). Hence, significant advances have been achieved, but undoubtedly time series clustering is still an active research area nowadays, with serious problems and challenges to address.

This introductory chapter is structured as follows. Given the importance of the dissimilarity notion between time series, this point is widely discussed in Section 1.2, and some popular and commonly used metrics are shortly described. An overview of the thesis highlighting motivation, structure and main contributions is provided in Section 1.3. Some preliminary concepts used throughout the dissertation are presented in Section 1.4.

1.2 Measuring dissimilarity between a pair of time series

A clustering procedure is strongly influenced by the dissimilarity principle inherent to the employed between-objects distance. Hence determining a proper dissimilarity measure between objects is a key issue in cluster analysis, and as mentioned, a particularly sensitive issue by dealing with time series data. Commonly used dissimilarities in conventional cluster ignore the temporal evolution of the series and may produce unsatisfactory results in a time series context. To address this problem, different dissimilarity criteria between series have been introduced in the literature. An overview of these dissimilarities can be seen in Montero and Vilar (2014a). According to the nature of the considered criteria, Montero and Vilar (2014a) classify the dissimilarities in well-defined categories, which are enumer-

ated below in order to shed some light on the most relevant approaches considered in the literature.

One group is formed by the *free-model distances*, mainly including distances between raw observations and those based on comparing features extracted from the original time series. Besides conventional distances like Minkowski and Fréchet distances, this category involves distances properly adjusted to be invariant to specific and typical distortions of temporal data such as local scaling (warping), phase, amplitude scaling, complexity, and so on (Batista et al., 2011). Dynamic time warping (DTW) (Berndt and Clifford, 1994) is surely the most commonly used metric within this distance type. As far as the measures using extracted features, many approaches have been explored, including distances based on comparing autocorrelations (Kovacic, 1998; Struzik and Siebes, 1999; Galeano and Peña, 2000; Caiado et al., 2006; D'Urso and Maharaj, 2009), cross-correlations (Golay et al., 2005; Chouakria and Nagabhushan, 2007), spectral features (Kakizawa et al., 1998; Vilar and Pértega, 2004; Pértega and Vilar, 2010; Casado de Lucas, 2010), wavelet coefficients (Chan and Fu, 1999; Popivanov and Miller, 2002; Chan et al., 2003; Zhang et al., 2006), and symbolic representations such as the SAX representation (symbolic aggregate approximation) (Lin et al., 2003), among others.

Other group involves the *model-based dissimilarities*, which assume specific underlying models and then assessing discrepancy between fitted models. The most common approach consists in assuming that the time series are generated by ARIMA processes (see e.g., Piccolo, 1990; Maharaj, 1996, 2000; Kalpakis et al., 2001, among others) although also alternative structures such as Markov chains (Ramoni et al., 2002) and hidden Markov models (Oates et al., 1999) have been considered.

Other measures are aimed at comparing levels of complexity of the time series, that is the amount of shared information by the two compared series. Two prominent approaches to evaluate complexity differences between a pair of time series are: (i) using algorithms based on data compression (see, e.g., Li et al., 2001; Keogh et al., 2004; Cilibrasi and Vitanyi, 2005; Keogh et al., 2007), and (ii) considering differences between permutation distributions (Brandmaier, 2012). This kind of measures have been intensively studied in Machine Learning and received less attention in the Statistics field.

Although most of the time series dissimilarities can be assigned to one of these three categories, this classification is not exhaustive at all. Sometimes the clustering objective suggests the use of alternative dissimilarities specifically designed to deal with the problem at hand. For instance, treating with time series, it is relatively simple to think on situations where the real interest of the clustering relies on the properties of the predictions at

a pre-specified future time. Note that time series with the same generating process might produce very different forecasts at a given horizon, and therefore a cluster partition generated from model- or feature-based dissimilarities could be inappropriate. Alonso et al. (2006), Vilar et al. (2010) and Vilar et al. (2013) focused on this idea and considered a notion of dissimilarity governed by the performance of future forecasts.

Related to the above consideration and given the broad range of available dissimilarities, other major issue arises in a natural way, namely to decide *which* dissimilarity measure should be used in a particular problem. Montero and Vilar (2014a) argue that this choice must mainly rely on the specific purpose of the clustering task, and only doing so, the cluster solution will admit an interpretation in terms of the grouping target. This argument is congruent with the non-supervised classification paradigm where the perception of a "good" classification could vary across users depending on the pursued target. Montero and Vilar (2014a) highlight this problematic with illustrative and valuable examples. For instance, sometimes the focus is to compare the geometric profiles of the series but in other situations the target is to identify similar generating processes. In the first case, which is quite common by dealing with short series or in situations with a small noise signal ratio, "shape-based" dissimilarities are required, i.e. dissimilarities emphasizing local differences for which conventional distances or complexity-based measures should behave properly. The second case requires "structure-based" dissimilarities aimed at capturing higher-level dynamic structures describing the global performance of the series. Feature- and model-based dissimilarities are expected to report better results in this framework. Nevertheless, the relevant issue is to establish the clustering purpose because the use of different metrics may lead to very different results. Montero and Vilar (2014a) illustrate this fact by using a simple and intuitive synthetic dataset of nine time series generated from three different patterns denoted by P1, P2 and P3. The nine profiles are depicted in Figure 1.2[1]. It is observed in panels (b) and (c) that different cluster partitions are attained when shape- and structure-based metrics are employed. Nevertheless both solutions can be reasonable according the pursued objective.

Establishing innovative time series dissimilarity criteria is one of the topics addressed in this dissertation. The focus is on the structure-based dissimilarities. Overall, this kind of dissimilarities assume regularity conditions for the series subjected to clustering, and users must be aware of it. For example, linearity and homoscedasticity are commonly required. Thus, one of the challenges is to introduce structure-based dissimilarities capable of performing reasonably well under very general conditions, showing robustness to different generating processes and with different distributional forms. To show the capability of

[1] Figure reproduced from Montero and Vilar (2014a) with permission from the authors.

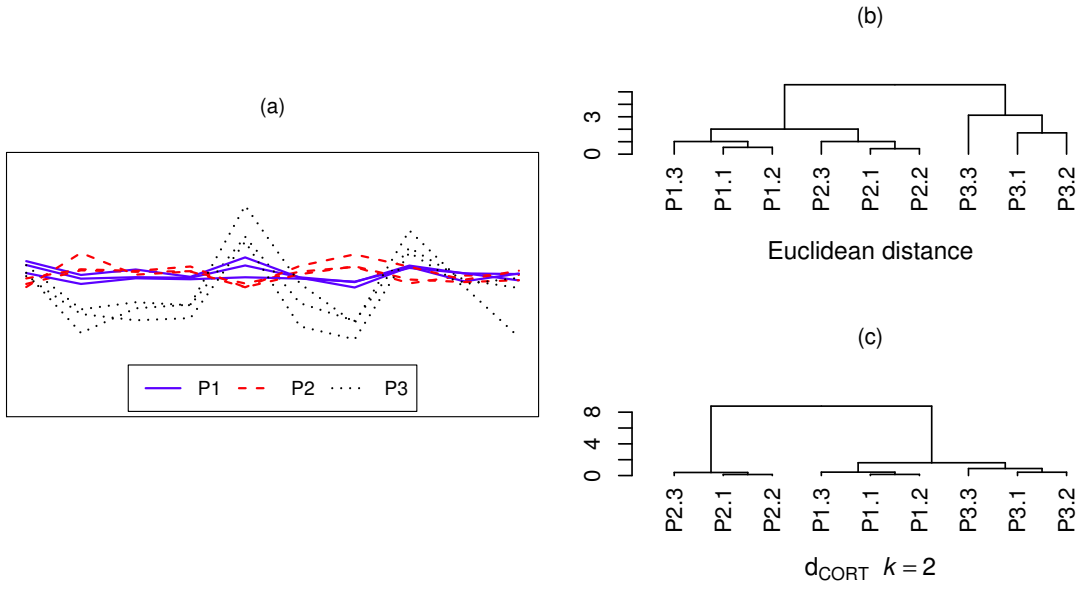

Figure 1.2: Realizations of 9 time series generated from different patterns P1, P2 and P3 (a), and dendrograms from clustering based on the Euclidean distance (shape-based dissimilarity) (b) and on a dissimilarity (d_{CORT} in **TSclust** package) considering temporal correlations (structure-based dissimilarity) (c).

these new approaches, experiments with series simulated from different scenarios will be carried out in order to compare the clustering results using different model-free and model-based metrics. In the following subsections, a brief description of the employed metrics is provided in order to present their main characteristics and also avoiding to introduce them in a reiterative way throughout the entire thesis. It is also worth to point out that a useful tool for practitioners is the R package **TSclust** (Montero and Vilar, 2014b) where most of the metrics enumerated along this section are available.

Herefater, $\boldsymbol{X}_t = (X_1, \ldots, X_T)^t$ and $\boldsymbol{Y}_t = (Y_1, \ldots, Y_T)^t$ denote partial realizations from two real-valued processes $X = \{X_t, t \in \mathbb{Z}\}$ and $Y = \{Y_t, t \in \mathbb{Z}\}$, respectively.

1.2.1 Model-free approaches

A natural approach to measure the dissimilarity between $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ is to replace the observed values by a feature vector of lower dimension and then evaluating a conventional distance between the extracted feature vectors. This intuitive approach presents some nice advantages, including: no assumptions on the generating processes are required, applicability to unbalanced serial realizations, and frequently low computational complexity. The

extracted features can be obtained either the time domain or the frequency domain. Some of the most commonly used dissimilarities belonging to this category are detailed below.

Autocorrelation-based distances

Several authors have considered measures based on the estimated autocorrelation functions (see e.g., Bohte et al., 1980; Galeano and Peña, 2000; Caiado et al., 2006; D'Urso and Maharaj, 2009).

Let $\hat{\boldsymbol{\rho}}_{\boldsymbol{X}_t} = (\hat{\rho}_{1,\boldsymbol{X}_t}, .., \hat{\rho}_{L,\boldsymbol{X}_t})^t$ and $\hat{\boldsymbol{\rho}}_{\boldsymbol{Y}_t} = (\hat{\rho}_{1,\boldsymbol{Y}_t}, .., \hat{\rho}_{L,\boldsymbol{Y}_t})^t$ be the estimated autocorrelation vectors of $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ respectively, for some L such that $\hat{\rho}_{i,\boldsymbol{X}_t} \approx 0$ and $\hat{\rho}_{i,\boldsymbol{Y}_t} \approx 0$ for $i > L$. Galeano and Peña (2000) define a distance between $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ as follows.

$$d_{ACF}\left(\boldsymbol{X}_t, \boldsymbol{Y}_t\right) = \sqrt{\left(\hat{\boldsymbol{\rho}}_{\boldsymbol{X}_t} - \hat{\boldsymbol{\rho}}_{\boldsymbol{Y}_t}\right)^t \boldsymbol{\Omega} \left(\hat{\boldsymbol{\rho}}_{\boldsymbol{X}_t} - \hat{\boldsymbol{\rho}}_{\boldsymbol{Y}_t}\right)},$$

where $\boldsymbol{\Omega}$ is a matrix of weights.

Some common choices of $\boldsymbol{\Omega}$ are:

(i) Consider uniform weights by taking $\boldsymbol{\Omega} = \boldsymbol{I}$. In such case d_{ACF} becomes the Euclidean distance between the estimated autocorrelation functions:

$$d_{ACFU}\left(\boldsymbol{X}_t, \boldsymbol{Y}_t\right) = \sqrt{\sum_{i=1}^{L} \left(\hat{\rho}_{i,\boldsymbol{X}_t} - \hat{\rho}_{i,\boldsymbol{Y}_t}\right)^2}.$$

(ii) Consider geometric weights decaying with the autocorrelation lag, so that d_{ACF} takes the form:

$$d_{ACFG}\left(\boldsymbol{X}_t, \boldsymbol{Y}_t\right) = \sqrt{\sum_{i=1}^{L} p(1-p)^i \left(\hat{\rho}_{i,\boldsymbol{X}_t} - \hat{\rho}_{i,\boldsymbol{Y}_t}\right)^2}, \text{ with } 0 < p < 1.$$

Analogous distances can be constructed by considering the partial autocorrelation functions (PACF's) instead of the ACF's. Hereafter, notation d_{PACFU} and d_{PACFG} will be used to denote the Euclidean distance between the estimated partial autocorrelation coefficients with uniform weights and with geometric weights decaying with the lag, respectively.

As mentioned, the frequency domain can be also used to define dissimilarities between time series. A short overview of the main concepts in spectral analysis of series is provided in Section 1.4.3, including definitions of spectral density and periodogram and some common criteria to build nonparametric estimates of the spectral density. These notions are used

below to introduce alternative dissimilarities between time series.

Periodogram-based distances

Let $\widehat{I}_{\boldsymbol{X}_t}(\lambda_k)$ and $\widehat{I}_{\boldsymbol{Y}_t}(\lambda_k)$ be the estimated periodograms of $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$, respectively, at frequencies $\lambda_k = 2\pi k/T$, $k = 1, \ldots, M$, with $M = [(T-1)/2]$.

Three dissimilarity measures based on periodograms were analyzed by Caiado et al. (2006).

(i) The Euclidean distance between the periodogram ordinates:

$$d_P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t\right) = \frac{1}{n}\sqrt{\sum_{k=1}^{n}\left(\widehat{I}_{\boldsymbol{X}_t}(\lambda_k) - \widehat{I}_{\boldsymbol{Y}_t}(\lambda_k)\right)^2}.$$

(ii) If we are not interested in the process scale but only on its correlation structure, better results can be obtained using the Euclidean distance between the normalized periodogram ordinates:

$$d_{NP}\left(\boldsymbol{X}_t, \boldsymbol{Y}_t\right) = \frac{1}{n}\sqrt{\sum_{k=1}^{n}\left(\widehat{NI}_{\boldsymbol{X}_t}(\lambda_k) - \widehat{NI}_{\boldsymbol{Y}_t}(\lambda_k)\right)^2},$$

where $\widehat{NI}_{\boldsymbol{X}_t}(\lambda_k) = \widehat{I}_{\boldsymbol{X}_t}(\lambda_k)/\widehat{\gamma}_{0,\boldsymbol{X}_t}$ and $\widehat{NI}_{\boldsymbol{Y}_t}(\lambda_k) = \widehat{I}_{\boldsymbol{Y}_t}(\lambda_k)/\widehat{\gamma}_{0,\boldsymbol{Y}_t}$ with $\widehat{\gamma}_{0,\boldsymbol{X}_t}$ and $\widehat{\gamma}_{0,\boldsymbol{Y}_t}$ being the sample variances of $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$, respectively.

(iii) As the variance of the periodogram ordinates is proportional to the spectrum value at the corresponding frequencies, it makes sense to use the logarithm of the normalized periodogram:

$$d_{LNP}\left(\boldsymbol{X}_t, \boldsymbol{Y}_t\right) = \frac{1}{n}\sqrt{\sum_{k=1}^{n}\left(\log\widehat{NI}_{\boldsymbol{X}_t}(\lambda_k) - \log\widehat{NI}_{\boldsymbol{Y}_t}(\lambda_k)\right)^2}.$$

Dissimilarity measures based on nonparametric spectral estimators

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

$$d_W\left(\boldsymbol{X}_t, \boldsymbol{Y}_t\right) = \frac{1}{4\pi}\int_{-\pi}^{\pi} W\left(\frac{f_{\boldsymbol{X}_t}(\lambda)}{f_{\boldsymbol{Y}_t}(\lambda)}\right) d\lambda, \tag{1.1}$$

where $f_{\boldsymbol{X}_t}$ and $f_{\boldsymbol{Y}_t}$ denote the spectral densities of $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$, respectively, and $W(\cdot)$ is a divergence function satisfying appropriate regular conditions to ensure that d_W has the quasi-distance property. Note that d_W is not a real distance because it is not symmetric and does not satisfy the triangle inequality. For clustering, it is more convenient to modify

the divergence function by setting $\tilde{W}(x) = W(x) + W(x^{-1})$.

In practice the spectra $f_{\boldsymbol{X}_t}$ and $f_{\boldsymbol{Y}_t}$ are unknown and has to be estimated. Three different versions of the d_W are obtained depending on how the estimation of this spectrum is carried out:

- $d_{W(LS)}$ when the spectra are replaced by the exponential transformation of local linear smoothers of the log periodograms, via least squares (see (1.20)).
- $d_{W(LK)}$ when the spectra are estimated by the exponential transformation of local linear smoothers of the log periodograms, by using the maximum local likelihood criterion (see (1.22)).
- $d_{W(DLS)}$ when the spectra are estimated by local linear smoothers of the periodogram, via least squares (see (1.23)).

An alternative nonparametric spectral dissimilarity measure introduced by Pértega and Vilar (2010) is also used throughout this thesis. This distance evaluates the integrated squared differences between nonparametric estimators of the log-spectra and it is given by

$$d_{ISD}\left(\boldsymbol{X}_t, \boldsymbol{Y}_t\right) = \int_{-\pi}^{\pi} \left(\hat{m}_{\boldsymbol{X}_t}(\lambda) - \hat{m}_{\boldsymbol{Y}_t}(\lambda)\right)^2 d\lambda,$$

where $\hat{m}_{\boldsymbol{X}_t}(\lambda)$ and $\hat{m}_{\boldsymbol{Y}_t}(\lambda)$ are local linear smoothers of the log-periodograms obtained using the maximum local likelihood criterion.

1.2.2 Model-based approaches

Model-based dissimilarity measures assume that the underlying models are generated from specific parametric structures. The main approach in the literature is to assume that the generating processes of $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ follow invertible ARIMA models. In such case, the idea is fitting an ARIMA model to each series and then measuring the dissimilarity between the fitted models.

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 $\hat{\mathbf{\Pi}}_{\boldsymbol{X}_t} = \left(\hat{\pi}_{1,\boldsymbol{X}_t}, \ldots, \hat{\pi}_{k_1,\boldsymbol{X}_t}\right)^t$ and $\hat{\mathbf{\Pi}}_{\boldsymbol{Y}_t} = (\hat{\pi}_{1,\boldsymbol{Y}_t}, \ldots, \hat{\pi}_{k_2,\boldsymbol{Y}_t})^t$ denote the vectors of AR(k_1) and AR(k_2) parameter estimations for $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$, respectively, then the Piccolo's distance takes the form

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

where $k = \max(k_1, k_2)$, $\hat{\pi}'_{j,\boldsymbol{X}_t} = \hat{\pi}_{j,\boldsymbol{X}_t}$, if $j \leq k_1$, and $\hat{\pi}'_{j,\boldsymbol{X}_t} = 0$ otherwise, and analogously $\hat{\pi}'_{j,\boldsymbol{Y}_t} = \hat{\pi}_{j,\boldsymbol{Y}_t}$, if $j \leq k_2$, and $\hat{\pi}'_{j,\boldsymbol{Y}_t} = 0$ otherwise.

Maharaj distance

For the class of invertible and stationary ARMA processes, Maharaj (1996) introduced two discrepancy measures based on hypotheses testing to determine whether or not two time series have significantly different generating processes. The first of these metrics is given by the test statistic

$$d_M\left(\boldsymbol{X}_t, \boldsymbol{Y}_t\right) = \sqrt{T}\left(\hat{\mathbf{\Pi}}'_{\boldsymbol{X}_t} - \hat{\mathbf{\Pi}}'_{\boldsymbol{Y}_t}\right)^t \hat{\boldsymbol{V}}^{-1}\left(\hat{\mathbf{\Pi}}'_{\boldsymbol{X}_t} - \hat{\mathbf{\Pi}}'_{\boldsymbol{Y}_t}\right),$$

where $\hat{\mathbf{\Pi}}'_{\boldsymbol{X}_t}$ and $\hat{\mathbf{\Pi}}'_{\boldsymbol{Y}_t}$ are the AR(k) parameter estimations of $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$, respectively, with k selected as in the Piccolo's distance, and $\hat{\boldsymbol{V}}$ is an estimator of $\boldsymbol{V} = \sigma^2_{\boldsymbol{X}_t}\boldsymbol{R}^{-1}_{\boldsymbol{X}_t}(k) + \sigma^2_{\boldsymbol{Y}_t}\boldsymbol{R}^{-1}_{\boldsymbol{Y}_t}(k)$, with $\sigma^2_{\boldsymbol{X}_t}$ and $\sigma^2_{\boldsymbol{Y}_t}$ denoting the variances of the white noise processes associated with $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$, and $\boldsymbol{R}_{\boldsymbol{X}_t}$ and $\boldsymbol{R}_{\boldsymbol{Y}_t}$ the sample covariance matrices of both series.

1.3 Overview of this thesis: Motivation, structure and contributions

In this dissertation, several new approaches to develop time series clustering are introduced. The main intention has been to contribute to the advancement of knowledge on this important topic by providing new tools (e.g. an innovative metric) but also discussing and comparing different methodological strategies (soft and hard paradigms, new clustering principles, robust approaches, and new algorithms designed to deal with time series). This section is aimed at enumerating the main motivations behind this thesis and also highlighting the major contributions.

The first motivation comes from considering a particular scenario of interest in the anal-

ysis of oscillatory phenomena. In fields such as medicine, biology and economics (among others), it is often required to clustering temporal oscillatory records in such a way that each group gathers together series with similar dominant periods of oscillation and also similar power at that dominant period. Indeed, the natural framework to face this problem is the frequency domain. Nevertheless, most of the dissimilarity measures introduced in the frequency domain have been designed to compare whole estimated spectra. This is not the natural approach here. In fact, two time series might eventually exhibit the main spectral peaks at the same frequency and with similar amplitudes, but having different spectral densities. Motivated by this argument, a clustering algorithm aimed at testing whether or not two time series significantly differ at their main spectral peak frequencies and amplitudes is presented in Chapter 2. In a nutshell, the proposed procedure consists of a two-stage algorithm combining ideas from the bootstrap method to test for a zero peak frequency difference proposed by Timmer et al. (1999) with the hierarchical clustering technique based on the resulting p-values developed by Maharaj (2000). The algorithm showed a good behavior in simulated scenarios, including standard linear and non-linear generating models characterized by reasonably separated dominant spectral peaks. In particular, the obtained results were clearly competitive with the ones from other procedures based on metrics relying on a different clustering purpose, and therefore vulnerable to produce erroneous partitions. Strengths and weaknesses of the proposed algorithm are discussed in Chapter 2 and its usefulness is illustrated by the application a real data set.

As argued in the above sections of this introductory chapter, the selection of a suitable dissimilarity between time series according the clustering purpose is basic. Although many dissimilarities have been proposed to clustering series with similar generating processes, most of them are restricted to work with linear models. As consequence of it, the clustering efficacy substantially decreases when these metrics are used to deal with more complex dependence structures (e.g. non-linear or heteroskedastic models). Indeed, this is expected by using model-based measures due to the model misspecification, but many feature-based dissimilarities also behave poorly because the extracted features are not able to properly characterize differences between the involved processes. Therefore, introducing a metric exhibiting a high capability to deal with a broad kind of processes constitutes a challenge in time series clustering. Classification of non-linear models and, above all, of heteroskedastic models is an issue of special interest due to the enormous importance of these models in many environmental and financial problems. With this purpose in mind, we propose a feature-based dissimilarity measure comparing sequences of estimated quantile autocovariances. Quantile autocovariances provide a much richer view into the serial dependence than other extracted features. They encompass a lot of appealing properties, including

robustness to the non-existence of moments, treating properly with heavy tailed marginal distributions, detecting nonlinear features and changes in conditional shapes, among others. Chapters 3 and 4 develop an extensive analysis of time series clustering procedures based on comparing quantile autocovariances.

The quantile autocovariance concept is firstly introduced in Chapter 3. Its properties and capability to time series clustering are presented and discussed via naive and illustrative examples. The asymptotic behavior of the quantile autocovariances is established, and then a dissimilarity measure between two time series based on comparing their estimated quantile autocovariances is formally stated. The rest of Chapter 3 focuses on assessing the behavior of this metric in hard clustering, i.e. using clustering procedures designed to assign each time series to exactly one cluster so that the resulting partition is formed by non-empty and disjoint subsets. Hierarchical and partitional algorithms are taken into consideration, and in both cases extensive simulation studies show that the proposed metric outperforms or is highly competitive with a range of dissimilarities reported in the literature, particularly exhibiting high capability to cluster time series generated from a broad range of dependence models and robustness against the kind of innovation distribution. Furthermore, two important additional issues are addressed in the development of this chapter, namely the determination of an automatic optimal selector of the lags and pairs of quantile levels required to construct the dissimilarity measure, and the estimation of the optimal number of clusters when this value is requested to execute a partitioning-based clustering approach. The algorithms introduced to solve both problems are properly tested by simulation obtaining again satisfactory results. Following the general structure of every chapter, Chapter 3 also includes the application of the proposed method to a specific study case involving financial time series.

Chapter 4 is completely devoted to fuzzy clustering approach. Likewise the hard clustering approaches, it is interesting to analyze the capability of the distance based on quantile autocovariances when soft clustering is carried out, i.e. when the cluster solution is permitted to include overlapping clusters so that some time series can exhibit temporal dynamics close to more than one cluster prototype. Soft clustering of time series has received much less attention in the literature and only some fuzzy approaches based on a few well known metrics have been explored. The promising results of the metric based on quantile autocovariances in hard clustering allow us to suspect that good results could also been obtained by performing soft clustering. Motivated by this intuition and the scarcity of results in this framework, a fuzzy C-medoids procedure using quantile autocovariance is proposed and its behavior is examined via simulations. In this case, the simulation scenarios add uncertainty to the classification procedure by generating variability over the parameters defining the

underlying processes and involving clusters with different levels of separation. Some time series were generated in such a way that their generating structures are equidistant from several clusters and hence they should present similar membership degrees for the corresponding clusters. In sum, the clustering task is substantially more complex and indeed the assessment criteria took into account the capability of the examined algorithms to detect the fuzzy nature of these equidistant series. The main conclusion from our analysis on simulated data was that the proposed approach reported the best results compared to alternative procedures. Again, the proposed procedure is free of problems related to the inaccurate estimation of the underlying parametric structures, and takes advantage of being simpler to implement and computationally lighter than the analyzed competitors. In this case, two comprehensive study cases considering air quality data and daily returns of stocks are subjected to clustering by using different fuzzy approaches to illustrate the behavior of the proposed methodology with real data.

The second part of Chapter 4 deals with other additional issue deserving particular attention: obtaining robust versions of the proposed fuzzy algorithm. This is a very important problem since the presence of time series presenting anomalous temporal behaviors could affect severely the performance of the clustering procedure. To address this problem, three different extensions of robust techniques (D'Urso and Giovanni, 2014) considering the metric based on quantile autocovariances are proposed, namely the metric approach (based on smoothing the distance), the noise approach (by introducing an artificial noise cluster) and the trimmed approach (by trimming away a small fraction of series). Simulations show the enormous importance of using robust techniques in presence of atypical series, the high capability of these techniques to alleviate the effect of anomalous, and an interesting comparative analysis between the different considered algorithms. In this setting, it is observed that the procedures are very sensitive to the input parameters required by each algorithm, but once again it is noticeable the excellent behavior of the metric using quantile autocovariances.

Besides the fuzzy approach, there are other alternatives techniques to perform soft clustering in the literature. Two well known techniques are the probabilistic D-clustering (Ben-Israel and Iyigun, 2008) and clustering based on mixed models (see e.g. Bouveyron and Brunet-Saumard, 2014). To the best of our knowledge, the former has not been employed to perform cluster analysis of time series, and the latter has been applied in a very limited way. More precisely, we are only aware of the work by Chen and Maitra (2011) where a model-based approach for clustering time series regression data is proposed by assuming that each mixture component follows a Gaussian autoregressive regression model of order p. Therefore, exploring new approaches considering probabilistic D-clustering and

mixed models to perform times series clustering is fairly of interest for several reasons. The probabilistic D-clustering is simple, requires a small number of cheap iterations and is insensitive to outliers. Approaches based on mixed models are more expensive in computational terms, but in contrast, they lead to membership degrees in an automatic way without pre-establishing a fuzziness parameter.

In Chapter 5 two new clustering procedures based on both the probabilistic D-clustering and mixed models are proposed. The first is constructed in a natural way by considering that the probability of cluster membership for an arbitrary time series is inversely proportional to the distance from the center of the cluster in question, when that distance is computed by using the estimated quantile autocovariances. The cluster centers may change so that the algorithm is carried out in an iterative manner until a stop rule determines the final clustering solution. It is expected that this probabilistic D-clustering takes advantage of the robust behavior of the metric based on quantile autocovariances. As far as the new clustering algorithm based on mixture models, the key idea is to take into account that the errors from the estimation of the smoothed log-periodogram follow a Gumbel distribution, i.e. with probability density function given by $\varphi(x) = \exp(x - \exp(x))$. Therefore, the values of an arbitrary log-periodogram are distributed by a mixture of these parametric distributions whose k-th coefficient represents the probability that the corresponding time series belongs to the k-th cluster. Next step consists of estimating the parameters of the mixture by maximizing the local log-likelihood function for all the collected log-periodograms, which is carried out by developing an Expectation-Maximization (EM) algorithm. In this case, the expectation step (E-step) requires an innovative criterion to compute the posterior probabilities in order to attain interpretable solutions in the context of soft clustering. It is also shown that the maximization of the complete log-likelihood in the M-step leads to closed-form expressions. Once the algorithms are properly described, a comparison with the fuzzy algorithm proposed in Chapter 4 is performed via simulation. Results reported from this simulation study show that the three examined soft procedures exhibit a satisfactory behavior, being capable to detect time series located between different clusters.

The main conclusions of this thesis are shortly enumerated in the last chapter, where some interesting open lines and additional challenges in the topic of time series clustering are also pointed out for further research.

1.4 Preliminary concepts

This section is devoted to establish some preliminary notions and tools which are of interest in the development of this dissertation. Specifically, a formal definition of the stationary concept, a short description of nonlinear models used in simulations later on, some basic results in spectral analysis and some useful tools to evaluate the quality of a cluster solution are presented in the following subsections.

1.4.1 Stationarity

All the clustering procedures developed throughout this thesis apply on strictly stationary time series. As it is well known, stationarity is the most important form of time-homogeneity used in time series analysis. Stationary property means time-invariance of the whole probability distribution of the data generating process (strict stationarity), or just of its first two moments (weak stationarity or simply stationarity).

Definition 1.4.1 (Stationarity) *The process* $\{X_t;\ t \in \mathbb{Z}\}$ *is said to be stationary if for all* $l, t \in \mathbb{Z}$, $\mathbb{E}(X_t) = \mu$ *and* $Cov(X_t, X_{t+l}) = \gamma(l)$, *with* $\gamma(0) < \infty$.

The terms "weakly stationary", "second-order stationary", "covariance stationary" and "wide-sense stationary" are also often used to refer to processes satisfying the above definition.

Definition 1.4.2 (Strict stationarity) *The process* $\{X_t;\ t \in \mathbb{Z}\}$ *is said to be strictly stationary if the random vectors* $(X_{t_1}, \ldots, X_{t_n})$ *and* $(X_{t_1+l}, \ldots, X_{t_n+l})$ *have the same joint distribution for any* $t_1, t_2, \ldots, t_n \in \mathbb{Z}\}$ *and for all integers* l *and* $n > 0$. *It can be written as*

$$(X_{t_1}, \ldots, X_{t_n}) \stackrel{d}{=} (X_{t_1+l}, \ldots, X_{t_n+l}),$$

where $\stackrel{d}{=}$ *means equal in distribution.*

If $\mathrm{Var}(X_t)$ is assumed to exist, then strict stationarity implies stationarity. While stationarity property is primarily used to deal with linear models, strict stationarity is often required in the context of nonlinear time series analysis. In particular, the consistency result for the estimates of the quantile autocovariances in Chapter 3 is obtained under the strict stationary assumption. Note that, by definition, all the finite-dimensional distributions for a Gaussian process are normal, and therefore a stationary Gaussian process is also strictly stationary. It is also worth to remark that many time series are nonstationary in practice, but they may be made stationary after some simple transformation, such as taking differ-

ences between consecutive observations, subtracting the estimated trend, etc. This kind of transformations will be performed in several applications with real data in this thesis.

1.4.2 Some nonlinear time series models

The framework of this thesis is not limited to cluster analysis of realizations from lineal time series. In fact, robustness against the generating model is one of the strengths of the proposed approaches. Considering nonlinear models is of great interest because of these models cover a much wider spectrum of potential dynamics for real time series data in many fields. On the other hand, the theory of nonlinear time series has received an increasing attention since the early and motivating monograph by Tong (1993). Significant advances have been attained and many well-studied parametric and nonparametric approaches to model nonlinear structures in time series are available nowadays. The monographs by Tong (1993) and Fan and Yao (2005) are key references to obtain a comprehensive background.

A range of popular nonlinear models have been considered in simulation studies developed throughout this thesis. For an easier and ordered reading, the used nonlinear models are shortly presented in this subsection and the particular constraints required to ensure their stationarity are also highlighted. Note that, in general, it is not simple to check whether a nonlinear time series is strictly stationarity. The common practice is to represent the series as a vector-valued Markov chain and to establish the geometrical ergodicity of the induced Markov chain (see Tjostheim, 1990; Tong, 1993, and references therein). Then, strict stationarity follows from the fact that an ergodic Markov chain is strictly stationary (Theorem 2.2 in Fan and Yao, 2005).

Indeed, there are many ways a process can be nonlinear, but our experiments focused on two main types of processes, namely parametric models for the conditional mean and parametric models for the conditional variance. The former represent the conditional mean function of the process as a nonlinear function of the past observations, keeping the conditional variance constant. The used models within this category are presented below. Notice that presentation is restricted to models of order one (with only one lag) because the experiments were limited to this case for the sake of simplicity.

In what follows, the stochastic process is denoted by X_t and $\{\varepsilon_t, t \in \mathbb{Z}\}$ represents a sequence of independent, identically distributed random variables with a positive density and finite first and second moments, and such that ε_t is independent of X_s, for all $s < t$.

Nonlinear autoregressive (NLAR) model. Nonlinear autoregression constitutes a very

general class of nonlinear processes where X_t is assumed to satisfy the model defined by

$$X_t = f(X_{t-1}) + \varepsilon_t, \ t \in \mathbb{Z}, \tag{1.2}$$

with $f : \mathbb{R}^2 \to \mathbb{R}$ being a function indexed by some parameters. For instance, experiments in Chapter 4 involve the NLAR(1) given by

$$X_t = \frac{0.3\,|X_{t-1}|}{3 + |X_{t-1}|} + \varepsilon_t. \tag{1.3}$$

The geometrical ergodicity of NLAR models is studied in detail by An and Huang (1996). In particular, the uniform boundedness of the nonlinear autoregressive function f ensures that X_t in (1.3) is geometrically ergodic and hence stationary (see Example 3.1 and Theorem 3.1 in An and Huang, 1996).

Threshold autoregressive (TAR) model. The TAR models capture the dynamic behavior by partitioning the real line with thresholds and considering a finite parametric model for each regime determined by these thresholds. They constitute a very important class of nonlinear models and have been studied in depth. The simplest form for a first-order TAR model with two regimes is given as

$$X_t = \phi_1 X_{t-1} I(X_{t-1} \leq r) + \phi_2 X_{t-1} I(X_{t-1} > r) + \varepsilon_t, \tag{1.4}$$

where $I(\cdot)$ denotes the indicator function and r is the threshold partitioning the real line. Petruccelli and Woolford (1984) establish that a necessary and sufficient condition for the geometrical ergodicity of the model (1.4) is $\phi_1 < 1$, $\phi_2 < 1$ and $\phi_1\phi_2 < 1$.

Exponential autoregressive (EXPAR) model. The EXPAR models introduced by Ozaki (1980) are particularly suitable to capture well-known features of nonlinear vibrations such as amplitude-dependent frequency, jump phenomena, and limit cycle behavior. The basic form of an EXPAR(1) model is

$$X_t = \left(\alpha + \beta \exp\left(-\delta X_{t-1}^2\right)\right) X_{t-1} + \varepsilon_t, \ \text{with } \delta > 0. \tag{1.5}$$

Example 10.4.3 in Amendola and Francq (2009) states that the model (1.5) is geometrically ergodic whenever $|\alpha| < 1$, whatever $\beta \in \mathbb{R}$.

Bilinear (BL) model. Bilinear models were introduced by Granger and Andersen (1978) and represent a natural way to introduce nonlinearity into a linear ARMA model by adding

product terms. The general formulation for a bilinear model of order (p, q, P, Q) is

$$X_t = \sum_{j=1}^{p} \beta_j X_{t-j} + \varepsilon_t + \sum_{k=1}^{q} \alpha_k \varepsilon_{t-k} + \sum_{j=1}^{P} \sum_{k=1}^{Q} \gamma_{jk} X_{t-j} \varepsilon_{t-k},$$

which is usually denoted by $\mathrm{BL}(p, q, P, Q)$. Note that X_t is linear in X_i as well as in ε_i and hence the name of 'bilinear'. Despite this intuitive definition, the analytical properties of bilinear models are less-understood than the ones of other nonlinear time series models. As in the above models, consider the first-order $\mathrm{BL}(1, 0, 1, 1)$ model given by

$$X_t = \beta_1 X_{t-1} + \varepsilon_t + \gamma_{11} X_{t-1} \varepsilon_{t-1}. \tag{1.6}$$

Pham and Tran (1981) prove that condition $\beta_1^2 + \sigma^2 \gamma_{11} < 1$, with $\sigma^2 = \mathrm{Var}\varepsilon_t$, implies the stationarity of the model (1.6).

Nonlinear mean average (NLMA) model. A very simple way to obtain a nonlinear model is to consider a nonlinear version of the moving average model, i.e.

$$X_t = \alpha_0 + \sum_i \alpha_i \varepsilon_{t-i} + \sum_i \sum_j \alpha_{ij} \varepsilon_{t-i} \varepsilon_{t-j} + \dots$$

Our experiments include a simple NLMA structure given by

$$X_t = \alpha_0 + \varepsilon_t + \alpha_1 \varepsilon_{t-1} + \alpha_{11} \varepsilon_{t-1}^2. \tag{1.7}$$

Since ε_t are i.i.d. variables, X_t in (1.7) is strictly stationary.

The listed models so far exhibit nonlinearity in the conditional mean but with constant conditional second moment. It is well known that for example the temporal dynamic of financial returns usually presents high volatility, i.e. the standard deviation of the financial returns shows large changes over time. The most popular approach for modelling time-varying conditional variance is to use the ARCH and GARCH models introduced by Engle (1982) and Bollerslev (1986), respectively. These models are particularly useful to capture some important *stylized features* of financial return series, including heavy-tailed errors and volatility clustering. Nevertheless, they fail to model other stylized features such as an asymmetric response of volatility. Motivated for this, extensions of the GARCH models (EGARCH, FIARCH, ARCH-M, ST-GARCH, GJR-GARCH, DT-GARCH,...) have been proposed in the literature (see Shephard, 1996, for a comprhensive survey on extended GARCH models). The particular models considered throughout this thesis are enumerated below.

Autoregressive conditional heteroskedastic (ARCH) model. An ARCH model of order p (≥ 1) has the form $X_t = \sigma_t \varepsilon_t$, where the variance at time t, σ_t^2, is conditional on the past observations according to

$$\sigma_t^2 = \gamma + \sum_{j=1}^{p} \beta_j X_{t-j}^2, \tag{1.8}$$

where $\gamma \geq 0$ and $\beta_j \geq 0$ are constants.

Theorem 4.3 in Fan and Yao (2005) states that $\sum_{j=1}^{p} \beta_j < 1$ is a necessary and sufficient condition for strict stationarity of the process X_t defined by (1.8).

Generalized autoregressive conditional heteroskedastic (GARCH) model. A GARCH model is the extension of an ARCH to include a moving average structure. This way, the conditional variance for a GARCH of order p (≥ 1) and q (≥ 0) follows the model

$$\sigma_t^2 = \gamma + \sum_{j=1}^{p} \beta_j X_{t-j}^2 + \sum_{j=1}^{q} \alpha_j \sigma_{t-j}^2 \tag{1.9}$$

where $\gamma \geq 0$, $\beta_j \geq 0$ and $\alpha_j \geq 0$ are constants.

The necessary and sufficient condition for strict stationarity of a GARCH time series is $\sum_{j=1}^{p} \beta_j + \sum_{j=1}^{q} \alpha_j < 1$ (see Theorem 4.4 in Fan and Yao, 2005).

Exponential GARCH (EGARCH) model. In the EGARCH model introduced by Nelson (1991), σ_t^2 takes the form

$$\ln\left(\sigma_t^2\right) = \gamma + \sum_{j=1}^{q} \alpha_j \ln\left(\sigma_{t-j}^2\right) + \sum_{j=1}^{s} g_j\left(\varepsilon_{t-j}\right), \text{ with } g_j(z) = \omega_j z + \lambda_j\left(|z| - \mathbb{E}(|z|)\right) \tag{1.10}$$

where parameters in (1.10) are not restricted to be nonnegative because the conditional volatility is always positive. Unlike the GARCH model, the form of σ_t^2 depends on both the size and the sign of the lagged ε_t by means of the functions $g_j(\cdot)$. This allows the EGARCH models to respond nonsymetrically to random shocks. Since $\{\varepsilon_t\}$ is i.i.d., $\{g_j(\varepsilon_t)\}$ is also i.i.d., and therefore X_t in (1.10) is strictly stationary if $\sum_{j=1}^{q} \alpha_j < 1$.

Glosten-Jagannathan-Runkle GARCH (GJR-GARCH) model. The GJR-GARCH model introduced by Glosten et al. (1993) can be interpreted as a special case of threshold model, where the conditional variance is formulated as follows for orders p (≥ 1) and q

(≥ 0).

$$\sigma_t^2 = \gamma + \sum_{j=1}^{p} \left[\beta_j + \lambda_j I\left(X_{t-j} < 0\right)\right] X_{t-j}^2 + \sum_{j=1}^{q} \alpha_j \sigma_{t-j}^2 \tag{1.11}$$

with all parameters in (1.11) being nonnegative constants. Note that the asymmetric volatility phenomenon is modeled in (1.11) by using dummy variables so that the impact on the conditional variance is different according to the past returns are positive or negative. Due to its simplicity, the GJR-GARCH model is very popular in the finance literature. For the case of $p = q = 1$, the conditions of stationarity for the GJR-GARCH model are $\gamma, \beta_1, \alpha_1 > 0$, $\beta_1 + \lambda_1 \geq 0$ and $\beta_1 + \alpha_1 + 0.5\lambda_1 < 1$ (see Table 2 in Chen et al., 2011).

1.4.3 Spectral estimation

In this section, some essential aspects of the spectral theory of stationary processes are described, paying attention to those concepts that will later be useful in the development of this thesis. A more detailed study of spectral analysis theory can be found in Priestley (1989) and Brillinger (1981).

Let $\boldsymbol{X} = \{X(t), t \in \mathbb{Z}\}$ an stationary process with zero mean and autocovariance function $\gamma(\cdot)$ such that

$$\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty.$$

The spectral density of an stochastic process $\{X_t\}$ is defined as the Fourier transform of the autocovariance function $\gamma(h) = E\left(X_t X_{t+h}\right)$, i.e.

$$f(\lambda) = \frac{1}{2\pi} \sum_{h} \gamma(h) e^{2\pi\lambda h}, \tag{1.12}$$

where the frequencies $\lambda \in (-\infty, \infty)$.

Since functions $\cos(\cdot)$ and $\sin(\cdot)$ have both period 2π, the spectral density is periodic with the same period, reason why it is enough to define it in the interval $(-\pi, \pi]$.

In practice, the theoretical spectral density function is unknown and it is necessary to obtain an approximation. A procedure to estimate the spectrum is to use the periodogram. Given $\boldsymbol{X}_T = (X_1, \ldots, X_T)$ a partial realization of the stationary process $\boldsymbol{X}$, the periodogram $I_T(\cdot)$ is defined as

$$I_T(\lambda) = \frac{1}{2\pi T}\left|\sum_{t=1}^{T} X_t e^{-i\lambda t}\right|^2, \tag{1.13}$$

where $\lambda \in (-\pi, \pi)$. Priestley (1989) proved that if λ_k denotes the Fourier frequencies, i.e. $\lambda_k = 2\pi k/T$, with $k = -M, \ldots, M$, $M = [(T-1)/2]$, then

$$I_T(\lambda_k) = \frac{1}{2\pi}\sum_{|h|<T} \widetilde{\gamma}(h) e^{-ih\lambda_k}, \tag{1.14}$$

where $\widetilde{\gamma}(h)$ is the sample autocovariance function associated to $\boldsymbol{X}_T$.

For $\lambda_k \notin \{-\pi, \pi\}$, the periodogram ordinates follow a distribution proportional to a chi-squared with 2 degrees of freedom, according to

$$I_T(\lambda_k) \sim \frac{1}{2} f(\lambda_k)\, \chi^2_2, \tag{1.15}$$

and thus expression (1.15) opens an approach to testing by the theoretical spectrum. Nevertheless the periodogram is not a consistent estimator for the spectrum.

To estimate the spectrum consistently an estimator based on replacing the spectra by the exponential transformation of local linear smoothers of the log-periodograms obtained via least squares (Fan and Kreutzberger, 1998) is used.

If $\boldsymbol{X}$ is a Gaussian linear process, it can be proved that the coordinates of the periodogram evaluated in the Fourier frequencies, $I_T(\lambda_k)$, are asymptotically distributed as an exponential of mean $f(\lambda_k)$ and are approximately independents. More formally, they follow the following heteroscedastic regression model:

$$I_T(\lambda_k) = f(\lambda_k)V_k + R_k \tag{1.16}$$

where f is the spectral density, R_k denotes an asymptotically null term and the V_k are variables with a standard exponential distribution and independent for all $k \neq 0$.

By applying a logarithmic transformation to the model (1.16) we have

$$Y_k = \log(I_n(\lambda_k)) = m(\lambda_k) + \varepsilon_k + k_k, \tag{1.17}$$

where $m(\lambda_k) = \log(f(\lambda_k))$, $\varepsilon_k = \log(V_k)$ are random variables iid with density function $\varphi(x) = \exp(x - \exp(x))$, and $r_k = \log\{1 + R_k/f(\lambda_k)V_k\}$ denotes an asymptotically null

term.

Since r_k is asymptotically null, it can be ignored in (1.17). Thus, if the regression model (1.17) is centered by subtracting $\mathbb{E}(\varepsilon_k) = C_0$, with C_0 being the Euler constant, and the term r_k is disregarded we have that

$$Y_k - C_0 = \log(I_n(\lambda_k)) - C_0 = m(\lambda_k) + (\varepsilon_k - C_0). \tag{1.18}$$

Fan and Gijbels (1996) propose up to three possible nonparametric approaches to estimate f. The first one is to smooth the logarithm of the periodogram using a least squares method. Applying the least squares method to model (1.18) in order to obtain the best local linear fit, the following estimator of the logarithm of the spectrum is attained

$$\widehat{m}_{LS}(\lambda) = \sum_{k=-M}^{M} w_k(\lambda)(Y_k - C_0), \tag{1.19}$$

where $w_k(\lambda)$ denotes the weights of the corresponding local linear fit. Then the estimator of the spectral density is obtained by back-transforming $\widehat{m}_{LS}$,

$$\widehat{f}(\lambda) = \widehat{f}_{LS}(\lambda) = \exp(m_{LS}(\lambda)), \tag{1.20}$$

which we refer to as smoothed log-periodogram.

The smoothed periodogram estimator $\widehat{m}_{LS}$ is not efficient due to the non-normality of the errors. The efficiency of the least squares method can be improved by using the maximum likelihood method. Assuming the model (1.18), for each λ the weighted log-likelihood is constructed as follows:

$$\mathcal{L}(a,b) = \sum_{k=1}^{M} \left[-\exp\left\{Y_k - a - b(\lambda_k - \lambda)\right\} + Y_k - a - b(\lambda_k - \lambda)\right] K_h(\lambda_k - \lambda), \tag{1.21}$$

where $K_h(\cdot) = K(\cdot/h)/h$.

Let $\widehat{a}$ and $\widehat{b}$ be the maximizer of (1.21). Then, the local likelihood estimator for $m(x)$ is $\widehat{m}_{LK} = \hat{a}$, and again the estimator for the spectral density is obtained by back-transforming $\widehat{m}_{LK}$,

$$\widehat{f}(\lambda) = \widehat{f}_{LK}(\lambda) = \exp(m_{LK}(\lambda)), \tag{1.22}$$

A third way to estimate the spectral density is to smooth directly the periodogram, that is

applying local linear smoothing directly to $\{\lambda_k, I_T(\lambda_k)\}$, which leads to

$$\widehat{f}(\lambda) = \widehat{f}_{DLS}(\lambda) = \sum_{k=1}^{M} K_n \left(\frac{\lambda - \lambda_k}{h} \right) I_T(\lambda_k), \tag{1.23}$$

1.4.4 Quality clustering indexes

In this section, some criteria related to the quality of a cluster solution are presented. Namely, two different indexes to estimate the number of clusters of a partition and an index of agreement to compare the true cluster partition with the experimental one.

Let $S = \{\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(n)}\}$ denote a set of n time series of length T and $\mathcal{E}_k = \{E_1, \ldots, E_k\}$ a given cluster partition of S.

One of the methods considered to estimate the number of clusters in S consists in maximizing the average Silhouette width, ASW, proposed by Kaufman and Rousseeuw (1990). Given the partition $\mathcal{E}_k$, ASW is defined by

$$\text{ASW} = \frac{1}{p} \sum_{i=1}^{p} \text{sil}(i),$$

where $\text{sil}(i)$ is called the Silhouette width for the ith individual series $\boldsymbol{X}^{(i)}$ and defined by

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

with $a(i)$ denoting the average of the distances between $\boldsymbol{X}^{(i)}$ and all other series in its cluster, and $b(i)$ the average of the distances between $\boldsymbol{X}^{(i)}$ and all series in the closest cluster (i.e. the second-best group for $\boldsymbol{X}^{(i)}$). By definition, a value of $\text{sil}(i)$ close to one indicates that $\boldsymbol{X}^{(i)}$ is very well clustered, a small value (around 0) means that $\boldsymbol{X}^{(i)}$ lies between two clusters, and a value close to -1 indicates placement in the wrong cluster. This way, ASW always takes values between -1 and 1 and provides an overall measure of how well series are clustered.

A commonly used index proposed by Krzanowski and Lai (1988) is also used to estimate the number of clusters. The objective is to select the value of k providing an optimal value for these functions or internal indexes. Specifically, given the partition $\mathcal{E}_k$, denote by B_k and W_k the $T \times T$ matrices of between and within k-clusters sums of squares and cross-products, respectively. Then, the mentioned indexes perform as follows.

The Krzanowski and Lai index (KL) calculates

$$\mathrm{KL}(k) = \frac{|diff_k|}{|diff_{k+1}|},$$

where $diff_k = (k-1)^{2/T} \operatorname{tr}(W_{k-1}) - k^{2/T} \operatorname{tr}(W_k)$. Likewise, the value of k maximizing $\mathrm{KL}(k)$, $k \geq 2$, is selected.

Finally, an index of agreement between the true cluster partition, $\mathcal{T} = \{T_1, \ldots, T_C\}$, and the experimental partition $\mathcal{R}$ in order to measure the quality of the clustering procedure is considered in chapters 2 and 3 of the dissertation. More specifically, the agreement index (Gavrilov et al., 2000; Liao, 2005) is defined by

$$Ind_1(\mathcal{T}, \mathcal{R}) = \frac{1}{C} \sum_{i=1}^{C} \max_{1 \leq j \leq C} Ind_1(T_i, R_j), \tag{1.24}$$

where

$$Ind_1(T_i, R_j) = \frac{2|T_i \cap R_j|}{|T_i| + |R_j|},$$

and $|V|$ denotes the cardinality of a set V. Index Ind_1 accounts for the number of series sharing a same cluster in both partitions, taking exactly the value 0 if both partitions are completely dissimilar and the value 1 if they are identical.

Chapter 2

Clustering based on frequencies and amplitudes of spectral peaks

Contents

2.1 Introduction

The spectral representation of a stationary process $X = \{X(t), t \in Z\}$ essentially decomposes X into a sum of sinusoidal components with random and uncorrelated coefficients. The spectral decomposition is thus, in the realm of the time series, a concept analogous to the Fourier representation of deterministic functions. The analysis of stationary processes in their spectral representation is usually called "analysis in the frequency domain" or "spectral analysis". While "time domain" analysis is based on the autocovariances function, spectral approach replaces the covariance matrix representation by its corresponding spectral density representation, which provides a different way of analyzing processes that may be more interesting and useful in some applications. A detailed study of spectral analysis theory can be found in specialized references (Brillinger, 1981; Priestley, 1989; Brockwell and Davis, 2002; Shumway and Stoffer, 2006).

The frequency domain approach provides an alternative paradigm to perform cluster analysis of time series, since the concept of dissimilarity between processes can be thought in terms of disparity between their spectral representations. A number of significant contributions have followed this approach by introducing metrics comparing the underlying spectral densities. Kakizawa et al. (1998) proposed a metric based on a general spectral disparity between two time series. In practice, the spectra are unknown and must be previously estimated. Vilar and Pértega (2004) studied the asymptotic properties of the metric proposed by Kakizawa et al. (1998) when the spectra 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 metric proposed by Kakizawa et al. (1998). Specifically, (a) replacing the spectra by local lineal smoothers of the periodograms obtained via least squares, (b) replacing the spectra by the exponential transformation of local linear smoothers of the log-periodograms obtained via least squares, and (c) proceeding as in (b) but here using the maximum local likelihood criterion to obtain the local linear smoothers. Also, other two alternative nonparametric spectral dissimilarity measures were introduced by Pértega and Vilar (2010). In both cases, the discrepancy measure is given by a nonparametric statistic originally introduced to check the equality of the log-spectra of two processes. 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. Pértega and Vilar (2010) introduced a modification of this test statistic in order to check the equality of two log-spectra. The second distance evaluates the integrated squared differences between nonparametric estimators of the log-spectra. Some of these distances have been briefly presented in Section 1.2 of Chapter 1.

Beyond the comparison of whole spectral densities, detecting differences between spectral peak frequencies is often a problem of major interest in medical, biological and economic applications. For instance, in clinical diagnosis different pathologies might be determined by deciding whether significant spectral peaks are located into different frequency ranges (Findley and Koller, 1987). Relevant information about activations and artifacts in functional magnetic resonance imaging (fMRI) data sets is sometimes obtained by determining the location of significant frequencies (Jarmasz and Somorjai, 2002). Motivated for this interest, we focus on developing a clustering algorithm aimed at partitioning the observed time series according to the location of their significant spectral peaks. More specifically, a two-stage clustering procedure based on comparing frequencies and magnitudes associated to the highest spectral peaks is presented in this chapter. In the first stage, the dissimilarity between each pair of series is evaluated in terms of the p-value associated to a bootstrap

test of equality of the frequencies where the spectral maxima are reached (Timmer et al., 1999). Based on the pairwise p-values matrix and following the clustering technique proposed by Maharaj (2000), a first cluster partition is built up. As it will be detailed later, the technique proposed by Maharaj proceeds in a similar way as an agglomerative hierarchical clustering starting from the p-values matrix, but here will only group together those series whose associated p-values are greater than a significance level pre-fixed by the user. In this first stage, each cluster brings together the series presenting the highest spectral peak at similar frequencies, but these peaks could exhibit different magnitudes. This fact accounts for a second stage of the clustering algorithm addressed to check if the areas under the spectral densities within each cluster differ in a local environment of the peak frequency. This task is separately carried out for each of the clusters generated at the first stage of the process. For each group, a new matrix of p-values coming from testing by equality of these local areas is constructed and used to perform again the hierarchical clustering procedure proposed by Maharaj (2000), thus obtaining the final cluster partition. Indeed, this procedure could be iteratively applied for the following significant spectral peaks.

The performed simulations showed the good performance of the proposed procedure, but it is important to notice about the limitations inherent to the method, particularly its high computational complexity and the need of introducing relevant input parameters. As it will be discussed in the section of conclusions, the recommendation is to consider this approach only when the clustering purpose focuses on splitting the set of time series into groups characterized by the location of their spectral peak frequencies. In a more general context where the interest is to classify the series according to the underlying processes, other metrics result more efficient.

The rest of this chapter is organized as follows. In Section 2.2, the two-stage clustering procedure based on comparing frequencies and magnitudes of the absolute peaks of the spectral densities is introduced and described in detail. The performance of the proposed clustering methodology is examined via simulations and compared to other alternative clustering approaches in Section 2.3. Section 2.4 shows an application on real data set involving economic time series, and the main conclusions are presented in Section 2.5.

2.2 The clustering procedure

Let $S = \left\{ \boldsymbol{X}^{(j)}; \boldsymbol{X}^{(j)} = \left(X_1^{(j)}, \ldots, X_T^{(j)} \right) \right\}$, for $j = 1, \ldots, n$, be a set of n realizations of time series of length T. The goal is to perform cluster analysis on S in such a way that each cluster brings together those series having the same location and magnitude for their

main spectral peaks. The proposed methodology consists on two stages.

The first stage focuses on the location problem, i.e. on checking whether the difference between the peak frequencies of the spectra of the series in S are or not significantly different from zero. The objective is to group series with the same spectral peak frequency.

The second stage separately applies to each of the clusters generated at the first stage, and consists on splitting each cluster into a new partition discriminating between series whose main peaks differ in power. In other words, the new clustering process is based on checking whether the areas under the spectral densities in a local neighborhood of the main peak frequencies differ significantly.

In both stages, pairwise dissimilarities are evaluated by means of the p-values from bootstrap tests for equality of the spectral features of interest, namely frequencies and powers for the main spectral peaks. The test procedures were proposed by Timmer et al. (1999). Note that the p-values associated with these tests can be used to measure the amount of dissimilarity between these spectral features: the smaller is the p-value, the larger is the discrepancy between them.

Once the $n \times n$ matrix of p-values is available, the hierarchical clustering algorithm based on p-values introduced by Maharaj (2000) is carried out. This algorithm works as follows. First, a significance level α is previously specified by the user. Then, the i-th series $\boldsymbol{X}^{(i)}$ will merge into a specific cluster formed by the m series $\left\{\boldsymbol{X}^{(j_1)}, \ldots, \boldsymbol{X}^{(j_m)}\right\}$ iff $p_{i,j_l} \geq \alpha$, for all $l = 1, \ldots, m$. Analogously, two clusters will be joined together if and only if the p values of all pairs of series across the two clusters are greater than α. Unlike the conventional hierarchical methods, this algorithm presents the advantage of providing automatically the number of clusters, which obviously depends on the prefixed significance level. Furthermore, the amount of compactness of each cluster can be evaluated by examining the p-values within each cluster.

A detailed description of the clustering procedure is provided below.

Stage 1:

1. Estimate the spectral density of each time series and, for each pair of series $\left(\boldsymbol{X}^{(i)}, \boldsymbol{X}^{(j)}\right)$, $i \neq j$, compute $\widehat{\Delta\lambda}_{ij} = \widehat{\lambda_i^p} - \widehat{\lambda_j^p}$, where $\widehat{\lambda_i^p}$ denotes the estimator of the main spectral peak frequency of the i-th series.

2. Generate B bootstrap resamples of the periodograms regarding that

$$I_T(\lambda_k) \sim \frac{1}{2} \widehat{f}\,(\lambda_k)\, \chi_2^2,$$

with $\lambda_k \in (-\pi, \pi)$ the k-th Fourier frequency and $\widehat{f}(\lambda_k)$ the estimator of the spectral density in λ_k.

3. For each set of resamples, the spectra are reestimated and the new differences between the peak frequencies are computed to obtain bootsrap replicates of the differences $\widehat{\Delta\lambda}_{ij}^{*}$.
4. Based on the bootstrap distribution of $\left(\widehat{\Delta\lambda}_{ij}^{*} - \widehat{\Delta\lambda}_{ij}\right)$ and the value of the statistic $\widehat{\Delta\lambda}_{ij}$, obtain the p-value p_{ij} to check the null hypothesis $H_0^{ij} : \lambda_i^p = \lambda_j^p$. The $n \times n$ matrix of p-values $\left(p_{ij}^{(1)}\right)$ is considered as dissimilarity matrix to develop hierarchical clustering.
5. According to a prefixed significance level α_1, the hierarchical clustering algorithm proposed by Maharaj (2000) is performed starting from the matrix $\left(p_{ij}^{(1)}\right)$.

Stage 2: For each cluster C generated in Stage 1, proceed as follows.

1. For each pair of series $\left(\boldsymbol{X}^{(i)}, \boldsymbol{X}^{(j)}\right)$ within the cluster C, estimate the area $\widehat{\Delta A}_{ij}$ between the estimated spectra on a local neighborhood Ω_C of the peak frequencies for the series in the cluster by computing:

$$\widehat{\Delta A}_{ij} = \int_{\Omega_C} \left|\widehat{f}_i(\lambda) - \widehat{f}_j(\lambda)\right| d\lambda,$$

2. Using the bootstrap resamples obtained in Stage 1, compute $\widehat{\Delta A}_{ij}^{*}$.
3. According to the bootstrap distribution of $\left(\widehat{\Delta A}_{ij}^{*} - \widehat{\Delta A}_{ij}\right)$, obtain the p-value $p_{ij}^{(2)}$ to check $H_0^{ij} : \Delta A_{ij} = 0$.
4. Based on the matrix $\left(p_{ij}^{(2)}\right)$ obtained in the above step, use again the hierarchical clustering algorithm proposed by Maharaj (2000) with a significance level α_2 set in this stage.

It is worth to highlight some remarks about the described algorithm. According to the common choices for the significance level of a test, the values of α_1 and α_2 will be set at 1%, although depending on the number of series involved in the clustering process some kind of adjustment for multiple testing could be carried out. As far as the estimation of the spectral densities required in steps 1 and 2 of Stage 1, any of the nonparametric approximations mentioned in the previous section can be used. In order to reduce the computational cost, a reasonable choice may be the local linear smoother of the log-periodogram computed by least squares. Note that the local linear fitting techniques present nice properties including a good performance at boundary points (other smoothing techniques suffer from the well

known "boundary effect"). This property is particularly useful here due to work on a compact support. In fact, the largest spectral peak might correspond to a frequency close to the boundary.

Other important remark refers to the determination of the local neighborhoods Ω_C in step 1 of Stage 2. Roughly speaking, for a given cluster C, Ω_C is the interval whose endpoints are the frequencies of half estimated power on left and right of the maximum peak. As all the series in C maximize the estimated power at the same (or similar) frequency, this interval should contain all the main peak frequencies in the cluster and provide a reasonable range for evaluating and comparing the curvatures at all peaks. More formally, assume that C is formed by m series, $C = \left\{\boldsymbol{X}^{(j_1)}, \ldots, \boldsymbol{X}^{(j_m)}\right\}$. For each series $\boldsymbol{X}^{(j_i)}$ in C, let λ_p^i be the frequency maximizing $\widehat{f}_{j_i}(\lambda_k)$, for all $k \in \{1, \ldots, N\}$, and denote by M_i the attained maximum. Construct the interval $\left[\lambda_l^i, \lambda_u^i\right]$ with endpoints are given by

$$\lambda_l^i = \max\left\{0, \max_{1\leq k\leq N}\left(\lambda_k \,/\, \widehat{f}_{j_i}(\lambda_k) \leq \frac{1}{2}M_i\right)\right\},$$

and

$$\lambda_u^i = \min\left\{0.5, \min_{1\leq k\leq N}\left(\lambda_k \,/\, \widehat{f}_{j_i}(\lambda_k) \geq \frac{1}{2}M_i\right)\right\}.$$

Finally, Ω_C is selected as the longest interval among all the $\left[\lambda_l^i, \lambda_u^i\right]$ intervals, for $i = 1, \ldots, m$.

Hereafter, the proposed procedure will be refered to SP algorithm.

2.3 Simulation study

In this section, we present the results from a numerical study designed to compare the behavior of a group of classic dissimilarity measures against the proposed method when they are used to cluster a group of observed time series.

2.3.1 Main features of the simulation study

Simulations were conducted to assess the performance of the SP algorithm compared to a wide selection of model-free dissimilarity measures, and considering two different classification setups, namely classification of (i) ARMA models and (ii) non-linear models. The generating models selected at each case are enumerated below.

Scenario 2.1 Classification of ARMA processes.

(a) AR(1) $X_t = 0.5X_{t-1} + \varepsilon_t$
(b) MA(1) $X_t = -0.9\varepsilon_{t-1} + \varepsilon_t$
(c) AR(2) $X_t = 0.3X_{t-1} - 0.6X_{t-2} + \varepsilon_t$
(d) MA(2) $X_t = 0.8\varepsilon_{t-1} - 0.6\varepsilon_{t-2} + \varepsilon_t$

Scenario 2.2 Classification of non-linear processes.

(a) TAR $X_t = 0.5X_{t-1}I\left(X_{t-1} \leq 0\right) - 2X_{t-1}I\left(X_{t-1} > 0\right) + \varepsilon_t$
(b) EXPAR $X_t = \left(0.3 - 10\exp\left(-X_{t-1}^2\right)\right)X_{t-1} + \varepsilon_t$
(c) MA $X_t = -0.4\varepsilon_{t-1} + \varepsilon_t$
(d) NLMA $X_t = -0.5\varepsilon_{t-1} + 0.8\varepsilon_{t-1}^2 + \varepsilon_t$
(e) Bilinear $X_t = \left(0.3 - 0.2\varepsilon_{t-1}\right)X_{t-1} + 1.0 + \varepsilon_t$

In all cases, process ε_t consisted of independent zero–mean Gaussian variables with unit variance. One hundred trials ($N = 100$) of this scheme were carried out for each scenario with three time series of length $T = 500$ generated from each model. Since all models are stationary in mean but present differences in scale, the series were previously normalized to have unit variance. The ARMA processes were generated using the R function *arima.sim* and the non-linear ones with self-programed code in R. A burn-in period of length 500 was considered in all cases, starting at $X_0 \sim \mathcal{N}(0, 1)$.

While clustering of linear models (Scenario 2.1) has been intensively studied and there are metrics specifically designed to deal with this kind of models, Scenario 2.2 introduces a major dificulty by including models with different conditional means that gradually depart from linearity. The models involved in Scenario 2.1 are similar to the ones previously considered by Maharaj (1996) by performing clustering of ARMA processes, and the models in Scenario 2.2 were used in a linearity test context by Tong and Yeung (1991).

To bring insight into the shapes of the true spectral density functions for the examined models, plots of the theoretical spectra for the ARMA models and large sample approximations to the corresponding spectra for the non-linear ones were obtained and depicted in Figure 2.1.

Plots in Figure 2.1 suggest that the *SP* algorithm should discriminate properly between the underlying processes. For the linear scenario, Figures 2.1(a), the theoretical patterns characterizing the clusters exhibit different profiles for the spectral densities and, in particular, well-separated peaks. The importance of the second stage in the *SP* algorithm is also evident in this scenario. Notice that the AR(2) and MA(2) models present peaks in frequencies close to each other, and therefore will be grouped together in the first stage.

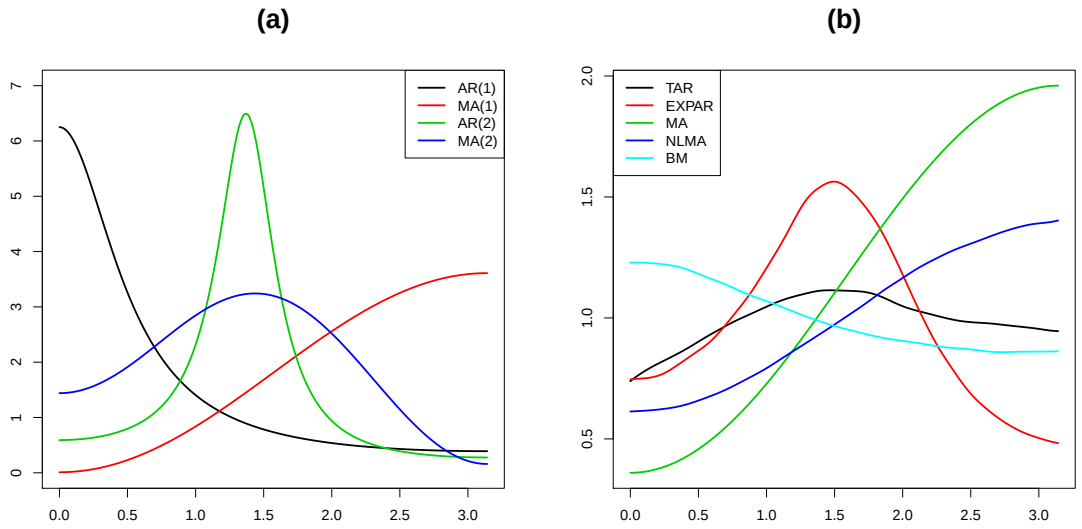

Figure 2.1: Theoretical spectral density functions for the models in the linear (a) scenario and large sample approximation of the spectral density functions for the models in the non-linear (b) scenario.

Nevertheless, they will be located into different clusters in the second stage when the areas on a local neighborhood are compared. Attending to the non-linear scenario, Figures 2.1 (b), discrimination between the EXPAR and BM models seems less hard than between the MA, TAR and NLMA models. Overall, the complexity of this scenario is greater and worse agreement indexes are expected.

For each data set generated, different metrics were considered specifically designed for time series clustering in order to compare the results with the proposed method.

- *Periodogram-based distances* (Caiado et al., 2006). In particular, the Euclidean distance between periodograms (d_P), ordinates of normalized periodograms (d_{NP}), log of peridograms (d_{LP}) and logarithm of normalized peridograms (d_{LNP}) were considered.

- *Autocorrelation-based distances* (Caiado et al., 2006). Direct and weighted Euclidean distances between simple and partial autocorrelations using a number of significant lags were taken into consideration, such as d_{ACFG}, d_{PACFG} and d_{PACFU} with $L = 10$ the number of significant lags considered. In particular, d_{ACFG} and d_{PACFU} were computed with $p = 0.05$.

- *Nonparametric dissimilarities in the frequency domain.* An spectral disparity measure defined as d_W in 1.1, where the densities ($d_{W(DLS)}$) were estimated by means of local

lineal smoothers of the periodograms, obtained via least squares (Kakizawa et al., 1998).

All of these metrics were compared with the proposed clustering algorithm. We denote by SP the spectral peak density method. For the computation of the algorithm, resamples of size $B = 200$ were generated. The estimation of the spectral densities required in the algorithm were carried out using the local linear smoother of the log-periodogram computed via least squares. Two different degrees of smoothing were tested for the implementation of the SP algorithm by considering $\hat{h} = \eta \hat{h}_{PI}$, with $\eta = \{1, 2\}$ and h_{PI} denoting the bandwidth selected via plug-in methodology. As for the significance levels used in the two stages of the algorithm, $\alpha_1 = \alpha_2 = 0.001$ were considered.

The SP algorithm automatically provides an estimate of the number of clusters that set up the partition. To make a fair comparison, two different criteria for determining the number of clusters are used with the rest of the metrics.

Starting from each dissimilarity matrix, a hierarchical clustering algorithm using average linkage method is applied. We consider two possible criteria for determining the number of clusters: (i) maximize the average silhouette coefficient (Kaufman and Rousseeuw, 1990) and (ii) maximize the Krzanowski-Lai index (Krzanowski and Lai, 1988). Both of them were defined in Section 1.4.

Other criteria for the selection of the number of clusters were considered but the best results were obtained using ASW and KL. A wide discussion on different methods for the selection of the number of clusters can be seen in Section 3.4 of Chapter 3.

The results of the cluster analysis were evaluated by comparing cluster solutions obtained experimentally with the true partition using the agreement index given by Liao (2005) (see Section 1.4).

2.3.2 Results

The results of the simulation study averaged over the $N = 100$ trials of the experiment are shown in Table 2.1.

According to Table 2.1, the algorithm based on the spectral peaks (SP) obtained the highest average scores in Scenario 2.2 and presented a little worse behaviour in Scenario 2.1. Nevertheless, this results are also competitive only being outperformed by metrics based on autocorrelations. Results improve when the value of the smoothing parameter is increased ($\eta = 2$). This seems reasonable because it makes the estimation of the spectrum

Table 2.1: Averages of the cluster similarity index and the number of clusters (between brackets) obtained from 100 trials of the simulation procedure for the classification of linear (Scenario 2.1) non-linear (Scenario 2.2) processes and each of the considered dissimilarity measures.

		Scenario 2.1		Scenario 2.2	
Method		ASW	KL	ASW	KL
Periodograms					
d_P		0.749 (2.36)	0.713 (5.21)	0.548 (2.22)	0.543 (7.38)
d_{LP}		0.750 (2.00)	0.701 (4.08)	0.554 (2.61)	0.536 (7.77)
d_{NP}		0.749 (2.36)	0.713 (5.21)	0.548 (2.22)	0.543 (7.38)
d_{LNP}		0.750 (2.00)	0.701 (4.08)	0.554 (2.61)	0.536 (7.77)
Autocorrelations					
d_{ACFG}		0.977 (3.79)	0.977 (3.93)	0.635 (2.22)	0.638 (2.37)
d_{PACFG}		0.911 (3.20)	0.962 (3.66)	0.621 (2.14)	0.642 (2.30)
d_{PACFU}		0.891 (3.04)	0.936 (3.43)	0.619 (2.10)	0.630 (2.31)
Non-parametric					
$d_{W(DLS)}$		0.916 (3.38)	0.921 (4.12)	0.671 (2.49)	0.685 (3.21)
Spectral peaks	η				
SP	1	0.863 (4.79)		0.685 (5.54)	
	2	0.933 (4.43)		0.743 (4.81)	

smoother and therefore, minimizes differences caused by errors in the estimation of the spectral peaks. Attending to the estimation of the real number of clusters, it can be seen that in both scenarios the total number is fairly well estimated with values really close to the real ones, being slightly more accurate in Scenario 2.2 reaching a value of 4.81 when $\eta = 2$ is considered.

The metrics based on autocorrelations performed very well in Scenario 2.1, with d_{ACFG} obtaining the best results in this scenario (0.977) regardless of the method employed for the estimation of the number of clusters, with an almost perfect estimation of the real number of clusters (3.93). The behavior of this metrics clearly worsened in Scenario 2.2 with really low classification indexes.

The non-parametric dissimilarity $d_{W(DLS)}$ produced very high average scores (0.921) in Scenario 2.1 with similar classification indexes despite of the system used for the selection of the number of clusters. As for Scenario 2.2 the behavior is similar to the autocorrelation-based metrics, the average scores clearly worsened with really low values, only achieving a value of 0.685 when the KL index is considered.

The remaining metrics based on periodograms produced the worst results in both scenarios. The periodograms are not able to separate properly the models considered in Scenario 2.2,

and only produced acceptable results in Scenario 2.1. Similar agreement indexes were obtained with the four versions of the periodogram metric being slightly better d_{LP} and d_{LNP}.

To shed light on what processes were more difficult to group, Tables 2.2 and 2.3 show the percentage of times that each of the processes was correctly grouped in each scenario.

Table 2.2: Percentage of times that the series of each ARMA process in Scenario 2.1 were correctly grouped in the experimental cluster solution.

	AR(1)	MA(1)	AR(2)	MA(2)
Periodograms				
d_P	41.00	43.00	8.00	17.00
d_{LP}	8.00	73.00	5.00	2.00
d_{NP}	41.00	43.00	8.00	17.00
d_{LNP}	8.00	73.00	5.00	2.00
Autocorrelations				
d_{ACFG}	98.00	99.00	84.00	84.00
d_{PACFG}	100.00	100.00	66.00	66.00
d_{PACFU}	100.00	100.00	43.00	43.00
Non-parametric				
$d_{W(DLS)}$	98.00	75.00	66.00	61.00
Spectral peaks				
SP	94.00	78.00	81.00	72.00

Table 2.2 shows that all the dissimilarities exhibited low ability to group correctly the series generated from the MA(1) and MA(2) processes, while the AR(1) and AR(2)) series form the most compact groups.

Table 2.3 corroborates the poor performance of all the metrics to cluster non-linear series. All the dissimilarities exhibited low ability to group correctly the series generated from all the processes except the EXPAR series which forms the most compact group. It also reveals that the SP algorithm showed the best performance, grouping correctly the series generated from the EXPAR model. Unlike SP, all the other measures were unable to detect homogeneity in the generating patterns of any of the series, and presented very poor success percentages. Definitively, the SP algorithm fairly outperformed the rest of dissimilarities in this setup.

Both Table 2.2 also corroborate the poor performance of the metrics based on periodograms cluster linear series and Table 2.3 shows the bad behavior of the metrics based on periodograms and autocorrelations when classifying non-linear series.

Finally, we record the number of correct clusters at each trial, i.e. the number of clusters

Table 2.3: Percentage of times that the series of each non-linear process in Scenario 2.2 were correctly grouped in the experimental cluster solution.

	TAR	EXPAR	MA	NLMA	BM
Periodograms					
d_P	0.00	4.00	3.00	2.00	2.00
d_{LP}	0.00	3.00	6.00	0.00	0.00
d_{NP}	0.00	4.00	3.00	2.00	2.00
d_{LNP}	0.00	3.00	6.00	0.00	0.00
Autocorrelations					
d_{ACFG}	0.00	18.00	3.00	0.00	4.00
d_{PACFG}	1.00	19.00	10.00	2.00	7.00
d_{PACFU}	0.00	14.00	6.00	0.00	5.00
Non-parametric					
$d_{W(DLS)}$	2.00	31.00	26.00	7.00	9.00
Spectral peaks					
SP	11.00	54.00	7.00	2.00	38.00

containing only the whole set of series with identical generating process. The distribution (in percentage) of this variable for each of the mentioned metrics and each scenario is depicted in Figure 2.2.

According to Figure 2.2 (a), the d_{ACFG} obtained the best result in Scenario 2.1, identifying the genuine solution of 4 clusters more times than the rest of dissimilarities, exactly 83% of the times. The proposed algorithm, SP , 56% of the times drawn out the four correct clusters becoming the third best metric only behind d_{ACFG} and d_{PACFG}. Also $d_{W(DLS)}$ obtained reasonable percentages of complete solutions, 43%. The metrics based on periodograms presented the poorest results in this scenario.

Figure 2.2 (b), show that the SP algorithm led to the best results in Scenario 2.2. While only d_{PACFG} was able to identify correctly the true solution of 5 clusters 1% of the times, the SP algorithm show a more consistent and better behavior identifying 3 clusters 8% of times. The proposed algorithm, usually identified one and two clusters (56% and 16%, respectively). Despite the fact that the d_{PACFG} detect one time the correct solution, 78% of times failed to detect any of the 5 models. The remaining dissimilarities yielded significantly worse results. Once again, the metrics based on periodograms presented the poorest results in this scenario. This results corroborate the difficulty of this specific scenario that was previously observed when we represented estimation of the theoretical spectral densities of each process.

The main limitation of the SP algorithm is the high computational complexity due to the

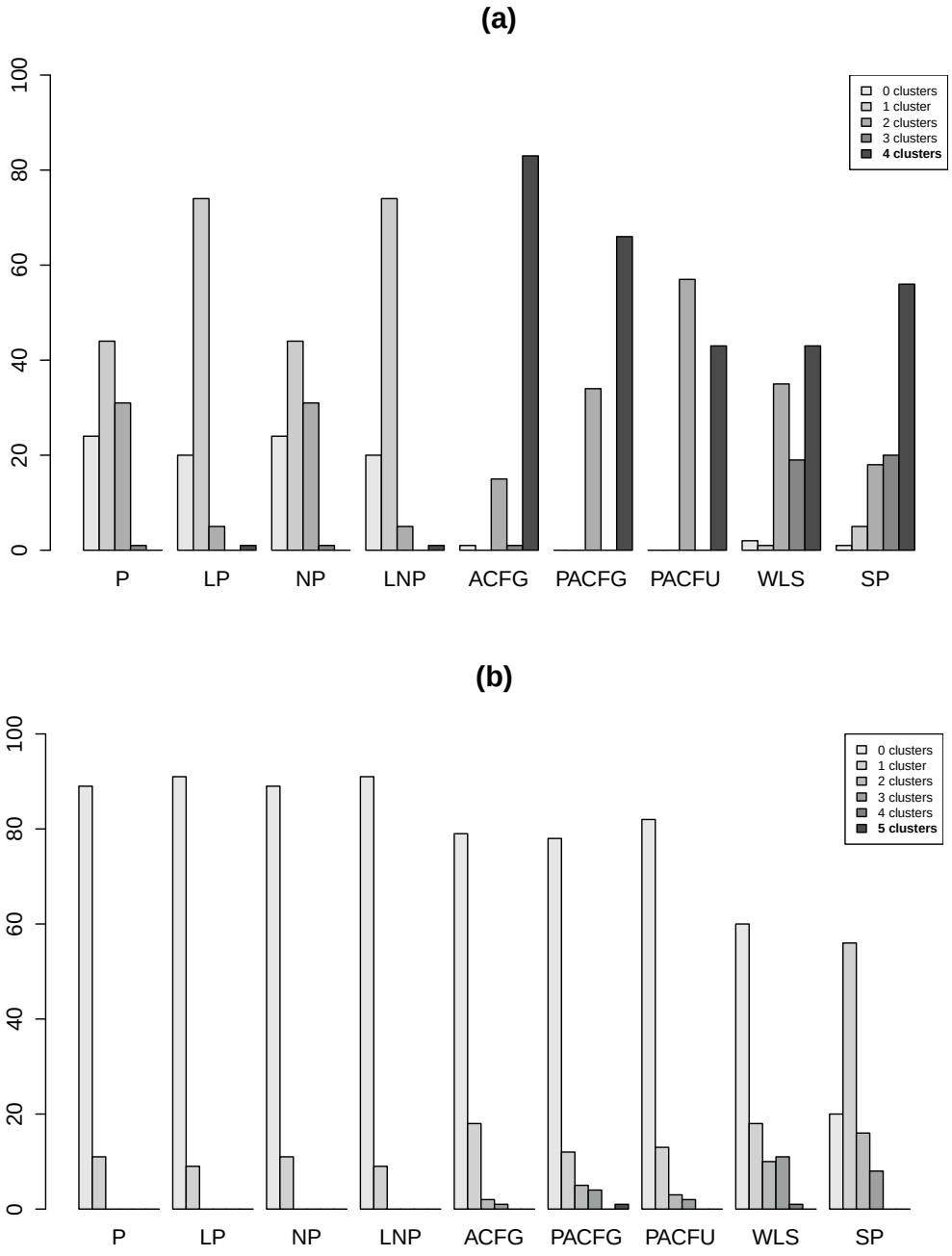

Figure 2.2: Distribution (in percentage) of the number of clusters identified at each iteration with different metrics for Scenarios 2.1 (a) and 2.2 (b). The true number of clusters at each scenario is shown in bold in the legends.

reiterated computation of both bootstrap resamples and numerical integration. To obtain accurate information about this, the computing time required for the SP algorithm at a particular iteration of the simulation has been measured. The algorithm was run on a

PC with the system specifications given by: Intel Core I7 - 3630QM processor, 2.4 Ghz CPU, 16 GB of RAM, Windows 10. For the linear scenario it took 17.74 minutes, while for the non-linear scenario 40.22 minutes were needed. Compared to the other distance-based models, these times are not competitive.

2.4 A case-study with real data

In this section we perform clustering on a real data example involving economic time series. The data set consists of a collection of 14 financial time series, each one recording the weekly bank share price (in euros) in the Spanish stock market over a period of two years (2001 and 2002).

The 14 banks considered are: Andalucía, Atlántico, BBVA, Banesto, Bankinter, Castilla, Crédito–Balear, Galicia, Guipuzcoano, Pastor, Santander, Valencia, Vasconia and Zaragozano. The measurements were recorded at the same time points for all series. In particular, each series consists on $T = 103$ weekly observations and each observation indicates the price per share taken on a Thursday. When Thursday fell on a public holiday, the observation was taken on a Wednesday.

Our purpose is to classify the 14 banks according to the maximum of their spectral densities. First of all, it is important to note that all series are non-stationary in mean. Following the usual approach each of the time series was transformed using logarithms and taking one regular difference. Graphs of the transformed series can be seen in Figure 2.3.

Just as in the simulation study, the spectral densities estimations were carried out by smoothing the associated periodogram using local polynomial estimation and a plug-in method to estimate the smoothing parameter. Figure 2.4 shows the representations of the periodograms and the spectral density for each bank.

Again, following the results in the simulations, we chose a significance level of $\alpha_1 = \alpha_2 = 0.01$ for the two stages of the SP algorithm and $B = 500$ bootstrap resamples of the periodograms were considered.

The clustering procedure lead to a 3 cluster solution. First group, $C_1 = \{\text{BBVA}, \text{Bankinter}, \text{Santander}\}$ forms a compact cluster, with some of the most important banks in Spain. All banks in this cluster, belong to IBEX-35 (which groups the 35 companies with the highest liquidity in the Spanish stock market) at the time the data was taken. The biggest cluster is the one formed by $C_3 = \{\text{Andalucía}, \text{Castilla}, \text{Crédito Balear}, \text{Galicia}, \text{Guipuzcoano}, \text{Valencia}, \text{Vasconia}\}$. As regards to the rest, a secondary cluster emerge from the classification:

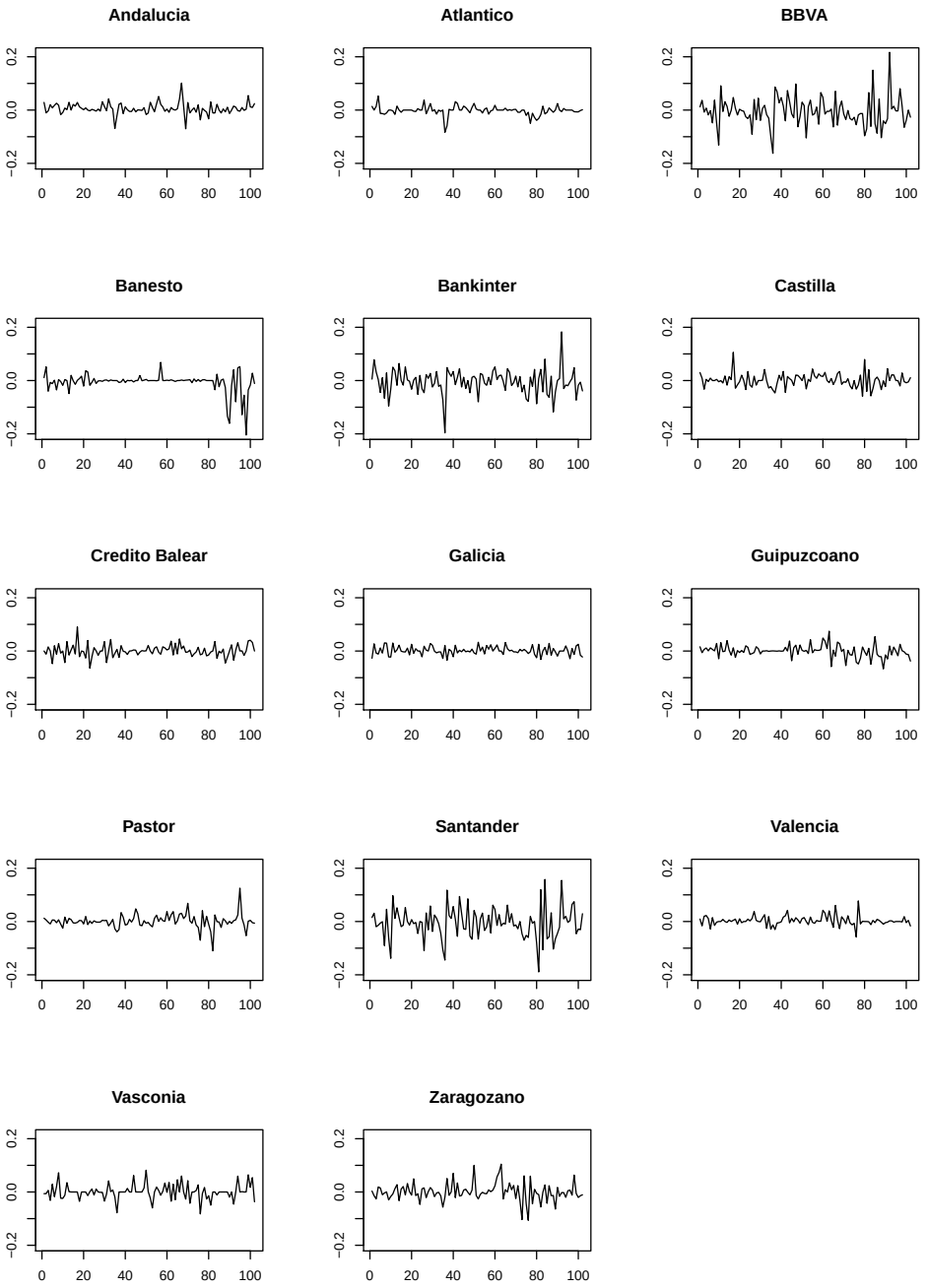

Figure 2.3: Transformed series of the weekly share price of the different banks in the Spanish stock market.

$C_2 = \{\text{Atlántico}, \text{Banesto}, \text{Pastor}, \text{Zaragozano}\}$.

To bring some insight into the classification, Figure 2.5 shows the banks spectral densities in each cluster. As it can be seen, banks in cluster C_2 (Figure 2.5 (b)) have the maximum

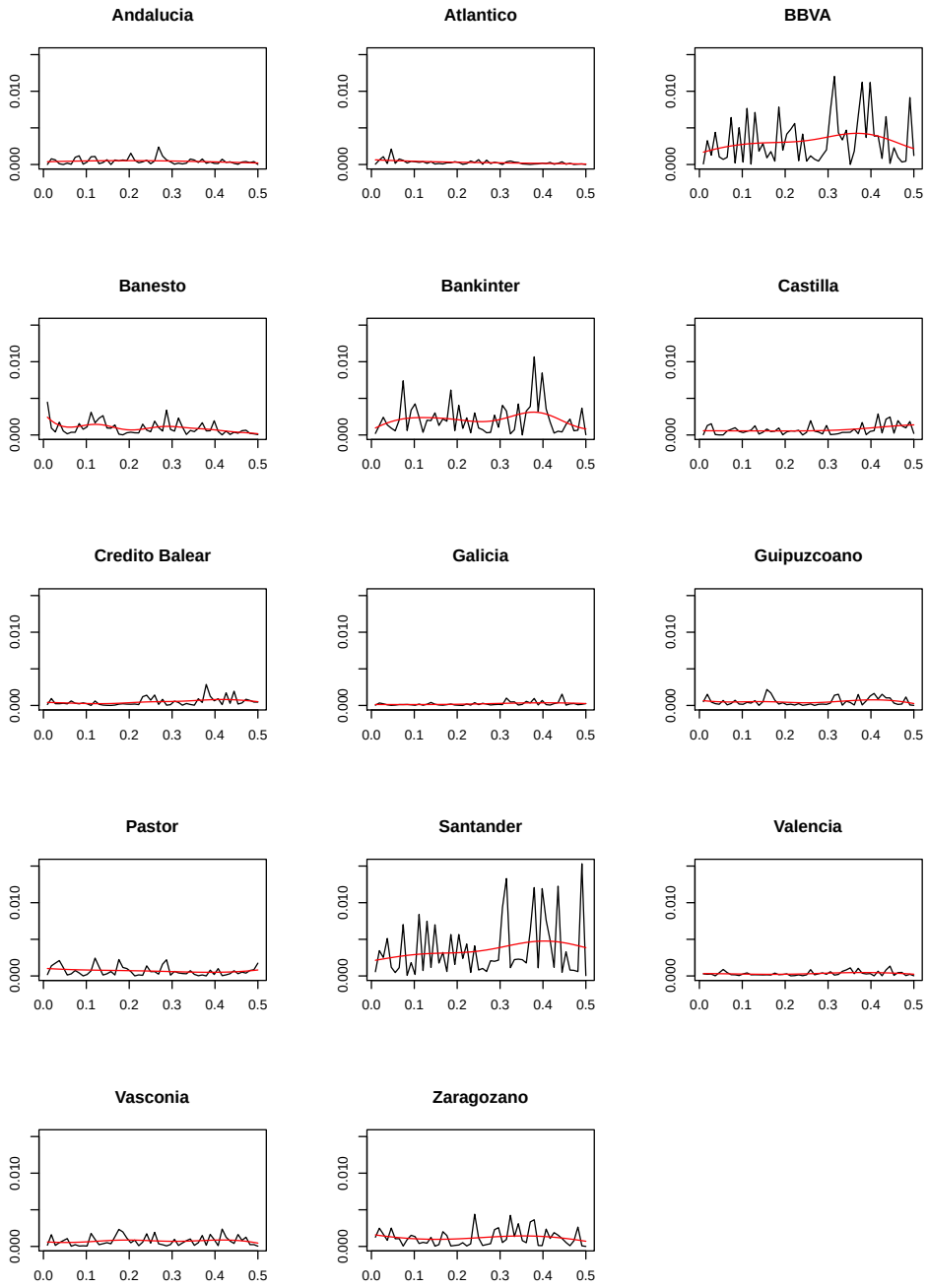

Figure 2.4: Periodograms and the spectral density of the weekly share price of the different banks in the Spanish stock market.

on the left bound of the interval while clusters C_1 and C_3 (Figures 2.5 (a) and (c)) have the peak around 0.4, but elements in cluster C_1 take higher values. Some of the banks in cluster C_3 (Figure 2.5 (c)) present spectral densities without any significative peak (their curves are almost flat). Also it is important to note that clusters C_1 and C_3 were not separated

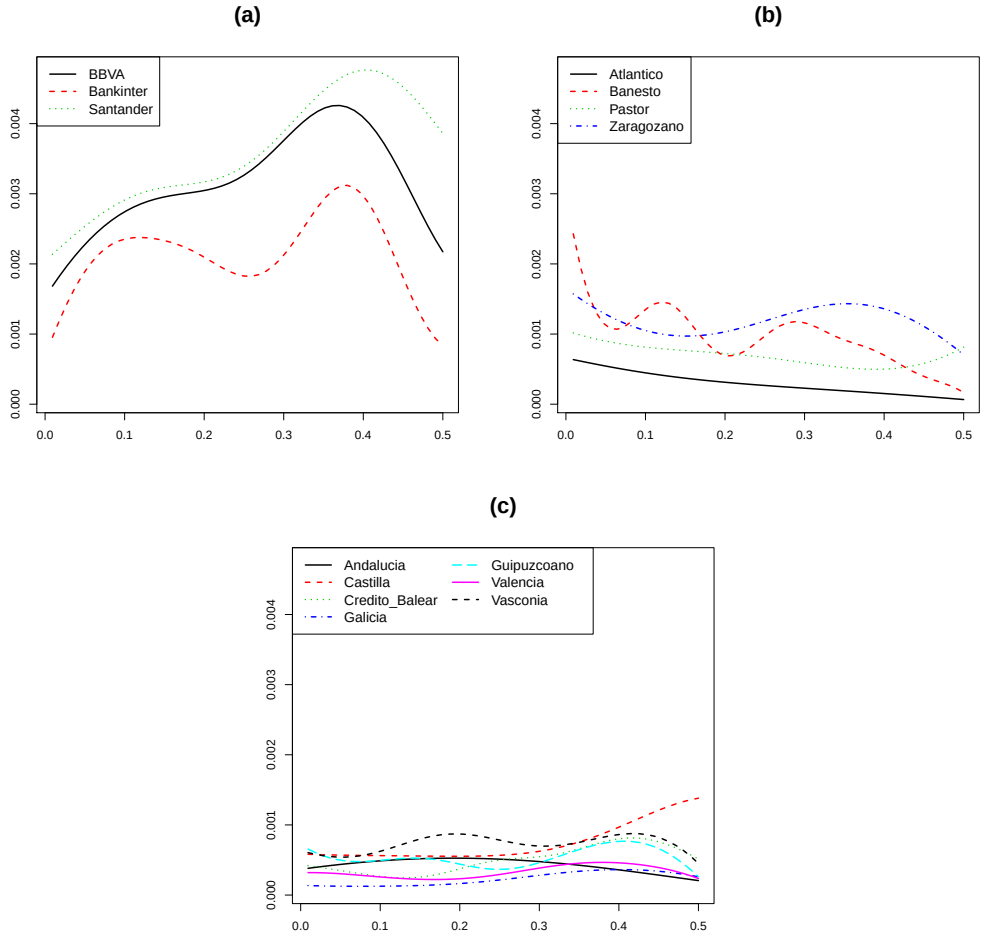

Figure 2.5: Spectral densities of the series of weekly share price of the different banks in the Spanish stock market in each cluster.

until the second stage of the algorithm came into play due to having the spectral peak at frequencies very close to each other.

2.5 Concluding remarks

A time series clustering procedure aimed at grouping series presenting the main spectral peaks at similar frequencies and with similar spectrum power has been introduced. This clustering principle is often of great interest in the analysis of oscillatory phenomena, where identifying serial realizations with similar/different predominant periods of oscillation and similarities/differences in terms of amplitude of these oscillations is a relevant issue. A typical example of application arises in the analysis of electrophysiological recordings like EEG signals, but also by investigating oscillatory signals in biological or financial problems.

Therefore, unlike the rest of chapters in this dissertation, the clustering purpose is not to grouping time series according the generating processes, and hence a comparison of the whole estimated spectra is not here the proper approach. Two spectra can be different but presenting similar dominant frequencies.

To attain the clustering target, a two-stage approach has been proposed. The key point consists of generating resamples of the periodograms based on the asymptotic χ^2 distribution, obtaining new estimated spectral densities, and hence: (i) measuring pairwise dissimilarities by means of the p-value from a bootstrap test of equality of peak frequencies (stage 1), and (ii) proceeding in a similar way checking by equality of spectrum power at the main peaks separately within each group formed in the first stage (stage 2). In both stages, a clustering procedure based on p-values is carried out.

The results from a simulation study showed a good performance of the proposed procedure, particularly in scenarios with non-linear series. Overall, the clustering behavior was highly competitive when compared to other approaches based on whole spectra or sequences of autocorrelations (which have been designed to attain a different clustering target). It was also observed that considering a high level of smoothing to estimate the spectral densities improves the classification. This criterion also allows to avoid the disruptive effect of non-significant peaks. Its application to a real study case involving financial series provides a cluster partition consistent with the main peaks and frequencies showed by the spectral estimates. In summary, the clustering procedure behaves in a promising way.

A number of strengths and weaknesses are inherent to the proposed clustering procedure. The most remarkable strength is that the procedure is specifically developed to attain the mentioned clustering target by directly aiming to identify peaks with similar frequency and amplitude. An additional advantage is that, by construction, the number of clusters is automatically determined once the significance levels for the hypothesis testing have been fixed. Other alternative procedures require to use some kind of criterion to establish the optimal number of clusters such as the average silhouette width or the Krazanowski-Lai index, among others. The main limitation is the high computational cost derived from the bootstrap procedures involved in both stages. In this sense, other clustering procedures are fairly preferred. As consequence of these considerations, the proposed procedure is particularly recommendable when the clustering purpose relies on the location and amplitude of the spectral peaks, the main peaks are clearly identifiable, and the number of series subjected to clustering is not too large.

Chapter 3

Clustering of time series based on quantile autocovariances

Contents

3.1 Introduction

The main motivation behind this chapter is to propose an innovative dissimilarity measure between time series in order to perform clustering governed by similarity between underlying dependence structures. The new measure should exhibit nice properties of robustness against the generating processes, thus enlarging the field of application to include complex scenarios but also producing competitive results in simpler scenarios. According to the

clustering purpose, a structure-based dissimilarity is required. Specifically, we focus on the feature–based approach, where the raw observations are replaced by a reduced number of features describing the temporal structure of the series, and then dissimilarity is evaluated in terms of these features. For instance, in the time domain, several authors have considered measures based on comparing estimations of simple or partial autocorrelation functions (Bohte et al., 1980; Caiado et al., 2006; D'Urso and Maharaj, 2009). Autocorrelations exhibit nice properties to discriminate between some kinds of processes (see Monte Carlo experiments in Caiado et al., 2006), but also present some weaknesses such as the lack of robustness to outliers and heavy tails or being unable to detect tail dependence. Note that heavy tails and non-existence of higher moments are distributional features frequently exhibited by, for instance, many financial time series (log-return series of stock indices, share prices, exchange rates, etc). In fact, several clustering approaches specifically developed to cluster financial time series have been currently introduced. For example, De Luca and Zuccolotto (2011) propose to use a tail dependence coefficient to group time series with an association between extremely low values, and D'Urso et al. (2013a) consider two fuzzy clustering procedures making use of GARCH models.

To overcome these limitations, we propose to measure dissimilarity comparing quantile autocovariance functions (see, e.g., Linton and Whang, 2007; Lee and Rao, 2012). For a given time series X_t, the quantile autocovariance function (QAF) is defined by means of the cross-covariances

$$\operatorname{cov}\left(I\left(X_t \leq x\right), I\left(X_{t+l} \leq y\right)\right),$$

where $I(\cdot)$ denotes the indicator function. The quantile autocovariances examine the general "pairwise" dependence structure (so-called serial dependence), i.e. the joint distribution of (X_t, X_{t+l}), thus allowing to account for sophisticated serial features that simple autocovariances are unable to detect. A detailed discussion on the advantages of the QAF and its representations in the frequency domain (quantile periodogram and quantile spectral density) compared to their respective classical counterparts can be seen in current references by Lee and Rao (2012), Hagemann (2013), Li (2014) and Dette et al. (2014). Furthermore, these works show the usefulness of the quantile versions in specific inference problems like testing for pairwise independence or for equality of serial dependence, and also modeling time series with time-dependent variance. Nevertheless, to the best of our knowledge, QAF has not been considered to perform time series clustering, even though it satisfies suitable properties to carry out this task, such as light computational complexity and robustness inherent to quantile methods. Moreover, unlike the usual autocovariance function, QAF is robust to the non-existence of moments, and thus a QAF-based dissimilarity should take advantage to discriminate between series generated from processes with different heavy-tailed

marginal distributions or following conditional heteroskedastic models.

The first objective in this chapter is introducing a QAF-based dissimilarity and then showing its high capability to cluster time series generated from a broad range of dependence models. We provide simulation results comparing this new metric with other alternative dissimilarities frequently used in time series clustering by using two different approaches: an hierarchical method in which each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy, and a partitioning around medoids (PAM) procedure (Kaufman and Rousseeuw, 1990), which returns a subset of series representative of the identified clusters (medoids). The attained results show the good behavior of the QAF-based metric compared to other commonly used dissimilarities. In particular, very good scores are reported by classifying heteroskedastic processes, which are frequently used with economic or financial indicators (Bauwens and Rombouts, 2007; Otranto, 2008; D'Urso et al., 2013a; Aielli and Caporin, 2014). Further, since Gaussian heteroskedastic models cannot often capture the asymmetry and leptokurtosis exhibited by some financial time series, e.g. log-return series of stock indices (Lazar and Alexander, 2006; Kipkoech, 2014), additional simulations based on heteroskedastic models with non-normal errors are performed attaining even better results.

An important issue in cluster analysis is to obtain an initial estimation for the number of clusters underlying the database. We propose to address this issue by adjusting the prediction-based resampling algorithm (so-called Clest) introduced by Dudoit and Fridlyand (2002). Clest is aimed to select the number of clusters k that provides the strongest evidence against the null hypothesis $H_0 : k = 1$. For each value of k, Clest evaluates the amount of reproducibility, say R_k, of the k-cluster solution combining ideas from supervised and unsupervised learning, and then examines whether the value of R_k is significantly larger than the expected one under the null hypothesis of no clusters. In the original procedure, the expected value for R_k under the null is approximated by resampling a multivariate uniform distribution. Nevertheless, this assumption is not reasonable when dependent data are considered. To overcome this drawback, the uniformity assumption under H_0 is marginally considered for each quantile autocovariance, i.e. the reference datasets are successively generated from univariate uniform distributions (Step 3 of the Clest algorithm in Section 3.4). The performance of this modified version of Clest algorithm and other existing methods is examined and compared by means of new Monte Carlo experiments. As it will be shown in Section 3.4.1, Clest algorithm produced accurate estimations of k and showed the most robust performance.

Other important contribution concerns the optimal selection of input parameters, i.e. establishing how many and which combinations of lags and quantile levels must be used to

define the QAF metric in order to optimize the clustering process. A proper adjustment of the variable selection algorithm proposed by Andrews and McNicholas (2014) for clustering and classification allows us to address this problem. Nevertheless, it is worth remarking that using a small number of quantiles with probability levels regularly spaced is enough to reach satisfactory results.

The rest of the chapter is organized as follows. Section 3.2 proposes to measure dissimilarity between a pair of observed series by comparing sequences of estimated quantile autocovariances. The estimation procedure is detailed, the asymptotic behavior established, and the dissimilarity defined and motivated. Section 3.3 focuses on the classification task following a hierarchical approach based on the introduced metric. The clustering behavior is analyzed throughout a simulation study where three classification scenarios featured by the kind of generating process are considered, namely linear, non-linear and conditional heteroskedastic models. The results with the proposed metric are compared with the ones obtained using other dissimilarity measures. The algorithm proposed to estimate the optimal number of clusters is described in Section 3.4, and its behaviour with finite samples is analyzed and compared with alternative procedures in a new simulation study. An application to cluster real time series involving volatility records of daily Euro exchange rates against other international currencies is presented in Section 3.5. Section 3.6 introduces an algorithm to select the optimal combinations of lags and pairs of quantile levels in order to perform clustering using the QAF-based dissimilarity. Section 3.7 focuses on the classification using a partitional aproach (PAM). Likewise the hierarchical procedure, the behavior in clustering of the QAF-based dissimilarity with the PAM procedure is examined in Section 3.7.1 by considering the same simulation scenarios but including different distributional forms for the errors. Finally, some concluding remarks are summarized in Section 3.8.

3.2 A dissimilarity measure between time series based on quantile autocovariances

3.2.1 The quantile autocovariance function

Let $X_1, \ldots, X_T$ be an observed stretch of a strictly stationary process $\{X_t;\ t \in \mathbb{Z}\}$. Denote by F the marginal distribution of X_t and by $q_\tau = F^{-1}(\tau)$, $\tau \in [0,1]$, the corresponding quantile function. Fixed $l \in \mathbb{Z}$ and an arbitrary couple of quantile levels $(\tau, \tau') \in [0,1]^2$, consider the cross covariance of the indicator functions $I\left(X_t \leq q_\tau\right)$ and $I\left(X_{t+l} \leq q_{\tau'}\right)$ given by

$$\gamma_l(\tau,\tau') = \operatorname{cov}\left\{I\left(X_t \leq q_\tau\right), I\left(X_{t+l} \leq q_{\tau'}\right)\right\} = \mathbb{P}\left(X_t \leq q_\tau, X_{t+l} \leq q_{\tau'}\right) - \tau\,\tau'. \tag{3.1}$$

Function $\gamma_l(\tau,\tau')$, with $(\tau,\tau') \in [0,1]^2$, is called *quantile autocovariance function (QAF) of lag l* and can be seen as a generalization of the classical autocovariance function. While the latter measures linear dependence between different lags by evaluating co-variability with respect to the average, the former studies the joint variability of the events $\{X_t \leq q_\tau\}$ and $\{X_{t+l} \leq q_{\tau'}\}$, i.e. examines how a part of the range of variation of X_t helps to predict whether the series will be below quantiles in a future time. By definition, QAF captures the sequential dependence structure of a time series, thus accounting for serial features related to the joint distribution of (X_t, X_{t+l}) that simple autocovariances cannot detect. Unlike the usual autocovariance function, QAF is well-defined even for processes with infinite moments and takes advantage from the local distributional properties inherent to the quantile methods, in particular showing a greater robustness against heavy tails, dependence in the extremes and changes in the conditional shapes (skewness, kurtosis), see Mikosch and Stărică (2000); Davis and Mikosch (2009); Lee and Rao (2012); Hagemann (2013); Li (2014); Dette et al. (2014). Based on these nice properties, QAF and its representations in the frequency domain (quantile periodogram and quantile spectral density) have been considered in several inference problems, including evaluation of directional predictability between time series (Linton and Whang, 2007; Han et al., 2016) and testing procedures for specific aspects of serial dependence such as interrelatedness, conditional homoscedasticity or conditional symmetry (Skaug and Tjøstheim, 1993; Hong, 2000; Kao et al., 2009).

An estimator of $\gamma_l(\tau,\tau')$ can be constructed replacing the theoretical quantiles by the corresponding empirical quantiles $\hat{q}_\tau$ and $\hat{q}_{\tau'}$ obtained from the observed realization $X_1,\ldots,X_T$. This way, the estimated QAF is given by

$$\hat{\gamma}_l(\tau,\tau') = \frac{1}{T-l}\sum_{t=1}^{T-l} I\left(X_t \leq \hat{q}_\tau\right) I\left(X_{t+l} \leq \hat{q}_{\tau'}\right) - \tau\,\tau', \tag{3.2}$$

where the empirical quantiles $\hat{q}_\alpha$, for $0 \leq \alpha \leq 1$, can be formally seen as the solution of the minimization problem (Koenker, 2005, page 7) given by

$$\hat{q}_\alpha = \arg\min_{q\in\mathbb{R}} \sum_{t=1}^{T} \rho_\alpha\left(X_t - q\right),$$

with $\rho_\alpha(x) = x(\alpha - I(x<0))$.

3.2.2 Asymptotic behavior

The asymptotic behavior of the sample quantile autocovariances defined by (3.2) is established in Theorem 3.2.1 of this section by following the asymptotic analysis developed by Han et al. (2016). Specifically, consider a two-dimensional strictly stationary process $\boldsymbol{X}_t = \{(X_{1t}, X_{2t})\,;\ t \in \mathbb{Z}\}$ with marginal distribution functions $F_i(\cdot)$ and quantiles $q_{i,\tau}$, for $i = 1, 2$ and $\tau \in (0, 1)$. Under general weak dependence conditions on $\boldsymbol{X}_t$, (Han et al., 2016, Th.1) obtain the asymptotic distribution of the sample cross-correlation between the events $I\left(X_{1t} \leq \hat{q}_{1,\tau}\right)$ and $I\left(X_{2(t+l)} \leq \hat{q}_{2,\tau'}\right)$, for arbitrary lag l and quantile levels τ and τ'. Adapting this result to the univariate setting and considering the non-normalized version of the mentioned cross-correlations, the limiting distribution stated in Theorem 3.2.1 for the sample quantile autocovariances is directly derived. First, some useful notation and the required assumptions are introduced.

Given an arbitrary lag l, let $\boldsymbol{\mathcal{A}} \equiv \mathcal{A}_t \times \mathcal{A}_{t+l}$ be a compact subset in $(0, 1)^2$, where $\mathcal{A}_t$ and $\mathcal{A}_{t+l}$ denote quantile ranges of interest for X_t and X_{t+l}, respectively. Denote by $F_l(\cdot, \cdot)$ the joint distribution of (X_t, X_{t+l}), and for $t = 1, \ldots, T$ and $(\tau, \tau') \in \boldsymbol{\mathcal{A}}$, consider the vector given by

$$\boldsymbol{\xi}_{t,l}\left(\tau, \tau'\right) = \left(I\left(X_t \leq q_\tau, X_{t+l} \leq q_{\tau'}\right) - F_l\left(q_\tau, q_{\tau'}\right),\, I\left(X_t \leq q_\tau\right) - \tau,\, I\left(X_{t+l} \leq q_{\tau'}\right) - \tau'\right)^t.$$

Now, define the three-dimensional mean-zero Gaussian process $\{\mathbb{B}_l\left(\tau, \tau'\right)\,;\ (\tau, \tau') \in (0, 1)^2\}$ having covariance matrix given by

$$\Gamma_l\left(\left(\tau_1, \tau_1'\right), \left(\tau_2, \tau_2'\right)\right) = \mathbb{E}\left(\mathbb{B}_l\left(\tau_1, \tau_1'\right) \mathbb{B}_l^t\left(\tau_2, \tau_2'\right)\right) = \sum_{t=-\infty}^{\infty} \operatorname{cov}\left(\boldsymbol{\xi}_{t,l}\left(\tau_1, \tau_1'\right), \boldsymbol{\xi}_{0,l}^t\left(\tau_2, \tau_2'\right)\right), \tag{3.3}$$

for $(\tau_i, \tau_i') \in \boldsymbol{\mathcal{A}}$, $i = 1, 2$.

The following conditions are assumed to hold.

A1. $\{X_t;\ t \in \mathbb{Z}\}$ is a strictly stationary and strongly mixing process with α-mixing coefficients satisfying $\alpha(n) = O(n^{-a})$, for $a > 1$.

A2. The marginal distribution $F(\cdot)$ has continuous density $f(\cdot)$, which is bounded away from 0 and ∞ at q_τ over $\tau \in \mathcal{A}_t \cup \mathcal{A}_{t+l}$.

A3. For any $\varepsilon > 0$ there exists a $\nu(\varepsilon)$ such that $\sup_{\tau \in \mathcal{A}_t \cup \mathcal{A}_{t+l}} \sup_{|s| \leq \nu(\varepsilon)} \left|f\left(q_\tau\right) f\left(q_\tau + s\right)\right| < \varepsilon$.

A4. The joint distribution $F_l(\cdot, \cdot)$ is continuously differentiable over the neighborhood of quantiles of interest.

Assumptions A1-A4 are mild regularity conditions and not too restrictive. While A1 entails a mixing condition for the dependence structure of X_t, A2 ensures that the quantile function is uniquely determined, and A2 and A4 impose enough smoothness and regularity for f and F_l, respectively. The weak convergence of the sample quantile autocovariance processes indexed by $(\tau, \tau') \in (0,1)^2$ is stated in Theorem 3.2.1 below.

Theorem 3.2.1 *Suppose that assumptions A1-A4 hold for a particular lag l. Then we have*

$$\sqrt{T}\left(\hat{\gamma}_l(\tau,\tau') - \gamma_l(\tau,\tau')\right) \Longrightarrow \boldsymbol{\lambda}^t_{l,(\tau,\tau')}\mathbb{B}_l\left(\tau,\tau'\right)$$

with

$$\boldsymbol{\lambda}_{l,(\tau,\tau')} = diag\left(1, \frac{1}{f\left(q_\tau\right)}, \frac{1}{f\left(q_{\tau'}\right)}\right)\begin{pmatrix} 1 \\ \nabla F_l\left(q_\tau, q_{\tau'}\right) \end{pmatrix}, \tag{3.4}$$

where $\nabla F_l\left(q_\tau, q_{\tau'}\right)$ denotes the gradient vector of $F_l(\cdot,\cdot)$ computed at $(q_\tau, q_{\tau'})$ and $\mathbb{B}_l\left(\tau,\tau'\right)$ is the above-mentioned zero-mean Gaussian process with covariance matrix given by (3.3).

Proof.

The convergence stated in Theorem 3.2.1 is established proceeding as in the proof of Theorem 1 of Han et al. (2016).

Consider an arbitrary lag l and a pair of levels of probability $(\tau,\tau') \in \boldsymbol{\mathcal{A}} \equiv \mathcal{A}_t \times \mathcal{A}_{t+l}$. According to the definition of quantile autocovariance in (3.1) and the corresponding estimator given in (3.2), we have

$$\sqrt{T}\left(\hat{\gamma}_l(\tau,\tau') - \gamma_l(\tau,\tau')\right) = \sqrt{T}\left[T^{-1}\sum_{t=1}^{T-l} I\left(X_t \leq \hat{q}_\tau, X_{t+l} \leq \hat{q}_{\tau'}\right) - F_l\left(q_\tau, q_{\tau'}\right)\right]. \tag{3.5}$$

Let $\mathbb{V}_{T,l}$ be the empirical process indexed by $(u,u')^T \in \mathbb{R}^2$ defined by

$$\mathbb{V}_{T,l}\left(u,u'\right) = T^{-1/2}\sum_{t=1}^{T-l}\varphi_t\left(u,u',l\right), \tag{3.6}$$

with $\varphi_t\left(u,u',l\right) = I\left(X_t \leq u, X_{t+l} \leq u'\right) - F_l\left(u,u'\right)$.

By adding and subtracting the term $T^{-1/2}F_l\left(\hat{q}_\tau, \hat{q}_{\tau'}\right)$ in (3.5), we obtain

$$\sqrt{T}\left(\hat{\gamma}_l(\tau,\tau') - \gamma_l(\tau,\tau')\right) = \mathbb{V}_{T,l}\left(\hat{q}_\tau, \hat{q}_{\tau'}\right) + \sqrt{T}\left[F_l\left(\hat{q}_\tau, \hat{q}_{\tau'}\right) - F_l\left(q_\tau, q_{\tau'}\right)\right]. \tag{3.7}$$

As $F_l(\cdot,\cdot)$ is differentiable by Assumption A4, the mean value expansion leads to

$$\sqrt{T}\left[F_l\left(\hat{q}_\tau, \hat{q}_{\tau'}\right) - F_l\left(q_\tau, q_{\tau'}\right)\right] = \nabla F_l\left(\overline{q}_\tau, \overline{q}_{\tau'}\right)^t \sqrt{T}\left(\hat{q}_\tau - q_\tau, \hat{q}_{\tau'} - q_{\tau'}\right), \tag{3.8}$$

uniformly in $(\tau, \tau') \in \boldsymbol{\mathcal{A}}$, where $\overline{q}_\alpha$ is between $\hat{q}_\alpha$ and q_α for $\alpha = \tau, \tau'$.

Under Assumptions A1-A4, similar arguments as those used in Theorem 7.3 of Rio (2000) allow to establish the weak convergence of $\mathbb{V}_{T,l}\left(u, u'\right)$ to the mean-zero Gaussian process $\mathbb{V}_{\infty,l}\left(u, u'\right)$ with covariance matrix given by

$$\begin{aligned}\Xi_l\left(\left(u_1, u_1'\right), \left(u_2, u_2'\right)\right) &= \mathbb{E}\left(\mathbb{V}_{\infty,l}\left(u_1, u_1'\right)\mathbb{V}_{\infty,l}\left(u_2, u_2'\right)\right) = \\ &= \sum_{t=-\infty}^{\infty} \operatorname{cov}\left[\varphi_t\left(u_1, u_1', l\right), \varphi_0\left(u_2, u_2', l\right)\right].\end{aligned} \tag{3.9}$$

Convergence from $\mathbb{V}_{T,l}$ to $\mathbb{V}_{\infty,l}$ is part of Lemma 1 in Han et al. (2016), and it is a key result in the proof. Note that when $\left(u_i, u_i'\right) = \left(q_{\tau_i}, q_{\tau_i'}\right)$, for $i = 1, 2$, then $\Xi_l\left(\left(u_1, u_1'\right), \left(u_2, u_2'\right)\right)$ is equivalent to the $(1,1)$-th element of the covariance matrix $\Gamma_l\left(\left(u_1, u_1'\right), \left(u_2, u_2'\right)\right)$ for the process $\mathbb{B}_l\left(\tau, \tau'\right)$ in (3.3).

Based on the mentioned Lemma 1, Han *et al.* prove that

$$\lim_{T\to\infty} P\left(\sup_{(\tau,\tau')\in\boldsymbol{\mathcal{A}}} \left|\, \mathbb{V}_{T,l}\left(\hat{q}_\tau, \hat{q}_{\tau'}\right) - \mathbb{V}_{T,l}\left(q_\tau, q_{\tau'}\right) \right|\right) = 0. \tag{3.10}$$

Combining (3.8) and (3.10), we can write (3.7) as follows.

$$\sqrt{T}\left(\hat{\gamma}_l(\tau, \tau') - \gamma_l(\tau, \tau')\right) = \mathbb{V}_{T,l}\left(q_\tau, q_{\tau'}\right) + \nabla F_l\left(q_\tau, q_{\tau'}\right)^t \sqrt{T}\left(\hat{q}_\tau - q_\tau, \hat{q}_{\tau'} - q_{\tau'}\right) + o_p(1), \tag{3.11}$$

uniformly in $(\tau, \tau') \in \boldsymbol{\mathcal{A}}$.

Define $\mathbf{W}_{T,l}\left(u_1, u_2\right) = \left(\mathbb{W}_{T,l}^{(1)}\left(u_1\right), \mathbb{W}_{T,l}^{(2)}\left(u_2\right)\right)^t$, where $\mathbb{W}_{T,l}^{(1)}\left(u_1\right) = \lim_{u_2\to\infty} \mathbb{V}_{T,l}\left(u_1, u_2\right)$ and $\mathbb{W}_{T,l}^{(2)}\left(u_2\right) = \lim_{u_1\to\infty} \mathbb{V}_{T,l}\left(u_1, u_2\right)$. Based on the Bahadur representation of the sample quantiles, it holds

$$\sqrt{T}\left(\hat{q}_\alpha - q_\alpha\right) = \frac{1}{f\left(q_\alpha\right)} \mathbb{W}_{T,l}^{(i)}\left(q_\alpha\right) + o_p(1), \tag{3.12}$$

for $i = 1, 2$ and uniformly in α. Expression (3.12) combined with (3.11) leads to

$$\sqrt{T}\left(\hat{\gamma}_l(\tau, \tau') - \gamma_l(\tau, \tau')\right) = \boldsymbol{\lambda}_{l,(\tau,\tau')}^t \boldsymbol{\nu}_{T,l}\left(\tau, \tau'\right) + o_p(1), \tag{3.13}$$

uniformly in $(\tau, \tau') \in \boldsymbol{\mathcal{A}}$, where $\boldsymbol{\nu}_{T,l}\left(\tau, \tau'\right) = \left[\mathbb{V}_{T,l}\left(q_\tau, q_{\tau'}\right), \mathbf{W}_{T,l}\left(q_\tau, q_{\tau'}\right)^t\right]^t$, and $\boldsymbol{\lambda}_{l,(\tau,\tau')}$ is given in (3.4).

Now, since the convergence of $\mathbb{V}_{T,l}(u,u')$ to $\mathbb{V}_{\infty,l}(u,u')$ leads to establish the finite dimensional distributions convergence of $\boldsymbol{\lambda}^t_{l,(\tau,\tau')}\boldsymbol{\nu}_{T,l}(\tau,\tau')$ over $(\tau,\tau') \in \boldsymbol{\mathcal{A}}$, it suffices to show the stochastic continuity of $\boldsymbol{\lambda}^t_{l,(\tau,\tau')}\boldsymbol{\nu}_{T,l}(\tau,\tau')$ to establish the convergence in Theorem 3.2.1. This can be attained by following exactly the same arguments in the proof of Theorem 1 in Han et al. (2016).

Based on the uniform boundedness of $\boldsymbol{\lambda}_{l,(\tau,\tau')}$ over $(\tau,\tau') \in \boldsymbol{\mathcal{A}}$, for any (α,α') and $(\beta,\beta') \in \boldsymbol{\mathcal{A}}$, we have

$$\begin{aligned}\left\|\boldsymbol{\lambda}^t_{l,(\alpha,\alpha')}\boldsymbol{\nu}_{T,l}(\alpha,\alpha') - \boldsymbol{\lambda}^t_{l,(\beta,\beta')}\boldsymbol{\nu}_{T,l}(\beta,\beta')\right\| \leq \\ C\left\|\boldsymbol{\nu}_{T,l}(\alpha,\alpha') - \boldsymbol{\nu}_{T,l}(\beta,\beta')\right\| + \left\|\boldsymbol{\lambda}_{l,(\alpha,\alpha')} - \boldsymbol{\lambda}_{l,(\beta,\beta')}\right\|\left\|\boldsymbol{\nu}_{T,l}(\alpha,\alpha')\right\|\end{aligned} \tag{3.14}$$

Let $\boldsymbol{\alpha} = (\alpha,\alpha')^t$ and $\boldsymbol{\beta} = (\beta,\beta')^t$ be two arbitrary elements in $\boldsymbol{\mathcal{A}}$ satisfying that $\|\boldsymbol{\alpha}-\boldsymbol{\beta}\| \leq \delta$, for some $\delta > 0$. Then, it necessarily follows that

$$\left\|(q_\alpha - q_\beta, q_{\alpha'} - q_{\beta'})^t\right\| \leq \tilde{\delta} = C_1\delta, \tag{3.15}$$

for some constant $C_1 > 0$. In fact, for an arbitrary coordinate $\star$, considering the definition of quantile and the Assumption A2, we have that

$$\alpha^\star - \beta^\star = \int_{q_{\beta^\star}}^{q_{\alpha^\star}} f(v)\,dv \geq |q_{\alpha^\star} - q_{\beta^\star}| \inf_{\gamma\in\mathcal{A}^\star} f(q_{\gamma^\star}) \geq |q_{\alpha^\star} - q_{\beta^\star}|\,C_2,$$

so that we can set $C_1 = C_2^{-1}$ in (3.15).

From (3.15) follows that

$$\begin{aligned}\sup_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\boldsymbol{\mathcal{A}},\,\|\boldsymbol{\alpha}-\boldsymbol{\beta}\|\leq\delta}\left\|\boldsymbol{\nu}_{T,l}(\alpha,\alpha') - \boldsymbol{\nu}_{T,l}(\beta,\beta')\right\| \leq \\ \sup_{\mathcal{U}(\tilde{\delta})}\left|\mathbb{V}_{T,l}(u,u') - \mathbb{V}_{T,l}(v,v')\right| + \sup_{\mathcal{U}(\tilde{\delta})}\left\|\mathbb{W}_{T,l}(u,u') - \mathbb{W}_{T,l}(v,v')\right\|\end{aligned}$$

where $\mathcal{U}(\tilde{\delta})$ is formed by the elements $\boldsymbol{u} = (u,u')^t$ and $\boldsymbol{v} = (v,v')^t$ in $\mathbb{R}^2$ such that $\|\boldsymbol{u}-\boldsymbol{v}\| \leq \tilde{\delta}$.

Now, from the stochastic equicontinuity of $\mathbb{V}_{T,l}(\cdot)$ and $\mathbb{W}_{T,l}(\cdot)$ follows that given positive constants η and ε, there exist a constant $\delta > 0$ such that

$$\lim_{T\to\infty} P\left(\sup_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\boldsymbol{\mathcal{A}},\,\|\boldsymbol{\alpha}-\boldsymbol{\beta}\|\leq\delta}\left\|\boldsymbol{\nu}_{T,l}(\alpha,\alpha') - \boldsymbol{\nu}_{T,l}(\beta,\beta')\right\| > \eta\right) < \varepsilon \tag{3.16}$$

Finally, Assumptions A2 and A4 ensure that $\sup_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\boldsymbol{\mathcal{A}},\,\|\boldsymbol{\alpha}-\boldsymbol{\beta}\|\leq\delta}\left\|\boldsymbol{\lambda}_{l,(\alpha,\alpha')} - \boldsymbol{\lambda}_{l,(\beta,\beta')}\right\| =$

$o(1)$. The convergence of $\mathbb{V}_{T,l}(\cdot)$ implies that $\sup_{(\alpha,\alpha')\in\mathcal{A}} \left\|\boldsymbol{\nu}_{T,l}\left(\alpha,\alpha'\right)\right\| = O_P(1)$. Both results together with (3.16) allow us to conclude the stochastic equicontinuity of $\boldsymbol{\lambda}^t_{l,(\tau,\tau')}\boldsymbol{\nu}_{T,l}\left(\tau,\tau'\right)$ in (3.15).

□

3.2.3 QAF-based dissimilarity

The great sensitivity of QAF to capture complex dynamic features also suggests high capability to discriminate between generating processes, and hence an interesting potential to be applied on clustering and classification problems (Lafuente-Rego and Vilar, 2016a). To illustrate this point, we have obtained the sample QAF and the sample ordinary autocovariances for series simulated from a Gaussian white noise process, a GARCH-type process and an exponential GARCH process with Gaussian innovations, respectively. Plots of the sample autocovariance function and $\hat{\gamma}_1(\tau,\tau')$, for $\tau = 0.1$, 0.5 and 0.9, are simultaneously depicted in Fig. 3.1 for the three series.

As the three processes are uncorrelated, the sample autocovariances in (a) are close to zero with differences simply due to the noise, and therefore the conventional autocovariances are not useful to discriminate between the generating processes. By contrast, QAF plots in panels (b)-(d) show structural differences enabling us to discriminate between the underlying processes. The graphs for the white noise are flat due to the independence, but this is not the case for the GARCH models, which are uncorrelated but not independent. For instance, the symmetry of the GARCH model produces a flat profile for $\hat{\gamma}_1(0.5,\cdot)$, indicating that if $\{X_t \leq q_{0.5}\}$ then $\{X_{t+1} \leq q_{0.5}\}$ and $\{X_{t+1} > q_{0.5}\}$ are events with equal probability. However, the asymmetry of the EGARCH model leads to a different profile for $\hat{\gamma}_1(0.5,\cdot)$ indicating that X_{t+1} likely takes values higher than X_t. On the other hand, unlike of the white noise, the heavy tails of the GARCH model are recognizable from $\hat{\gamma}_1(0.1,\cdot)$ and $\hat{\gamma}_1(0.9,\cdot)$ since large and small values at time t tend to remain that way at time $t+1$. In short, this simple example involving GARCH processes brings insight into the potential of the quantile autocovariances to detect distinct underlying processes, providing a more comprehensive understanding on the dependence structure than the traditional autocovariances.

These considerations strongly support the idea of measuring dissimilarity between a pair of times series $\boldsymbol{X}_t^{(1)}$ and $\boldsymbol{X}_t^{(2)}$ by comparing estimates of their quantile autocovariances over a common range of selected quantiles, such as we propose in Lafuente-Rego and Vilar (2016a). Specifically, each time series $\boldsymbol{X}_t^{(u)}$, $u = 1, 2$, is characterized by means of the

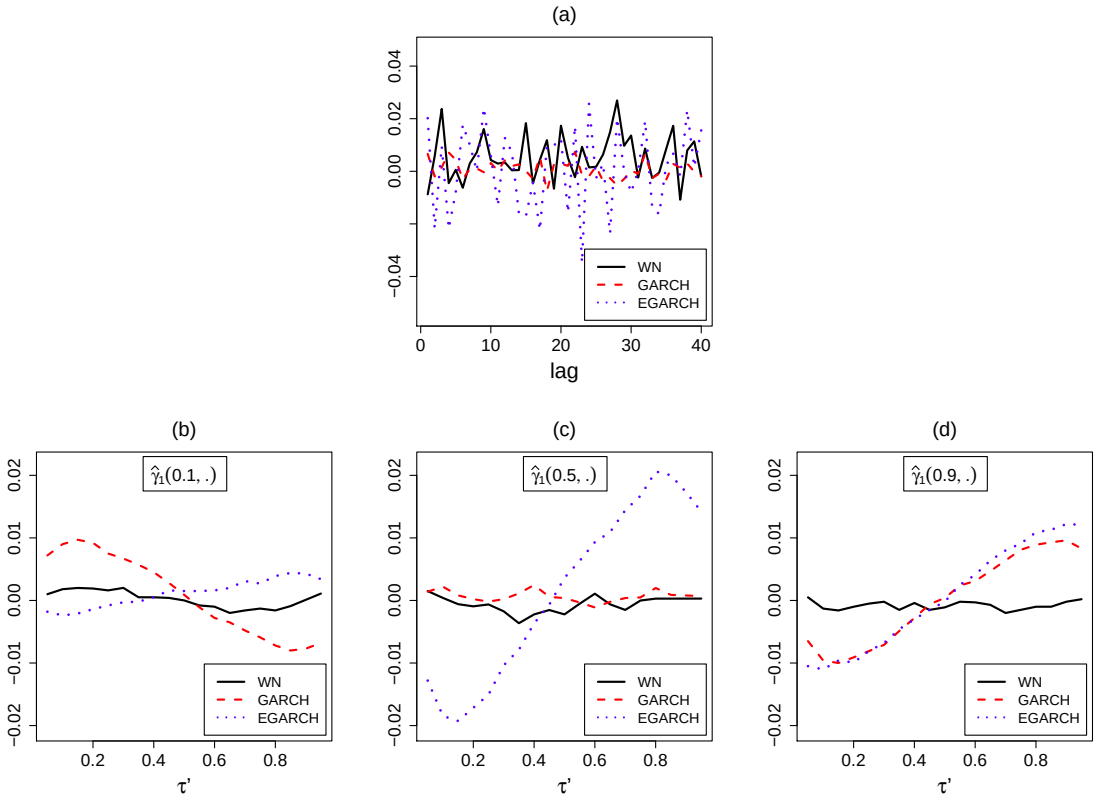

Figure 3.1: Sample autocovariances (a) and sample quantile autocovariances $\hat{\gamma}_1(\tau, \tau')$ for $\tau = 0.1$ (b), 0.5 (c) and 0.9 (d), obtained from simulated realizations of a Gaussian white noise process, a GARCH-type process and an exponential GARCH with Gaussian innovations.

vector $\mathbf{\Gamma}^{(u)}$ constructed as follows. For prefixed ranges of L lags, $l_1, \ldots, l_L$, and r quantile levels, $0 < \tau_1 < \ldots < \tau_r < 1$, the vector $\mathbf{\Gamma}^{(u)}$ is given by

$$\mathbf{\Gamma}^{(u)} = \left(\mathbf{\Gamma}_{l_1}^{(u)}, \ldots, \mathbf{\Gamma}_{l_L}^{(u)}\right), \tag{3.17}$$

where each $\mathbf{\Gamma}_{l_i}^{(u)}$, $i = 1, \ldots, L$, consists of a vector of length r^2 formed by re-arranging by rows the elements of the $r \times r$ matrix

$$\left(\hat{\gamma}_{l_i}^{(u)}(\tau_j, \tau_{j'})\right)_{j,j'=1\ldots,r}, \tag{3.18}$$

being $\hat{\gamma}$ the sample quantile autocovariance given in (3.2). This way, the dissimilarity between $\boldsymbol{X}_t^{(1)}$ and $\boldsymbol{X}_t^{(2)}$ is defined as the squared Euclidean distance between the corresponding representations $\mathbf{\Gamma}^{(1)}$ and $\mathbf{\Gamma}^{(2)}$, i.e.

$$d_{QAF}\left(\boldsymbol{X}_t^{(1)}, \boldsymbol{X}_t^{(2)}\right) = \left\|\boldsymbol{\Gamma}^{(1)} - \boldsymbol{\Gamma}^{(2)}\right\|^2 = \sum_{i=1}^{L}\sum_{j=1}^{r}\sum_{j'=1}^{r}\left(\hat{\gamma}_{l_i}^{(1)}(\tau_j, \tau_{j'}) - \hat{\gamma}_{l_i}^{(2)}(\tau_j, \tau_{j'})\right)^2 \quad (3.19)$$

By definition, d_{QAF} belongs to the class of dissimilarities based on comparing features extracted of the series instead of directly comparing the observed series. Others authors have proposed feature-based dissimilarities considering distances between simple or partial autocorrelations (Bohte et al., 1980; Galeano and Peña, 2000; D'Urso and Maharaj, 2009), ARMA representations (Piccolo, 1990; Maharaj, 1996, 2000), periodograms or log–periodograms (Caiado et al., 2006), cepstral coefficients Maharaj and D'Urso (2011) and other spectral features (Vilar and Pértega, 2004; Pértega and Vilar, 2010), among others. Obviously, all of these dissimilarities take advantage from the properties of the considered feature, and analogously d_{QAF} inherits the nice properties of the quantile autocovariances. In particular, the quantile autocovariance function is able to capture many types of serial dependence (including models with zero autocorrelation or exhibiting tail dependence) and exhibits robustness against outliers and heavy tails. From a practical point of view, it is also worthy remarking that d_{QAF} presents an efficient implementation at a very low cost in terms of computing time. Further, by construction, d_{QAF} can be evaluated on time series with unequal length. All of these interesting properties suggest that d_{QAF} has an enormous potential to perform time series clustering, and our results will corroborate this fact throughout a broad simulation study considering hierarchical and partitional cluster analysis.

To gain some insight into the usefulness of d_{QAF} in time series clustering, an illustrative example is presented below. Consider three different scenarios formed by two groups of fifteen simulated series of length $T = 500$. Each group is generated from different processes. Specifically, the confronted processes are:

Scenario A: A Gaussian white noise process against an AR-type process:

WN $X_t \sim \mathcal{N}(0,1)$
AR $Y_t = 0.5Y_{t-1} + \varepsilon_t$, $\varepsilon_t \sim \mathcal{N}(0,1)$

Scenario B. A Gaussian white noise process against a GARCH-type process:

WN $X_t \sim \mathcal{N}(0,1)$
GARCH $Y_t = \sigma_t\varepsilon_t$, $\sigma_t^2 = 0.1 + 0.7Y_{t-1}^2 + 0.2\sigma_{t-1}^2$, $\varepsilon_t \sim \mathcal{N}(0,1)$

Scenario C. An AR-type process against a GARCH-type process:

$$\begin{aligned} &\text{AR} && X_t = 0.1 + 0.5X_{t-1} + \varepsilon_t \\ &\text{GARCH} && Y_t = 0.1 + 0.5Y_{t-1} + a_t \\ & && a_t = \sigma_t\varepsilon_t,\ \sigma_t^2 = 0.1 + 0.7a_{t-1}^2 + 0.2\sigma_{t-1}^2,\ \varepsilon_t \sim \mathcal{N}(0,1) \end{aligned}$$

In this simple example, we focus on examining the pairwise distances between series. For each scenario, the distances between all pairs of series were obtained using two metrics, namely d_{QAF} and the squared Euclidean distance between autocorrelations (denoted by d_{ACF}). The first ten lags were used to compute d_{ACF}, while $r = 3$ quantiles of levels 0.1, 0.5 and 0.9 and only one lag ($L = 1$) were used to obtain d_{QAF}. The averages of the pairwise distances within each group and between series from different groups are reported in Table 3.1.

Table 3.1: Averages of pairwise distances for series within and between groups of Scenarios A, B and C.

	d_{QAF}	d_{ACF}
Scenario A (WN vs AR)		
Within WN group	0.0244205	0.0717162
Within AR group	0.0209015	0.1387937
Between groups	0.1126435	0.4254393
Scenario B (WN vs GARCH)		
Within WN group	0.0247152	0.0749827
Within GARCH group	0.0260225	0.1804724
Between groups	0.0393069	0.1290689
Scenario C (AR vs GARCH)		
Within AR group	0.0247897	0.1141849
Within GARCH group	0.0290222	0.2393515
Between groups	0.0352631	0.1742244

It is observed that d_{QAF} seems to present a high discriminatory power in the three considered scenarios. Note that by working with d_{QAF}, the average distance between groups is substantially greater than the ones within groups, presenting ratios around 4.6, 1.5 and 1.2 for Scenarios A, B and C, respectively. By contrast, d_{ACF} is unable to separate the processes forming Scenarios B and C, where GARCH models are included. Although the processes in these scenarios exhibit different dynamics, they present similar correlograms and therefore a based-autocorrelation metric easily fails to discriminate them. As expected, d_{ACF} properly works in Scenario A, where correlated and uncorrelated series are faced. Nevertheless, d_{QAF} also produces competitive results in this simple scenario, thus showing a promising property of flexibility to deal with different generating models.

Results from this simple example suggest that a more accurate clustering could be obtained if the quantile autocovariance function is used to set up the dissimilarity matrix. To support this intuition, a simulation study involving a broad range of different models and a number of existing dissimilarities is presented in the following section.

3.3 Hierarchical clustering based on quantile autocovariances: A simulation study

This section is devoted to examine the behavior of d_{QAF} in hierarchical clustering by means of simulated experiments. As mentioned, we must have in mind that the grouping principle is to bring together series with the same generating process.

An agglomerative hierarchical clustering procedure using the complete linkage method was carried out, although other linkage techniques provided similar results.

Simulations were designed to be able of comparing the performance of d_{QAF} with a wide selection of model-free and model–based dissimilarity measures. Specifically, three different classification setups were considered, namely classification of (i) ARMA models, (ii) non-linear models, and (iii) several structures of conditional heteroskedasticity. The generating models selected at each case are enumerated below.

Scenario 3.1 Classification of ARMA processes.

(a) AR(1) $X_t = 0.9X_{t-1} + \varepsilon_t$
(b) MA(1) $X_t = -0.7\varepsilon_{t-1} + \varepsilon_t$
(c) AR(2) $X_t = 0.3X_{t-1} - 0.1X_{t-2} + \varepsilon_t$
(d) MA(2) $X_t = 0.8\varepsilon_{t-1} - 0.6\varepsilon_{t-2} + \varepsilon_t$
(e) ARMA(1,1) $X_t = 0.8X_{t-1} + 0.2\varepsilon_{t-1} + \varepsilon_t$

Scenario 3.2 Classification of non-linear processes.

(a) NLMA $X_t = -0.5\varepsilon_{t-1} + 0.8\varepsilon_{t-1}^2 + \varepsilon_t$
(b) EXPAR $X_t = \left[0.3 - 10\,\exp\left(-X_{t-1}^2\right)\right] X_{t-1} + \varepsilon_t$
(c) TAR $X_t = 0.5X_{t-1}I\left(X_{t-1} \leq 0\right) - 2X_{t-1}I\left(X_{t-1} > 0\right) + \varepsilon_t$

To make more complex the clustering task in this scenario, we have added series generated from the following linear model.

(d) MA $X_t = -0.4\varepsilon_{t-1} + \varepsilon_t$

Scenario 3.3 Classification of conditional heteroskedastic processes. Consider the linear model $X_t = 0.5a_{t-1} + a_t$, with the error term satisfying $a_t = \sigma_t\varepsilon_t$, where the variance

at time t, σ_t^2, is conditional on observations at $t-1$ by means of some of the following models.

(a) ARCH $\sigma_t^2 = 0.2 + 0.95a_{t-1}^2$

(b) GARCH $\sigma_t^2 = 0.2 + 0.05a_{t-1}^2 + 0.9\sigma_{t-1}^2$

(c) GJR–GARCH $\sigma_t^2 = 0.2 + (0.05 + 1.2N_{t-1})\, a_{t-1}^2 + 0.1\sigma_{t-1}^2$, with $N_{t-1} = I\,(a_{t-1} < 0)$

Likewise the above scenario, we included a linear model MA(1) given by

(d) MA $X_t = 0.5\varepsilon_{t-1} + \varepsilon_t$

In all cases, process ε_t consisted of independent zero–mean Gaussian variables with unit variance. The linear and non-linear processes were generated as in Section 2.3, and the heteroskedastic ones using self-programed code in R. Again, a burn-in period of length 500 was used starting from $X_0 \sim \mathcal{N}(0, 1)$.

While clustering of linear models (Scenario 3.1) has been intensively studied and there are metrics specifically designed to deal with this kind of models, Scenario 3.2 introduces a major difficulty by including models with different conditional means that gradually depart from linearity. Scenario 3.3 proposes a more challenging task by involving models with non-constant volatility. The autoregressive conditional heteroskedasticity models ARCH and GARCH are able to capture both time-varying volatility clustering and some amount of fat-tailedness of the distribution, features frequently exhibited for returns on assets. Unlike of the GARCH models, the Glosten-Jagannathan-Runkle GARCH (GJR-GARCH) models allow to capture asymmetric effects on the conditional variance due to positive or negative past values, taking into account the leverage effect observed in many financial series.

As far as the dissimilarities to be compared, our selection must take into account the clustering purpose. We are not interested in measuring proximity between geometric profiles of series, thus shape-based dissimilarities (e.g. L_p distances) are not useful here because of clustering would be governed by local fluctuations, that is by the noise. Our aim is to bring together series generated from the same model. Hence, the selected metrics must capture differences between high level dynamic structures, which describe the global performance of the series. Given the parametric models chosen to set up the simulation scenarios, it is expected that some commonly used model- and feature-based distances work fine in our experiments, at least in Scenario 3.1. We decided to examine a wide range of dissimilarities, including measures comparing: estimated autocorrelations and partial autocorrelations, cross-correlations, periodograms, nonparametric spectral estimators, fitted ARMA models and cepstral coefficients, among other. All of these measures were computed using the R

package **TSclust** (see Chapter 1). We limit our report to the set of dissimilarities producing the best results in our numerical experiments, which are enumerated below.

- *Periodogram-based distances* (Caiado et al., 2006). Euclidean distances between periodograms, log—periodograms, normalized periodograms and log-–normalized periodograms were checked, reporting in this section the results for the Euclidean distance between log–periodograms, denoted by d_{LP}.

- *Autocorrelation-based distances* (Caiado et al., 2006). Direct and weighted Euclidean distances between simple and partial autocorrelations using a number of significant lags were taken into consideration. Results showed here correspond to the weighted Euclidean distance between partial autocorrelations (d_{PACFG}) based on a number of 10 lags and with weights ω_i decaying with the lag in the form $\omega_i = \pi(1-\pi)^i$, with $\pi = 0.5$.

- *Model–based distances.* The AR metric introduced by Maharaj (1996) and denoted by d_M.

- *Nonparametric dissimilarities in the frequency domain.* Although several metrics were considered within this group, we focus on the results attained with the integrated squared difference between estimated log-spectra (d_{ISD}) proposed by Pértega and Vilar (2010).

All of these metrics were compared with the proposed metric d_{QAF}. In our experiments, $r = 3$ quantiles of levels 0.1, 0.5 and 0.9 and only one lag ($L = 1$, with $l_1 = 1$) were considered to compute d_{QAF}. Note that except for two models in Scenario 3.1, all the remaining models present one significant lag, thus accounting for our choice $L = 1$.

Each pairwise dissimilarity matrix is then processed by an agglomerative hierarchical clustering algorithm using the complete linkage method.

The Monte Carlo study was conducted as follows. For each scenario, five time series of length $T = 200$ for the linear and non–linear setups and length $T = 1000$ for the case of conditionally heteroskedastic series are generated from each model, thus providing a sample set of labeled series available to perform clustering. Larger realizations were necessary with heteroskedastic models in order to estimate the quantile autocovariances with higher accuracy. Pairwise dissimilarity matrices are obtained for each set of series using the dissimilarities summarized below.

The algorithm output was the resulting partition, let us say $\mathcal{R} = \{R_1, \ldots, R_C\}$. Next step consisted in measuring the quality of the clustering procedure by means of two indexes

of agreement between the true cluster partition, $\mathcal{T} = \{T_1, \ldots, T_C\}$, and the experimental partition $\mathcal{R}$. Note that, according to the clustering target, each element T_i in $\mathcal{T}$ is a cluster formed by all the series generated from the same model, and hence the true partition is known. The two selected criteria take into account this fact and are described below.

The first considered agreement index (Gavrilov et al., 2000; Liao, 2005), Ind_1, was defined in Section 1.4. The second index, Ind_2, is the well-known adjusted Rand index (Hubert and Arabie, 1985), a corrected-for-chance version of the Rand index (Rand, 1971) which computes the proportion of pairs of series that are located together in the same or different clusters for both partitions. The adjusted Rand index modifies the Rand index in such a way that its expected value is equal to zero when the partitions are picked up at random (according to a generalized hypergeometric model) and the number of series in the clusters remain fixed. Likewise Ind_1, the maximum value of Ind_2 is 1 and it is attained when partitions agree perfectly. Nevertheless, the adjusted Rand index typically takes values substantially lower than other agreement indexes, even occasionally negative values, and it is known to exhibit a greater sensitivity on the cluster stability than other indexes.

Besides Ind_1 and Ind_2, we have also calculated a third index (Ind_3) using the one-nearest-neighbour (1-NN) classifier evaluated by leave-one-out cross-validation. Specifically, Ind_3 returns the proportion of series correctly classified when each series has been assigned to the element of $\mathcal{T}$ containing the nearest series according to the considered dissimilarity. Notice that Ind_3 does not evaluate the clustering algorithm, but providing insight into the efficacy of each of the used dissimilarities. This evaluation criterion has been intensively used in a broad range of pattern recognition applications, including time series clustering (see e.g. Keogh and Kasetty, 2003).

This simulation procedure was replicated $N = 100$ times for each scenario, and the cluster similarity indexes obtained with each dissimilarity were averaged over the 100 trials.

According to results in Table 3.2, the dissimilarity based on quantile autocovariances d_{QAF} produced the highest average scores in Scenarios 3.2 and 3.3, and presented worse behaviour in Scenario 3.1. d_{QAF} always led to clustering quality indexes above 0.9 in Scenario 3.2, with values GI and loo1NN very close to 1. With the ARMA series, d_{QAF} outperformed the metrics based on simple autocorrelations and periodograms, with quality indexes reasonably high but lower than the ones obtained with the rest of dissimilarities.

As expected, the metric based on ARMA models, d_M, is obviously affected by model misspecification, and hence it performed well in Scenario 3.1 but produced poor results in Scenarios 3.2 and 3.3.

The non-parametric dissimilarity, d_{ISD}, performed fairly well in Scenarios 3.1 and 3.2. This

Table 3.2: Averages and standard deviations (in brackets) of the cluster similarity indexes obtained from 100 trials of the simulation procedure for Scenarios 3.1, 3.2 and 3.3 and each of the considered dissimilarity measures.

	Scenario 3.1			Scenario 3.2			Scenario 3.3		
Measure	Ind_1	Ind_2	Ind_3	Ind_1	Ind_2	Ind_3	Ind_1	Ind_2	Ind_3
d_{LP}	0.763	0.614	0.742	0.713	0.501	0.675	0.417	0.006	0.225
	(.060)	(.107)	(.086)	(.110)	(.165)	(.106)	(.056)	(.056)	(.073)
d_{PACFG}	0.927	0.857	0.935	0.667	0.397	0.613	0.429	0.043	0.252
	(.071)	(.114)	(.058)	(.093)	(.136)	(.146)	(.058)	(.066)	(.105)
d_M	0.902	0.842	0.959	0.680	0.453	0.746	0.416	0.045	0.273
	(.094)	(.135)	(.047)	(.094)	(.135)	(.114)	(.053)	(.053)	(.108)
d_{ISD}	0.910	0.847	0.943	0.916	0.826	0.919	0.424	0.061	0.280
	(.083)	(.109)	(.048)	(.079)	(.130)	(.075)	(.052)	(.064)	(.107)
d_{QAF}	0.817	0.683	0.802	0.961	0.917	0.980	0.751	0.604	0.724
	(.062)	(.086)	(.060)	(.061)	(.101)	(.032)	(.053)	(.070)	(.100)

measure takes advantage of its nonparametric nature, being free of the linearity restriction, and hence its good behaviour. Nevertheless, the results worsened substantially by classifying heteroskedastic models. In fact, d_{QAF} noticeably outperforms d_{ISD} in Scenario 3.3.

The remaining metrics based on autocorrelations and periodograms produced worse results, corresponding the worst indexes to the periodogram-based measure. Unlike the quantile autocovariances, the PACF is not able to separate properly the models considered in Scenarios 3.2 and 3.3 and only produced good results in Scenario 3.1. This result corroborates the intuition suggested from the illustrative example considered at the end of Section 3.2.3.

In order to illustrate graphically the above comments, Figure 3.2 shows boxplots based on the cluster similarity indexes from the 100 simulated trials.

Boxplots in Figure 3.2(b) and (c) corroborate the good performance of d_{QAF} in Scenarios 3.2 and 3.3. In Scenario 3.3 (Figure 3.2(c)), d_{QAF} clearly appears like the best performed dissimilarity regardless of the considered index. In Scenario 3.2 (Figure 3.2(b)), with nonlinear models, the nonparametric dissimilarity d_{ISD} also attains very good average scores. Nevertheless, compared to the nonparametric competitors, d_{QAF} presents smaller standard deviations. Furthermore, d_{QAF} seems to take a substantial advantage in this scenario as the `loo1NN` index is considered, which is especially interesting because this goodness-of-assignment criterion directly evaluates the efficacy of the dissimilarity measure regardless of the considered clustering algorithm. On the other hand, unlike d_{ISD}, dissimilarity d_{QAF} is computationally efficient, thus enabling us to perform clustering on large databases including very long series. For instance, d_{ISD} involves numerical integration of differences between local linear smoothers computed by maximum local likelihood, which implies to

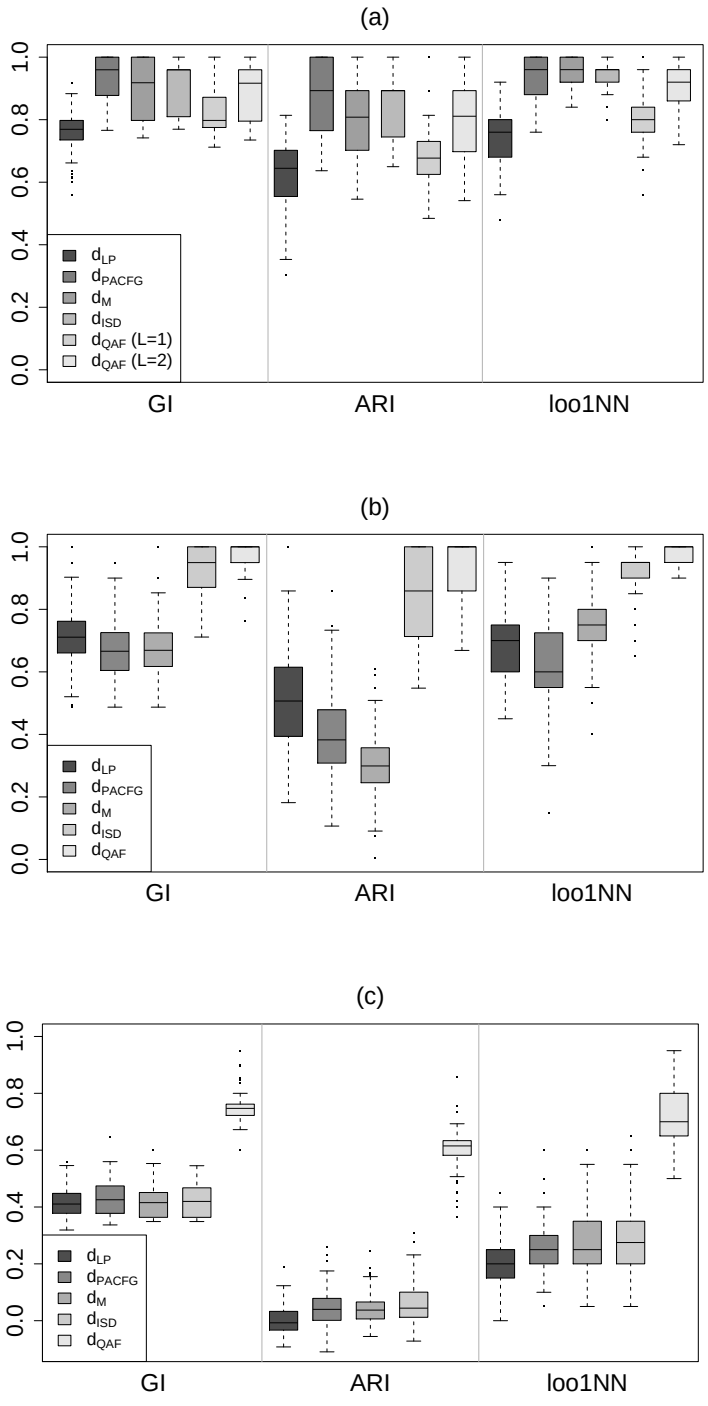

Figure 3.2: Boxplots of the cluster similarity indexes obtained from 100 trials of the simulation procedure for Scenarios 3.1 (a), 3.2 (b) and 3.3 (c) and a relevant subset of the dissimilarity measures.

solve repeatedly an optimization problem in two variables. This computational complexity could do unfeasible to perform clustering on large databases. Thus, computational efficiency is an additional strength of our proposal respect to the main competitors in Scenarios 3.2 and 3.3.

Boxplots in Figure 3.2(a) confirm that the worst performance of d_{QAF} occurs in Scenario 3.1, with linear models. Here, partial autocorrelations and Maharaj's distance (specifically designed to deal with this kind of processes) work fine. Nevertheless, Figure 3.2(a) also shows a noticeable improvement of d_{QAF} when two lags ($L = 2$) are used to construct this dissimilarity, which is not surprising because two of the models in Scenario 3.1 exhibit two significant lags. This way, d_{QAF} attains competitive scores in its worst scenario as the number of lags is correctly established.

To gain further insight into the clustering procedure with each metric, the experimental solutions were individually examined. We record the number of correct clusters at each trial, i.e. the number of clusters containing only the whole set of series with identical generating process. The distribution (in percentage) of this variable for each of the mentioned metrics and each scenario is depicted in Figure 3.3.

According to Figure 3.3 (a), the Maharaj distance, d_M, obtained the best result in Scenario 3.1, identifying the genuine solution of 5 clusters more times than the rest of dissimilarities, exactly 35% of the times. This fact corroborates the good clustering behaviour by using model-based metrics when the model is adequately specified. Among the model-free dissimilarities, the non-parametric d_{ISD} and the metric based on partial autocorrelations d_{PACF} obtained reasonable percentages of complete solutions, 21% and 28%, respectively. Nevertheless, three correct clusters were frequently determined with these metrics. The proposed metric, d_{QAF}, usually identified two clusters, although often moved between one and three, and just one percent of the times drawn out the five correct clusters. The metric based on periodograms presented the poorest results in this scenario.

Figure 3.3 (b) shows that the proposed metric d_{QAF} led to the best results in Scenarios 3.2, identifying the largest number of correct clusters. Clustering non-linear processes, d_{QAF} was able to obtain the whole solution around 59% of the trials, mixing some series of two different processes in the remaining iterations to form only two correct clusters. The remaining dissimilarities yielded significantly worse results. Only the non-parametric dissimilarity d_{ISD} was able to generate the full correct solution at any iterations (26%).

Figure 3.3 (c) shows the real complexity of heteroskedastic scenarios. None of the presented metrics were able to correctly classify all groups on any occasion. Only the QAF–based metric was able to correctly identify in some case two of the four clusters. The rest of the

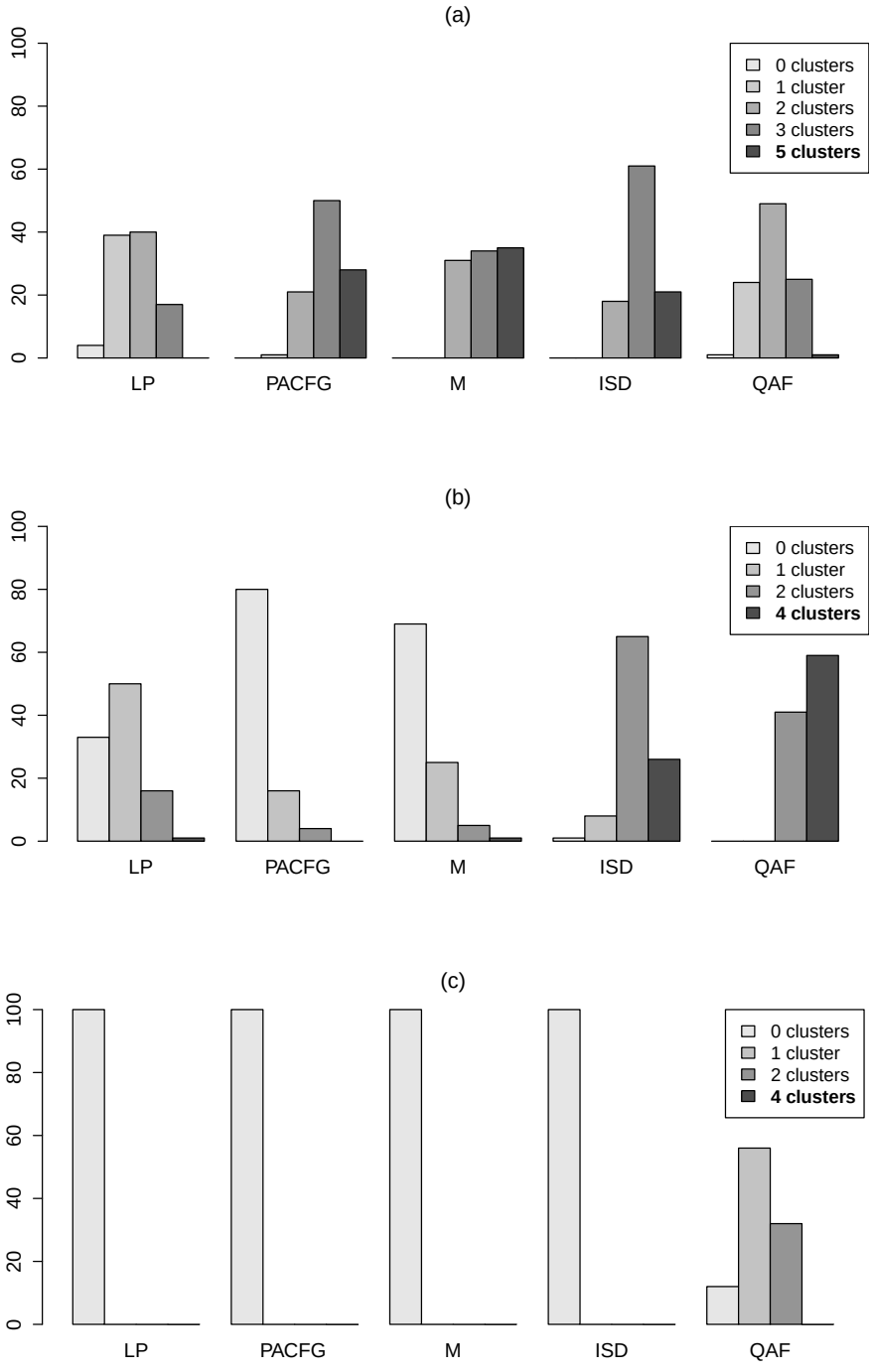

Figure 3.3: Distribution (in percentage) of the number of clusters identified at each iteration with different metrics for Scenarios 3.1 (a), 3.2 (b) and 3.3 (c). The true number of clusters at each scenario is shown in bold in the legends.

presented metrics clearly failed in the classification of heteroskedastic time series.

3.4 A procedure to estimate the optimal number of clusters

In this section, we address the problem of estimating the number of clusters. As the dissimilarity based on quantile autocovariances reported good results in our clustering experiments, we also consider this dissimilarity principle to determine the number of clusters. Specifically, we adopt the prediction–based resampling algorithm, so-called Clest, introduced by Dudoit and Fridlyand (2002), but carrying out slight modifications in order to use d_{QAF} and tackle the dependence of the quantile autocovariances.

Clest is aimed to select the value of k, $2 \leq k \leq K$, with $K \leq p$ denoting the maximum possible of clusters, that provides the strongest evidence against the null hypothesis of no clusters, $H_0 : k = 1$. For each value of k, Clest determines partitions of size k using supervised and unsupervised learning and evaluates the amount of agreement, say R_k, between both partitions. High agreement indexes mean high capability to reproduce the cluster structure. Then, a resampling procedure is used to examine whether the value of R_k is significantly larger than the expected one under a suitable distribution when $k = 1$. The value of k yielding the largest significance is established as the estimated number of clusters, $\hat{k}$.

To adapt the Clest algorithm to our framework including time series, some adjustments might be appropriate. First, the unsupervised partitions involved in Clest should use a proper dissimilarity between time series. In this point, we propose to use the dissimilarity based on quantile autocovariances d_{QAF}, i.e. the cluster partitions are based on the estimated values $\hat{\gamma}_l(\tau, \tau')$ given in (3.1). On the other hand, the expected value for R_k under the null is approximated in the original procedure by resampling a multivariate uniform distribution. In our framework, the new variables $\hat{\gamma}_l(\tau, \tau')$ exhibit a strong dependence, and for this reason we propose to obtain replicates using the uniformity assumption marginally for each quantile autocovariance. This way, the support of the distribution for each autocovariance is the range of the estimated values for that autocovariance. The version of the Clest algorithm including these adjustments is outlined below.

For each k, $2 \leq k \leq K$, perform steps 1-4 below.

Step 1. Repeat the following B times:

1. Randomly split the original set of time series S into two non-overlapping sets, a learning set $\mathcal{L}^b$ and a test set $\mathcal{T}^b$.

2. Apply the clustering procedure based on d_{QAF} to the learning set $\mathcal{L}^b$. Let $\mathcal{P}\left(\mathcal{L}^b\right)$ be the obtained k-cluster solution.

3. Classify each series of the test set $\mathcal{T}^b$ using linear discriminant analysis based on the learning partition $\mathcal{P}\left(\mathcal{L}^b\right)$. Denote by $\mathcal{P}_{pred}\left(\mathcal{T}^b\right)$ the partition of the test set obtained from this supervised learning approach.

4. Apply the clustering procedure based on d_{QAF} to the test set $\mathcal{T}^b$. Let $\mathcal{P}_{clust}\left(\mathcal{T}^b\right)$ be the resulting partition.

5. Compute an index of agreement $s_{k,b}$ between $\mathcal{P}_{pred}\left(\mathcal{T}^b\right)$ and $\mathcal{P}_{clust}\left(\mathcal{T}^b\right)$, the partitions generated by supervised (prediction) and unsupervised (clustering) approaches, respectively. A range of external indexes to measure the amount of agreement between two partitions is available in the literature. Here, following the original proposal by Dudoit and Fridlyand (2002), the Fowlkes and Mallows index (Fowlkes and Mallows, 1983) has been considered.

Step 2. Compute the similarity statistic for the k-cluster partition by means of $R_k = \text{median}\,(s_{k,1}, s_{k,2}, \ldots, s_{k,B})$. The null hypothesis $H_0 : k = 1$ will be checked by using R_k as test statistic.

Step 3. Obtain B_0 resamples of the quantile autocovariances matrix under $H_0 : k = 1$. As the columns of this matrix are dependent, resamples of each column are separately generated from an uniform distribution with support determined by the range of the column. For each generated dataset, repeat the procedure described in Steps 1 and 2 to obtain B_0 similarity statistics $R_{k,1}, R_{k,2}, \ldots, R_{k,B_0}$. Based on this set of statistics, compute $\overline{R}_k = \dfrac{1}{B_0}\displaystyle\sum_{b=1}^{B_0} R_{k,b}$ and $p_k = \dfrac{1}{B_0}\#\{R_{k,b} \geq R_k : 1 \leq b \leq B_0\}$.

Step 4. Denote by $d_k = R_k - \overline{R}_k$ the difference between the observed similarity statistic and its estimated expected value under $H_0 : k = 1$. Then, define the set K^- as

$$K^- = \{2 \leq k \leq K : p_k \leq p_{max},\, d_k \geq d_{min}\}, \tag{3.20}$$

where p_{max} and d_{min} are preset thresholds. If K^- is empty, estimate the number of clusters as $\hat{k} = 1$. Otherwise, take $\hat{k} = \text{argmax}_{k \in K^-}\, d_k$, i.e., select the number of clusters $\hat{k}$ corresponding to the largest significant difference d_k.

3.4.1 Comparing procedures for estimating the number of clusters on simulated data

The performance of this adjusted version of Clest was compared with five existing methods using different simulated scenarios. Besides Scenarios 3.1, 3.2 and 3.3 considered in Section 3.3, three new scenarios without underlying clustering structure were generated to evaluate the procedures under the null hypothesis $H_0 : k = 1$. The examined methods, the selected scenarios and the main features of the simulation study are described below.

As before, $S = \{\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(n)}\}$ denotes a set of n time series of length T and $\mathcal{E}_k = \{E_1, \ldots, E_k\}$ a given cluster partition of S. One of the methods considered to estimate the number of clusters in S consists in maximizing the average Silhouette width, ASW, proposed by Kaufman and Rousseeuw (1990) (see Section 1.4).

Three commonly used indexes proposed by Krzanowski and Lai (1988), Caliński and Harabasz (1974) and Hartigan (1975) are also considered. Roughly speaking, these so-called internal indexes are functions of between- and within-clusters sums of squares. In all cases, the objective is to select the value of k providing an optimal value for these functions or internal indexes. Specifically, given the partition $\mathcal{E}_k$, denote by B_k and W_k the $T \times T$ matrices of between and within k-clusters sums of squares and cross-products, respectively. The Krzanowski and Lai index (KL) was defined in Section 1.4. The remaining mentioned indexes perform as follows.

The Caliński and Harabasz index (CH) is defined as

$$\mathrm{CH}(k) = \frac{\mathrm{tr}(B_k)/(k-1)}{\mathrm{tr}(W_k)/(p-k)},$$

where tr denotes the trace of a matrix. The value of k maximizing $\mathrm{CH}(k)$, $k \geq 2$, is selected.

The Hartigan index (Hart) is given by

$$\mathrm{Hart}(k) = (p-k-1)\left(\frac{\mathrm{tr}(W_k)}{\mathrm{tr}(W_{k-1})} - 1\right),$$

and the estimated number of clusters corresponds to the smallest $k \geq 1$ satisfying $\mathrm{Hart}(k) \leq 10$.

The last considered procedure is the Gap method proposed by Tibshirani et al. (2001). Gap method is based on comparing the within-clusters sum of squares W_k with its expected value under a reference null distribution (usually the uniform distribution with support the range of observed values). Specifically, B reference datasets generated under the null hypothesis are subjected to clustering, and values $\mathrm{tr}(W_k^1), \ldots, \mathrm{tr}(W_k^B)$ are computed from

each of obtained partitions. Then, the following values are calculated: (i) the estimated Gap statistic, given by

$$\text{Gap}(k) = \frac{1}{B}\sum_{b=1}^{B} \log\left(\text{tr}(W_k^b)\right) - \log\left(\text{tr}(W_k)\right),$$

(ii) the standard deviation $\text{sd}(k)$ of $\log\left(\text{tr}(W_k^b)\right)$, $1 \leq b \leq B$, and (iii) the value of $s(k) = \text{sd}(k)\sqrt{1+1/B}$. The estimated number of clusters is the smallest $k \geq 1$ such that $\text{Gap}(k) \geq \text{Gap}(k+1) - s(k+1)$.

All of these methods were compared with the adjusted version of Clest through six simulated scenarios, including Scenarios 3.1, 3.2 and 3.3 described in Section 3.3, with $k = 4$ or 5 underlying clusters, consisting of fifty series of length $T = 500$, for the linear and non–linear scenarios, and $T = 1500$, for the heteroskedastic scenario, and three new scenarios with unclustered data ($k = 1$). Each of the new scenarios consisted of fifty series of length $T = 500$ generated from the same process. The selected processes are:

Scenario 3.4 AR process: $X_t = 0.9X_{t-1} + \varepsilon_t$.

Scenario 3.5 EXPAR process: $X_t = \left[0.3 - 10\exp\left(-X_{t-1}^2\right)\right] X_{t-1} + \varepsilon_t$.

Scenario 3.6 ARCH process: $X_t = \mu_t + a_t$, with $\mu_t \sim$ MA(1) and $a_t = \sigma_t\varepsilon_t$, with $\sigma_t^2 = 0.1 + 0.8a_{t-1}^2$.

The error ε_t consisted in all cases of independent zero–mean Gaussian variables with unit variance. The specific parameters required by the Clest algorithm were established as follows: as many learning-test iterations as reference datasets, namely $B = B_0 = 25$, the maximum number of clusters was $K = 7$, the size of each learning set was $2n/3$, and the thresholds required to construct K^- in (3.20) were $p_{max} = 0.05$ and $d_{min} = 0.05$.

Results from our Monte Carlo study are based on $N = 100$ trials for each of the six simulated scenarios. Note that the procedures ASW, CH and KL do not have, by definition, the ability to estimate the presence of only one cluster, and our experiments showed that these methods generally identified two clusters in Scenarios 3.4, 3.5 and 3.6. For this reason, the simulation results are separately presented for scenarios with $k > 1$ and $k = 1$, omitting the methods ASW, CH and KL when $k = 1$. Figures 3.4 and 3.5 display barplots representing the percentage of trials for which a given method estimated correctly the number of underlying clusters in scenarios with $k > 1$ and $k = 1$, respectively. For a more detailed analysis, the distribution of the number of clusters estimated with each method for each scenario are provided in Tables 3.3 and 3.4.

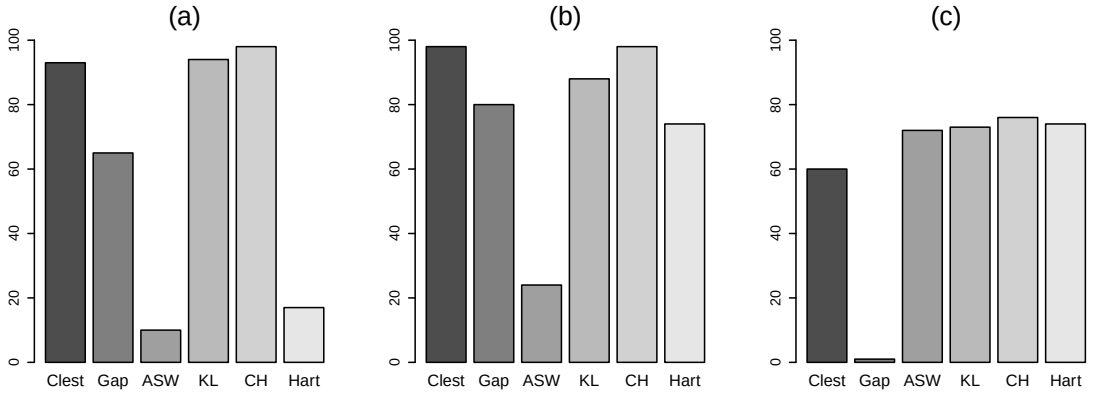

Figure 3.4: Percentage of simulations for which the number of clusters was correctly estimated with each of the considered methods in Scenario 3.1, -k=5- (a), Scenario 3.2 -k=4- (b), and Scenario 3.3 -k=4 (c).

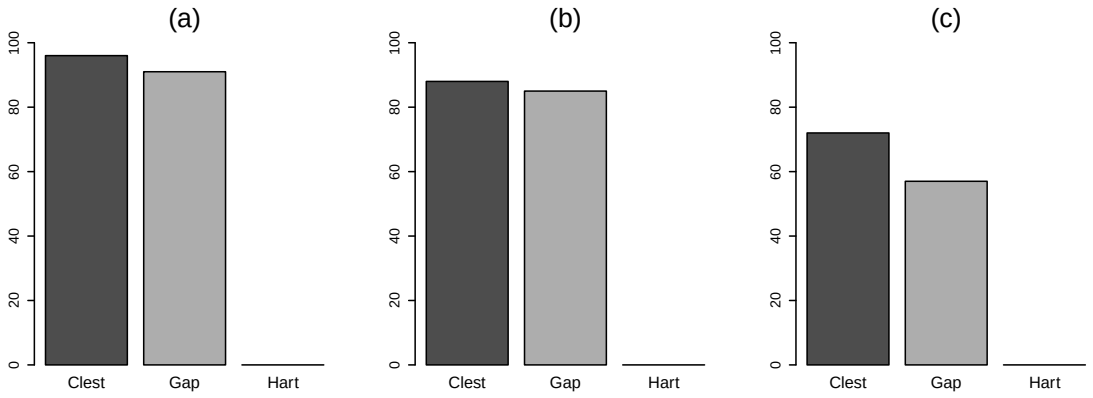

Figure 3.5: Percentage of simulations for which the number of clusters was correctly estimated with each of the considered methods in scenarios without underlying clustering structure ($k = 1$), namely, Scenarios 3.4 (a), 3.5 (b), and 3.6 (c).

From Figure 3.4 and Table 3.3, it is observed that Clest, CH and KL gave uniformly very good results over the scenarios with $k > 1$. The three methods identified the correct number of clusters above 90% of the times with both linear and non-linear time series, and between 60 and 75% in the Scenario 3.3 with heteroskedastic series. Hart index performed particularly well in Scenario 3.3, but presented a worse behaviour in the non-linear scenario and fairly failed with linear series. Gap method places in an intermediate location, achieving reasonably good results in Scenarios 3.1 and 3.2, but performing poorly in Scenario 3.3. Lastly, the criterion based on the Silhouette width was uniformly the worst method, only

Table 3.3: Distribution of the estimated number of clusters for the considered methods in Scenarios 3.1 (k=5), 3.2 (k=4) and 3.3 (k=4). The true number of clusters is denoted by asterisk and the modes for the 100 estimates are indicated in bold for each method.

Method	Number of clusters, $\hat{k}$						
Scenario 3.1	1	2	3	4	5*	6	7
Clest	0	0	0	2	**93**	5	0
Gap	0	0	0	0	**65**	15	20
ASW	–	12	**78**	10	0	0	0
CH	–	0	0	0	**98**	2	0
KL	–	0	0	1	**94**	5	0
Hart	0	0	12	**71**	17	0	0
Scenario 3.2	1	2	3	4*	5	6	7
Clest	0	0	0	**98**	2	0	0
Gap	0	0	0	**80**	11	7	2
ASW	–	0	**76**	24	0	0	0
CH	–	0	0	**98**	2	0	0
KL	–	0	0	**88**	2	4	6
Hart	0	0	26	**74**	0	0	0
Scenario 3.3	1	2	3	4*	5	6	7
Clest	0	0	0	**60**	33	5	2
Gap	**74**	24	1	1	0	0	0
ASW	–	0	17	**72**	10	1	0
CH	–	0	10	**72**	14	0	0
KL	–	0	7	**73**	9	5	6
Hart	0	0	20	**74**	6	0	0

Table 3.4: Distribution of the estimated number of clusters for the considered methods in scenarios without underlying clustering structure ($k = 1$). The modes for the 100 estimates are indicated in bold for each method.

Method	Number of clusters, $\hat{k}$						
Scenario 3.4	1*	2	3	4	5	6	7
Clest	**96**	2	2	0	0	0	0
Gap	**91**	9	0	0	0	0	0
Hart	0	0	**39**	23	16	9	13
Scenario 3.5	1*	2	3	4	5	6	7
Clest	**88**	3	4	2	1	2	
Gap	**85**	12	2	1	0	0	
Hart	0	0	**58**	13	11	6	12
Scenario 3.6	1*	2	3	4	5	6	7
Clest	**72**	3	5	3	9	5	3
Gap	**57**	32	10	1	0	0	0
Hart	0	0	**42**	27	17	9	5

working reasonably well in Scenario 3.3.

With regard to the scenarios under the null hypothesis of no cluster structure in the data ($k = 1$), Figure 3.5 and Table 3.4 show that Clest was always the best method, outperform-

ing Gap method in the three considered scenarios. In Scenario 3.6, Clest was somewhat less efficient, but fairly outperformed Gap. The Hartigan index was not able to detect the lack of clustering structure.

In sum, the Monte Carlo study allows us to conclude that Clest procedure yielded good results in all considered scenarios, being a competitive method when a clustering structure is present and the best one to detect the lack of cluster structure. Only Gap seems to show similar robustness, but with worse success rates and tending to overestimate the number of clusters in some scenarios.

3.5 A case study: Clustering series of daily returns of Euro exchange rates

The dissimilarity based on the quantile autocovariance function d_{QAF} is used to perform clustering on a real data example involving time series of exchange rate. Specifically, our database contains the daily closing values of Euro exchange rates against twenty-eight international currencies. The sample period spans from 1st January 2010 to 28th February 2014, thus resulting serial realizations of length $T = 1520$. All data are sourced from the website of the Bank of Italy[1]. Note that all series are non-stationary in mean, as expected for this type of series and, therefore, the series of nominal exchange rates are transformed to obtain series of daily returns, i.e. series formed by the first differences of the natural logarithm of the nominal exchange rates. These new series are depicted in Figure 3.6.

Here, our concern is not to achieve a correct model specification or accurate predictions for the series of exchange rate returns, but classifying them into homogeneous groups characterized by similar dependence structure. Likewise other financial time series, exchange rate returns exhibit empirical statistical regularities, so-called "stylized facts", which are crucial to perform a proper analysis. The most common stylized facts include: heavy tails and a peaked center compared to the normal distribution, volatility clustering (periods of low volatility mingle with periods of high volatility), leverage effects (returns are negatively correlated with volatility) and autocorrelation at much longer horizons than one would expect. The GARCH models have been widely used (see, e.g. Taylor, 1986) to deal with these peculiar features. For example, D'Urso et al. (2013a) have proposed two fuzzy clustering procedures based on GARCH fittings. In particular, a similar dataset, with shorter observation period, was used to illustrate the merits of their approaches. Our proposal is to

[1] https://www.bancaditalia.it/banca_centrale/cambi/rif;internal&action=_setlanguage.action?LANGUAGE=en

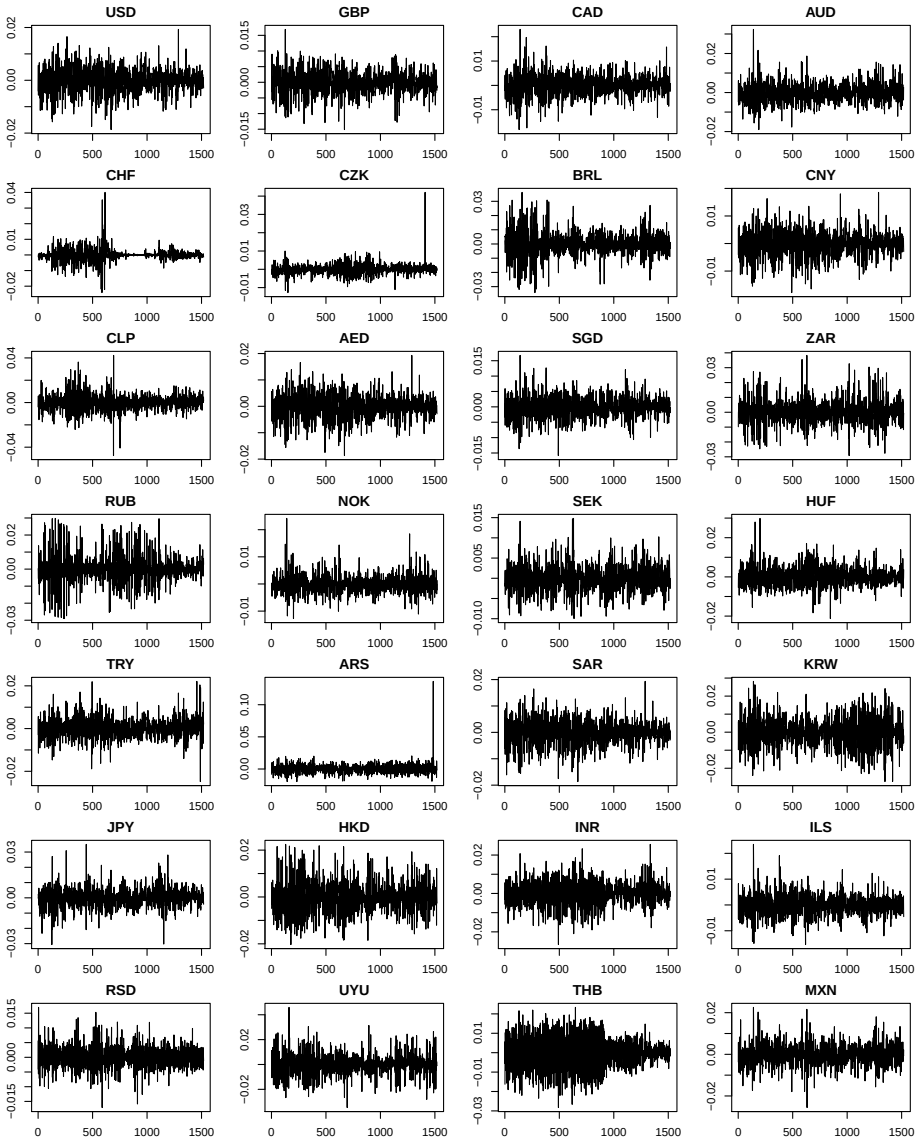

Figure 3.6: Daily returns of Euro exchange against against 28 currencies.

take advantage from the high capability of the quantile autocovariance functions to detect these stylized facts and performing cluster analysis based on d_{QAF}. In fact, d_{QAF} fairly yielded the best results classifying non-linear and heteroskedastic processes in simulations of Section 3.3. This approach allows us to overcome the need of obtaining suitable GARCH fittings, which is not per se the objective, and to attain an efficient implementation.

The 28 series of exchange rates returns were subjected to hierarchical clustering based on the proposed dissimilarity d_{QAF}. Just as in simulations, $r = 3$ quantiles of levels 0.1, 0.5 and 0.9, and one lag ($L = 1$, with $l_1 = 1$) have been considered to compute d_{QAF}. Figure 3.7 shows the obtained dendrogram with the complete linkage method. Dendogram

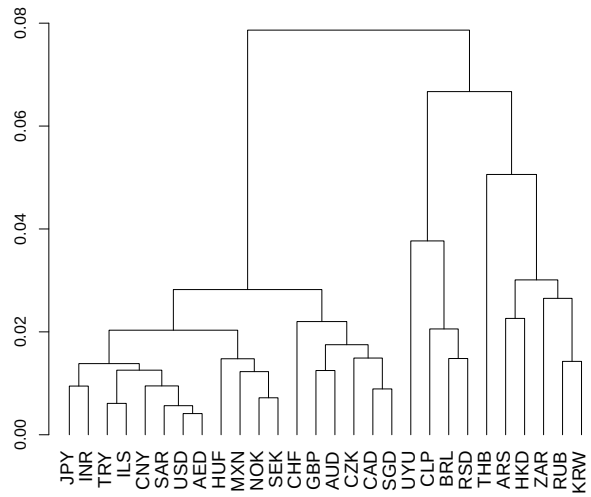

Figure 3.7: Complete linkage dendrogram based on d_{QAF} for series of daily exchange rates returns.

in Figure 3.7 suggests the existence of three major groups, although the exchange rate of the Thailand currency (EUR/THB) might also be seen as an isolated point and hence a four-cluster solution determined. The optimal number of clusters was estimated by means of the adjusted Clest algorithm introduced in Section 3.4. Setting the algorithm inputs as $K = 7$, $B = B_0 = 50$, learning subset of size 18 and $p_{max} = d_{min} = 0.05$, the Clest algorithm led to $\hat{k} = 3$, thus corroborating the intuitive solution derived from the dendogram.

The three-cluster solution involves a particularly large cluster, C1, formed by 18 exchange rates. It is observed that C1 groups the Euro exchange rates against the major international currencies and those linked to the US dollar, such as the Canadian dollar (CAD) and the Great Britain pound (GBP), among others. The two other clusters, C2 and C3, are formed by 4 and 6 memberships, respectively. While C2 is quite homogeneous by including three South American currencies (Brazilian real -BRL-, Uruguayan peso -UYU- and Chilean peso -CLP-), C3 $\equiv$ {South African rand (ZAR), Russian ruble (RUB), Argentine peso (ARS), South Korean won (KRW), Thailand baht (THB) and Hong Kong dollar (HKD)} is the most heterogeneous cluster by involving Euro exchange rates against Asian, European, South American and African currencies.

By studying this kind of series, an issue of great interest is the exchange rate volatility. Volatility provides an idea on the fluctuations of the exchange rates over a given period, and it is usually measured from the conditional variance of these movements. High volatility implies high chance of a large rate change. The obtained cluster solution brings insight into

the patterns of underlying volatility structure. In fact, Figure 3.8 depicts the conditional volatility of the medoids of the three-cluster solution. For each cluster, the medoid has been determined by selecting the membership minimizing the average dissimilarity to all the series in the cluster. Figure 3.8 fairly shows different shapes for the conditional volatility at each cluster. Clusters C1 and C2 have lower levels of fluctuation, although the currencies within C1 show higher stability than the ones in C2. The most heterogeneous cluster C3 includes the series with the highest levels of fluctuation.

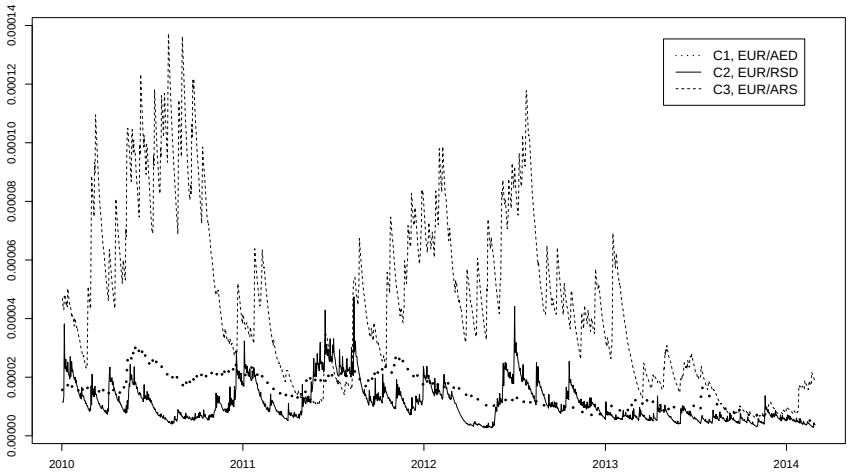

Figure 3.8: Conditional volatility of the medoid of each cluster.

It is also worth remarking that similar cluster solutions are obtained as the average and complete linkages are used, thus validating the stability of the encountered grouping. In both cases, the most distinctive feature is that the Uruguayan peso (UYU) constitutes an isolated point at the late stages of the hierarchical processes. Apart from it, the existence of three groups formed by the same exchange rates (with minimum differences in the two smallest groups) is observed.

3.6 Optimal selection of lags and quantile levels for clustering

In this section, the problem of the selection of the optimal number of parameters for the correct computation of the metric d_{QAF} is addressed. According to definition (3.19), computation of d_{QAF} requires setting a number of input parameters, namely the number L of significant lags and the set of quantile levels $\{\tau_1, \ldots, \tau_r\}$. Since our target is to use this metric to perform time series clustering, our concern is to determine how many and which combinations of lags and quantile levels $(l_i, \tau_j, \tau_{j'})$ must be considered to optimize the clustering process. The light computational complexity of d_{QAF} enables us to employ a

reasonably large number of lags and quantiles without a significant loss of efficiency. Nevertheless, working with a large set of inputs does not necessarily improve the clustering performance. In practice, introducing non–significant lags or very close quantiles means to supply noise to the classification process, thus generating worse results. Therefore, the goal is simple: starting from a preset grid of input parameters, determining a reasonably small subset that conveys the more relevant information on the underlying clustering structure.

To reach this goal, we follow a novel approach proposed by Andrews and McNicholas (2014). In a general context, Andrews and McNicholas introduce a variable selection stepwise algorithm for clustering and classification (called VSCC) based on determining the variables that simultaneously minimize the within–group variance and maximize the between–group variance. Indeed, if the variables have been standardized to have the same variance, then minimization of the within–group variance also implies the maximization of the between–group variance. Besides this criterion in terms of 'within' and 'between' variances, the algorithm imposes that the correlation between the selected variables drops below a threshold. The purpose is to ignore highly correlated variables, which does not provide new information and may introduce noise. The correlation threshold is not a prefixed value but a sliding threshold allowed to be larger as the within–group variance is small. Specifically, if S denotes the subset of selected variables at a particular step of the algorithm, then a new variable s is added to S if for all $r \in S$ we have

$$|\rho_{sr}| < 1 - \mathcal{W}_s^{\alpha} \tag{3.21}$$

where ρ_{sr} is the correlation between the variables s and r, $\mathcal{W}_s$ denotes the within–group variance for the variable s, and α is a preset parameter determining the shape of the relationship between the whitin–group variance and the between–variable correlation. Note that the smaller α, more stringent is the correlation threshold.

Compared to other variable selection techniques in clustering, the VSCC algorithm is intuitive, competitive and more computationally efficient. Based on these arguments, we decided to adapt this algorithm to address the optimal selection of lags and quantiles in order to perform clustering using d_{QAF}.

Consider a set of n realizations of time series subjected to clustering. Starting from a grid of r regularly spaced quantile levels and a number L of lags, an initial set of vectors of length Lr^2, $\mathbf{\Gamma}^{(u)}$ for $u = 1, \ldots, n$, is computed according to (3.17), i.e. the observed series are replaced by vectors of estimated quantile autocovariances. These vectors are arranged by rows in a $n \times Lr^2$ matrix, $\boldsymbol{A}$, whose columns represent the variables used to perform clustering. The dissimilarity between two series $X^{(u)}$ and $X^{(v)}$ is given by the

squared Euclidean distance between the rows u and v of $\boldsymbol{A}$, that is $d_{QAF}\left(X_t^{(u)}, X_t^{(v)}\right) = \left\|\boldsymbol{\Gamma}^{(u)} - \boldsymbol{\Gamma}^{(v)}\right\|^2$, and the clustering procedure relies on this dissimilarity criterion.

Therefore, we start with Lr^2 variables characterized by combinations of a lag and a pair of quantile levels $\left(l_i, \tau_j, \tau_j'\right)$, and our intention is to apply the VSCC algorithm to obtain an optimal selection of these combinations. First, the columns of $\boldsymbol{A}$ are standardized to have zero mean and unit variance, which allows us to concentrate our attention on minimizing the within-group variance. Then, the VSCC procedure is carried out as follows.

Step 1. Set the initial grid of r regularly spaced quantile levels and L lags, the number C of clusters and the value of α governing the relationship (3.21).

Step 2. Perform a partitional clustering procedure of the set of time series based on the matrix $\boldsymbol{A} = \left(\boldsymbol{\Gamma}^{(u)\,t}\right)_{1\leq u\leq n}$ of estimated quantile autocovariances and d_{QAF}.

Step 3. For each column of $\boldsymbol{A}$, i.e. for each combination $\left(l_i, \tau_j, \tau_j'\right)$, with $i = 1, \ldots, L$ and $j, j' = 1, \ldots, r$, compute the within–group variance $\mathcal{W}_{\left(l_i,\tau_j,\tau_j'\right)}$ defined by

$$\mathcal{W}_{\left(l_i,\tau_j,\tau_j'\right)} = \frac{1}{n}\sum_{c=1}^{C}\sum_{s=1}^{n} z_{sc}\left(\hat{\gamma}_{l_i}^{(s)}(\tau_j,\tau_{j'}) - \bar{\hat{\gamma}}_{l_i}^{(c)}(\tau_j,\tau_{j'})\right)^2,$$

where z_{sc} is the group membership indicator function and $\bar{\hat{\gamma}}_{l_i}^{(c)}(\tau_j,\tau_{j'})$ is the average of the corresponding estimated quantile autocovariances over the group c, that is $\bar{\hat{\gamma}}_{l_i}^{(c)}(\tau_j,\tau_{j'}) = \frac{1}{n}\sum_{s=1}^{n} z_{sc}\hat{\gamma}_{l_i}^{(s)}(\tau_j,\tau_{j'})$. By dealing with a hard partition, $z_{sc} = I\left(\boldsymbol{\Gamma}^{(s)} \in c\right)$ is the indicator function taking the value 1 if the series s belongs to the cluster c and 0 otherwise. In the case of a fuzzy partition, z_{sc} takes the value 1 for the cluster c where the series s presents the highest membership degree and 0 otherwise.

Step 4. Sort $\mathcal{W}_{\left(l_i,\tau_j,\tau_j'\right)}$ in ascending order. Denote this sorted list by $\mathbf{W} = \left\{\mathcal{W}_{(1)}, \ldots, \mathcal{W}_{(Lr^2)}\right\}$ and the combination of lag and quantile levels associated with $\mathcal{W}_{(k)}$ by $\left(l_{i_k}, \tau_{j_k}, \tau_{j_k}'\right)$.

Step 5. $\mathcal{W}_{(1)}$ minimizes $\mathbf{W}$ and hence $\left(l_{i_1}, \tau_{j_1}, \tau_{j_1}'\right)$ is automatically placed into the subset of selected variables, denoted by S. Set count $k = 2$.

Step 6. If $|\rho_{sk}| < 1 - \mathcal{W}_{(2)}^{\alpha}$, for all $s \in S$, then the combination associated with $\mathcal{W}_{(2)}$, $\left(l_{i_2}, \tau_{j_2}, \tau_{j_2}'\right)$ is added to S.

Step 7. While $k < Lr^2$, set $k = k + 1$ an return to Step 6. Then end algorithm.

In the rest of this chapter, the VSCC procedure is taken into consideration. In our experiments, we have considered up to five different values for α, namely $\alpha = 1, \ldots, 5$, exactly as proposed in Andrews and McNicholas (2014). Overall, the most stringent choice $\alpha = 1$ led to a small number of variables and good clustering behavior. Furthermore, in Section 3.7.1, the results from a sensitivity analysis conducted to test the efficacy of the proposed procedure against the use of different sequences of quantile levels are presented.

3.7 Partitioning around medoids clustering based on quantile autocovariances

In this section, we extend the analysis to cover partitioning-based clustering methods. Assuming the existence of C clusters and starting from an initial partition, these methods proceed by iteratively relocating objects between clusters until an optimal partition is attained. At each iteration, C cluster centers (usually referred to as prototypes or centroids) are estimated and a reassignment of objects based on the updated centers is carried out. The most popular partitioning-based algorithm is the C-means procedure, where the centroids are the means of objects in the clusters and the objective is to minimize the within-cluster squared error. Nevertheless, C-means is not a proper choice in our framework because the average of quantile autocovariances does not necessarily characterizes a time series model. For instance, if ARMA or GARCH models are considered, then there are no guarantees that the centroids represent one of these models. In fact, the resulting centroid could not satisfy the constraints required on the coefficients defining these models. This way the centroids may be "fictitious" time series, which leads to serious drawbacks. First, the distance d_{QAF} between observed time series objects and centroids could not be properly defined. On the other hand, time series clustering is often aimed at finding "representative" time series for each cluster, let us say a set of C patterns summarizing the different underlying dynamics, and again this is not guaranteed and the resulting centroids could fail in providing a suitable characterization of the cluster dynamics. A natural way to overcome these drawbacks is to perform a C-medoids-based algorithm where the prototypes are restricted to be chosen among the data points. The goal is to find C representative objects minimizing the average dissimilarity of all objects to their closest representative object. This way, d_{QAF} (or whatever is the selected distance) directly determines the efficacy of the clustering. In fact, the C-medoid algorithms can be run using the pairwise distances without requiring the data records. Unlike the C-means procedure, where optimization involves minimizing within-group variance and maximizing between-group variance, and therefore a L_2 analysis, the C-medoids-based algorithms are L_1 methods and therefore more robust to outliers and

noise. Regarding these nice properties, we have carried out an extensive simulation study using the standard version of the well-known PAM algorithm (from "Partitioning Around Medoids", Kaufman and Rousseeuw, 1990) which is currently available in R code.

3.7.1 Simulation study

A second set of simulations was conducted to assess the performance of d_{QAF} using a PAM algorithm considering the same three classification setups as in Section 3.3, namely Scenarios 3.1, 3.2, 3.3.

In this case, the error process ε_t consisted of iid variables following different distributions, namely Gaussian innovations with unit variance, Student-t innovations with 1 degree of freedom, and exponential Exp(0.75) innovations. Using these distributions, we intend to assess the behavior of the clustering procedure also when kurtosis or skewness are present. The processes of every scenario were generated as in Section 3.3 but considering starting points from an Student-t with 1 degree of freedom and an Exp(0.75) for the scenarios considering heavy-tailed and non-symmetric innovations, respectively.

The behavior of the partitioning procedure based on d_{QAF} was compared with its counterpart based on the metrics considered in 3.3. The quantile levels for the computation of d_{QAF} were determined by means of the variable selection algorithm VSCC introduced in Section 3.6. As a starting point, the VSCC algorithm was implemented over a grid of regularly spaced quantile levels formed by all the combinations $(0.05j, 0.05j')$, with j and j' ranging from 1 to 19.

For each scenario, five time series of equal length T were generated from each model, thus providing a sample set of labeled series available to perform clustering. The experiments were carried out for three different series lengths, $T = 250$, 500 and 1000. Note that all models are stationary in mean but they present differences in scale. To avoid that these differences dominate the clustering, the series were previously normalized to have unit variance.

Each set of simulated series was subjected to partitional clustering using the PAM algorithm together with each of the studied metrics. The algorithm inputs were the true number C of clusters, the pairwise dissimilarity matrix, and the initial C medoids, which were randomly determined among all the series.

Again, the quality of the clustering procedure is evaluated comparing the experimental cluster solution with the true cluster partition using three different agreement measures based on known "ground-truth", namely the Gavrilov index (Gavrilov et al., 2000), the ad-

justed Rand index (Hubert and Arabie, 1985), and one-nearest-neighbour (1-NN) classifier evaluated by leave-one-out cross-validation (Keogh and Kasetty, 2003). All of them have been previously presented in Section 3.3.

The simulation procedure was replicated $N = 100$ times for each scenario and the obtained indexes were averaged over the 100 trials. The averages and standard deviations (in brackets) obtained for the different lengths of the series are reported in Tables 3.5, 3.6 and 3.7, including results for the considered innovation distributions.

In the case of Gaussian innovations, the dissimilarity based on quantile autocovariances d_{QAF} led to the highest average scores in clustering of non-linear and heteroskedastic models, Scenarios 3.2 and 3.3 respectively. In fact, the results in these scenarios based on d_{QAF} were substantially better than the ones obtained with the rest of metrics for the three indexes. In the linear framework (Scenario 3.1), d_{QAF} produced reasonably high indexes although a little worse than d_{PACFG}, d_M and d_{ISD}. It is worthy to point out the outstanding behavior of d_{QAF} in clustering of non-linear models, with average agreement indexes always above 0.985 for the smallest length, and exactly 1 with series of length 1000 for all the experiments (since $sd = 0$). Beyond the efficacy of PAM algorithm, the scores of the 1-NN classifier (Ind_3) illustrate the high capability of d_{QAF} to discriminate between these processes, fairly superior to the other metrics. The worst clustering results were obtained in the Scenario 3.3, thus showing the complexity of clustering heteroskedastic structures. Only d_{QAF} is able to draw out good classification rates in this complex clustering framework, specially with large series. Note even that, except for d_{QAF}, the clustering results do not improve as the length of the series increases.

Similar conclusions derive from the results obtained with non-symmetric and heavy-tailed disturbances, although d_{QAF} reported additional nice properties. First, d_{QAF} was again the best-performed metric in Scenarios 3.1 and 3.3, increasing the average quality indexes with respect to the Gaussian setting. Specially noteworthy was the improvement with heavy-tailed innovations where significantly high scores are now attained. The rest of metrics presented different behaviors in these two scenarios. While they exhibited an improvement with heavy–tailed innovations (although fairly below d_{QAF}), their results substantially worsened with non-symmetric innovations, thus concluding that the asymmetry has an important influence over these metrics. Finally, except for d_{QAF}, all the metrics are affected by asymmetry and kurtosis when classifying ARMA models, particularly the non–parametric dissimilarity d_{ISD}. On the contrary, d_{QAF} presents better results, being very close to the best metrics also in the linear scenario.

In short, our numerical experiments illustrate the good performance of the proposed metric

Table 3.5: Indexes of clustering quality in the Monte-Carlo simulation with series of length $T = 250$.

Dissimilarity	Scenario 3.1			Scenario 3.2			Scenario 3.3		
	Ind_1	Ind_2	Ind_3	Ind_1	Ind_2	Ind_3	Ind_1	Ind_2	Ind_3
Gaussian innovations									
d_{LP}	0.729	0.596	0.736	0.483	0.109	0.381	0.414	-0.002	0.204
	(.098)	(.102)	(.092)	(.069)	(.106)	(.110)	(.046)	(.062)	(.097)
d_{PACFG}	0.873	0.775	0.883	0.714	0.444	0.683	0.437	0.017	0.233
	(.070)	(.098)	(.077)	(.088)	(.138)	(.114)	(.061)	(.073)	(.105)
d_M	0.875	0.811	0.937	0.749	0.554	0.789	0.428	0.018	0.258
	(.102)	(.139)	(.068)	(.095)	(.145)	(.132)	(.051)	(.066)	(.095)
d_{ISD}	0.906	0.821	0.908	0.726	0.517	0.722	0.421	0.027	0.256
	(.063)	(.095)	(.070)	(.094)	(.137)	(.117)	(.052)	(.067)	(.100)
d_{QAF}	0.767	0.600	0.709	0.995	0.986	0.997	0.622	0.268	0.496
	(.090)	(.100)	(.117)	(.014)	(.039)	(.014)	(.070)	(.078)	(.109)
Non–symmetric innovations									
d_{LP}	0.741	0.593	0.742	0.522	0.263	0.438	0.509	0.181	0.413
	(.083)	(.079)	(.107)	(.059)	(.103)	(.120)	(.101)	(.144)	(.114)
d_{PACFG}	0.876	0.779	0.884	0.627	0.401	0.560	0.625	0.398	0.619
	(.073)	(.103)	(.070)	(.089)	(.091)	(.137)	(.103)	(.166)	(.120)
d_M	0.888	0.817	0.945	0.631	0.433	0.623	0.569	0.286	0.599
	(.092)	(.126)	(.054)	(.093)	(.078)	(.156)	(.089)	(.145)	(.145)
d_{ISD}	0.896	0.802	0.911	0.586	0.412	0.614	0.598	0.359	0.691
	(.062)	(.089)	(.058)	(.086)	(.082)	(.131)	(.079)	(.133)	(.117)
d_{QAF}	0.830	0.728	0.838	1.000	1.000	1.000	0.663	0.409	0.575
	(.060)	(.071)	(.087)	(.000)	(.000)	(.000)	(.084)	(.103)	(.150)
Heavy–tailed innovations									
d_{LP}	0.614	0.514	0.714	0.416	0.070	0.535	0.483	0.138	0.391
	(.043)	(.050)	(.104)	(.076)	(.122)	(.107)	(.051)	(.076)	(.103)
d_{PACFG}	0.872	0.815	0.953	0.723	0.556	0.834	0.424	0.050	0.351
	(.097)	(.124)	(.051)	(.046)	(.077)	(.084)	(.059)	(.069)	(.087)
d_M	0.838	0.772	0.965	0.738	0.576	0.866	0.438	0.062	0.367
	(.087)	(.111)	(.037)	(.073)	(.110)	(.093)	(.064)	(.083)	(.100)
d_{ISD}	0.743	0.641	0.905	0.652	0.428	0.757	0.481	0.144	0.458
	(.100)	(.111)	(.060)	(.086)	(.146)	(.105)	(.055)	(.067)	(.101)
d_{QAF}	0.830	0.735	0.846	0.998	0.996	0.999	0.676	0.392	0.600
	(.067)	(.084)	(.087)	(.009)	(.024)	(.007)	(.077)	(.110)	(.123)

in partitional clustering for a wide range of time series models. Specifically, d_{QAF} outperformed the rest of analyzed metrics in clustering of non-linear and heteroskedastc models, but was also highly competitive in clustering of linear models. Furthermore, unlike the rest of metrics, quality of the clustering results based on d_{QAF} showed robustness to the kind of disturbance distribution.

Table 3.6: Indexes of clustering quality in the Monte-Carlo simulation with series of length $T = 500$.

Dissimilarity	Scenario 3.1			Scenario 3.2			Scenario 3.3		
	Ind_1	Ind_2	Ind_3	Ind_1	Ind_2	Ind_3	Ind_1	Ind_2	Ind_3
Gaussian innovations									
d_{LP}	0.804	0.674	0.812	0.530	0.188	0.442	0.406	-0.002	0.206
	(.093)	(.115)	(.100)	(.070)	(.107)	(.122)	(.046)	(.048)	(.088)
d_{PACFG}	0.949	0.899	0.951	0.858	0.694	0.862	0.428	0.014	0.249
	(.066)	(.107)	(.052)	(.091)	(.152)	(.078)	(.051)	(.069)	(.103)
d_M	0.952	0.928	0.977	0.905	0.812	0.922	0.416	0.020	0.253
	(.085)	(.118)	(.041)	(.103)	(.171)	(.074)	(.046)	(.067)	(.118)
d_{ISD}	0.952	0.896	0.958	0.869	0.748	0.895	0.405	0.016	0.246
	(.047)	(.084)	(.044)	(.108)	(.170)	(.078)	(.039)	(.057)	(.125)
d_{QAF}	0.855	0.730	0.832	0.999	0.999	1.000	0.714	0.459	0.643
	(.089)	(.131)	(.099)	(.005)	(.014)	(.000)	(.075)	(.135)	(.117)
Non–symmetric innovations									
d_{LP}	0.775	0.637	0.789	0.556	0.342	0.459	0.554	0.263	0.495
	(.083)	(.092)	(.096)	(.079)	(.101)	(.116)	(.098)	(.135)	(.111)
d_{PACFG}	0.953	0.903	0.955	0.678	0.458	0.660	0.657	0.463	0.742
	(.056)	(.092)	(.051)	(.107)	(.120)	(.130)	(.075)	(.141)	(.111)
d_M	0.956	0.934	0.985	0.714	0.544	0.744	0.521	0.226	0.681
	(.081)	(.113)	(.027)	(.124)	(.135)	(.140)	(.072)	(.119)	(.120)
d_{ISD}	0.945	0.888	0.960	0.664	0.492	0.770	0.647	0.436	0.797
	(.054)	(.089)	(.042)	(.099)	(.104)	(.137)	(.070)	(.119)	(.090)
d_{QAF}	0.877	0.791	0.916	1.000	1.000	1.000	0.777	0.592	0.737
	(.084)	(.117)	(.062)	(.000)	(.000)	(.000)	(.109)	(.146)	(.154)
Heavy–tailed Errors									
d_{LP}	0.620	0.523	0.727	0.435	0.100	0.553	0.503	0.175	0.419
	(.042)	(.050)	(.101)	(.089)	(.140)	(.133)	(.045)	(.047)	(.103)
d_{PACFG}	0.934	0.900	0.984	0.757	0.611	0.890	0.415	0.040	0.379
	(.091)	(.126)	(.030)	(.080)	(.117)	(.070)	(.052)	(.051)	(.077)
d_M	0.901	0.860	0.987	0.756	0.611	0.607	0.905	0.046	0.388
	(.105)	(.141)	(.023)	(.079)	(.124)	(.062)	(.051)	(.055)	(.082)
d_{ISD}	0.784	0.704	0.942	0.658	0.445	0.810	0.503	0.180	0.507
	(.128)	(.154)	(.051)	(.101)	(.161)	(.079)	(.045)	(.041)	(.099)
d_{QAF}	0.877	0.796	0.897	1.000	1.000	1.000	0.726	0.523	0.731
	(.083)	(.111)	(.078)	(.000)	(.000)	(.000)	(.088)	(.125)	(.125)

3.7.2 The role of the lag number in the computation of d_{QAF}

A sensitivity analysis experimenting with different sequences of regularly spaced quantile levels was conducted, including a comparison with the results based on the variable selection VSCC algorithm. This way, we intend to analyze the effect of the selection of quantile levels on the clustering results. Table 3.8 reports the averages of the cluster similarity indexes

Table 3.7: Indexes of clustering quality in the Monte-Carlo simulation with series of length $T = 1000$.

Dissimilarity	Scenario 3.1			Scenario 3.2			Scenario 3.3		
	Ind_1	Ind_2	Ind_3	Ind_1	Ind_2	Ind_3	Ind_1	Ind_2	Ind_3
Gaussian innovations									
d_{LP}	0.843	0.727	0.864	0.575	0.267	0.512	0.418	0.013	0.225
	(.084)	(.104)	(.084)	(.088)	(.133)	(.120)	(.052)	(.057)	(.073)
d_{PACFG}	0.995	0.988	0.994	0.976	0.936	0.978	0.420	0.015	0.252
	(.018)	(.040)	(.016)	(.038)	(.091)	(.041)	(.050)	(.065)	(.104)
d_M	0.998	0.995	0.998	0.989	0.973	0.987	0.426	0.044	0.273
	(.009)	(.023)	(.011)	(.031)	(.063)	(.032)	(.051)	(.060)	(.107)
d_{ISD}	0.986	0.965	0.991	0.979	0.945	0.977	0.416	0.032	0.300
	(.027)	(.060)	(.019)	(.033)	(.082)	(.037)	(.046)	(.055)	(.106)
d_{QAF}	0.926	0.845	0.920	1.000	1.000	1.000	0.765	0.605	0.716
	(.070)	(.118)	(.069)	(.000)	(.000)	(.000)	(.058)	(.083)	(.105)
Non–symmetric innovations									
d_{LP}	0.855	0.746	0.877	0.388	0.053	0.305	0.584	0.315	0.543
	(.084)	(.114)	(.093)	(.057)	(.117)	(.082)	(.094)	(.144)	(.127)
d_{PACFG}	0.991	0.978	0.992	0.726	0.552	0.812	0.657	0.476	0.832
	(.022)	(.050)	(.023)	(.114)	(.142)	(.094)	(.063)	(.125)	(.106)
d_M	0.997	0.995	0.996	0.781	0.674	0.920	0.518	0.217	0.750
	(.023)	(.033)	(.016)	(.149)	(.175)	(.073)	(.063)	(.107)	(.091)
d_{ISD}	0.979	0.949	0.990	0.692	0.564	0.856	0.658	0.448	0.875
	(.032)	(.070)	(.022)	(.120)	(.118)	(.079)	(.074)	(.124)	(.080)
d_{QAF}	0.931	0.876	0.953	1.000	1.000	1.000	0.890	0.777	0.861
	(.082)	(.126)	(.050)	(.000)	(.000)	(.000)	(.084)	(.128)	(.111)
Heavy–tailed innovations									
d_{LP}	0.615	0.526	0.740	0.426	0.086	0.554	0.520	0.199	0.459
	(.044)	(.043)	(.101)	(.084)	(.134)	(.153)	(.043)	(.039)	(.113)
d_{PACFG}	0.972	0.959	0.996	0.805	0.670	0.928	0.411	0.035	0.389
	(.069)	(.096)	(.012)	(.104)	(.153)	(.059)	(.047)	(.045)	(.090)
d_M	0.941	0.919	0.998	0.759	0.612	0.926	0.412	0.033	0.402
	(.093)	(.125)	(.010)	(.082)	(.123)	(.057)	(.053)	(.048)	(.088)
d_{ISD}	0.789	0.708	0.948	0.665	0.453	0.837	0.519	0.201	0.540
	(.112)	(.131)	(.046)	(.098)	(.159)	(.092)	(.043)	(.039)	(.087)
d_{QAF}	0.957	0.920	0.963	1.000	1.000	1.000	0.702	0.515	0.766
	(.068)	(.107)	(.056)	(.000)	(.000)	(.000)	(.083)	(.110)	(.126)

obtained from 100 trials of the simulation procedure for Scenarios 3.1, 3.2 and 3.3, with Gaussian innovations, series of length $T = 250$, and the metric d_{QAF} based on the following combinations of quantile levels: (i) $\boldsymbol{\tau}_1 = (0.1, 0.5, 0.9)$, (ii) $\boldsymbol{\tau}_2 = (0.1, 0.3, 0.5, 0.7, 0.9)$, and (iii) $\boldsymbol{\tau}_3 = (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)$.

Table 3.8 reveals that results get better as the number of quantiles is increased. Never-

Table 3.8: Influence of the selection of quantile levels

Vector of	Scenario 3.1			Scenario 3.2			Scenario 3.3		
quantile levels	Ind_1	Ind_2	Ind_3	Ind_1	Ind_2	Ind_3	Ind_1	Ind_2	Ind_3
$\boldsymbol{\tau}_1$	0.766	0.583	0.688	0.986	0.961	0.989	0.595	0.231	0.535
$\boldsymbol{\tau}_2$	0.753	0.586	0.696	0.993	0.982	0.996	0.607	0.256	0.537
$\boldsymbol{\tau}_3$	0.767	0.600	0.716	0.993	0.982	0.997	0.610	0.251	0.560
VSCC	0.767	0.600	0.709	0.995	0.986	0.997	0.622	0.268	0.496

theless, no large differences are observed, and acceptable results are reached using only three quantile levels ($\boldsymbol{\tau}_1$). Except for Ind_3 in Scenario 3.3, the VSCC algorithm leads to the highest scores, with the additional advantage of determining a proper trade-off between number of quantile levels and clustering quality on the basis of an objective criterion. Anyway, results from Table 3.8 suggest that d_{QAF} should produce satisfactory results in clustering with a small number of regularly spaced quantile levels. Although this is indeed a noteworthy property, it is also worth remarking that d_{QAF} is computationally efficient because of an increase in the number quantiles does not mean a substantial cost in terms of computing time.

3.8 Concluding remarks

In this chapter, our motivation has been to introduce an efficient dissimilarity measure with a high capability to cluster series generated from a broad range of dependence models. With this objective in mind, a metric based on quantile autocovariance functions (d_{QAF}) has been proposed. Quantile autocovariances provide valuable insight into the serial dependence and present a much richer view than other extracted features about the underlying dependence structure. Robustness to nonexistence of moments and capability to deal with heavy–tailed marginal distributions, to analyze dependence of extreme values, and to detect nonlinear features and changes in conditional shapes are appealing properties of the quantile autocovariances, which suggest their usefulness to classify a wide range of time series models. This intuition has been illustrated by means of a motivating example addressed to discriminate between realizations of Gaussian white noise, GARCH and exponential GARCH processes.

To perform hard clustering, we focus in two different approaches. First an agglomerative hierarchical clustering algorithm with complete linkage was considered. An extensive simulation study showed that the proposed dissimilarity produces satisfactory results by performing cluster analysis on different types of processes. In complex scenarios in-

cluding conditional heteroskedastic processes, d_{QAF} led to the best results compared to a range of representative dissimilarities introduced in the literature. In fact, apart from d_{QAF}, none of the remaining examined dissimilarities showed acceptable results by clustering heteroskedastic processes, thus emphasizing the usefulness of d_{QAF} in this framework. Dissimilarity d_{QAF} also produced very good results by clustering non-linear processes, attaining so high scores as the ones obtained by the non-parametric dissimilarities, which are particularly suitable to tackle this kind of processes. Only classifying linear models d_{QAF} showed worse behaviour than some dissimilarities specifically designed to deal with linear processes. Nevertheless, clearly competitive scores can be also attained in this scenario if the tuning parameters required to construct d_{QAF} are properly adjusted. In short, the dissimilarity based on quantile autocovariance functions seems to show great flexibility to properly work with different types of underlying processes. Furthermore, unlike other dissimilarities, the proposed metric satisfies additional properties specially useful in time series clustering. Specifically, d_{QAF} presents an efficient implementation at a very low cost in terms of computing time and can be applied to series of unequal length.

Also a partitioning around C-medoids technique was considered to perform clusering analysis, and the results of the simulations have shown once again the good performance of the C-crisp model based on the squared Euclidean distance between sample quantile autocovariances, d_{QAF}. Compared to other distances measuring discrepancy between generating models and other extracted features, our approach led to the best classification rates by grouping non-linear and heteroskedastic models in well–separated clusters. Likewise the hierarchical approach, our proposal produced competitive success rates, by clustering linear models in general close to the ones obtained with metrics specifically designed to deal with ARMA models. The C-crisp model combined with d_{QAF} also exhibited a remarkable property of robustness against the kind of innovation distribution, unlike the rest of examined metrics which have been noticeably affected by skewness and kurtosis.

In order to provide an automatic tool for clustering, an iterative algorithm to select the lags and quantile levels optimizing the clustering process has been introduced. Overall, a small number of quantiles is selected, and a further sensitivity analysis has illustrated that a few quantiles are enough to obtain satisfactory results.

The problem of estimating the optimal number of clusters has been also addressed. A range of existing procedures have been compared in a new simulation study considering different generating processes. The prediction-based resampling algorithm Clest proposed by Dudoit and Fridlyand (2002), properly adjusted to use the dissimilarity d_{QAF}, produced good results in all considered scenarios. When a clustering structure was present, the adjusted version of Clest led to accurate estimations of the true number of clusters, ranking

among the best-performed methods, and it was clearly the best procedure to detect the lack of clusters.

For illustrative purposes, the proposed methodology has been applied to classify the time series of Euro exchange rates against 28 international currencies. The Clest algorithm identified three clusters, and the 3-cluster hierarchical solution based on d_{QAF} allowed us to characterize each cluster in terms of the different volatility structures of their elements. Therefore, we have taken here advantage of the capability of the quantile autocovariance function to discriminate between non-linear and heteroskedastic models, which cannot be accounted for by other structure-based dissimilarities.

Part of the material developed in the present chapter, including the introduction of the QAF metric for hierarchical clustering of time series, the solution proposed to address the optimal selection of the number of clusters and the case study have been published in Lafuente-Rego and Vilar (2016a).

Chapter 4

Fuzzy clustering of time series based on quantile autocovariances. Robust approaches.

Contents

4.1 Introduction

In cluster analysis, attending to the cluster assignment, two different paradigms are usually considered depending on whether a "hard" or "soft" partition is constructed. Traditional clustering methods assign each data object to exactly one cluster, thus producing a hard partition of the data into non-empty and disjoint subsets. This approach can result too

rigid in situations with data objects equidistant from two or more groups or in presence of overlapping clusters. Fuzzy cluster techniques (Döring et al., 2006; D'Urso, 2015) provide a more versatile approach by allowing gradual membership of data objects to clusters. In the resulting soft partition, the objects can belong to several clusters with specific membership levels indicating the amount of confidence in the assignment of each data to the clusters. Adoption of the fuzzy logic in time series clustering is interestingly motivated by some authors. D'Urso and Maharaj (2009) and D'Urso et al. (2013b, 2015a) argue that the dynamic of a time series may change over time in such a way that it could belong to distinct clusters during different periods of time, i.e. in a fuzzy way. Aielli and Caporin (2013) motivate a soft clustering based on mixture models arguing that whether similarity is based on estimated dynamic parameters, then the error estimation generates variability causing overlapping clusters. Although hard methods have received greater attention in the time series clustering literature, a number of recent contributions have adopted the fuzzy approach combined with different dissimilarity criteria between series, including distances based on autocorrelation functions (D'Urso and Maharaj, 2009), features extracted in the frequency domain such as normalized periodogram and its logarithm, and cepstral coefficients (Maharaj and D'Urso, 2011), autoregressive approximations (D'Urso et al., 2013b), wavelet analysis (Tseng et al., 2010; Maharaj et al., 2010; D'Urso and Maharaj, 2012) and estimated GARCH coefficients (D'Urso et al., 2013a, 2016).

This chapter is aimed at assessing the behavior of the distance based on estimated quantile autocovariances (QAF) in partitional clustering of time series by considering a fuzzy approach. Again, we assume that the target is to group series according to the underlying dependence structures, i.e. similarity between series is measured in terms of similarity between generating processes. As mentioned, in this framework, the use of a metric robust to the generating mechanism of the series is necessary to attain a proper cluster solution, and the QAF-based distance introduced in the above chapter reported very satisfactory results in hard clustering. Therefore, the motivation is clear: a fuzzy clustering algorithm considering this metric would be expected to show a proper behavior. Furthermore, the problem of dealing with anomalous fuzzy data is also addressed in the second part of the present chapter. Anomalous time series might have a disruptive effect on the clustering process, and hence the use of robust fuzzy clustering models is of great interest in practice.

The first contribution in this chapter consists of introducing a novel fuzzy procedure to cluster time series. We adopt a fuzzy C-medoids approach where the QAF-based metric is considered to compute distances between series and medoids. In this way, the proposed approach inherits the advantages of the fuzzy methods (flexibility to describe complex cluster structures with overlapping clusters), the partitioning around medoids technique (selection

of particular series representing the underlying cluster patterns) and the QAF-based metric (high capability to discriminate between a broad range of dependence structures). Once the fuzzy algorithm is introduced, its behavior is evaluated via simulations. Here, our experiments mainly focused on the classification of heteroskedastic models, a complex scenario but frequently realistic when analyzing financial, industrial or environmental indicators, among others. The capability of the proposed model to clustering GARCH models is examined, and its performance is tested against two fuzzy clustering algorithms considering GARCH-based dissimilarities (D'Urso et al., 2013a), and therefore specifically designed to work in the simulated scenario. The fuzzy clustering algorithm is applied to two study cases considering air quality data and daily returns of stocks to illustrate its usefulness in practice.

The second contribution deals with the problem of the detection and neutralization of outliers. Overall, the presence of anomalous data can prevent from correctly identifying the hidden clustering structure of the data at hand, and hence introducing robust fuzzy methods is a valuable issue. In fact, performing fuzzy clustering in presence of anomalous data is a very interesting line of research in the clustering literature, and different approaches to face this problem have been proposed. Dave (1991) and Dave and Sen (1997) attain robustness by creating a fictitious cluster called noise cluster where all the outliers are assigned. Kim et al. (1996) use the least trimmed squares technique applied to prototype-based clustering algorithms such as the C-means and the fuzzy C-Means to make them robust. Winkler et al. (2011) present a fuzzy clustering algorithm with polynomial fuzzifier function in connection with M-estimators using a normalization function of the robust weights. Wu and Yang (2002) propose to use a new metric (exponential metric) which is more robust to the existence of outliers. An overview of several robust fuzzy methods can be seen in Klawonn and Höppner (2009).

In the time series framework and regarding the clustering purpose in mind, a time series is considered as an outlier when exhibits an atypical dynamic behavior, which substantially differs from the rest of identified prototypes. Three robust versions of the fuzzy C-medoids clustering algorithm for the classification of time series based on comparing estimated sequences of quantile autocovariances are introduced and compared. Specifically, (i) QAF-based exponential fuzzy C-medoids clustering, (ii) QAF-based fuzzy C-medoids clustering with noise cluster, and (iii) QAF-based trimmed fuzzy C-medoids clustering. The first model uses a robust metric to neutralize and smooth the effect of outliers, the second one is aimed at detecting outliers by classifying them into a noise cluster, and with the third method the model achieves its robustness by trimming away a certain fraction of the furthest time series. All of these models are robust extensions of the QAF-based fuzzy

C-medoids clustering model introduced in the first part of the chapter. Recent works have followed analogous robust approaches but using the AR distance (D'Urso et al., 2013b, 2015b, 2017) and distances considering underlying heteroskedastic models (D'Urso et al., 2016). To gain insight on the capability of the proposed robust models, all of these procedures were compared by means of an extensive simulation study including ARMA and GARCH models and in the presence of outliers. Obviously the alternative procedures take advantage of being specifically constructed to discriminate between these processes, and hence we can obtain a realistic measure of the capability of the QAF-based procedures. The usefulness and effectiveness of the proposed robust fuzzy models is also highlighted by considering an application in finance.

The rest of the chapter is organized as follows. The proposed C–medoids fuzzy clustering algorithm based on the estimated quantile autocovariances is described in Section 4.2. Section 4.3 presents some results from a simulation study conducted to analyze its performance under different generating processes, including linear and conditional heteroskedastic processes. Unlike the previous chapters, the simulation scenarios are now characterized by the fuzzy nature of the clusters by introducing additional uncertainty on the parameters defining the generating processes. Two applications on real data involving time series of air quality data and daily returns of stocks in IBEX-35 are carried out in Section 4.4. The robust models based on the C–medoids fuzzy clustering algorithm considering d_{QAF} are described and discussed in Section 4.5, and the results from simulations are presented in Section 4.6. Illustrative application of these robust fuzzy approaches on an study case is shown in Section 4.7, and some concluding remarks are summarized in Section 4.8.

4.2 QAF-based fuzzy C-medoids clustering model

Consider a set S of n realizations of univariate time series $\left\{\boldsymbol{X}_t^{(1)}, \ldots, \boldsymbol{X}_t^{(n)}\right\}$ subjected to clustering, and denote by $\boldsymbol{\Gamma} = \left\{\boldsymbol{\Gamma}^{(1)}, \ldots, \boldsymbol{\Gamma}^{(n)}\right\}$ the corresponding vectors of estimated quantile autocovariances computed as defined in (3.17). Assume that all vectors $\boldsymbol{\Gamma}^{(i)}$ have the same length Lr^2, being L and r the numbers of lags and quantile levels considered for all the series, respectively. This way, the pairwise d_{QAF} distances between two arbitrary series can be computed according to (3.19). In this framework, we propose to perform partitional fuzzy clustering on S by means of the QAF-based Fuzzy C-Medoids Clustering model (QAF-FCMdC), which aims at finding the subset of $\boldsymbol{\Gamma}$ of size C, $\widetilde{\boldsymbol{\Gamma}} = \left\{\widetilde{\boldsymbol{\Gamma}}^{(1)}, \ldots, \widetilde{\boldsymbol{\Gamma}}^{(C)}\right\}$, and the $n \times C$ matrix of fuzzy coefficients $\boldsymbol{\Omega} = (u_{ic})$, $i = 1, \ldots, n$, $c = 1, \ldots, C$, that lead to solve the minimization problem:

$$\begin{cases} \min\limits_{\widetilde{\mathbf{\Gamma}},\boldsymbol{\Omega}} \sum\limits_{i=1}^{n}\sum\limits_{c=1}^{C} u_{ic}^{m} \left\| \mathbf{\Gamma}^{(i)} - \widetilde{\mathbf{\Gamma}}^{(c)} \right\|^2 \\ \text{subject to: } \sum\limits_{c=1}^{C} u_{ic} = 1 \text{ and } u_{ic} \geq 0, \end{cases} \tag{4.1}$$

where $u_{ic} \in [0,1]$ represents the membership degree of the i-th series in the c-th cluster, $\widetilde{\mathbf{\Gamma}}^{(c)}$ is the vector of quantile autocovariances associated to the medoid series for the cluster c, and $m > 1$ is a weighting exponent that controls the fuzziness of the partition. Constraints on u_{ic} are standard requirements in fuzzy clustering. In particular, that the sum of the membership degrees for each series equals 1 implies that all of them contribute with the same weight to the clustering process. Parameter m determines the level of fuzziness introduced in the clustering procedure. In the naive case $m = 1$, we have $u_{ic} = 1$ if the i-th series is the medoid for the cluster c and 0 otherwise so that the crisp version of the procedure is obtained. As the value of m increases, the boundaries between clusters become softer and therefore the classification is fuzzier.

In a nutshell, the aim of QAF-FCMdC model is to determine a fuzzy partition into C clusters such that the QAF-distance between the clusters and their prototypes is minimized. Likewise the crisp approach, the clustering quality strongly depends on the capability of d_{QAF} to identify different dependence structures, but now the non-stochastic uncertainty inherent to the assignment of series to clusters is incorporated to the procedure by means of the membership degrees.

An iterative algorithm that alternately optimizes the membership degrees and the medoids is used to solve the optimization problem in (4.1). First, the membership degrees are optimized for a set of fixed medoids. The iterative solutions for the membership degrees take the form (Höppner et al., 1999):

$$u_{ic} = \left[\sum_{c'=1}^{C} \left(\frac{\left\| \mathbf{\Gamma}^{(i)} - \widetilde{\mathbf{\Gamma}}^{(c)} \right\|^2}{\left\| \mathbf{\Gamma}^{(i)} - \widetilde{\mathbf{\Gamma}}^{(c')} \right\|^2} \right)^{\frac{1}{m-1}} \right]^{-1}, \quad \text{for } i = 1,\ldots,n \text{ and } c = 1,\ldots,C. \tag{4.2}$$

Then, based on the membership degrees obtained from (4.2), the C series minimizing (4.1) are selected as new medoids. This two-step procedure is iterated until there is no change in the medoids or a maximum number of iterations is achieved. The initial values for the medoids are usually determined at random, but the procedure is very sensitive to an unsuitable choice. In this case, the initial set of medoids was obtained after running a hard PAM algorithm based on the QAF dissimilarity.

The QAF-based fuzzy C-medoids clustering algorithm (QAF-FCMdC) is implemented as outlined in Algorithm 1.

Algorithm 1 The QAF-based Fuzzy C-Medoids Clustering Algorithm (QAF–FCMdC)

Fix C, m and $max.iter$
Set $iter = 0$
Pick the initial medoids $\widetilde{\mathbf{\Gamma}} = \left\{\widetilde{\mathbf{\Gamma}}^{(1)}, \ldots, \widetilde{\mathbf{\Gamma}}^{(C)}\right\}$
repeat
    Set $\widetilde{\mathbf{\Gamma}}_{\text{OLD}} = \widetilde{\mathbf{\Gamma}}$ {Store the current medoids}
    Compute u_{ic}, $i = 1, \ldots, n$, $c = 1, \ldots, C$, using (4.2)
    For each $c \in \{1, \ldots, C\}$, determine the index $j_c \in \{1, \ldots, n\}$ satisfying:
    $$j_c = \underset{1 \leq j \leq n}{\operatorname{argmin}} \sum_{i=1}^{n} u_{ic}^m \left\| \mathbf{\Gamma}^{(i)} - \mathbf{\Gamma}^{(j)} \right\|^2$$
    return $\widetilde{\mathbf{\Gamma}}^{(c)} = \mathbf{\Gamma}^{(j_c)}$, for $c = 1, \ldots, C$ {Update the medoids}
    $iter \leftarrow iter + 1$
until $\widetilde{\mathbf{\Gamma}}_{\text{OLD}} = \widetilde{\mathbf{\Gamma}}$ or $iter = max.iter$

4.3 Assessing the behavior of the QAF–FCMdC model: A simulation study

A simulation study was conducted to evaluate the performance of the proposed QAF–FCMdC algorithm. We intended to recreate fuzzy scenarios with different time series models, including realizations of AR, ARCH and GARCH processes. In all cases, the base scenario consisted of two clusters with five series each, let us say $\mathcal{C}_1$ and $\mathcal{C}_2$, and one additional time series located at equal distance from both clusters. Moreover, we add uncertainty to the classification procedure by two ways: (i) introducing variability over the parameters defining the underlying model for each cluster, and (ii) considering different levels of separation between the clusters. Variability within clusters was generated by drawing out the parameters at random according to uniform distributions with different support for each cluster. The distance between clusters is given by the distance between the means of the uniform distributions. The specific scenarios and the generation schemes for each scenario are described in detail in Table 4.1.

For all scenarios, innovations ε_t follow a Gaussian distribution with zero mean and unit variance. Compared to scenarios denoted by B, scenarios A exhibit greater distance between the clusters $\mathcal{C}_1$ and $\mathcal{C}_2$, and hence less balanced memberships are expected. In other terms, the five series generated from one specific cluster should group all together with membership degrees more markedly close to one in scenarios A. As far as the time series

Table 4.1: Simulation scenarios for evaluation of the QAF-FCMdC algorithm

Generating process	Scenario	Elements and structure
Scenario 4.1: Autoregressive processes AR(1)		
$X_t = \phi X_{t-1} + \varepsilon_t$	4.1.A	Cluster $\mathcal{C}_1$: 5 series with $\phi \sim U(0, 0.2)$ Cluster $\mathcal{C}_2$: 5 series with $\phi \sim U(0.8, 1)$ One equidistant series with: $\phi = 0.5$
	4.1.B	Cluster $\mathcal{C}_1$: 5 series with $\phi \sim U(0.2, 0.4)$ Cluster $\mathcal{C}_2$: 5 series with $\phi \sim U(0.6, 0.8)$ One equidistant series with: $\phi = 0.5$
Scenario 4.2: Autoregressive conditional heteroskedastic processes ARCH(1)		
$X_t = \sigma_t \varepsilon_t$, with $\sigma_t^2 = 0.1 + \alpha X_{t-1}^2$	4.2.A	Cluster $\mathcal{C}_1$: 5 series with $\alpha \sim U(0, 0.1)$ Cluster $\mathcal{C}_2$: 5 series with $\alpha \sim U(0.9, 1)$ One equidistant series with: $\alpha = 0.5$
	4.2.B	Cluster $\mathcal{C}_1$: 5 series with $\alpha \sim U(0, 0.2)$ Cluster $\mathcal{C}_2$: 5 series with $\alpha \sim U(0.8, 1)$ One equidistant series with: $\alpha = 0.5$
Scenario 4.3: General autoregressive conditional heteroskedastic processes GARCH(1,1)		
$X_t = \sigma_t \varepsilon_t$, with $\sigma_t^2 = 0.1 + \alpha X_{t-1}^2 + 0.1\sigma_{t-1}^2$	4.3.A	Cluster $\mathcal{C}_1$: 5 series with $\alpha \sim U(0, 0.15)$ Cluster $\mathcal{C}_2$: 5 series with $\alpha \sim U(0.85, 0.9)$ One equidistant series with: $\alpha = 0.5$
	4.3.B	Cluster $\mathcal{C}_1$: 5 series with $\alpha \sim U(0.1, 0.2)$ Cluster $\mathcal{C}_2$: 5 series with $\alpha \sim U(0.8, 0.9)$ One equidistant series with: $\alpha = 0.5$

located at an intermediate situation between $\mathcal{C}_1$ and $\mathcal{C}_2$, it is expected that these series belong simultaneously to the two clusters showing membership degrees close to 0.5.

The QAF–FCMdC algorithm was compared with other fuzzy clustering models based on alternative dissimilarities. For Scenario 4.1, a fuzzy C-medoids algorithm considering Euclidean distances between estimated autoregressive representations was used. According to the fuzzy approach based on features extracted of the time series, a fuzzy model can be formalized as solution of the general optimization problem:

$$\min: \ \mathcal{F}_m\left(\widetilde{\mathbf{\Phi}}, \mathbf{\Omega}\right) = \sum_{i=1}^{n}\sum_{c=1}^{C} u_{ic}^m \left\|\mathbf{\Phi}^{(i)} - \widetilde{\mathbf{\Phi}}^{(c)}\right\|^2, \ \text{subject to:} \ \sum_{c=1}^{C} u_{ic} = 1 \text{ and } u_{ic} \geq 0, \tag{4.3}$$

where $\mathbf{\Phi}^{(i)}$ represents the vector of estimated features for the i-th series, $i = 1, \ldots, n$, and $\widetilde{\mathbf{\Phi}}$ denotes an arbitrary subset of C vectors $\mathbf{\Phi}^{(i)}$ denoted by $\widetilde{\mathbf{\Phi}}^{(c)}$, $c = 1, \ldots, C$. The subset $\widetilde{\mathbf{\Phi}}$ minimizing the objective function $\mathcal{F}_m$ involves the solution with the C medoids or prototype time series. The solution is iteratively reached by optimizing alternately medoids and membership degrees. At each iteration, the membership degrees for fixed medoids are obtained by using the Lagrangian multipliers method, resulting the update formula:

$$u_{ic} = \left[\sum_{c'=1}^{C} \left(\frac{\left\| \mathbf{\Phi}^{(i)} - \tilde{\mathbf{\Phi}}^{(c)} \right\|^2}{\left\| \mathbf{\Phi}^{(i)} - \tilde{\mathbf{\Phi}}^{(c')} \right\|^2} \right)^{\frac{1}{m-1}} \right]^{-1}, \quad \text{for } i = 1, \ldots, n \text{ and } c = 1, \ldots, C. \tag{4.4}$$

The use of different features in the class of fuzzy clustering models defined by (4.3) and (4.4) leads to distinct fuzzy algorithms. We have considered various features including single and partial autocorrelations but our best results were obtained by using autoregressive representations, which leads to the following fuzzy model.

- AR-FCMdC: Fuzzy C-medoids clustering model based on the AR-metric (D'Urso et al., 2013c). When the extracted features $\mathbf{\Phi}^{(i)}$, $i = 1, \ldots, n$, are the autoregressive representations of the time series, we take into consideration the distance introduced by Piccolo (1990) to deal with ARIMA models (Pértega and Vilar, 2010; Piccolo, 1990; Maharaj, 2000; Liao et al., 2008; Vilar et al., 2009). Each series $X_t^{(i)}$ is identified by the AR(∞) operator approximating its ARIMA structure. In practice, the truncated AR(∞) representations are used, and thus $X_t^{(i)}$ is characterized by the vector of AR(r_i) parameter estimates, $\widehat{\boldsymbol{\pi}}^{(i)} = \left(\widehat{\pi}_1^{(i)}, \ldots, \widehat{\pi}_{r_i}^{(i)} \right)$, where the r_i significant lags are obtained by means of a model selection criterion such as Akaike's Information Criterion (AIC). Then $\mathbf{\Phi}^{(i)} \equiv \boldsymbol{\pi}^{(i)}$, and we have in (4.3) and (4.4):

$$\left\| \mathbf{\Phi}^{(i)} - \widetilde{\mathbf{\Phi}}^{(c)} \right\|^2 = \sum_{u=1}^{r_{ic}} \left(\widehat{\pi}_u^{(i)} - \widetilde{\widehat{\pi}}_u^{(c)} \right)^2, \tag{4.5}$$

  where $r_{ic} = \max(r_i, r_c)$. When $r_i \neq r_c$, the shortest AR coefficient vector is completed by adding zeros up to have two vectors with the same length.

Scenarios 4.2 and 4.3 involve conditionally heteroskedastic models and the clustering task is substantially more complex due to the peculiar features exhibited for these processes (results in experiments of Chapter 3 illustrate this assertion). As in Scenario 4.1, it is desirable to examine our procedure against fuzzy clustering models based on suitable distances regarding the underlying heteroskedastic structures. At this aim, we select two partitioning around medoids algorithms based on GARCH modeling recently proposed by D'Urso et al. (2013a). Both models rely on distances employing the autoregressive representation of a GARCH(p,q) process. More precisely, the GARCH(p,q) model allows to model the serial dependence of the volatility by assuming that $X_t = \sigma_t \varepsilon_t$, where the innovations ε_t are independent and identically distributed variables and the squared disturbances σ_t^2 satisfy

the following ARMA(p,q) representation:

$$\sigma_t^2 = \text{Var}\left(X_t \,|\mathcal{J}_{t-1}\right) = \gamma + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2, \tag{4.6}$$

where $\mathcal{J}_{t-1} = \sigma\left(X_{t-1}, X_{t-2}, \ldots\right)$ represents the information available up to time $(t-1)$, $\gamma > 0$, $0 \leq \alpha_i < 1$ and $0 \leq \beta_j < 1$, for $i = 1, \ldots, p$ and $j = 1, \ldots, q$, and $(\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j) < 1$. Based on expression (4.6), two distance measures between heteroskedastic processes are introduced and plugged into the fuzzy C-medoids clustering model as described below.

- GARCH-FCMdC: Fuzzy C-medoids clustering model based on the AR distance between GARCH approximations (D'Urso et al., 2013a). Starting from (4.6) and after some algebra, it can be shown that

$$X_t^2 = \gamma + \sum_{i=1}^{p^\star} \left(\alpha_i + \beta_i\right) X_{t-i}^2 + \sum_{j=1}^{q} \beta_j \eta_{t-j} + \eta_t, \tag{4.7}$$

  with $p^\star = \max(p, q)$, $\alpha_i = 0$ for $i > p$, $\beta_i = 0$ for $i > q$, and $\eta_t = X_t^2 - \sigma_t^2$ a zero-mean error uncorrelated with the past. Equation (4.20) establishes an ARMA($p^\star$,q) representation for X_t^2, which can be approximated by an AR(∞) structure with autoregressive coefficients π_u^G given by

$$\pi_u^G = \left(\alpha_u + \beta_u\right) + \sum_{j=1}^{\min(q,u)} \beta_j \pi_{u-j}^G$$

  where $\pi_0^G = -1$, $\alpha_u = 0$ for $u > p$, and $\beta_u = 0$ for $u > q$. Then, GARCH-FCMdC model proceeds in the same line as AR–FCMdC but using estimators of these new autoregressive coefficients to compute the AR distance, i.e. replacing $\left(\widehat{\pi}_1^{(i)}, \ldots, \widehat{\pi}_{r_i}^{(i)}\right)$ by $\left(\widehat{\pi}_1^{G,(i)}, \ldots, \widehat{\pi}_{r_i}^{G,(i)}\right)$ in (4.5).

- GARCH-FCMdCC: Fuzzy C-medoids clustering model based on the Caiado and Crato distance between GARCH approximations (D'Urso et al., 2013a). Caiado and Crato (2007) proposed an alternative approach to measure distance between GARCH models by taking into account the covariance between the fitted GARCH coefficients. Specifically, the distance between a pair of series $\boldsymbol{X}_t^{(u)}$ and $\boldsymbol{X}_t^{(v)}$ is defined by

$$d_{GARCH}\left(\boldsymbol{X}_t^{(u)}, \boldsymbol{X}_t^{(v)}\right) = \left(\boldsymbol{L}^{(u)} - \boldsymbol{L}^{(v)}\right)^t \left(\boldsymbol{V}^{(u)} + \boldsymbol{V}^{(v)}\right)^{-1} \left(\boldsymbol{L}^{(u)} - \boldsymbol{L}^{(v)}\right), \tag{4.8}$$

  where $\boldsymbol{L}^{(u)} = \left(\widehat{\boldsymbol{\alpha}}^{(u)}, \widehat{\boldsymbol{\beta}}^{(u)}\right)^t$ is the estimated vector of parameters in the GARCH

representation (4.6) for $\boldsymbol{X}_t^{(u)}$, and $\boldsymbol{V}^{(u)} = \mathbb{E}\left(\boldsymbol{L}^{(u)}\boldsymbol{L}^{(u),t}\right)$ denotes the corresponding covariance matrix between the estimated parameters. Based on d_{GARCH}, the GARCH-FCMdCC model is derived by solving the minimization problem:

$$\min_{\widetilde{\boldsymbol{L}},\boldsymbol{\Omega}} \sum_{i=1}^{n}\sum_{c=1}^{C} u_{ic}^{m}\left(\boldsymbol{L}^{(i)}-\widetilde{\boldsymbol{L}}^{(c)}\right)^{t}\left(\boldsymbol{V}^{(i)}+\widetilde{\boldsymbol{V}}^{(c)}\right)^{-1}\left(\boldsymbol{L}^{(i)}-\widetilde{\boldsymbol{L}}^{(c)}\right),$$

$$\text{subject to } \sum_{c=1}^{C} u_{ic}=1 \text{ and } u_{ic}\geq 0,$$

where $\widetilde{\boldsymbol{L}}$ is a subset of cardinality C of estimated GARCH vectors for the series in study.

By comparing QAF-FCMdC with AR-FCMdC in Scenario 4.1 and with GARCH-FCMdC and GARCH-FCMdCC in Scenarios 4.2 and 4.3, we choose competitors based on distances properly adjusted to the dependence structures from each simulated scenario, and therefore valuable insight into the usefulness of the proposed model and its robustness to the generating process should be obtained.

The experiments were carried out with different lengths for the time series, namely $T = 250$, 500 and 1000 for Scenario 4.1, and $T = 500$, 1000 and 2000 for Scenarios 4.2 and 4.3. The size of the series is increased for the heteroskedastic scenarios to face the high variability of the estimated GARCH parameters. Based on a controlled simulation experiment, Aielli and Caporin (2013) assert that the standard quasi maximum likelihood GARCH estimates obtained from simulated realizations of a GARCH(1,1) process are characterized by higher dispersion for smaller sample sizes. These arguments account for choosing large sample sizes by treating with this kind of processes, and in fact the chosen lengths are commonly used in the literature (Aielli and Caporin, 2013; Bauwens and Rombouts, 2007; Otranto, 2010). Furthermore, large sample sizes are also usual in applications. A typical example of heteroskedastic series are the financial time series, which usually include longer sequences formed by daily or intra-daily data. Some experiments were also performed in Scenarios 4.2 and 4.3 with short series, but the results were poor, being particularly affected the fuzzy algorithms using GARCH-based distances due to inaccurate estimations of the GARCH parameters.

The fuzziness parameter m also has an important role, and in practice its value must be determined in advance. As already mentioned, $m = 1$ leads to a crisp partition, but very large values for m are not recommendable. Kamdar and Joshi (2000) argue that very high values for m may imply to lose mobility of the medoids because all membership degrees would become very small except the one corresponding to the current medoid, which always

equals 1 within its cluster. To our knowledge, there are no theoretical arguments supporting an optimal choice of m (see discussion in Yang et al., 2008). A popular choice is $m = 2$, although based on different heuristic arguments various authors suggest that the value of the proper level of fuzziness should be between 1.5 and 2.5 (Pal and Bezdek, 1995; Hall et al., 1992; Cannon et al., 1986). An interesting discussion on this point including related references is given in Section 3.1.6 of Maharaj and D'Urso (2011). We were very interested in checking the effect of moving the fuzzifier m, and based on the previous considerations we decided to take the values $m = 1.5$, 2.0, 2.2 and 2.5, which is also a consistent choice with other recent experimental studies (Maharaj and D'Urso, 2011; de A.T. de Carvalho et al., 2006).

The number of clusters was set at $C = 2$, and hence the equidistant series are forced to belong simultaneously to both clusters. At all scenarios, ten sets of 100 simulations were carried out. For each set was first calculated the percentage of times in which time series were correctly classified, and then these success rates were averaged over the ten replications. At each trial, the correct classification occurs when the five series generated from the model defining $\mathcal{C}_1$ are located together in one cluster, the five series coming from $\mathcal{C}_2$ are grouped together in another cluster, and the single series generated from an equidistant model is simultaneously located in both clusters. Since grouping is performed in a fuzzy framework, a cut-off value for the membership degrees must be fixed to decide when a time series is assigned to a specific cluster or to both clusters simultaneously. Our assignment rule was to place the i-th series into the c-th if $u_{ic} > 0.7$. In other case, the series is simultaneously located in the two clusters because of its membership degrees are reasonably similar (both of them between 0.3 and 0.7). It is worthy remarking that the chosen cut-off point is compatible with the indications suggested in the literature (see e.g. D'Urso and Maharaj, 2009; D'Urso et al., 2013b; Maharaj and D'Urso, 2011; Maharaj et al., 2010; D'Urso and Giordani, 2006; Dembélé and Kastner, 2003).

The average percentages of correct classification were obtained with all the fuzzy models in order to be compared. In the case of the QAF-FCMdC model, the distance d_{QAF} between estimated quantile autocovariances was evaluated over a grid of regularly spaced quantile levels formed by all the combinations $(0.05j, 0.05j')$, with j and j' ranging from 1 to 19. Concerning the GARCH-based models, it is important to remark that the right number of GARCH parameters was provided as an input in the computation of d_{AR} and d_{GARCH}. Indeed, this is a substantial advantage in favour of these models since the significant number of GARCH parameters must be estimated in real scenarios. Table 4.2 shows the results for Scenario 4.1.

The influence of the fuzziness parameter m is evident from Table 4.2. The value $m = 1.5$

Table 4.2: Average percentage of correct classification in Scenario 4.1

		Scenario 4.1.A			Scenario 4.1.B		
	Algorithm	$T = 250$	$T = 500$	$T = 1000$	$T = 250$	$T = 500$	$T = 1000$
$m = 1.5$	AR-FCMdC	31.5	45.1	52.0	17.2	26.6	31.5
	QAF-FCMdC	29.6	34.3	35.2	9.7	17.9	23.2
$m = 2.0$	AR-FCMdC	67.5	83.9	93.0	22.5	45.2	59.9
	QAF-FCMd	69.1	76.2	77.6	28.7	44.1	58.3
$m = 2.2$	AR-FCMdC	76.9	91.8	97.5	21.1	44.3	65.2
	QAF-FCMdC	80.8	84.3	88.0	33.8	51.2	67.5
$m = 2.5$	AR-FCMdC	84.0	97.1	99.7	16.6	40.1	65.6
	QAF-FCMdC	88.8	93.7	96.1	34.1	56.0	75.7

produced uniformly the worst percentages, and it is observed that the results seem to improve progressively when m increases. Note that using a high fuzzifier means to smooth the boundary between clusters, thus making more difficult to separate them. In particular, a reasonably high value for m implies a more uniform distribution of the membership degrees, thus benefiting the correct classification of the equidistant series. As expected, the success rates substantially decreased for the Scenario 4.1.B. By increasing the level of proximity between clusters, both procedures are more sensitive to the noise and frequently the equidistant series present a membership degree $u_{ic} > 0.7$ for some c, thus producing an important number of failed trials and reducing noticeably the global success rate (even though the series of each cluster are really well-classified). Lastly, Table 4.2 also shows that QAF-FCMdC produces competitive results compared to AR-FCMdC. In spite of AR-FCMdC is designed to deal with ARMA series, the proposed algorithm QAF-FCMdC exhibited a similar performance, drawing out a little worse success rates in the easiest Scenario 4.1.A, but somewhat higher ones in the most difficult Scenario 4.1.B, where the clusters are closer each other.

Now, we focus on Tables 4.3 and 4.4 where the results using ARCH(1) (Scenarios 4.2) and GARCH(1,1) (Scenarios 4.3) processes are presented, respectively.

The above considerations on the effect of the fuzziness parameter m also apply in these scenarios, and it is observed that the success rates improve when m increases. In the simplest Scenario 4.2, involving ARCH(1) models, the GARCH-based algorithms outperform the QAF-based procedure when $T = 1000$. Nevertheless, by dealing with series of length $T = 2000$, QAF-FCMdC is clearly competitive, exhibiting better behavior than GARCH-FCMdCC and only a little worse than GARCH-FCMdC. As expected, again the success rates in Scenario 4.2.B are worse then the ones in Scenario 4.2.A. For the largest length of series ($T = 5000$) and regardless of the amount of separability of the clusters (Scenar-

Table 4.3: Average percentage of correct classification in Scenario 4.2

		Scenario 4.2.A			Scenario 4.2.B		
	Algorithm	$T = 1000$	$T = 2000$	$T = 5000$	$T = 1000$	$T = 2000$	$T = 5000$
$m = 1.5$	GARCH-FCMdC	32.5	48.5	64.3	30.1	40.1	64.4
	GARCH-FCMdCC	20.4	17.8	10.1	20.4	21.8	20.0
	QAF-FCMdC	15.8	26.9	44.0	12.1	19.4	44.2
$m = 2.0$	GARCH-FCMdC	71.9	86.3	86.6	64.2	76.2	86.0
	GARCH-FCMdCC	47.1	51.3	53.4	48.5	51.2	53.1
	QAF-FCMdC	47.6	70.0	88.2	36.9	57.6	88.2
$m = 2.2$	GARCH-FCMdC	79.7	91.2	87.7	72.5	82.7	87.1
	GARCH-FCMdCC	56.5	65.7	69.3	56.9	62.0	70.7
	QAF-FCMdC	58.9	81.5	94.9	42.4	69.8	95.2
$m = 2.5$	GARCH-FCMdC	88.2	93.9	87.7	78.9	88.5	87.5
	GARCH-FCMdCC	70.6	82.4	85.2	67.6	75.4	84.4
	QAF-FCMdC	59.8	89.5	99.0	37.0	79.0	99.2

Table 4.4: Average percentage of correct classification in Scenario 4.3

		Scenario 4.3.A			Scenario 4.3.B		
	Algorithm	$T = 1000$	$T = 2000$	$T = 5000$	$T = 1000$	$T = 2000$	$T = 5000$
$m = 1.5$	GARCH-FCMdC	17.9	18.3	19.3	14.5	18.5	19.4
	GARCH-FCMdCC	5.0	13.5	15.0	5.3	11.9	14.1
	QAF-FCMdC	12.9	23.8	39.2	10.4	18.7	40.0
$m = 2.0$	GARCH-FCMdC	39.5	47.1	52.1	33.6	45.6	51.9
	GARCH-FCMdCC	4.7	15.9	33.5	4.5	11.5	27.3
	QAF-FCMdC	38.9	66.2	84.4	30.1	56.9	84.5
$m = 2.2$	GARCH-FCMdC	46.0	59.5	62.2	40.9	54.5	64.0
	GARCH-FCMdCC	3.6	9.3	32.2	3.3	7.4	22.4
	QAF-FCMdC	47.5	77.7	93.9	32.7	70.0	93.8
$m = 2.5$	GARCH-FCMdC	51.4	76.5	81.6	48.8	70.7	82.7
	GARCH-FCMdCC	2.7	2.8	20.2	1.9	3.4	15.7
	QAF-FCMdC	43.3	84.9	98.6	29.4	76.2	98.1

ios 4.2.A or 4.2.B), QAF-FCMdC is fairly the best procedure by attaining percentages of correct classification moving from 88% to 99%, while its competitors produced percentages always below 88%. Therefore, in spite of the GARCH-based algorithms take advantage of knowing the underlying parametric dependence structure (also the number of significant parameters in our simulations), QAF-FCMdC showed a similar behavior when high values of T were considered, and particularly excellent results (percentages of correct classification close to 100%) for series of length $T = 5000$.

This good performance of the proposed fuzzy algorithm is still more noticeable in Scenario 4.3, with GARCH(1,1) models. In fact, Table 4.4 shows that QAF-FCMdC produced

very similar (slightly lower) results as the ones obtained in Table 4.3, thus exhibiting an interesting robustness to the generating models. On the contrary, the fuzzy algorithms based on the GARCH metrics were strongly affected by the high variability in the estimation procedure of the GARCH parameters, corresponding to the GARCH-FCMdCC model the worst behavior. It must be noted that GARCH-FCMdCC requires the additional estimation of the covariance structure between GARCH parameters. The fuzzy approach based on the distance between quantile autocovariances is free of determining the underlying parametric structure and takes advantage of its enormous potential to detect complex types of dependence. These arguments account for the best results achieved by the proposed algorithm in this scenario. Only with $T = 1000$, GARCH-FCMdC seems to outperform QAF-FCMdC, but here is worthy to point out the importance of the length of the series in these heteroskedastic scenarios. In fact, the success rates for $T = 1000$ were substantially lower than using lengths 2000 and 5000, and always below 50%. Thus it is evident that larger lengths should be used. In short, although the good bahavior of the QAF-based distance in clustering of heteroskedastic series was already observed by performing hard cluster analysis, the results presented in this section also illustrate how the fuzzy nature of time series presenting features intermediate between different conditionally heteroskedastic models is well-captured by a fuzzy algorithm based on d_{QAF}.

4.4 Applications

In this section, two study cases considering air quality data and daily returns of stocks are presented to illustrate the usefulness of the fuzzy C-medoids clustering algorithm based on quantile autocovariances. In both applications, results from different fuzzy clustering models are discussed and compared to obtain a valuable insight into the behavior of our proposal.

4.4.1 Application to air quality data

The first study case is related to the non-supervised classification of geographical zones in terms of their temporal records of air pollutants. Specifically, we have considered time series of daily averages of concentrations of nitrogen dioxide (NO_2) and ozone (O_3), from 1st November 2006 to 31th December 2009. All data are sourced from the official website of the Air Quality Monitoring Network of Madrid Community [1].

For monitoring of emission levels, the Community of Madrid, which is an autonomous

[1] http://gestiona.madrid.org/azul_internet/run/j/AvisosAccion.icm

community of Spain, has a Network Control Air Quality consisting of a set of 21 fixed automatic stations and two mobile reference laboratories (a mobile unit and a bus). These 23 stations provide data on the air pollutant concentration along time and are distributed on 7 homogeneous areas that can be grouped in urban areas (Madrid, Urban North, Urban South and Henares Corridor) and rural areas (Northern Sierra, Alberche basin and Tajuña basin. We have used information extracted from 19 of the 23 stations (four stations were discarded because the database was not complete), namely Alcalá de Henares, Alcobendas, Torrejón de Ardoz, Arganda del Rey, Rivas Vaciamadrid, Leganés, Fuenlabrada, Móstoles, Aranjuez, Valdemoro, Majadahonda, Colmenar Viejo, Collado Villalba, Guadalix de la Sierra, El Atazar, S. Martín de Valdeiglesias, Villa del Prado, Villarejo de Salvanés and Orusco de Tajuña. Figure 4.1 shows the geographical distribution of the stations forming the network.

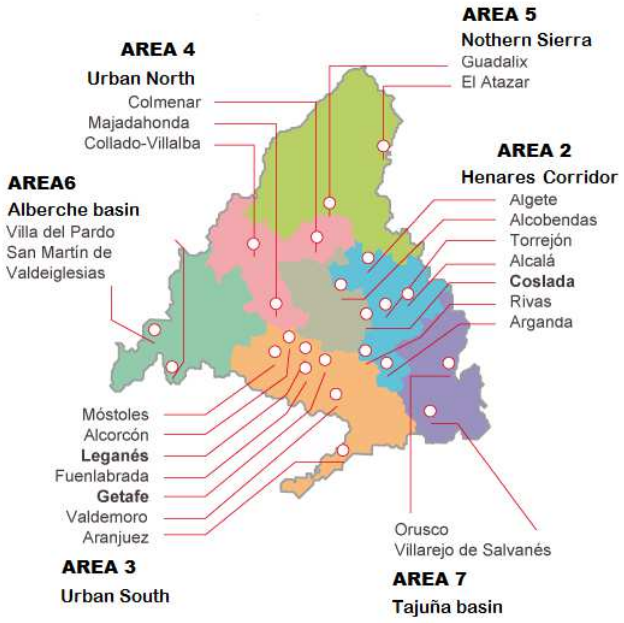

Figure 4.1: Location of the stations forming the Air Quality Monitoring Network of Madrid Community.

Several studies have revealed serious health effects associated with the continuous exposure to high concentrations of nitrogen dioxide and ozone, and for this reason we have focused on them. While some works have considered the problem of checking by significant differences between the mean levels of these pollutants on different areas of the community (see e.g. Estévez-Pérez and Vilar, 2013), our concern is to analyze the capability of the fuzzy clustering approach to identify locations with similar daily changes in levels of NO_2 and O_3. Nevertheless, it is important to remark that our motivation is only to illustrate the use-

fulness of the proposed fuzzy algorithm, without seeking to give any type of environmental implications.

The 19 time series available are formed by $T = 1,154$ records and are non-stationary in mean, thus we proceeded to transform them taking one regular difference. Figures 4.2 and 4.3 show plots of the transformed series for the levels of O_3 and NO_2, respectively. It is observed that the variance is not constant over time. There are periods of time in which the series strongly fluctuate, while in others fluctuation is less marked. On the other hand, it is reasonable to think that a fuzzy behavior might be present, with time series sharing features of different and well-defined patterns of daily changes of concentrations of NO_2 and O_3. According to our simulations results, the proposed fuzzy C-medoids clustering algorithm, QAF-FCMdC, should lead to a proper fuzzy partition of the stations. Just as in simulations, distance d_{QAF} was computed by considering one lag ($L = 1$, with $l_1 = 1$) and a grid of quantile levels formed by all the combinations $(0.05j, 0.05j')$, with $j, j' \in \{1, \ldots, 19\}$. For the purpose of comparison, the AR-FCMdC and GARCH-FCMdC fuzzy algorithms were also carried out. Given the underlying heteroskedasticity (particularly evident in the NO_2 case), the latter is expected to produce better results than the former.

A fuzzy extension of the classical silhouette width criterion was used to determine the optimal number of clusters. This fuzzy version takes into consideration the membership degrees matrix and consists in selecting the number of clusters maximizing the so-called Fuzzy Silhouette Width (Campello and Hruschka, 2006), defined by

$$\mathrm{FSW} = \frac{\sum_{i=1}^{n} (u_{ir} - u_{iv})^{\alpha} s_i}{\sum_{i=1}^{n} (u_{ir} - u_{iv})^{\alpha}}$$

where s_i is the standard silhouette width for the i-th element, u_{ir} and u_{iv} are the first and the second largest elements of the i-th row of the fuzzy partition matrix and $\alpha \geq 0$ is a weighting coefficient. This way, FSW provides a weighted average of the individual silhouette widths, thus permitting to underweight series belonging to overlapping clusters. The value $\alpha = 1$ is commonly considered, and it was also used in our application.

Figure 4.4 shows the values of the standard (crisp) and fuzzy silhouette indexes for a range of partition sizes using the QAF-FCMdC algorithm. In all cases, the existence of two major groups is concluded. Focusing on the fuzzy approach, the highest FSW indexes were 0.854 for O_3 and 0.773 for NO_2, corresponding to partitions of two clusters in both cases. Note that these high values suggest a strong clustering structure. On the contrary, considering more than two clusters substantially reduces the values of the FSW indexes, particularly in the case of NO_2. On the other hand, a two-cluster partition is intuitively consistent with a

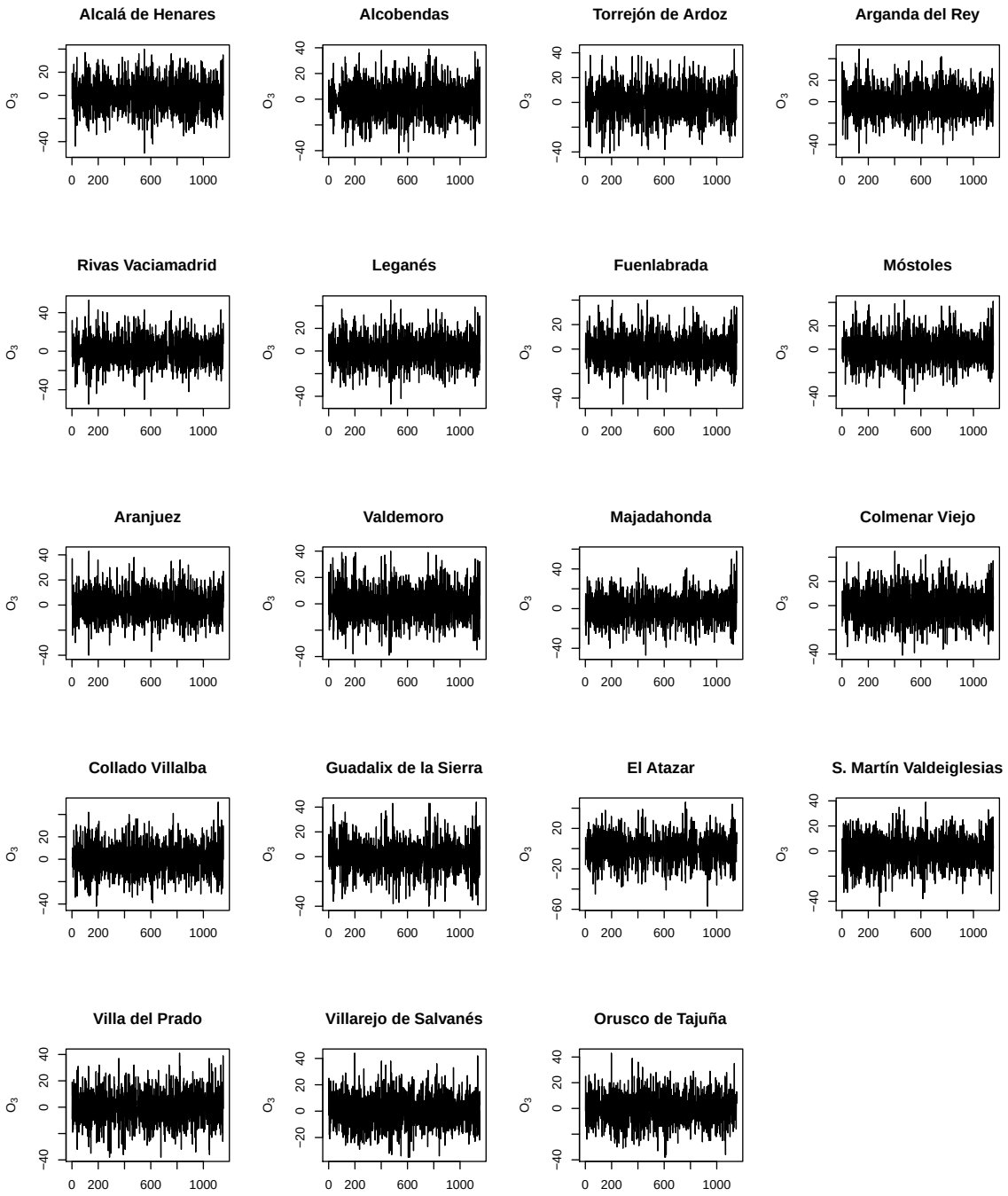

Figure 4.2: Daily series of O_3 levels transformed by taking one regular difference.

natural grouping of the stations according to their location in urban or rural areas. Based on these arguments we decide to set the number of clusters at $C = 2$.

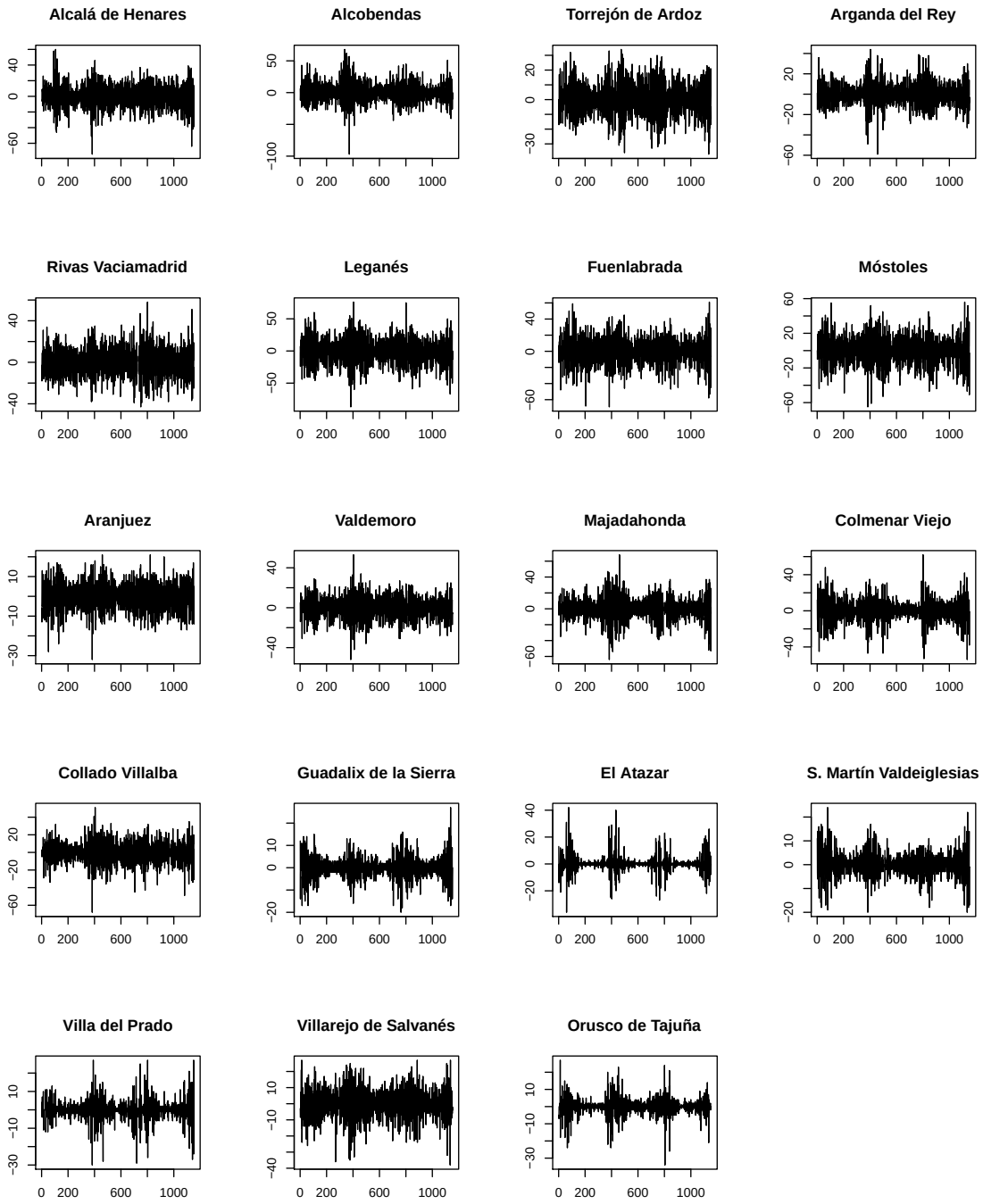

Figure 4.3: Daily series of NO_2 levels transformed by taking one regular difference.

The 2-cluster solutions for the series of daily changes in levels of O_3 using a fuzziness parameter $m = 2$ are shown in Table 4.5. For each single series, the shaded cells enhance the highest membership degrees obtained with each procedure, i.e. the cluster assignments

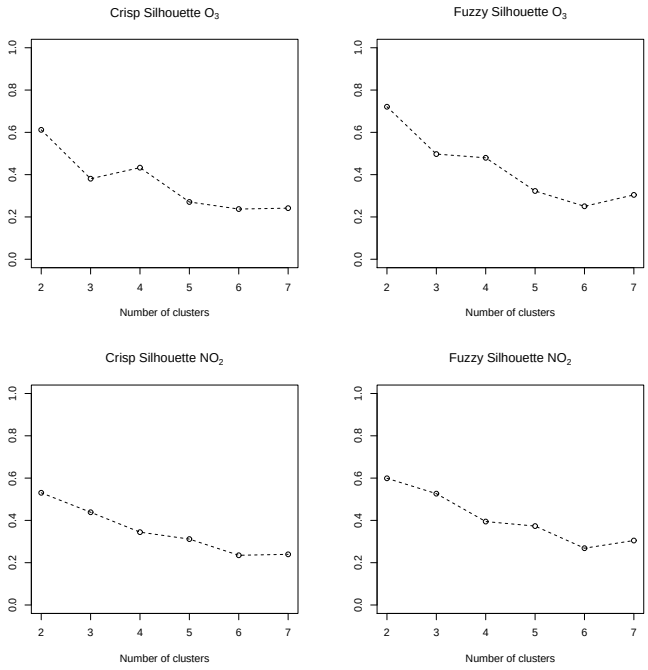

Figure 4.4: Crisp and fuzzy silhouette width values for a different number partitions using QAF–FCMdC.

from a crisp perspective. Stations with both membership degrees within $(0.3, 0.7)$ are fuzzy allocated between the two clusters and their memberships are shown in bold font.

In essence, the model QAF-FCMdC produces the expected classification by grouping the series of daily changes in O_3 according to the kind of location where they were monitored, i.e. stations placed in urban (cluster $\mathcal{C}_1$) and rural (cluster $\mathcal{C}_2$) areas. The group $\mathcal{C}_2$ brings together all the stations located in rural areas with memberships always above 0.798, but also including the stations of Aranjuez, Majadahonda and Colmenar Viejo. Actually, Aranjuez presents a vague location which might be explained because, despite being in an urban area, Aranjuez is located far from the rest of stations, just in the boundary of the Community. Also, in terms of ozone records, Majadahonda is set as a suburban location (see website of the Air Quality Monitoring Network of Madrid Community), which might account for its allocation in $\mathcal{C}_2$. All the stations in cluster $\mathcal{C}_1$ belong to urban areas presenting very high membership degrees for this cluster.

The results obtained with the models AR-FCMdC and GARCH-FCMdC cannot be meaningfully interpreted, at least in terms of rural and urban locations. While the model GARCH-FCMdC draws out a solution where just one cluster gathers almost all the series with memberships very close to one, the model AR-FCMdC identifies the two areas

Table 4.5: Membership degrees in clustering of the daily change series in levels of O_3 ($C = 2$ and $m = 2.2$).

		AR-FCMdC		GARCH-FCMdC		QAF-FCMdC	
Station	Area	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_1$	$\mathcal{C}_2$
Alcalá de Henares	Urban	0.740	0.260	0.992	0.008	0.961	0.039
Alcobendas	Urban	0.787	0.213	1.000	0.000	0.891	0.109
Torrejón de Ardoz	Urban	0.819	0.181	0.973	0.027	0.809	0.191
Arganda del Rey	Urban	0.724	0.276	1.000	0.000	0.832	0.168
Rivas Vaciamadrid	Urban	0.708	0.292	0.971	0.029	0.893	0.107
Leganés	Urban	**0.526**	**0.474**	0.993	0.007	0.954	0.046
Fuenlabrada	Urban	1.000	0.000	0.985	0.015	1.000	0.000
Móstoles	Urban	**0.511**	**0.489**	0.998	0.002	0.920	0.080
Aranjuez	Urban	**0.634**	**0.366**	0.932	0.068	**0.356**	**0.644**
Valdemoro	Urban	0.791	0.209	1.000	0.000	0.934	0.066
Majadahonda	Urban	**0.556**	**0.444**	1.000	0.000	0.225	0.775
Colmenar Viejo	Urban	**0.529**	**0.471**	0.995	0.005	0.000	1.000
Collado Villalba	Urban	**0.532**	**0.468**	0.995	0.005	0.781	0.219
Guadalix de la Sierra	Rural	0.000	1.000	0.000	1.000	0.191	0.809
El Atazar	Rural	**0.383**	**0.617**	**0.549**	**0.451**	0.132	0.868
San Martín de Valdeiglesias	Rural	0.299	0.701	0.988	0.012	0.074	0.926
Villa del Pardo	Rural	**0.312**	**0.688**	0.932	0.068	0.202	0.798
Villarejo de Salvanés	Rural	**0.449**	**0.551**	0.991	0.009	0.117	0.883
Orusco de Tajuña	Rural	**0.469**	**0.531**	0.950	0.050	0.112	0.888

but in a very fuzzy manner in most of the series. A simple way to visualize and compare the 2-cluster solutions obtained with the three models is provided by Figure 4.5, where the membership degrees for cluster $\mathcal{C}_1$ are depicted and the final assignment indicated.

The 2-cluster solutions for the series of daily changes in levels of NO_2 are shown in Table 4.6 and Figure 4.6. Note that GARCH-FCMdC and QAF-FCMdC lead to a very similar partition. With both models the distribution of the stations in rural and urban areas is still more evident than in the case of the ozone records. Among the rural locations, only Villarejo de Salvanés presents a pattern congruent with the urban locations, exhibiting the highest membership degree for $\mathcal{C}_1$ with both algorithms. Again Majadahonda station is unexpectedly placed into $\mathcal{C}_2$ and the only discrepancy is the classification of Colmenar Viejo. In contrast, AR-FCMdC again increases the fuzziness of the resulting partition, which is fairly no congruent with the grouping in urban and rural areas. It is worth noting that $C = 2$ is also the value maximizing the FSW index when AR-FCMdC is used, but in this case FSW = 0.66, substantially lower then the values 0.83 and 0.77 obtained with GARCH-FCMdC and QAF-FCMdC, respectively.

The main conclusions from this study case can be summarized as follows. Our fuzzy clustering approach, QAF-FCMdC, led to partitions with a meaningful interpretation for the two considered pollutants by grouping almost all stations according to their urban

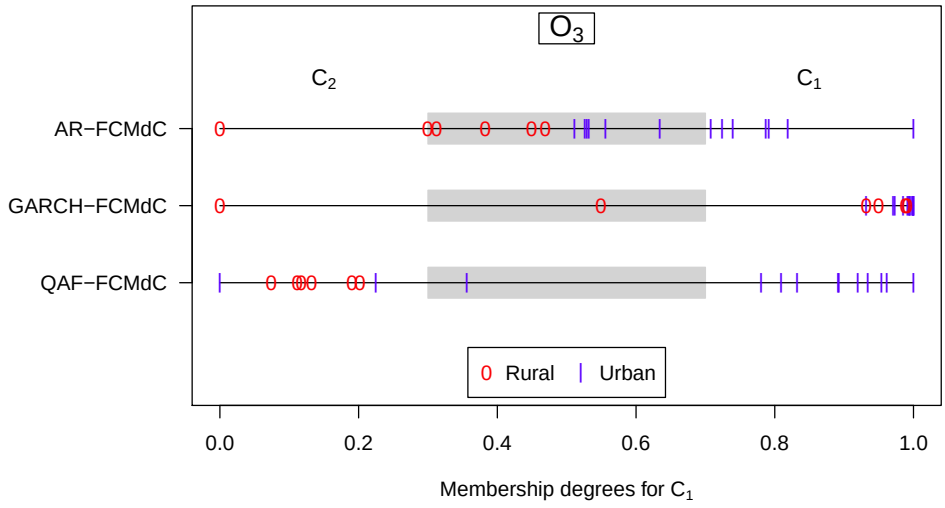

Figure 4.5: Membership degrees for cluster $\mathcal{C}_1$ in clustering of the daily changes in levels of O_3

Table 4.6: Membership degrees in clustering of the daily change series in levels of NO_2 ($C = 2$ and $m = 2.2$).

		AR-FCMdC		GARCH-FCMdC		QAF-FCMdC	
Station	Area	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_1$	$\mathcal{C}_2$
Alcalá de Henares	Urban	**0.591**	**0.409**	0.982	0.018	0.799	0.201
Alcobendas	Urban	**0.512**	**0.488**	0.850	0.150	0.866	0.134
Torrejón de Ardoz	Urban	**0.480**	**0.520**	0.965	0.035	1.000	0.000
Arganda del Rey	Urban	**0.598**	**0.402**	0.841	0.159	0.808	0.192
Rivas Vaciamadrid	Urban	0.197	0.803	0.828	0.172	0.791	0.209
Leganés	Urban	0.840	0.160	1.000	0.000	**0.679**	**0.321**
Fuenlabrada	Urban	0.872	0.128	1.000	0.000	0.812	0.188
Móstoles	Urban	1.000	0.000	0.815	0.185	0.794	0.206
Aranjuez	Urban	0.000	1.000	0.894	0.106	0.797	0.203
Valdemoro	Urban	0.726	0.274	0.975	0.025	0.828	0.172
Colmenar Viejo	Urban	0.778	0.222	0.249	0.751	0.760	0.240
Majadahonda	Urban	0.828	0.172	0.063	0.937	0.286	0.714
Collado Villalba	Urban	0.840	0.160	0.999	0.001	0.848	0.152
Guadalix de la Sierra	Rural	0.737	0.263	0.000	1.000	0.000	1.000
El Atazar	Rural	**0.303**	**0.697**	0.082	0.918	0.275	0.725
San Martín de Valdeiglesias	Rural	0.729	0.271	0.162	0.838	0.141	0.859
Villa del Pardo	Rural	0.710	0.290	0.120	0.880	0.218	0.782
Villarejo de Salvanés	Rural	0.287	0.713	0.957	0.043	0.823	0.177
Orusco de Tajuña	Rural	0.270	0.730	0.171	0.829	0.262	0.738

or rural location. The approach based on GARCH approximations, GARCH-FCMdC, performed in a similar way for the NO_2 series, but produced an unexpected and anomalous

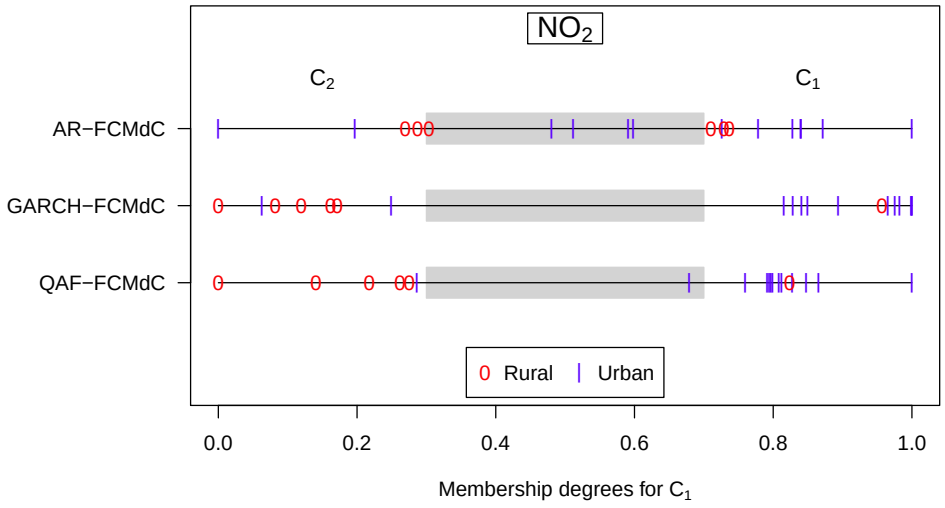

Figure 4.6: Membership degrees for cluster $\mathcal{C}_1$ in clustering of the daily changes in levels of O_3

partition for the ozone records, thus showing less accuracy than QAF-FCMdC. Since the observed series exhibit heteroskedasticity, the AR-FCMdC approach relies on misspecified models, which might account for the obtained partitions, with lower quality indexes and hardly interpretable for the two air pollutants. Lastly, it is worth enhancing that all the procedures have determined some series showing fuzzy nature, which supports the usefulness of the fuzzy approach.

4.4.2 Application to daily stocks returns in IBEX-35 index

The second application considers daily returns of stocks included in the IBEX-35, which groups the thirty-five companies with the highest liquidity and trading volume in the Spanish stock market. Specifically, we manage a database formed by the daily returns of twenty-four stocks located in the TOP-30 ranking according to the finance section of the Yahoo website[2]. The period of observation of the series spans from 1st January 2008 to 19th December 2016, thus resulting realizations of length $T = 2337$. The daily adjusted closing prices for all the stocks were sourced from the mentioned website and used to obtain the daily returns by considering the first differences of their natural logarithms. The time series are depicted in Figure 4.7.

[2] `https://finance.yahoo.com/`

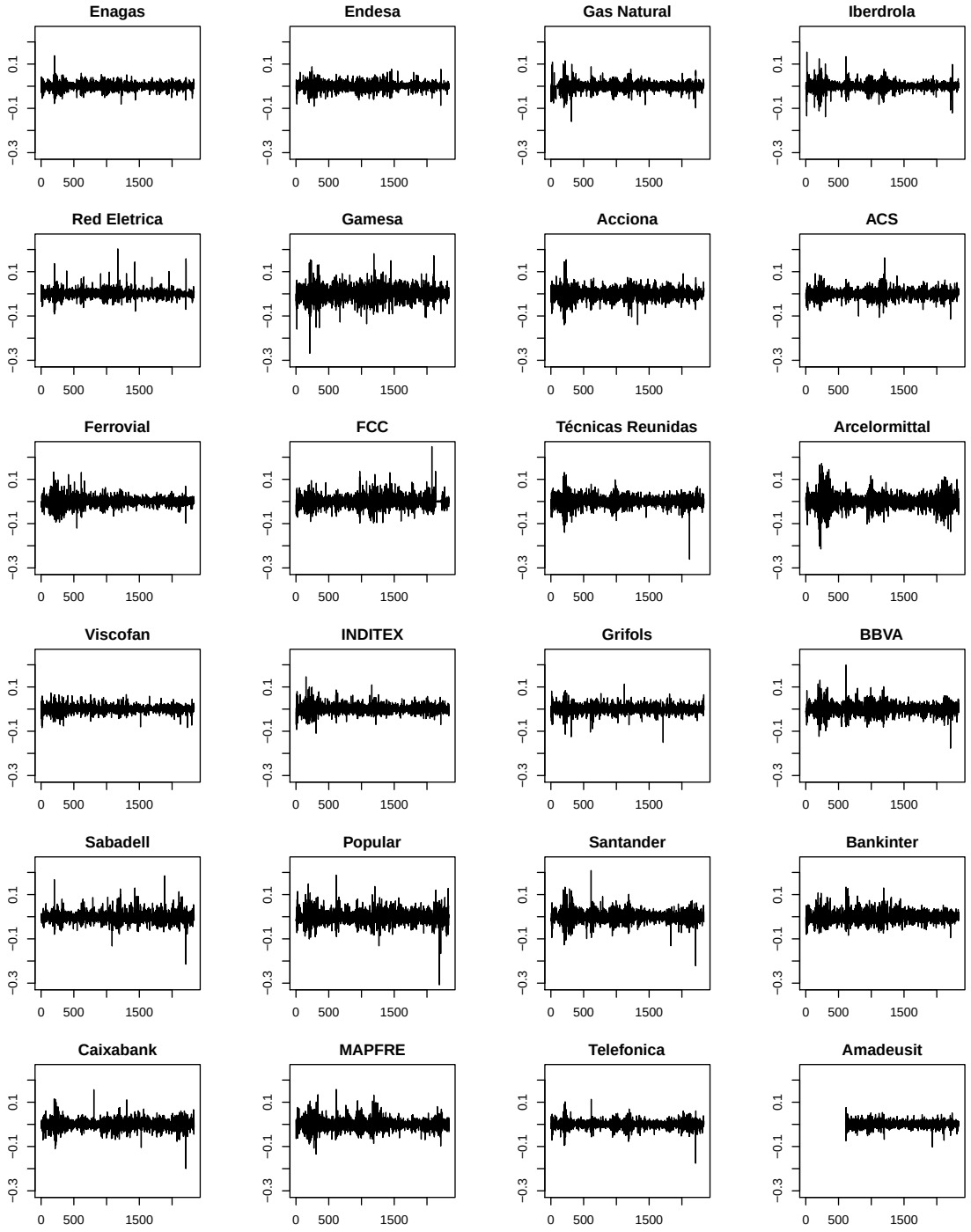

Figure 4.7: Daily returns of 24 stocks included in the IBEX-35. Sample period: 1st January 2008 to 19th December 2016

Heteroskedasticity is again observed although less pronounced for several stocks. The largest fuzzy silhouette widths are obtained with a partition in three clusters ($C = 3$)

but with different strength for each model. While QAF-FCMdC drawn out a value FSW = 0.636, the models AR-FCMdC and GARCH-FCMdC led to lower FSW indexes of 0.424 and 0.293, respectively. Therefore, all the models suggest the existence of three clusters but QAF-FCMdC indicates in a more conclusive way that a well-defined cluster structure lies behind data. It is worthy noting that the best FSW is reached using $m = 1.5$ for AR-FCMdC, while $m = 2$ is the fuzziness level producing the highest FSW for the GARCH- and QAF-based models. As in this application we lack of an intuitive idea on the underlying partition, we decided to corroborate our result by using an alternative index. Specifically, we calculated the Xie-Beni index (Xie and Beni, 1991) which is given by the ratio between the total variance and the minimum separation between clusters so that the optimal C is reached when this ratio is minimized. The minimum values of the Xie-Beni index corresponded to $C = 3$, with values 0.4804, 0.5537 and 0.6890 for QAF-FCMdC, AR-FCMdC and GARCH-FCMdC, respectively, again concluding that a 3-cluster solution seems the most adequate and that QAF-FCMdC produces the best-defined partition. Based on these arguments, cluster analysis using the three fuzzy models and setting $C = 3$ was carried out. The resulting membership degrees are shown in Table 4.7. As in previous application, the shaded cells enhance the highest membership degrees with each procedure and the stocks allocated in a fuzzy way between two or three clusters are indicated with memberships in bold font.

The 3-cluster solution generated by the QAF-FCMdC model identifies a large cluster, $\mathcal{C}_1$, gathering together most of the stocks, including the ones of the sectors of Energy and Materials, Industry and Construction (except for Arcelormittal-MTE), and also the three banks with the highest capitalization level in the Financial services sector, namely BBVA, Santander-SAN and Caixabank-CABK. The cluster $\mathcal{C}_3$ groups the company Arcelormittal-MTE together with the smaller banks Banco Popular-POP, Banco Sabadell-SAB and Bankinter-BKT, although SAB and BKT could be allocated in $\mathcal{C}_1$ by exhibiting similar memberships for both clusters. The cluster $\mathcal{C}_2$ groups together two important companies of the consumer goods industry (Viscofan-VIS and Inditex-ITX), the only insurance company (Mapfre-MAP), and a technological company related to the travel sector (Amadeus-AMS). In sum, the fuzzy partition provided by the QAF-FCMdC model seems to be congruent with features like company size and business area. Nevertheless, our concern is not to obtain conclusions in financial terms such as searching proper model specifications or accurate predictions for the daily return series. These targets go beyond the scope of this work. Our motivation is to illustrate the capability of the proposed fuzzy clustering approach to identify similar dependence structures. In this sense, a relevant point by treating with daily returns is to analyze their dispersion, i.e. the underlying volatility patterns. To bring

Table 4.7: Membership degrees in clustering of the daily returns of 24 stocks included in the IBEX-35 ($C = 3$, $m = 1.5$ for AR–FCMdC and $m = 2$ for GARCH–FCMdC and QAF–FCMdC).

		AR–FCMdC			GARCH–FCMdC			QAF–FCMdC		
		$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$
Sector: Energy										
Enagás	ENG	0.142	0.002	0.856	**0.482**	**0.447**	0.071	0.558	0.275	0.167
Endesa	ELE	0.996	0.000	0.004	0.038	**0.546**	**0.416**	0.684	0.072	0.244
Gas Natural	GAS	1.000	0.000	0.000	0.031	0.943	0.026	0.946	0.013	0.041
Iberdrola	IBE	0.193	0.023	0.784	0.997	0.002	0.001	**0.579**	0.052	**0.369**
Red Eléctrica	REE	0.869	0.029	0.102	0.532	0.278	0.190	0.814	0.054	0.132
Sector: Materials, Industry and Construction										
Gamesa	GAM	0.616	0.020	0.364	0.001	0.003	0.996	0.703	0.079	0.218
Acciona	ANA	0.930	0.001	0.069	0.000	0.001	0.999	1.000	0.000	0.000
ACS	ACS	1.000	0.000	0.000	1.000	0.000	0.000	0.816	0.020	0.164
Ferrovial	FER	0.269	0.003	0.728	0.031	0.324	0.645	0.754	0.095	0.151
FCC	FCC	1.000	0.000	0.000	0.821	0.145	0.034	0.821	0.025	0.154
Técnicas Reunidas	TRE	**0.340**	**0.230**	**0.430**	0.014	0.105	0.881	0.710	0.147	0.143
Arcelormittal	MTS	0.000	1.000	0.000	0.001	0.003	0.996	0.084	0.020	0.896
Sector: Consumer goods										
Viscofan	VIS	0.204	0.049	0.747	0.027	0.095	0.878	0.087	0.849	0.064
Inditex	ITX	0.239	0.052	0.709	0.083	0.210	0.707	0.000	1.000	0.000
Grifols	GRF	**0.429**	0.022	**0.549**	0.008	0.054	0.938	**0.539**	0.070	**0.391**
Sector: Financial services										
BBVA	BBVA	0.876	0.009	0.115	0.063	0.895	0.042	0.860	0.017	0.123
Santander	SAN	0.863	0.002	0.135	0.000	1.000	0.000	0.801	0.076	0.123
Caixabank	CABK	0.052	0.002	0.946	0.000	0.000	1.000	0.860	0.037	0.103
Banco Sabadell	SAB	0.963	0.007	0.030	0.007	0.030	0.963	**0.316**	0.168	**0.516**
Banco Popular	POP	0.999	0.000	0.001	0.001	0.997	0.002	0.000	0.000	1.000
Bankinter	BKT	0.829	0.004	0.167	0.024	0.867	0.109	**0.501**	0.040	**0.459**
Sector: Insurance										
Mapfre	MAP	0.096	0.014	0.890	0.002	0.013	0.985	0.118	0.821	0.061
Sector: Technology and Telecommunications										
Telefónica	TEF	0.325	0.054	0.621	0.970	0.023	0.007	0.955	0.006	0.039
Amadeus	AMS	0.000	0.000	1.000	0.038	**0.597**	**0.365**	0.185	0.708	0.107

insight into this issue, nonparametric approximations of the variance between returns were obtained. The estimated volatility curves grouped according the three clusters identified with the QAF-FCMdC model are depicted in Figure 4.8.

Note that all the curves in Figure 4.8 (a) present a very similar fluctuation pattern, with some bumps of different size in similar periods of time. The curves in Figure 4.8 (b) corresponding to the cluster $\mathcal{C}_2$ are characterized by a flat profile throughout the second half of the sample period. In fact, only Mapfre-MAP showed a few periods with moderate rise in the level of volatility. The cluster $\mathcal{C}_3$ brings together stocks exhibiting a marked pickup in volatility in the last year, particularly Arcelormittal-MTE and Banco Popular-POP. This effect is less evident for Banco Sabadell-SAB, which could account for its vague allocation

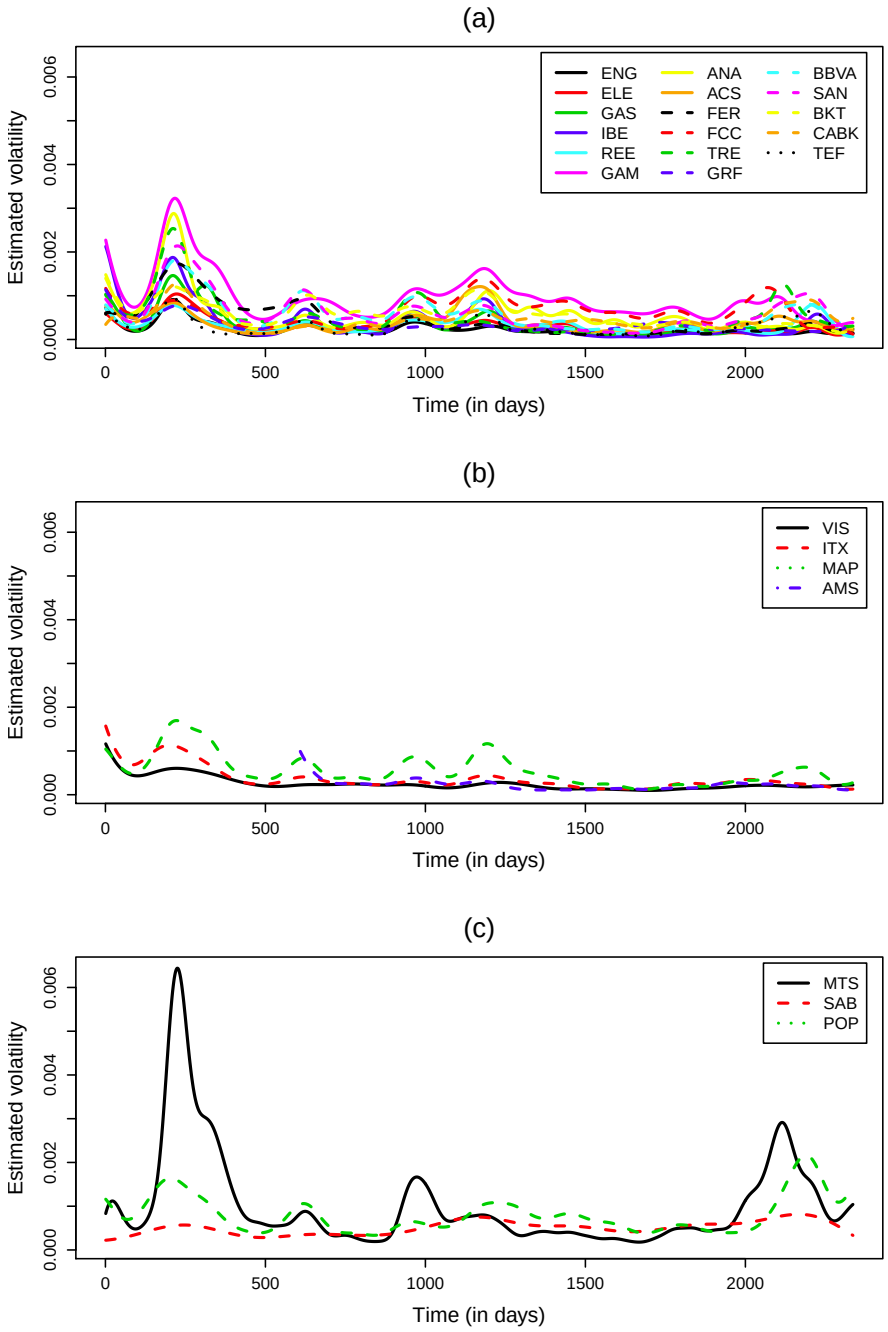

Figure 4.8: Nonparametric estimators of the volatility for the daily returns of the 24 analyzed stocks grouped according to the cluster solution provided by the QAF–FCMdC model: $\mathcal{C}_1$ (a), $\mathcal{C}_2$ (b) and $\mathcal{C}_3$ (c)

in this cluster (with a membership of 0.516). It is worthy noting that Arcelormittal-MTE presented a sharp rising of volatility during the first year and a half of the sample period, fairly above the rest of the analyzed stocks. This significant behavior might determine the

atypical character of this time series. Overall, Figure 4.8 allows us to describe representative volatility patterns for each of the clusters determined by the QAF-FCMdC model.

The AR-FCMdC model led to a reasonable cluster solution although a bit less intuitive than the one obtained with QAF-FCMdC. For instance, it is unexpected that the companies Enagás-ENG, Iberdrola-IBE and Ferrovial-FER are separated from the rest of companies belonging to the same market sectors. Also, according to the volatility profiles showed in Figures 4.8 (a) and (b) for the banks, it seems inappropriate to locate all of them together at the same cluster, particularly the Banco Popular-POP. These arguments support the better clustering quality indexes obtained by the QAF-FCMdC model. Lastly, the GARCH-FCMdC model produced a cluster partition hardly interpretable and substantially different from those obtained with the two other procedures. As in the case of the ozone records in the first application, the GARCH approximations have not been accurate enough to properly discriminate between the generating models.

4.5 Robust fuzzy clustering based on quantile autocovariances

Following the literature on fuzzy clustering, several techniques have been introduced to increase robustness of algorithms for clustering of object data (see for example Dave, 1991; Wu and Yang, 2002; D'Urso et al., 2013b, 2015b, 2017).

To tackle the problem of dealing with outliers, we propose three different types of robustification of the QAF-FCMdC model, namely:

- QAF-based Exponential Fuzzy C-Medoids Clustering model (QAF-FCMdC-Exp)
- QAF-based Fuzzy C-Medoids Clustering with Noise Cluster model (QAF-FCMdC-NC)
- QAF-based Trimmed Fuzzy C-Medoids Clustering model (QAF-TrFCMdC)

Each model face the presence of outliers in a different way. The QAF-FCMdC-Exp neutralizes the effect of the outliers by using a robust distance measure, the QAF-FCMdC-NC achieves its robustness by assigning potential outliers into an artificial cluster (the so-called noise cluster), and the QAF-TrFCMdC model is aimed at identifying a certain fraction of the furthest time series and trimmed them away from the classification process.

Just as in Section 4.2, let $S = \{\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(n)}\}$ be a set of n observed time series subjected to clustering. Denote by $\boldsymbol{\Gamma}^{(i)}$ the vector of quantile autocovariances estimated from the i-th

observed series, for $1 \leq i \leq n$. The distance between $\boldsymbol{X}^{(i)}$ and $\boldsymbol{X}^{(j)}$ is characterized in terms of the distance between $\boldsymbol{\Gamma}^{(i)}$ and $\boldsymbol{\Gamma}^{(j)}$, for $i, j \in \{1, \ldots, n\}$.

4.5.1 QAF-based Exponential Fuzzy C-Medoids Clustering model

Wu and Yang (2002) observed that the solution of the objective function in fuzzy clustering based on the Euclidean metric can be written as a weighted sum of the observed data with weights all equal to 1 regardless of the level of variability in the data set. This way, the result could be strongly influenced by the presence of outliers. To overcome this limitation, they propose to use a more robust metric, the so-called "exponential distance". This metric is aimed at giving different weights to each data object depending on whether it is or not considered as outlier. In essence, a small weight is assigned to outliers, while a large weight is given to data objects laying close to the bulk of the data set. This way a more robust metric is obtained. Note that this idea also applies in a time series context when the considered distance is defined by the Euclidean distance between estimated feature vectors. In particular, the exponential distance based on the estimated quantile autocovariances is defined by

$$d_{i,i'} = \left[1 - \exp\left\{-\beta \left\|\boldsymbol{\Gamma}^{(i)} - \boldsymbol{\Gamma}^{(i')}\right\|^2\right\}\right]^{\frac{1}{2}}, \tag{4.9}$$

where β is a positive constant.

To obtain a suitable value of the parameter β, Wu and Yang (2002) suggest to take the inverse of the variability in the data set. This way, a large value of β is obtained in presence of low dispersion, which means a lower weight for potential outliers (i.e. distant objects) than in the case of high dispersion.

Considering the exponential distance defined in (4.9), the QAF-based Fuzzy C-Medoids Clustering with Exponential Distance (QAF-FCMdC-Exp) is based on the new objective function defined by

$$\begin{cases} \min_{\tilde{\Gamma},\Omega} \sum_{i=1}^{n} \sum_{c=1}^{C} u_{ic}^{m} \left[1 - \exp\left\{-\beta \left\|\boldsymbol{\Gamma}^{(i)} - \widetilde{\boldsymbol{\Gamma}}^{(c)}\right\|^2\right\}\right] \\ \text{subject to: } \sum_{c=1}^{C} u_{ic} = 1 \text{ and } u_{ic} \geq 0, \end{cases} \tag{4.10}$$

where $m > 1$ is a weighting exponent that controls the fuzziness of the obtained partition and u_{ic} indicates the membership degree of the i-th unit in the c-th cluster.

In this case, following the results in Wu and Yang (2002), the iterative solutions for the

membership degrees are given by:

$$u_{ic} = \left[\sum_{c'=1}^{C} \left(\frac{1 - \exp\left\{ -\beta \left\| \mathbf{\Gamma}^{(i)} - \widetilde{\mathbf{\Gamma}}^{(c)} \right\|^2 \right\}}{1 - \exp\left\{ -\beta \left\| \mathbf{\Gamma}^{(i)} - \widetilde{\mathbf{\Gamma}}^{(c')} \right\|^2 \right\}} \right)^{\frac{1}{m-1}} \right]^{-1}, \quad \text{for } i = 1, \ldots, n \text{ and } c = 1, \ldots, C. \tag{4.11}$$

As for the selection of the β parameter, the idea of adopting the inverse of the variability is taken into account so that β is calculated by

$$\beta = \left[\frac{1}{n} \sum_{i=1}^{n} \left\| \mathbf{\Gamma}^{(i)} - \widetilde{\mathbf{\Gamma}}_2^{(c)} \right\|^2 \right]^{-1}, \tag{4.12}$$

where $\widetilde{\mathbf{\Gamma}}^{(c)}$ corresponds to the index c satisfying $c = \underset{1 \leq i' \leq n}{\operatorname{argmin}} \sum_{i''=1}^{n} \left\| \mathbf{\Gamma}^{(i'')} - \mathbf{\Gamma}^{(i')} \right\|^2$.

Based on the membership degrees obtained from (4.11), the C series minimizing (4.10) are selected as new medoids. This two-step procedure is iterated until there is no change in the medoids or a maximum number of iterations is achieved.

Just as in the standard QAF-FCMdC algorithm, the initial set of medoids is selected using a hard PAM algorithm based on the QAF dissimilarity. This criterion is applied to all robust fuzzy algorithms considered in this chapter. The QAF-based Exponential Fuzzy C-Medoids Clustering model (QAF-FCMdC-Exp) is implemented as outlined in Algorithm 2.

Algorithm 2 The QAF-based Exponential Fuzzy C-Medoids Clustering model (QAF-FCMdC-Exp)

Fix C, m and *max.iter*
Compute β using 4.12
Set $iter = 0$
Pick the initial medoids $\widetilde{\mathbf{\Gamma}} = \left\{ \widetilde{\mathbf{\Gamma}}^{(1)}, \ldots, \widetilde{\mathbf{\Gamma}}^{(C)} \right\}$
repeat
    Set $\widetilde{\mathbf{\Gamma}}_{\text{OLD}} = \widetilde{\mathbf{\Gamma}}$ {Store the current medoids}
    Compute u_{ic}, $i = 1, \ldots, n$, $c = 1, \ldots, C$, using (4.11)
    For each $c \in \{1, \ldots, C\}$, determine the index $j_c \in \{1, \ldots, n\}$ satisfying:
    $j_c = \underset{1 \leq j \leq n}{\operatorname{argmin}} \sum_{i=1}^{n} u_{ic}^m \left[1 - \exp\left\{ -\beta \left\| \mathbf{\Gamma}^{(i)} - \mathbf{\Gamma}^{(j)} \right\|^2 \right\} \right]$
    return $\widetilde{\mathbf{\Gamma}}^{(c)} = \mathbf{\Gamma}^{(j_c)}$, for $c = 1, \ldots, C$ {Update the medoids}
    $iter \leftarrow iter + 1$
until $\widetilde{\mathbf{\Gamma}}_{\text{OLD}} = \widetilde{\mathbf{\Gamma}}$ or $iter = max.iter$

In short, the QAF-FCMdC-Exp model does not discriminate the time series outliers. Its aim is to smooth the effect of these anomalous series by adjusting their influence with proper weights. The result is that the memberships of the anomalous are similarly distributed across the clusters but the true clustering structure is not seriously affected by their presence.

4.5.2 QAF-based Fuzzy C-Medoids Clustering with Noise Cluster model

The proposed noise fuzzy clustering algorithm is a version of the one proposed by Dave (1991) and Dave and Sen (1997) to overcome sensitivity of the classical Fuzzy C- Medoids algorithms in the presence of noisy data, based on an idea introduced by Ohashi (1984).

The key idea is to neutralize the negative effect of the outliers by classifying them in an artificial cluster (the noise cluster). The noise cluster is characterized by a virtual prototype that has a constant and sufficiently large distance ("noise distance") from all the remaining series. This way, the anomalous series are separated and located into the noise cluster and the true cluster structure is not altered in the classification process. A single time series belongs to a real cluster if its distance from a medoid is lower than the noise distance, let us say δ, otherwise, the time series belongs to the noise cluster.

The QAF-based Fuzzy C-Medoids Clustering with Noise Cluster (QAF-FCMdC-NC) can be formalized as follows:

$$\begin{cases} \min_{\tilde{\Gamma},\Omega} \sum_{i=1}^{n} \sum_{c=1}^{C-1} u_{ic}^{m} \left\| \mathbf{\Gamma}^{(i)} - \widetilde{\mathbf{\Gamma}}^{(c)} \right\|^2 + \sum_{i=1}^{n} \delta^2 \left(1 - \sum_{c=1}^{C-1} u_{ic} \right)^m \\ \text{subject to: } \sum_{c=1}^{C} u_{ic} = 1 \text{ and } u_{ic} \geq 0. \end{cases} \tag{4.13}$$

where u_{ic} are the membership degrees, $m > 1$ is the fuzziness parameter and δ is the noise distance, to be set in advance. Obviously, in the prior formulation, the C-th cluster is the noise cluster and the first $C-1$ clusters identify the real underlying clusters.

The iterative solutions for the membership degrees are given by:

$$u_{ic} = \left[\sum_{c'=1}^{C} \left(\frac{\left\| \mathbf{\Gamma}^{(i)} - \widetilde{\mathbf{\Gamma}}^{(c)} \right\|^2}{\left\| \mathbf{\Gamma}^{(i)} - \widetilde{\mathbf{\Gamma}}^{(c')} \right\|^2} \right)^{\frac{1}{m-1}} + \left(\frac{\left\| \mathbf{\Gamma}^{(i)} - \widetilde{\mathbf{\Gamma}}^{(c)} \right\|^2}{\delta^2} \right)^{\frac{1}{m-1}} \right]^{-1}, \tag{4.14}$$

for $i = 1, \ldots, n$ and $c = 1, \ldots, C$.

Indeed, an important point is to select a suitable value for the noise distance δ. If δ is too small, a large portion of the data set will receive a high degree of membership to the noise cluster, identifying as atypical series that actually are not. Otherwise, if the value of δ is too large, then none of the time series is going to be placed into the noise cluster. Therefore, a suitable threshold is required and in fact the procedure is very sensitive to this choice. Unfortunately, the proper selection of this parameter is still an open problem in the literature. Following the recommendation by D'Urso et al. (2013b), an expression on δ can be obtained by doing

$$\delta^2 = \lambda \left[\frac{1}{n\,(C-1)} \sum_{i=1}^{n} \sum_{c=1}^{C-1} \left\| \mathbf{\Gamma}^{(i)} - \widetilde{\mathbf{\Gamma}}^{(c)} \right\|^2 \right] \tag{4.15}$$

where λ is a scale multiplier to be selected depending on the nature of data.

For the selection of the value λ, Cimino et al. (2005) suggest to proceed as follows. First, the fuzzy clustering is run with decreasing values of λ on a prefixed grid. The percentage of series located into the noise cluster for each value of λ is recorded. By the definition of the noise distance in (4.15), these percentages increase when λ decreases. Then, the value selected for λ is the one where an abrupt change of slope (elbow) is observed.

Algorithm 3 QAF-based Fuzzy C-Medoids Clustering with Noise Cluster model (QAF-FCMdC-NC)

1: Fix $C-1$, m and $max.iter$
2: Set $iter = 0$
3: Pick the initial medoids $\widetilde{\mathbf{\Gamma}} = \left\{ \widetilde{\mathbf{\Gamma}}^{(1)}, \ldots, \widetilde{\mathbf{\Gamma}}^{(C-1)} \right\}$
4: **repeat**
5: Set $\widetilde{\mathbf{\Gamma}}_{\text{OLD}} = \widetilde{\mathbf{\Gamma}}$ {Store the current medoids}
6: Compute δ using 4.15.
7: Compute u_{ic}, $i = 1, \ldots, n$, $c = 1, \ldots, C$, using (4.14)
8: For each $c \in \{1, \ldots, C-1\}$, determine the index $j_c \in \{1, \ldots n\}$ satisfying:
$$j_c = \underset{1 \le j \le n}{\operatorname{argmin}} \sum_{i=1}^{n} u_{ic}^m \left\| \mathbf{\Gamma}^{(i)} - \widetilde{\mathbf{\Gamma}}^{(j)} \right\|^2$$
9: **return** $\widetilde{\mathbf{\Gamma}}^{(c)} = \mathbf{\Gamma}^{(j_c)}$, for $c = 1, \ldots, C-1$ {Update the medoids}
10: $iter \leftarrow iter + 1$
11: **until** $\widetilde{\mathbf{\Gamma}}_{\text{OLD}} = \widetilde{\mathbf{\Gamma}}$ or $iter = max.iter$

As usual, the algorithm works in an iterative approach. Based on the membership degrees obtained from (4.14), the C series minimizing (4.13) are selected as new medoids. This two-step procedure is iterated until there is no change in the medoids or a maximum number of iterations is achieved.

The QAF-based Fuzzy C-Medoids Clustering with Noise Cluster model (QAF-FCMdC-NC)

is implemented as outlined in Algorithm 3.

4.5.3 QAF-based Trimmed Fuzzy C-Medoids Clustering model

The last robust version of the QAF-based Fuzzy C-Medoids Clustering model is introduced in this section, namely the QAF-based Trimmed Fuzzy C-Medoids Clustering model (QAF-TrFCMdC). In this case the model achieves its robustness by trimming away a proportion of time series that are more distant from the medoids representing the cluster partition.

Given a trimming size α, which ranges between 0 and 1, the QAF-TrFCMdC can be formalized as the following minimization problem:

$$\begin{cases} \min_{Y,\Omega} \sum_{i=1}^{H(\alpha)} \sum_{c=1}^{C} u_{ic}^{m} \left\| \mathbf{\Gamma}^{(i)} - \widetilde{\mathbf{\Gamma}}^{(c)} \right\|^2 \\ \text{subject to: } \sum_{c=1}^{C} u_{ic} = 1 \text{ and } u_{ic} \geq 0. \end{cases} \tag{4.16}$$

where u_{ic} is the membership degree of the i-th time series to the c-th cluster, $m > 1$ is the fuzziness parameter and Y ranges on all the subsets of the set of the p sequences of estimated quantile autocovariances of size $H(\alpha) = [p(1-\alpha)]$. Notice that if $\alpha = 0$, then none of the series is trimmed away from the process and the standard non-robust QAF-FCMdC version of the procedure is obtained. All non-trimmed time series are classified according to the QAF-FCMdC model.

Just as in the QAF-FCMdC model, the local optimal solution for the estimation of the membership degrees is:

$$u_{ic} = \left[\sum_{c'=1}^{C} \left(\frac{\left\| \mathbf{\Gamma}^{(i)} - \widetilde{\mathbf{\Gamma}}^{(c)} \right\|^2}{\left\| \mathbf{\Gamma}^{(i)} - \widetilde{\mathbf{\Gamma}}^{(c')} \right\|^2} \right)^{\frac{1}{m-1}} \right]^{-1} \tag{4.17}$$

where i ranges in the subset of the non-trimmed series and $c = 1, \ldots, C$.

To determine the trimming ratio α, i.e. the fraction of time series to be trimmed, the following approach is considered. By replacing the expression of the u_{ic} (4.2) into (4.1), we obtain:

$$\sum_{i=1}^{p} \left[\sum_{c=1}^{C} \left(\left\| \mathbf{\Gamma}^{(i)} - \widetilde{\mathbf{\Gamma}}^{(c)} \right\|^2 \right)^{1/(1-m)} \right]^{1-m} = \sum_{i=1}^{p} h_i \tag{4.18}$$

where

$$h_i = \left[\sum_{c=1}^{C} \left(\left\| \mathbf{\Gamma}^{(i)} - \widetilde{\mathbf{\Gamma}}^{(c)} \right\|_2^2 \right)^{1/(1-m)} \right]^{1-m} . \tag{4.19}$$

The value h_i provides the distance from each series to all the medoids. Therefore, based on these values h_i, it is feasible to identify the subset Y by selecting the $H(\alpha)$ time series closest to the medoids. The value of $H(\alpha) < p$ is chosen depending on how many time series we would like to eliminate in the clustering process.

An unsuitable selection of the trimming ratio will result in an overestimation of the number of outliers. In practice, the choice of α is carried out by minimizing a quality clustering index over a grid of possible values. The Xie-Beni index (Xie and Beni, 1991) or the Kwon index (Kwon, 1998) are frequently used.

Based on the membership degrees obtained from (4.17), the C series minimizing (4.16) are selected as new medoids. A new two-step procedure is iterated until there is no change in the medoids or a maximum number of iterations is achieved.

The QAF-based Trimmed Fuzzy C-Medoids Clustering model (QAF-TrFCMdC) is implemented as outlined in Algorithm 4.

Algorithm 4 QAF–based Trimmed Fuzzy C-Medoids Clustering model (QAF–TrFCMdC)

Fix C, m, α and $max.iter$
Set $iter = 0$
Pick the initial medoids $\widetilde{\mathbf{\Gamma}} = \left\{ \widetilde{\mathbf{\Gamma}}^{(1)}, \ldots, \widetilde{\mathbf{\Gamma}}^{(C)} \right\}$
repeat
    Identify the subset Y made of the $H(\alpha) = [p(1-\alpha)]$ series closest to the medoids
    Set $\widetilde{\mathbf{\Gamma}}_{\text{OLD}} = \widetilde{\mathbf{\Gamma}}$ {Store the current medoids}
    Compute u_{ic}, $i = 1, \ldots, n$, $c = 1, \ldots, C$, using (4.17)
    For each $c \in \{1, \ldots, C\}$, determine the index $j_c \in \{1, \ldots, n\}$ satisfying:
    $$j_c = \underset{1 \le j \le p}{\operatorname{argmin}} \sum_{i=1}^{H(\alpha)} u_{ic}^m \left\| \mathbf{\Gamma}^{(i)} - \mathbf{\Gamma}^{(j)} \right\|^2$$
    return $\widetilde{\mathbf{\Gamma}}^{(c)} = \mathbf{\Gamma}^{(j_c)}$, for $c = 1, \ldots, C$ {Update the medoids}
    $iter \leftarrow iter + 1$
until $\widetilde{\mathbf{\Gamma}}_{\text{OLD}} = \widetilde{\mathbf{\Gamma}}$ or $iter = max.iter$

4.6 Assessing the behavior of the robust versions of the QAF–FCMdC model: A simulation study

This section reports some results from a broad simulation study conducted to evaluate the clustering performance and accuracy of the proposed methods compared with standard procedures and other robust models based on different metrics. To gain insight into robustness to the generating models, simulation scenarios considering different time series setups were recreated, namely scenarios involving linear, non-linear and conditionally heteroskedastic models. At each of these setups, we start with a base scenario formed by two well-separated clusters $\mathcal{C}_1$ and $\mathcal{C}_2$ including four time series each, and then the base scenario is successively contaminated with the presence of one and two outlier time series ($\mathcal{O}_1$ and $\mathcal{O}_2$). The specific scenarios and the generation schemes for each scenario are described below.

Clustering of linear models

L.1 Four time series simulated from each of the AR(1) model $X_t = 0.5X_{t-1} + \varepsilon_t$ (cluster $\mathcal{C}_1$) and the MA(1) model $X_t = \varepsilon_t - 0.5\varepsilon_{t-1}$ (cluster $\mathcal{C}_2$).

L.2 The base scenario L.1 plus one outlier time series $\mathcal{O}_1$ simulated from a Gaussian white noise process.

L.3 The scenario L.2 and an additional outlier time series $\mathcal{O}_2$ simulated from the ARMA(1,1) model $X_t = -0.9X_{t-1} + \varepsilon_t + 0.3\varepsilon_{t-1}$.

Clustering of non–linear models

NL.1 Four time series simulated from an exponential autoregressive model of the form

$$X_t = \left(0.3 - 10\exp(-X_{t-1}^2)\right) X_{t-1} + \varepsilon_t \text{ (cluster } \mathcal{C}_1),$$

and four time series simulated from the bilinear model given by

$$X_t = 0.6X_{t-1} - 0.8X_{t-2} + \varepsilon_t + 0.5\varepsilon_{t-1} + 0.8\varepsilon_{t-1}X_{t-1} \text{ (cluster } \mathcal{C}_2).$$

NL.2 The base scenario NL.1 plus one outlier time series $\mathcal{O}_1$, which consisted of one realization from the non–linear autoregressive model given by

$$X_t = 0.3|X_{t-1}|(3 + |X_{t-1}|)^{-1} + \varepsilon_t.$$

NL.3 The scenario NL.2 plus an additional outlier time series $\mathcal{O}_2$ generated from the non–linear moving average model given by

$$X_t = -0.1\varepsilon_{t-1} + 0.3\varepsilon_{t-1}^2 + \varepsilon_t.$$

Clustering of conditional heteroskedastic models

CH.1 A base scenario formed by two clusters with four time series each generated from ARCH processes $X_t = \sigma_t\varepsilon_t$, where $\sigma_t^2 = 0.1 + \phi X_{t-1}^2$ with $\phi = 0.05$ for cluster $\mathcal{C}_1$ and $\phi = 0.95$ for cluster $\mathcal{C}_2$.

CH.2 The base scenario CH.1 plus one outlier time series $\mathcal{O}_1$ simulated from an exponential GARCH model where the conditional variance is modeled by

$$\ln(\sigma_t^2) = 0.1 + 0.3\varepsilon_{t-1} + 0.7\left[|\varepsilon_{t-1}| - \mathbb{E}(|\varepsilon_{t-1}|)\right].$$

CH.3 The scenario CH.2 plus a second outlier time series $\mathcal{O}_2$ simulated from a GJR–GARCH model of the form

$$\sigma_t^2 = 0.1 + \left[0.1 + 0.6\, I\left(X_{t-1} < 0\right)\right] X_{t-1}^2 + 0.1\sigma_{t-1}^2.$$

In all cases, the error process ε_t consisted of iid variables following a zero-mean Gaussian distribution with unit variance. To bring insight into the shapes of the true quantile autocovariance functions for the examined models, plots of large sample approximations to these functions were obtained. Specifically, one hundred series of size 1000 were generated from each model and the corresponding sample quantile autocovariances averaged over the 100 replicates. For each $\tau \in \{0.1, 0.5, 0.9\}$, plots of the points $\{\hat{\gamma}\left(\tau, 0.05\, i\right),\ i = 1, \ldots, 19\}$ joined by lines are shown in Figure 4.9.

Plots in Figure 4.9 illustrate the capability of the quantile autocovariances to discriminate between the underlying processes. For the linear and non-linear scenarios (Figures 4.9(a) and (b), respectively), the theoretical patterns characterizing clusters and outliers exhibit very different curves of quantile autocovariances. As far as the heteroskedastic scenario (Figure 4.9(c)), discrimination between clusters and outliers is also evident if a joint assessment of the plots over the three quantiles is carried out.

Another graphical way to visualize both the spatial structure of the generating models and the separability between groups is to perform a multidimensional scaling (MDS) based on the pairwise QAF-dissimilarity matrix. For each scenario, 50 and 20 time series were generated from each of the models defining the clusters and the outlier processes, respectively.

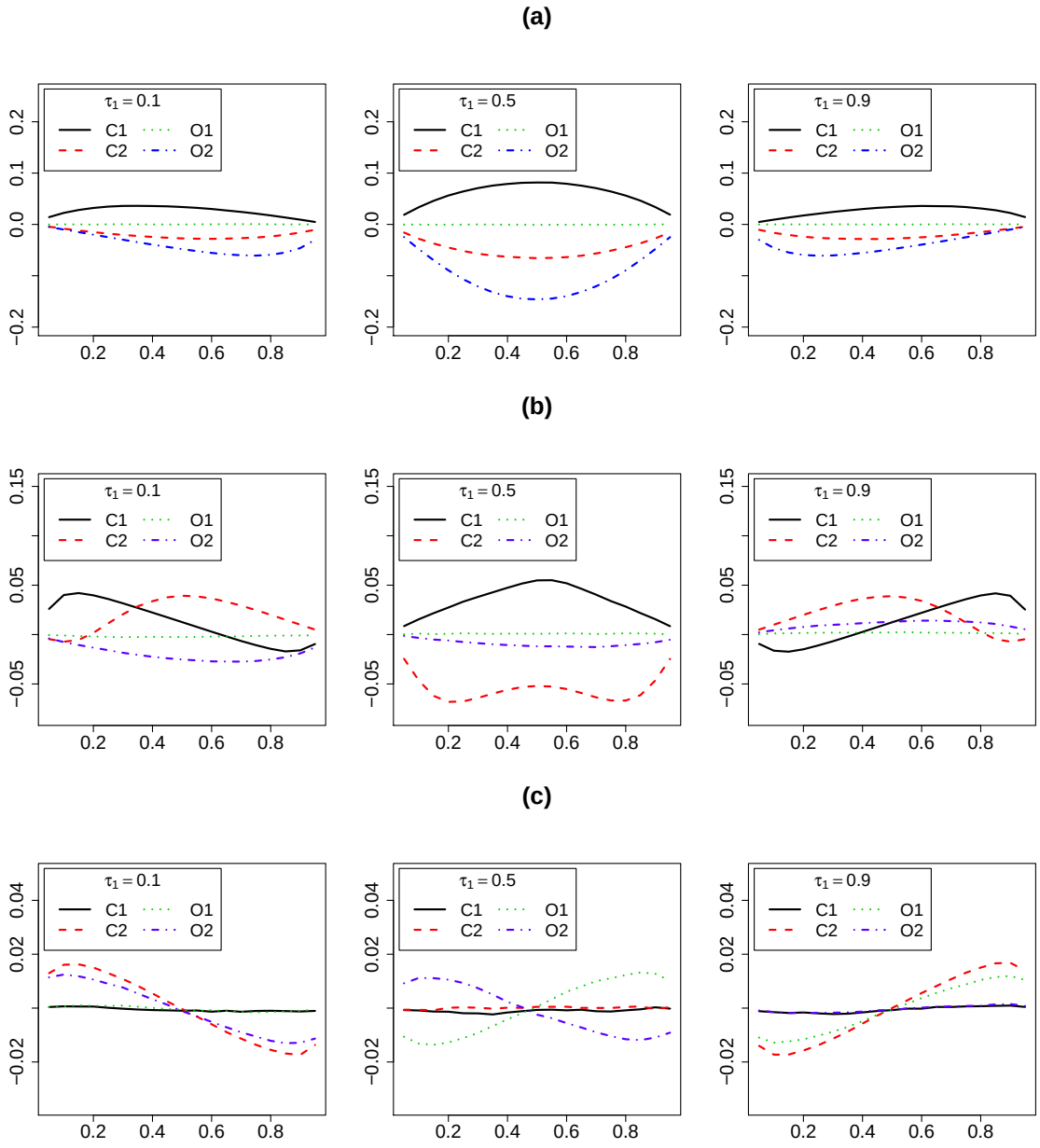

Figure 4.9: Large sample approximation of the quantile autocovariances for the models in the linear (a), non–linear (b) and heteroskedastic (c) scenarios.

The reason to generate 40 outliers was simply to examine the variability of these realizations. Then, a two-dimensional scaling based on these realizations was carried out and the corresponding coordinate matrices are displayed in Figure 4.10. Note that two different lengths of series were considered, namely $T = 150$ and $T = 250$ for the linear and non–linear models, and $T = 1500$ and $T = 2500$ for the case of conditionally heteroskedastic

series. As already mentioned, larger realizations are necessary with heteroskedastic models in order to estimate the quantile autocovariances with higher accuracy. Indeed, this limitation also affects other metrics considered in this setup. For instance, estimates for the ARCH/GARCH coefficients are required to measure discrepancy between fitted models, and poor clustering results are obtained if small sample sizes are used.

The spatial configurations of the MDS coordinates in Figure 4.10 show that the series forming the clusters $\mathcal{C}_1$ (red) and $\mathcal{C}_2$ (black) are grouped into two compact and well-separated clusters, while the outlier time series $\mathcal{O}_1$ (green) and $\mathcal{O}_2$ (blue) tend to be placed at an intermediate location between the clusters, except for the linear scenario where the second outlier, $\mathcal{O}_2$, is situated closer to cluster $\mathcal{C}_2$. Note that the non-linear models selected to generate outlier realizations produce overlapping clusters, while the linear and heteroskedastic models lead to separated groups, although also reasonably equidistant from $\mathcal{C}_1$ and $\mathcal{C}_2$ in the heteroskedastic scenarios. In short, Figure 4.10 reveals that the QAF metric should provide a useful approach to discriminate between the considered clusters and to detect the outlier time series. As expected, by increasing the length of the time series the gap between groups is more pronounced and, therefore, it will be easier to discriminate between them.

To assess the effectiveness of the proposed approaches in presence of outliers, each simulated dataset was subjected to clustering using the QAF-based fuzzy C-medoids clustering model (QAF-FCMdC) and the robust versions QAF-FCMdC-Exp, QAF-FCMdC-NC and QAF-TrFCMdC. The performance of the QAF metric was also examined by comparison with fuzzy C-medoids algorithms using other distances between fitted models. For the scenarios including ARMA models, we consider the AR distance introduced by Piccolo (Piccolo, 1990), which computes the Euclidean distance between estimated coefficients of truncated AR(∞) representations. This way, the fuzzy C-medoids clustering model based on the AR metric, AR-FCMdC (D'Urso et al., 2013c), and the corresponding robust versions AR-FCMdC-Exp (D'Urso et al., 2015b), AR-FCMdC-NC (D'Urso et al., 2013b) and AR-TrFCMdC (D'Urso et al., 2017) were carried out in Scenarios L.1, L.2 and L.3. Analogously, a metric based on the autoregressive representations of GARCH(p,q) processes was employed with the heteroskedastic models. More precisely, a GARCH(p,q) process satisfies $X_t = \sigma_t \varepsilon_t$, where the innovations ε_t are iid variables and the squared conditional variance σ_t^2 follows an ARMA(p,q) model with parameters $(\delta, \alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q)$. It can be shown that

$$X_t^2 = \delta + \sum_{i=1}^{p^\star} (\alpha_i + \beta_i) X_{t-i}^2 + \sum_{j=1}^{q} \beta_j \eta_{t-j} + \eta_t, \tag{4.20}$$

with $p^\star = \max(p, q)$, $\alpha_i = 0$ for $i > p$, $\beta_i = 0$ for $i > q$, and $\eta_t = X_t^2 - \sigma_t^2$ a zero-mean error uncorrelated with the past. Equation (4.20) establishes an ARMA($p^\star$,q) representation for

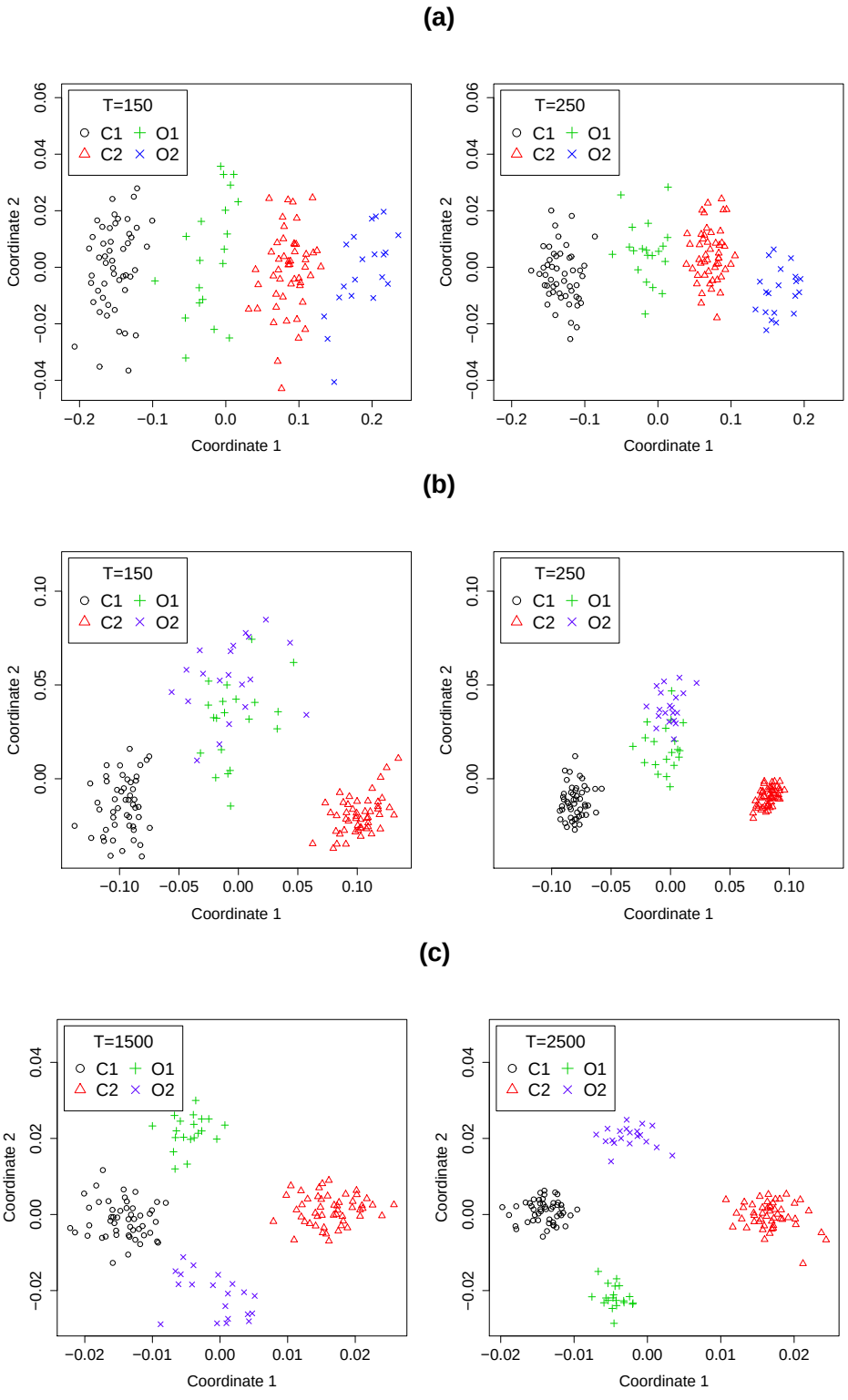

Figure 4.10: Two-dimensional scaling configurations based on the QAF distance from the simulated linear (a), non–linear (b) and heteroskedastic (c) models.

X_t^2, which can be approximated by an AR(∞) structure with autoregressive coefficients π_u^G given by

$$\pi_u^G = (\alpha_u + \beta_u) + \sum_{j=1}^{\min(q,u)} \beta_j \pi_{u-j}^G$$

where $\pi_0^G = -1$, $\alpha_u = 0$ for $u > p$, and $\beta_u = 0$ for $u > q$. At this point, the GARCH distance is defined by the Euclidean distance between estimators of these new autoregressive coefficients. Based on the GARCH distance, the counterpart fuzzy algorithms GARCH-FCMdC (D'Urso et al., 2013a), GARCH-FCMdC-Exp, GARCH-FCMdC-NC and GARCH-TrFCMdC were carried out in the Scenarios CH.1, CH.2 and CH.3. In sum, we examined the performance of competitors using tailor-made distances for the lineal and heteroskedastic scenarios. Unlike the QAF-based models, it is expected that these model-based approaches get worse in case of model misspecification. However, their use in proper scenarios provide us valuable insight into the robustness of the QAF distance against the generating processes.

According to our clustering aim, the performance and accuracy of each algorithm is evaluated in terms of the percentage of times in which the series generated from the same process are grouped together in the same cluster, with membership degrees close to one for that cluster. Robustness in presence of outliers is examined by analysing the effect of the anomalous series on the membership degrees in the final partition, and also by reporting the percentage of times that the outliers are identified when the noise cluster and the trimmed models are used.

The number of clusters was set at $C = 2$. For each of the nine scenarios, 10 sets of 100 simulations were carried out and subjected to fuzzy clustering with the described algorithms. For each of these 100 trials, the percentage of times that all the series were correctly classified was computed, and then the average percentage of correct classification over the 10 sets was taken as measure of clustering accuracy of the algorithm.

Due to deal with fuzzy models, it was necessary to specify cut-off values to decide when a specific realization is assigned to a particular cluster. In the baseline scenarios, with no anomalous series, the i-th time series is assigned to the c-th cluster if its fuzzy membership degree is $u_{ic} > 0.6$. In the scenarios with data contaminated with outliers, the anomalous series were identified following different criteria according to the employed model. By using the noise cluster models, an outlier is considered to be correctly classified when it is assigned to the noise cluster, i.e. if $u_{ic_{NC}} > 0.6$, with c_{NC} denoting the index of the noise cluster. By performing the standard fuzzy algorithms and the robust versions based on the exponential metric, we assume that the algorithm correctly handles the outliers when their membership degrees are reasonably similar for the two clusters, specifically both of them belonging to

the $(0.3, 0.7)$ interval. Lastly, in the case of the trimmed fuzzy, we checked whether the true outliers are trimmed units in the process. It is worthy remarking that these criteria and the selected cut-off values are compatible with the recommendations suggested in the literature ((see e.g. D'Urso et al., 2013b, 2015b).

In our experiments, three quantiles of levels 0.1, 0.5 and 0.9 and only one lag ($L = 1$, with $l_1 = 1$) were considered to compute the fuzzy algorithms based on the QAF dissimilarity. Certainly, increasing the number of quantiles does not mean an important cost in terms of computing time due to the computational efficiency of the QAF metric. Nevertheless, it was observed that three quantiles were enough to provide satisfactory results. To compute the AR and the GARCH distances, the order k of the truncated AR(∞) approximations was determined by the AIC.

In all scenarios, we perform the fuzzy clustering models for several values of the fuzziness parameter m, which has a great influence in the clustering results. While small values of m, close to one, result in partitions with a low level of fuzziness that is with membership degrees close to 1 and 0, large values of m increase the amount of overlapping and the membership degrees are more homogeneously spread across the clusters. Using $m = 1.5$ or $m = 2$ are two popular choices in the literature but, to our knowledge, a theoretically justifiable optimality criterion to select m has not been provided yet. In our experience, high values of m, let us say $m \geq 2$, result in a poor clustering behavior when dealing with the noise cluster based algorithms (this point is discussed later). Based on the previous considerations, we decided to use the values $m = 1.3$, 1.5 and 2.

As already mentioned, suitable choices of the parameters λ and β are also essential to reach satisfactory results. In fact, it was observed that the optimal selection of these parameters clearly depends on the value considered for m. Therefore, we proceeded to execute our simulations over a range of equally spaced values of λ and β, and the parameters retained were the ones maximizing the percentage of correct classification for each m. All the results reported hereafter correspond to this optimal selection of inputs for the algorithms. This way, we intend to perform fair comparisons, free of the effect of an inappropriate selection of the parameters.

The average percentages of correct classification obtained with the different models in the linear scenarios are shown in Table 4.8.

As expected, the standard algorithms show a very good behavior in Scenario L.1 without outliers. The two clusters are well-separated and both AR and QAF metrics are able to correctly classify all the series. Also the robust versions FCMdC-Exp and FCMdC-NC work fine in this setup. Adding outlier times series fairly has a disruptive effect on the results,

Table 4.8: Average percentages of correct classification for the simulated linear scenarios

		Scenario L.1: no outliers			Scenario L.2: 1 outlier			Scenario L.3: 2 outliers		
	Model	$m = 1.3$	$m = 1.5$	$m = 2$	$m = 1.3$	$m = 1.5$	$m = 2$	$m = 1.3$	$m = 1.5$	$m = 2$
$T = 150$	*AR–based*									
	AR-FCMdC	100.0	100.0	100.0	20.7	37.4	77.4	0.0	0.0	0.0
	AR-FCMdC-Exp	100.0	100.0	99.7	76.2	81.4	88.7	58.2	61.2	68.0
	AR-FCMdC-NC	100.0	100.0	99.5	46.6	35.0	7.4	32.5	23.9	4.9
	AR-TrFCMdC	–	–	–	68.5	66.0	54.5	57.1	55.4	52.0
	QAF-based									
	QAF-FCMdC	100.0	100.0	100.0	14.4	28.3	58.3	0.0	0.0	0.0
	QAF-FCMdC-Exp	100.0	100.0	100.0	86.5	85.7	90.3	77.3	79.0	84.0
	QAF-FCMdC-NC	100.0	100.0	99.9	67.7	56.7	12.7	55.6	45.7	9.8
	QAF-TrFCMdC	–	–	–	88.7	89.0	86.6	84.4	83.9	80.9
$T = 250$	*AR–based*									
	AR-FCMdC	100.0	100.0	100.0	21.8	42.7	81.4	0.0	0.0	0.0
	AR-FCMdC-Exp	100.0	100.0	100.0	99.7	99.7	99.9	83.5	86.4	89.3
	AR-FCMdC-NC	100.0	100.0	100.0	98.3	97.0	84.8	61.9	51.7	18.2
	AR-TrFCMdC	–	–	–	99.3	99.3	99.8	84.7	85.0	81.7
	QAF-based									
	QAF-FCMdC	100.0	100.0	100.0	20.9	38.8	71.1	0.0	0.0	0.0
	QAF-FCMdC-Exp	100.0	100.0	100.0	96.9	97.2	97.9	93.2	93.4	94.9
	QAF-FCMdC-NC	100.0	100.0	100.0	94.7	92.5	65.1	82.5	77.4	42.0
	QAF-TrFCMdC	–	–	–	99.2	99.5	99.6	96.4	96.6	96.9

which is clearly more pronounced with two outliers. In particular, AR-FCMdC and QAF-FCMdC present unsatisfactory success percentages for the three values of m, specially in Scenario L.3 were they failed always at correctly identify both outliers. Actually the non-anomalous series are always well-classified and the failures are caused by the outliers, which are seldom identified. For this reason the best results are reached for the highest value of m, since high values for m imply softer boundaries between clusters, and hence the memberships assigned to the outliers are closer to 0.5. To illustrate these assertions, we have randomly selected one set of 100 trials from the Scenario L.3 and calculated the means and standard deviations of the membership degrees for $m = 2$ and $T = 250$ (the most favorable scenario). The results are displayed in Table 4.9. It is observed that the eight non-atypical series are always well-grouped. $\mathcal{O}_1$ present average memberships (highlighted in magenta) very close to the cut-off values (0.3 and 0.7) and standard deviations large enough to account for a non-negligible number of failures, while $\mathcal{O}_2$ is always assigned to cluster $\mathcal{C}_2$. Smaller values of m led to average memberships higher (lower) than 0.7 (0.3), thus generating worse results.

Regardless of the considered distance, the robust versions based on the exponential metric and the trimmed approach substantially outperform the standard models. With one outlier

Table 4.9: Mean and standard deviation (in brackets) of membership degrees computed from one randomly selected set of 100 trials in Scenario L.3, with $T = 250$ and $m = 2$.

	QAF-FCMdC		QAF-FCMdC-Exp		QAF-FCMdC-NC		
	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_1$	$\mathcal{C}_2$	NC
X_1	0.982	0.018	0.882	0.118	0.813	0.012	0.175
	(.015)	(.015)	(.089)	(.089)	(.147)	(.009)	(.138)
X_2	0.981	0.019	0.877	0.123	0.796	0.013	0.190
	(.021)	(.021)	(.148)	(.148)	(.154)	(.012)	(.142)
X_3	0.982	0.018	0.866	0.134	0.819	0.012	0.169
	(.018)	(.018)	(.147)	(.147)	(.142)	(.010)	(.133)
X_4	0.981	0.019	0.866	0.134	0.784	0.014	0.202
	(.018)	(.018)	(.142)	(.142)	(.158)	(.012)	(.147)
X_5	0.024	0.976	0.127	0.873	0.014	0.802	0.185
	(.025)	(.025)	(.159)	(.159)	(.012)	(.153)	(.142)
X_6	0.024	0.976	0.125	0.875	0.013	0.812	0.175
	(.022)	(.022)	(.130)	(.130)	(.010)	(.133)	(.123)
X_7	0.020	0.980	0.127	0.873	0.011	0.837	0.152
	(.022)	(.022)	(.133)	(.133)	(.009)	(.131)	(.122)
X_8	0.023	0.977	0.125	0.875	0.013	0.805	0.182
	(.021)	(.021)	(.137)	(.137)	(.010)	(.141)	(.131)
X_9	0.446	0.554	0.456	0.544	0.139	0.231	0.630
	(.166)	(.166)	(.062)	(.062)	(.052)	(.087)	(.051)
X_{10}	0.095	0.905	0.454	0.546	0.024	0.238	0.738
	(.031)	(.031)	(.050)	(.050)	(.002)	(.114)	(.112)

(L.2) and realizations of length $T = 250$, both models produced excellent success rates, between 97% and 100%. The results were somewhat worse with two outliers (Scenario L.3) but also satisfactory, particularly using the QAF distance (scores always above 95% and 92% with QAF-FCMdC-Exp and QAF-TrFCMdC, respectively). The standard fuzzy versions of the AR and QAF metrics failed when trying to classify the second outlier $\mathcal{O}_2$ since, as showed in Figure 4.9 (a), it is always closer to $\mathcal{C}_2$. The averages and standard deviations of the membership degrees highlighted in blue in Table 4.9 corroborate the high capability of QAF-FCMdC-Exp to identify the outlier time series. It is also remarkable that the robust QAF-based models performed somewhat better than the AR–based ones despite handling ARMA models. For instance, Figure 4.11 shows the evolution of the percentages of correct classification for AR-FCMdC-Exp and QAF-FCMdC-Exp as function of β in Scenario L.3 with $T = 250$. Besides getting insight into the optimal values for β, Figure 4.11 allows us to conclude that QAF-FCMdC-Exp is preferable to AR-FCMdC-Exp for the three values of m if a suitable choice of β is considered.

As far as the trimmed approach is concerned, Table 4.10 shows that the QAF distance was more efficient than the AR one in removing the true outlier time series in Scenarios L.2 and L.3.

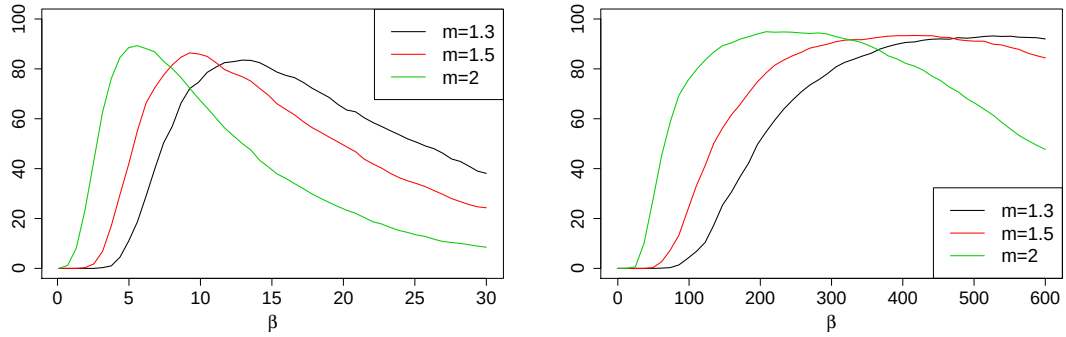

Figure 4.11: Average percentage of correct classification as a function of β by using AR–FCMdC-Exp (left panel) and QAF-FCMdC-Exp (right panel) models in Scenario L.3 with $T = 250$.

Table 4.10: Average percentage of the number of correctly trimmed outliers by using AR–TrFCMdC and QAF-TrFCMdC in the linear scenarios L.2 and L.3.

			Scenario L.2	Scenario L.3	
		Model	1 outlier	1 outlier	2 outliers
$T = 150$	$m = 1.3$	AR–TrFCMdC	68.5	18.8	62.1
		QAF-TrFCMdC	88.7	8.6	76.5
	$m = 1.5$	AR–TrFCMdC	66.0	20.0	59.3
		QAF-TrFCMdC	89.0	7.4	77.7
	$m = 2.0$	AR–TrFCMdC	54.5	23.7	50.3
		QAF-TrFCMdC	86.6	7.8	74.0
$T = 250$	$m = 1.3$	AR–TrFCMdC	99.3	5.3	89.3
		QAF-TrFCMdC	99.2	1.8	92.4
	$m = 1.5$	AR–TrFCMdC	99.3	5.3	88.2
		QAF-TrFCMdC	99.5	1.7	92.5
	$m = 2.0$	AR–TrFCMdC	99.8	11.9	80.9
		QAF-TrFCMdC	99.6	2.1	95.1

The fuzzy models based on the noise cluster, AR–FCMdC-NC and QAF-FCMdC-NC, reported good results but worse than the ones obtained with the other robust algorithms. In particular, the percentage of success substantially decayed with $m = 2$ and in presence of two outliers. The reason is again that a more balanced distribution of the membership degrees occurs as m increases, thus making more difficult to assign the outliers to the noise cluster. For illustrative purpose only, let us briefly come back to Table 4.9. As required, the highest average memberships of the outliers with QAF-FCMdC-NC (highlighted in orange) correspond to the noise cluster. Nevertheless they are not significantly greater than

the cut-off value, 0.6, and therefore an important number of trials draw out erroneous classification. Likewise Figure 4.11, we have depicted the evolution of the percentages of correct classification by using AR–FCMdC-NC and QAF-FCMdC-NC as function of λ in Figure 4.12. The poor rates of correct classification with $m = 2$ are evident for all λ, thus concluding that the only way to improve the results is using a less stringent cut-off value. Comparison of the two panels in Figure 4.12 also highlights the superiority of the QAF distance to develop the noise cluster fuzzy model.

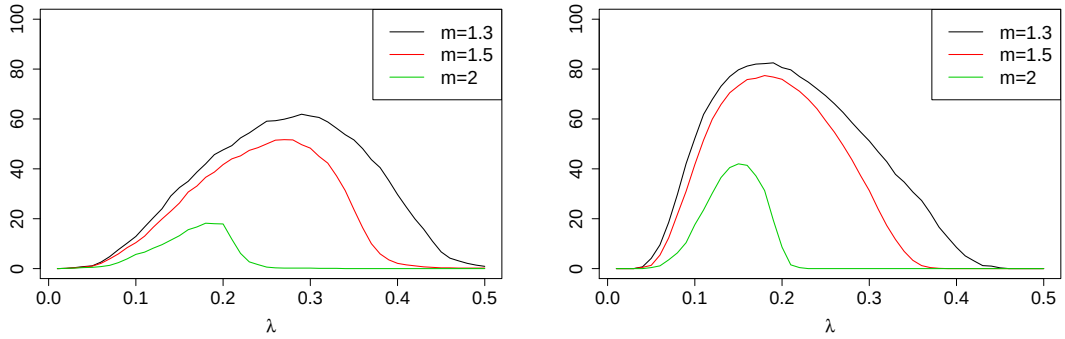

Figure 4.12: Average percentage of correct classification as a function of λ by using AR–FCMdC-NC (left panel) and QAF-FCMdC-NC (right panel) models in Scenario L.3 with $T = 250$.

As far as the scenarios NL.1, NL.2 and NL.3, including non–linear models, the most noticeable fact was the excellent performance showed by the QAF-based models. Although the models based on the AR distance were considered in our experiments, model misspecification heavily affected the results and they have been omitted. Table 4.11 reports the simulation results for the three non–linear scenarios using the QAF distance and Table 4.12 exhibits means and standard deviations of memberships for an arbitrary set of 100 trials in Scenario NL.3. The percentages of correct classification are higher than in the linear scenarios in these new setups for all models and values of m, particularly by working with the shortest series ($T = 150$). The average percentage of times in which QAF-TrFCMdC trimmed the true outlier in Scenario NL.2 was always above 97.8% for $T = 150$ and 99.6% for $T = 250$, while in Scenario NL.3 the two true outliers were detected above 97% and 99.2% for $T = 150$ and $T = 250$, respectively. It is also significant the improvement of the results for the robust model based on the noise cluster. The average membership degrees reported in Table 4.12 for the outliers time series and graphs in Figure 4.13 help us to understand this improvement.

Simulation results from the heteroskedastic scenarios CH.1, CH.2 and CH.3 based on the

Table 4.11: Average percentages of correct classification for the simulated non–linear scenarios

		Scenario NL.1: no outliers			Scenario NL.2: 1 outlier			Scenario NL.3: 2 outliers		
	Model	$m = 1.3$	$m = 1.5$	$m = 2$	$m = 1.3$	$m = 1.5$	$m = 2$	$m = 1.3$	$m = 1.5$	$m = 2$
$T = 150$	*QAF-based*									
	QAF-FCMdC	100	100	100	27.2	41.8	73.5	7.8	22.3	60.1
	QAF-FCMdC-Exp	100	100	100	95.9	96.2	97.9	94.8	95.4	97.0
	QAF-FCMdC-NC	100	100	100	87.5	79.9	26.6	81.5	73.2	21.6
	QAF-TrFCMdC	–	–	–	98.0	97.8	97.9	97.1	97.4	97.0
$T = 250$	*QAF-based*									
	QAF-FCMdC	100	100	100	25.0	42.7	76.9	14.4	33.6	71.0
	QAF-FCMdC-Exp	100	100	100	99.6	99.8	100	99.6	99.6	99.7
	QAF-FCMdC-NC	100	100	100	96.9	95.0	67.1	96.5	93.5	68.1
	QAF-TrFCMdC	–	–	–	99.6	99.6	99.8	99.2	99.2	99.5

Table 4.12: Mean and standard deviation (in brackets) of membership degrees computed from one randomly selected set of 100 trials in Scenario NL.3, with $T = 250$ and $m = 2$.

	QAF-FCMdC		QAF-FCMdC-Exp		QAF-FCMdC-NC		
	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_1$	$\mathcal{C}_2$	NC
X_1	0.977	0.023	0.899	0.101	0.832	0.014	0.153
	(.017)	(.017)	(.084)	(.084)	(.139)	(.012)	(.127)
X_2	0.975	0.025	0.894	0.106	0.819	0.016	0.165
	(.019)	(.019)	(.078)	(.078)	(.120)	(.011)	(.109)
X_3	0.977	0.023	0.884	0.116	0.815	0.016	0.169
	(.020)	(.020)	(.085)	(.085)	(.136)	(.012)	(.124)
X_4	0.978	0.022	0.877	0.123	0.814	0.016	0.170
	(.022)	(.022)	(.081)	(.081)	(.134)	(.012)	(.122)
X_5	0.013	0.987	0.066	0.934	0.009	0.891	0.100
	(.012)	(.012)	(.056)	(.056)	(.007)	(.091)	(.084)
X_6	0.012	0.988	0.064	0.936	0.009	0.897	0.094
	(.013)	(.013)	(.061)	(.061)	(.008)	(.094)	(.086)
X_7	0.010	0.990	0.064	0.936	0.009	0.896	0.095
	(.011)	(.011)	(.052)	(.052)	(.008)	(.087)	(.079)
X_8	0.014	0.986	0.066	0.934	0.009	0.893	0.098
	(.014)	(.014)	(.059)	(.059)	(.008)	(.101)	(.093)
$\mathcal{O}_1$	0.583	0.417	0.523	0.477	0.210	0.153	0.637
	(.132)	(.132)	(.040)	(.040)	(.070)	(.042)	(.042)
$\mathcal{O}_2$	0.525	0.475	0.504	0.496	0.159	0.147	0.695
	(.099)	(.099)	(.017)	(.017)	(.040)	(.027)	(.031)

GARCH and QAF distances are shown in Tables 4.13–4.15 and Figures 4.14–4.15. As already mentioned, conditional heteroskedasticity induces a more complex scenario because of the simulated realizations from GARCH processes are characterized by high dispersion for small sample sizes (Aielli and Caporin, 2013). Table 4.13 corroborates this feature since success rates comparable to the ones obtained in the linear and non–linear scenarios are

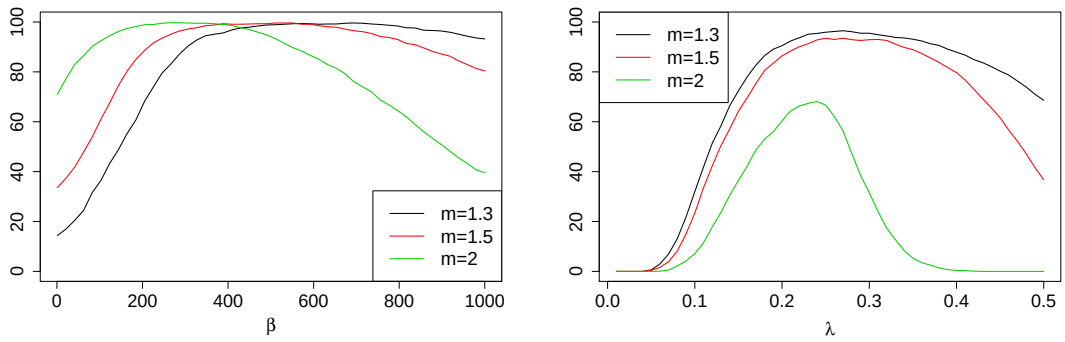

Figure 4.13: Average percentage of correct classification in Scenario NL.3 with $T = 250$ for QAF-FCMdC-Exp (left panel) and QAF-FCMdC-NC (right panel) models as function of β and λ, respectively.

Table 4.13: Average percentages of correct classification for the simulated conditional heteroskedastic scenarios

		Scenario CH.1: no outliers			Scenario CH.2: 1 outlier			Scenario CH.3: 2 outliers		
	Model	$m = 1.3$	$m = 1.5$	$m = 2$	$m = 1.3$	$m = 1.5$	$m = 2$	$m = 1.3$	$m = 1.5$	$m = 2$
$T = 1500$	*GARCH-based*									
	GARCH-FCMdC	63.0	63.0	62.8	11.8	22.8	43.1	2.4	5.8	21.7
	GARCH-FCMdC-Exp	63.0	63.0	62.6	53.8	53.7	53.0	45.8	45.1	41.5
	GARCH-FCMdC-NC	63.0	62.9	61.9	50.7	53.6	50.0	0.2	45.2	41.0
	GARCH-TrFCMdC	–	–	–	58.3	58.0	57.7	50.5	50.5	50.5
	QAF-based									
	QAF-FCMdC	99.5	98.9	96.9	34.7	56.9	86.4	10.2	29.2	75.8
	QAF-FCMdC-Exp	99.5	98.9	96.9	84.4	88.0	89.9	75.7	79.1	83.2
	QAF-FCMdC-NC	99.5	98.6	85.7	53.9	29.3	0.5	36.8	17.5	0.1
	QAF-TrFCMdC	–	–	–	87.5	85.5	79.0	76.4	73.7	66.4
$T = 2500$	*GARCH-based*									
	GARCH-FCMdC	69.8	69.8	69.5	20.3	33.8	58.1	3.0	10.8	34.1
	GARCH-FCMdC-Exp	69.8	69.8	69.4	63.1	63.2	63.8	59.6	59.9	59.9
	GARCH-FCMdC-NC	69.8	69.8	69.4	62.6	63.3	62.0	58.7	58.4	57.0
	GARCH-TrFCMdC	–	–	–	66.7	66.7	66.7	63.1	63.1	63.1
	QAF-based									
	QAF-FCMdC	99.9	100.0	100.0	30.5	57.1	93.8	7.9	29.0	81.6
	QAF-FCMdC-Exp	100.0	100.0	100.0	96.0	97.4	98.0	90.9	92.9	93.9
	QAF-FCMdC-NC	99.9	100.0	98.4	83.2	66.7	5.8	67.0	48.4	2.6
	QAF-TrFCMdC	–	–	–	98.8	98.2	97.1	95.4	95.0	91.9

only attained with $T = 2500$. It is worthy mentioning that these sample sizes are frequently considered in the literature by working with heteroskedastic processes. Notice also that the membership degrees for the non–anomalous series in Table 4.14 are moderately further from 0 and 1 than in previous analyses, thus emphasizing the major difficulty of clustering

under heteroskedasticity. In fact, non–anomalous series were sometimes missclassified using the GARCH distance. This assertion is easily understood by comparing the outputs in Tables 4.13 and 4.15. It is observed in Table 4.15 that GARCH-TrFCMdC and QAF-TrFCMdC present similar percentages of success by trimming the true outliers, but in contrast QAF-TrFCMdC exhibits higher average percentages of correct classification in Table 4.13.

Again the main conclusion is that the QAF-based models fairly outperform the GARCH-based ones. While the latter are affected by the inaccurate estimation of the GARCH parameters, the former take advantage of the capability of the QAF distance to detect changes in conditional shapes and to deal with heavy-tailed marginal distributions. As in above scenarios, the robust models, particularly QAF-FCMdC–Exp and QAF-TrFCMdC, led to the best results in presence of outliers regardless of the fuzziness parameter. In this case, the model based on the noise cluster showed worse results, specially in Scenario CH.3 with two outiler time series (see Table 4.14 and Figure 4.15).

Table 4.14: Mean and standard deviation (in brackets) of membership degrees computed from one randomly selected set of 100 trials in Scenario CH.3, with $T = 2500$ and $m = 2$.

	QAF-FCMdC		QAF-FCMdC-Exp		QAF-FCMdC-NC		
	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_1$	$\mathcal{C}_2$	NC
X_1	0.888	0.112	0.813	0.187	0.781	0.088	0.130
	(.020)	(.020)	(.020)	(.020)	(.020)	(.016)	(.008)
X_2	0.886	0.114	0.808	0.192	0.780	0.087	0.133
	(.023)	(.023)	(.015)	(.015)	(.015)	(.013)	(.006)
X_3	0.884	0.116	0.808	0.192	0.773	0.087	0.139
	(.026)	(.026)	(.014)	(.014)	(.022)	(.019)	(.008)
X_4	0.892	0.108	0.808	0.192	0.777	0.092	0.131
	(.021)	(.021)	(.016)	(.016)	(.028)	(.023)	(.011)
X_5	0.125	0.875	0.218	0.782	0.097	0.750	0.153
	(.022)	(.022)	(.011)	(.011)	(.015)	(.022)	(.011)
X_6	0.126	0.874	0.223	0.777	0.093	0.746	0.162
	(.022)	(.022)	(.016)	(.016)	(.016)	(.022)	(.012)
X_7	0.127	0.873	0.221	0.779	0.098	0.735	0.167
	(.018)	(.018)	(.010)	(.010)	(.020)	(.022)	(.008)
X_8	0.122	0.878	0.211	0.789	0.087	0.763	0.149
	(.024)	(.024)	(.017)	(.017)	(.015)	(.022)	(.009)
$\mathcal{O}_1$	0.573	0.427	0.516	0.484	0.302	0.222	0.476
	(.011)	(.011)	(.002)	(.002)	(.009)	(.005)	(.006)
$\mathcal{O}_2$	0.565	0.435	0.517	0.483	0.313	0.233	0.454
	(.005)	(.005)	(.003)	(.003)	(.004)	(.003)	(.004)

Table 4.15: Average percentage of the number of correctly trimmed outliers by using GARCH-TrFCMdC and QAF-TrFCMdC in the heterokedastic scenarios CH.2 and CH.3.

			Scenario CH.2	Scenario CH.3	
		Model	1 outlier	1 outlier	2 outliers
$T = 1500$	$m = 1.3$	GARCH-TrFCMdC	86.6	9.6	70.4
		QAF-TrFCMdC	87.6	9.6	76.4
	$m = 1.5$	GARCH-TrFCMdC	85.9	9.8	70.0
		QAF-TrFCMdC	86.0	10.2	73.7
	$m = 2.0$	GARCH-TrFCMdC	85.4	9.2	70.3
		QAF-TrFCMdC	81.9	12.3	67.6
$T = 2500$	$m = 1.3$	GARCH-TrFCMdC	93.8	5.7	81.8
		QAF-TrFCMdC	98.8	1.5	95.4
	$m = 1.5$	GARCH-TrFCMdC	93.6	5.6	81.8
		QAF-TrFCMdC	98.2	1.7	95.0
	$m = 2.0$	GARCH-TrFCMdC	93.1	5.3	82.2
		QAF-TrFCMdC	97.2	2.5	92.0

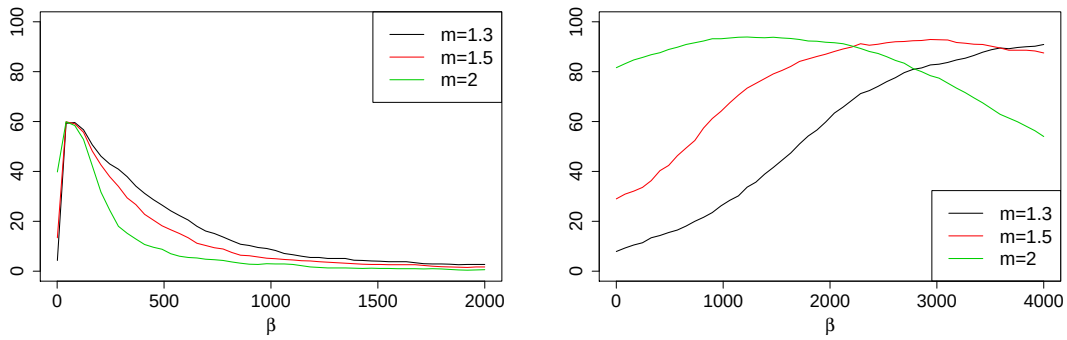

Figure 4.14: Average percentage of correct classification as a function of β by using GARCH-FCMdC-Exp (left panel) and QAF-FCMdC-Exp (right panel) models in Scenario CH.3 with $T = 2500$.

4.7 A case study: Clustering series of daily returns of Euro exchange rates

This section is devoted to show the effectiveness in practical situations of the robust fuzzy models based on quantile autocovariances. An specific application involving realizations of financial time series is performed. Our analysis is not aimed at deriving economic implications, but at illustrating the usefulness of the proposed clustering approaches to identify homogeneous groups with similar underlying temporal patterns and isolated series exhibit-

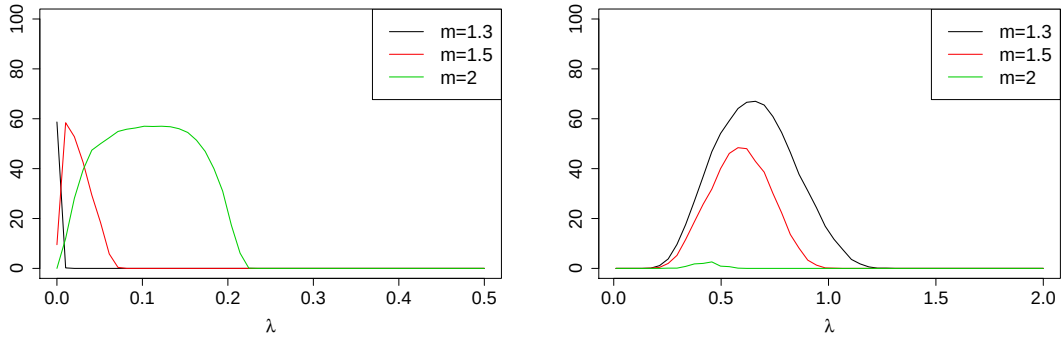

Figure 4.15: Average percentage of correct classification as a function of λ by using GARCH-FCMdC-NC (left panel) and QAF-FCMdC-NC (right panel) models in Scenario CH.3 with $T = 2500$.

ing atypical dynamic behaviors.

The database used in this section is the same as that used in Section 3.5 of Chapter 3, which consists of a set of series of the daily closing values of Euro exchange rates against twenty-eight international currencies, collected from 1st January 2010 to 28th February 2014 ($T = 1520$).

Just as in simulations, the metric d_{QAF} was constructed using three quantiles of levels 0.1, 0.5 and 0.9 and one lag ($L = 1$, with $l_1 = 1$). Also, in line with the range of values considered for the fuzziness parameter m in simulations, we select the values $m = 1.3$ and $m = 1.7$. Both values produced very similar results, in particular drawing the same number of outlier time series for all the considered fuzzy models. For it, only the results for $m = 1.7$ are here included.

A two-dimensional scaling (MDS) based on the pairwise QAF-dissimilarity matrix was carried out to gain insight both the spatial structure of the Euro exchange rates and the level of separability between groups. The corresponding coordinate matrices are displayed in Figure 4.16.

Figure 4.16 shows the existence of a reasonably compact cluster formed by eighteen series including the Euro exchange rates against the major international currencies and those linked to the US dollar, such as the Canadian dollar (CAD) and the Great Britain pound (GBP), among others. The remaining ten objects are more spread out. At least the Uruguayan peso (UYU) and the Thailand baht (THB) appear to be isolated, well-separated of the remaining currencies, and they could be identified as anomalous data. South African rand (ZAR), Argentine peso (ARS), Brazilian real (BLR), Serbian dinar (RSD), and Chilean

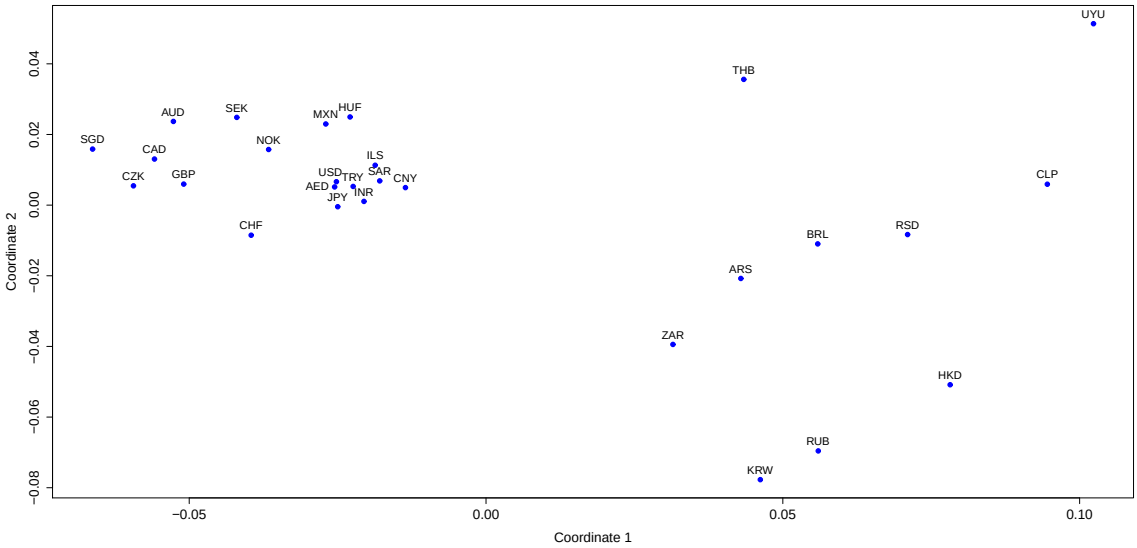

Figure 4.16: Two-dimensional scaling configurations based on the QAF-distance for the daily returns of Euro exchange against against 28 currencies.

peso (CLP) are placed close to each other, and they could constitute another cluster. The three remaining currencies are somewhat separated from the latter group and they could be joined to this group or form a third cluster. For comparison purpose, a two-dimensional scaling based on the AR metric was also performed. The resulting plot exhibits a configuration with much greater dispersion and without identifying well-separated groups, which is quite unrealistic in the analysed problem. These worse results are indeed expected because of the AR metric relies on autoregressive fits fairly inappropriate to model the heteroskedastic series in study.

Two different criteria to determine the optimal number of clusters C were considered, namely those values of C minimizing the Xie-Beni (Xie and Beni, 1991) and Kwon (Kwon, 1998) indexes. To simplify the definition of both indexes, let us denote by H_{ic} the squared Euclidean distance between the sequence of estimated quantile autocovariances for the i-th series and the average sequence for the c-th cluster, that is

$$H_{ic} = \sum_{k=1}^{L} \sum_{j=1}^{r} \sum_{j'=1}^{r} \left(\hat{\gamma}_{l_k}^{(i)}(\tau_j, \tau_{j'}) - \overline{\hat{\gamma}}_{l_k}^{(c)}(\tau_j, \tau_{j'}) \right)^2 . \tag{4.21}$$

The Xie-Beni index for a partition into C clusters is defined as the ratio between the total variance and the minimum separation between clusters, i.e.

$$XB(C) = \frac{\sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^{m} H_{ic}}{I \min_{c \neq c'} H_{cc'}}. \tag{4.22}$$

Note that minimizing the numerator of $XB(\cdot)$ in (4.22) is the goal of the QAF-FCMdC algorithm. On the other hand, the denominator measures how separated are the clusters, thus the Xie-Beni index decreases with the separation between clusters.

The Kwon index provides a correction of the Xie-Beni index by penalizing the decreasing tendency when the number of clusters becomes very large and close to the number of time series. Specifically, the Kwon index is defined as follows.

$$K(C) = \frac{\sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^{m} H_{ic} + \frac{1}{C}\sum_{c=1}^{C}\sum_{c'=1}^{C} H_{cc'}}{I \min_{c \neq c'} H_{cc'}}. \tag{4.23}$$

The values obtained for both indexes using the QAF-FCMdC model are depicted in Figure 4.17. In both cases the lowest value is attained for $C = 2$ clusters, with a substantial increase when three or more clusters are considered. Similar results were obtained by using the robust versions of the model, and therefore both criteria lead to conclude the existence of two major groups.

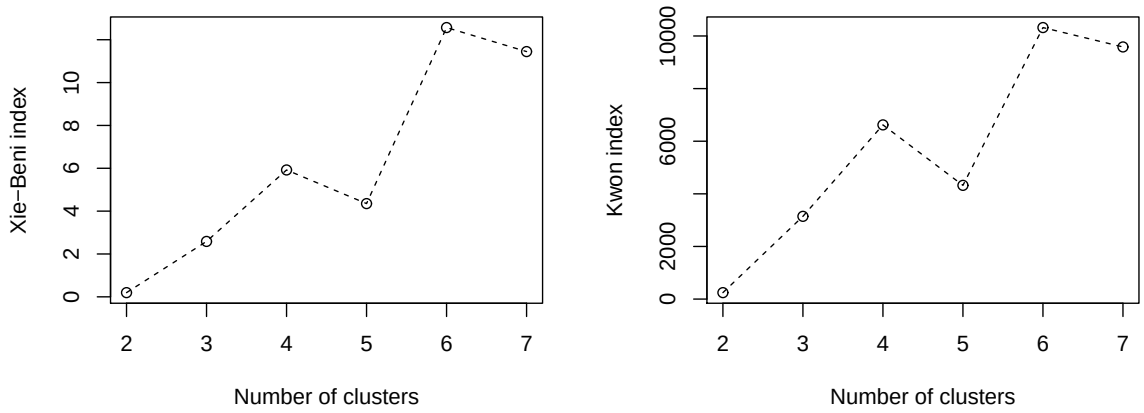

Figure 4.17: Xie–Beni and Kwon index values for different sizes of partition using QA-–FCMdC.

The value for the parameter β required by the QAF-FCMdC-Exp model was determined using (4.12) in Section 4.5, resulting $\beta = 1095.649$. To set δ in the QAF-FCMdC-NC model, we follow the approach suggested by Cimino et al. (2005), which consists of successively executing the fuzzy QAF-FCMdC-NC algorithm for decreasing values of δ, recording the percentage of series assigned to the noise cluster, and selecting the value of δ producing an abrupt change of slope (elbow) in this percentage. The idea is gradually reducing δ until a proper threshold is found out because of excessively small values of δ lead to assign non-anomalous objects into the cluster noise. According to this criterion, $\delta = 0.4$ was selected. As far as the QAF-TrFCMdC model, the trimming ratio minimizing the Xie-Beni and the Kwon indexes over a grid of possible values for α was considered as the optimal choice, resulting $\alpha = 0.1621$, i.e. five time series were trimmed.

Table 4.16 shows the membership degrees obtained by using the standard and the robust fuzzy methods. For each single series, the shaded cells enhance the highest membership degrees obtained with each procedure, i.e. the cluster assignments from a crisp perspective. The memberships showed in bold font for a particular robust procedure indicate time series identified as outlier. The currencies' names in bold font refer to series identified as outliers by the three robust methods. When only one or two robust procedures achieved that conclusion, the currency is written in italic font.

Overall, the obtained partition with the standard fuzzy model QA-FCMdC is consistent with the plot displayed in Figure 4.16. The medoid time series are the Emiratri dirham (AED), for the most compact cluster (C_2) grouping eighteen currencies, and the Brazilian real (BR) for the cluster C_1 exhibiting higher spread. It is noticeable that most of the currencies are assigned to one cluster with high membership degrees ($u_{ic} \geq 0.7$), the only exception being the Thailand baht (THB), which was located in C_1 with membership 0.639. Nevertheless, Figure 4.16 suggests that THB is too far from the time series forming C_2 and hence the Thailand baht should be considered as an outlier. In short, QAF-FCMdC seems to work reasonably fine, but it does not allow us to identify currencies showing an atypical behavior.

The partition obtained with QAF-FCMdC-Exp determines the existence of four outlier time series by splitting their membership degrees uniformly across the clusters, namely the Uruguayan peso (UYU), the Thailand baht (THB), the South Korean won (KRW) and the Russian ruble (RUB). These four currencies are also allocated together into the noise cluster with memberships $u_{in_C} > 0.6$ when the QAF-FCMdC-NC model is considered. The South African rand (ZAR) and the Hong Kong dollar (HKD) are also added to the noise cluster on the basis of much weaker memberships, particularly the former with memberships for $\mathcal{C}_1$ and the noise cluster hardly discernible, 0.458 and 0.463, respectively. Note that

Table 4.16: Membership degrees for the fuzzy clustering models based on quantile autocovariances by considering a two-cluster partition.

	QAF-FCMdC		QAF-FCMdC-Exp		QAF-FCMdC-NC			QAF-TrFCMdC		
	C_1	C_2	C_1	C_2	C_1	C_2	NC	C_1	C_2	Tr
USD	0.001	0.999	0.100	0.900	0.034	0.801	0.165	0.012	0.988	N
GBP	0.030	0.970	0.087	0.913	0.016	0.807	0.177	0.036	0.964	N
CAD	0.030	0.970	0.084	0.916	0.013	0.813	0.174	0.045	0.955	N
AUD	0.040	0.960	0.077	0.923	0.013	0.828	0.159	0.044	0.956	N
CHF	0.042	0.958	0.145	0.855	0.037	0.697	0.266	0.025	0.975	N
CZK	0.040	0.960	0.111	0.889	0.018	0.755	0.228	0.044	0.956	N
BRL	1.000	0.000	1.000	0.000	1.000	0.000	0.000	0.938	0.062	N
CNY	0.019	0.981	0.171	0.829	0.070	0.684	0.246	0.028	0.972	N
CLP	0.946	0.054	0.802	0.198	0.573	0.024	0.403	0.953	0.047	N
AED	0.000	1.000	0.085	0.915	0.028	0.829	0.143	0.012	0.988	N
SGD	0.044	0.956	0.159	0.841	0.022	0.652	0.326	0.067	0.933	N
ZAR	0.799	0.201	0.720	0.280	0.458	0.079	**0.463**	0.691	0.309	N
RUB	0.809	0.191	**0.636**	**0.364**	0.281	0.044	**0.674**	–	–	**Y**
NOK	0.021	0.979	0.000	1.000	0.000	1.000	0.000	0.013	0.987	N
SEK	0.045	0.955	0.017	0.983	0.003	0.963	0.034	0.030	0.970	N
HUF	0.066	0.934	0.059	0.941	0.017	0.877	0.106	0.012	0.988	N
TRY	0.012	0.988	0.085	0.915	0.031	0.832	0.137	0.004	0.996	N
ARS	0.911	0.089	0.858	0.142	0.705	0.038	0.257	0.693	0.307	N
SAR	0.004	0.996	0.140	0.860	0.053	0.730	0.218	0.019	0.981	N
KRW	0.744	0.256	**0.591**	**0.409**	0.214	0.051	**0.735**	–	–	**Y**
JPY	0.015	0.985	0.113	0.887	0.035	0.770	0.194	0.018	0.982	N
HKD	0.904	0.096	0.723	0.277	0.426	0.031	**0.543**	–	–	**Y**
INR	0.020	0.980	0.103	0.897	0.033	0.790	0.177	0.027	0.973	N
ILS	0.022	0.978	0.076	0.924	0.029	0.853	0.119	0.000	1.000	N
RSD	0.964	0.036	0.894	0.106	0.768	0.020	0.212	1.000	0.000	N
UYU	0.812	0.188	**0.581**	**0.419**	0.193	0.039	**0.767**	–	–	**Y**
THB	0.639	0.361	**0.582**	**0.418**	0.239	0.105	**0.655**	–	–	**Y**
MXN	0.045	0.955	0.058	0.942	0.017	0.877	0.106	0.022	0.978	N

consideration of these isolated objects modifies the $\mathcal{C}_2$ medoid, now resulting the Norwegian krone (NOK) which seems to be a more representative prototype than AED in Figure 4.16. The fuzzy QAF-TrFCMdC model draw out very similar results. Considering a trimmed ratio of $\alpha = 0.1621$, five Euro exchange currencies are trimmed away, namely the same four outliers identified by the other two robust methods plus HKD. Actually, a small reduction of the trimmed ratio allows to cancel this additional outlier so that in essence the three robust methods allows us to obtain similar conclusions.

4.8 Concluding remarks

In this chapter we have shown that the sample quantile autocovariances are an useful tool to perform soft partitional clustering of times series when the target is to group series generated from the same stochastic process.

Soft partitional clustering has been considered by introducing a fuzzy C-medoids clustering model for time series based on the sample quantile autocovariances (QAF-FCMdC). Fuzzy paradigm enriches the cluster solution by permitting overlapping clusters, i.e. identifying time series with dynamics close to more than one prototype. To evaluate the QAF-FCMdC algorithm, we have carried out numerical experiments including clusters with different levels of separability and time series equidistant from several clusters. Our assessment criterion took into account the capability of the examined algorithms to detect the fuzzy nature of the equidistant series. Regardless of the considered models and compared with other fuzzy algorithms based on distances between estimates of the underlying parametric structures, the proposed fuzzy algorithm produced good results. Overall, QAF-FCMdC reported better results in the most complex scenarios, where the clusters are closer each other. The most noticeable differences in favour of QAF-FCMdC were observed by clustering GARCH(1,1) processes, particularly for large sample sizes. In this framework, the GARCH-based algorithms were affected by the inaccurate estimation of the GARCH structure. By contrast, QAF-FCMdC is free of determining the underlying parametric structure and takes advantage of the capability of the quantile autocovariances to detect changes in conditional shapes, thus permitting to discriminate between volatility structures and identify series showing fuzzy behavior. Furthermore, QAF-FCMdC can be applied to series with different lengths, and it is simple to implement and computationally lighter than the analyzed competitors. According to these properties, the proposed fuzzy algorithm is a promising tool to be applied in many situations where it is unrealistic to assume homoscedasticity, such as we have illustrated by means of a specific case-study.

The fuzzy approach based on the QAF metric introduced in the first part of this chapter has been published in Lafuente-Rego and Vilar (2016b), and a more comprehensive and detailed study encompassing both soft and hard partitional approaches is available in the paper by Vilar et al. (2017).

Other additional issue dealt with in this chapter was to obtain robust versions of the fuzzy QAF-FCMdC algorithm to neutralize the effect of anomalous fuzzy series. Three different generalizations of the robustness techniques considered by D'Urso and Giovanni (2014), namely the metric approach by smoothing the distance (QAF-FCMdC-Exp), the noise approach by introducing an artificial noise cluster (QAF-FCMdC-NC) and the trimmed approach by trimming away a small fraction of series (QAF-TrFCMdC), were introduced. For the evaluation of these techniques a broad simulation study was considered. Just as happened with the QAF-FCMdC, the fuzzy robust proposals produced good results regardless of the considered models and compared with other fuzzy algorithms based on distances between estimates of the underlying parametric structures. The proposed robust

models work fine and produce very satisfactory results in presence of outliers when an optimal selection of the input parameters is made. The real clustering structure is not altered since the fuzzy models are able to neutralize the effect of the anomalous series. When the robust versions are compared, a slight improvement is observed by using QAF-FCMdC-Exp and QAF-TrFCMdC, but it is relevant to emphasize that the noise approach can report similar results by correctly handling the combination of the fuzziness parameter and the noise distance. Overall, all the robust procedures are particularly sensitive to the choice of the input parameters. An specific application involving realizations of financial time series allowed to illustrate the usefulness of the proposed clustering approaches to identify series exhibiting atypical dynamic behaviors.

Chapter 5

Soft clustering of time series: New approaches based on mixture models and D-probabilistic techniques

Contents

5.1 Introduction

Besides fuzzy clustering approaches, other clustering algorithms belonging to the domain of soft computing have been proposed and successfully applied in the past decades. In this chapter we confront the fuzzy QAF-FCMdC algorithm against two soft classification alternatives, namely the probabilistic D-clustering and an approach based on finite mixture models using the Expectation-Maximization (EM) algorithm.

A possible via by performing model-based clustering is to consider that the underlying distribution has the form of a suitable finite mixture of parametric distributions, where each mixture component describes the probabilistic nature of a specific group in the dataset (Fraley and Raftery, 2002; Melnykov and Maitra, 2010; Chen and Maitra, 2011). In the time

series setting, this approach is not simple due to the high dimensionality of the data objects. Wong and Li (2000) consider a first-order autoregressive Gaussian mixture model to time series data, and later Chen and Maitra (2011) extend this model to include information from explanatory variables and consider more general p-th order autoregressive time series. Both procedures work in the time domain and take advantage of the reasonably simple form ($p+1$ free parameters) of the covariance matrix specifying the dispersion of an AR(p) with common marginal variance. Is spite of it, the traditional maximum likelihood approach of estimating the parameters using the expectation-maximization (EM) algorithm is here computationally demanding due to the high dimensionality of the problem. In fact, a novel conditional maximization algorithm is proposed to speed up the process and obtain a more efficient implementation.

These works motivate the need of developing alternative methods to perform clustering of time series based on mixture models. In this line, we propose to look at the frequency domain and consider the asymptotic representation of the log-periodogram by means of a nonparametric regression model with log-exponentially distributed errors. Assuming that the time series within the same cluster are characterized by a specific spectral density, a nonparametric finite mixture of univariate regression models with known probability distribution is available. Estimation of the mixture model involves nonparametric approximations of the log-periodograms for each cluster and estimators of the probabilities of belonging to the clusters. To obtain these estimators, a local-likelihood estimation procedure (Tibshirani and Hastie, 1987) is carried out by implementing an EM algorithm (Dempster et al., 1977). As it is well known, the EM algorithm alternates between two different stages. At the $(s+1)$-th iteration, the expectation (E) step calculates the expected value of the unobserved variables indicating the probabilities of each times series to belong to every cluster (latent variables), using the conditional distribution at the current parameter values obtained at the end of the s-th iteration. In the maximization (M) step, the centers of the clusters and the prior probabilities are computed by maximizing the expected log-likelihood built on the E-step. The algorithm iterates until convergence is achieved. As it will be detailed later, the usual E-step needs to be here modified to obtain a proper solution. Unlike fuzzy and probabilistic D-clustering approaches, clustering based on mixture models does not require to fix a metric to measure dissimilarity between time series and reports a soft partition without specifying a fuzziness parameter such as fuzzy procedures do.

The probabilistic D-clustering (Ben-Israel and Iyigun, 2008) is based on the idea that the probability of cluster membership at any point is inversely proportional to the distance from the center of the cluster in question. Given an arbitrary data object x, the basic principle of this algorithm is to assume that $d_k(x)p_k(x) = \text{cte}$ (depending on x), for all cluster $\mathcal{C}$,

where $d_k(x)$ and $p_k(x)$ denote the distance from x to the center of $\mathcal{C}$ and the probability that x is a member of $\mathcal{C}$, respectively. This way, the closer to the center of a cluster, the higher probability of belonging to that cluster. Unlike the model-based approaches, selecting a proper metric is here very important to obtain a satisfactory partition. Results in Chapters 3 and 4 support the idea of using the distance based on quantile autocovariances in probabilistic D-clustering of time series. This intuition is fully confirmed in the numerical experiments carried out in this chapter. Likewise the mixture models approach, the probabilistic D-clustering algorithm prevents specifying a fuzziness parameter. Fuzziness is automatically determined in terms of distances to the different cluster centers. Indeed, this is a very nice property given the noticeable influence of the fuzziness parameter observed in prior chapters.

The present chapter is structured as follows. Section 5.2 addresses cluster analysis of time series in the frequency domain based on nonparametric mixture models using the expectation maximization algorithm. The estimation procedure is described in detail and the modification required in the E-step is discussed and motivated. Section 5.3 proposes to perform times series clustering using the probabilistic D-clustering algorithm by plug-in the distance based on the estimated quantile autocovariances introduced in Chapter 3. The performance of both soft clustering approaches is analysed and compared to the fuzzy QAF-FCMdC model throughout a simulation study in Section 5.4, and the chapter ends with a summary of the main conclusions in Section 5.5.

5.2 A nonparametric mixture model for time series clustering

Let S be a set of n realizations of univariate stationary time series with zero mean denoted by $\boldsymbol{X}_t^{(i)} = \left\{X_1^{(i)}, \ldots, X_{T_i}^{(i)}\right\}$, for $i = 1, \ldots, n$. Let us assume for simplicity $T_i = T$, for all i. Consider the corresponding spectral representations via the log-periodograms $I_k^{(i)}$, $i = 1, \ldots, n$, evaluated at the Fourier frequencies λ_k, $k = 1, \ldots, M$, with $M = [(T-1)/2]$. According to Section 1.4.3 in the introductory chapter, for each time series the sequence of centered log-periodograms $Y_k^i = \log(I_k^i) - C_0$, with $C_0 = -0.57721$ being the Euler's constant, approximately admits the nonparametric regression model given by

$$Y_k^i = m^i(\lambda_k) + \epsilon_k^i,$$

where $m^i(\cdot) = \log(f^i(\cdot))$ denotes the logarithm of the spectral density for the i-th series, and the errors ϵ_k^i are asymptotically i.i.d. with probability density function $\varphi(\lambda) = \exp(\lambda - \exp(\lambda))$.

Assuming the existence of C homogeneous groups for the n series, i.e. the existence of C different spectral densities, $\mathbf{f} = \{f^1(\cdot), \ldots, f^C(\cdot)\}$, then the set S of observed time series satisfies

$$Y_k^i = m^c(\lambda_k) + \epsilon_k^i, \quad \text{for } i = 1, \ldots, n,\ k = 1, \ldots, M \ \text{ and } c = 1, \ldots, C. \tag{5.1}$$

Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_C)^t$ be the vector of prior probabilities for each time series into each cluster, i.e. $\pi_c = \mathbb{P}\left(\boldsymbol{X}_t^{(i)} \in \text{group } c\right)$, for all $i = 1, \ldots, n$, and $c = 1, \ldots, C$.

Denote by $\Theta = \{\pi_1, \ldots, \pi_{C-1}, m^1(\cdot), \ldots, m^C(\cdot)\}$ the set of unknown parameters and functions determining the probabilistic structure of the observed n time series. From (5.1), it is concluded that the probability density function of the errors, let us say $g(\cdot)$, can be written as

$$g\left(\epsilon_k^i/\Theta\right) = \sum_{c=1}^{C} \pi_c\, \varphi\left(Y_k^i - m^c(\lambda_k)\right), \quad \text{for } i = 1, \ldots, n, \text{ and } \ k = 1, \ldots, M. \tag{5.2}$$

Equation (5.2) establishes that the density of the errors from the nonparametric regression models (5.1) has the form of a finite mixture of distributions whose c-th coefficient represents the probability that the corresponding time series belongs to the c-th cluster. According to (5.2), the likelihood of the set of unknown parameters and log-spectra, Θ, given the data in hand, $Y \equiv \{\left(\lambda_k, Y_k^i\right),\ k = 1, \ldots, M,\ i = 1, \ldots, n\}$, is given by

$$L(\Theta/Y) = \prod_{i=1}^{n} \prod_{k=1}^{M} \sum_{c=1}^{C} \pi_c\, \varphi\left(Y_k^i - m^c(\lambda_k)\right),$$

and the corresponding log-likelihood by

$$\mathcal{L}(\Theta/Y) = \log L(\Theta/Y) = \sum_{i=1}^{n} \sum_{k=1}^{M} \log\left(\sum_{c=1}^{C} \pi_c \varphi\left(Y_k^i - m^c(\lambda_k)\right)\right).$$

Nevertheless, the elements $m^c \in \Theta$ are actually functions, which suggests to address the problem as a local optimization problem assuming that the log-spectra are smooth. Thus, nonparametric kernel approximations for $m^c(\cdot)$ can be obtained by maximizing the local log-likelihood function instead of the log-likelihood function. Regarding the Jensen inequality

for concave functions, the local log-likelihood function takes the form

$$\begin{aligned}\ell(\Theta/Y)(\lambda) &= \sum_{i=1}^{n}\sum_{k=1}^{M}\log\left(\sum_{c=1}^{C}\pi_c\,\varphi\left(Y_k^i - m^c(\lambda)\right)\right)K_h(\lambda_k-\lambda)\\ &\geq \sum_{i=1}^{n}\sum_{k=1}^{M}\sum_{c=1}^{C}\log\left(\pi_c\varphi\left(Y_k^i - m^c(\lambda)\right)\right)K_h(\lambda_k-\lambda),\end{aligned}$$

where $K_h(\cdot) = \frac{1}{h}K\left(\frac{\cdot}{h}\right)$ is a kernel function $K(\cdot)$ rescaled with a bandwidth h. Maximization of the local-likelihood function $\ell(\Theta/Y)(\cdot)$ is carried out by using an Expectation-Maximization (EM) algorithm. It is worth to notice that kernel regression is here performed instead of local linear regression (as e.g. in Chapter 1) in order to yield closed-form solutions in the M-step of the EM algorithm.

In the EM framework, the mixture model problem is formulated as an incomplete data problem. The observed data are considered to be incomplete since each data has associated an unobserved value, or latent variable, specifying the mixture component to which this data belongs. To formulate the problem in terms of complete data, labels $(z_{i1},\ldots,z_{iC})$, $c=1,\ldots,C$, are assigned to the i-th series, for all $i=1,\ldots,n$, where $z_{ic}=1$ if the time series belongs to cluster c and 0 otherwise. Hereafter, Z will denote the $(n\times C)$-matrix which i-th row is the vector $\boldsymbol{Z}^{(i)} = (z_{i1},\ldots,z_{iC})^t$, with $z_{ic} = \mathbb{1}_{\{\boldsymbol{X}_t^{(i)}\in \text{ group } c\}}$. Thus, the "complete data" are $\left\{\boldsymbol{X}_t^{(i)}, \boldsymbol{Z}^{(i)}\right\}$, and the local log-likelihood with complete data takes the form

$$\ell(\Theta/Y,Z)(\lambda) = \sum_{i=1}^{n}\sum_{c=1}^{C} z_{ic}\sum_{k=1}^{M}\log\{\pi_c\,\varphi\left(Y_k^i - m^c(\lambda)\right)\}K_h(\lambda_k-\lambda).$$

The expected value of the labels $\{z_{ic}\}$ conditional on the most recent estimators of Θ (estimates for π and m^c obtained in the above M-step) are calculated and iteratively updated in the expectation step (E-step).

The $(s+1)$-th iteration of the EM procedure is detailed below. At the end of the s-th iteration, estimates $\Theta_s = \{\pi_1^{(s)},\ldots,\pi_{C-1}^{(s)}, m^{1^{(s)}}(\cdot),\ldots,m^{C^{(s)}}(\cdot)\}$ are available. Then the E- and M-steps proceed as follows.

E-step According to estimates from the iteration s, we have

$$z_{ic}^{(s+1)} = \mathbb{E}\left(z_{ic}/\Theta_s, Y\right) = \mathbb{P}\left(\boldsymbol{X}_t^{(i)}\in \text{group } c\,/\,\Theta_s, Y\right),$$

for each $c=1,\ldots,C$ and $i=1,\ldots,n$. The standard approach to estimate this expectation

is to use the Bayes' rule,

$$
\begin{aligned}
z_{ic}^{(s+1)} &= \frac{\pi_c^{(s)} \prod_{k=1}^{M} \varphi\left(Y_k^i - m^{c^{(s)}}(\lambda_k)\right)}{\sum_{c'=1}^{C} \pi_{c'}^{(s)} \prod_{k=1}^{M} \varphi\left(Y_k^i - m^{c'^{(s)}}(\lambda_k)\right)} \\
&= \frac{\pi_c^{(s)} \prod_{k=1}^{M} \exp\left(Y_k^i - m^{c^{(s)}}(\lambda_k) - \exp\left(Y_k^i - m^{c^{(s)}}(\lambda_k)\right)\right)}{\sum_{c'=1}^{C} \pi_{c'}^{(s)} \prod_{k=1}^{M} \exp\left(Y_k^i - m^{c'^{(s)}}(\lambda_k) - \exp\left(Y_k^i - m^{c'^{(s)}}(\lambda_k)\right)\right)}
\end{aligned}
\tag{5.3}
$$

for $i = 1, \ldots, n$ and $c = 1, \ldots, C$.

Even though expression (5.3) provides a closed solution for the estimation of z_{ic}, some problems arose when tests on simulated data were carried out. These problems are intrinsically related to the heavy tails of the product of exponential distributions, which results in values arbitrarily close to zero of the numerator of $z_{ic}^{(s+1)}$ in (5.3) for all c different from the true cluster. This way, whether one time series is equidistant from *all* the clusters, then there is always one cluster (the *nearest* cluster) receiving a membership value equals to 1. Apart from an unstable membership assignment, this behavior is not desirable at all in soft clustering, where one would expect membership degrees uniformly distributed over the clusters.

Let us see as a simple simulated experiment allows to illustrate graphically the mentioned problem, and simultaneously suggests a way of overcoming this hurdle. Consider a scenario with two clusters C_1 and C_2 formed by five series plus an equidistant time series. The series of C_1 and C_2 are generated from ARMA(1,1) structures with autoregressive parameters $\phi_1 = \theta_1 = 0.5$ and $\phi_2 = \theta_2 = -0.5$, respectively, while the equidistant series is a realization of Gaussian white noise. All the series have length $T = 5000$. Denote by Y^{eq} the centered log-periodograms for the equidistant series and by Y^{C_1} and Y^{C_2} the averages of the centered log-periodograms for the series in C_1 and C_2, respectively. Based on the true log-spectra $m^1(\cdot)$ and $m^2(\cdot)$ for the models defining C_1 and C_2, the errors $\varepsilon_k^{1,1} = Y_k^{C_1} - m^1(\lambda_k)$, $\varepsilon_k^{eq,1} = Y_k^{eq} - m^1(\lambda_k)$, $\varepsilon_k^{2,2} = Y_k^{C_2} - m^2(\lambda_k)$, and $\varepsilon_k^{eq,2} = Y_k^{eq} - m^2(\lambda_k)$ are calculated. Plots of density estimates for these error sequences are shown in Figure 5.1. Specifically, estimates for $\varepsilon_k^{1,1}$ and $\varepsilon_k^{eq,1}$ are given in Figure 5.1(a), and estimates for $\varepsilon_k^{2,2}$ and $\varepsilon_k^{eq,2}$ in Figure 5.1(b).

As expected, the estimated densities for $\varepsilon_k^{1,1}$ and $\varepsilon_k^{2,2}$ (black lines) correctly approximate the Gumble probability density $\varphi(\lambda)$, but the estimated densities for $\varepsilon_k^{eq,1}$ and $\varepsilon_k^{eq,2}$ (red

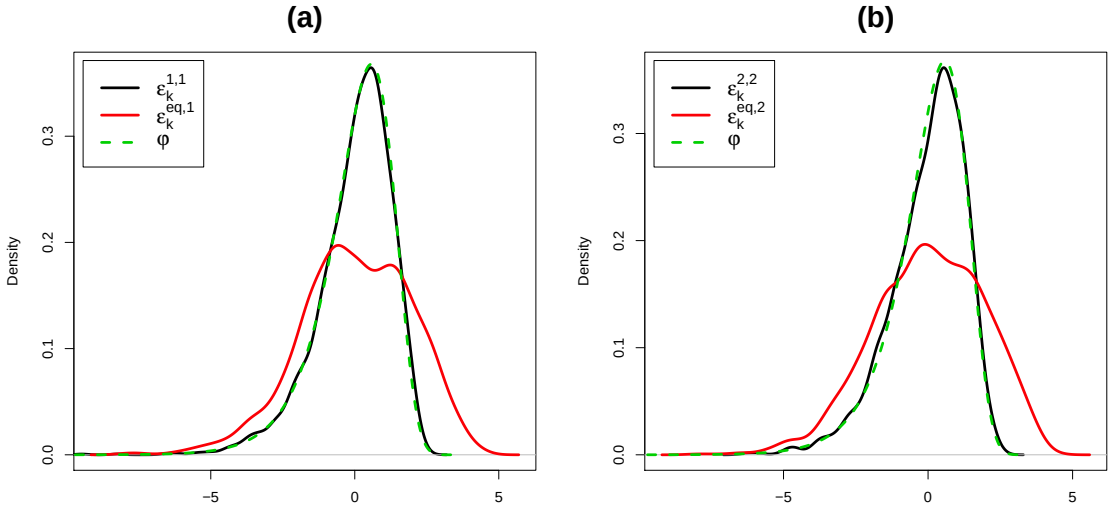

Figure 5.1: Density estimates of the errors for a series equidistant from two clusters (red lines) against density estimates of the errors for the two centroids (black lines) and the reference Gumbel density (green dashed lines).

lines) are fairly different from $\varphi(\lambda)$, presenting heavier tails. These tails produce the above mentioned effect of obtaining numerators of $z_{eq,1}^{(s+1)}$ and $z_{eq,2}^{(s+1)}$ very close to zero so that either $z_{eq,1}^{(s+1)}$ or $z_{eq,2}^{(s+1)}$ will take the value 1 when the ratio is calculated.

Figure 5.1 also suggests that computing the distance between a kernel density estimator based on the errors $Y_k^i - m^c(\lambda_k)$ and the density $\varphi(\lambda)$ provides a useful criterion to check how plausible is that the i-th series belongs to the cluster c. Notice that the red and green densities show similar distances in Figures 5.1(a) and (b), thus reproducing the equidistance from the two clusters.

Based on these comments, a new approach to estimate $\mathbb{P}(\Theta_s, Y|\boldsymbol{X}_t^{(i)} \in \text{group } c)$ is proposed below. For each series $\boldsymbol{X}_t^{(i)}$, $i = 1, \ldots, n$, compute kernel density estimates $\widetilde{\varphi}_c^i$ based on the errors $Y_k^i - m^c(\lambda_k)$, for $c = 1, \ldots, C$. Then, we define

$$\mathbb{P}(\Theta_s, Y|\boldsymbol{X}_t^{(i)} \in \text{group } c) = P_{ic} = \frac{1/\text{KLD}(\varphi, \widetilde{\varphi}_c^i)}{\sum_{c'=1}^{C} 1/\text{KLD}(\varphi, \widetilde{\varphi}_{c'}^i)}, \tag{5.4}$$

where $\text{KLD}(\cdot)$ denotes the Kullback-Leibler divergence between two probability distributions (Kullback and Leibler, 1951). Actually, KLD is not a metric. It is always nonnegative

and equals zero if and only if the two distributions are identical, but it is not symmetric and also it does not satisfy the triangle inequality. Nevertheless, this fact is not important here because the main concern is to measure the information lost when the estimated densities $\widetilde{\varphi}_c^i$ are used to approximate the reference density φ. In other words, the roles played by $\widetilde{\varphi}_c^i$ and φ are different. Anyway, any other distance between distributions could be used. Lastly note that the Kullback-Leibler divergence takes values between 0 and ∞ so that we adopt the criterion of setting $P_{ic} = 1$ if $\mathrm{KLD}(\varphi, \widetilde{\varphi}_c^i) = 0$ and $P_{ic} = 0$ when $\mathrm{KLD}(\varphi, \widetilde{\varphi}_c^i) = \infty$.

Once the P_{ic} are calculated, the posterior probabilities are defined by

$$z_{ic}^{(s+1)} = \frac{\pi_c P_{ic}}{\sum_{c'=1}^{C} \pi_{c'} P_{ic'}}. \tag{5.5}$$

M-step The M-step provides updated parameter estimates $\Theta_{(s+1)}$ by maximizing the expected complete local log-likelihood function with the values for the latent variables $z_{ic}^{(s+1)}$ obtained in the E-step. A regularly spaced grid of frequencies is selected for λ, $\lambda \in \{\gamma_1, \gamma_2, \ldots, \gamma_r\}$, and then the objective function has the form

$$\begin{aligned}
\ell(\Theta/Y, Z)(\lambda) &= \sum_{i=1}^{n}\sum_{c=1}^{C} z_{ic}^{(s+1)} \sum_{k=1}^{M} \log\{\pi_c\, \varphi\left(Y_k^i - m^c(\lambda)\right)\} K_h(\lambda_k - \lambda) \\
&= \sum_{i=1}^{n}\sum_{c=1}^{C} z_{ic}^{(s+1)} \left\{ \log\, \pi_c + \sum_{k=1}^{M} \log\left\{\varphi(Y_k^i - m^c(\lambda))\right\} K_h(\lambda_k - \lambda) \right\} \\
&= \underbrace{\sum_{i=1}^{n}\sum_{c=1}^{C} z_{ic}^{(s+1)} \log\, \pi_c}_{(A)} \\
&\quad + \underbrace{\sum_{i=1}^{n}\sum_{c=1}^{C} z_{ic}^{(s+1)} \sum_{k=1}^{M} \exp\left\{Y_k^i - m^c\left(\lambda\right) - \exp\left\{Y_k^i - m^c\left(\lambda\right)\right\}\right\} K_h(\lambda_k - \lambda),}_{(B)}
\end{aligned}$$

for $\lambda = \gamma_j$, $j = 1, \ldots, r$. In our numerical experiments, the Fourier frequencies λ_j have been chosen to constitute the frequency grid $\{\gamma_1, \gamma_2, \ldots, \gamma_r\}$, so that $r = M$.

Optimization is carried out by maximizing the terms A and B separately. Concerning the term A, optimization is made by using the Lagrange multiplier procedure. The constrained

optimization problem is given by

$$\max_{\pi} \sum_{i=1}^{n} \sum_{c=1}^{C} z_{ic}^{(s+1)} \log \pi_c, \quad \text{subject to } \sum_{c=1}^{C} \pi_c = 1, \quad \pi_c \geq 0 \quad \text{for } c = 1, \dots C,$$

so that the Lagrangian function takes the form

$$R(\pi, \beta) = \sum_{i=1}^{n} \sum_{c=1}^{C} z_{ic} \log \pi_c + \beta \left(\sum_{c=1}^{C} \pi_c - 1 \right),$$

where β denotes the unknown Lagrange multiplier. To obtain the critical points of $R(\pi, \beta)$, the system of simultaneous equations below involving the partial derivatives respect to π_c and β equal to zero must be solved.

$$\frac{\partial R}{\partial \pi_c} = \frac{1}{\pi_c} \sum_{i=1}^{n} z_{ic}^{(s+1)} + \beta = 0,$$
$$\frac{\partial R}{\partial \beta} = \sum_{c=1}^{C} \pi_c - 1 = 0.$$

Solutions are given by $\pi_c^{(s+1)} = -\frac{1}{\beta} \sum_{i=1}^{n} z_{ic}^{(s+1)}$ and $\widehat{\beta} = -\frac{1}{\sum_{c=}^{C} \sum_{i=1}^{n} z_{ic}^{(s+1)}}$, and therefore

$$\pi_c^{(s+1)} = \frac{\sum_{i=1}^{n} z_{ic}^{(s+1)}}{\sum_{c=1}^{C} \sum_{i=1}^{n} z_{ic}^{(s+1)}}. \tag{5.6}$$

On the other hand, maximization of the term B is directly calculated by setting to zero the first derivative with respect to $m^c(\lambda)$ and finding , resulting the estimators

$$\begin{aligned} m_c^{(s+1)}(\lambda) &= \log \left[\frac{\sum_{i=}^{n} z_{ic}^{(s+1)} \sum_{k=1}^{M} \exp\left(Y_k^i\right) K_h\left(\lambda_k - \lambda\right)}{\sum_{i=1}^{n} z_{ic}^{(s+1)} \sum_{k=1}^{M} K_h\left(\lambda_k - \lambda\right)} \right] \\ &= \log \left(\sum_{i=1}^{n} w_{ic}^{(s+1)} \widehat{f}^{i,(s+1)}(\lambda) \right), \end{aligned} \tag{5.7}$$

for $c = 1, \dots, C$ and λ in the selected grid, where $w_{ic}^{(s+1)} = z_{ic}^{(s+1)} / \sum_{i=1}^{n} z_{ic}^{(s+1)}$ and $\widehat{f}^{i,(s+1)}(\lambda)$ is the Nadaraya-Watson estimate of the spectrum with smoothing parameter

h and kernel K. Direct plug-in methodology was used to estimate the smoothing parameter h, as described by Ruppert et al. (1995).

It is worth to emphasize that the maximization of the complete local log-likelihood in the M-step leads to closed-form expressions to update centroids and prior probabilities, which results in lower computational complexity.

These two steps of the EM algorithm are iteratively applied until a stopping criterion is satisfied. Several options to determine this criterion may be selected. In this case, the stopping rule has been that the log-likelihood of data does not increase significantly, that is

$$\frac{\log L(\Theta_{s+1}, Y) - \log L(\Theta_s, Y)}{|\log L(\Theta_s, Y)|} < \epsilon$$

for some prefixed and sufficiently small value $\epsilon > 0$, or alternatively having reached a maximum number of iterations. Once the EM algorithm has converged, the values z_{ic}, for $c = 1, \ldots, C$, provide the sequence of membership degrees for the i-th time series, $i = 1, \ldots, n$. Indeed, the EM procedure requires initial values for the prior probabilities π_c and the centroids $m_c(\cdot)$, $c = 1, \ldots, C$. Our proposal is to run a hard PAM algorithm based on a suitable dissimilarity for time series, and then determining the initial structure for Θ using the resulting partition. Thus, π_c^0 is given by the ratio of time series located in the c-th group, and $m_c^0(\cdot)$ is determined by averaging the spectral smoothers for those series within the group c. Since the comparison between series is made in the frequency domain, it is reasonable to perform the PAM algorithm using a dissimilarity measure defined in this framework, e.g. the $d_{W(LS)}$ dissimilarity.

In summary, the mixture models EM algorithm is implemented as outlined in Algorithm 5.

Algorithm 5 Mixture models EM algorithm

1: Fix C, $\epsilon > 0$ and $max.iter$
2: Set $iter = 0$
3: Based on a partition generated with the PAM algorithm, determine an initial set of centers $m^1, \ldots, m^c$ and prior probabilities $\pi_1, \ldots, \pi_c$, i.e. Θ
4: **repeat**
5: Set $\Theta_{OLD} = \Theta$.
6: Compute z_{ic}, $i = 1, \ldots, n$, $c = 1, \ldots, C$, using (5.5) {E-step}
7: Update the values of Θ using (5.6) and (5.7) {M-step}
8: $iter \leftarrow iter + 1$
9: **until** $\dfrac{\log L(\Theta, Y) - \log L(\Theta_{OLD}, Y)}{|\log L(\Theta_{OLD}, Y)|} < \epsilon$ or $iter = max.iter$

5.3 Probabilistic D-clustering

Consider a set S of n realizations of univariate time series $\left\{\boldsymbol{X}_t^{(1)}, \ldots \boldsymbol{X}_t^{(n)}\right\}$ and a partition of S into C clusters $\mathcal{C}_1, \ldots, \mathcal{C}_C$, each one of the clusters being represented by a center $\mathbf{c}_1, \ldots, \mathbf{c}_C$. In probabilistic D-clustering, the clustering criterion is metric, which means that each series is assigned to the cluster represented by the nearest center. After the assignment is completed, the centers are recalculated and the series are reassigned according to the new centers. The algorithm keeps iterating until convergence is achieved.

Denote by $d(\boldsymbol{X}_t^{(i)}, \mathbf{c}_k)$ the distance of the series $\boldsymbol{X}_t^{(i)}$ to the center $\mathbf{c}_k$ and by $p_k\left(\boldsymbol{X}_t^{(i)}\right)$ the probability that the series $\boldsymbol{X}_t^{(i)}$ is a member of $\mathcal{C}_k$. There are several ways to model the relationship between distances and probabilities in the literature. A simple criterion was proposed by Ben-Israel and Iyigun (2008), which consists in assuming that

$$p_k\left(\boldsymbol{X}_t^{(i)}\right) d\left(\boldsymbol{X}_t^{(i)}, \mathbf{c}_k\right) = \text{constant}, \tag{5.8}$$

for each series $\boldsymbol{X}_t^{(i)} \in S$ and each center $\mathbf{c}_k$, $k = 1, \ldots, C$, where the constant in (5.8) depends on $\boldsymbol{X}_t^{(i)}$. Under this criterion, the probability that a series $\boldsymbol{X}_t^{(i)}$ belongs to a cluster $\mathcal{C}_k$ increases as the distance of the series to the center of the cluster decreases. From assumption (5.8), Ben-Israel and Iyigun (2008) easily proved that the probabilities $p_k\left(\boldsymbol{X}_t^{(i)}\right)$ can be written as

$$p_k\left(\boldsymbol{X}_t^{(i)}\right) = \frac{\prod_{j \neq k} d\left(\boldsymbol{X}_t^{(i)}, \mathbf{c}_j\right)}{\sum_{k'=1}^{C} \prod_{j \neq k'} d\left(\boldsymbol{X}_t^{(i)}, \mathbf{c}_j\right)}, \quad k = 1, \ldots, C.$$

At the end of the iterative process, the sequence of probabilities $p_k\left(\boldsymbol{X}_t^{(i)}\right)$, $k = 1, \ldots, C$, identifies a sequence of membership degrees for the i-th series $\boldsymbol{X}_t^{(i)}$, thus providing with a soft clustering partition.

The probabilistic D-clustering approach has been introduced in a general way for arbitrary data objects and using the squared Euclidean distance between data and centers. To our knowledge, it has not been considered in time series clustering. However, the working principle is versatile to the choice of the metric d, and therefore it can be easily adapted to deal with time series by selecting a suitable metric between series. In order to take advantage of the nice properties of the metric based on the estimated sequences of quantile autocovariances, d_{QAF}, we have implemented the probabilistic D-clustering algorithm based

on d_{QAF} to perform soft cluster analysis of time series, i.e. the membership probabilities are given by

$$p_k\left(\boldsymbol{X}_t^{(i)}\right) = \frac{\prod_{j\neq k} d_{QAF}\left(\boldsymbol{X}_t^{(i)}, \mathbf{c}_j\right)}{\sum_{k'=1}^{C}\prod_{j\neq k'} d_{QAF}\left(\boldsymbol{X}_t^{(i)}, \mathbf{c}_j\right)}, \quad k = 1, \ldots, C. \tag{5.9}$$

Note that d_{QAF} is simply the squared Euclidean distance between feature vectors of the data objects so that the optimality properties established by Ben-Israel and Iyigun (2008) hold in this new framework.

Ben-Israel and Iyigun (2008) obtained the iterative update of the centers by considering the minimization problem:

$$f\left(\mathbf{c}_1, \ldots, \mathbf{c}_C\right) = \sum_{i=1}^{n}\sum_{k=1}^{C} d_{QAF}\left(\boldsymbol{X}_t^{(i)}, \mathbf{c}_k\right) p_k\left(\boldsymbol{X}_t^{(i)}\right)^2, \tag{5.10}$$

proving that the minimizers are given by

$$\mathbf{c}_k = \sum_{i=1}^{n} \frac{v_k\left(\boldsymbol{X}_t^{(i)}\right)}{\sum_{j=1}^{n} v_k\left(\boldsymbol{X}_t^{(j)}\right)} \Gamma^{(i)}, \tag{5.11}$$

where $\Gamma^{(i)}$ are the estimated sequences of quantile autocovariances for the series $\boldsymbol{X}_t^{(i)}$ and

$$v_k(\boldsymbol{X}_t^{(i)}) = \frac{p_k(\boldsymbol{X}_t^{(i)})^2}{d_{QAF}(\boldsymbol{X}_t^{(i)}, \mathbf{c}_k),} \tag{5.12}$$

for $k = 1, \ldots, C$.

The QAF-based probabilistic D-clustering is implemented as outlined in Algorithm 6.

This algorithm has been implemented in R using the function *PDclust* in the package **FPDclustering**.

5.4 Simulation study

A simulation study was conducted to evaluate the performance of the proposed soft clustering alternatives. We intended to recreate fuzzy scenarios with different time series models, including realizations of linear and non-linear processes. Just like in Section 4.3, a base

Algorithm 6 QAF-based probabilistic D-clustering

Fix C, $\varepsilon > 0$ and $max.iter$
Set $iter = 0$
Pick an initial set of centers $\mathbf{c}_1, \ldots, \mathbf{c}_C$
repeat
    Set $\widetilde{\mathbf{c}}_k = \mathbf{c}_k$, $k = 1, \ldots, C$.
    Compute $p_k(X_t^{(i)})$, $i = 1, \ldots, n$, $k = 1, \ldots, C$, using (5.9)
    Update th centers $\mathbf{c}_k$, $k = 1, \ldots, C$ using (5.11) and (5.12)
    $iter \leftarrow iter + 1$
until $\sum_{k=1}^{C} \|\mathbf{c}_i - \widetilde{\mathbf{c}}_i\| < \varepsilon$ or $iter = max.iter$

scenario consisting of two clusters $\mathcal{C}_1$ and $\mathcal{C}_2$ with five series each was considered, and complexity was then increased by adding one additional realization located at equal distance from both clusters. Uncertainty was also added to the classification procedure by introducing variability over the parameters defining the underlying model for each cluster. The specific scenarios and the generation schemes for each scenario are properly specified in Table 5.1.

Table 5.1: Simulation scenarios for numerical comparison of the three different soft clustering procedures.

Generating process	Scenario	Elements and structure
Scenario 5.1: Soft clustering of ARMA(1,1) processes		
$X_t = \phi X_{t-1} + \theta\varepsilon_{t-1} + \varepsilon_t$	5.1.A	$\mathcal{C}_1$: 5 series with $\phi, \theta \sim U(0.4, 0.6)$ $\mathcal{C}_2$: 5 series with $\phi, \theta \sim U(-0.6. - 0.4)$
	5.1.B	Scenario 5.1.A plus one equidistant series with $\phi = \theta = 0$
Scenario 5.2: Soft clustering of non-linear moving average processes NLMA		
$X_t = \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-1}^2 + \varepsilon_t$	5.2.A	$\mathcal{C}_1$: 5 series with $\theta_1, \theta_2 \sim U(0.4, 0.6)$ $\mathcal{C}_2$: 5 series with $\theta_1, \theta_2 \sim U(-0.6. - 0.4)$
	5.2.B	Scenario 5.2.A plus one equidistant series with $\theta_1 = \theta_2 = 0$

In all cases, innovations ε_t follow a Gaussian distribution with zero mean and unit variance. Scenarios labeled with the letter A are formed by well-separated groups, while scenarios labeled with B are contaminated with a realization of Gaussian white noise, which is equidistant from both clusters. The five series generated from one specific cluster should group all together with membership degrees more markedly close to one in scenarios A. In scenarios B, the realization located at an intermediate place between $\mathcal{C}_1$ and $\mathcal{C}_2$, is expected to belong simultaneously to the two clusters thus showing membership degrees close to 0.5.

To bring insight into the capability of the mixture models algorithm to discriminate between the underlying processes, Figure 5.2 shows density kernel estimates based on the errors $Y_k^i - m^c(\lambda_k)$ for a series of each cluster and for the equidistant one, $m^c(\lambda_k)$ denoting the log-periodogram for the centroid of the c-th group, $c = 1, 2$. For the linear scenario, Figure 5.2(a), it is observed that the more far away the generating processes the more distant the density estimates from the theoretical density (ϕ). Thus, the equidistant realization always exhibits a density estimate (green curve) located into an intermediate situation regardless of the considered centroid. Similar conclusions are drawn for the non-linear scenario, Figure 5.2(b), only that in this case the classification is harder since the curves are closer to each other.

The probabilistic D-clustering and the mixture models algorithms were compared with the fuzzy QAF-FCMdC algorithm proposed in Chapter 4. Three quantiles of levels 0.1, 0.5 and 0.9 and only one lag ($L = 1$, with $l_1 = 1$) were considered to compute the fuzzy and the probabilistic D-clustering algorithms based on the QAF dissimilarity. The fuzziness parameter m for the implementation of QAF-FCMdC was set to $m = 2.5$. The experiments were carried out with different lengths for the time series, namely $T = 1000$ for Scenarios 5.1, and $T = 1500$ for Scenarios 5.2. The size of the series is increased for the non-linear scenario since it was seen in Figure 5.2 that is a much more complicated scenario.

The number of clusters was set at $C = 2$, and hence the equidistant series are forced to belong simultaneously to both clusters. At all scenarios, ten sets of 100 simulations were carried out. The means and standard deviations of the membership degrees averaged over the 10 sets were taken as measure of clustering accuracy of the algorithms.

The averages and standard deviations of the membership degrees obtained with the different models in the linear scenarios are shown in Tables 5.2 and 5.3. It is observed that in the baseline scenario with no equidistant series (Table 5.2), the ten series are always well-grouped, with probabilities greater than 0.94 to be assigned to the correct cluster for each algorithm. Similar results were obtained with the three algorithms although slightly better with QAF-FCMdC. The low standard deviations in all cases support the right performance of the algorithms. Attending to the scenario with the equidistant series (Table 5.3), the ten non-atypical series were again well-grouped with similar low standard deviations. The equidistant series is correctly detected with the three soft clustering algorithms by taking memberships close to 0.5, although in this case, the standard deviation were higher than the ones for the regular series.

Similar results were obtained in Scenarios 5.2.A (Table 5.4) and 5.2.B (Table 5.5) considering NLMA processes. Again QAF-FCMdC performed sligthly better, but the alternative

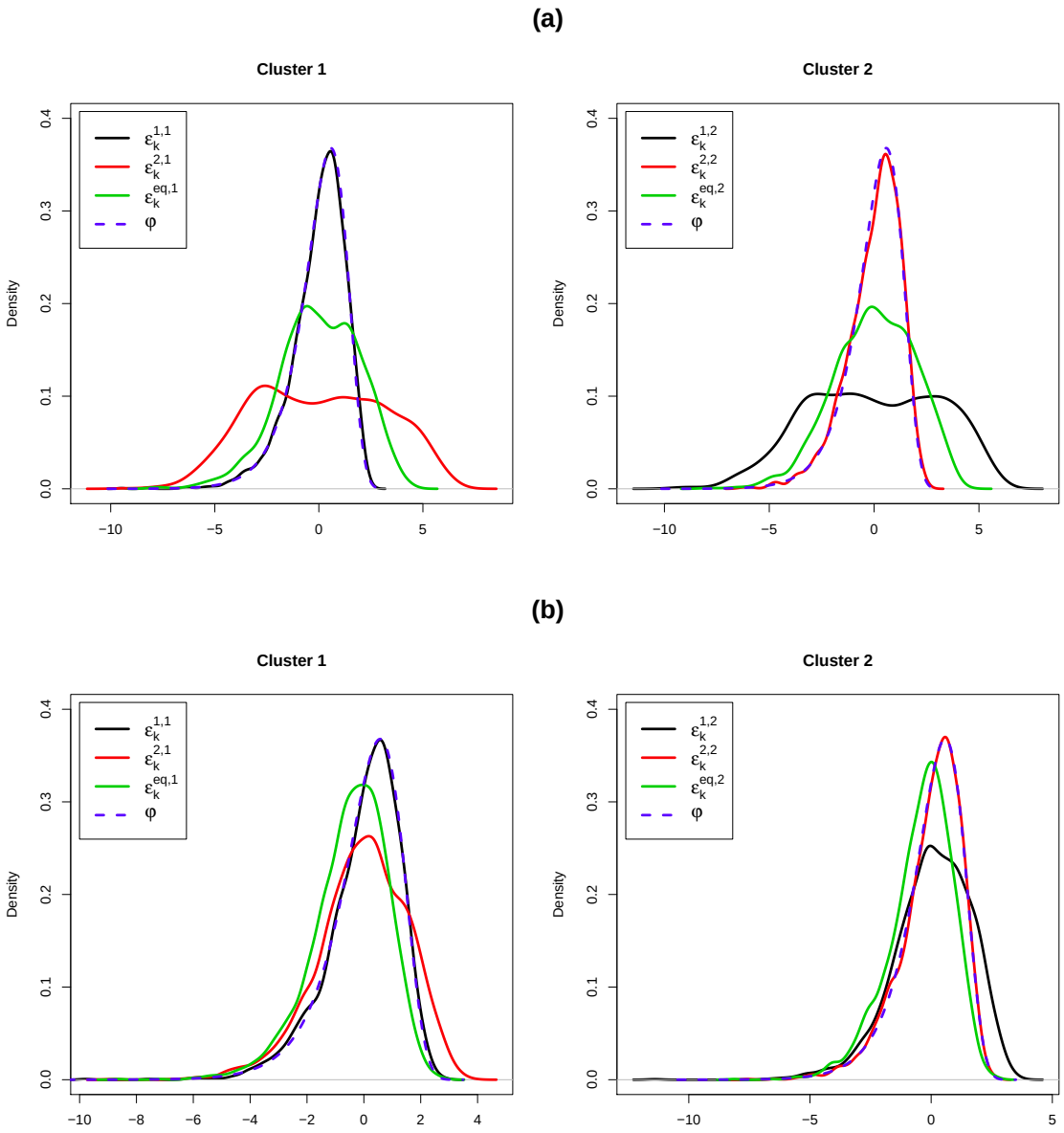

Figure 5.2: Density estimates of the errors for the centroids of clusters C_1 and C_2 (black and red lines respectively), the equidistant series (green lines) and the reference Gumbel density (blue dashed lines) for Scenarios 5.1.B (a) and 5.2.B (b).

proposals led to excellent scores as well. In this case, all the standard deviations are higher than in Scenarios 5.1, especially for the mixture models algorithm. This result is somehow expected given the error densities depicted in Figure 5.2 (b), which are fairly closer each other than in the linear case.

It is also important to make some consideration about the computational times for the

Table 5.2: Average percentage of correct classification in Scenario 5.1.A

	QAF-FCMdC		QAF-PDclust		MM-EM	
	C_1	C_2	C_1	C_2	C_1	C_2
Cluster 1						
X_1	0.987 (.010)	0.013 (.010)	0.968 (.017)	0.032 (.017)	0.951 (.062)	0.049 (.062)
X_2	0.987 (.010)	0.013 (.010)	0.967 (.018)	0.033 (.018)	0.951 (.051)	0.049 (.051)
X_3	0.987 (.010)	0.013 (.010)	0.967 (.018)	0.033 (.018)	0.949 (.061)	0.051 (.061)
X_4	0.987 (.010)	0.013 (.010)	0.967 (.018)	0.033 (.018)	0.950 (.060)	0.050 (.060)
X_5	0.987 (.010)	0.013 (.010)	0.967 (.018)	0.033 (.018)	0.952 (.057)	0.048 (.057)
Cluster 2						
X_6	0.014 (.011)	0.986 (.011)	0.033 (.019)	0.967 (.019)	0.051 (.052)	0.949 (.052)
x_7	0.014 (.011)	0.986 (.011)	0.033 (.019)	0.967 (.019)	0.051 (.060)	0.949 (.060)
x_8	0.013 (.010)	0.987 (.010)	0.033 (.017)	0.967 (.017)	0.051 (.065)	0.949 (.065)
x_9	0.014 (.011)	0.986 (.011)	0.033 (.019)	0.967 (.019)	0.048 (.058)	0.952 (.058)
x_{10}	0.013 (.010)	0.987 (.010)	0.033 (.017)	0.967 (.017)	0.046 (.044)	0.954 (.044)

Table 5.3: Average percentage of correct classification in Scenario 5.1.B

	QAF-FCMdC		QAF-PDclust		MM-EM	
	C_1	C_2	C_1	C_2	C_1	C_2
Cluster 1						
X_1	0.985 (.011)	0.015 (.011)	0.967 (.018)	0.033 (.018)	0.952 (.067)	0.048 (.067)
X_2	0.985 (.011)	0.015 (.011)	0.966 (.019)	0.034 (.019)	0.952 (.067)	0.048 (.067)
X_3	0.986 (.011)	0.014 (.011)	0.968 (.018)	0.032 (.018)	0.954 (.062)	0.046 (.062)
X_4	0.986 (.011)	0.014 (.011)	0.967 (.018)	0.033 (.018)	0.952 (.071)	0.048 (.071)
X_5	0.986 (.011)	0.014 (.011)	0.967 (.018)	0.033 (.018)	0.954 (.065)	0.046 (.065)
Cluster 2						
X_6	0.015 (.011)	0.985 (.011)	0.033 (.019)	0.967 (.019)	0.047 (.064)	0.953 (.064)
X_7	0.015 (.011)	0.985 (.011)	0.034 (.018)	0.966 (.018)	0.046 (.066)	0.954 (.066)
X_8	0.014 (.011)	0.986 (.011)	0.033 (.018)	0.967 (.018)	0.047 (.058)	0.953 (.058)
X_9	0.014 (.011)	0.986 (.011)	0.032 (.018)	0.968 (.018)	0.044 (.057)	0.956 (.057)
X_{10}	0.014 (.010)	0.986 (.010)	0.032 (.017)	0.968 (.017)	0.044 (.062)	0.956 (.062)
Equidistant						
$\mathcal{O}_1$	0.499 (.039)	0.501 (.039)	0.499 (.028)	0.501 (.028)	0.499 (.103)	0.501 (.103)

examined procedures. Even though the EM algorithm produces closed-form expressions for the estimates, the algorithm based on mixture models is expected to be computationally more complex due to it involves the estimation of the spectral density. To obtain accurate information about this point, the computing times at one arbitrary iteration of the simulation have been measured for the three algorithms. The algorithms were executed on a PC with the system specifications given by: Intel Core I7 - 3630QM processor, 2.4 Ghz CPU, 16 GB of RAM, Windows 10. Considering the linear scenario, the algorithm based on mixture models took 4.12 seconds in completing an iteration, while the QAF-FCMdC model took nearly 0.045 seconds and the QAF-PDclust 0.026 seconds.

Table 5.4: Average percentage of correct classification in Scenario 5.2.A

	QAF-FCMdC		QAF-PDclust		MM-EM	
	C_1	C_2	C_1	C_2	C_1	C_2
Cluster 1						
X_1	0.962 (.028)	0.038 (.028)	0.930 (.032)	0.070 (.032)	0.911 (.097)	0.089 (.097)
X_2	0.961 (.027)	0.039 (.027)	0.928 (.032)	0.072 (.032)	0.914 (.096)	0.086 (.096)
X_3	0.962 (.027)	0.038 (.027)	0.930 (.031)	0.070 (.031)	0.915 (.088)	0.085 (.088)
X_4	0.962 (.027)	0.038 (.027)	0.930 (.031)	0.070 (.031)	0.908 (.120)	0.092 (.120)
X_5	0.961 (.027)	0.039 (.027)	0.930 (.032)	0.070 (.032)	0.915 (.083)	0.085 (.083)
Cluster 2						
X_6	0.038 (.027)	0.962 (.027)	0.069 (.033)	0.931 (.033)	0.093 (.109)	0.907 (.109)
X_7	0.036 (.027)	0.964 (.027)	0.067 (.032)	0.933 (.032)	0.087 (.083)	0.913 (.083)
X_8	0.037 (.028)	0.963 (.028)	0.069 (.034)	0.931 (.034)	0.090 (.098)	0.910 (.098)
X_9	0.037 (.026)	0.963 (.026)	0.069 (.031)	0.931 (.031)	0.091 (.090)	0.909 (.090)
X_{10}	0.036 (.027)	0.964 (.027)	0.069 (.031)	0.931 (.031)	0.091 (.109)	0.909 (.109)

Table 5.5: Average percentage of correct classification in Scenario 5.2.B

	QAF-FCMdC		QAF-PDclust		MM-EM	
	C_1	C_2	C_1	C_2	C_1	C_2
Cluster 1						
X_1	0.962 (.027)	0.038 (.027)	0.929 (.032)	0.071 (.032)	0.899 (.108)	0.101 (.108)
X_2	0.962 (.028)	0.038 (.028)	0.930 (.032)	0.070 (.032)	0.897 (.110)	0.103 (.110)
X_3	0.960 (.028)	0.040 (.028)	0.928 (.033)	0.072 (.033)	0.897 (.119)	0.103 (.119)
X_4	0.961 (.025)	0.039 (.025)	0.929 (.030)	0.071 (.030)	0.905 (.089)	0.095 (.089)
X_5	0.961 (.026)	0.039 (.026)	0.928 (.031)	0.072 (.031)	0.897 (.104)	0.103 (.104)
Cluster 2						
X_6	0.037 (.026)	0.963 (.026)	0.069 (.031)	0.931 (.031)	0.097 (.106)	0.903 (.106)
X_7	0.038 (.028)	0.962 (.028)	0.071 (.033)	0.929 (.033)	0.100 (.116)	0.900 (.116)
X_8	0.039 (.027)	0.961 (.027)	0.071 (.033)	0.929 (.033)	0.095 (.095)	0.905 (.095)
X_9	0.038 (.028)	0.962 (.028)	0.070 (.033)	0.930 (.033)	0.099 (.107)	0.901 (.107)
X_{10}	0.040 (.027)	0.960 (.027)	0.071 (.033)	0.929 (.033)	0.094 (.092)	0.906 (.092)
Equidistant						
$\mathcal{O}_1$	0.515 (.060)	0.485 (.060)	0.511 (.045)	0.489 (.045)	0.502 (.074)	0.498 (.074)

5.5 Concluding remarks

Two different approaches to perform soft partitional clustering of times series have been introduced in this chapter and compared to the fuzzy model proposed in Chapter 4. Both of them consider paradigms broadly studied in soft cluster analysis of static data objects, namely cluster based on finite mixture models, where the mixture of underlying distributions is trained by the expectation-maximization (EM) algorithm, and the probabilistic-D clustering where the memberships are assumed to be inversely proportional to the distances from the cluster centroids. Nevertheless, these approaches have received much less atten-

tion in time series clustering in spite of exhibiting interesting properties, thus motivating the present study.

Our proposal of clustering algorithm considering the mixture models paradigm works in the frequency domain and relies on the idea that the log-periodogram ordinates admit a non-parametric regression model whose errors follow approximately a Gumbel distribution. This algorithm presents nice properties. Unlike other proposals in the literature, it is not limited to deal with AR(p) processes and takes advantage of the flexibility of the nonparametric regression to model complex shapes of spectral densities (including for example stationary non-linear models). Although the iterative determination of the membership degrees in the E-step requires a strategy different from the usual approach with normal mixtures to obtain a good clustering performance, it is noteworthy that the proposed EM algorithm produces closed-form solutions, which means feasible computational times. Beyond these properties, the proposed algorithm presents the properties inherent to the use of mixture models in clustering, including indeed that a soft partition is obtained without requiring to select a fuzziness parameter and a particular dissimilarity measure between time series.

As far as the probabilistic-D clustering, the key issue is to determine a proper metric between time series and our proposal consisted in using the Euclidean distance between sequences of estimated quantile autocovariances d_{QAF}, which has been introduced and studied in detail in Chapters 3 and 4. This selection is supported by the excellent results and robustness property showed by d_{QAF} in our experiments.

The performance of the new soft clustering algorithms was then examined via simulation and compared to the fuzzy QAF-FCMdC algorithm. Different scenarios including linear and non-linear models were considered and the assessment criterion took into account the capability of the algorithms to detect the fuzzy nature of series located between different clusters. Regardless of the considered models, all the algorithms drew out excellent results being always capable to detect the equidistant series, and it was observed that the fuzzy algorithm performed slightly better when compared to the other two procedures. It is noticeable that the probabilistic-D clustering, without needing to determine the level of fuzziness, led to results very close to the fuzzy algorithm, which again shows the high discriminatory power of the d_{QAF} dissimilarity. On the other hand, the algorithm based on mixed models is still more flexible by omitting the selection of a metric. In short, there is no an absolute winner algorithm because each of them exhibits different and valuable properties. In our opinion, the proposed algorithms could be used in a complementary way in order to help the users to check for the validity of the cluster partition.

Future work

This thesis has presented new approaches to perform hard and soft cluster analysis of time series in both frequency and time domains. The behavior of the proposed procedures has been carefully examined throughout extensive simulation studies involving a range of generated processes with different complexity levels. Compared to alternative clustering algorithms available in the literature, the new approaches have reported satisfactory results and have been shown to be useful in different applications. A summary of the main contributions of this research is given in Section 1.3 of Chapter 1. Nevertheless, there are indeed many interesting issues to be considered in further research. Some of these open lines are shortly pointed out below.

The metric based on estimated quantile autocovariances has shown a valuable robustness against the the underlying models, but consistency of the sample quantile autocovariances has been established assuming strictly stationary processes, which can be a constraint in practice. Although stationarity is a quite common requirement in time series clustering, introducing suitable approaches to encompass non-stationary models has great interest in applications, particularly when the series in study are not easy to be transformed or such transformation does not make sense.

On the other hand, notice that quantile autocovariances are well-defined for time series taking ordinal values. Therefore, it is worth analyzing the behavior of the proposed procedures in clustering of temporal sequences of ordinal data or mixed-type (metric-ordinal) data, which typically arises in social stratification and generally in social science (Hennig and Liao, 2013). Furthermore, the potential exhibited by d_{QAF} in clustering allows us to guess its usefulness to perform supervised classification of time series, and this point could be properly explored given the importance of this topic in applicatins.

Also related to the quantile autocovariance notion, although we have tackled the clustering task in the time domain, an alternative approach to be addressed in future works is to consider the frequency domain by using a distance between proper estimators of the *quantile*

spectral densities (Lee and Rao, 2012) defined by

$$G(x,y,\omega) = \frac{1}{2\pi}\sum_{l}\operatorname{cov}\left\{I(X_0 \le x), I(X_l \le y)\right\}\exp(il\omega).$$

The quantile spectral density $G(x,y,\omega)$ specifies the frequency decomposition of the quantile autocovariances so that it can be seen as the cross spectral density of the bivariate time series $(I(X_t \le x), I(X_t \le y))$. Likewise the quantile autocovariances, $G(x,y,\omega)$ provides all the information about the serial dependence structure but now from the spectral point of view. Different approaches to estimate the quantile spectral density considering L_1 and L_2 procedures and their asymptotic properties have been provided in several works (Lee and Rao, 2012; Hagemann, 2013; Li, 2014; Dette et al., 2014). In particular, Lee and Rao (2012) propose to check the equality of serial dependence of two stationary time series by using the test statistic given by

$$\mathcal{P}_T = \frac{1}{T}\sum_{k=1}^{T}\int\left|\widehat{G}_{1,T}(x,y,\omega_k) - \widehat{G}_{2,T}(x,y,\omega_k)\right|^2 dF(x)dF(y)$$

where ω_k denote the k-th Fourier frequency, $\widehat{G}_{1,T}$ and $\widehat{G}_{2,T}$ are the quantile spectral density estimators and F is any distribution function. This way, $\mathcal{P}_T$ statistic might be considered as an innovative spectral dissimilarity measure between two time series.

Another interesting issue to take into consideration consists in extending the fuzzy C-medoids model based on the QAF metric by taking different weights for each pair of quantile levels and lags. The purpose is to give a greater weight to those combinations contributing with much more discriminatory information. In this framework, the minimization problem takes the form

$$\begin{cases} \min \displaystyle\sum_{i=1}^{n}\sum_{c=1}^{C} u_{ic}^{m}\sum_{l=1}^{L}\left[\sum_{j=1}^{r}\sum_{j'=1}^{r}\left[w_{jj'}\left(\widehat{\gamma}_l^{(i)}(\tau_j,\tau_j') - \widehat{\gamma}_l^{(c)}(\tau_j,\tau_j')\right)\right]^2\right] \\ \text{subject to: } \displaystyle\sum_{c=1}^{C} u_{ic} = 1,\ u_{ic} \ge 0 \text{ and } \sum_{j=1}^{r}\sum_{j'=1}^{r} w_{jj'} = 1, \end{cases} \tag{5.13}$$

where $\widehat{\gamma}_l(\tau,\tau')$, with $(\tau,\tau') \in [0,1]^2$, are the sequences of estimated quantile autocovariances of lag l.

Appendix A

Resumen en castellano

En esta tesis, se introducen nuevos enfoques para realizar clustering de series temporales. La intención principal ha sido contribuir al avance del conocimiento sobre este importante tema proporcionando nuevas herramientas (por ejemplo, una métrica innovadora), pero también discutiendo y comparando diferentes estrategias metodológicas (paradigmas suaves (soft) y duros (hard), nuevos principios de agrupamiento, enfoques robustos y nuevos algoritmos diseñados para tratar con series de tiempo).

Esta sección tiene como objetivo enumerar las principales motivaciones detrás de esta tesis y también destacar las principales contribuciones.

Capítulo 1: Introducción

El clustering de series de tiempo tiene como objetivo dividir un conjunto de realizaciones parciales de series temporales en diferentes categorías o clusters. La partición se realiza de tal manera que series en el mismo cluster son más similares entre sí que las que están en diferentes clusters. Es un problema central en muchos campos y es hoy en día un área de investigación activa en una amplia gama de campos tales como finanzas y economía, medicina, ingeniería, física, reconocimiento de patrones, entre muchos otros. Estos argumentos explican el creciente interés en este tema que ha dado lugar a un gran número de contribuciones.

Una cuestión crucial en el análisis cluster de series de tiempo es determinar una medida adecuada para evaluar la disimilitud entre dos series de tiempo. A diferencia del cluster convencional con datos estáticos, las series temporales son intrínsecamente dinámicas, con estructuras de autocorrelación subyacentes y, por lo tanto, la búsqueda de similitudes debe ser gobernada por el comportamiento de la serie durante sus períodos de observación.

Aunque la selección de una métrica adecuada desempeña un papel clave, hay dificultades adicionales que deben abordarse en el clustering de series de tiempo. Por ejemplo, muchas aplicaciones de clustering en la vida real implican un gran número de series muy largas, es decir, uno se enfrenta a un problema de alta dimensionalidad. Por lo tanto, algoritmos que trabajen directamente sobre las serie podrían llegar a ser ineficientes, o simplemente inviables. Para superar el problema de la alta dimensionalidad, nos centraremos a lo largo de toda la tesis en un enfoque basado en características, donde los datos en bruto son reemplazados por un vector de menor dimensión formado por las características extraídas que representan la estructura dinámica de cada serie, obteniendo un ahorro significativo en el tiempo de cálculo. De esta manera, la disimilitud entre series temporales se mide en términos de la discrepancia entre esas representaciones.

Además, cuando se trabaja con algoritmos partitivos, el concepto de centroide es particularmente complejo. Como es bien sabido, los centroides son objetos representativos de los clusters y a veces el objetivo del proceso de clasificación es identificar estos prototipos en lugar de realizar una clasificación exacta. En el ámbito de las series de tiempo, un centroide determina un patrón temporal específico y es a menudo importante tener una visión de estos patrones para realizar predicciones o establecer diferencias entre comportamientos temporales. Sin embargo, se debe tener cuidado al definir correctamente el centroide cuando se trata con series temporales.

Otros puntos a considerar en el análisis cluster de series temporales están en efecto relacionados con la naturaleza de la serie en estudio, el propósito final de la clasificación y la complejidad computacional de los procedimientos empleados. Ciertamente, una distancia adecuada para tratar series generadas a partir de modelos lineales puede ser inapropiada para hacer frente a modelos no lineales, y un algoritmo de clúster diseñado para discriminar entre procesos estacionarios difícilmente será útil para agrupar series que muestren tendencias similares.

En resumen, el alto nivel de complejidad y particularidades asociadas a la clasificación de series de tiempo junto con su enorme interés en una amplia gama de aplicaciones, explican el gran foco de atracción que este tema ha tenido en las últimas décadas en investigación, principalmente en los campos de Estadística, Minería de Datos e Inteligencia Artificial. Por tanto, se han logrado avances significativos, pero sin duda el clustering de series de tiempo sigue siendo un área de investigación activa en la actualidad, con serios problemas y desafíos a abordar.

Capítulo 2: Clustering basado en frecuencias y amplitudes de picos espectrales

La principal motivación detrás de este capítulo viene de considerar un escenario de particular interés en el análisis de los fenómenos oscilatorios. En campos como la medicina, la biología y la economía (entre otros), a menudo se requiere agrupar registros oscilatorios temporales de tal manera que cada cluster reúna series con períodos de oscilación dominantes similares y también potencia similar en ese período dominante. De hecho, el marco natural para enfrentar este problema es el dominio de la frecuencia. Sin embargo, la mayoría de las métricas introducidas en el dominio de la frecuencia han sido diseñadas para comparar espectros totales estimados. Este no es el enfoque natural aquí. De hecho, dos series temporales podrían eventualmente exhibir los principales picos espectrales en la misma frecuencia y con amplitudes similares, pero con diferentes densidades espectrales. Motivados por este interés, nos centramos en el desarrollo de un algoritmo cluster con el objetivo de dividir las series temporales observadas en función de la ubicación de sus picos espectrales significativos. Más específicamente, en este capítulo se presenta un procedimiento cluster en dos etapas basado en la comparación de frecuencias y magnitudes asociadas a los picos espectrales más altos. En la primera etapa, la métrica entre cada par de series se evalúa en términos del p-valor asociado a un contraste bootstrap de igualdad de frecuencias donde se alcanzan los máximos espectrales (Timmer et al., 1999). Basado en la matriz de p-valores obtenida y siguiendo la técnica cluster propuesta por Maharaj (2000), se obtiene una primera partición del conjunto de series. La técnica propuesta por Maharaj procede de manera similar a una algoritmo jerárquico aglomerativo a partir de la matriz p-valores, pero solo agrupará aquellas series cuyos p-valores asociados sean mayores que un nivel de significación prefijado de antemano. En esta primera etapa, cada cluster agrupa las series que presentan el pico espectral más alto en frecuencias similares, pero estos picos pueden presentar magnitudes diferentes. Este hecho justifica una segunda etapa del algoritmo cluster dirigida a comprobar si las áreas bajo las densidades espectrales dentro de cada cluster difieren en un entorno local de la frecuencia pico. Esta tarea se lleva a cabo por separado para cada uno de los clusters generados en la primera etapa del proceso. Para cada grupo, se construye una nueva matriz de p-valores procedente de contrastar la igualdad de estas áreas locales y ésta se utiliza para aplicar de nuevo el procedimiento de agrupamiento jerárquico propuesto por Maharaj (2000), obteniendo así la partición definitiva. Este procedimiento podría aplicarse iterativamente para los siguientes picos espectrales significativos.

Las simulaciones realizadas muestran el buen comportamiento del procedimiento propuesto, pero es importante remarcar las limitaciones inherentes al método, particularmente su alta

complejidad computacional y la necesidad de introducir parámetros de entrada relevantes. La recomendación es considerar este enfoque sólo cuando el propósito sea dividir un conjunto de series de tiempo en grupos caracterizados por la ubicación de sus frecuencias pico espectrales. En un contexto más general donde el interés sea clasificar las series según los procesos subyacentes, otras métricas resultan más eficientes.

Capítulo 3: Clustering de series temporales basado en autocovarianzas cuantiles

La selección de una métrica adecuada entre series de tiempo según el propósito de agrupamiento es básica. Aunque se han propuesto muchas métricas para la clasificación de series con procesos generadores similares, la mayoría de ellos están restringidos a trabajar con modelos lineales. Como consecuencia de ello, la eficacia de la clasificación disminuye sustancialmente cuando estas métricas se utilizan para trabajar con estructuras de dependencia más complejas (por ejemplo, modelos no lineales o heterocedásticos). De hecho, este mal comportamiento se espera utilizando métricas basadas en modelos debido a la falta de especificación del modelo, pero muchas métricas basadas en características extraídas de las series también se comportan mal porque dichas características no son capaces de caracterizar adecuadamente las diferencias entre los procesos involucrados en el proceso de clasificación. Por lo tanto, la introducción de una métrica que exhiba una alta capacidad para hacer frente a un amplio tipo de procesos constituye un desafío en el análisis cluster de series de tiempo. La clasificación de modelos no lineales y, sobre todo, de modelos heterocedásticos es un tema de especial interés debido a la enorme importancia de estos modelos en muchos problemas ambientales y financieros. Con este propósito en mente, proponemos una métrica basada en características que compara secuencias de autocovarianzas cuantiles estimadas. Las autocovarianzas cuantiles proporcionan una visión mucho más rica de la dependencia de las series que otras características extraídas. Éstas abarcan muchas propiedades interesantes, incluyendo robustez frente a la inexistencia de momentos, trabajar de manera correcta con distribuciones marginales con colas pesadas, detección de características no lineales y cambios en formas condicionales, entre otros. En particular, los capítulos 3 y 4 desarrollan un extenso análisis de los procedimientos de clustering de series temporales basado en la comparación de las autocovarianzas cuantiles.

El concepto de autocovariancia cuantil se introduce en primer lugar en el Capítulo 3. Sus propiedades y capacidad para el clustering de series de tiempo se presentan y se discuten a través de ejemplos simples e ilustrativos. Se establece el comportamiento asintótico de las autocovarianzas cuantiles y se define formalmente una métrica entre dos series temporales

basada en la comparación de sus autocovarianzas cuantiles estimadas (QAF).

Proporcionamos resultados de simulación comparando esta nueva métrica con otras alternativas frecuentemente usadas en el análisis cluster de series de tiempo usando dos enfoques diferentes: un método jerárquico en el que cada observación comienza en su propio cluster y los pares de clusters se combinan a medida que se sube en la jerarquía, y un procedimiento de partición en torno a medoides (PAM) (Kaufman and Rousseeuw, 1990), que devuelve un subconjunto de series representativas de los clusters identificados (medoides). Los resultados obtenidos muestran el buen comportamiento de la métrica QAF en comparación con otras disimilitudes comúnmente utilizadas. En particular, muy buenos índices de clasificación se obtienen en la clasificación de procesos heterocedásticos, que se utilizan con frecuencia en indicadores económicos o financieros (Bauwens and Rombouts, 2007; Otranto, 2008; D'Urso et al., 2013a; Aielli and Caporin, 2014). Además, puesto que los modelos heterocedásticos gausianos no pueden capturar frecuentemente la asimetría y la leptokurtosis expuestas por algunas series temporales financieras, p.e. series de log-retornos de índices bursátiles (Lazar and Alexander, 2006; Kipkoech, 2014), se realizan simulaciones adicionales basadas en modelos heterocedásticos con errores no normales que logran resultados aún mejores.

Una cuestión importante en análisis cluster es obtener una estimación inicial del número de clusters subyacentes a la base de datos. Se propone abordar este problema mediante la adaptación de un algoritmo de remuestreo basado en predicción (llamado Clest) introducido por Dudoit and Fridlyand (2002). Clest tiene como objetivo seleccionar el número de clusters k que proporciona la evidencia más fuerte contra la hipótesis nula $H_0 : k = 1$. Para cada valor de k, Clest evalúa la cantidad de reproducibilidad, denotada por R_k, de la solución k-cluster combinando ideas de aprendizaje supervisado y no supervisado y luego examina si el valor de R_k es significativamente mayor que el esperado bajo la hipótesis nula. En el procedimiento original, el valor esperado para R_k bajo la hipótesis nula se aproxima mediante el remuestreo de una distribución uniforme multivariante. Sin embargo, esta suposición no es razonable cuando se consideran datos dependientes. Para solucionar este inconveniente, la suposición de uniformidad en H_0 se considera marginalmente para cada autocovarianza cuantil. El comportamiento de esta versión modificada del algoritmo de Clest y otros métodos existentes en la literatura se examina y compara mediante una nueva simulación, ibteniéndose que el algoritmo Clest produce buenas estimaciones del número de clusters y mostró el rendimiento más robusto.

Otra contribución importante se refiere a la selección óptima de los parámetros de entrada, es decir, establecer cuántas y qué combinaciones de retardos y niveles de cuantiles deben utilizarse para definir la métrica QAF con el fin de optimizar el proceso de clustering. Una modificación del algoritmo de selección de variables propuesto por Andrews and McNicholas

(2014) para clustering y clasificación nos permite abordar este problema. Sin embargo, vale la pena remarcar que el uso de un pequeño número de cuantiles con niveles de probabilidad regularmente espaciados es suficiente para alcanzar resultados satisfactorios.

Siguiendo la estructura general de cada capítulo, el Capítulo 3 también incluye la aplicación del método propuesto a un estudio específico que involucra series temporales financieras.

Capítulo 4: Clustering fuzzy de series temporales basado en autocovarianzas cuantiles. Enfoques robustos.

Este capítulo tiene como objetivo evaluar el comportamiento de la distancia basada en las autocovarianzas cuantiles estimadas (QAF) en clustering partitivo de series de tiempo considerando un enfoque fuzzy. De nuevo, suponemos que el objetivo es agrupar las series de acuerdo a sus estructuras de dependencia subyacentes, es decir, la similitud entre series se mide en términos de similitud entre los procesos generadores. El uso de una métrica robusta al proceso generador de la serie es necesario para lograr una solución cluster adecuada, y la distancia basada en QAF introducida en el capítulo anterior reportó resultados muy satisfactorios en clustering hard. Por lo tanto, la motivación es clara: se espera que un algoritmo de clustering fuzzy que considera esta métrica muestre un comportamiento adecuado. Además, en la segunda parte del presente capítulo se aborda también el problema de tratar datos fuzzy anómalos. Las series temporales anómalas pueden tener un efecto disruptivo sobre el proceso de clustering y, por tanto, el uso de modelos robustos de clustering fuzzy es de gran interés en la práctica.

La primera contribución en este capítulo consiste en introducir un nuevo procedimiento fuzzy para agrupar series temporales. Adoptamos un enfoque fuzzy C-medoides donde se considera que la métrica QAF para calcular las distancias entre series y medoides. De esta manera, el enfoque propuesto hereda las ventajas de los métodos fuzzy (flexibilidad para describir estructuras cluster complejas con clusters superpuestos), la técnica de partición e torno a medoides y la métrica basada en QAF (alta capacidad para discriminar entre una amplia gama de estructuras de dependencia). Una vez introducido el algoritmo fuzzy, su comportamiento se evalúa mediante un estudio de simulación. Los experimentos se centraron principalmente en la clasificación de modelos heterocedásticos, un escenario complejo pero frecuentemente realista al analizar indicadores financieros, industriales o ambientales, entre otros. Se examina la capacidad del modelo propuesto para clasificar modelos GARCH, y su comportamiento se evalúa enfrentándolo a dos algoritmos de clustering fuzzy que consideran disimilaridades basadas en modelos GARCH (D'Urso et al., 2013a) y, por lo tanto, específicamente diseñados para trabajar en el escenario simulado. El algoritmo clustering

fuzzy se aplica a dos bases de datos reales considerando datos de calidad del aire y retornos diarios de índices bursátiles para ilustrar su utilidad en la práctica.

La segunda contribución trata el problema de la detección y neutralización de valores atípicos. En general, la presencia de datos anómalos puede impedir identificar correctamente la estructura cluster subyacente, por lo que la introducción de métodos fuzzy robustos es un tema importante. En el marco de series de tiempo, una serie temporal se considera como un valor atípico cuando exhibe un comportamiento dinámico atípico, que difiere sustancialmente del resto de prototipos identificados. Para abordar este problema, se proponen tres diferentes extensiones de técnicas fuzzy robustas considerando la métrica basada en las autocovarianzas cuantiles. Específicamente, (i) clustering fuzzy C-medoides exponencial basado en la métrica QAF, (ii) clustering fuzzy C-medoides basado en la métrica QAF con cluster ruido y (iii) clustering fuzzy C-medoides truncado basado en la métrica QAF. El primer modelo utiliza una métrica robusta para neutralizar y suavizar el efecto de los valores atípicos, el segundo está enfocado en detectar los valores atípicos clasificándolos en un cluster ruido y con el tercer método el modelo logra su robustez truncando una cierta fracción de las series de tiempo más lejanas. Todos estos modelos son extensiones robustas del modelo de clustering fuzzy C-medoides basado en la métrica QAF, introducido en la primera parte del capítulo. Existen trabajos recientes que han seguido enfoques robustos análogos pero usando la distancia AR (D'Urso et al., 2013b, 2015b, 2017) y métricas que consideran modelos heteroscedásticos subyacentes (D'Urso et al., 2016). Para obtener información sobre la capacidad de los modelos robustos propuestos, todos estos procedimientos se compararon mediante un extenso estudio de simulación que incluía modelos ARMA y GARCH en presencia de valores atípicos. Obviamente, los procedimientos alternativos se aprovechan de estar específicamente construidos para discriminar entre estos procesos, y por lo tanto podemos obtener una medida realista de la capacidad de los procedimientos basados en la métrica QAF. La utilidad y la eficacia de los modelos fuzzy robustos propuestos se resalta también considerando una aplicación en el campo de las finanzas.

Capítulo 5: Clustering soft de series temporales: Nuevos enfoques basados en modelos mixtos y técnicas D-probabilísticas

Además del enfoque fuzzy, existen en la literatura otras técnicas alternativas para llevar a cabo clustering soft. Dos técnicas bien conocidas son el D-clusterig probabilístico (Ben-Israel and Iyigun, 2008) y el clustering basado en modelos mixtos (ver por ejemplo Bouveyron and Brunet-Saumard, 2014). Hasta donde sabemos, el primero no ha sido empleado para llevar a cabo análisis cluster de series temporales, y el segundo se ha aplicado

de una manera muy limitada. En concreto, sólo somos conscientes del trabajo de Chen and Maitra (2011) donde se propone un enfoque basado en modelos para la agrupación de datos de regresión en series temporales suponiendo que cada componente de la mixtura sigue un modelo de regresión autorregresivo gaussiano de orden p. Por lo tanto, la exploración de nuevos enfoques considerando D-clustering probabilístico y modelos mixtos para realizar clasificación de series de tiempo es de interés por varias razones. El clustering D-probabilístico es simple, requiere un pequeño número de iteraciones baratas y es insensible a valores extremos.

En el Capítulo 5 se proponen dos nuevos procedimientos basados en modelos mixtos y D-clustering probabilístico. El primero propone examinar el dominio de frecuencia y considerar la representación asintótica del log-periodograma por medio de un modelo de regresión no paramétrico con errores log-exponencialmente distribuidos. Suponiendo que las series temporales dentro de un mismo cluster se caracterizan por una densidad espectral específica, se puede definir una mistura finita no paramétrica de modelos de regresión univariante con una distribución de probabilidad conocida. La estimación del modelo mixto implica aproximaciones no paramétricas de los log-periodogramas para cada cluster y estimaciones de las probabilidades de pertenencia a los clusters. Para obtener estos estimadores, se lleva a cabo un procedimiento de estimación de verosimilitud local (Tibshirani and Hastie, 1987) mediante la implementación de un algoritmo EM (Dempster et al., 1977). Como es bien conocido, el algoritmo EM alterna entre dos etapas diferentes. En la $(s+1)$-ésima iteración, la etapa de expectación (E) calcula el valor esperado de las variables no observadas indicando las probabilidades de cada serie temporal de pertenecer a cada cluster (variables latentes), utilizando la distribución condicional de los valores actuales de los parámetros obtenidos al final de la iteración s-ésima. En la etapa de maximización (M), los centros de los clusters y las probabilidades a priori se calculan maximizando la log-verosimilitud esperada construida en la etapa E. El algoritmo itera hasta lograr la convergencia. En este caso, el paso E habitual requiere un criterio innovador para calcular las probabilidades a posteriori con el fin de alcanzar soluciones interpretables en el contexto del clustering soft. A diferencia de los enfoques fuzzy y de D-clustering probabilístico, el clustering basado en modelos mixtos no requiere fijar una métrica para medir la disimilitud entre series de tiempo y devuelve una partición soft sin especificar un parámetro de fuzziness como los procedimientos fuzzy.

El D-clustering probabilístico (Ben-Israel and Iyigun, 2008) se basa en la idea de que la probabilidad de pertenencia a un cluster en cualquier punto es inversamente proporcional a la distancia desde el centro del cluster en cuestión. Dado un objeto arbitrario x, el principio básico de este algoritmo es asumir que $d_k(x)p_k(x) = \text{cte}$ (dependiendo de x), para todo

cluster $\mathcal{C}$, donde $d_k(x)$ y $p_k(x)$ indican la distancia de x al centro de $\mathcal{C}$ y la probabilidad de que x sea un miembro de $\mathcal{C}$, respectivamente. De esta manera, cuanto más cerca del centro de un cluster, mayor es la probabilidad de pertenecer a ese cluster. A diferencia de los enfoques basados en modelos, seleccionar una métrica adecuada aquí es muy importante para obtener una partición satisfactoria. Los resultados en los Capítulos 3 y 4 apoyan la idea de usar la distancia basada en las autocovarianzas cuantiles en el algoritmo D-clustering probabilístico de series temporales. Esta intuición está plenamente confirmada en los experimentos numéricos realizados en este capítulo. Del mismo modo el enfoque de modelos mixtos, el algoritmo D-clustering probabilístico no necesita especificar un parámetro de fuzziness. Lo difusa que será la clasificación se determina automáticamente en términos de distancias a los diferentes centros de los clusters. De hecho, esta es una propiedad muy interesante, dada la notable influencia del parámetro de fuzziness observada en capítulos anteriores.

Bibliography

Aielli, G. P. and Caporin, M. (2013). Fast clustering of GARCH processes via gaussian mixture models. *Math. Comput. Simul.*, 94:205–222.

Aielli, G. P. and Caporin, M. (2014). Variance clustering improved dynamic conditional correlation mgarch estimators. *Comput. Stat. Data Anal.*, 76:556–576.

Alonso, A. M., Berrendero, J. R., Hernández, A., and Justel, A. (2006). Time series clustering based on forecast densities. *Comput. Stat. Data Anal.*, 51(2):762–776.

Amendola, A. and Francq, C. (2009). *Concepts and tools for Nonlinear Time-Series Modelling*, pages 377–427. John Wiley & Sons, Ltd.

An, H. Z. and Huang, F. C. (1996). The geometrical ergodicity of nonlinear autoregressive models. *Stat. Sin.*, 6(4):943–956.

Andrews, J. L. and McNicholas, P. D. (2014). Variable selection for clustering and classification. *J. Classif.*, 31(2):136–153.

Batista, G., Wang, X., and Keogh, E. (2011). A complexity-invariant distance measure for time series. In *SDM 11, Proceedings of the eleventh SIAM International Conference on Data Mining*, pages 699–710.

Bauwens, L. and Rombouts, J. V. K. (2007). Bayesian clustering of many Garch models. *Econom. Rev.*, 26(2-4):365–386.

Ben-Israel, A. and Iyigun, C. (2008). Probabilistic d-clustering. *J. Classif.*, 25(1):5.

Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *KDD Workshop*, pages 359–370.

Bohte, Z., Cepar, D., and Košmelj, K. (1980). Clustering of time series. In Barritt, M. and Wishart, D., editors, *Compstat 80, Proceedings in Computational Statistics*, pages 587–593, Vienna. Physica-Verlag, Heidelberg.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *J. Econom.*, 31(3):307–327.

Bouveyron, C. and Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: A review. *Comput. Stat. Data Anal.journal*, 71:52 – 78.

Brandmaier, E. M. (2012). *Permutation distribution clustering and structural equation model trees*. PhD thesis, Universitat des Saarlandes, Saarbruken, Germany.

Brillinger, D. R. (1981). *Time series analysis: data analysis and theory*. Holt, Rinehart and Winston.

Brockwell, P. and Davis, R. (2002). *Introduction to time series and forecasting*. Number v. 1. Springer.

Caiado, J. and Crato, N. (2007). A GARCH-based method for clustering of financial time series: International stock markets evidence. In Skiadas, C. H., editor, *Recent Advances in Stochastic Modeling and Data Analysis*, pages 542–551. World Scientific Publishing, Singapore.

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

Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Commun. Stat.-Simul. Comput.*, 3(1):1–27.

Campello, R. and Hruschka, E. (2006). A fuzzy extension of the sihouette width criterion for cluster analysis. *Fuzzy Sets Syst.*, 157(21):2858–2875.

Cannon, R. L., Davè, J. V., and Bezdek, J. C. (1986). Efficient implementation of the fuzzy c-means clustering algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(2):248–255.

Casado de Lucas, D. (2010). *Classification techniques for time series and functional data*. PhD thesis.

Chan, F., Fu, A., and Yu, C. (2003). Haar wavelets for efficient similarity search of time-series: With and without time warping. *Knowledge and Data Engineering, IEEE Transactions on*, 15(3):686–705.

Chan, K.-P. and Fu, A.-C. (1999). Efficient time series matching by wavelets. In *Data Engineering, 1999. Proceedings, 15th International Conference on*, pages 126–133.

Chen, C., So, M., and Liu, F.-C. (2011). A review of threshold time series models in finance. *Stat. Interface*, 4:167–181.

Chen, W. and Maitra, R. (2011). Model-based clustering of regression time series data via apecm—an aecm algorithm sung to an even faster beat. *Stat. Anal. Data Min.*, 4(6):567–578.

Chouakria, A. D. and Nagabhushan, P. N. (2007). Adaptive dissimilarity index for measuring time series proximity. *Adv. Data Anal. Classif.*, 1(1):5–21.

Cilibrasi, R. and Vitanyi, P. M. (2005). Clustering by compression. *IEEE Trans. Inf. Theor.*, 51(4):1523–1545.

Cimino, M., Frosini, G., Lazzerini, B., and Marcelloni, F. (2005). On the noise distance in robust fuzzy c-means. In *Proceedings of World Academy of Science, Engineering and Technology*, pages 361–364.

Corduas, M. and Piccolo, D. (2008). Time series clustering and classification by the autoregressive metric. *Comput. Stat. Data Anal.*, 52(4):1860–1872.

Dave, R. N. (1991). Characterization and detection of noise in clustering. *Pattern Recognit. Lett.*, 12(11):657 – 664.

Dave, R. N. and Sen, S. (1997). Noise clustering algorithm revisited. In *1997 Annual Meeting of the North American Fuzzy Information Processing Society - NAFIPS (Cat. No.97TH8297)*, pages 199–204.

Davis, R. A. and Mikosch, T. (2009). The extremogram: A correlogram for extreme events. *Bernoulli*, 15(4):977–1009.

de A.T. de Carvalho, F., Tenório, C. P., and Junior, N. L. C. (2006). Partitional fuzzy clustering methods based on adaptive quadratic distances. *Fuzzy Sets Syst.*, 157(21):2833 – 2857.

De Luca, G. and Zuccolotto, P. (2011). A tail dependence-based dissimilarity measure for financial time series clustering. *Adv. Data Anal. Classif.*, 5(4):323–340.

Dembélé, D. and Kastner, P. (2003). Fuzzy c-means method for clustering microarray data. *Bioinformatics*, 19(8):973–980.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *J. R. Stat. Soc. Series B Stat. Methodol.*, 39(1):1–38.

Dette, H., Hallin, M., Kley, T., and Volgushev, S. (2014). Of copulas, quantiles, ranks and spectra: An l_1-approach to spectral analysis. Arxiv e-prints. `arXiv:1111.7205v2`.

Döring, C., Lesot, M.-J., and Kruse, R. (2006). Data analysis with fuzzy clustering methods. *Comput. Stat. Data Anal*, 51(1):192 – 214.

Douzal-Chouakria, A., Diallo, A., and Giroud, F. (2009). Adaptive clustering for time series: Application for identifying cell cycle expressed genes. *Comput. Statist. Data Anal.*, 53(4):1414 – 1426.

Dudoit, S. and Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biol.*, 3(7):research0036.1–research0036.21.

D'Urso, P. (2015). Fuzzy clustering. In Hennig, C., Meila, M., Murtagh, F., and Rocci, R., editors, *Handbook of Cluster Analysis*, Handbooks of Modern Statistical Methods, pages 545–573. Chapman and Hall/CRC, London.

D'Urso, P., Cappelli, C., Lallo, D. D., and Massari, R. (2013a). Clustering of financial time series. *Physica A*, 392(9):2114–2129.

D'Urso, P., De Giovanni, L., and Massari, R. (2015a). Time series clustering by a robust autoregressive metric with application to air pollution. *Chemometrics Intell. Lab. Syst.*, 141:107–124.

D'Urso, P., De Giovanni, L., and Massari, R. (2016). Garch-based robust fuzzy clustering of time series. *Fuzzy Sets Syst.* (in press).

D'Urso, P. and Giordani, P. (2006). A weighted fuzzy c-means clustering model for fuzzy data. *Comput. Stat. Data Anal*, 50(6):1496 – 1523.

D'Urso, P. and Giovanni, L. D. (2014). Robust clustering of imprecise data. *Chemometrics Intell. Lab. Syst.*, 136:58–80.

D'Urso, P., Giovanni, L. D., and Massari, R. (2015b). Time series clustering by a robust autoregressive metric with application to air pollution. *Chemometr. Intell. Lab.*, 141(15):107–124.

D'Urso, P., Giovanni, L. D., Massari, R., and Lallo, D. D. (2013b). Noise fuzzy clustering of time series by the autoregressive metric. *Metron*, 71(3):217–-243.

D'Urso, P., Lallo, D. D., and Maharaj, E. A. (2013c). Autorregresive model-based fuzzy clustering and its application for detecting information redundancy in air pollution monitoring networks. *Soft Comput.*, 17(1):83–-131.

D'Urso, P. and Maharaj, E. A. (2009). Autocorrelation-based fuzzy clustering of time series. *Fuzzy Sets Syst.*, 160(24):3565–3589.

D'Urso, P. and Maharaj, E. A. (2012). Wavelets–based clustering of multivariate time series. *Fuzzy Sets Syst.*, 193:33–61.

D'Urso, P., Massari, R., Cappelli, C., and Giovanni, L. D. (2017). Autoregressive metric-based trimmed fuzzy clustering with an application to {PM10} time series. *Chemometr. Intell. Lab*, 161:15 – 26.

Engle, R. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007.

Estévez-Pérez, G. and Vilar, J. A. (2013). Functional anova starting from discrete data: an application to air quality data. *Environ. Ecol. Stat.*, 20(3):495–517.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*. Monographs on Statistics and Applied Probability. Chapman & Hall.

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

Fan, J. and Yao, Q. (2005). *Nonlinear time series: Nonparametric and parametric methods*. Springer Series in Statistics. Springer, New York.

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

Findley, L. and Koller, W. (1987). Essential tremor: A review. *Neurology*, 37(7):1194–1197. cited By 90.

Fowlkes, E. B. and Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *J. Am. Stat. Assoc.*, 78(383):553–569.

Fraley, C. and Raftery, A. (2002). Model-based clustering, discriminant analysis, and density estimation. *JASA*, 97(458):611–631.

Fu, T. C. (2011). A review on time series data mining. *Eng. Appl. Artif. Intell.*, 24(1):164–181.

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

García-Magariños, M. and Vilar, J. A. (2015). A framework for dissimilarity-based partitioning clustering of categorical time series. *Data Min. Knowl. Discov.*, 29(2):466–502.

Gavrilov, M., Anguelov, D., Indyk, P., and Motwani, R. (2000). Mining the stock market (extended abstract): Which measure is best? In *Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'00, pages 487–496, New York, USA. ACM.

Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *J. Finance*, 48(5):1779–1801.

Golay, X., Kollias, S., Stoll, G., Meier, D., Valavanis, A., and Boesiger, P. (2005). A new correlation-based fuzzy logic clustering algorithm for fmri. *Magn. Reson. Med.*, 40(2):249–260.

Granger, C. W. J. and Andersen, A. P. (1978). *An introduction to bilinear time series models*, volume 8 of *Angewandre Statistik und Okonometrie*. Gottingen: Vandenhoeck & Ruprecht.

Grimaldi, S. (2004). Linear parametric models applied to daily hydrological series. *J. Hydrol. Eng.*, 9(5):383–391.

Hagemann, A. (2013). Robust spectral analysis. Arxiv e-prints. `arXiv:1111.1965v1`.

Hall, L. O., Bensaid, A. M., Clarke, L. P., Velthuizen, R. P., Silbiger, M. S., and Bezdek, J. C. (1992). A comparison of neural network and fuzzy clustering techniques in segmenting magnetic resonance images of the brain. *IEEE Trans. Neural Netw.*, 3(5):672–682.

Han, H., Linton, O., Oka, T., and Whang, Y.-J. (2016). The cross-quantilogram: Measuring quantile dependence and testing directional predictability between time series. *J. Econom.*, 193(1):251–270.

Hartigan, J. A. (1975). *Clustering algorithms*. John Wiley & Sons, Inc., New York, NY, USA, 99th edition.

Hennig, C. and Liao, T. F. (2013). How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *Appl. Stat.-J. R. Stat. Soc.*, 62(3):309–369.

Hong, Y. (2000). Generalized spectral tests for serial dependence. *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, 62(3):557–574.

Höppner, F., Klawonn, F., Kruse, R., and Runkler, T. (1999). *Fuzzy cluster analysis: Methods for classification, data analysis and image recognition*. Wiley, Chichester, UK.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *J. Classif.*, 2(1):193–218.

Jarmasz, M. and Somorjai, R. L. (2002). Exploring regions of interest with cluster analysis (eroica) using a spectral peak statistic for selecting and testing the significance of fmri activation time-series. *Artif. Intell. Med.*, 25(1):45–67.

Kakizawa, Y., Shumway, R. H., and Taniguchi, M. (1998). Discrimination and clustering for multivariate time series. *J. Amer. Statist. Assoc.*, 93(441):328–340.

Kalpakis, K., Gada, D., and Puttagunta, V. (2001). Distance measures for effective clustering of arima time-series. In Cercone, N., Lin, T. Y., and Wu, X., editors, *Proceedings 2001 IEEE International Conference on Data Mining*, pages 273–280. IEEE Comput. Soc.

Kamdar, T. and Joshi, A. (2000). On creating adaptive web servers using weblog mining. Tr-cs-00-05, Departament of Computer Science and Electrical Engineering, University of Maryland, Baltimore County, USA.

Kao, S.-C., Ganguly, A. R., and Steinhaeuser, K. (2009). Motivating complex dependence structures in data mining: A case study with anomaly detection in climate. In Saygin, Y., Yu, J. X., Kargupta, H., Wang, W., Ranka, S., S.Yu, P., and Wu, X., editors, *2009 IEEE International Conference on Data Mining Workshops*, pages 223–230, Los Alamitos, CA, USA. IEEE Computer Society.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. Wiley series in probability and mathematical statistics. John Wiley & Sons, Inc., New York, 9th edition.

Keogh, E. and Kasetty, S. (2003). On the need for time series data mining benchmarks: A survey and empirical demonstration. *Data Min. Knowl. Discov.*, 7(4):349–371.

Keogh, E., Lonardi, S., and Ratanamahatana, C. A. (2004). Towards parameter-free data mining. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 206–215, New York, NY, USA. ACM.

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

Kim, J., Krishnapuram, R., and Davé, R. (1996). Application of the least trimmed squares technique to prototype-based clustering. *Pattern Recognit. Lett.*, 17(6):633 – 641.

Kipkoech, R. T. (2014). Modeling volatility under normal and student-t distributional assumptions (a case study of the kenyan exchange rates). *American Journal of Applied Mathematics and Statistics*, 2(4):179–184.

Klawonn, F. and Höppner, F. (2009). *Fuzzy cluster analysis from the viewpoint of robust statistics*, pages 439–455. Springer Berlin Heidelberg, Berlin, Heidelberg.

Koenker, R. (2005). *Quantile regression*. Econometric Society Monographs. Cambridge University Press.

Kovacic, Z. J. (1998). Classification of time series with applications to the leading indicator selection. In *Data Science, Classification and Related Methods - Proceedings of the Fifth Conference of the International Federation of Classification Societies (IFCS-96)*, pages 204–207, Kobe, Japan.

Krzanowski, W. J. and Lai, Y. T. (1988). A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, 44(1):23–34.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86.

Kwon, S. H. (1998). Cluster validity index for fuzzy clustering. *Electron. Lett.*, 34(22):2176–2177.

Lafuente-Rego, B. and Vilar, J. A. (2016a). Clustering of time series using quantile autocovariances. *Adv. Data Anal. Classif.*, 10(3):391–415.

Lafuente-Rego, B. and Vilar, J. A. (2016b). *Fuzzy Clustering of Series Using Quantile Autocovariances*, pages 49–64. Springer International Publishing, Cham.

Lazar, E. and Alexander, C. (2006). Normal mixture GARCH(1,1): Applications to exchange rate modelling. *J. Appl. Econometr.*, 21(3):307–336.

Lee, J. and Rao, S. (2012). The quantile spectral density and comparison based tests for nonlinear time series. Unpublished manuscript, Department of Statistics, Texas A&M University, College Station, U.S.A. `arXiv:1112.2759v2`.

Li, C., Biswas, G., Dale, M., and Dale, P. (2001). Building models of ecological dynamics using hmm based temporal data clustering - a preliminary study. In *Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis*, IDA '01, pages 53–62, London, UK, UK. Springer-Verlag.

Li, T.-H. (2014). Quantile periodograms. *J. Am. Stat. Assoc.*, 107(498):765–776.

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

Liao, T. W., Tang, F., Qu, J., and Blau, P. (2008). Grinding wheel condition monitoring with boosted minimum distance classifiers. *Mech. Syst. Signal Proc.*, 22(1):217–232.

Lin, J., Keogh, E., Lonardi, S., and Chiu, B. (2003). A symbolic representation of time series, with implications for streaming algorithms. In *In Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 2–11. ACM Press.

Linton, O. and Whang, Y.-J. (2007). The quantilogram: With an application to evaluating directional predictability. *J. Econom.*, 141(1):250–282.

Maharaj, E. A. (1996). A significance test for classifying ARMA models. *J. Statist. Comput. Simulation*, 54(4):305–331.

Maharaj, E. A. (2000). Cluster of time series. *J. Classif.*, 17(2):297–314.

Maharaj, E. A. and D'Urso, P. (2011). Fuzzy clustering of time series in the frequency domain. *Inf. Sci.*, 181(7):1187 – 1211.

Maharaj, E. A., D'Urso, P., and Galagedera, D. U. (2010). Wavelet-based fuzzy clustering of time series. *J. Classif.*, 27(2):231–275.

Melnykov, V. and Maitra, R. (2010). Finite mixture models and model-based clustering. *Statist. Surv.*, 4:80–116.

Mikosch, T. and Stărică, C. (2000). Limit theory for the sample autocorrelations and extremes of a GARCH(1,1) process. *Ann. Stat.*, 28(5):1427–1451.

Montero, P. and Vilar, J. A. (2014a). **TSclust**: An R package for time series clustering. *J. Stat. Softw.*, 62(1):1–43.

Montero, P. and Vilar, J. A. (2014b). **TSclust***: Time series clustering utilities.* R package version 1.2.1.

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A wew approach. *Econometrica*, 59(2):347–70.

Oates, T., Firoiu, L., and Cohen, P. (1999). Clustering time series with hidden markov models and dynamic time warping. In *Proceedings of the IJCAI-99 workshop on neural, symbolic and reinforcement learning methods for sequence learning*, pages 17–21.

Ohashi, Y. (1984). Fuzzy clustering and robust estimation. In *9th Meeting SAS Users Group Int.*

Otranto, E. (2008). Clustering heteroskedastic time series by model-based procedures. *Comput. Stat. Data Anal.*, 52(10):4685–4698.

Otranto, E. (2010). Identifying financial time series with similar dynamic conditional correlation. *Comput. Stat. Data Anal.*, 54(1):1–15.

Ozaki, T. (1980). Non-linear time series models for non-linear random vibrations. *J. Appl. Probab.*, 17(1):84–93.

Pal, N. R. and Bezdek, J. C. (1995). On cluster validity for the fuzzy c-means model. *IEEE Trans. Fuzzy Syst.*, 3(3):370–379.

Pértega, S. and Vilar, J. A. (2010). Comparing several parametric and nonparametric approaches to time series clustering: A simulation study. *J. Classif.*, 27(3):333–362.

Petruccelli, J. D. and Woolford, S. W. (1984). A threshold ar(1) model. *J. Appl. Probab.*, 21(2):270–286.

Pham, T. D. and Tran, L. T. (1981). On the first-order bilinear time series model. *J. Appl. Probab.*, 18(3):617–627.

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

Popivanov, I. and Miller, R. J. (2002). Similarity search over time-series data using wavelets. In *Proceedings 18th International Conference on Data Engineering*, pages 212–221.

Priestley, M. (1989). *Spectral analysis and time series.* Number v. 1-2 in Probability and mathematical statistics. Academic Press.

Ramoni, M., Sebastiani, P., and Cohen, P. (2002). Bayesian clustering by dynamics. *Mach. Learn.*, 47(1):91–121.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.*, 66(336):846–850.

Rio, E. (2000). *Théorie asymptotique des processus aléatoires faiblement dépendants*, volume 31 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer, Berlin.

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.*, 90(432):1257–1270.

Shephard, N. (1996). *Statistical aspects of ARCH and stochastic volatility*, pages 1–67. Chapman & Hall, London, (edited by d.r. cox, david v. hinkley and ole e. barndorff-neilsen) edition.

Shumway, R. H. and Stoffer, D. S. (2006). *Time series analysis and its applications : with R examples*. Springer texts in statistics. Springer, New York.

Skaug, H. J. and Tjøstheim, D. (1993). A nonparametric test of serial independence based on the empirical distribution function. *Biometrika*, 80(3):591–602.

Struzik, Z. R. and Siebes, A. (1999). *The Haar wavelet transform in the time series similarity paradigm*, pages 12–22. Springer Berlin Heidelberg, Berlin, Heidelberg.

Taylor, S. (1986). *Modelling time series*. John Wiley & Sons.

Tibshirani, R. and Hastie, T. (1987). Local likelihood estimation. *J. Amer. Statist. Assoc.*, 82(398):559–567.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, 63:411–423.

Timmer, J., Lauk, M., Vach, W., and Lücking, C. (1999). A test for a difference between spectral peak frequencies. *Comput. Stat. Data Anal.*, 30:45–55.

Tjostheim, D. (1990). Non-linear time series and markov chains. *Adv. Appl. Probab.*, 22(3):587–611.

Tong, H. (1993). *Non-linear time series: A dynamical system approach*. Dynamical System Approach. Clarendon Press.

Tong, H. and Yeung, I. (1991). On tests for self-exciting threshold autoregressive-type non-linearity in partially observed time series. *J. R. Stat. Soc. Ser. C-Appl. Stat.*, 40(1):43–62.

Tseng, C.-E., Peng, C.-Y., Chang, M.-W., Yen, J.-Y., Lee, C.-K., and Huang, T.-S. (2010). Novel approach to fuzzy-wavelet ECG signal analysis for a mobile device. *J. Med. Syst.*, 34(1):71–81.

Vilar, J. A., Alonso, A. M., and Vilar, J. M. (2010). Non-linear time series clustering based on non-parametric forecast densities. *Comput. Statist. Data Anal.*, 54(11):2850–2865.

Vilar, J. A., Lafuente-Rego, B., and D'Urso, P. (2017). Quantile autocovariances: A powerful tool for hard and soft partitional clustering of time series. *Fuzzy Sets Syst.*

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

Vilar, J. A., Vilar, J. A., and Vilar, J. M. (2013). Time series clustering based on nonparametric multidimensional forecast densities. *Electron. J. Stat.*, 7:1019–1046.

Vilar, J. M., Vilar, J. A., and Pértega, S. (2009). Classifying time series data: A nonparametric approach. *J. Classif.*, 26(1):3–28.

Winkler, R., Klawonn, F., and Kruse, R. (2011). Fuzzy clustering with polynomial fuzzifier function in connection with m-estimators. *Appl. Comput. Math.*, 10:146–163.

Wong, C. S. and Li, W. K. (2000). On a mixture autoregressive model. *J. R. Stat. Soc. Series B Stat. Methodol.*, 62(1):95–115.

Wu, K.-L. and Yang, M.-S. (2002). Alternative c-means clustering algorithms. *Patern Recogn.*, 35:2267–2278.

Xie, X. L. and Beni, G. (1991). A validity measure for fuzzy clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(8):841–847.

Yang, M.-S., Wu, K.-L., Hsieh, J.-N., and Yu, J. (2008). Alpha-cut implemented fuzzy clustering algorithms and switching regressions. *IEEE Trans. Sys. Man Cyber. Part B*, 38(3):588–603.

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):305–319.