Advances in Data Analysis and Classification (2020) 14:117–146
https://doi.org/10.1007/s11634-019-00365-8

REGULAR ARTICLE

A fragmented-periodogram approach for clustering big data time series

Jorge Caiado[1] · Nuno Crato[1,2] · Pilar Poncela[2,3]

Received: 6 September 2018 / Revised: 2 June 2019 / Accepted: 5 June 2019 / Published online: 14 June 2019

Abstract

We propose and study a new frequency-domain procedure for characterizing and comparing large sets of long time series. Instead of using all the information available from data, which would be computationally very expensive, we propose some regularization rules in order to select and summarize the most relevant information for clustering purposes. Essentially, we suggest to use a fragmented periodogram computed around the driving cyclical components of interest and to compare the various estimates. This procedure is computationally simple, but able to condense relevant information of the time series. A simulation exercise shows that the smoothed fragmented periodogram works in general better than the non-smoothed one and not worse than the complete periodogram for medium to large sample sizes. We illustrate this procedure in a study of the evolution of several stock markets indices. We further show the effect of recent financial crises over these indices behaviour.

Keywords Big data · Fragmented periodogram · Spectral clustering · Smoothed periodogram · Time series clustering

Mathematics Subject Classification 62H30 · 62M10 · 62M15

1 Introduction

The big data revolution is now offering researchers and analysts new possibilities and new challenges. This is particularly true with time series, as for many domains we now have access to many and very long time series related to a given domain of interest.

✉ Pilar Poncela
pilar.poncela@uam.es

[1] Cemapre and The University of Lisbon, Lisbon, Portugal

[2] European Commission, Joint Research Centre, Ispra, Italy

[3] Univ Autónoma de Madrid, Madrid, Spain

This happens in areas as diverse as astronomy, geophysics, medicine, social media, and finance.

In astronomy, for instance, we now have long and diverse series of star magnitude and spectra, radio-astronomy signals, asteroid position measurements, and other records. In medicine, we have very long and multiple records of physical activity indicators, heart rate, and other biological features. In social media and social studies, we have long records of human interactions, from administrative data to internet activities. In finance, we have tic-by-tic data of asset prices from many markets and firms.

The diversity and length of data available to researchers leads to particular challenges when comparing and clustering time series. For these tasks it is usually not possible to use traditional methods of analysing, estimating models, and comparing features, as these methods imply computing and inverting extremely large matrices.

In this paper, we propose a spectral method of synthesizing and comparing time series characteristics which is nonparametric and focused on the periodic behaviour of the time series. This method does not imply the computation of the full periodograms, but only of the periodogram components around the frequencies of interest. It then proceeds to comparing the periodogram ordinates for the various time series and grouping them with common clustering methods. We call it a *fragmented-periodogram* approach.

This method is somehow inspired by a procedure to predict tides due to William Thomson, later knighted Lord Kelvin. Almost 150 years ago, this famous scientist devised the first computationally achievable method for successfully predicting the tidal behaviour at any port for which a sufficiently long historical data would be available (Thomson 1881). For this purpose, he computed the amplitude and phase of just the cyclical components known to be of interest from astronomical reasonings.

We also check the advantages of smoothing the fragmented periodograms since it reduces the variance of the differences we compare: when two time series are different, the differences show up in the smoothed periodograms; when they are similar, the smoothed versions should be closer since the variance of the difference of the smoothed periodograms diminishes.

Clustering time series would be based on these new regularized periodograms. We will check the performance of our new proposal through some simulations. We will also apply the method to real financial time series, in particular, to the returns' volatility series, and check the effects of the Great Recession and of the European sovereign debt crisis in the financial integration of the stock markets.

The plan for the rest of this paper is as follows. In Sect. 2 we review the main traditional methods used to group time series. In Sect. 3 we discuss how time series characterization can be simplified in the presence of significative periodic components by using the corresponding spectral estimates. We introduce what we call the fragmented periodogram. In Sect. 4 we discuss the distribution of the fragmented periodogram ordinates and show that smoothing reduces the variance of the periodogram differences. In Sect. 5 we present finite-sample simulation evidence about the properties of the new clustering methods when we vary both the time dimension T and the cross sectional dimension N. In Sect. 6 we illustrate the procedure with three examples. Firstly, we cluster five data series both with the full and the fragmented periodograms. Secondly, we show how the 2007/2008 crisis changed the clustering of the main world financial indices. Thirdly, we show how the sovereign debt crisis changed the clustering of the main national stock indices. In Sect. 7 we present some conclusions.

2 Traditional clustering of time series

Various methods have been developed for comparing and clustering time series. One class of such methods is based on statistics for the time series co-movements. This can be done through structure analysis of covariances applied to Principal Component Analysis (PCA) and factor models analysis, among other methods. Dynamic factor models have gained a great popularity in recent times since they are widely used, for instance, in economics and finance. They allow to model common movements in time series fluctuations by decomposing the variance-covariance matrix of the observed series into the sum of a reduced-rank matrix (the one related to the common movements), and a (quasi) idiosyncratic part. Peña and Box (1987) analyzed dynamic factor models with stationary time series, Peña and Poncela (2006) extended the analysis to nonstationary time series, and Lam et al. (2011) further extended it to large data sets. Additionally, PCA methods have been heavily used in recent times to consistently estimate the common factors in dynamic factor models when both the time series dimension T and the cross dimension N tend to infinity [for stationary time series see Forni et al. (2000, 2005) or Stock and Watson (2002); for nonstationary time series see Bai and Ng (2004)]. Doz et al. (2011, 2012) extended these double asymptotic results for dynamic factor models estimated in state space form through the Kalman filter. For surveys of the literature on this topic see, for instance, Bai and Ng (2008) and Stock and Watson (2011).

In contrast to this class of methods, there are others that do not use any type of information regarding the co-movement of the series. Consider, for instance, studies in the classification of heart-beat patterns: the major interest of these studies is to characterize different types of patterns, without considering any correlation among records (Yang et al. 2011). In financial time series we may also be interested in market reaction to different shocks in different times. For this purpose, it may be helpful to compare and cluster time series according to their type of behaviour only. In order to do so, it is necessary to construct measures of similarity or dissimilarity among the time series. Preferably, but not necessarily, these measures should be distances, i.e., respect the identity, symmetry, and triangle inequality properties.

The first of this type of methods was introduced by Piccolo (1990). It is simply the Euclidean distance between the coefficients of the autoregressive representation of the time series under consideration. Consider two time series allowing AR representations, x_t and y_t, with t integer-valued, i.e.,

$$\begin{aligned} x_t &= \pi_{1,x} x_{t-1} + \pi_{2,x} x_{t-2} + \pi_{3,x} x_{t-3} + \cdots + \varepsilon_t \\ y_t &= \pi_{1,y} y_{t-1} + \pi_{2,y} y_{t-2} + \pi_{3,y} y_{t-3} + \cdots + \varepsilon_t \end{aligned}$$

Then, Piccolo's (1990) distance is given by

$$d(x, y) = \sqrt{\sum_{j=1}^{\infty}(\pi_{j,x} - \pi_{j,y})^2} \tag{1}$$

If the autoregressive coefficients are square summable, then this distance exists. However, even if the series are not stationary, it is possible to rely on truncated AR representations, which are always possible for empirical time series. Piccolo's method requires the estimation of a model and the computation of its AR coefficients. The same is true for other similar methods such as the residual-based metric of Tong and Dabas (1990) and the truncated AR expansions distance measure proposed by Maharaj (1996).

Early nonparametric methods for clustering time series were based on the sample autocorrelation distance (see, Galeano and Peña 2000). Assume, as above, we have two time series x_t and y_t. Assume also that $\hat{\boldsymbol{\rho}}_x = (\hat{\rho}_{x,1}, ..., \hat{\rho}_{x,r})$ and $\hat{\boldsymbol{\rho}}_y = (\hat{\rho}_{y,1}, ..., \hat{\rho}_{y,r})$ are the vectors of estimated autocorrelation coefficients for each series, and that for some r, $\hat{\rho}_{i,j} \cong 0$ for $j > r$ and $i = x, y$. Following Galeano and Peña (2000), a distance measure between x and y maybe defined by

$$d(x, y) = \sqrt{(\hat{\boldsymbol{\rho}}_x - \hat{\boldsymbol{\rho}}_y)'\Omega(\hat{\boldsymbol{\rho}}_x - \hat{\boldsymbol{\rho}}_y)}, \tag{2}$$

where Ω is a diagonal matrix of positive elements. This matrix may naturally give more weight to the lower order coefficients. If Ω is the identity matrix, then this measure corresponds simply to the Euclidean distance between the autocorrelation vectors of the two series x and y.

Later, Caiado et al. (2006) introduced frequency domain methods, which are also nonparametric as they do not require the estimation of any particular model. Asymptotically, these methods are equivalent to autocorrelation methods, but they may highlight and conveniently extract different information from the time series under consideration.

Spectral methods for comparing time series benefit from early work by Coates and Diggle (1986), who developed periodogram-based nonparametric tests for the hypothesis that two independent time-series are realizations of the same stationary process, and by Diggle and Fisher (1991), who developed similar tests by using the cumulative periodograms. Since then, spectral methods have found considerable interest in the literature.

Following Caiado et al. (2006), spectral methods essentially use the periodogram defined for each frequency $w_j = \frac{2\pi j}{T}$, $j = 1, ..., [T/2]$, ($[z]$ denotes the integer part of z, and T the number of time points)

$$I_x(w_j) = T^{-1}\left|\sum_{t=1}^{T} x_t e^{-itw_j}\right|^2 \tag{3}$$

and compute some type of distance between time series x and y such as

$$d(x, y) = \sqrt{\sum_{j=1}^{[T/2]} \left(P_x(\omega_j) - P_y(\omega_j)\right)^2}, \tag{4}$$

where P stands for the periodogram I, for the normalized periodogram $\gamma_0^{-1} I$, where γ_0 is the variance of the series, or for the log-normalized periodogram $\ln(\gamma_0^{-1} I)$.

In a simulation study, Caiado et al. (2006) show that using the normalized periodogram in (2) works quite well for distinguishing between nonstationary and near-nonstationary time series. An interpolated periodogram based metric was later proposed by Caiado et al. (2009) for handling series of unequal length.

The surveys by Liao (2005) and Caiado et al. (2015) review some of the most important developments on time series clustering and applications.

3 Cutting down the computations for large data sets clustering

In order to cut down the computations when the number of observations is large, we will propose to classify time series based only on that part of the series dynamics that define their main fluctuations; so, instead of computing the whole periodogram, we will compute it only for specific intervals. For instance, in macroeconomics, if we want to classify a set of time series according to their business cycle, we would filter the periodogram keeping only the frequencies that generate a cycle, for example, between 1 and 4 years. In finance, Corsi (2009) introduced the Heterogeneous Autoregressive (HAR) model for the stochastic volatility, a cascade of autoregressions on the squares of the returns at different frequencies (daily, weekly, monthly). This way one can reproduce the main features observed in financial returns such as long memory, fat tails and self-similarity. See Fig. 1 for an illustration of these features with the German stock index DAX. Let us assume that we have a time series process that on the levels or on the squares exhibit seasonality at weekly and monthly frequency, not presenting important oscillations at any other frequency. As an illustration, consider daily time series p_t close to a random walk and define the returns as $r_t = \ln p_t - \ln p_{t-1}$, which will be close to a non-correlated noise. An HAR(3) model for the (realized) volatility $RV_t = r_t^2$ would be given by

$$RV_{t+1}^{(d)} = c + \beta^{(d)} RV_t^{(d)} + \beta^{(w)} RV_t^{(w)} + \beta^{(m)} RV_t^{(m)} + \omega_{t+1} \tag{5}$$

where d stands for daily, $w = 5$ for weekly, $m = 21$ for monthly, and ω_t is a white noise. Following Corsi (2009), we can approximate the weekly and monthly terms respectively by $RV_t^{(w)} = \frac{1}{5}\sum_{i=1}^{5} RV_{t-i}^{(d)}$ and $RV_t^{(m)} = \frac{1}{21}\sum_{i=1}^{21} RV_{t-i}^{(d)}$. The HAR model is able to pick up the fact that the participants of financial markets respond to uncertainty at different horizons. As recently recognized in Bollerslev et al. (2018) this specification has become a benchmark to compare forecasts of realized volatility.

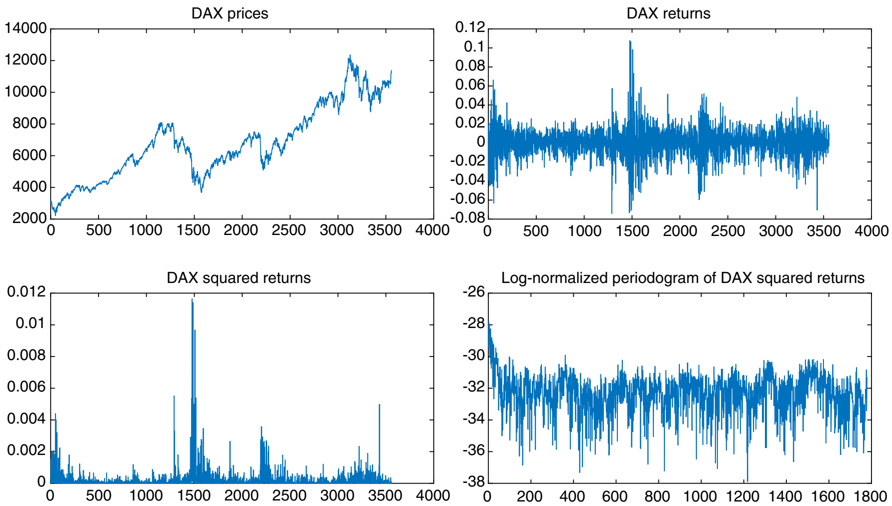

Fig. 1 German stock index DAX series illustrates the issues regarding the analysis we are performing. Price values are nonstationary and returns display typical volatility clustering; squared returns display typical time dependence with some dominant frequencies, the periodogram displays a sequence of peaks revealing these dominant frequencies

Our proposal consists on computing the periodogram *only* around the frequencies s of interest. But we may not choose frequency-domain symmetric intervals with frequencies $\pm\, h$ around the desired one s, since this would render time-domain asymmetric intervals.

Assume, for instance, that we are dealing with daily time series and are interested only in the annual cycle. For time series of length $T = 3556$ (one of the dimensions that we use in our empirical financial applications), we will select cycles of around 252 working days, that is, fluctuations at the annual frequency corresponding to $j_s = T/252 \cong 14$. As the time series might not behave exactly in the same way, we would like to keep an interval of frequencies around the one of interest, s. In this case, if we keep frequencies in the interval $[I_{s-10}; I_{s+10}] = [I_4; I_{24}]$, we will capture cycles between 148 and 889 working days. Although the interval is symmetric $[I_{s-10}; I_{s+10}]$ around the frequency of interest, it is asymmetric in the number of days we consider around 252. Therefore, we will use intervals that give us cycles of $\pm h$ days around the one of interest, i.e., intervals of the form $[I(\frac{2\pi}{s+h'}); \; I(\frac{2\pi}{s-h'})]$.

For instance, consider a time symmetric filter for monthly seasonality in daily time series around the monthly frequency s that reflects oscillations with a period of 22 working days. The time symmetric bandpass $[I(\frac{2\pi}{s+h'}); \; I(\frac{2\pi}{s-h'})]$ for $h' = 2$ will pick up all the cycles between 22 ± 2 days or between 20 and 24 days. In summary, we propose to use time-domain symmetric intervals (although frequency domain asymmetric). Figure 2 illustrates this issue.

We believe this choice is intervals adequate for many types of analyses, namely for economic and financial time series comparisons. Time domain intervals are easier to understand for most problems, and so they are easier to assess and chose in order to isolate the frequencies of interest.

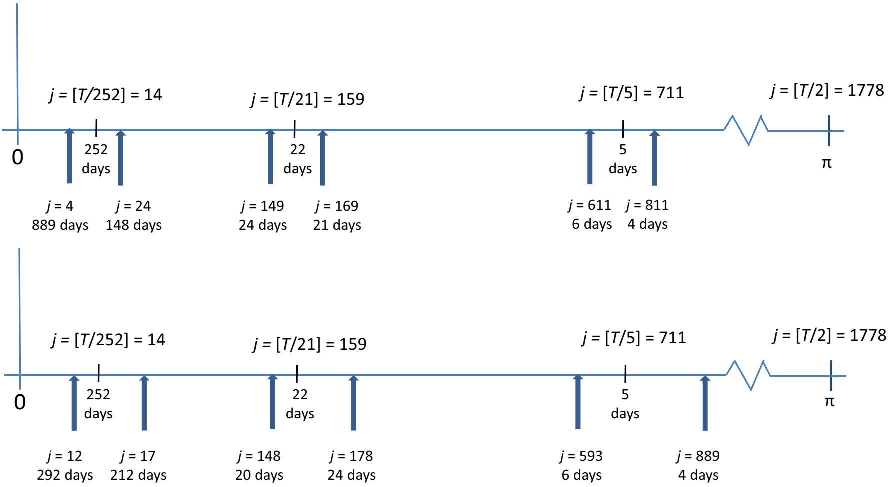

Fig. 2 Illustration of limits for fragmented periodogram frequencies to use, with two windows: frequency-domain symmetric (top) and time-domain symmetric (bottom). We clearly observe that symmetry in one domain implies asymmetry in the other domain

Finally, the amount of ordinates in the periodogram used at high frequencies is much larger than that used at low frequencies. Therefore, we would make the bandwidth dependent on the frequency we want to isolate: the higher the frequency, the larger the bandwidth. To highlight the dependence of the bandwith on the frequency of interest we will denote the h' parameter as h_s.

The choice of a reasonable parameter h_s is very flexible and can be adapted to the problem under consideration. There are a few conditions to take into account. Firstly, the longer the time series the larger h_s can be. Secondly, as we are limited by the computer memory and speed we should chose a small fraction of the frequencies. Thirdly, when we know the frequencies of interest we should try as much as possible to isolate them, i.e., to have small and non-overlapping windows. Taking these issues into consideration, we experimented various window lengths and got very reasonable results when h_s is proportional to s.

For the application developed in Sect. 6 of this paper, we have selected $h_s = [\frac{s}{4}]$. The seasonal periods are $s = 5, 21, 252$ days. Define the lower and upper bounds, l and u, as

$$l = \left\lfloor \frac{T}{s + h_s} \right\rfloor \quad \text{and} \quad u = \left\lceil \frac{T}{s - h_s} \right\rceil, \tag{6}$$

respectively, where $\lfloor\,\rfloor$ denotes the floor function and $\lceil\,\rceil$ the ceiling function. For instance, if the number of time points is $T = 2000$, then the frequency corresponding to the weekly cycle is $j_s = 400$, $h_s = 1$, and the upper and lower bounds are $l = 333$ and $u = 500$. So we will evaluate the periodogram for the interval $[I_{333}; I_{500}]$, picking up 77 frequencies to the left and 100 to the right of the weekly seasonal frequency. For $s = 21$, that corresponds to $j_s = 95$, $h_s = 5$

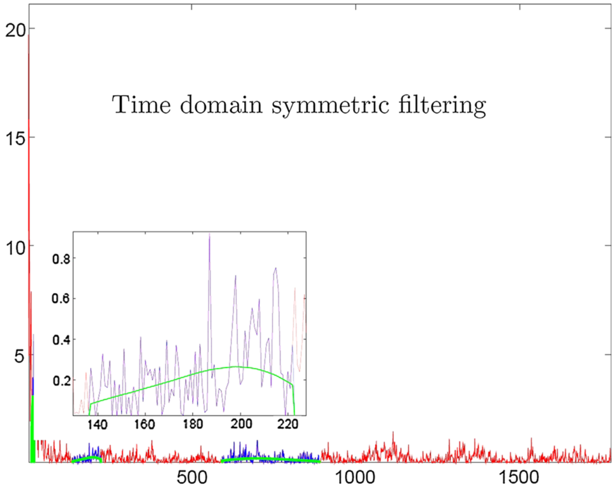

Fig. 3 A fragmented periodogram compared with the full periodogram and the filtered frequencies of interest. The inset highlights part of the smoothed fragmented periodogram

and we will compute the periodogram ordinates for the interval $[I_{76}; I_{125}]$. Finally, for the annual cycle $s = 252$ that corresponds to $j_s = 7$, $h_s = 63$ and we will keep the frequencies in the interval $[I_6; I_{10}]$. In all cases, notice that this asymmetry of the interval in the frequency domain leads to symmetric intervals in the time domain.

4 Why smoothing?

It is well known that the periodogram is not a consistent estimator of the spectral density, but it can be made consistent by smoothing the ordinates with an appropriate moving window (see, e.g., Brockwell and Davis 1991).

The purpose of this paper, however, is not to get estimates but to compare time series by using the periodogram. In this context, the usefulness of smoothing is not immediate: When comparing periodogram ordinates at a given set of frequencies, many differences between each spectral density and the corresponding periodogram ordinate are added or averaged, and so the noise in the values may level out for the final comparison. This means that the comparison is already a means of using grouped estimates.

However, by simulation and using a few theoretical results we found out that proper smoothing is justifiable and very useful. In reality, smoothing the periodogram generally improves significantly time series clustering results (Fig. 3).

In this section, we carry on some theoretical arguments for the usefulness of smoothing. We draw from early work of Coates and Diggle (1986) and Diggle and Fisher (1991). For simplicity, we assume a rectangular smoothing scheme. The smoothed spectral estimates will be denoted by $\hat{I}^k$, where k is the number of frequencies used; the periodogram ordinates will be denoted by I^{frag}, which will be just the periodogram ordinates $I(\omega)$ if the frequency ω is used, i.e., included in the fragmented periodogram, and zero otherwise. For a rectangular window we thus have:

$$\hat{I}^k(\omega_j) = \frac{1}{2M+1} \sum_{i=-M}^{M} I^{frag}(\omega_{j-i}).$$

Recall that in this case $k = 2M + 1$.

For frequencies not overlapping zero or π, the normalized periodogram filtering estimates approximate a χ^2 distribution:

$$I^k(\omega_j)2k/f(\omega_j) \dot{\sim} \chi^2_{2k}.$$

Then, assuming the time series X and Y are uncorrelated but have identical second-order properties with $f(\cdot)$ as their common spectral density, the normalized difference of the periodogram smoothed estimates follows asymptotically a Variance-Gamma (VG) distribution with parameters $\lambda = 2k/2$, $\alpha = 1/2$, $\beta = \mu = 0$:

$$(I^k_X(\omega_j) - I^k_Y(\omega_j))2k/f(\omega_j) \dot{\sim} VG.$$

From the assumptions and the properties of the VG, we immediately get

$$\begin{aligned} \mathrm{E}[2k(I^k_X(\omega_j) - I^k_Y(\omega_j))/f(\omega_j)] &= 0 \\ \mathrm{Var}[2k(I^k_X(\omega_j) - I^k_Y(\omega_j))/f(\omega_j)] &= \frac{2\lambda(\alpha^2 + \beta^2)}{(\alpha - \beta)^2(\alpha + \beta)^2} = 8k. \end{aligned}$$

Then,

$$\mathrm{Var}\left[\frac{\hat{I}^k_X(\omega_j) - \hat{I}^k_Y(\omega_j)}{f(\omega_j)}\right] = \frac{2}{k} = \frac{2}{2M+1} \to 0, \quad (M \to \infty).$$

Similar results hold when comparing log-periodograms. Again, using results in Coates and Diggle (1986) and Diggle and Fisher (1991), we know this difference follows a logistic distribution. For clarity, we can write both spectral densities f_X and f_Y, although we are assuming that $f_X = f_Y = f$. For each used periodogram ordinate, the difference D will then have the asymptotic distribution

$$\begin{aligned} D = \ln I_X(\omega_j) - \ln I_Y(\omega_j) &\dot{\sim} \mathrm{logistic}\{\ln f_X(\omega_j) - \ln f_Y(\omega_j), 1\} \\ &\dot{\sim} \mathrm{logistic}\{0, 1\}. \end{aligned}$$

and so from the properties of the logistic it follows that $\mathrm{E}[D] = 0$ and $\mathrm{Var}[D] = \pi^2/3 \nrightarrow 0$, where we stress the fact that the variance of each ordinate difference converges to a nonzero constant.

As we have assumed that $f_X = f_Y$, for smoothed periodogram ordinates not overlapping 0 or π we obtain the distribution of e^D

$$e^D = \frac{I_X^k(\omega_j)}{I_Y^k(\omega_j)} \sim \frac{f_X(\omega_j)}{f_Y(\omega_j)} F_{2k,2k} \sim F_{2k,2k}$$

Then $\mathrm{E}\left[e^D\right] = k/(k-2) \to 1$ and $\mathrm{Var}\left[e^D\right] = \frac{2k^2(2k-2)}{k(k-2)^2(k-4)} \sim \frac{1}{k} \to 0$, from which we obtain the distribution of the log-smoothed periodogram

$$D = \ln I_X^k(\omega_j) - \ln I_Y^k(\omega_j) \sim 2 \ln \mathrm{FisherZ}(2k, 2k).$$

Now, from the well-known distribution of the FisherZ, we get

$$\mathrm{E}[\ln I_X^k(\omega_j) - \ln I_Y^k(\omega_j)] \to 2 \cdot \frac{1}{2}\left(\frac{1}{2k} - \frac{1}{2k}\right) = 0$$
$$\mathrm{Var}[\ln I_X^k(\omega_j) - \ln I_Y^k(\omega_j)] \to 4 \cdot \frac{1}{2}\left(\frac{1}{2k} + \frac{1}{2k}\right) \to 0.$$

We conclude that for both the periodogram and the log-periodogram, smoothing reduces the variance of the differences. This constitutes an argument in favour of smoothing the fragmented periodogram: time series with similar spectral densities should appear closer and time series with different spectral densities should be distinguishable.

We propose the following procedure to compute the smoothed periodogram:

1. Compute the periodogram, normalized periodogram or log normalized periodogram only for the ordinates in the intervals

$$\left[\frac{2\pi}{s+h_s}; \frac{2\pi}{s-h_s}\right] \tag{7}$$

   where h_s depends on the frequencies s of interest.
2. Smooth the fragmented periodogram, normalized periodogram or log normalized periodogram P^{frag}. We suggest two most popular smoothers, the Bartlett and the rectangular ones, given in Eqs. (8) and (9) respectively, although, in principle, any other smoother can be used.

$$\hat{P}_j^k = \frac{1}{M} \sum_{i=-M}^{M} \left(1 - \frac{|i|}{M}\right) P_{j-i}^{frag} \tag{8}$$

$$\hat{P}_j^k = \frac{1}{2M+1} \sum_{i=-M}^{M} P_{j-i}^{frag} \tag{9}$$

   where $k = 2M + 1$.

In the next section, we assess and confirm the advantage of smoothing by means of an extensive simulation study.

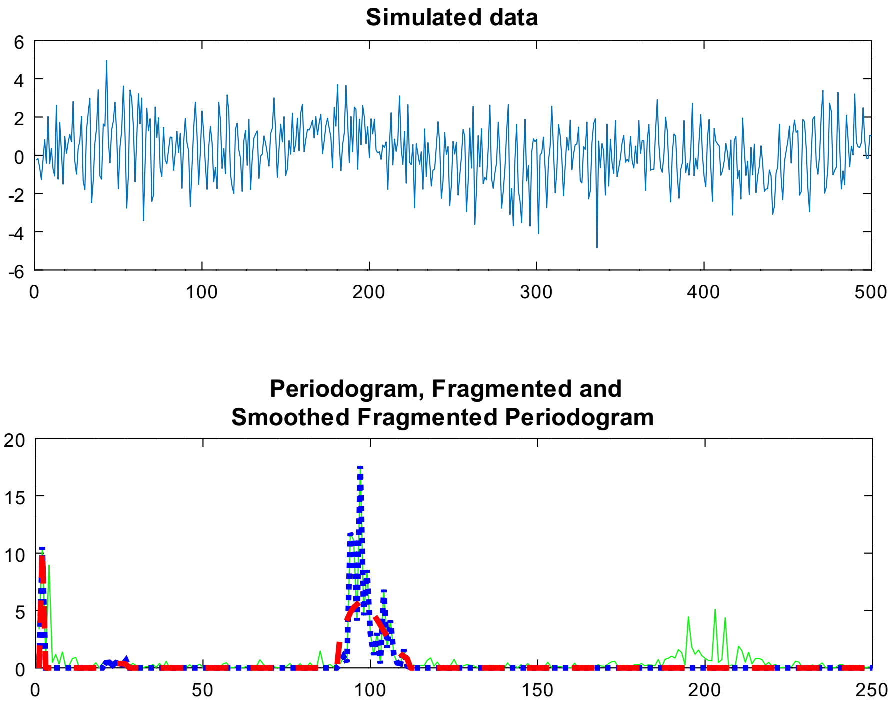

Fig. 4 Top panel: simulated data. Bottom panel: periodogram (full line, in green), fragmented periodogram (dot dashed line, in blue), and smoothed fragmented periodogram (dashed line, in red) (colour figure online)

5 Simulation exercise

In order to check the performance of our proposed procedure we have performed several simulation exercises. For $N = 4$ time series we have simulated the following model

$$y_t = \phi_w y_{t-w} + \phi_m y_{t-m} + \phi_a y_{t-a} + \epsilon_t$$

where ϕ_i, $i = w, m, a$ is the autoregressive parameter associated to the weekly, monthly and annual cycles, respectively, and ϵ_t is white noise. To get an idea of the simulated data, Fig. 4 plots the generated time series in one replication when the AR parameters take the values $\phi_w = 0.7$, $\phi_i = 0.1$, $i = m, a$ and $var(\epsilon_t) = 1$. This is one of the models that we will use in this set of simulations as it will be shown in Table 3. The model reflects the case where the weekly cycle is very noticeable while the monthly and annual cycles are less marked. The sample size is $T = 500$ time points and the plot shows the time series in the top panel and the periodogram and its fragmented and smoothed fragmented versions in the bottom panel. Notice that we only use a small percentage of the ordinates of the periodogram in its fragmented versions (smoothed or not). Notice as well that we do not even use the information provided by the harmonics in the fragmented versions.

In each run, we generated N time series divided into 2 groups considering two different data generating processes each time we simulate a replica. First, we consider that the weekly, monthly and annual cycles might not been of exactly the same length, so we have varied the lags w, m, a in the two data sets. In our first experiment we have considered that for the first group of series the lags associated to the weekly, monthly and annual seasonal cycles were 5, 21 and 252, respectively, while for the second group they were 4, 25 and 300. Therefore, in our first set of simulations we will be comparing series generated by the model

$$y_t = \phi_w y_{t-5} + \phi_m y_{t-21} + \phi_a y_{t-252} + \varepsilon_t$$

versus those generated by

$$y_t = \phi_w y_{t-4} + \phi_m y_{t-25} + \phi_a y_{t-300} + \varepsilon_t.$$

This might be due, for instance, to different bank holidays. We have generated the 2 sets of series according to the previous models and computed the log normalized periodogram and the fragmented and the smoothed fragmented versions. We use the log normalized periodogram based on the results of Caiado et al. (2006). For the smoothed version, we have used the Bartlett filter given in (8) with $M = \frac{u-l}{2}$, with u and l as defined in Eqs. (5) and (6) and h_s taking values 0.5, 3 and 45 for the annual, monthly and weekly cycles, respectively. We have used these values of h_s in order to keep approximately the same percentage of ordinates (a little bit more than 11% in all cases) in the fragmented versions of the periodogram. Then, we have clustered the time series using the Euclidean distance of the 3 versions of the log normalized periodogram. For brevity, we will simply denote the log normalized periodogram as just the periodogram from now on. We have run 1000 replications for each set of parameters and sample sizes $T = 500, 1000, 2000, 5000$ and 10,000.

All the tables in this section are read in the same way. The first column shows a triple of numbers (T, p, f) being the first one T the sample size, the second one p the number of ordinates of the periodogram ($T/2$ for T even) and the third one the number of ordinates that we use in the fragmented and smoothed fragmented periodograms, which are the same. Then the table shows three panels of 3 columns each one of them, where we show the percentage of times that the N time series are correctly clustered for each sample size T when using the whole periodogram and the fragmented and smoothed fragmented versions. Notice that our fragmented versions use a small part of ordinates of the full periodogram (around 11%). We want to check that if using only a small fraction of the information (ordinates in the periodogram), we could get classification results comparable to those obtained when using the whole periodogram. We would also like to see the effect of smoothing the fragmented periodogram before clustering. The overall effect of smoothing the periodogram is to reduce the variance of the spectral estimates differences as the number of ordinates used for smoothing increases. The effect of smoothing should be larger, the larger the interval we use for smoothing. In this sense, we would like to point out that the number of ordinates associated to the weekly cycle in the fragmented periodogram is larger than that related to the monthly cycle and both larger than that used in the annual cycle. In fact, for the smallest sample size $T = 500$ used in the simulations, the number of ordinates of the fragmented and smoothed fragmented periodograms associated to the annual cycle is just 1 and, of course, both coincide in that ordinate. The effect of smoothing should be larger, the larger the interval we use.

Table 1 illustrates the results of our first simulation exercise, in which we have used three different sets of parameters for generating three stationary processes. The first column shows the sample size and information regarding the number of ordinates used. Columns 2–4, 5–7 and 8–10 the percentage of correctly classified cases for the 3 sets of parameters used in the simulations. For a given set of parameters, the results always improve with the sample size T. Notice that given the period of the annual cycle in both sets (the lags used for simulating the yearly cycle were 252 and 300, respectively, in the 2 groups), the annual cycle is hardly reproduced for the smallest sample size, $T = 500$, so the results improve with the sample size as we can detect more times the annual cycle. Moreover, for a given sample size, the best results are always given by the full periodogram. However the smoothed periodogram, that uses a small percentage of information, follows closely the outcome of the full periodogram for moderate to large sample sizes. The fragmented periodogram gives the worst results of all the procedures but it works quite well from $T = 2000$ (with, for instance, already 92% of correctly classified replications for the first set of parameters). As smoothing decreases the variance of the periodogram, we could detect differences in periodograms more easily using the smoothed version than just the fragmented version.

In the second set of simulations we assign the same value to the autoregressive parameters associated to each one of the seasonal cycles, so all the cycles rely on the same value of the parameters and it is not the difference in the value of the parameters what drives the good or bad performance of the method. Table 2 shows the results in the left, center and right panels when all the autoregressive parameters take the value of .3, .2 and .1, respectively. Notice that as the magnitude of the autoregressive parameters decreases, the processes are closer to white noise. In the extreme case in which all the autoregressive parameters took the value of zero, all the time series would be white noise and then we would have only one group of series instead of two.

The ordering in the performance of the methods is maintained in Table 2: the full periodogram works better in classifying the time series, followed by the smoothed fragmented periodogram and finally by the fragmented. As expected, results are better in the left panel, then in the center panel and, finally, they worsen in the right panel as we are diminishing the differences between the two sets of time series getting all of them closer to the same white noise process. Also, and as expected, given a column in the table, results improve as we move down, that is, as we increase the sample size T. Notice that for $T = 2000$, results are quite good for all methods when $\phi_i = 0.3$, $i = w, m, a$.

In the next set of simulations, we will maintain the same lag for all the autoregressive processes and the differences would rely in the value of only one parameter. In particular, we will simulate $N/2$ series from each one of the following processes.

$$y_t = \phi_5 y_{t-5} + 0.1 y_{t-21} + 0.1 y_{t-252} + \varepsilon_t$$

Table 1 Percentage of times that the $N = 4$ series were correctly clustered when the 2 sets of series differ in the lags used to define the weekly, monthly and annual cycles

(T, p, f)	$\phi_w = .4, \phi_m = .3, \phi_a = -.2$			$\phi_w = -.4, \phi_m = .3, \phi_a = .2$			$\phi_w = .4, \phi_m = -.3, \phi_a = -.2$		
	All	Frag	Smooth	All	Frag	Smooth	All	Frag	Smooth
(500, 250, 29)	49.8	27.6	34.3	47.5	23.3	32.5	46.5	22.1	29.5
(1000, 500, 56)	99.3	65.5	78.7	98.9	54.9	77.9	99.0	56.3	75.2
(2000, 1000, 112)	100.0	92.0	97.0	100.0	85.5	96.8	100.0	86.4	96.2
(5000, 2500, 279)	100.0	100.0	100.0	100.0	99.8	100.0	100.0	99.8	100.0
(10,000, 5000, 558)	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

First group, lags = 5, 21 and 252; second group, lags = 4, 25 and 300

Table 2 Percentage of times that the $N = 4$ series were correctly clustered when the 2 sets of series differ in the lags used to define the weekly, monthly and annual cycles

(T, p, f)	$\phi_w = \phi_m = \phi_a = 0.3$			$\phi_w = \phi_m = \phi_a = 0.2$			$\phi_w = \phi_m = \phi_a = 0.1$		
	All	Frag	Smooth	All	Frag	Smooth	All	Frag	Smooth
(500, 250, 29)	38.1	20.3	25.5	20.7	12.7	13.7	11.8	9.9	12.5
(1000, 500, 56)	97.7	59.2	64.0	60.9	22.9	27.9	17.0	11.6	14.4
(2000, 1000, 112)	100.0	91.7	92.6	94.5	42.4	45.2	31.2	14.0	18.5
(5000, 2500, 279)	100.0	99.9	100.0	100.0	75.2	93.3	53.6	19.6	37.6
(10,000, 5000, 558)	100.0	100.0	100.0	100.0	88.1	100.0	76.1	24.2	60.3

First group, lags = 5, 21 and 252; second group, lags = 4, 25 and 300. Left panel: $\phi_i = 0.3$; center panel $\phi_i = 0.2$; right panel $\phi_i = 0.1, i = w, m, a$

versus

$$y_t = \phi_5^* y_{t-5} + 0.1 y_{t-21} + 0.1 y_{t-252} + \varepsilon_t,$$

so the only difference would be the in the value of the autoregressive parameter associated to the weekly seasonal cycle. Table 3 presents the results when the difference between the two parameters are high, medium, and low.

As expected, the results are better for all methods when the differences between the two parameters are large (left panel), they worsen when they are of medium size (center panel) and they further diminish when they are low (right panel). As usual, the results improve with the sample size T for all methods and are close to 100% of corrected classified cases for sample sizes from $T = 2000$ when the differences in the parameters are large (left panel). Notice, however, that the ordering of the methods is not the same as in previous tables and for the smallest sample size, $T = 500$, clustering based on the smoothed fragmented periodogram works best, followed by clustering based on the full periodogram and, finally, by that based on the fragmented one. This happens for the 3 sets of parameters. It seems that clustering based on smoothing fragmented periodograms always improves over clustering over plain fragmented periodograms as it decreases the variance of the difference of the spectral estimates periodogram at each smoothed ordinate but notice that for $T = 500$ it also works better than the full periodogram. It could not be due to the fact the discarded information was only noise since the fragmented always works worst than the full periodogram, so we conclude that smoothing in the weekly cycle (the one associated to the largest number of ordinates of the three cycles), reduces drastically the variance of the smoothed ordinates of the periodogram. For instance, for $T = 500$, the percentage of correctly classified cases when using the smoothed periodogram for the parameters in the center panel (medium size of the differences between the parameters) is about twice that percentage of just the fragmented periodogram.

We repeat the exercise when the difference in the parameters are given in the monthly cycle and simulate the processes

$$y_t = 0.1 y_{t-5} + \phi_{21} y_{t-21} + 0.1 y_{t-252} + \varepsilon_t$$

versus

$$y_t = 0.1 y_{t-5} + \phi_{21}^* y_{t-21} + 0.1 y_{t-252} + \varepsilon_t.$$

Table 4 presents the results when the differences in the monthly parameter are big (left panel), medium (center panel) and small (right panel). Notice that, for a given sample size T, the monthly cycle is repeated fewer times than the weekly cycle so the results could worsen with respect to differences in the weekly cycle.

The overall results are similar to those of Table 3, although the percentages of correctly classified cases are lower for the smaller sample sizes. Here, however, the usual ordering about the performance of the methods is maintained and we do not see a surpass in the classification based on the smoothed fragmented periodogram over that based on the full periodogram, perhaps, due to the fact that the number of ordinates

Table 3 Percentage of times that the $N = 4$ series were correctly clustered when the two sets of series differ in the autoregressive parameter of the weekly cycle

(T, p, f)	$\phi_5, \phi_5^* = 0.7, 0.1$			$\phi_5, \phi_5^* = 0.7, 0.3$			$\phi_5, \phi_5^* = 0.7, 0.5$		
	All	Frag	Smooth	All	Frag	Smooth	All	Frag	Smooth
(500, 250, 29)	60.0	38.7	65.6	40.0	22.6	46.0	18.0	14.4	21.9
(1000, 500, 56)	99.2	64.3	95.1	90.4	40.8	77.6	40.9	17.8	37.2
(2000, 1000, 112)	100.0	88.2	99.7	99.7	59.4	96.0	77.2	22.1	60.0
(5000, 2500, 279)	100.0	99.2	100.0	100.0	86.9	99.9	99.5	34.1	85.0
(10,000, 5000, 558)	100.0	100.0	100.0	100.0	97.2	100.0	100.0	50.8	99.6

Table 4 Percentage of times that the $N = 4$ series were correctly clustered when the 2 sets of series differ in the autoregressive parameter of the monthly cycle

(T, p, f)	$\phi_{21}, \phi_{21}^* = 0.7, 0.1$			$\phi_{21}, \phi_{21}^* = 0.7, 0.3$			$\phi_{21}, \phi_{21}^* = 0.7, 0.5$		
	All	Frag	Smooth	All	Frag	Smooth	All	Frag	Smooth
(500, 250, 29)	57.4	26.6	33.9	37.1	19.6	21.8	17.8	11.3	13.3
(1000, 500, 56)	99.5	60.3	79.1	86.5	38.4	55.0	40.5	17.2	26.6
(2000, 1000, 112)	100.0	90.8	98.2	99.6	63.8	90.4	77.2	28.5	54.3
(5000, 2500, 279)	100.0	99.7	100.0	100.0	94.0	99.8	98.9	50.1	88.1
(10,000, 5000, 558)	100.0	100.0	100.0	100.0	99.8	100.0	100.0	73.5	99.5

of the monthly cycle in the fragmented periodogram is smaller than that related to the weekly cycle, so the benefits from smoothing are smaller as well.

Overall, the previous set of simulations shows the effects of increasing T for small N. Firstly, for very large (T, p, f) sizes the three methods used to cluster time series (full, fragmented and smoothed fragmented) give similar results. However, as the size T decreases there can be great differences. Secondly, smoothing the fragmented periodogram always gives better results that just using the plain fragmented periodogram since the variance of the difference of the periodograms at each ordinate goes to zero with the smoothing parameter. Thirdly, differences due to the weekly cycle are better detected than differences due to the monthly cycle, as the former are reproduced a larger number of times in each periodogram. Additionally, the higher the interval associated to each seasonal frequency the higher is the seasonal frequency, so smoothing can work through a larger set of ordinates. That is, the higher the frequency associated to the seasonal cycle that is different in the two data sets, the better classification results we obtain.

We complete our simulation exercise increasing the cross section dimension N so as to having results varying both dimensions in our problem, N and T. As we have mentioned in Sect. 2, related literature in the context of principal components suggest that when both N and T go to infinity, principal components can be used to consistently estimate dynamic factor models. However, in spite of the aforementioned theoretical results, Boivin and Ng (2006) for dynamic factor models estimated through principal components and Poncela and Ruiz (2015) for those estimated through the Kalman filter question this in empirical models. Simulation studies in Poncela and Ruiz (2015) suggest that choosing N between 20 and 30 is a good compromise between the increasing complexity of the models with N and the reduction in the uncertainty when estimating the common factors that the asymptotic theoretical results suggest when both N and T tend to infinity. To check the role of both dimensions in our case we also run additional simulations for $N = 20$ and $N = 100$, while varying T as in the previous tables, $T = 500, 1000, 2000, 5000, 10,000$. In particular, $N/2$ series were simulated from each of two processes. The N series were then grouped into two clusters and we computed the number of series correctly classified into both clusters. This was repeated 1000 times. A conversation formula is then used to determine the percentage of success on a scale ranging from 0 to 100%. Tables 5 and 6 show the results for $N = 20$ and $N = 100$, respectively, for the same two processes.

Overall, we can conclude that the effect of increasing N from 4 to 20 renders better classification results while further increasing N to 100 worsens the results, except if the time dimension T is large enough to assure success. Medium size problems in the cross dimension seem to point to a good compromise between increasing complexity with N and retaining enough information for clustering.

Now, we are going to increase the cross section dimension N so as it would be difficult to handle the classification problem with more traditional methods. In particular, we will consider $N = 1000$ and $N = 10,000$ time series. The first thing we need to do is to we use a non-hierarchical clustering procedure since it is less time consuming. More specifically, we will use the k-means algorithm, which is more suitable than hierarchical clustering for large amounts of data. Tables 7 and 8 show the results for $N = 1000$ and $N = 10,000$, respectively, for the same two processes that we have

Table 5 Percentage of times that the $N = 20$ series were correctly clustered when the 2 sets of series differ in the lags used to define the weekly, monthly and annual cycles

(T, p, f)	$\phi_w, \phi_m, \phi_a = 0.4, 0.3, -0.2$			$\phi_w, \phi_m, \phi_a = 0.4, -0.3, 0.2$			$\phi_w, \phi_m, \phi_a = 0.4, -0.3, -0.2$		
	All	Frag	Smooth	All	Frag	Smooth	All	Frag	Smooth
(500, 250, 29)	62.8	24.2	33.2	65.2	20.4	31.6	66.2	21.6	31.4
(1000, 500, 56)	100.0	77.2	84.6	100.0	60.6	82.6	100.0	67.8	80.6
(2000, 1000, 112)	100.0	99.0	99.6	100.0	96.2	98.8	100.0	85.8	98.2
(5000, 2500, 279)	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
(10,000, 5000, 558)	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

First group, lags = 5, 21 and 252; second group, lags = 4, 25 and 300

Table 6 Percentage of times that the $N = 100$ series were correctly clustered when the 2 sets of series differ in the lags used to define the weekly, monthly and annual cycles

(T, p, f)	$\phi_w, \phi_m, \phi_a = 0.4, 0.3, -0.2$			$\phi_w, \phi_m, \phi_a = 0.4, -0.3, 0.2$			$\phi_w, \phi_m, \phi_a = 0.4, -0.3, -0.2$		
	All	Frag	Smooth	All	Frag	Smooth	All	Frag	Smooth
(500, 250, 29)	18.8	6.4	8.8	20.4	7.2	12.0	22.4	7.6	10.6
(1000, 500, 56)	98.4	32.4	67.2	97.6	28.4	71.4	96.2	25.2	66.6
(2000, 1000, 112)	100.0	84.6	95.8	100.0	70.8	96.4	100.0	62.8	93.6
(5000, 2500, 279)	100.0	100.0	100.0	100.0	98.4	100.0	100.0	100.0	100.0
(10,000, 5000, 558)	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

First group, lags = 5, 21, and 252; second group, lags = 4, 25, and 300

Table 7 Percentage of times that the $N = 1000$ series were correctly clustered

(T, p, f)	$\phi_w, \phi_m, \phi_a = 0.4, 0.3, -0.2$			$\phi_w, \phi_m, \phi_a = 0.4, -0.3, 0.2$			$\phi_w, \phi_m, \phi_a = 0.4, -0.3, -0.2$		
	All	Frag	Smooth	All	Frag	Smooth	All	Frag	Smooth
(500, 250, 28)	97.9	86.1	91.6	98.8	84.2	89.2	97.3	82.3	89.2
(1000, 500, 56)	99.8	98.5	99.7	99.3	97.8	99.2	99.4	97.4	98.9
(2000, 1000, 112)	100.0	99.9	100.0	99.9	99.9	99.9	100.0	99.8	99.9
(5000, 2500, 279)	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
(10,000, 5000, 558)	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Table 8 Percentage of times that the $N = 10,000$ series were correctly clustered

(T, p, f)	$\phi_w, \phi_m, \phi_a = 0.4, 0.3, -0.2$			$\phi_w, \phi_m, \phi_a = 0.4, -0.3, 0.2$			$\phi_w, \phi_m, \phi_a = 0.4, -0.3, -0.2$		
	All	Frag	Smooth	All	Frag	Smooth	All	Frag	Smooth
(500, 250, 28)	91.7	86.1	86.3	89.4	83.6	86.1	90.7	83.3	86.4
(1000, 500, 56)	99.7	96.5	98.7	98.7	91.3	94.8	99.4	97.1	97.6
(2000, 1000, 112)	100.0	99.8	99.9	100.0	99.9	99.9	99.9	99.9	99.9
(5000, 2500, 279)	100.0	100.0	100.0	100.0	98.4	100.0	100.0	100.0	10.00
(10,000, 5000, 558)	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

used in Tables 5 and 6. The main conclusions are maintained through this new set of simulations. We observe that all procedures work quite well. As before, the smoothed fragmented periodogram works better than just the fragmented version. Smoothing helps. Notice that we only use around 11.2% of the ordinates in the fragmented and smoothed fragmented versions and the results are quite similar. When we also increase the time dimension T the results get even closer and from $T = 2000$ onwards we can hardly see the differences.

6 Real data applications

As illustrative examples, we provide here a simple but convincing illustration and two simple case studies of country financial clustering evolution. In finance, cluster analysis of time series has become an important task, as researchers and investors are interested in identifying similarities in financial assets for investments and risk management purposes.

For instance, Otranto (2010) proposed a clustering algorithm to compare groups of homogeneous time series in terms of dynamic conditional correlations. Caiado and Crato (2010) introduced a volatility-based metric for cluster analysis of stock returns using the information about the estimated parameters in the threshold GARCH model. Bastos and Caiado (2014) proposed a metric for clustering financial time series based on the distance between variance ratio statistics computed for individual series.

6.1 A simple illustration

In the illustrative example presented below, we use spectral analysis and classic multidimensional scaling to construct a configuration of five stock markets (DAX and HDAX from Germany, Italia All-Share and MIB, and PSI20 Index from Portugal) in a two-dimensional space. We purposely use two time series from Italy, two from Germany, and one from Portugal.

We all know these three countries have different characteristics and expect them to be separated. We also expect that financial indices from the same country to appear close.

When using all periodogram ordinates, results shown in Fig. 5 are reassuring about the efficiency of the spectral method as indices appear as expected. The two Italian indices appear close together, the two German indices appear close together, and the three countries appear separated. Now, instead of using all periodogram ordinates, we perform the same analysis using the fragmented periodogram ordinates only. It is very revealing and reassuring that not much comparative information is lost. Essentially, with a significantly reduced computational effort we get the same information, as we can see in Fig. 6.

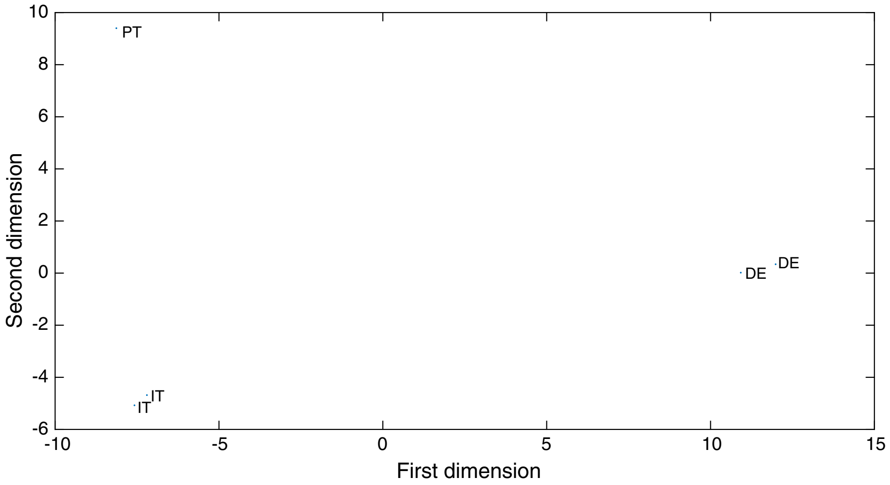

Fig. 5 Five stocks analysed with a two-dimensional scaling map using with the full periodograms

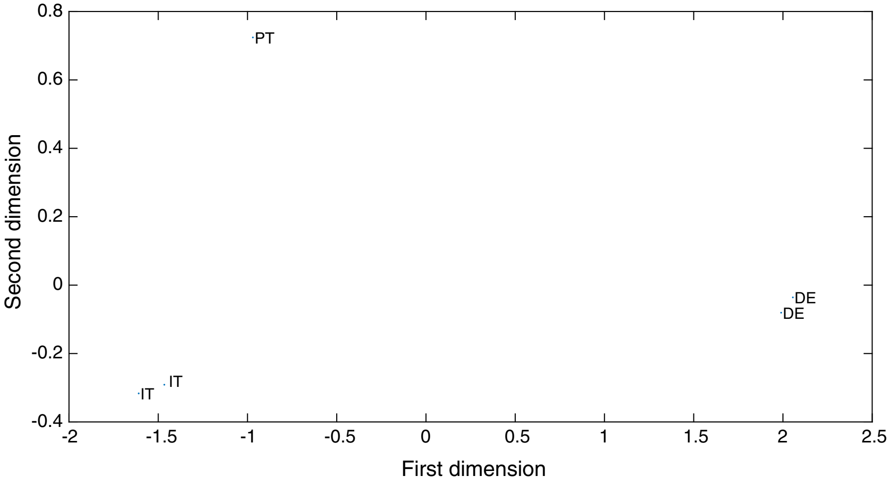

Fig. 6 Five stocks analysed with a two-dimensional scaling map using the fragmented periodograms

6.2 A two-period comparison of European stock markets

We now use the fragmented-periodogram approach to identify similarities among 44 European stock markets. Data correspond to closing prices and cover the period from January 2003 to December 2016. We divide the analysis into two periods: (i) before the sovereign debt crisis (2nd January 2003–30th June 2011) and (ii) after the sovereign debt crisis (1th July 2011–31th December 2016).

Figures 7 and 8 show the two-dimensional scaling maps for these two periods. In *blue* we show series from distressed countries or from countries that have experienced a significant deterioration (Ireland, Greece, Spain, Italy and Portugal). In *red*

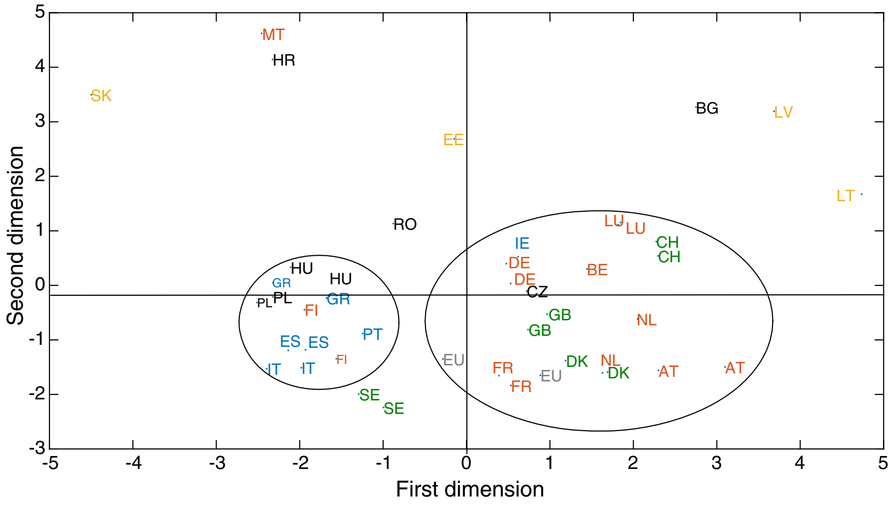

Fig. 7 European stocks for the years 2001–2011, i.e., before the sovereign debt crisis, analysed with a two-dimensional scaling map using the fragmented periodograms

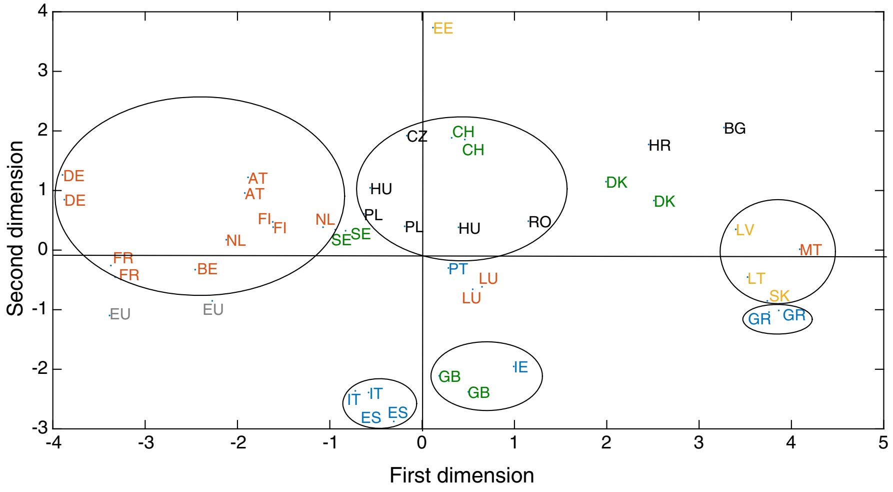

Fig. 8 European stocks for the years 2011–2016, i.e., after the sovereign debt crisis, analysed with a two-dimensional scaling map using the fragmented periodograms

we show series from the Euro area core (Austria, Belgium, Finland, France, Germany, Luxembourg, and Netherlands). In *green* we show series from non-distressed but non euro-area core countries (Denmark, Great Britain, Sweden, and Switzerland). In *orange* we show series from the East Euro area (Estonia, Latvia, Lithuania, Slovakia). In *black* we show series from the East euro-area (Bulgaria, Czech Republic, Hungary, Poland, Romania, and Croatia). In *grey* we show global European stock indexes (eurostock50 and stxe600).

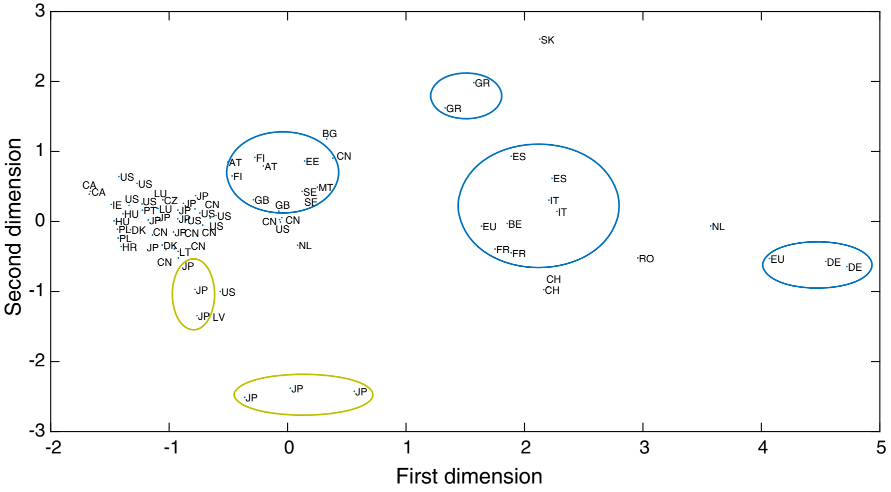

Fig. 9 International stocks before the financial crisis (2003–2007) analysed with a two-dimensional scaling map using the fragmented periodograms

Figure 7, corresponds to period (i) and Fig. 8 to period (ii). Comparing the two graphs it becomes clear that this technique sheds light over the evolution of the markets. In the first period, most non-distressed euro-area and non-euro-area countries are together while euro-area distressed countries are close together in a clear cluster. In the second period, after the sovereign debt crisis, the separation is clearer: euro-area core countries remain close to each other but far from non euro-area countries UK and Denmark; Greece is further separated from the other euro are distressed countries that improved their situation during this period (Portugal, Spain, Italy and Ireland).

6.3 A two-period comparison of worldwide stock markets

We now use the same technique for clustering 79 free float adjusted market capitalization equity indices constructed by Morgan Stanley Capital International (MSCI). Data used in the analysis consists on daily index prices from January 2003 to December 2016. Again, we divide the analysis into two periods: (i) before the financial crisis (2003–2007) and (ii) after the financial crisis (2008–2016).

Figures 9 and 10 show two-dimensional scaling maps for the two subperiods. Again, it becomes clear by comparing the two graphs that this technique sheds light over the evolution of the markets. In the first period, we do not see a clear separation: a large cluster contains Asian, North America and European stock markets. In the second period we see that the crisis forced the countries do individualize their paths: we can identify a cluster containing most Japanese and Chinese stock markets at the left-hand side of the map, a cluster containing most Canadian and American stock markets at the right-hand side, and a large cluster containing most European markets at the middle.

All these examples show how this fragmented periodogram clustering technique is able to illuminate interesting characteristics of the time series under consideration

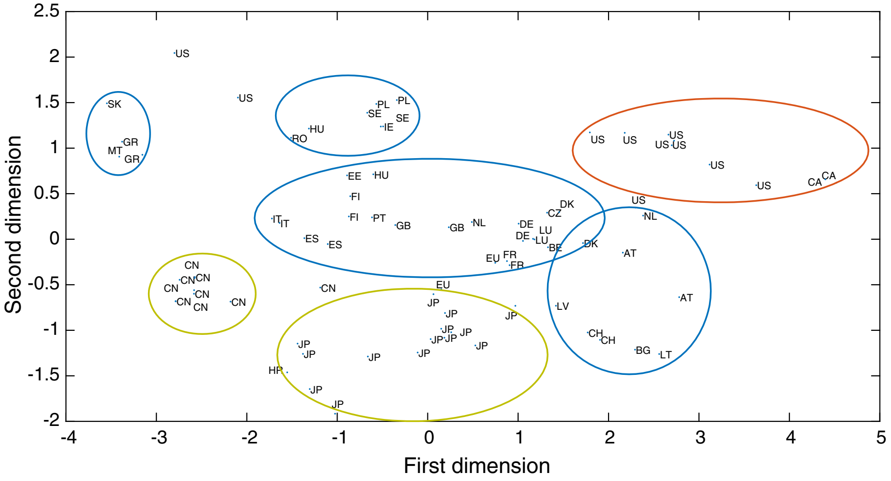

Fig. 10 International stocks after the financial crisis (2008–2016) analysed with a two-dimensional scaling map using the fragmented periodograms

7 Conclusions

We have suggested a method for comparing and clustering time series, which is suitable to face large data sets and should provide insights to new big data challenges. The method works in the frequency domain, uses a computationally simple and very parsimonious approach, and is nonparametric, as it does not rely on fitting any specific model. The proposal is somehow inspired by the ideas for tide predicting set forth by Lord Kelvin in the mid of the nineteenth century, but it can be used for any type of time series, particularly those with known dominating frequencies.

In its essence, our method computes the periodogram ordinates only in narrow vicinities of the frequencies of interest, obtaining what we call a fragmented-periodogram. Then, it proceeds to smooth these ordinates in order to obtain spectral estimates. With these estimates, the spectral differences have a reduced variance and provide an improved way for comparing the time series. Finally, we apply standard clustering methods for grouping the time series under consideration.

By using some standard results we were able to theoretically show that smoothing the fragmented periodogram reduces the variance of the spectral estimates differences and thus provides a more reliable means for comparing time series.

By means of simulations, we have contrasted our method with one using the full periodogram and shown that the fragmented periodogram obviously looses some information but not much. We further show that smoothing the fragmented periodogram improves the method, giving results that are often as reliable as those from the full periodogram. We finally show that the method is able to successfully characterize, compare, and differentiate time series, thus providing a basis for clustering. Its success depends on the existence of long time series—but it was devised exactly to deal with these cases.

To illustrate the usefulness of our method we presented two financial studies. In the first one, we studied 44 European stock market series both before and after the sovereign debt crisis. In the second one, we studied 79 worldwide equity indices both before and after the 2008 financial crisis. In both cases, our method was able to find significant and meaningful changes in the clustering of the time series under consideration.

The method is applicable to any time series domain, from geophysics to finance—it only needs the presence of some type of cyclical behaviour. It is particularly suited to large data time series, and should provide useful for big-data clustering challenges as it provides a computationally feasible way of comparing a large number of very long time series.

Acknowledgements Jorge Caiado and Nuno Crato have been supported through Project CEMAPRE-UID/MULTI/00491/2019 financed by FCT/MCTES through national funds. This paper was finished when the Pilar Poncela returned to Universidad Autonoma de Madrid. Partial financial support from the Spanish Ministry of Economy and Competitiveness, Project ECO2015-70331-C2-1R and from Comunidad de Madrid, Project MadEco-CM S2015/HUM-3444 is acknowledged. Pilar Poncela, fellow of the UC3M-BS Institute of Financial Big Data (IFIBID), also thanks the Institute. Finally, we also thank Michela Nardo for providing the data.

References

Bai J, Ng S (2004) A PANIC attack on unit roots and cointegration. Econometrica 72:1127–1177

Bai J, Ng S (2008) Large dimensional factor analysis. Found Trends Econom 3:89–163

Bastos JA, Caiado J (2014) Clustering financial time series with variance ratio statistics. Quant Financ 14:2121–2133

Boivin J, Ng S (2006) Are more data always better for factor analysis? J Econom 132:169–194

Bollerslev T, Hood B, Lasse H, Pedersen LH (2018) Risk everywhere: modeling and managing volatility. Rev Financ Stud 31:2729–2773

Brockwell PJ, Davis RA (1991) Time series: theory and methods, 2nd edn. Springer, New York

Caiado J, Crato N (2010) Identifying common dynamic features in stock returns. Quant Financ 10:797–807

Caiado J, Crato N, Peña D (2006) A periodogram-based metric for time series classification. Comput Stat Data Anal 50:2668–2684

Caiado J, Crato N, Peña D (2009) Comparison of time series with unequal length in the frequency domain. Commun Stat Simul Comput 38:527–540

Caiado J, Maharaj EA, D'Urso P (2015) Time series clustering. In: Henning C, Meila M, Murtagh F, Rocci R (eds) Handbook of cluster analysis. CRC Press, Boca Raton, pp 241–263

Coates DS, Diggle PJ (1986) Tests for comparing two estimated spectral densities. J Time Ser Anal 7:7–20

Corsi F (2009) Heterogeneous autoregressive model of realized volatility (HAR-RV). J Financ Econom 7:174–196

Diggle PJ, Fisher NI (1991) Nonparametric comparison of cumulative periodograms. Appl Stat 40:423–434

Doz C, Giannone D, Reichlin L (2011) A two step estimator for large approximate dynamic factor models. J Econom 164(1):188–205

Doz C, Giannone D, Reichlin L (2012) A quasi maximum likelihood approach for large approximate dynamic factor models. Rev Econ Stat 94:1014–1024

Forni M, Hallin M, Lippi M, Reichlin L (2000) The generalized dynamic factor model: identification and estimation. Rev Econ Stat 82:540–554

Forni M, Hallin M, Lippi M, Reichlin L (2005) The generalized dynamic factor model: one-sided estimation and forecasting. J Am Stat Assoc 100:830–839

Galeano P, Peña D (2000) Multivariate analysis in vector time series. Resenhas 4:383–404

Lam C, Yao Q, Bathia N (2011) Estimation of latent factors using high-dimensional time series. Biometrika 98:901–918

Liao TW (2005) Clustering of time series data: a survey. Pattern Recognit 38:1857–1874

Maharaj EA (1996) A significance test for classifying ARMA models. J Stat Comput Simul 54:305–331

Otranto E (2010) Identifying financial time series with similar dynamic conditional correlation. Comput Stat Data Anal 54(1):1–15

Peña D, Box GEP (1987) Identifying a simplifying structure in time series. J Am Stat Assoc 82:836–843

Peña D, Poncela P (2006) Non-stationary dynamic factor analysis. J Stat Plan Inference 136:237–257

Piccolo D (1990) A distance measure for classifying ARIMA models. J Time Ser Anal 11:152–164

Poncela P, Ruiz E (2015) More is not always better: back to the Kalman filter in dynamic factor models. In: Shephard N, Koopman SJ (eds) Unobserved components and time series econometrics. Oxford University Press, Oxford

Stock JH, Watson MW (2002) Forecasting using principal components from a large number of predictors. J Am Stat Assoc 97:1169–1179

Stock JH, Watson MW (2011) Dynamic factor models. In: Clements MP, Hendry DF (eds) Oxford handbook of economic forecasting. Oxford University Press, Oxford

Thomson William (1881) The tide gauge, tidal harmonic analyser, and tide predicter. Proc Inst Civ Eng 65:2–25

Tong H, Dabas P (1990) Cluster of time series models: an example. J Appl Stat 17:187–198

Yang AC, Tsai S-J, Hong C-J, Wang C, Chen T-J, Liou Y-J et al (2011) Clustering heart rate dynamics is associated with β-adrenergic receptor polymorphisms: analysis by information-based similarity index. PLoS ONE 6(5):e19232

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.