engineering proceedings

MDPI

Proceeding Paper

Robust Methods for Soft Clustering of Multidimensional Time Series †

Ángel López-Oriona [1,*], Pierpaolo D'Urso [2], José A. Vilar [1,3] and Borja Lafuente-Rego [1]

1 Research Group MODES, Research Center for Information and Communication Technologies (CITIC), University of A Coruña, 15071 A Coruña, Spain; jose.vilarf@udc.es (J.A.V.); borja.lafuente@udc.es (B.L.-R.)
2 Department of Economics, Sapienza University of Rome, Piazzale Aldo Moro 5, 00185 Rome, Italy; pierpaolo.durso@uniroma1.it
3 Technological Institute for Industrial Mathematics (ITMATI), 15782 Santiago de Compostela, Spain
* Correspondence: oriona38@hotmail.com
† Presented at the 4th XoveTIC Conference, A Coruña, Spain, 7–8 October 2021.

Abstract: Three robust algorithms for clustering multidimensional time series from the perspective of underlying processes are proposed. The methods are robust extensions of a fuzzy C-means model based on estimates of the quantile cross-spectral density. Robustness to the presence of anomalous elements is achieved by using the so-called metric, noise and trimmed approaches. Analyses from a wide simulation study indicate that the algorithms are substantially effective in coping with the presence of outlying series, clearly outperforming alternative procedures. The usefulness of the suggested methods is also highlighted by means of a specific application.

Keywords: multidimensional time series; fuzzy C-means; unsupervised learning

Citation: López-Oriona, Á.; D'Urso, P.; Vilar, J.A.; Lafuente-Rego, B. Robust Methods for Soft Clustering of Multidimensional Time Series. *Eng. Proc.* **2021**, *7*, 60. https://doi.org/10.3390/engproc2021007060

Academic Editors: Joaquim de Moura, Marco A. González, Javier Pereira and Manuel G. Penedo

Published: 12 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Clustering of time series is a pivotal problem in statistics with several applications [1,2]. Generally, the goal is to divide collection of unlabelled time series into uniform groups so that intra-cluster similarity is maximized wheres the inter-cluster similarity is minimized. Most of the current techniques deal with univariate time series (UTS), while clustering of multidimensional time series (MTS) has received limited attention. This paper proposes three robust clustering methods for MTS. All of them are aimed at neutralizing the effect of outlying series while detecting the underlying grouping structure.

2. Robust Clustering Methods for Multivariate Time Series

Let $\{\boldsymbol{X}_t,\, t \in \mathbb{Z}\} = \{(X_{t,1},\ldots,X_{t,d}),\, t \in \mathbb{Z}\}$ be a d-variate real-valued strictly stationary stochastic process. Let F_j the marginal distribution function of $X_{t,j}$, $j = 1,\ldots,d$, and let $q_j(\tau) = F_j^{-1}(\tau)$, $\tau \in [0,1]$, 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 $I\{X_{t,j_1} \leq q_{j_1}(\tau)\}$ and $I\left\{X_{t+l,j_2} \leq q_{j_2}(\tau')\right\}$

$$\gamma_{j_1,j_2}(l,\tau,\tau') = \mathrm{Cov}\left(I\{X_{t,j_1} \leq q_{j_1}(\tau)\}, I\left\{X_{t+l,j_2} \leq q_{j_2}(\tau')\right\}\right), \tag{1}$$

for $1 \leq j_1, j_2 \leq d$. Taking $j_1 = j_2 = j$, the function $\gamma_{j,j}(l,\tau,\tau')$, with $(\tau,\tau') \in [0,1]^2$, so-called quantile autocovariance function (QAF) of lag l, generalizes the traditional autocovariance function.

For the multivariate process $\{\boldsymbol{X}_t,\, t \in \mathbb{Z}\}$, we can consider the $d \times d$ matrix $\boldsymbol{\Gamma}(l,\tau,\tau') = \left(\gamma_{j_1,j_2}(l,\tau,\tau')\right)_{1\leq j_1,j_2\leq d}$, which simultaneously gives information about both the cross-dependence (when $j_1 \neq j_2$) and the serial dependence (since there is a lag l).

Under appropriate summability conditions (mixing conditions), we can define the the Fourier transform of the cross-covariances. In this regards, the *quantile cross-spectral density* is given by

$$\mathfrak{f}_{j_1,j_2}(\omega,\tau,\tau') = (1/2\pi)\sum_{l=-\infty}^{\infty}\gamma_{j_1,j_2}(l,\tau,\tau')e^{-il\omega}, \quad (2)$$

for $1 \leq j_1, j_2 \leq d$, $\omega \in \mathbb{R}$ and $\tau, \tau' \in [0,1]$. Note that $\mathfrak{f}_{j_1,j_2}(\omega,\tau,\tau')$ is complex-valued.

The quantile cross-spectral density contains information about the general dependence patterns of a given stochastic process. For a specific realization of the process, this quantity can be consistently estimated by means of the so-called smoothed CCR-periodogram, $\hat{G}^{j_1,j_2}_{T,R}(\omega,\tau,\tau')$, proposed by [3].

Based on previous remarks, a simple dissimilarity measure between two realizations of the d-variate process (MTS) can be defined as follows. Given the i-th MTS, $\boldsymbol{X}_t^{(i)}$, consider the set $G^{(i)} = \{\hat{G}^{j_1,j_2}_{T,R}(\omega,\tau,\tau'), j_1, j_2 = 1,\ldots,d, \omega \in \Omega, \tau,\tau' \in \mathcal{T}\}$, where Ω is the set of Fourier frequencies and $\mathcal{T} = \{0.1, 0.5, 0.9\}$. Let $\boldsymbol{\Psi}^{(i)}$ be the vector formed by concatenating the elements of the set $G^{(i)}$. The dissimilarity measure between the series $\boldsymbol{X}_t^{(1)}$ and $\boldsymbol{X}_t^{(2)}$ is defined as the Euclidean distance between the complex vectors $\boldsymbol{\Psi}^{(1)}$ and $\boldsymbol{\Psi}^{(2)}$. We call this dissimilarity d_{QCD}.

The dissimilarity d_{QCD} is used to develop three robust fuzzy clustering methods. All of them assume that we want to group n MTS into C clusters, and are based on the traditional fuzzy C-means clustering algorithm. They look for the set of centroids $\overline{\boldsymbol{\Psi}} = \{\overline{\boldsymbol{\Psi}}^{(1)},\ldots,\overline{\boldsymbol{\Psi}}^{(C)}\}$, and the $n \times C$ matrix of fuzzy coefficients, $\boldsymbol{U} = (u_{ic})$, $i = 1,\ldots,n$, $c = 1,\ldots,C$, which define the solution of a given minimization problem. The quantity u_{ic} represents the membership degree of the i-th MTS in the c-th cluster. The minimization problem for the first method is the following:

$$\min_{\overline{\boldsymbol{\Psi}},\boldsymbol{U}} \sum_{i=1}^{n}\sum_{c=1}^{C} u_{ic}^{m}\left[1 - \exp\left\{-\beta\left\|\boldsymbol{\Psi}^{(i)} - \overline{\boldsymbol{\Psi}}^{(c)}\right\|_2^2\right\}\right] \text{ w.r.t } \sum_{c=1}^{C} u_{ic} = 1 \text{ and } u_{ic} \geq 0,$$

where β is an hyperparameter that needs to be set in advance and m is a parameter which determines the fuzziness of the partition, frequently called the fuziness parameter.

The exponential distance is used in the previous model because it is capable of neutralizing the effect of outlying series by spreading out their membership degrees between the different clusters [4].

The second robust procedure follows the noise cluster approach, and takes into account the following minimization problem:

$$\min_{\overline{\boldsymbol{\Psi}},\boldsymbol{U}} \sum_{i=1}^{n}\sum_{c=1}^{C-1} u_{ic}^{m}\left\|\boldsymbol{\Psi}^{(i)} - \overline{\boldsymbol{\Psi}}^{(c)}\right\|_2^2 + \sum_{i=1}^{n}\delta^2\left(1 - \sum_{c=1}^{C-1} u_{ic}\right)^m \text{ w.r.t. } \sum_{c=1}^{C} u_{ic} = 1 \text{ and } u_{ic} \geq 0,$$

where $\delta > 0$ is the a parameter known as the noise distance, which has to be specified in advance.

The previous model includes C groups, but only $(C-1)$ are "real" clusters. The noise cluster is artificially created for outlier identification purposes. The aim is to locate the outliers and place them in the noise cluster, which is represented by a fictitious prototype that has a constant distance from every MTS (the noise distance, δ).

The third technique can be expressed by means of the minimization problem:

$$\min_{\boldsymbol{Y},\boldsymbol{U}} \sum_{i=1}^{H(\alpha)}\sum_{c=1}^{C} u_{ic}^{m}\left\|\boldsymbol{\Psi}^{(i)} - \overline{\boldsymbol{\Psi}}^{(c)}\right\|^2 \text{ w.r.t. } \sum_{c=1}^{C} u_{ic} = 1 \text{ and } u_{ic} \geq 0.$$

where $\boldsymbol{Y}$ ranges on all the subsets of $\boldsymbol{\Psi} = \{\boldsymbol{\Psi}^{(1)}, \ldots, \boldsymbol{\Psi}^{(n)}\}$ of size $H(\alpha) = \lfloor n(1-\alpha) \rfloor$. The model attains its robustness by removing a certain proportion of the series and requires the specification of the fraction α of the data to be trimmed.

The three previously presented robust models have been analysed by means of a broad simulation study containing a wide variety of generating processes. Two alternative dissimilarities were taken into account for comparison purposes [5,6]. In all cases, the three proposed algorithms outperformed the competitors.

3. Application to real data

The three techniques proposed in Section 2 were applied to perform clustering in a real MTS database. Specifically, we considered daily stock returns and trading volume of the top 20 companies of the S&P 500 index, thus obtaining 20 bivariate MTS. Table 1 shows the membership degrees of the series concerning the trimmed approach.

Table 1. Membership degrees for the top 20 companies in the S&P 500 index by considering the trimmed approach and a 6-cluster partition.

Company	C_1	C_2	C_3	C_4	C_5	C_6
AAPL	0.083	0.146	0.299	**0.365**	0.066	0.041
MSFT	0.107	0.049	0.213	**0.356**	0.099	0.176
AMZN	**0.865**	0.017	0.051	0.032	0.010	0.025
GOOGL	**0.682**	0.032	0.092	0.128	0.025	0.040
GOOG	**0.902**	0.010	0.031	0.028	0.008	0.022
FB	0.002	**0.983**	0.006	0.004	0.003	0.002
TSLA	0.023	0.012	0.056	**0.885**	0.013	0.010
BRK.B	-	-	-	-	-	-
V	0.004	0.014	0.015	0.017	**0.941**	0.009
JNJ	0.004	0.015	0.019	0.013	**0.937**	0.013
WMT	-	-	-	-	-	-
JPM	0.002	0.001	0.003	0.003	0.002	**0.989**
MA	0.005	0.006	**0.968**	0.010	0.005	0.006
PG	0.015	0.012	0.028	0.016	0.019	**0.909**
UNH	0.006	**0.924**	0.026	0.013	0.022	0.008
DIS	0.020	0.038	**0.772**	0.099	0.042	0.030
NVDA	0.025	0.020	0.085	**0.804**	0.043	0.024
HD	-	-	-	-	-	-
PYPL	0.155	**0.301**	0.297	0.115	0.057	0.075
BAC	0.076	0.086	0.225	0.067	0.060	**0.485**

The symbols in bold correspond to the companies which were trimmed away, Berkshire Hathaway (BRK.B), Walmart (WMT) and Home Depot (HD). Similar clustering solutions were obtained with the remaining two methods.

4. Conclusions

This work proposes three robust methods to perform fuzzy clustering of MTS. They are based on the so-called exponential, noise and trimmed ideas. Each approach attains robustness to outlying series in a different way. The three procedures have been presented and assessed through a wide simulation study, substantially outperforming alternative approaches. A real data application has been also carried out in order to show the usefulness of the presented techniques.

Acknowledgments: This research has been supported by MINECO (MTM2017-82724-R and PID2020-113578RB-100), the Xunta de Galicia (ED431C-2020-14), and "CITIC" (ED431G 2019/01).

References

1. Liao, T.W. Clustering of time series data—A survey. *Pattern Recognit.* **2005**, *38*, 1857–1874. [CrossRef]
2. Aghabozorgi, S.; Shirkhorshidi, A.S.; Wah, T.Y. Time-series clustering—A decade review. *Inf. Syst.* **2015**, *53*, 16–38. [CrossRef]

3. Baruník, J.; Kley, T. Quantile coherency: A general measure for dependence between cyclical economic variables. *Econom. J.* **2019**, *22*, 131–152. [CrossRef]
4. Wu, K.L.; Yang, M.S. Alternative c-means clustering algorithms. *Pattern Recognit.* **2002**, *35*, 2267–2278. [CrossRef]
5. D'Urso, P.; Maharaj, E.A. Autocorrelation-based fuzzy clustering of time series. *Fuzzy Sets Syst.* **2009**, *160*, 3565–3589. [CrossRef]
6. D'Urso, P.; Maharaj, E.A. Wavelets-based clustering of multivariate time series. *Fuzzy Sets Syst.* **2012**, *193*, 33–61. [CrossRef]