Bernoulli **27**(1), 2021, 469–501
https://doi.org/10.3150/20-BEJ1246

A similarity measure for second order properties of non-stationary functional time series with applications to clustering and testing

ANNE VAN DELFT[1] and HOLGER DETTE[2]

[1] *Department of Statistics, Columbia University, 1255 Amsterdam Avenue, New York, NY 10027, USA. E-mail: anne.vandelft@columbia.edu*
[2] *Ruhr-Universität Bochum, Fakultät für Mathematik, 44780 Bochum, Germany. E-mail: holger.dette@rub.de*

Due to the surge of data storage techniques, the need for the development of appropriate techniques to identify patterns and to extract knowledge from the resulting enormous data sets, which can be viewed as collections of dependent functional data, is of increasing interest in many scientific areas. We develop a similarity measure for spectral density operators of a collection of functional time series, which is based on the aggregation of Hilbert–Schmidt differences of the individual time-varying spectral density operators. Under fairly general conditions, the asymptotic properties of the corresponding estimator are derived and asymptotic normality is established. The introduced statistic lends itself naturally to quantify (dis)-similarity between functional time series, which we subsequently exploit in order to build a spectral clustering algorithm. Our algorithm is the first of its kind in the analysis of non-stationary (functional) time series and enables to discover particular patterns by grouping together 'similar' series into clusters, thereby reducing the complexity of the analysis considerably. The algorithm is simple to implement and computationally feasible. As a further application, we provide a simple test for the hypothesis that the second order properties of two non-stationary functional time series coincide.

Keywords: clustering; functional data; local stationarity; spectral analysis; time series

1. Introduction

The surge in data storage techniques over the past two decades has led to more and more data sets that are almost continuously recorded on their domain of definition. The development of tools to model these type of data is the main focus of functional data analysis. In functional data analysis, the variables of interest are perceived as random smooth functions that vary on a continuum D, that is, $X(\tau)$, $\tau \in D$. While the intrinsically infinite variation of such random functions can be considered a rich source of information, extracting relevant information to identify patterns becomes more and more a challenge. This is especially the case when the data are collected sequentially over time and the curves exhibit serial dependence, that is, when the data set consists of a collection of *d functional time series*, $\{X_{i,t}(\tau) : \tau \in D\}_{t\in\mathbb{Z}, i\in 1,\dots,d}$. This type of data arises naturally in a wide range of scientific disciplines such as astronomy, biology, finance, meteorology, medicine or yet engineering (see, for example, Aston and Kirch [3], Zhang and Shao [65], Tavakoli and Panaretos [55] for applications in brain imaging, molecular biophysics and climatology, respectively).

In addition, in most real-world applications, the second order characteristics of time series change gradually over time. In meteorology, the distribution of the daily records of temperature, precipitation and cloud cover for a region – viewed as three related functional surfaces – may change over time due to global climate changes. Other relevant examples appear in the study of cognitive functions such as high-resolution recordings from local field potentials, EEG and MEG or from the financial industry

1350-7265 © 2021 ISI/BS

where implied volatility of an option as a function of moneyness changes over time. The development of appropriate exploratory techniques that allow to discover patterns or anomalies is therefore of foremost interest for this type of data.

The most widely used technique for this preliminary step of data exploration is known as *cluster analysis*. Clustering is concerned with partitioning a data set into a set of disjoint homogeneous groups (clusters) of realizations. Unlike supervised learning, clustering does not rely on prior knowledge of the groups or on building classifiers based upon a training set. It is therefore especially meaningful when little is known about the nature of the process and the data set is large.

A large body of literature on clustering (and related learning techniques) of ordinary time series has been published. Depending on the goal of the application, clustering algorithms can differ in a variety of aspects such as the representation of the data, how similarities are measured, and in the way clusters are constructed. The first two aspects are known to be crucial in terms of efficiency and accuracy of the solution and this is what most research focuses on (see Section 5 of Aghabozorgi, Sirkhorshidi and Wah [2], for a full taxonomy of the different aspects of clustering time series). For instance, in parametric approaches clusters are usually built based upon similarity of their estimated parameters (see, e.g., Kalpakis, Gada and Puttagunta [40], Corduas and Piccolo [15]), some of which take a Bayesian approach (Bauwens and Rombouts [5], Frühwirth-Schnatter and Kaufmann [26], Juárez and Steel [38]). Nonparametric methods are often based on comparing similarity of the estimated power spectra, which is a research topic in its own right (see, e.g., Coates and Diggle [14], Eichler [21], Dette [19], Dette and Hildebrandt [20], Jentsch and Pauly [37]). This approach is among others taken in Kakizawa, Shumway and Taniguchi [39], Savvides, Promponas and Fokianos [52], Fokianos and Promponas [25], Holan and Ravishanker [31], and also in Euán, Ombao and Ortega [22], who consider clustering time series based on a total variation distance between spectral densities. A wavelet-based approach can be found in Vlachos *et al.* [62].

Clustering and classification methods have also been extended to non-stationary time series. For example, Sakiyama and Taniguchi [51] use the framework of locally stationary time series (Dahlhaus [16]) for clustering while Chandler and Polonik [10] use it to develop a shape-based approach discriminant analysis. Another branch of literature focuses on piecewise stationary processes using Smooth Localized Complex EXponentials (SLEX) transforms – which were introduced by Ombao *et al.* [47] – or variations thereof (see, e.g., Huang, Ombao and Stoffer [33], Harvill, Kohli and Ravishanker [29] for clustering approaches and Böhm *et al.* [6] for classification of multivariate series).

In contrast to the Euclidean case, the literature on cluster analysis for functional data is less rich. Several methods have been developed for clustering i.i.d. functional data (Jacques and Preda [36], Chamroukhi and Nguyen [9], and references therein). A popular technique is to first reduce dimension by projecting the curves onto a basis of finite dimension and then to apply a standard classical clustering algorithm such as k-means (see Peng and Müller [49], Abraham *et al.* [1]). More recently, Delaigle, Hall and Pham [18] consider clustering the i.i.d. curves from two populations based upon differences in mean by applying a weighted k-means algorithm on a carefully chosen univariate projection of the data. Alternatively, nonparametric methods have been proposed that use specific distances for functional data (Ferraty and Vieu [23], Ieva *et al.* [34]) as well as parametric (Bayesian) approaches that assume the data are drawn from a particular probability distribution (see Jacques and Preda [35], Heard, Holmes and Stephens [30], among others). Another interesting recent approach that deals with sparsely clustered i.i.d. functional data is given by Floriello and Vitelli [24].

Despite of the vast amount of literature available on various data structures, existing methods are inappropriate to cluster possibly non-stationary functional time series. Compared to the i.i.d. case, the intrinsically infinite nature of the underlying functions is much more complex due to the fact that the variation in the process is not *static* as there is serial correlation between the functions. Hence, a clustering approach must be able to capture the complex within-curves dynamics as well as the between-curve dynamics. At the same time, it needs to be efficient to apply because each element of a functional time series is intrinsically high dimensional. Furthermore, we must take into account the fact that these dynamics are not necessarily temporally constant. In this article, we address this problem from several perspectives. We develop a new measure to compare the second order properties of non-stationary functional time series. This measure is based on the aggregation of Hilbert–Schmidt differences of the individual time-varying spectral density operators (see van Delft and Eichler [61]). Under fairly general conditions, the asymptotic properties of the corresponding estimator are derived and asymptotic normality is established. We then use this methodology for two purposes. Firstly, we consider this measure and its estimate to develop a new spectral clustering algorithm for functional time series, which are allowed to be non-stationary and do not require structural modeling assumptions such as linearity. Our algorithm is novel in the sense that, not only is it the first of its kind for exploratory analysis of functional time series, but moreover because – to the best of our knowledge – spectral clustering has also not yet been considered for Euclidean-valued time series. The underlying principle of spectral clustering is to reformulate the problem into a graph partitioning problem (see Figure 3). Geometrical properties of graphs can be conveniently described by the spectral properties of the corresponding graph Laplacian (see, e.g., Chung [13], Chung and Radcliffe [12], von Luxburg [63]). Using these properties, we can detect clusters in non-convex regions which classical clustering techniques may not be able to find. Furthermore, it can be solved efficiently via classical linear algebra operations. Broadly speaking, one can view spectral clustering as a dimension reduction technique that enhances the clustering properties before the actual clustering step. We will show that our introduced measure of similarity provides a meaningful basis for the adjacency matrix underlying the graph Laplacian. Secondly, we use this measure to develop a particularly simple level α-test for the hypothesis of equality of two time-varying spectral density operators, which uses the quantiles of the standard normal distribution.

The structure of this paper is as follows. We first introduce necessary notation and background on the type of processes considered in this paper. We then define a measure of similarity for a pair of functional time series and derive a consistent estimator to construct an empirical adjacency matrix. In Section 3, the spectral clustering algorithm is discussed in detail and it is shown that the algorithm based upon an empirical graph Laplacian – a transformation of the estimated adjacency matrix – is consistent. In Section 4, we discuss the application of hypothesis testing, whereas in Section 5 we study the properties of our algorithm in finite samples. The proofs of the main statements are provided in the Appendix. Several auxiliary results and an illustration of the clustering method by means of an application to high-resolution meteorological data can be found in the Online Supplement (van Delft and Dette [60]).

2. A measure of similarity

In this section, we introduce a measure of similarity for functional time series which is appropriate to use as a basis for a similarity matrix to cluster functional time series.

2.1. Notation

First, let us introduce some necessary notation. For a separable Hilbert space $\mathcal{H}$, we denote the inner product as $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{C}$ and its induced norm by $\| \cdot \|$. The Banach space of bounded linear operators $A : \mathcal{H} \to \mathcal{H}$ with operator norm $|||A|||_\infty = \sup_{\|x\| \le 1} \|Ax\|$ shall be denoted by $\mathcal{L}(\mathcal{H})$, the adjoint of $A \in \mathcal{L}(\mathcal{H})$ by $A^\dagger$. $A \in \mathcal{L}(\mathcal{H})$ is called self-adjoint if $A = A^\dagger$ and nonnegative definite if

$\langle Ax, x\rangle \geq 0$ for each $x \in \mathcal{H}$. If well-defined, we denote the trace of $A \in \mathcal{L}(\mathcal{H})$ by $\mathrm{Tr}(A)$. A compact operator $A \in \mathcal{L}(\mathcal{H})$ belongs to the Schatten class of order $1 \leq p < \infty$, denoted by $A \in S_p(\mathcal{H})$, if $|||A|||_p^p = \sum_{j\geq 1} s_j^p(A) < \infty$, where $\{s_j(A)\}_{j\geq 1}$ are the singular values of A. Operators that belong to the Banach spaces $(S_1(\mathcal{H}), |||\cdot|||_1)$ or $(S_2(\mathcal{H}), |||\cdot|||_2)$ will be referred to as trace-class operators and Hilbert–Schmidt operators, respectively. We remark in particular that $(S_2(\mathcal{H}), |||\cdot|||_2)$ is a Hilbert space with the inner product given by $\langle A, B\rangle_{\mathrm{HS}} = \mathrm{Tr}(AB^\dagger) = \sum_{j\geq 1}\langle Ae_j, Be_j\rangle$ for each $A, B \in S_2(\mathcal{H})$ and $\{e_j\}_{j\geq 1}$ an orthonormal basis of $\mathcal{H}$. For $f, g \in \mathcal{H}$, we define the tensor product $f \otimes g : \mathcal{H} \otimes \mathcal{H} \to \mathcal{H}$ as the bounded linear operator

$$(f \otimes g)v = \langle v, g\rangle f \quad \forall v \in \mathcal{H}.$$

We consider the Hilbert space $\mathcal{H} = L^2_{\mathbb{C}}([0, 1])$ of equivalence classes of square integrable measurable functions $f : [0, 1] \to \mathbb{C}$ with inner product $\int_0^1 f(\tau)\overline{g(\tau)}\, d\tau$, where the complex conjugate of $x \in \mathbb{C}$ is denoted as usual by $\overline{x}$. Additionally, we denote $\mathcal{H}_{\mathbb{R}} = L^2_{\mathbb{R}}([0, 1])$. Since the mapping $\mathcal{T} : \mathcal{H} \otimes \mathcal{H} \to S_2(\mathcal{H})$ defined by the linear extension of $\mathcal{T}(f \otimes g) = f \otimes \overline{g}$ is an isometric isomorphism, it defines a Hilbert–Schmidt operator with the kernel in $L^2_{\mathbb{C}}([0, 1]^2)$ given by $(f \otimes g)(\tau, \sigma) = f(\tau)\overline{g(\sigma)}$ for each $\tau, \sigma \in [0, 1]$ in an L^2-sense. We refer to Section S1 of the Online Supplement for further details and background.

2.2. A measure of similarity for functional time series

The main object of this paper is a collection of d stochastic processes $\{X_{i,t,T}\}_{t=1,\dots,T;T\in\mathbb{N}}$ $i = 1, \dots, d$, for fixed d, that take values in $\mathcal{H}_{\mathbb{R}}$. This is is without loss of generality since the theory holds for any separable Hilbert space. We tacitly assume that the $X_{i,t,T}$ are zero-mean random elements defined on some common probability space $(\Omega, \mathcal{B}, \mathbb{P})$ with $\mathbb{E}\|X_{i,t,T}\|_2^2 < \infty$. We call such processes weakly stationary if the second order dynamics are invariant under time translations and hence can be described via a sequence of lag h covariance operators $C_h := \mathbb{E}(X_{t+h} \otimes X_t) = \mathbb{E}(X_h \otimes X_0)$, $\forall t, h \in \mathbb{Z}$, which are elements of $S_1(\mathcal{H})$. In this paper, the second order dynamics are assumed to be well-defined but are moreover allowed to *change over time*. Processes of this type fit the framework of *locally stationary functional time series* as defined in van Delft and Eichler [61] – which extends the concept of local stationarity (Dahlhaus [16]) to the function space – and encompasses weakly stationary functional processes as a subclass. Asymptotic properties of these processes can be described by so-called infill-asymptotics, such that, as $T \to \infty$, we obtain more and more observations at a local level.

Definition 2.1. A process $\{X_{t,T}\}_{t=1,\dots,T;T\in\mathbb{N}}$ is called *functional locally stationary* if, for all rescaled times $u = t/T \in [0, 1]$, there exists an $\mathcal{H}_{\mathbb{R}}$-valued strictly stationary process $\{X_t^{(u)} : t \in \mathbb{Z}\}$ such that

$$\|X_{t,T} - X_t^{(u)}\|_2 \leq \left(\left|\frac{t}{T} - u\right| + \frac{1}{T}\right) P_{t,T}^{(u)} \quad \text{a.s.} \tag{2.1}$$

for all $1 \leq t \leq T$, where $\{P_{t,T}^{(u)}\}_{t=1,\dots,T;T\in\mathbb{N}}$ is a positive real-valued process such that for some $\rho > 0$ and $C < \infty$ the process satisfies $\mathbb{E}(|P_{t,T}^{(u)}|^\rho) < C$ for all t and T and uniformly in $u \in [0, 1]$.

What we will exploit throughout this paper is that the full second order dynamics of the triangular array $\{X_{t,T}\}_{t=1,\dots,T;T\in\mathbb{N}}$ are then completely and uniquely characterized by the *time-varying spectral density operator*

$$\mathcal{F}_{u,\omega} := \frac{1}{2\pi}\sum_{h\in\mathbb{Z}} C_{u,h} e^{-\mathrm{i}\omega h}, \tag{2.2}$$

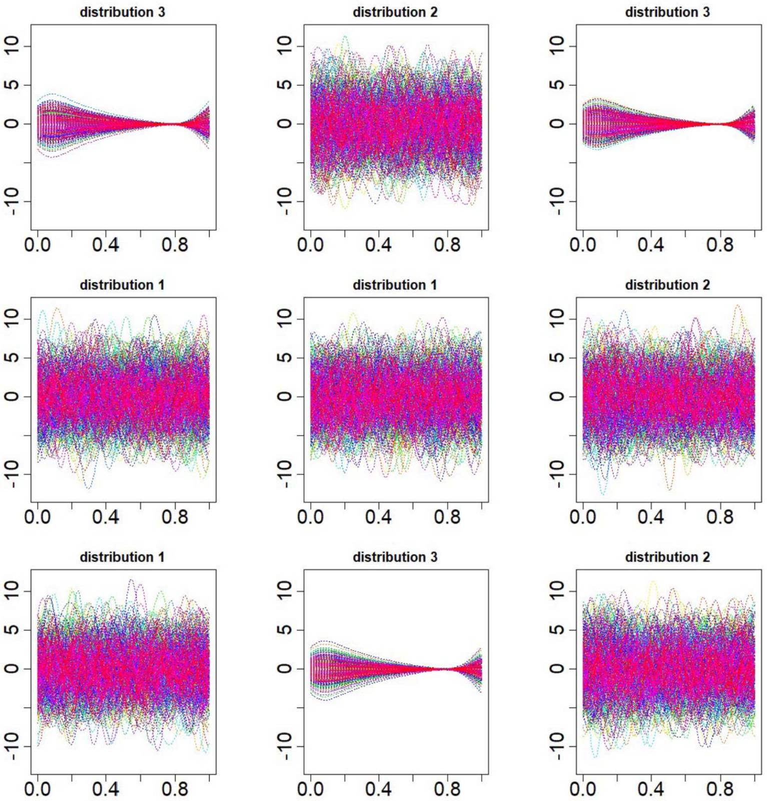


Figure 1. Functional time series from 3 different distributions. Each series $\{X_{i,t}(\tau) : \tau \in [0, 1]\}_{t=1}^{T}$, $i = 1, \ldots, 9$ consists of $T = 512$ curves on the domain $[0, 1]$.

where $C_{u,h} = \mathbb{E}(X_{t+h}^{(u)} \otimes X_t^{(u)})$ is the *local lag h covariance operator* at time u of the approximating process $\{X_t^{(u)} : t \in \mathbb{Z}\}$. For processes that satisfy equation (2.1) and $\sum_{h\in\mathbb{Z}} |||C_{u,h}|||_p < \infty$, (2.2) is a well-defined nonnegative definite hermitian element of $S_p(\mathcal{H})$ for each $u \in [0, 1]$, $\omega \in [-\pi, \pi]$. If the process is in fact weakly stationary, we can drop the dependence on local time and thus $\mathcal{F}_{u,\omega} \equiv \mathcal{F}_\omega$. This uniquely characterizing object of a functional time series lends itself naturally as a basis for a measure of similarity.

Example. We have generated 90 zero-mean functional time series uniformly from three groups (these are the models I, II, III described in detail in Section 5), which should be clustered according to their second order properties. Each series consists of $T = 512$ functions and exemplary we depict 9 series in Figure 1 (three from each distribution) to visualize the difficulties of this task. The second order structure of a stationary processes can be captured in $\mathbb{Z} \times \mathcal{H}_\mathbb{R} \times \mathcal{H}_\mathbb{R}$ in the time domain or $[-\pi, \pi] \times \mathcal{H} \times \mathcal{H}$ in the frequency domain and is hard to inspect visually. This is even more problematic for non-stationary functional time series. As can be seen, a visual inspection of the curves alone makes an assessment of the second order properties almost impossible for 6 of them, while 3 of them appear more obvious to distinguish from the other 6.

More specifically, let $\mathcal{F}^{(i_1)}_{u,\omega}$ and $\mathcal{F}^{(i_2)}_{u,\omega}$ denote the time-varying spectral density operator of processes $\{X_{i_1,t,T}\}_{t=1,\dots,T;T\in\mathbb{N}}$ and $\{X_{i_2,t,T}\}_{t=1,\dots,T;T\in\mathbb{N}}$, respectively. As a measure of pairwise similarity between two functional time series we consider

$$\mathcal{A}_{i_1,i_2} := \frac{\int_0^1 \int_{-\pi}^{\pi} |||\mathcal{F}^{(i_1)}_{u,\omega} - \mathcal{F}^{(i_2)}_{u,\omega}|||_2^2 \, d\omega \, du}{\int_0^1 \int_{-\pi}^{\pi} |||\mathcal{F}^{(i_1)}_{u,\omega}|||_2^2 + |||\mathcal{F}^{(i_2)}_{u,\omega}|||_2^2 \, d\omega \, du}. \tag{2.3}$$

Clearly, if this distance is zero then processes $\{X_{i_1,t,T}\}_{t=1,\dots,T;T\in\mathbb{N}}$ and $\{X_{i_2,t,T}\}_{t=1,\dots,T;T\in\mathbb{N}}$ must have the same second order properties and hence must belong to the same cluster. While other distance metrics than the Hilbert–Schmidt distance metric could be considered, distance metrics used in existing literature on Euclidean-valued data do not necessary lend themselves naturally to be generalized to infinite dimensional spaces. Furthermore, the embedding of the operators into a Hilbert space is beneficial in our context as it allows us to exploit geometrical properties and notions, both for theory and computation.

Proposition 2.1. *For processes that adhere to Definition 2.1, the distance $\mathcal{A}_{i_1,i_2}$ takes values in the interval* $[0, 1]$.

Proof. Observe that we can write the numerator as

$$\int_0^1 \int_{-\pi}^{\pi} |||\mathcal{F}^{(i_1)}_{u,\omega}|||_2^2 + |||\mathcal{F}^{(i_2)}_{u,\omega}|||_2^2 \, d\omega \, du - \int_0^1 \int_{-\pi}^{\pi} \langle \mathcal{F}^{(i_1)}_{u,\omega}, \mathcal{F}^{(i_2)}_{u,\omega} \rangle_{\mathrm{HS}} - \int_0^1 \int_{-\pi}^{\pi} \langle \mathcal{F}^{(i_2)}_{u,\omega}, \mathcal{F}^{(i_1)}_{u,\omega} \rangle_{\mathrm{HS}}.$$

The claim therefore follows from an application of the Cauchy Schwarz inequality and from the fact that the last two terms are nonnegative. For the latter, observe that $\mathcal{F}^{(i_1)}_{u,\omega}$ and $\mathcal{F}^{(i_2)}_{u,\omega}$ are Hermitian and nonnegative definite. Separability of $\mathcal{H}$ therefore ensures these have a real-valued discrete spectrum. More specifically, these operators admit an eigendecomposition with nonnegative eigenvalues, say $\{\nu^{(i_1)}_{u,\omega,j}\}_{j\geq 1}$ and $\{\nu^{(i_2)}_{u,\omega,k}\}_{k\geq 1}$, respectively. The composite operator $\mathcal{F}^{(i_1)}_{u,\omega}\mathcal{F}^{(i_2)}_{u,\omega}$ is also a well-defined element of $S_1(\mathcal{H})$ and hence has a finite trace. Using the properties of the Hilbert–Schmidt inner-product and the tensor product, we find

$$\langle \mathcal{F}^{(i_1)}_{u,\omega}, \mathcal{F}^{(i_2)}_{u,\omega} \rangle_{\mathrm{HS}} = \mathrm{Trace}\big(\mathcal{F}^{(i_1)}_{u,\omega}\mathcal{F}^{(i_2)}_{u,\omega}\big) = \sum_{j,k=1}^{\infty} \nu^{(i_1)}_{u,\omega,j} \nu^{(i_2)}_{u,\omega,k} \big|\langle \phi^{(i_2)}_{u,\omega,k}, \phi^{(i_1)}_{u,\omega,j} \rangle\big|^2 \geq 0,$$

where $\{\phi^{(i_1)}_{u,\omega,j}\}_{j\geq 1}$ and $\{\phi^{(i_2)}_{u,\omega,k}\}_{k\geq 1}$ denote the eigenfunctions of $\mathcal{F}^{(i_1)}_{u,\omega}$ and $\mathcal{F}^{(i_2)}_{u,\omega}$, respectively. □

The local scaling via the denominator is an essential aspect for its usage in a spectral clustering procedure. Differences in scales can lead spectral clustering to fail. Most similarity graphs rely upon a global scaling parameter of which the optimal value is difficult to determine and which can highly affect the clustering performance (see von Luxburg [63]). By accounting for local scales, our method avoids this issue. We discuss this further in Section 5.

Squared L^2-distances are quite popular in statistical inference for time series. In the context of functional data analysis van Delft, Characiejus and Dette [59] recently used an (unweighted) L^2-distance to measure deviations from stationarity. We emphasize however that the measure (2.3) is of a very different nature. On the one hand, it does not vanish for stationary time series as it is constructed to compare the second order structure of two possibly non-stationary functional time series. On the other hand, and more importantly, it uses a scaling which is of particular importance for the clustering procedure constructed below. As a consequence, the investigation of the stochastic properties of corresponding estimators of (2.3) is by no means trivial (see Appendix A and B and the Online Supplement for more details).

For the consistent estimation of $\mathcal{A}_{i_1,i_2}$ we split the sample into M blocks with N elements inside each of these blocks so that $T = MN = M(T)N(T)$ for each $T \in \mathbb{N}$, where $M \in \mathbb{N}$ and N is an even number. M and N correspond to the number of terms used in a Riemann sum approximating the integrals in (2.3) with respect to du and $d\omega$ and therefore they have to be reasonable large. The functional discrete Fourier transform (fDFT) at time point u, is a random function with values in $L^2_{\mathbb{C}}([0,1])$ defined by

$$D_i^{u,\omega} := \frac{1}{\sqrt{2\pi N}} \sum_{s=0}^{N-1} X_{i,\lfloor uT \rfloor - N/2+s+1,T} e^{-\mathrm{i}\omega s}. \tag{2.4}$$

The local periodogram tensor for the i-th time series is given by

$$I_i^{u,\omega} := D_i^{u,\omega} \otimes D_i^{u,\omega}, \tag{2.5}$$

for $i = 1, \dots, d$. We base our estimator upon a linear combination of the following Hilbert–Schmidt inner products

$$F_{i_1 i_2} = \frac{1}{T} \sum_{j=1}^{M} \sum_{k=1}^{\lfloor N/2 \rfloor} \langle I_{i_1}^{u_j,\omega_k}, I_{i_2}^{u_j,\omega_{k-1}} \rangle_{\mathrm{HS}}. \tag{2.6}$$

In particular, a suitable and (symmetric) estimator for the distance (2.3) is given by

$$\hat{\mathcal{A}}_{i_1,i_2} := \frac{F_{i_1 i_1} + F_{i_2 i_2} - F_{i_1 i_2} - F_{i_2 i_1}}{(F_{i_1 i_1} + F_{i_2 i_2})}. \tag{2.7}$$

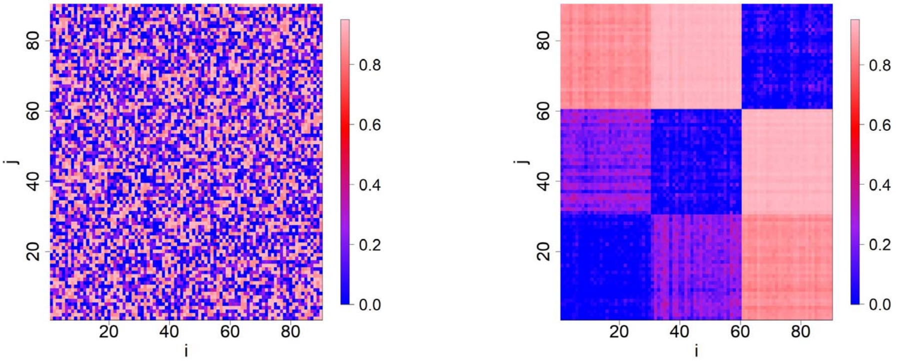
To ease notation, we provide empirical quantities with $\hat{\cdot}$. The dependence of these quantities on T is therefore implicit. We obtain under suitable regularity conditions, which are postponed to Section 4, the following result.

Theorem 2.1 (consistency). *Suppose Assumption 4.1 with $m = 8$ and Assumption 4.2 hold. Then*

$$\hat{\mathcal{A}}_{i_1,i_2} - \mathcal{A}_{i_1,i_2} = O_p(T^{-1/2}).$$

We remark that under suitable moment conditions the estimator is moreover asymptotically multivariate normal (Theorem 4.1) and therefore lends itself for other statistical applications such as a test for equality of time-varying spectral density operators. We shall briefly discuss this application together with more details on the statistic $\hat{\mathcal{A}}_{i_1,i_2}$ in Section 4. In the next section, we define a similarity graph based upon the measure (2.7) and introduce a spectral clustering algorithm to cluster the functional time series.

Remark 2.1. The current approach is based on clustering the different series by means of similarity of the *full* second order structure, that is, over all time and frequency. However, the result can be shown to hold true when we restrict this to a given time-frequency band, provided the sample length is split appropriately (see Assumption 4.2). By replacing the scaling factor in (2.6) with $1/N$ and dropping the sum over j, we can moreover obtain a $\sqrt{N}$-consistent estimator at a given time point, that is, for the measure in (2.3) with the outer integrals removed. This can be shown in spirit of Theorem A.1. However, in order to make the estimator pointwise consistent in frequency direction, one has to smooth the local periodogram tensor over a certain frequency band using a kernel smoother which requires an



(a) *without information on the clusters* (b) *with information on the clusters*

Figure 2. Heat map of the value of the estimates $\hat{\mathcal{A}}_{i,j}$ of $\mathcal{A}_{i,j}$ plotted for all pairs $i, j = 1, \dots, 90$ in random order (left); and ordered by cluster (right).

additional tuning parameter (see, e.g., Tavakoli and Panaretos [55], van Delft [58], and references therein).

Example (continued). For 6 of the 9 functional time series depicted in Figure 1, it is difficult to visually distinguish the series and their second order properties. This becomes infeasible for all 90 series. However, we can use the measure $\hat{\mathcal{A}}$ to identify similarities. A heat map of the corresponding estimates for all 90 time series are displayed in Figure 2(a). As can be seen, the empirical measure gives the pairs of time series different weights ranging between 0 and 1, but it is difficult to identify any structure. For the sake comparison, we ordered the times series assuming knowledge of the clusters in the right part of Figure 2(b). We observe that the similarity measure $\hat{\mathcal{A}}$ makes the clusters visible.

3. Spectral clustering of functional time series

In this section, we develop a spectral clustering algorithm which consists of a several steps. We start by translating the problem of clustering d functional time series into k clusters into a graph partitioning problem using the previously defined similarity measure. Secondly, we construct a spectral embedding using an empirical graph Laplacian, which is shown to have spectral properties that converge to those of the population graph Laplacian. We then use the embedded points to cluster the data by means of a k-means algorithm and subsequently show that the number of misclustered points converges to zero as $T \to \infty$.

3.1. Construction of the graph

We construct an undirected similarity graph $G = (V, E)$, where V denotes the set of vertices and E denotes the set of edges. Denote the set $[d] := \{1, \dots, d\}$. To each family of random curves, $X_i := \{X_{i,t,T}\}_{t,T}$, $i \in [d]$, we relate a node $v_i \in V$ and describe the similarity between node v_{i_1} and v_{i_2} via nonnegative weights on the edges. These weights are given by the *empirical adjacency matrix* which is defined as the following transformation of the matrix $\hat{\mathcal{A}} = \{\hat{\mathcal{A}}_{i_1,i_2} : i_1, i_2 = 1, \dots, d\}$ in (2.7);

$$\hat{W} = e^{-\hat{\mathcal{A}}} \in \mathbb{R}^{d \times d}. \tag{3.1}$$

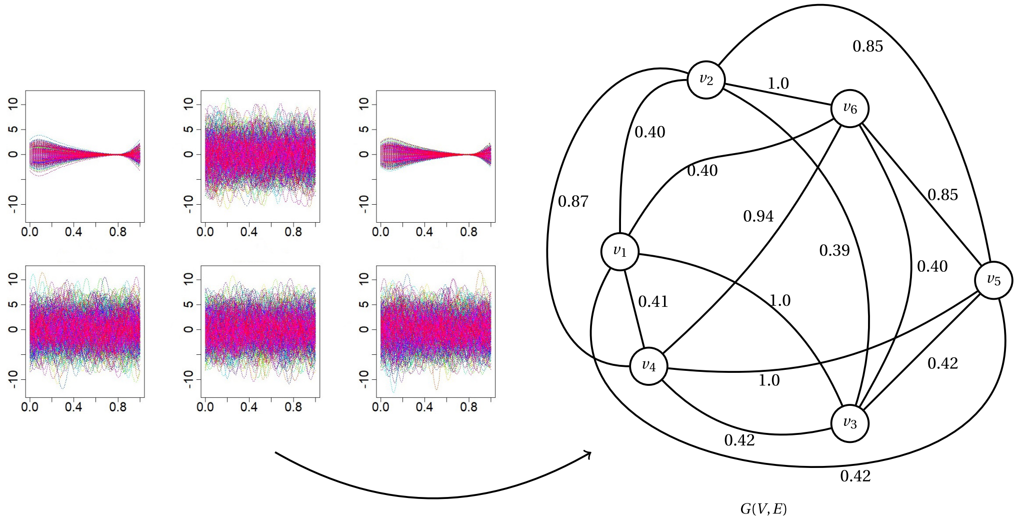


Figure 3. Illustration of the map from the space of functional time series (left) into a graph partitioning problem (right). The first 6 processes of Figure 1 are depicted in the left graph. The right graph gives the corresponding undirected similarity graph, where node v_i corresponds to process i. The value on the edge between node v_{i_1} and v_{i_2}, $i_1 \neq i_2 \in \{1, \dots, 6\}$ correspond to entry (i_1, i_2) of the matrix $\hat{W}$ in (3.1).

This is illustrated in Figure 3 for the first 6 processes depicted in Figure 1. The clustering problem then becomes equivalent to partitioning the similarity graph into connected components such that nodes with pairwise high weights on the edges are put into to the same component while nodes with low weights are put into different components.

We note that the theory developed in this paper is applicable for any continuous weight function. However, and as already mentioned below Proposition (2.1), the function (3.1) does not rely *explicitly* upon specification of a global scaling parameter. This is an important advantage compared to classical spectral clustering approaches, where the most common choice for the similarity graph is the classical Gaussian similarity function, and for which it is well-known the method can be very sensitive to the global scaling parameter (see, e.g., von Luxburg [63]). It will be demonstrated in Section 5.2 that our method is robust to the choice of global scaling parameters.

3.2. The spectral embedding

The main ingredient to our algorithm is the empirical graph Laplacian

$$\hat{L} = I - \hat{D}^{-1/2}\hat{W}\hat{D}^{-1/2}, \tag{3.2}$$

where I denotes the identity matrix, $\hat{D} = \operatorname{diag}(\{\hat{D}_i\}_{i=1}^{d})$ denotes the *degree matrix* of $\hat{W}$ which carries the degree of vertex v_i as its i-th diagonal element, that is, $\hat{D}_i = \sum_{j=1}^{d} \hat{W}_{i,j}$. This Laplacian can be viewed as a perturbed version of the unknown population Laplacian

$$L = I - D^{-1/2} W D^{-1/2}, \tag{3.3}$$

where $W = e^{-\mathcal{A}}$ is the population adjacency matrix and D is the corresponding degree matrix. Note that we use a normalized Laplacian instead of the matrix $\tilde{L} = D - W$. This has several advantages and especially shows a better performance if d is large (von Luxburg, Belkin and Bousquet [64]) and can also be applied to irregular graphs, i.e., graphs of which the vertices have different degrees.

There exists a whole field dedicated to the study of the matrices L and $\tilde{L}$, called *spectral graph theory* (see, e.g., Chung [13]) and in the following we briefly summarize the important properties necessary for our approach. A good summary for statistical purposes can be found in the tutorial of von Luxburg [63]. The matrix L in (3.3) is symmetric and positive-definite and therefore has an eigendecomposition, say $L = U^\top \Lambda U$ with $\Lambda = \operatorname{diag}(\{\lambda_i\}_{i=1}^d) \in \mathbb{R}_{\geq 0}^{d\times d}$. In the case where $G = (V, E)$ is an undirected weighted graph with nonnegative weights, its spectrum provides information on the connectivity of the graph. More specifically, 0 is an eigenvalue of the matrix L and its multiplicity, say k, equals the number of connected components $G_1, \dots, G_k$ in the graph G. Furthermore, the eigenspace of the eigenvalue 0 is spanned by the vectors $D^{1/2}\mathbb{1}_{G_i} i = 1, \dots, k$, where $\mathbb{1}_{G_i} \in \mathbb{R}^d$ denotes the indicator vector on component G_i (see, e.g., Chung [13]).

To understand the usefulness of these properties, suppose for a moment that our graph has exactly k disconnected components where the nodes belonging to different components are infinitely far apart, that is, have zero weight. If we collect the k eigenvectors that belong to the eigenvalue 0 of the matrix L and subsequently row-normalize this matrix, we obtain a matrix of indicator vectors

$$\mathcal{U} := [\mathbb{1}_{G_1}, \dots, \mathbb{1}_{G_k}] \in \mathbb{R}^{d\times k}. \tag{3.4}$$

Per row of $\mathcal{U}$, there will be exactly one nonzero element indicating the component to which it belongs to. In practice, one never encounters the ideal situation that the nullspace of the empirical graph Laplacian is perfectly spanned by (scaled) indicator vectors because the empirical similarity graph, by construction, consists of only one connected component. Nevertheless, the information about the structure of k clusters is still completely contained in the eigenvectors that belong to the smallest k eigenvalues of empirical graph Laplacian $\hat{L}$. In particular, these eigenvectors provide a relaxation solution to the *Normalized minimal cut* problem (Shi and Malik [53]), which has the objective to partition the graph into k disjoint components by 'cutting' it at the edges of which the total sum of normalized weights (relative to the volume of the partitioning component) is minimized. In order to exploit this information, we embed the infinite-dimensional processes X_i into the space spanned by the k eigenvectors that belong to the k smallest eigenvalues by representing the i-th process by the i-th row of $\mathcal{U}$. The embedded points then provide a representation of X_i in $\mathbb{R}^k$ of which the clustering properties are enhanced. As a result, a simple algorithm such as k-means can be applied to the embedded points to identify the clusters.

To be precise, let $\hat{\mathcal{U}}_{\cdot,1}, \dots, \hat{\mathcal{U}}_{\cdot,k} \in \mathbb{R}^d$ denote the row-normalized k eigenvectors of the empirical graph Laplacian $\hat{L}$ defined in (3.2) corresponding to the smallest k eigenvalues. Note that the matrix $\hat{L}$ is an estimator of the Laplacian (3.3) and shares some of the nice properties of the matrix L. If the estimator $\hat{L}$ is consistent (as shown below), one can expect that the k smallest eigenvalues (counted with their multiplicities) are close to 0. However, since $\hat{L}$ is a 'perturbed' Laplacian, its eigenvectors corresponding to eigenvalues with multiplicity larger than 1 can only be identified up to an orthogonal rotation, while those with multiplicity 1 can be identified up to a sign. As explained below, row normalization avoids this additional source of unidentifiability among the eigenvectors corresponding to the k smallest eigenvalues. These row-normalized eigenvectors can therefore be uniquely identified. In order to guarantee that the embedding of our data as the rows of the matrix

$$\hat{\mathcal{U}} := [\hat{\mathcal{U}}_{\cdot,1}, \dots, \hat{\mathcal{U}}_{\cdot,k}] \in \mathbb{R}^{d\times k}, \tag{3.5}$$

is meaningful for clustering we will show that $\hat{L}$ is 'close' to L measured by a suitable norm such that the spectral properties of $\hat{L}$ converge to their population counterparts. The approach that we take to establish consistency of the empirical Laplacian as T tends to infinity is based on perturbation theory comparing the matrices $\hat{L}$ and L (see also Ng, Jordan and Weiss [46], Rohe, Chatterjee and Yu [50]). The proofs are technical and rely on several auxiliary results which are relegated to the Appendix. Using consistency of $\hat{\mathcal{A}}$ and the symmetry of L, we can show that the distance in operator norm between $\hat{L}$ and L converges to zero in probability.

Lemma 3.1. *Under the conditions of Theorem 2.1*

$$\forall \varepsilon > 0, \quad \lim_{T\to\infty} \mathbb{P}\big(|||\hat{L} - L|||_\infty > \varepsilon\big) = 0.$$

To analyze the concentration of $\hat{\mathcal{U}}$, we use a slight modification of the classical Davis–Kahan Sin Θ theorem (Stewart and Sun [54]). This theorem provides an upper bound on the sinus of the principal angles between the eigenspaces $\hat{U} O$ – for some orthonormal rotation matrix $O \in \mathbb{R}^{k\times k}$ – and U, in terms of the spectral grap δ, the dimension of the space k and on the size of the perturbation $|||H|||_\infty$. Rather, Lemma B.1 in the Appendix can be used to bound the Euclidean distance between the unnormalized empirical eigenvectors $\hat{U}$ and their population counterparts U up to rotation. For the row-normalized matrix $\hat{\mathcal{U}}$, we avoid the additional source of unidentifiability caused by the rotation matrix. We derive the following result.

Lemma 3.2. *The matrix $\hat{\mathcal{U}}$ defined in* (3.5) *satisfies*

$$\|\hat{\mathcal{U}} - \mathcal{U}\|_2 \le \frac{4\sqrt{k}}{\sqrt{\min_i \|U_{i,\cdot}\|_2^2}} \frac{|||\hat{L} - L|||_\infty}{\lambda_{k+1}} \le 4\sqrt{k}\sqrt{\frac{\mathcal{C}_{\max}}{\min_i D_i}} \frac{|||\hat{L} - L|||_\infty}{\lambda_{k+1}}$$

where λ_{k+1} is the $(k+1)$-th smallest eigenvalue of L and where $\mathcal{C}_{\max} = \max_i \sum_{i_1,i_2\in G_i} W_{i_1,i_2}$. Hence, under the conditions of Lemma 3.1,

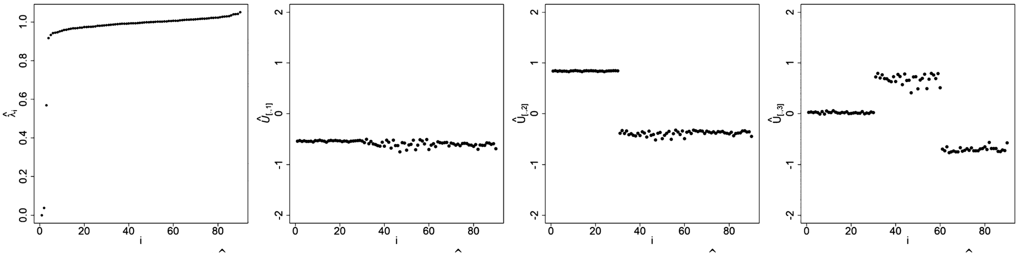
$$\forall \varepsilon > 0, \quad \lim_{T\to\infty} \mathbb{P}\big(\|\hat{\mathcal{U}} - \mathcal{U}\|_2 > \varepsilon\big) = 0.$$

These results thus justify to use the rows $\hat{\mathcal{U}}_{1,\cdot}, \dots, \hat{\mathcal{U}}_{d,\cdot}$ of (3.5) to embed the d functional time series. Each embedded point $\hat{\mathcal{U}}_{i,\cdot}$ then represents a process X_i (or node) in k dimensions, where these k dimensions can be seen as the features. Based on this representation, one can cluster the data using k-means. This step is analyzed next. We remark that our treatment of the spectral embedding by means of row-normalized eigenvectors of the graph Laplacian is therefore similar to Ng, Jordan and Weiss [46], who first investigated the use of a symmetric Laplacian in the context of spectral clustering.

Example (continued). For the example, the eigenvalues in Figure 4(a) indicate the empirical graph Laplacian has one connected component. This is not surprising as we have a fully connected graph. The first eigenvector is therefore approximately constant, which can be seen in Figure 4(b), while the second and third are approximately constant on the clusters. The second eigenvector (Figure 4(c)) allows to separate the first 3 series from the rest, while the third eigenvector indicates to separate the last three from the first 6 series (Figure 4(d)).

3.3. Clustering the embedded points using k-means

In this section, we analyze the final step where k-means is applied to the embedded points $\hat{\mathcal{U}}_{1,\cdot}, \dots, \hat{\mathcal{U}}_{d,\cdot} \in \mathbb{R}^k$. We show that the k-means algorithm clusters with high probability the data correctly. More



(a) *eigenvalues of* $\hat{L}$ (b) *eigenvector* $\hat{\mathcal{U}}_{\cdot,1}$ (c) *eigenvector* $\hat{\mathcal{U}}_{\cdot,2}$ (d) *eigenvector* $\hat{\mathcal{U}}_{\cdot,3}$

Figure 4. The spectrum of $\hat{L}$ and the three row-normalized eigenvectors corresponding to the three smallest eigenvalues $\hat{\lambda}_1 \leq \hat{\lambda}_2 \leq \hat{\lambda}_3$.

specifically, we derive a non-asymptotic bound on the number of points that are misclustered and show under regularity conditions that this converges to zero as $T \to \infty$.

The k-means objective aims to partition the d embedded points $\hat{\mathcal{U}}_{1,\cdot}, \ldots, \hat{\mathcal{U}}_{d,\cdot} \in \mathbb{R}^k$ into k clusters $\{C_1, \ldots, C_k\}$ in such a way that the pairwise squared deviations of points within the same cluster is minimized. The algorithm thus returns the centroids $\{c_1^\star, \ldots, c_k^\star\}$ from the objective function

$$\min_{\{c_1,\ldots,c_k\}\subset\mathbb{R}^k} \sum_i \min_j \|\hat{\mathcal{U}}_{i,\cdot} - c_j\|_2^2.$$

The data points $\hat{\mathcal{U}}_{i,\cdot}$ and $\hat{\mathcal{U}}_{j,\cdot}$ are then put in the same cluster if $c_i^\star = c_j^\star$. More formally, but equivalently, for $\hat{\mathcal{U}}$ defined in (3.5) the k-means algorithm should return a matrix $C^\star \in \mathbb{R}^{d\times k}$ with at most k unique rows such that

$$C^\star = \underset{C\in\mathcal{M}(d,k)}{\arg\min} \|\hat{\mathcal{U}} - C\|_2^2. \tag{3.6}$$

where $\mathcal{M}(d,k) = \{M \in \mathbb{R}^{d\times k} : M \text{ has at most } k \text{ distinct rows}\}$.

To analyze the algorithm, we need to define what we mean with a point being correctly clustered. Let $C^\star$ as in (3.6), that is, the matrix returned from applying the k-means algorithm to $\hat{\mathcal{U}}$. Intuitively, a point $\hat{\mathcal{U}}_{i,\cdot}$ is correctly clustered if there is no other row of the population matrix $\mathcal{U}$ defined in (3.4) that is closer to $C_{i,\cdot}^\star$ than the i-th row. Using this intuition, we can provide a meaningful definition of the set of correctly clustered point (Lemma B.2) and hence of its complement set. The next theorem gives a non-asymptotic bound on the number of misclustered points.

Theorem 3.1. *Assume the graph has k components. Then the cardinality of the set of misclustered points, denoted by* $|\Sigma|$, *satisfies*

$$|\Sigma| \leq \iota k \frac{1}{\min_i \|U_{i,\cdot}\|_2^2} \frac{|||\hat{L} - L|||_\infty^2}{\lambda_{k+1}^2} \leq \iota k \frac{\mathcal{C}_{\max}}{\min_i D_i} \frac{|||\hat{L} - L|||_\infty^2}{\lambda_{k+1}^2} \tag{3.7}$$

for some constant $\iota > 0$ *and where* $\mathcal{C}_{\max} = \max_i \sum_{i_1\in G_i} \sum_{i_2\in G_i} W_{i_1,i_2}$ *denotes the maximum sum of all entries of any of the components. If the conditions underlying Theorem* 2.1 *hold, then* $|\Sigma|$ *converges to zero in probability as* $T \to \infty$.

This upper bound implies that the probability of misclustering is affected by various properties of the (data-induced) similarity graph. Firstly, one can see it is an increasing function of the number

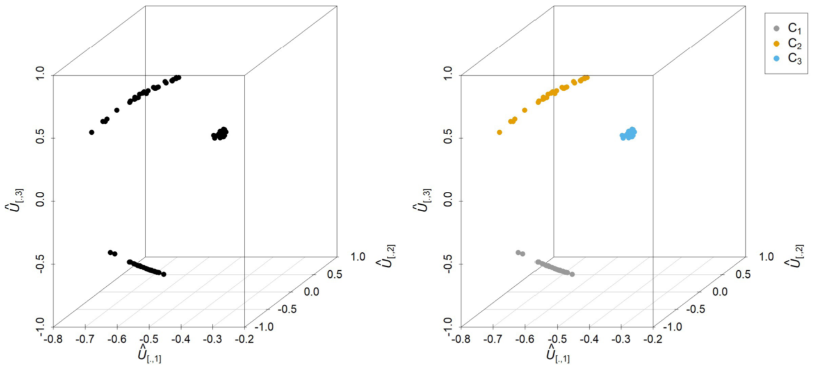


Figure 5. The embedded points $\hat{\mathcal{U}}_{[1,\cdot]}, \dots, \hat{\mathcal{U}}_{[90,\cdot]}$ (left) and the result of applying k-means to these points where the color indicates the cluster the point belongs to (right).

of clusters k. Secondly, it is a decreasing function of the minimum degree. In particular, the second inequality implies that for an unbalanced graph, as measured by the maximal sum of all entries of any of the groups relative to the minimal degree $\min_i D_i$, the probability to miscluster has a less tight upper bound. In other words, if the graph contains isolated vertices and has many points that are highly concentrated, we can expect a higher probability to miscluster. Thirdly, it is a decreasing function of the distance between the zero eigenvalues and the first nonzero eigenvalue λ_{k+1}. Finally, it is affected by the accuracy of $\hat{L}$ as an estimator for L.

Example (continued). For the example, the embedded points are shown in the left graph of Figure 5. Applying k-means identifies all clusters exactly, where the colors indicate the cluster the respective data points belong to. The data have in fact been represented in such an effective way that a simple k-means algorithm can 'easily' identify the clusters.

3.4. The choice of *k*

So far, we have considered the case where the number of clusters k is known. In the remaining part of this section we briefly discuss a data driven choice of k. This problem has received much attention within the clustering literature and a wide range of methods has been developed (see Gordon [27], Theodoridis and Koutroumbas [56], for overviews) and compared (see Milligan and Cooper [45], Tibshirani, Walther and Hastie [57], and references therein). Most methods to pick k are formulated in terms of optimizing a relative criterion. For different choices of k, the quality of the clustering structure is evaluated according to some measure, such as the intra-cluster dispersion. The number of clusters is then specified to be the value of k for which this criterion is optimized.

There is however no universally optimal method because the definition of 'optimal' number of clusters can highly depend on the application and on the method used to cluster to data. If the data is well separated and there are clear distinguishable clusters then there are several successful methods that will correctly detect the underlying clusters. However, in noisy data sets with overlapping clusters, different methods will detect different number of clusters. In the case of spectral clustering, an intuitive alternative would be to pick the number of clusters k such that the first $\lambda_1, \dots, \lambda_k$ eigenvalues of $\hat{L}$ are small but λ_{k+1} is 'relatively' large. This 'eigengap' heuristic can be justified through the spectral properties of the population graph Laplacian and Lemma 3.1. In practice, such approaches are also known to be very sensitive to the construction of the similarity graph and can quickly fail unless the data is well separated. Deemed therefore an unsolvable problem, it is not uncommon to apply multiple criteria and pick the criteria that works best for the particular problem at hand (see Charrad *et al.* [11], for an implementation of various criteria).

A thorough development of a new method would be beyond the scope of this paper. In our empirical study in Section 5, we investigate the performance of available methods to choose k for our particular algorithm. Additionally, we consider two variations of the eigengap heuristic each with a different interpretation of a ‘relatively large’ eigengap.

4. Testing for equality

Besides from the application of the similarity measure as a basis for spectral clustering of functional time series, a well-defined limiting distribution allows it to be meaningful in a variety of statistical applications and in particular for the construction of hypothesis tests. The problem to detect similarities or to compare time series is of interest in a wide range of scientific fields and for classical time series a variety of methods have been proposed (see references in the introduction and examples therein). In case of function-valued time series, the literature is less well developed. Horváth, Hušková and Rice [32] proposed a procedure to test the hypothesis that two sets of functional data are identical and independently distributed using the sum of L^2-distances of the sequence of correlation functions. Tavakoli and Panaretos [55] instead proposed a test between two stationary functional time series based upon the Hilbert–Schmidt norm of the difference of the sample spectral density operators restricted to a Hilbert–Schmidt space of finite dimension. Bootstrap-based methods to test for equality of mean functions or covariance operators are proposed in Paparoditis and Sapatinas [48] and, more recently, Leucht, Paporoditis and Sapatinas [44] discussed a test for the equality of spectral density operators for linear functional time series.

To the best of our knowledge, no procedure is available that allows to test for similarities between functional time series of which the second order structure is allowed to be time-dependent. In this section, we develop such a test using the previously defined similarity measure $\mathcal{A}_{i_1,i_2}$ in (2.3). For the sake of brevity, we restrict ourselves to the case of two functional time series and consider for a fixed pair (i_1, i_2) (with $i_1 \neq i_2$) the hypothesis

$$H_0 : \mathcal{F}^{(i_1)}_{u,\omega} \equiv \mathcal{F}^{(i_2)}_{u,\omega} \quad \text{a.e. on } [-\pi, \pi] \times [0, 1] \tag{4.1}$$

versus

$$H_a : \mathcal{F}^{(i_1)}_{u,\omega} \neq \mathcal{F}^{(i_2)}_{u,\omega} \text{ on a subset of } [-\pi, \pi] \times [0, 1] \text{ of positive Lebesgue measure.} \tag{4.2}$$

The similarity measure in (2.3) lends itself quite naturally to test this hypothesis, that is, we can equivalently formulate the hypothesis as

$$H_0 : \mathcal{A}_{i_1,i_2} = 0 \quad \text{versus} \quad H_a : \mathcal{A}_{i_1,i_2} > 0. \tag{4.3}$$

By Theorem 2.1, the statistic $\hat{\mathcal{A}}_{i_1,i_2}$ defined in (2.7) is a consistent estimator of the normalized distance $\mathcal{A}_{i_1,i_2}$. Therefore, it is reasonable to reject the null hypothesis for large values of the estimator $\hat{\mathcal{A}}_{i_1,i_2}$. In order to derive the distributional properties of $\hat{\mathcal{A}}_{i_1,i_2}$, we require the following assumptions on the functional processes $\{X_{i,t,T} : t \in \mathbb{Z}\}_{T\in\mathbb{N}}$, $i = 1, \dots, d$.

Assumption 4.1. Assume $\{X_{i,t,T} : t \in \mathbb{Z}\}_{T\in\mathbb{N}}$, $i \in [d]$, are d locally stationary zero-mean stochastic processes taking values in $\mathcal{H}_{\mathbb{R}}$ and, for even $m \in \mathbb{N}$, let $\kappa_{m;t_1,\dots,t_{m-1}} : L^2([0, 1]^{m/2}) \to L^2([0, 1]^{m/2})$ be a positive operator independent of T such that, for all $j = 1, \ldots, m-1$ and some $\ell \in \mathbb{N}$,

$$\sum_{t_1,\ldots,t_{m-1} \in \mathbb{Z}} \left(1 + |t_j|^{\ell}\right) |||\kappa_{m;t_1,\ldots,t_{m-1}}|||_1 < \infty. \tag{4.4}$$

Let us denote

$$Y_{i,t}^{(T)} = X_{i,t,T} - X_{i,t}^{(t/T)} \quad \text{and} \quad Y_{i,t}^{(u,v)} = \frac{X_{i,t}^{(u)} - X_{i,t}^{(v)}}{(u-v)} \tag{4.5}$$

for $T \geq 1$, $1 \leq t \leq T$ and $u, v \in [0,1]$ such that $u \neq v$. Suppose furthermore that the m-th order joint cumulant tensors satisfy

(i) $|||\mathrm{Cum}(X_{i_1,t_1,T}, \ldots, X_{i_{m-1},t_{m-1},T}, Y_{i_m,t_m}^{(T)})|||_1 \leq \frac{1}{T} |||\kappa_{m;t_1-t_m,\ldots,t_{m-1}-t_m}|||_1$;

(ii) $|||\mathrm{Cum}(X_{i_1,t_1}^{(u_1)}, \ldots, X_{i_{m-1},t_{m-1}}^{(u_{m-1})}, Y_{i_m,t_m}^{(u_m,v)})|||_1 \leq |||\kappa_{m;t_1-t_m,\ldots,t_{m-1}-t_m}|||_1$;

(iii) $\sup_u |||\mathrm{Cum}(X_{i_1,t_1}^{(u)}, \ldots, X_{i_{m-1},t_{m-1}}^{(u)}, X_{i_m,t_m}^{(u)})|||_1 \leq |||\kappa_{m;t_1-t_m,\ldots,t_{m-1}-t_m}|||_1$;

(iv) $\sup_u |||\frac{\partial^{\ell}}{\partial u^{\ell}} \mathrm{Cum}(X_{i_1,t_1}^{(u)}, \ldots, X_{i_{m-1},t_{m-1}}^{(u)}, X_{i_m,t_m}^{(u)})|||_1 \leq |||\kappa_{m;t_1-t_m,\ldots,t_{m-1}-t_m}|||_1$.

We remark that if the process is in fact stationary, the dependence on localized time u drops. Furthermore, these assumptions are intricately related to the existence of moments (see Lemma S2.1 of the online supplement). As explained in Section 2, the estimator requires splitting the sample $T \in \mathbb{N}$ as $T = N(T)M(T)$, where $N(T)$ defines the resolution in frequency of the local fDFT and $M(T)$ controls the number of non-overlapping local fDFT's in (2.6). Since these correspond to the number of terms used in a Riemann sum approximating the integrals with respect to du and $d\omega$ they have to be sufficiently large. We assume

Assumption 4.2. $M \to \infty$, $N \to \infty$ as $T \to \infty$, such that

$$N/M \to \infty \quad \text{and} \quad N/M^3 \to 0.$$

The number of elements in the blocks grows therefore must grow faster than the number of blocks, but slower than the cube number of blocks. Observe that the uncertainty principle implies that accuracy of estimation is limited by the reciprocal relationship that exists between time and frequency resolution. A high degree of nonstationarity would benefit from more blocks to resolve the position of the energy dispersion which, for a fixed length of data, means less data is available to resolve the peak in frequency direction, and vice versa. Smoothness of the mapping $(u,\omega) \mapsto \mathcal{F}_{u,\omega}$ in both directions affects the sensitivity to the specified parameter. The choice of the number of blocks is carefully discussed in van Delft, Characiejus and Dette [59], who used localized integrated periodogram operators as a basis for a stationarity test, and showed that the resulting procedure is fairly robust with respect to the choice of M and N. The following result gives the asymptotic distribution of $\hat{\mathcal{A}}_{i_1,i_2}$.

Theorem 4.1. *Suppose that Assumption 4.1 with* $m \geq 1$ *and Assumption 4.2 hold. Then,*

$$\left\{\sqrt{T}(\hat{\mathcal{A}}_{i_1,i_2} - \mathcal{A}_{i_1,i_2})\right\}_{\{i_1,i_2 \in [d]\}} \to \mathcal{N}(\mathbf{0}, \mathbf{\Gamma}) \quad \text{as } T \to \infty,$$

where $\mathbf{0} \in \mathbb{R}^d$ *and* $\mathbf{\Gamma}$ *is a positive definite element of* $\mathbb{R}^{d \times d}$.

Under the null hypothesis, the asymptotic variance reduces to a very succinct form in case the processes are moreover independent.

Theorem 4.2. *Suppose that Assumption 4.1 with $m \geq 1$ and Assumption 4.2 hold and suppose that for $i_1 \neq i_2 \in [d]$ the functional time series $\{X_{i_1,t,T}\}$ and $\{X_{i_2,t,T}\}$ are independent. Then, under the null hypothesis $H_0 : \mathcal{F}^{i_1}_{u,\omega} \equiv \mathcal{F}^{i_2}_{u,\omega}$, we have*

$$\sqrt{T}\hat{\mathcal{A}}_{i_1,i_2} \to \mathcal{N}\big(0, \sigma^2_{H_0}\big) \quad \textit{as } T \to \infty,$$

where the asymptotic variance is given by

$$\sigma^2_{H_0} = 4\pi \frac{\int_{-\pi}^{\pi}\int_0^1 \|\!|\mathcal{F}^{i_1}_{u,\omega}|\!\|_2^4 \, du \, d\omega}{(\int_{-\pi}^{\pi}\int_0^1 \|\!|\mathcal{F}^{i_1}_{u,\omega}|\!\|_2^2 \, du \, d\omega)^2} \text{‘}. \tag{4.6}$$

Let $I_p^{u_j,\omega_k} = (I_{i_1}^{u_j,\omega_k} + I_{i_2}^{u_j,\omega_k})/2$ be the pooled periodogram operator evaluated at u_j and ω_k. The asymptotic variance under the null can be estimated by

$$\hat{\sigma}^2_{H_0} = \frac{2}{3T}\sum_{j=1}^{M}\sum_{k=1}^{\lfloor N/2\rfloor}\big(\langle I_p^{u_j,\omega_k}, I_p^{u_j,\omega_{k-1}}\rangle_{\mathrm{HS}}\big)^2 \Big/ \left(\frac{2}{T}\sum_{j=1}^{M}\sum_{k=1}^{\lfloor N/2\rfloor}\langle I_p^{u_j,\omega_k}, I_p^{u_j,\omega_{k-1}}\rangle_{\mathrm{HS}}\right)^2. \tag{4.7}$$

Lemma 4.1. *Under the conditions of Theorem 4.2, the estimator defined in (4.7) satisfies*

$$\hat{\sigma}^2_{H_0} \xrightarrow{p} \sigma^2_{H_0} \quad (T \to \infty).$$

Consequently, a test for the hypothesis (4.1) can be based upon rejecting the null if

$$\hat{\mathcal{A}}_{i_1,i_2} > \frac{\hat{\sigma}^2_{H_0}}{\sqrt{T}} z_{1-\alpha}, \tag{4.8}$$

where $z_{1-\alpha}$ denotes the $(1-\alpha)$-quantile of the standard normal distribution. It follows from Theorem 4.2 that this test has asymptotic level α under the null hypothesis $\mathcal{A}_{i_1,i_2} = 0$. Moreover, under the alternative $\mathcal{A}_{i_1,i_2} > 0$, we obtain from Theorem 4.1 that the left hand side of (4.8) converges to the positive constant $\mathcal{A}_{i_1,i_2}$ in probability while the right hand side converges to 0. Therefore, the test is also consistent. The finite sample performance of this test is studied in a simulation study at the end of Section 5.

Remark 4.1. We emphasize that Theorem 4.2 still holds in case the series are dependent but the expression of the asymptotic variance is slightly more involved. It can however still be estimated similar in spirit to (4.7). See Appendix A for more details.

5. Simulation study

An application of the new clustering method to high-resolution meteorological data can be found in the Online Supplement. In this section, we study the performance in finite samples by means of a simulation study in a mixture of stationary and non-stationary models. We vary the number of clusters k and the number of observations per cluster n as well as the models we include. Furthermore, we investigate the algorithm both in the scenario of the number of true clusters being known and in the scenario that this number is unknown. In the latter case, we obtain an additional source of variability as the number of clusters needs to be estimated from the data. Finally, we consider the effect on the simulations of applying a scaling factor η to the similarity matrix. Before we discuss how to determine k, we start by introducing the simulation setting and data-generating processes.

5.1. Simulation setting

We simulate functional autoregressive and functional moving average via their basis representation as follows. A p-th order (time-varying) functional autoregressive process (tvFAR(p)), $\{X_t, t \in \mathbb{Z}\}$ can be defined as

$$X_t = \sum_{t'=1}^{p} A_{t,t'}(X_{t-t'}) + \epsilon_t, \tag{5.1}$$

where $A_{t,1}, \dots, A_{t,p}$ are time-varying autoregressive operators on $\mathcal{H}_{\mathbb{R}}$ and $\{\epsilon_t\}_{t\in\mathbb{Z}}$ is a sequence of mean zero innovations taking values in $\mathcal{H}_{\mathbb{R}}$. To generate such processes, let $\{\psi_l\}_{l\geq 1}$ be a Fourier basis of $\mathcal{H}_{\mathbb{R}}$. By means of a basis expansion, one can show that (see, e.g., Aue and van Delft [4]) the first $L_{\max}$ coefficients of (5.1) are generated using the p-th order vector autoregressive, VAR(p), process

$$\widetilde{X}_t = \sum_{t'=1}^{p} \widetilde{A}_{t,t'} \widetilde{X}_{t-t'} + \widetilde{\epsilon}_t,$$

where $\widetilde{X}_t := (\langle X_t, \psi_1\rangle, \dots, \langle X_t, \psi_{L_{\max}}\rangle)^\top$ is the vector of basis coefficients and the (l, l')-th entry of $\widetilde{A}_{t,j}$ is given by $\langle A_{t,j}(\psi_l), \psi_{l'}\rangle$ and $\widetilde{\epsilon}_t := (\langle \epsilon_t, \psi_1\rangle, \dots, \langle \epsilon_t, \psi_{L_{\max}}\rangle)^T$. The entries of the matrix $\widetilde{A}_{t,j}$ are generated as $N(0, \nu_{l,l'}^{(t,j)})$ with $\nu_{l,l'}^{(t,j)}$ specified below. To ensure stationarity or existence of a causal solution the norms $\kappa_{t,j}$ of $A_{t,j}$ are required to satisfy certain conditions (see Bosq [7] for stationary and van Delft and Eichler [61] for local stationary functional time series, respectively). We also consider the following time-varying functional moving-average process or order 1:

$$X_{t,T} = B_1(\epsilon_t) - \frac{1}{2}\left(1 + b\cos\left(2\pi\frac{t}{T}\right)\right)B_2(\epsilon_{t-1}), \tag{5.2}$$

where B_1 and B_2 are bounded linear operators on $L^2([0, 1])$ and $b \in \mathbb{R}$. Similarly as above, we use a basis expansion and generate data from the model

$$\widetilde{X}_{t,T} = \widetilde{B}_1\widetilde{\epsilon}_t - \frac{1}{2}\left(1 + b\cos\left(2\pi\frac{t}{T}\right)\right)\widetilde{B}_2\widetilde{\epsilon}_{t-1},$$

where $\widetilde{X}_{t,T} = (\langle X_{t,T}, \psi_1\rangle, \dots, \langle X_{t,T}, \psi_{L_{\max}}\rangle)^T$ is the vector of basis coefficients, the (l, l')-th entry of $\widetilde{B}_1$ and $\widetilde{B}_2$ are given by $\langle B_1(\psi_l), \psi_{l'}\rangle$ and $\langle B_2(\psi_l), \psi_{l'}\rangle$, respectively and $\widetilde{\epsilon}_t$ is as above.

We consider the following data generating processes:

(I) The functional i.i.d. process $\{\epsilon_t\}_{t=1}^T$ with coefficient variances $\mathbb{E}\langle \epsilon_t, \psi_l\rangle^2 = \exp(-(l-1)/10)$;

(II) The FAR(2) $\{X_t\}_{t=1}^T$ with operators specified by variances $\nu_{l,l'}^{(1)} = \exp(-l - l')$ and $\nu_{l,l'}^{(2)} = 1/(l + l'^{3/2})$ with norms $\kappa_1 = 0.75$ and $\kappa_2 = -0.4$ and innovations as in (I);

(III) The MA(1) with $b = 0$ and operators specified by variances $\nu_{l,l'}^{(1)} = \exp(-l - l')$;

(IV) The tvFAR(1) with operator specified by variances $\nu_{l,l'}^{(t,1)} = \nu_{l,l'}^{(1)} = \exp(-l - l')$ and norm $\kappa_1 = 0.8$, and innovations are as in (I) but with a multiplicative time-varying variance

$$\sigma^2(t) = \cos\left(\frac{1}{2} + \cos\left(\frac{2\pi t}{T}\right) + 0.3\sin\left(\frac{2\pi t}{T}\right)\right);$$

(V) The tvFAR(2) with operators as in (IV), but with time-varying norm $\kappa_{1,t} = 1.8\cos(1.5 - \cos(\frac{4\pi t}{T}))$ and constant norm $\kappa_2 = -0.81$ and innovations as in (I);

(VI) A FAR(2) with structural break;

- for $t \leq 3T/8$, the operators are as in (II) with norms $\kappa_1 = 0.7$ and $\kappa_2 = 0.2$, with innovations as in (I);
- for $t > 3T/8$, the operators are as in (II) with norms $\kappa_1 = 0$ and $\kappa_2 = -0.2$, with innovations as in (I) but with coefficient variances $\mathrm{Var}(\langle \epsilon_t, \psi_l \rangle) = 2\exp((l-1)/10)$.

Our simulation consists of the following balanced settings:

Setting 1: $k = 3$ with models I, II and III, where the replications per cluster are taken $n = 10$ and $n = 30$;
Setting 2: $k = 3$ with models IV, V and VI, where the replications per cluster are taken $n = 10$ and $n = 30$;
Setting 3: $k = 6$ with models I–VI, where the replications per cluster are taken $n = 10$ and $n = 30$ and $n = 50$.

In order to investigate the performance in case of imbalanced scenarios, we investigate two additional settings where the replications per cluster are specified via a permutation of the six models $\pi : (I, II, III, IV, V, VI) \to (I, II, III, IV, V, VI)$ and take the replications per cluster in each setting as:

Setting 4: The replications per cluster are taken as $n_\pi = (20, 20, 30, 30, 40, 40)$ with permutation $\pi_1 \cdots \pi_6 = 143625$;
Setting 5: The replications per cluster are taken as $n_{\pi_i} = \lfloor 20 \times 1.25^{(\pi_i - 1)} \rfloor$, $i = 1, \dots, 6$, with permutation $\pi_1 \cdots \pi_6 = 263415$.

Per set-up, we run 500 simulations for both $T = 256$ with $M = 8$ and $T = 512$ with $M = 16$.

For the choice of k we investigated a subset of well-known classical criteria that have been demonstrated to work well in aforementioned comparison studies on classical clustering: the Silhouette Index (Kaufman and Rousseeuw [41]), the CH Index (Caliński and Harabasz [8]), the Hartigan Index (Hartigan [28]) and the KL Index (Krzanowski and Lai [42]). We respectively refer to these in the tables as ‘sil’,‘ch’, ‘hartigan’ and ‘kl’. Because these cannot be applied to the spectral embedding directly, the respective criteria were constructed using the similarity graph $\hat{W}$. Additionally, we considered two variations of the eigengap heuristic each with a different interpretation of ‘relatively large’ eigengap. Let $0 = \hat{\lambda}_1 \leq \cdots \leq \hat{\lambda}_d$ be the estimated eigenvalues of $\hat{L}$ in ascending order

1. ‘Relgap’: define the relative contribution of the k-th eigenvalue as

$$\rho_k = \frac{(\hat{\lambda}_k - \hat{\lambda}_{k-1})}{\hat{\lambda}_k}.$$

   Then the rule is to pick $k^\star$ as the largest k for which the k-th contribution is still larger than a threshold value that is allowed to depend on the scaling parameter η of the graph (see (5.3) below)

$$k^\star = \max_k \big\{ k \in \{1, \dots, k_{\max}\} : \rho_k \leq 0.01\eta \big\}$$

2. ‘sd1gap’: let

$$\sigma(\hat{\lambda}_{[-1:k]}) = \frac{1}{d-k} \sum_{j=k+1}^{d} \left(\hat{\lambda}_j - \frac{1}{d-k} \sum_{j=k+1}^{d} \hat{\lambda}_j \right)^2$$

be the squared deviation from the mean excluding the first k eigenvalues. Then the rule is to pick $k^\star$ as the largest k for which the k-th gap is still larger than $\sigma(\hat{\lambda}_{[-1:k]})$, that is,

$$k^\star = \max_k \big\{ k \in \{1, \dots, k_{\max}\} : \hat{\lambda}_{k+1} - \hat{\lambda}_k \geq \sigma(\hat{\lambda}_{[-1:k]}) \big\}$$

For all criteria, the maximum numbers of clusters to consider was set to $k_{\max} = 15$.

5.2. Simulation results

Table 1 provides the average k chosen according to the different criteria in each setting, while the corresponding percentage of misclustered points averaged over simulations are given in Table 2. From the first row for $T = 256$ and $T = 512$ of Table 2, we find that our algorithm does very well if the true k is known; it has a very low percentage of misclustered points across the different settings. Based upon the percentage of misclustered points, the most difficult model is clearly setting 3 with $T = 256$ and $n = 10$. This finding is in accordance with the upper bound in Theorem 3.1, which is less tight for larger k and for lower estimation precision of $\hat{L}$.

If the true k is not known, we obtain higher percentages of misclustered points where the percentages appear to be caused by the selection method for k. As can be seen from Table 1, the CH Index does best in determining the true number of clusters, while the KL index does worst when the number of true clusters increases. Both the Silhouette Index and the Hartigan Index tend to pick k more conservatively. The two rules based upon the estimated eigenvalues of the graph Laplacian – 'Relgap' and 'sd1gap' – appear competitive with the CH index, except in setting 2 for $n = 10$. We observe a clear overall improvement as both n and T increase. It appears in particular that the eigenvalue-based methods suffer

Table 1. Chosen k per method averaged over simulations (standard deviation in brackets)

	Setting 1		Setting 2		Setting 3		
method	$n = 10$	$n = 30$	$n = 10$	$n = 30$	$n = 10$	$n = 30$	$n = 50$
				$T = 256$			
true k	3	3	3	3	6	6	6
sil	2 (0)	2 (0)	3 (0.1)	3 (0)	5.3 (0.8)	5.2 (0.5)	5.2 (0.4)
ch	3 (0.5)	3.0 (0.2)	3.1 (0.2)	3 (0)	6.4 (0.6)	6.1 (0.3)	6.0 (0.2)
kl	2.2 (1.4)	2.0 (0.7)	3.1 (1)	3.2 (1.4)	10.4 (3.1)	11.1 (3.5)	10.8 (3.6)
hart	3.0 (0.2)	3 (0)	3 (0)	3 (0)	4.8 (0.9)	4.9 (0.7)	4.9 (0.6)
Relgap	3.0 (0.2)	3 (0)	3.9 (1.1)	3.1 (0.3)	5.1 (0.4)	5.2 (0.4)	5.2 (0.4)
sd1gap	3.1 (0.3)	3.0 (0.2)	3.8 (1.2)	3.6 (1)	5.7 (0.9)	6.2 (0.7)	6.4 (0.9)
				$T = 512$			
true k	3	3	3	3	6	6	6
sil	2 (0)	2 (0)	3 (0)	3 (0)	5.9 (0.4)	6.0 (0.2)	6.0 (0.1)
ch	3.1 (0.4)	3 (0)	3 (0)	3 (0)	6.5 (0.7)	6.2 (0.4)	6.1 (0.3)
kl	2.4 (1.7)	2.2 (1)	3.1 (0.7)	3.4 (2.1)	9.9 (3.3)	10.5 (3.5)	9.5 (3.5)
hart	3 (0)	3 (0)	3 (0)	3 (0)	5.3 (0.7)	5.4 (0.6)	5.4 (0.5)
Relgap	3 (0)	3 (0)	3.6 (0.8)	3.0 (0.2)	5.5 (0.6)	5.9 (0.3)	6.0 (0.1)
sd1gap	3.1 (0.4)	3.0 (0.1)	3.9 (1.2)	3.4 (0.8)	6.4 (0.9)	6.3 (0.6)	6.2 (0.5)

Table 2. Percentage of misclustered points of the spectral clustering algorithm averaged over simulations (standard deviation in brackets)

	Setting 1		Setting 2		Setting 3		
method	$n = 10$	$n = 30$	$n = 10$	$n = 30$	$n = 10$	$n = 30$	$n = 50$
				$T = 256$			
true	0.1 (0.6)	0.1 (0.2)	0.0 (0.1)	0 (0)	3.2 (6.6)	0.3 (1.1)	0.1 (0.2)
sil	33.3 (0)	33.3 (0)	0.03 (0.4)	0 (0)	13.2 (10.2)	13.0 (8.2)	13.9 (6.9)
ch	4.7 (10.4)	1.4 (6.6)	0.5 (2.6)	0 (0)	5.1 (6.8)	0.8 (2.2)	0.3 (1.1)
kl	33.3 (6.4)	33.5 (2.4)	0.6 (6)	1.45 (9.4)	25.6 (14.7)	25.9 (15.6)	22.6 (14.4)
hart	0.6 (2.6)	0.1 (0.2)	0.0 (0.1)	0 (0)	19.8 (14.6)	18.9 (12.5)	19.0 (10.7)
Relgap	0.6 (2.4)	0.1 (0.2)	10.8 (11.9)	1.3 (4.2)	15.75 (5.6)	14.5 (6.3)	13.7 (6.7)
sd1gap	1.3 (4)	0.5 (2.3)	9.1 (12.5)	6.6 (10.9)	12.33 (7.5)	3.0 (5.5)	2.8 (5.2)
				$T = 512$			
true	0 (0)	0 (0)	0 (0)	0 (0)	0.23 (1.7)	0.01 (0.1)	0 (0)
sil	33.3 (0)	33.3 (0)	0.03 (0.4)	0 (0)	2.7 (5.8)	0.8 (3.5)	0.4 (2.4)
ch	1.7 (4.6)	0 (0)	0.4 (2.1)	0 (0)	3.4 (4.4)	1.1 (2.6)	0.6 (2)
kl	28.0 (14.5)	30.6 (10.2)	1.7 (10.3)	2.6 (12.5)	20.9 (16.8)	21.4 (16.3)	15.6 (14.4)
hart	0 (0)	0 (0)	0 (0)	0 (0)	12.3 (12.1)	10.0 (9.4)	9.6 (8.7)
Relgap	0.02 (0.4)	0 (0)	6.9 (9.7)	0.4 (2.2)	8.7 (8.4)	2 (5.4)	0.2 (1.7)
sd1gap	1.1 (4.4)	0.1 (1.4)	10.2 (12.6)	5.0 (9.7)	4.2 (6)	1.9 (3.9)	1.4 (3.4)

from more variation when n and T are small. This is intuitive, since the choice of k directly depends upon the estimation precision of the spectral properties, which can be expected to be more sensitive to estimation error for small T and n (see also Theorem 3.1). From the results in Table 2, we find the CH index to perform best in combination with our algorithm. It is most stable across the different settings and has the lowest percentage of misclustered points, which appears to be a direct consequence of the fact that this index estimated the true number of clusters best and that our algorithm exhibits lower error conditional upon knowing the correct number of clusters.

The results for the two imbalanced settings, setting 4 and 5, are given in Table 3. Due to space constraints, we only report the results for the true number of clusters and for the CH Index. As expected from the discussion below Theorem 3.1, we observe that a larger imbalance in replications per cluster leads in principal to a higher percentage of misclustered points. In the very imbalanced scenario of setting 5, where the replications per cluster vary according to an exponential growth rate, the percentage is considerably higher for $T = 256$ compared to setting 3 and 4 but the error clearly drops fast as the sample size increases; the percentages do not appear much worse than in the balanced scenario with $n = 10$ for $T = 512$.

Finally, we investigate the robustness of our method with respect to scaling. As explained in Section 2, we apply a local scaling to our measure in order to avoid the well-known sensitivity problem to the specification of a global scaling parameter. Indeed, methods of which the similarity graph relies on such a global parameter can lead the spectral clustering to fail; it is usually not clear how to specify this parameter and differences in scales might be amplified. Scaling pairwise ensures the similarity graph does not have the latter issue. To verify the robustness, we consider our clustering algorithm as explained in Section 3 but with an additional scaling parameter $\eta > 0$ in the construction of the adjacency

Table 3. Percentage of misclustered points and specified k for various values of η averaged over simulations (standard deviation in brackets) for imbalanced scenarios

	% of misclustered points		average chosen k	
method	Setting 4	Setting 5	Setting 4	Setting 5
		$T = 256$		
true	0.8 (3)	16.6 (7.1)	6	6
ch	1.0 (3.3)	16.0 (7.2)	6.0 (0.1)	6.0 (0.3)
		$T = 512$		
true	0.1 (0.8)	3.2 (6.7)	6	6
ch	0.2 (1.4)	3.3 (6.7)	6.0 (0.2)	6.0(0.2)

Table 4. Percentage of misclustered points and specified k for various values of η averaged over simulations (standard deviation in brackets) for $T = 512$, $n = 30$

	% of misclustered points				average chosen k			
meth.	$\eta = 0.5$	$\eta = 2.5$	$\eta = 5$	$\eta = 10$	$\eta = 0.5$	$\eta = 2.5$	$\eta = 5$	$\eta = 10$
				Setting 1: $T = 512$, $n = 30$				
true	0 (0)	0 (0)	0 (0)	0 (0)	3 (0)	3 (0)	3 (0)	3 (0)
sil	33.3 (0)	33.3 (0)	2.1 (8.2)	0 (0)	2 (0)	2 (0)	2 (0)	2.9 (0.2)
ch	0.0 (0.6)	0 (0)	0 (0)	0 (0)	3 (0)	3 (0)	3 (0)	3 (0)
kl	33.0 (5.1)	2.8 (9.3)	0 (0)	0.3 (4.3)	2.1 (0.9)	2.2 (1)	2.9 (0.3)	3 (0)
hart	0 (0)	0 (0)	0 (0)	0 (0)	3 (0)	3 (0)	3 (0)	3 (0)
Relgap	0 (0)	0.0 (0.5)	0.0 (0.5)	0.0 (0.5)	3 (0)	3 (0)	3 (0)	3 (0)
sd1gap	0.1 (0.8)	0.9 (3.5)	1.1 (3.9)	0.6 (2.7)	3 (0.1)	3.0 (0.1)	3.1 (0.3)	3.1 (0.3)
				Setting 2: $T = 512$, $n = 30$				
true	0 (0)	0 (0)	0 (0)	0 (0)	3 (0)	3 (0)	3 (0)	3 (0)
sil	0 (0)	0 (0)	0 (0)	0 (0)	3 (0)	3 (0)	3 (0)	3 (0)
ch	0 (0)	0 (0)	0 (0)	0 (0)	3 (0)	3 (0)	3 (0)	3 (0)
kl	0.4 (4.6)	2.4 (11.2)	1.6 (9.4)	0.4 (4.6)	3.1 (0.7)	3.4 (2.1)	3.4 (2)	3.3 (1.7)
hart	0 (0)	0 (0)	0 (0)	0 (0)	3 (0)	3 (0)	3 (0)	3 (0)
Relgap	0.1 (1.2)	2.3 (5.8)	6.7 (9.3)	7.1 (8.5)	3.0 (0.1)	3.0 (0.2)	3.2 (0.5)	3.6 (0.9)
sd1gap	4.8 (9.5)	5.5 (9.7)	4.6 (8.7)	2.9 (5.9)	3.4 (0.8)	3.4 (0.8)	3.5 (0.9)	3.4 (0.8)
				Setting 3: $T = 512$, $n = 30$				
true	0.0 (0.1)	0 (0)	0.0 (0.9)	0.1 (0.9)	6 (0)	6 (0)	6 (0)	6 (0)
sil	1.0 (3.9)	0.0 (0.7)	0.0 (0.9)	0.1 (0.9)	6.0 (0.2)	6.0 (0.2)	6 (0)	6 (0)
ch	1.7 (3.2)	0.3 (1.4)	0.0 (0.9)	0.1 (0.9)	6.2 (0.5)	6.2 (0.4)	6.0 (0.2)	6 (0)
kl	22.7 (16.4)	13.9 (15.5)	7.4 (12.2)	4.6 (10.5)	10.5 (3.4)	10.5 (3.5)	9.2 (3.6)	7.8 (3.1)
hart	11.5 (8)	7.2 (13.5)	0.0 (0.9)	0.1 (0.9)	5.3 (0.5)	5.4 (0.6)	5.6 (0.8)	6 (0)
Relgap	7.2 (8.3)	0.6 (2.2)	3.2 (3.9)	3.9 (4.4)	5.6 (0.5)	5.9 (0.3)	6.1 (0.3)	6.6 (0.8)
sd1gap	1.4 (3.4)	4.5 (5.1)	6.0 (5.6)	4.2 (5.1)	6.2 (0.5)	6.3 (0.6)	6.8 (1)	7.2 (1.3)

matrix. That is, we consider the simulations but with

$$\hat{W} = e^{-\eta \hat{A}} \in \mathbb{R}^{d \times d}, \tag{5.3}$$

where $\eta = \{0.5, 2.5, 5, 10\}$. We remark that for any $\eta > 0$ the theory derived in Section 3 remains true (see also Appendix B). The results in Table 2 correspond to the choice $\eta = 1$ and in Table 4 we present the four alternative choices. Because of space constraints, we only report these for the specification $T = 512$ and $n = 30$. The results for the other cases show a very similar picture and are available upon request. It can be observed from the first row for each of the settings that the outcomes are fairly similar when k is known. Variation again therefore seems mostly caused by the way k is chosen, where we find the Silhouette index is sensitive as well as the KL index, but also the eigengap heuristics show some sensitivity for $\eta = 10$. Overall, we may conclude that our method seems capable to detect the correct number of clusters, despite the highly complicated nature of the data. The numerical study moreover suggests that the CH index should be used to find the numbers of clusters if these are unknown (see also our data application in the Online Supplement).

5.3. Testing for equality

We conclude this section with a small investigation of the proposed asymptotic α-level test in (4.8) for the hypothesis of equality of (possibly) time-varying spectral density operators. To investigate the finite samples properties of the test, we performed a simulation study which includes the previously defined stationary and non-stationary models with parameter specifications for $T = 256$ and $T = 512$ with blocks fixed to $M = 16$. The pairwise rejection percentages at the 5% and 10% over 1000 replications are provided in Table 5 and Table 6, where the diagonal elements correspond to the null hypothesis. We observe a good approximation of the nominal level, albeit with model *II–IV* a bit undersized. Given the relatively small value of N, it is reasonable to believe that this is caused by finite sample bias present in the pooled variance estimator and that, for these models, more data is required in order to reflect the asymptotic independence of the local fDFT at lagged frequencies. The off-diagonal shows good power overall, with both model I and IV appearing more difficult to distinguish from model II. Power of the test clearly improves with increasing sample size. Regardless of the second order properties being time-varying or not, it appears therefore that the quantiles of the normal distribution are well-captured for the various models if H_0 is true, while good power is observed under H_A.

Appendix A: Distributional properties of the similarity measure

Due to space constraints, we provide here only the main steps of the proof. Auxiliary statements and necessary background are relegated to the Online Supplement. In the following, for random variables Y_0, Y_1, Y_2, Y_3 let $\mathrm{cum}_{m_0,m_1,m_2,m_3}(Y_0, Y_1, Y_2, Y_3)$ denote the joint cumulant

$$\mathrm{cum}(\underbrace{Y_0, \dots, Y_0}_{m_0 \text{ times}}, \underbrace{Y_1, \dots, Y_1}_{m_1 \text{ times}}, \underbrace{Y_2, \dots, Y_2}_{m_2 \text{ times}}, \underbrace{Y_3, \dots, Y_3}_{m_3 \text{ times}}),$$

where $0 \le m_i \le m$, $i = 0, 1, 2, 3$ s.t. $\sum_{i=0}^{3} m_i = m$. Using the results in Section S2, we can derive the order of the higher order joint cumulants of elements defined in (2.6).

Table 5. Rejection percentages of the pairwise equality test (4.8) at the 10% (top); and 5% level for $T = 256$

	I	II	III	IV	V	VI
I	11.4	74.2	100	97.9	99.9	100
II	74.2	8.3	100	75.4	95.9	100
III	100	100	5.7	100	100	100
IV	97.9	75.4	100	10.3	97.7	100
V	99.9	95.9	100	97.7	10.5	99.8
VI	100	100	100	100	99.8	11.1

	I	II	III	IV	V	VI
I	5.6	61.3	100	95.0	99.6	100
II	61.3	3.3	100	61.7	91.3	99.9
III	100	100	1.9	99.8	100	100
IV	95.0	61.7	99.8	3.9	95.4	100
V	99.6	91.3	100	95.4	3.9	99.4
VI	100	99.9	100	100	99.4	5.2

Table 6. Rejection percentages of the pairwise equality test (4.8) at the 10% (top); and 5% level for $T = 512$

	I	II	III	IV	V	VI
I	10.8	96.6	100	100	100	100
II	96.6	8.0	100	96.1	99.9	100
III	100	100	6.0	100	100	100
IV	100	96.1	100	8.5	100	100
V	100	99.9	100	100	9.8	100
VI	100	100	100	100	100	10.0

	I	II	III	IV	V	VI
I	5.4	93.1	100	99.9	100	100
II	93.1	3.2	100	92.0	99.7	100
III	100	100	2.1	100	100	100
IV	99.9	92.0	100	3.2	99.9	100
V	100	99.7	100	99.9	3.6	100
VI	100	100	100	100	100	4.7

Theorem A.1. *If Assumption 4.1 is satisfied then for finite m*

$$T^{m/2}\,\mathrm{cum}_{m_0,m_1,m_2,m_3}(\hat{F}_{i_1,i_2},\hat{F}_{i_3,i_4},\hat{F}_{i_5,i_6},\hat{F}_{i_7,i_8})$$
$$= \frac{1}{T^{m/2}}\sum_{k_1,\dots,k_m=1}^{\lfloor N/2\rfloor}\sum_{j_1,\dots,j_m=1}^{M}\mathrm{Tr}\Bigg(\sum_{\boldsymbol{P}=P_1\cup\cdots\cup P_G} S_{\boldsymbol{P}}\Big(\bigotimes_{g=1}^{G}\mathrm{cum}\big(D_{i_p}^{u_{j_p},\omega_{k_p}}\,|\,p\in P_g\big)\Big)\Bigg) = O\big(T^{1-m/2}\big),$$

uniformly in $0 \le m_i \le m$ s.t. $\sum_{i=0}^{3} m_i = m$.

Proof of Theorem A.1. For a fixed partition $P = \{P_1, \dots, P_G\}$, let the cardinality of set P_g be denoted by $|P_g| = \mathscr{C}_g$. By (S2.3) of Corollary S2.1 and Lemma S2.3 an upperbound of (S2.2) is given by

$$O\left(T^{-m/2} \sum_{k_1,\dots,k_m=1}^{\lfloor N/2\rfloor} \sum_{j_1,\dots,j_m=1}^{M} \prod_{g=1}^{G} \frac{1}{N^{\mathscr{C}_g/2-1}} M^{-\delta_{\{\exists p_1,p_2\in P_g:|j_{p_1}-j_{p_2}|>1\}}}\right). \tag{A.1}$$

Similar to Lemma 4.3 of van Delft, Characiejus and Dette [59], we can show inductively that the indecomposability of the array (S2.2) and the behavior of the joint cumulants of the local fDFT's at different midpoints imply this is at most of order

$$O\big(N^{m/2} M^{-m/2} E^m M N^{-2m+G}\big) = O\big(T^{1-m/2} N^{G-m-1}\big).$$

Thus, partitions of size $G \le m+1$ will vanish as $T \to \infty$. For $G \ge m+2$, indecomposability of the array requires to stay on the frequency manifold (see equation (S2.4) of Corollary S2.1) and therefore imposes additional restrictions in frequency direction. It can be shown that for a partition of size $G = m + r_1 + 1$ with $r_1 \ge 1$ of the array (S2.2) only partitions with at least r_1 restrictions in frequency direction are indecomposable if $m > 2$, while if $m = 2$ there must be at least 1 restriction in frequency direction. Consequently, the joint cumulant is at most of order $O(T^{1-n/2} N^{m+r_1+1-m-1} N^{-r_1}) = O(T^{1-n/2})$. □

Proof of Theorem 2.1. Using Theorem S2.1 with $n = 1$ implies

$$\mathbb{E}F_{i_1,i_2} = \frac{1}{T} \sum_{k=1}^{\lfloor N/2\rfloor} \sum_{j=1}^{M} \operatorname{Tr}\big(\mathbb{E}\big[D_{i_1}^{u_{j_1},\omega_{k_1}} \otimes D_{i_1}^{u_{j_1},-\omega_{k_1}} \otimes D_{i_2}^{u_{j_1},-\omega_{k_1-1}} \otimes D_{i_2}^{u_{j_1},\omega_{k_1-1}}\big]\big).$$

Rewriting this expectation in cumulant tensors, we get

$$\begin{aligned}
\mathbb{E}F_{i_1,i_2} = {}& \frac{1}{T} \sum_{k=1}^{\lfloor N/2\rfloor} \sum_{j=1}^{M} \operatorname{Tr}(S_{1234} \operatorname{Cum}\big(\big(D_{i_1}^{u_{j_1},\omega_{k_1}}, D_{i_1}^{u_{j_1},-\omega_{k_1}}, D_{i_2}^{u_{j_1},-\omega_{k_1-1}}, D_{i_2}^{u_{j_1},\omega_{k_1-1}}\big)\big) \\
&+ \frac{1}{T} \sum_{k=1}^{\lfloor N/2\rfloor} \sum_{j=1}^{M} \operatorname{Tr}\big(S_{1234}\big(\operatorname{Cum}\big(D_{i_1}^{u_{j_1},\omega_{k_1}}, D_{i_1}^{u_{j_1},-\omega_{k_1}}\big) \otimes \operatorname{Cum}\big(D_{i_2}^{u_{j_1},-\omega_{k_1-1}}, D_{i_2}^{u_{j_1},\omega_{k_1-1}}\big)\big)\big) \\
&+ \frac{1}{T} \sum_{k=1}^{\lfloor N/2\rfloor} \sum_{j=1}^{M} \operatorname{Tr}\big(S_{1324}\big(\operatorname{Cum}\big(D_{i_1}^{u_{j_1},\omega_{k_1}}, D_{i_2}^{u_{j_1},-\omega_{k_1-1}}\big) \otimes \operatorname{Cum}\big(D_{i_1}^{u_{j_1},-\omega_{k_1}}, D_{i_2}^{u_{j_1},\omega_{k_1-1}}\big)\big)\big) \\
&+ \frac{1}{T} \sum_{k=1}^{\lfloor N/2\rfloor} \sum_{j=1}^{M} \operatorname{Tr}\big(S_{1423}\big(\operatorname{Cum}\big(D_{i_1}^{u_{j_1},\omega_{k_1}}, D_{i_2}^{u_{j_1},\omega_{k_1-1}}\big) \otimes \operatorname{Cum}\big(D_{i_1}^{u_{j_1},-\omega_{k_1}}, D_{i_2}^{u_{j_1},-\omega_{k_1-1}}\big)\big)\big).
\end{aligned}$$

By Corollary S2.1 and Lemma S2.2, we thus find

$$\mathbb{E}F_{i_1,i_2} = \frac{1}{T} \sum_{k=1}^{\lfloor N/2\rfloor} \sum_{j=1}^{M} \big\langle \mathcal{F}_{u_j,\omega_k}^{i_1}, \mathcal{F}_{u_j,\omega_{k-1}}^{i_2} \big\rangle_{\mathrm{HS}} + O\left(\frac{1}{M^2}\right) + O\left(\frac{1}{N}\right).$$

Hence,

$$\lim_{N,M\to\infty} \mathbb{E}\, F_{i_1,i_2} = \frac{1}{2\pi}\int_0^{\pi}\int_0^1 \langle \mathcal{F}^{i_1}_{u,\omega}, \mathcal{F}^{i_2}_{u,\omega}\rangle_{\mathrm{HS}}\, \mathrm{d}u\, \mathrm{d}\omega = \frac{1}{4\pi}\int_{-\pi}^{\pi}\int_0^1 \langle \mathcal{F}^{i_1}_{u,\omega}, \mathcal{F}^{i_2}_{u,\omega}\rangle_{\mathrm{HS}}\, \mathrm{d}u\, \mathrm{d}\omega.$$

Secondly, we have for any $i_1, i_2, i_3, i_4 \in \{1, \dots, d\}$

$$T\,\mathrm{Cov}(F_{i_1,i_2}, F_{i_3,i_4}) = T\,\mathrm{Cum}(F_{i_1,i_2}, \overline{F_{i_3,i_4}})$$

Hence, if Assumption 4.1 is satisfied with $m = 8$, Theorem A.1 implies this term is of order $O(1)$. summarizing, we find F_{i_1,i_1}, F_{i_2,i_2}, F_{i_1,i_2} and F_{i_2,i_1} are asymptotically unbiased and jointly convergence in probability. The continuous mapping theorem establishes then that $\hat{\mathcal{A}}_{i_1,i_2}$ is a $\sqrt{T}$-consistent estimator of $\mathcal{A}_{i_1,i_2}$ for any $i_1, i_2, \in \{1 \dots, d\}$. □

Proof of Theorem 4.1. If Assumption 4.1 holds for all moments, then Theorem A.1, yields that for $m > 2$

$$T^{m/2}\,\mathrm{cum}_{m_0,m_1,m_2,m_3}(F_{i_1,i_2}, F_{i_3,i_4}, F_{i_5,i_6}, F_{i_7,i_8}) \to 0 \quad \text{as } T \to \infty,$$

from which asymptotic joint normality of F_{i_1,i_2}, F_{i_3,i_4}, F_{i_5,i_6}, F_{i_7,i_8} follows, i.e., we have

$$\sqrt{T}\begin{pmatrix} 4\pi F_{i_1,i_1} - \mathbb{E}(F_{i_1,i_1}) \\ 4\pi F_{i_2,i_2} - \mathbb{E}(F_{i_2,i_2}) \\ 4\pi F_{i_1,i_2} - \mathbb{E}(F_{i_1,i_2}) \\ 4\pi F_{i_2,i_1} - \mathbb{E}(F_{i_1,i_1}) \end{pmatrix} \longrightarrow \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}),$$

where

$$\mathbf{\Sigma} = \begin{pmatrix} \mathrm{Var}(F_{i_1,i_1}) & \mathrm{Cov}(F_{i_1,i_1}, F_{i_2,i_2}) & \mathrm{Cov}(F_{i_1,i_1}, F_{i_1,i_2}) & \mathrm{Cov}(F_{i_1,i_1}, F_{i_2,i_1}) \\ \mathrm{Cov}(F_{i_1,i_1}, F_{i_2,i_2}) & \mathrm{Var}(F_{i_2,i_2}) & \mathrm{Cov}(F_{i_2,i_2}, F_{i_1,i_2}) & \mathrm{Cov}(F_{i_2,i_2}, F_{i_2,i_1}) \\ \mathrm{Cov}(F_{i_1,i_1}, F_{i_1,i_2}) & \mathrm{Cov}(F_{i_2,i_2}, F_{i_1,i_2}) & \mathrm{Var}(F_{i_1,i_2}) & \mathrm{Cov}(F_{i_1,i_2}, F_{i_2,i_1}) \\ \mathrm{Cov}(F_{i_1,i_1}, F_{i_2,i_1}) & \mathrm{Cov}(F_{i_2,i_2}, F_{i_2,i_1}) & \mathrm{Cov}(F_{i_1,i_2}, F_{i_2,i_1}) & \mathrm{Var}(F_{i_2,i_1}) \end{pmatrix}. \tag{A.2}$$

To derive from this the distribution of $\hat{\mathcal{A}}_{i_1,i_2}$, consider the function $g : \mathbb{R}^4 \to \mathbb{R}$

$$g(x_1, x_2, x_3, x_4) = 1 - \frac{x_3}{(x_1 + x_2)} - \frac{x_4}{(x_1 + x_2)}$$

of which the gradient is given by

$$\nabla g^{\top}(\boldsymbol{x}) = \begin{pmatrix} x_3(x_1+x_2)^{-2} + x_4(x_1+x_2)^{-2} \\ x_3(x_1+x_2)^{-2} + x_4(x_1+x_2)^{-2} \\ -(x_1+x_2)^{-1} \\ -(x_1+x_2)^{-1} \end{pmatrix} = \frac{1}{(x_1+x_2)}\begin{pmatrix} \dfrac{x_3+x_4}{(x_1+x_2)} \\ \dfrac{x_3+x_4}{(x_1+x_2)} \\ -1 \\ -1 \end{pmatrix}.$$

Then since we can write

$$\hat{\mathcal{A}}_{i_1,i_2} = g(F_{i_1,i_1}, F_{i_2,i_2}, F_{i_1,i_2}, F_{i_2,i_1}) = 1 - \frac{F_{i_1,i_2}}{(F_{i_1,i_1} + F_{i_2,i_2})} - \frac{F_{i_2,i_1}}{(F_{i_1,i_1} + F_{i_2,i_2})},$$

the Delta method implies, that as $T \to \infty$,

$$\big\{\sqrt{T}(\hat{\mathcal{A}}_{i_1,i_2} - \mathcal{A}_{i_1,i_2})\big\}_{\{i_1,i_2 \in [d]\}} \to \mathcal{N}\big(\mathbf{0}, \nabla g^\top(\boldsymbol{x})\boldsymbol{\Sigma}\nabla g(\boldsymbol{x})\big), \tag{A.3}$$

where for fixed $i_1, i_2 \in \{1, \dots, d\}$,

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} \frac{1}{4\pi}\int_{-\pi}^{\pi}\int_0^1 \langle \mathcal{F}^{i_1}_{u,\omega}, \mathcal{F}^{i_1}_{u,\omega}\rangle_{\text{HS}}\,\mathrm{d}u\,\mathrm{d}\omega \\ \frac{1}{4\pi}\int_{-\pi}^{\pi}\int_0^1 \langle \mathcal{F}^{i_2}_{u,\omega}, \mathcal{F}^{i_2}_{u,\omega}\rangle_{\text{HS}}\,\mathrm{d}u\,\mathrm{d}\omega \\ \frac{1}{4\pi}\int_{-\pi}^{\pi}\int_0^1 \langle \mathcal{F}^{i_1}_{u,\omega}, \mathcal{F}^{i_2}_{u,\omega}\rangle_{\text{HS}}\,\mathrm{d}u\,\mathrm{d}\omega \\ \frac{1}{4\pi}\int_{-\pi}^{\pi}\int_0^1 \langle \mathcal{F}^{i_2}_{u,\omega}, \mathcal{F}^{i_1}_{u,\omega}\rangle_{\text{HS}}\,\mathrm{d}u\,\mathrm{d}\omega \end{pmatrix}, \tag{A.4}$$

and $\boldsymbol{\Sigma}$ is defined in (A.2). Derivation of the covariance structure is tedious and relegated to the Online Supplement. □

Appendix B: Analysis of the spectral clustering algorithm

B.1. Consistency of $\hat{L}$ for L

Proof of Lemma 3.1. From Theorem 2.1, we have that $\hat{\mathcal{A}} \in \mathbb{R}^{d\times d}$ is a $\sqrt{T}$-consistent estimator of the distance measure $\mathcal{A}$. The continuous mapping theorem therefore implies that $\hat{W}$ is consistent, that is, A simple calculation shows that, as $T \to \infty$,

$$\mathbb{P}\big(\|\hat{W} - W\|_\infty \geq \varepsilon\big) \leq \mathbb{P}\Big(d \max_{i,j} |\hat{W}_{i,j} - W_{i,j}| \geq \varepsilon\Big) \to 0. \tag{B.1}$$

Similarly,

$$\begin{aligned} \mathbb{P}\Big(\max_i |D_i - \hat{D}_i| \geq \varepsilon\Big) &= \mathbb{P}\Big(\max_i \Big|\sum_j \hat{W}_{i,j} - \sum_j \hat{W}_{i,j}\Big| \geq \varepsilon\Big) \\ &\leq \mathbb{P}\Big(d \max_{i,j} |\hat{W}_{i,j} - W_{i,j}| \geq \varepsilon\Big) \to 0. \end{aligned} \tag{B.2}$$

We use the decomposition

$$\begin{aligned} \hat{L} - L &= \hat{D}^{-1/2}\hat{W}\hat{D}^{-1/2} - D^{-1/2}\hat{W}D^{-1/2} + D^{-1/2}\hat{W}D^{-1/2} - D^{-1/2}WD^{-1/2} \\ &= \big(\hat{D}^{-1/2} - D^{-1/2}\big)\hat{W}\hat{D}^{-1/2} + D^{-1/2}\hat{W}\big(\hat{D}^{-1/2} - D^{-1/2}\big) + D^{-1/2}(\hat{W} - W)D^{-1/2} \\ &= \big(I - D^{-1/2}\hat{D}^{1/2}\big)\hat{D}^{-1/2}\hat{W}\hat{D}^{-1/2} + \big(D^{-1/2}\hat{D}^{1/2}\big)\hat{D}^{-1/2}\hat{W}\hat{D}^{-1/2}\big(I - \hat{D}^{1/2}D^{-1/2}\big) \\ &\quad + D^{-1/2}(\hat{W} - W)D^{-1/2} \end{aligned}$$

and bound these terms separately. Note that as D and $\hat{D}$ are degree matrices, they are diagonal with nonnegative entries. We therefore have

$$|||I - D^{-1/2}\hat{D}^{1/2}|||_\infty = \max_i \left|1 - \sqrt{\frac{\hat{D}_i}{D_i}}\right| \leq \max_i \left|1 - \frac{\hat{D}_i}{D_i}\right| \leq \max_i \frac{|D_i - \hat{D}_i|}{\min_i D_i}.$$

The triangle inequality gives

$$|||D^{-1/2}\hat{D}^{1/2}|||_\infty = |||I - \big(I - D^{-1/2}\hat{D}^{1/2}\big)|||_\infty \leq 1 + \max_i \frac{|D_i - \hat{D}_i|}{\min_i D_i}.$$

Additionally, since $\hat{D}_i = \sum_j \hat{W}_{i,j}$ it follows that $|||\hat{D}^{-1/2}\hat{W}\hat{D}^{-1/2}|||_\infty = 1$. Furthermore, $|||D^{-1/2}(\hat{W} - W)D^{-1/2}|||_\infty \leq \frac{1}{\min_i D_i}|||\hat{W} - W|||_\infty$. Therefore,

$$|||\hat{L} - L|||_\infty \leq \frac{\max_i |D_i - \hat{D}_i|}{\min_i D_i}\left(2 + \frac{\max_i |D_i - \hat{D}_i|}{\min_i D_i}\right) + \frac{1}{\min_i D_i}|||\hat{W} - W|||_\infty.$$

Consequently, (B.1) and (B.2) imply

$$\forall \varepsilon > 0, \quad \lim_{T\to\infty} \mathbb{P}\big(|||\hat{L} - L|||_\infty > \varepsilon\big) = 0. \qquad \square$$

B.2. Concentration of $\hat{\mathcal{U}}$

We shall use Lemma 3.1 to analyze the concentration of $\hat{\mathcal{U}}$. We first need the following auxiliary lemma.

Lemma B.1. *Let $S \subset \mathbb{R}$ an interval. Let $A, H \in \mathbb{R}^{d\times d}$ be two symmetric matrices and let $\hat{A} = A + H$ denote a perturbed version of A. Denote $\hat{Q}$ and Q be orthornormal matrices of dimension $\mathbb{R}^{d\times k}$ whose column spaces equal the eigenspace of $\hat{A}$ and A, respectively. Then there exists an orthonormal rotation matrix $O \in \mathbb{R}^{k\times k}$ such that*

$$\|\hat{Q} - QO\|_2 \leq \frac{\sqrt{2k}|||H|||_\infty}{\delta}$$

where $\delta = \min\{|\lambda - s| : \lambda$ eigenvalue of $A, \lambda \notin S, s \in S\}$.

Proof of Lemma B.1. Using the singular value decomposition, we can find orthonormal matrices P_1 and P_2 such that the singular values of $Q^\top\hat{Q}$ are exactly the cosines of the principal angles Θ, that is, we can find P_1 and P_2 such that $Q^\top\hat{Q} = P_1\Sigma P_2^\top$ where the diagonal of Σ contains the principal angles between the column space of $\hat{Q}$ and Q. Define the rotation matrix O as $O = P_1 P_2^\top$. Then, by definition of the Frobenius norm, the orthonormality of $\hat{Q}$ and Q

$$\begin{aligned}\|\hat{Q} - QO\|_2^2 &= \mathrm{Tr}\big((\hat{Q} - QO)^\top(\hat{Q} - QO)\big) = 2k - 2\,\mathrm{Tr}\big(OQ^\top\hat{Q}\big)\\ &= 2k - 2\,\mathrm{Tr}(\cos\Theta) = 2k - 2\sum_{i=1}^k \cos\theta_i\\ &\leq 2k - 2\sum_{i=1}^k \cos(\theta)_i^2 = 2k - 2k + 2\sum_{i=1}^k \sin(\theta)_i^2 = 2\|\sin\Theta\|_2^2.\end{aligned}$$

The classical Davis–Kahan theorem (Davis and Kahan [17]) then yields

$$\|\hat{Q} - QO\|_2^2 \le 2\|\sin\Theta\|_2^2 \le 2\frac{\|H\|_2^2}{\delta^2}.$$

Finally, since $\|H\|_2^2 \le k \max_j |\lambda_j^H|^2 = k|||H|||_\infty^2$, we obtain

$$\|\hat{Q} - QO\|_2 \le \sqrt{2k}\frac{|||H|||_\infty}{\delta}. \qquad \square$$

Corollary B.1. *There exists an orthonormal rotation matrix $O \in \mathbb{R}^{k\times k}$ such that*

$$\|\hat{U} - UO\|_2 \le \frac{2\sqrt{k}|||\hat{L} - L|||_\infty}{\lambda_{k+1}}$$

where λ_{k+1} is the $(k+1)$-th smallest eigenvalue of L.

Proof of Corollary B.1. By construction, $\hat{L}$ and L are symmetric and it is clear that we can view $\hat{L}$ as a perturbed version of L. Additionally, the columns of $\hat{U}$ and U contain the eigenvectors that correspond to the k smallest eigenvalues of $\hat{L}$ and L, respectively. It follows therefore directly from Lemma B.1 that

$$\|\hat{U} - UO\|_2 \le \frac{2\sqrt{k}|||\hat{L} - L|||_\infty}{\delta} \le \frac{2\sqrt{k}|||\hat{L} - L|||_\infty}{\lambda_{k+1}}.$$

The last inequality is a consequence of the following observation. The matrix L has exactly k zero eigenvalues. Hence if we take $S = [0, \epsilon)$ for arbitray small $\epsilon > 0$ or actually the singleton $S = \{0\}$, then the first k eigenvalues of L all belong to S. The smallest distance between eigenvalues that belong to S and that do not belong to S is thus given by $|0 - \lambda_{k+1}|$. Hence, $\delta = \lambda_{k+1}$. □

Proof of Lemma 3.2. We note that by definition we have $\hat{\mathcal{U}}_{i,\cdot} = \frac{\hat{U}_{i,\cdot}}{\|\hat{U}_{i,\cdot}\|_2}$ and $\mathcal{U}_{i,\cdot} = \frac{(UO)_{i,\cdot}}{\|U_{i,\cdot}\|_2}$. Therefore standard linear algebra shows

$$\begin{aligned}
\|\hat{\mathcal{U}} - \mathcal{U}\|_2^2 &= \sum_{i=1}^d \left\| \frac{\hat{U}_{i,\cdot}}{\|\hat{U}_{i,\cdot}\|_2} - \frac{(UO)_{i,\cdot}}{\|U_{i,\cdot}\|_2} \right\|_2^2 \\
&\le 2\sum_{i=1}^d \left\| \frac{\hat{U}_{i,\cdot}(\|U_{i,\cdot}\|_2 - \|\hat{U}_{i,\cdot}\|_2)}{\|\hat{U}_{i,\cdot}\|_2\|U_{i,\cdot}\|_2} \right\|_2^2 + \left\| \frac{\hat{U}_{i,\cdot} - (UO)_{i,\cdot}}{\|U_{i,\cdot}\|_2} \right\|_2^2 \\
&= 2\sum_{i=1}^d \frac{|\|U_{i,\cdot}\|_2 - \|\hat{U}_{i,\cdot}\|_2|^2}{\|U_{i,\cdot}\|_2^2} + \frac{\|\hat{U}_{i,\cdot} - (UO)_{i,\cdot}\|_2^2}{\|U_{i,\cdot}\|_2^2} \\
&\le 4\sum_{i=1}^d \frac{\|\hat{U}_{i,\cdot} - (UO)_{i,\cdot}\|_2^2}{\|U_{i,\cdot}\|_2^2} \\
&\le \frac{4}{\min_i \|U_{i,\cdot}\|_2^2} \left\|\hat{U} - (UO)\right\|_2^2 = \frac{4}{\min_i D_i} \left\|\hat{U} - (UO)\right\|_2^2.
\end{aligned}$$

The last equality follows since U collects eigenvectors of the form $\sqrt{D}\mathbb{1}_{C_l}$ for $l = 1, \dots, k$, where $\mathbb{1}_{C_l} \in \mathbb{R}^d$ denotes the indicator vector that equals 1 if point i belongs to component C_l. This means in particular that U has exactly one nonzero entry per row. A trivial lower bound on $\min_i D_i$ can be given by

$$\min_i \|U_{i,\cdot}\|_2^2 \geq \frac{\min_i D_i}{\mathcal{C}_{\max}}$$

where $\mathcal{C}_{\max} = \max_i \sum_{i_1 \in G_i} \sum_{i_2 \in G_i} W_{i_1,i_2}$. Hence, using Corollary B.1 and Lemma 3.1

$$\|\hat{\mathcal{U}} - \mathcal{U}\|_2 \leq 4\sqrt{k}\sqrt{\frac{\mathcal{C}_{\max}}{\min_i D_i}} \frac{\|\hat{L} - L\|_\infty}{\lambda_{k+1}} \to 0 \quad \text{as } T \to \infty.$$

□

B.3. Analyzing the k-means step

Using the properties of the row-normalized eigenvectors of L, we proceed by providing a definition of the set of points that are clustered correctly and then derive a bound on the complement set (see also Rohe, Chatterjee and Yu [50], Lei and Rinaldo [43]).

Lemma B.2. *Assume the graph has k components. Let $C^\star$ defined in* (3.6) *and $\mathcal{U}$ defined in* (3.4). *Then, the set of correctly clustered points is defined as the complement of the set*

$$\Sigma = \left\{ i : \left\|C^\star_{i,\cdot} - \mathcal{U}_{i,\cdot}\right\|_2 \geq \frac{1}{\sqrt{2}} \right\}. \tag{B.3}$$

Proof of Lemma B.2. By construction and using the properties of the Laplacian, $\mathcal{U}$ has exactly one 1 per row. All other entries in that row are zero. In total, there are k distinct rows which are orthonormal. Therefore, $\|\mathcal{U}_{i,\cdot} - \mathcal{U}_{j,\cdot}\|_2 = 0$ if the embedded points i and j belong to the same component and $\|\mathcal{U}_{i,\cdot} - \mathcal{U}_{j,\cdot}\|_2 = \sqrt{2}$ if they belong to different components. At the same time, $\|C^\star_{i,\cdot} - C^\star_{j,\cdot}\|_2 = 0$ if and only if the algorithm has clustered i, j in the same cluster. So let, i and j belong to Σ^{c}. Minkowski's inequality yields

$$\|\mathcal{U}_{i,\cdot} - \mathcal{U}_{j,\cdot}\|_2 \leq \left\|\mathcal{U}_{i,\cdot} - C^\star_{i,\cdot}\right\|_2 + \left\|C^\star_{i,\cdot} - C^\star_{j,\cdot}\right\|_2 + \left\|C^\star_{j,\cdot} - \mathcal{U}_{j,\cdot}\right\|_2 \leq 2\frac{1}{\sqrt{2}} = \sqrt{2}$$

if and only if i and j are clustered in the same cluster. Otherwise, we have a contradiction. Additionally, since $C^\star \in \mathcal{M}(d,k)$ clusters cannot be split. Hence, points in Σ^{c} must be correctly clustered □

Proof of Theorem 3.1. First note that $\mathcal{U} \in \mathcal{M}(d,k)$ since it has exactly k distinct rows. Consequently,

$$\operatorname*{arg\,min}_{C \in \mathcal{M}(d,k)} \|\hat{\mathcal{U}} - C\|_2^2 = \left\|\hat{\mathcal{U}} - C^\star\right\|_2^2 \leq \|\hat{\mathcal{U}} - \mathcal{U}\|_2^2,$$

$$\begin{aligned}
|\Sigma| = \sum_{i \in \Sigma} 1 &\leq \sum_{i \in \Sigma} 2\left\|C^\star_{i,\cdot} - \mathcal{U}_{i,\cdot}\right\|_2^2 \leq 2\left\|C^\star - \mathcal{U}\right\|_2^2 \\
&\leq 4\left(\left\|C^\star - \hat{\mathcal{U}}\right\|_2^2 + \|\hat{\mathcal{U}} - \mathcal{U}\|_2^2\right) \\
&= 8\|\hat{\mathcal{U}} - \mathcal{U}\|_2^2.
\end{aligned}$$

The result now follows from Lemma 3.2. □

Acknowledgements

This work has been supported in part by the Collaborative Research Center "Statistical modeling of nonlinear dynamic processes" (SFB 823, Teilprojekt A1, A7, C1) of the German Research Foundation (DFG). The authors would like to thank three anonymous referees and the editor for their constructive comments on an earlier version of this manuscript.

Supplementary Material

Online Supplement to "A similarity measure for second order properties of non-stationary functional time series with applications to clustering and testing" (DOI: 10.3150/20-BEJ1246SUPP; .pdf). This supplement contains additional background and technical results necessary to complete the proofs of statements in the main body of the paper. In Section S3 of the supplement, the clustering method is moreover illustrated by means of an application to high-resolution meteorological data.

References

[1] Abraham, C., Cornillon, P.A., Matzner-Løber, E. and Molinari, N. (2003). Unsupervised curve clustering using B-splines. *Scand. J. Stat.* **30** 581–595. MR2002229 https://doi.org/10.1111/1467-9469.00350

[2] Aghabozorgi, S., Sirkhorshidi, A.S. and Wah, Y.W. (2015). Time-series clustering – a decade review. *Inform. Sci.* **53** 16–38.

[3] Aston, J.A.D. and Kirch, C. (2012). Detecting and estimating changes in dependent functional data. *J. Multivariate Anal.* **109** 204–220. MR2922864 https://doi.org/10.1016/j.jmva.2012.03.006

[4] Aue, A. and van Delft, A. (2020). Testing for stationarity of functional time series in the frequency domain. *Ann. Statist.* **48** 2505–2547. https://doi.org/10.1214/19-AOS1895

[5] Bauwens, L. and Rombouts, J.V.K. (2007). Bayesian clustering of many Garch models. *Econometric Rev.* **26** 365–386. MR2364366 https://doi.org/10.1080/07474930701220576

[6] Böhm, H., Ombao, H., von Sachs, R. and Sanes, J. (2010). Classification of multivariate nonstationary signals: The SLEX-shrinkage approach. *J. Statist. Plann. Inference* **140** 3754–3763. MR2674163 https://doi.org/10.1016/j.jspi.2010.04.040

[7] Bosq, D. (2000). *Linear Processes in Function Spaces: Theory and Applications. Lecture Notes in Statistics* **149**. New York: Springer. MR1783138 https://doi.org/10.1007/978-1-4612-1154-9

[8] Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Commun. Stat.* **3** 1–27. MR0375641 https://doi.org/10.1080/03610927408827101

[9] Chamroukhi, F. and Nguyen, H.D. (2019). Model-based clustering and classification of functional data. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **9** e1298. https://doi.org/10.1002/widm.1298

[10] Chandler, G. and Polonik, W. (2006). Discrimination of locally stationary time series based on the excess mass functional. *J. Amer. Statist. Assoc.* **101** 240–253. MR2268042 https://doi.org/10.1198/016214505000000899

[11] Charrad, M., Ghazzali, N., Boiteau, V. and Niknafs, A. (2014). NbClust: An R package for determining the relevant number of clusters in a data set. *J. Stat. Softw.* **61** 1–36.

[12] Chung, F. and Radcliffe, M. (2011). On the spectra of general random graphs. *Electron. J. Combin.* **18** Paper 215, 14. MR2853072

[13] Chung, F.R.K. (1997). *Spectral Graph Theory. CBMS Regional Conference Series in Mathematics* **92**. Washington, DC: Published for the Conference Board of the Mathematical Sciences; Providence, RI: Amer. Math. Soc. MR1421568

[14] Coates, D.S. and Diggle, P.J. (1986). Tests for comparing two estimated spectral densities. *J. Time Series Anal.* **7** 7–20. MR0832349 https://doi.org/10.1111/j.1467-9892.1986.tb00482.x

[15] Corduas, M. and Piccolo, D. (2008). Time series clustering and classification by the autoregressive metric. *Comput. Statist. Data Anal.* **52** 1860–1872. MR2418476 https://doi.org/10.1016/j.csda.2007.06.001

[16] Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *Ann. Statist.* **25** 1–37. MR1429916 https://doi.org/10.1214/aos/1034276620

[17] Davis, C. and Kahan, W.M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.* **7** 1–46. MR0264450 https://doi.org/10.1137/0707001

[18] Delaigle, A., Hall, P. and Pham, T. (2019). Clustering functional data into groups by using projections. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 271–304. MR3928143 https://doi.org/10.1111/rssb.12310

[19] Dette, H. (2009). Bootstrapping frequency domain tests in multivariate time series with an application to comparing spectral densities. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 831–857. MR2750097 https://doi.org/10.1111/j.1467-9868.2009.00709.x

[20] Dette, H. and Hildebrandt, T. (2012). A note on testing hypotheses for stationary processes in the frequency domain. *J. Multivariate Anal.* **104** 101–114. MR2832189 https://doi.org/10.1016/j.jmva.2011.07.002

[21] Eichler, M. (2008). Testing nonparametric and semiparametric hypotheses in vector stationary processes. *J. Multivariate Anal.* **99** 968–1009. MR2405101 https://doi.org/10.1016/j.jmva.2007.06.003

[22] Euán, C., Ombao, H. and Ortega, J. (2018). Spectral synchronicity in brain signals. *Stat. Med.* **37** 2855–2873. MR3832245 https://doi.org/10.1002/sim.7695

[23] Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice. Springer Series in Statistics*. New York: Springer. MR2229687

[24] Floriello, D. and Vitelli, V. (2017). Sparse clustering of functional data. *J. Multivariate Anal.* **154** 1–18. MR3588554 https://doi.org/10.1016/j.jmva.2016.10.008

[25] Fokianos, K. and Promponas, V.J. (2012). Biological applications of time series frequency domain clustering. *J. Time Series Anal.* **33** 744–756. MR2969908 https://doi.org/10.1111/j.1467-9892.2011.00758.x

[26] Frühwirth-Schnatter, S. and Kaufmann, S. (2008). Model-based clustering of multiple time series. *J. Bus. Econom. Statist.* **26** 78–89. MR2422063 https://doi.org/10.1198/073500107000000106

[27] Gordon, A.D. (1999). *Classification*, 2nd ed. London: Chapman and Hall–CRC. MR0637465

[28] Hartigan, J.A. (1975). *Clustering Algorithms. Wiley Series in Probability and Mathematical Statistics*. New York: Wiley. MR0405726

[29] Harvill, J.L., Kohli, P. and Ravishanker, N. (2017). Clustering nonlinear, nonstationary time series using BSLEX. *Methodol. Comput. Appl. Probab.* **19** 935–955. MR3683978 https://doi.org/10.1007/s11009-016-9528-1

[30] Heard, N.A., Holmes, C.C. and Stephens, D.A. (2006). A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: An application of Bayesian hierarchical clustering of curves. *J. Amer. Statist. Assoc.* **101** 18–29. MR2252430 https://doi.org/10.1198/016214505000000187

[31] Holan, S.H. and Ravishanker, N. (2018). Time series clustering and classification via frequency domain methods. *Wiley Interdiscip. Rev.: Comput. Stat.* **10** e1444, 15. MR3873674 https://doi.org/10.1002/wics.1444

[32] Horváth, L., Hušková, M. and Rice, G. (2013). Test of independence for functional data. *J. Multivariate Anal.* **117** 100–119. MR3053537 https://doi.org/10.1016/j.jmva.2013.02.005

[33] Huang, H.-Y., Ombao, H. and Stoffer, D.S. (2004). Discrimination and classification of nonstationary time series using the SLEX model. *J. Amer. Statist. Assoc.* **99** 763–774. MR2090909 https://doi.org/10.1198/016214504000001105

[34] Ieva, F., Paganoni, A.M., Pigoli, D. and Vitelli, V. (2013). Multivariate functional clustering for the morphological analysis of electrocardiograph curves. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **62** 401–418. MR3060623 https://doi.org/10.1111/j.1467-9876.2012.01062.x

[35] Jacques, J. and Preda, C. (2014). Model-based clustering for multivariate functional data. *Comput. Statist. Data Anal.* **71** 92–106. MR3131956 https://doi.org/10.1016/j.csda.2012.12.004

[36] Jacques, J. and Preda, C. (2014). Functional data clustering: A survey. *Adv. Data Anal. Classif.* **8** 231–255. MR3253859 https://doi.org/10.1007/s11634-013-0158-y

[37] Jentsch, C. and Pauly, M. (2015). Testing equality of spectral densities using randomization techniques. *Bernoulli* **21** 697–739. MR3338644 https://doi.org/10.3150/13-BEJ584

[38] Juárez, M.A. and Steel, M.F.J. (2010). Model-based clustering of non-Gaussian panel data based on skew-*t* distributions. *J. Bus. Econom. Statist.* **28** 52–66. MR2650600 https://doi.org/10.1198/jbes.2009.07145

[39] Kakizawa, Y., Shumway, R.H. and Taniguchi, M. (1998). Discrimination and clustering for multivariate time series. *J. Amer. Statist. Assoc.* **93** 328–340. MR1614589 https://doi.org/10.2307/2669629

[40] Kalpakis, K., Gada, D. and Puttagunta, V. (2001). Distance measures for effective clustering of arima time-series. In *Proceedings of the* 2001 *IEEE International Conference on Data Mining*, *San Jose* 273–280.

[41] Kaufman, L. and Rousseeuw, P.J. (1990). *Finding Groups in Data*: *An Introduction to Cluster Analysis*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. New York: Wiley. A Wiley-Interscience Publication. MR1044997 https://doi.org/10.1002/9780470316801

[42] 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** 23–34. MR0931626 https://doi.org/10.2307/2531893

[43] Lei, J. and Rinaldo, A. (2015). Consistency of spectral clustering in stochastic block models. *Ann. Statist.* **43** 215–237. MR3285605 https://doi.org/10.1214/14-AOS1274

[44] Leucht, A., Paporoditis, E. and Sapatinas, T. (2018). Testing equality of spectral density operators for functional linear processes. arXiv:1804.03366.

[45] Milligan, G.W. and Cooper, M.C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika* **50** 159–179.

[46] Ng, A., Jordan, S. and Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems* 14 (T. Dietterich, S. Becker and Z. Ghahramani, eds.) MIT Press.

[47] Ombao, H.C., Raz, J.A., von Sachs, R. and Malow, B.A. (2001). Automatic statistical analysis of bivariate nonstationary time series. *J. Amer. Statist. Assoc.* **96** 543–560. MR1946424 https://doi.org/10.1198/016214501753168244

[48] Paparoditis, E. and Sapatinas, T. (2016). Bootstrap-based testing of equality of mean functions or equality of covariance operators for functional data. *Biometrika* **103** 727–733. MR3551795 https://doi.org/10.1093/biomet/asw033

[49] Peng, J. and Müller, H.-G. (2008). Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. *Ann. Appl. Stat.* **2** 1056–1077. MR2516804 https://doi.org/10.1214/08-AOAS172

[50] Rohe, K., Chatterjee, S. and Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39** 1878–1915. MR2893856 https://doi.org/10.1214/11-AOS887

[51] Sakiyama, K. and Taniguchi, M. (2004). Discriminant analysis for locally stationary processes. *J. Multivariate Anal.* **90** 282–300. MR2081780 https://doi.org/10.1016/j.jmva.2003.08.002

[52] Savvides, A., Promponas, V.J. and Fokianos, K. (2008). Clustering of biological time series by cepstral coefficients based distances. *Pattern Recognit.* **41** 2398–2412.

[53] Shi, J. and Malik, J. (2002). Nomalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **22** 888–905.

[54] Stewart, G.W. and Sun, J.G. (1990). *Matrix Perturbation Theory*. *Computer Science and Scientific Computing*. Boston, MA: Academic Press. MR1061154

[55] Tavakoli, S. and Panaretos, V.M. (2016). Detecting and localizing differences in functional time series dynamics: A case study in molecular biophysics. *J. Amer. Statist. Assoc.* **111** 1020–1035. MR3561926 https://doi.org/10.1080/01621459.2016.1147355

[56] Theodoridis, S. and Koutroumbas, K. (2008). *Pattern Recognition*, 4th ed. New York: Academic Press.

[57] 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. MR1841503 https://doi.org/10.1111/1467-9868.00293

[58] van Delft, A. (2020). A note on quadratic forms of stationary functional time series under mild conditions. *Stochastic Process. Appl.* **130** 4206–4251. MR4102264 https://doi.org/10.1016/j.spa.2019.12.002

[59] van Delft, A., Characiejus, V. and Dette, H. (2019). A nonparametric test for stationarity in functional time series. *Statist. Sinica.* To appear. https://doi.org/10.5705/ss.202018.0320

[60] van Delft, A. and Dette, H. (2020). Supplement to "A similarity measure for second order properties of non-stationary functional time series with applications to clustering and testing." https://doi.org/10.3150/20-BEJ1246SUPP

[61] van Delft, A. and Eichler, M. (2018). Locally stationary functional time series. *Electron. J. Stat.* **12** 107–170. MR3746979 https://doi.org/10.1214/17-EJS1384

[62] Vlachos, M., Lin, J., Keogh, E. and Gunopulos, D. (2003). A wavelet-based anytime algorithm for k-means clustering of time series. In *Proc*. *Workshop on Clustering High Dimensionality Data and Its Applications*, *San Francisco*.

[63] von Luxburg, U. (2007). A tutorial on spectral clustering. *Stat*. *Comput*. **17** 395–416. MR2409803 https://doi.org/10.1007/s11222-007-9033-z

[64] von Luxburg, U., Belkin, M. and Bousquet, O. (2008). Consistency of spectral clustering. *Ann*. *Statist*. **36** 555–586. MR2396807 https://doi.org/10.1214/009053607000000640

[65] Zhang, X. and Shao, X. (2015). Two sample inference for the second-order property of temporally dependent functional data. *Bernoulli* **21** 909–929. MR3338651 https://doi.org/10.3150/13-BEJ592

Received June 2019 and revised June 2020