Received: 4 May 2021 | Accepted: 16 June 2021

DOI: 10.1002/for.2801

RESEARCH ARTICLE

WILEY

Modeling interval trendlines: Symbolic singular spectrum analysis for interval time series

Miguel de Carvalho[1] | **Gabriel Martos**[2]

[1]School of Mathematics, University of Edinburgh, Edinburgh, UK

[2]Department of Mathematics and Statistics, Universidad Torcuato Di Tella, Buenos Aires, Argentina

Correspondence Miguel de Carvalho, School of Mathematics, The University of Edinburgh, James Clerk Maxwell Building, Peter Guthrie Tait Road, Edinburgh EH9 3FD, UK.
Email: miguel.decarvalho@ed.ac.uk

Funding information
Fundação para a Ciência e a Tecnologia, Grant/Award Number: UID/MAT/00006/2019; International Research and Partnership Fund (Develop ing Countries), Grant/Award Number: PTDC/MAT-STA/28649/2017; University Of Edinburgh; INTERSTATA (Interdisciplinary Statistics in Action); FCT (Fundação para a Ciência e a Tecnologia, Portugal)

Abstract

In this article we propose an extension of singular spectrum analysis for interval-valued time series. The proposed methods can be used to decompose and forecast the dynamics governing a set-valued stochastic process. The resulting components on which the interval time series is decomposed can be understood as interval trendlines, cycles, or noise. Forecasting can be conducted through a linear recurrent method, and we devised generalizations of the decomposition method for the multivariate setting. The performance of the proposed methods is showcased in a simulation study. We apply the proposed methods so to track the dynamics governing the Argentina Stock Market (MERVAL) in real time, in a case study over a period of turbulence that led to discussions of the government of Argentina with the International Monetary Fund.

KEYWORDS

decomposition of interval-valued time series, interval data, interval-valued signal, set-valued stochastic process, singular spectrum analysis, symbolic data analysis

1 | INTRODUCTION

Modeling and forecasting time series with singular spectrum analysis (SSA) has received considerable attention in recent forecasting literature (de Carvalho & Martos, 2020; Hassani & Mahmoudvand, 2013; Khan & Poskitt, 2017; Mahmoudvand & Rodrigues, 2018). The rising popularity of the methods stems from the fact that SSA—along with its multivariate version—is naturally tailored for both forecasting and decomposing univariate or multivariate nonstationary time series into a set of principal components, which can be interpreted as trends, cyclical components, or noise. Applications of SSA in practice include predicting inflation dynamics, tracking business cycles, and forecasting industrial production, among others, which can be found in the papers above and references therein. For a time series $\mathbf{y} = (a_1, \ldots, a_n)$, a key step on which SSA relies is on the singular value decomposition of a matrix $\boldsymbol{Y}$ containing rolling windows of length l, that is

$$\boldsymbol{Y} = \begin{bmatrix} a_1 & a_2 & \cdots & a_k \\ a_2 & a_3 & \cdots & a_{k+1} \\ \vdots & \vdots & \ddots & \vdots \\ a_l & a_{l+1} & \cdots & a_n \end{bmatrix} = \sum_{i=1}^{d} \sqrt{\lambda_i} \mathbf{u}_i \mathbf{v}_i^{\mathrm{T}}. \qquad (1)$$

Here $\lambda_1 \geq \cdots \geq \lambda_l$ and $\mathbf{u}_1, \ldots, \mathbf{u}_l$ are respectively the eigenvalues and the eigenvectors of $\boldsymbol{Y}\boldsymbol{Y}^{\boldsymbol{T}}$, and $\mathbf{v}_i = \boldsymbol{Y}^{\mathrm{T}}\mathbf{u}_i/\sqrt{\lambda_i}$ with $d = \max\{i \in \{1, \ldots, l\} : \lambda_i > 0\}$.

One of the main goals of this paper is to develop SSA methods to model and forecast interval time series, $\mathbf{y} = ([a_1, b_1], \ldots, [a_n, b_n])$. This paper thus contributes to both the literature on singular spectrum analysis as well as that on interval time series. There has been an increasing interest on interval time series as can be seen from González-Rivera and Arroyo (2012), González-Rivera and Lin (2013), Rodrigues and Salish (2015), Lin and González-Rivera (2016), and Wang et al. (2016). Interval time series are natural for settings where the interest is on modeling the dynamics of a range of values, such as, for instance, in financial time series where one is interested in modeling the interval of prices during a trading session (low, high). It is by now well known that naive 'midpoint' analyses discard the so-called internal or within variation (Le-Rademacher & Billard, 2012), which in the case of the financial example mentioned earlier corresponds to ignoring intra-day variation. As recently discussed by Sun et al. (2018), advantages of interval time series over a point-valued analysis include the facts that they (i) can be used to learn about both trends and volatilities—whereas point-valued approaches often lead to informational losses; (ii) lead to more effienct inferences; and (iii) account for variation over time, while avoiding undesirable noises from high-frequency point-valued data.

A main methodological goal of this article is on developing univariate and multivariate singular spectrum analysis methods for interval time series that can be used for modeling, decomposing, and forecasting. We underscore that our approach does not consist of a bivariate point-valued model, but rather will build over ideas, concepts, and methods from a relatively new field of Statistics known as symbolic data analysis (Billard & Diday, 2003; Billard, 2006), so to take on board the interval-valued nature of the data. The proposed approach pioneers the development of a symbolic version of singular spectrum analysis, and it will disentangle the dynamics of an interval time series into a sequence of set-valued stochastic processes (Kisielewicz, 2013) that can be interpreted as components underlying trends, regular movements, and noise. Hence, the proposed methods can be used to decompose and forecast the dynamics governing a set-valued stochastic process.

The article is organized as follows. In the next section we introduce symbolic singular spectrum analysis. A simulation study is reported in Section 3. An application to stock market data can be found in Section 4. The supplementary materials provide supporting numerical experiments, and the `R` (R Development Core Team, 2016) package `ASSA` (de Carvalho & Martos, 2018) can be used to implement the proposed methods.

2 | SYMBOLIC SINGULAR SPECTRUM ANALYSIS

2.1 | Preparations

Below, the unit of analysis will be an interval-valued time series, $\mathbf{y} = ([a_1, b_1], \ldots, [a_n, b_n])$. For modeling $\mathbf{y}$ using singular spectrum analysis, we resort to a special type of block matrix to which we refer to as a matrix of ordered pairs. Below, a matrix $\mathbf{Y}$ is said to be an $k \times l$ matrix of ordered pairs if its elements are ordered pairs, that is

$$\mathbf{Y} = \begin{bmatrix} (a_{1,1}, b_{1,1}) & \cdots & (a_{1,k}, b_{1,k}) \\ \vdots & \cdots & \vdots \\ (a_{l,1}, b_{l,1}) & \cdots & (a_{l,k}, b_{l,k}) \end{bmatrix}.$$

Throughout, sums and differences between such matrices should be understood as their respective pointwise Minkowski-type counterparts, respectively, defined as

$$\mathbf{A}+\mathbf{B} = \begin{bmatrix} (a_{1,1}+c_{1,1}, b_{1,1}+d_{1,1}) & \cdots & (a_{1,k}+c_{1,k}, b_{1,k}+d_{1,k}) \\ \vdots & \ddots & \vdots \\ (a_{l,1}+c_{l,1}, b_{l,1}+d_{l,1}) & \cdots & (a_{l,k}+c_{l,k}, b_{l,k}+d_{l,k}) \end{bmatrix},$$

$$\mathbf{A}\text{-}\mathbf{B} = \begin{bmatrix} (a_{1,1}-c_{1,1}, b_{1,1}-d_{1,1}) & \cdots & (a_{1,k}-c_{1,k}, b_{1,k}-d_{1,k}) \\ \vdots & \ddots & \vdots \\ (a_{l,1}-c_{l,1}, b_{l,1}-d_{l,1}) & \cdots & (a_{l,k}-c_{l,k}, b_{l,k}-d_{l,k}) \end{bmatrix},$$

where

$$\mathbf{A} = \begin{bmatrix} (a_{1,1}, b_{1,1}) & \cdots & (a_{1,k}, b_{1,k}) \\ \vdots & \ddots & \vdots \\ (a_{l,1}, b_{l,1}) & \cdots & (a_{l,k}, b_{l,k}) \end{bmatrix},$$

$$\mathbf{B} = \begin{bmatrix} (c_{1,1}, d_{1,1}) & \cdots & (c_{1,k}, d_{1,k}) \\ \vdots & \ddots & \vdots \\ (c_{l,1}, d_{l,1}) & \cdots & (c_{l,k}, d_{l,k}) \end{bmatrix}.$$

The components of the resulting matrices $\mathbf{A} + \mathbf{B}$ and $\mathbf{A} - \mathbf{B}$ can be naturally mapped into interval data via the set-valued function $\phi(x,y) = [\min\{x,y\}, \max\{x,y\}]$; note that $\phi(a_{i,j} + c_{i,j}, b_{i,j} + d_{i,j}) = [a_{i,j} + c_{i,j}, b_{i,j} + d_{i,j}]$ and $\phi(a_{i,j} - c_{i,j}, b_{i,j} - d_{i,j}) = [\min\{a_{i,j} - c_{i,\ j}, b_{i,j} - d_{i,j}\}, \max\{a_{i,j} - c_{i,j}, b_{i,j} - d_{i,j}\}]$ for $1 \le i \le l$, and $1 \le j \le k$ respectively. To

compute the norm of $\mathbf{Y}$, we use the following variant of the Frobenious norm:

$$||\mathbf{Y}||_{\mathrm{C}} = \frac{1}{\sqrt{2}} \left\{ \sum_{i=1}^{l} \sum_{j=1}^{k} (a_{i,j}^2 + b_{i,j}^2) \right\}^{1/2}, \quad (2)$$

for a matrix of ordered pairs $\mathbf{Y} = \{(a_{i,j}, b_{i,j})\}$; note that if $a_{i,j} = b_{i,j}$ we recover the standard Frobenious norm. Of particular interest for our developments is the class of what we will refer to as Hankel matrices of ordered pairs. An $l \times k$ matrix of ordered pairs is said to be an Hankel matrix (of ordered pairs) if its elements coincide on the antidiagonals $i+j=s$ for any $2 \le s \le l+k$. The notation $\mathscr{H}_{l,k}$ will be used throughout to denote the space of all $l \times k$ Hankel matrices of ordered pairs.

2.2 | Interval-valued singular spectrum analysis (IVSSA)

Let $\mathbf{y} = ([a_1, b_1], \ldots, [a_n, b_n])$ be an interval-valued time series. Interval-valued singular spectrum analysis (IVSSA), to be proposed below, can be regarded as an extension of singular spectrum analysis (Golyandina & Zhigljavsky, 2013) to be used for decomposing an interval-valued time series into components, and for learning about interval trendlines from data. IVSSA entails two phases, namely decomposition and reconstruction, and each of these phases includes two steps. The decomposition includes the steps of embedding and symbolic singular value decomposition, which we discuss below.

EMBEDDING. IVSSA starts by organizing the original interval-valued time series of interest, $\mathbf{y} = ([a_1, b_1], \ldots, [a_n, b_n])$, into a trajectory matrix $\boldsymbol{Y}$, that is, a matrix of ordered pairs whose columns consist of rolling windows of length l, as follows

$$\boldsymbol{Y} = \begin{bmatrix} (a_1, b_1) & (a_2, b_2) & \cdots & (a_k, b_k) \\ (a_2, b_2) & (a_3, b_3) & \cdots & (a_{k+1}, b_{k+1}) \\ \vdots & \vdots & \ddots & \vdots \\ (a_l, b_l) & (a_{l+1}, b_{l+1}) & \cdots & (a_n, b_n) \end{bmatrix}, \quad (3)$$

where l is set by the user and $k = n - l + 1$. Here l is a signal–noise separation parameter that plays a similar role to that of the bandwidth in nonparametric regression. Since all elements over the diagonal $i+j = \text{const}$ are equal, then $\boldsymbol{Y} \in \mathscr{H}_{l,k}$.

SYMBOLIC SINGULAR VALUE DECOMPOSITION. In the second step we perform a symbolic singular value decomposition of the trajectory matrix resorting to the so-called covariance matrix for symbolic data, as defined in Billard (2007) and Le-Rademacher and Billard (2012). Let $\lambda_1^{\mathrm{S}} \ge \cdots \ge \lambda_l^{\mathrm{S}}$ be the eigenvalues and $\mathbf{u}_1^{\mathrm{S}}, \ldots, \mathbf{u}_l^{\mathrm{S}}$ be the eigenvectors corresponding to matrix $\boldsymbol{S} \in \mathbb{R}^{l \times l}$, with entries given by

$$[\boldsymbol{S}]_{jj'} = s_{jj'} = \frac{1}{6} \sum_{i=1}^{k} \Big\{ 2a_{j+i-1} a_{j'+i-1} + a_{j+i-1} b_{j'+i-1} + b_{j+i-1} a_{j'+i-1} + 2b_{j+i-1} b_{j'+i-1} \Big\},$$

where $s_{jj'}$, for $1 \le j \le l$ and $1 \le j' \le l$, up to a constant, represent the estimated covariance between interval data in rows j and j' on $\boldsymbol{Y}$ (Billard & Le-Rademacher, 2012, Equation 3). We resort on the eigenvectors and eigenvalues of $\boldsymbol{S}$ to decompose the trajectory matrix $\boldsymbol{Y}$ as follows

$$\boldsymbol{Y} = \sum_{i=1}^{d} \boldsymbol{Y}_i, \quad (4)$$

where $\boldsymbol{Y}_i = \sqrt{\lambda_i^{\mathrm{S}}} \mathbf{u}_i^{\mathrm{S}} (\mathbf{v}_i^{\mathrm{S}})^{\mathrm{T}}$, $\mathbf{v}_i^{\mathrm{S}} = \boldsymbol{Y}^{\mathrm{T}} \mathbf{u}_i^{\mathrm{S}} / \sqrt{\lambda_i^{\mathrm{S}}}$ and $d = \max\{i \in \{1, \ldots, l\} : \lambda_i^{\mathrm{S}} > 0\}$. Notice that

$$\mathbf{v}_i^{\mathrm{S}} = \frac{1}{\sqrt{\lambda_i^{\mathrm{S}}}} \left[\left(\sum_{j=1}^{l} u_{i,j}^{\mathrm{S}} a_j, \ \sum_{j=1}^{l} u_{i,j}^{\mathrm{S}} b_j \right), \ldots, \left(\sum_{j=1}^{l} u_{i,j}^{\mathrm{S}} a_{j+k-1}, \ \sum_{j=1}^{l} u_{i,j}^{\mathrm{S}} b_{j+k-1} \right) \right]^{\mathrm{T}};$$

thus, the resulting matrices $\boldsymbol{Y}_i$ are matrices of ordered pairs, for $i = 1, \ldots, d$. Next we discuss the reconstruction phase, which involves the steps of grouping components and diagonal averaging.

GROUPING. Not all terms in Equation (4) contain relevant information on the interval trendline, and hence we retain only a subset $I \subset \{1, \ldots, d\}$ to compute $\boldsymbol{Y}_I = \sum_{i \in I} \boldsymbol{Y}_i$. The goal of this step is on disentangling the signal from noise, assuming that $\boldsymbol{Y} = \boldsymbol{Y}_I + \boldsymbol{\varepsilon}$, being $\boldsymbol{Y}_I$ the signal and $\boldsymbol{\varepsilon} = \sum_{i \notin I} \boldsymbol{Y}_i$ the noise on the data. To learn about I, in Appendix A.2 we show how the periodogram-based method of de Carvalho and Martos (2020) can be extended to an interval-valued time series context by devising a periodogram for interval-valued time series. The proposed extension is based on the analysis of the periodogram of an interval-valued time series of residuals (termed below as Hausdorff residuals). The strengths and limitations with such periodogram-based approach will be numerically examined in Section 3.

DIAGONAL AVERAGING. In this step we average over all the elements of the (anti)diagonal $i+j = const$ of $\boldsymbol{Y}_I$ so to

obtain an Hankel matrix of ordered pairs, from where our interval trendline indicator results. The following proposition provides the formal justification for this step; see Appendix A.1 for a proof.

Proposition 1. Let $\mathscr{H}_{l,k}$ be the space of all $l \times k$ Hankel matrices of ordered pairs. Let $\mathbf{Y} = \{(a_{i,j}, b_{i,j})\}$ be an $l \times k$ matrix of ordered pairs. Then,

$$\mathbf{H}^* = \{(\alpha^*_{i,j}, \beta^*_{i,j})\} = \arg\min_{\mathbf{H} \in \mathscr{H}_{l,k}} \|\mathbf{Y} - \mathbf{H}\|_C,$$

where $(\alpha^*_{i,j}, \beta^*_{i,j}) = (1/n_s \sum_{i+j=s} a_{i,j}, 1/n_s \sum_{i+j=s} b_{i,j})$, with n_s denoting the number of (i,j) such that $i+j=s$, with $i \in \{1,\ldots,l\}$ and $j \in \{1,\ldots,k\}$.

Thus, following Proposition 1, we construct our interval trendline indicator by averaging the matrix of ordered pairs $\boldsymbol{Y}_I$ over the antidiagonals $i+j=s$. Let $(a^I_{ij}, b^I_{ij}) = [\mathbf{Y}_I]_{ij}$ then for $s=2$, $\tilde{y}_1 = \phi(a^I_{11}, b^I_{11})$; $s=3$, yields $\tilde{y}_2 = \phi(a^I_{12} + a^I_{21}, b^I_{12} + b^I_{22})/2$; $s=4$, yields $\tilde{y}_3 = \phi(a^I_{13} + a^I_{22} + a^I_{31}, b^I_{13} + b^I_{22} + b^I_{31})/3$; etc. Extending this simple construct, we build our interval trendline indicator through the map

$$\begin{aligned} \widetilde{\mathbf{y}} &= \{[\widetilde{a}_t, \widetilde{b}_t]\}_{t=1}^{n} = \overline{\mathbb{D}}(\boldsymbol{Y}_I) \\ &\equiv \left\{ \frac{1}{n_2} \phi\left(\sum_{i+j=2} a^I_{ij}, \sum_{i+j=2} b^I_{ij} \right), \ldots, \right. \\ &\quad \left. \frac{1}{n_{n+1}} \phi\left(\sum_{i+j=n+1} a^I_{ij}, \sum_{i+j=n+1} b^I_{ij} \right) \right\}, \end{aligned} \tag{5}$$

with n_s denoting the number of (i,j) such that $i+j=s$, with $i \in \{1,\ldots,l\}$ and $j \in \{1,\ldots,k\}$.

2.3 | Selected comments on forecasting with IVSSA

Similarly to SSA for time series analysis, forecasting can be here conducted via a recurrent forecasting algorithm (Golyandina & Zhigljavsky, 2013, Chapter 3). The recurrent forecasting method relies on an autoregressive-type assumption that specifies that the i*th* interval observation $y_i = [a_i, b_i]$ is a combination of the preceding $l-1$ observations, so that for all $i \geq l$ it holds that

$$\begin{aligned} [a_i, b_i] &= \phi(\alpha_1(a_{i-1}, b_{i-1}) + \cdots + \alpha_{l-1}(a_{i-l+1}, b_{i-l+1})) \\ &= \phi\left(\sum_{j=1}^{l-1} \alpha_j a_{i-j}, \sum_{j=1}^{l-1} \alpha_j b_{i-j} \right), \end{aligned} \tag{6}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{l-1})$ is a vector of coefficients. The specification in (6) can then be used for forecasting. For example, the one-step forecast, $[\hat{a}_{n+1}, \hat{b}_{n+1}]$, is a combination of the most recent $l-1$ interval-valued signals, that is

$$\begin{aligned} [\hat{a}_{n+1}, \hat{b}_{n+1}] &= \phi(\alpha_1(\widetilde{a}_n, \widetilde{b}_n) + \cdots + \alpha_{l-1}(\widetilde{a}_{n-l+2}, \widetilde{b}_{n-l+2})) \\ &= \phi\left(\sum_{j=1}^{l-1} \alpha_j \widetilde{a}_{n+1-j}, \sum_{j=1}^{l-1} \alpha_j \widetilde{b}_{n+1-j} \right). \end{aligned}$$

And the out-of-sample forecasts corresponding to the time periods $n+2$, $n+3, \ldots$, are obtained using the previous formula recursively. The question of forecasting via (6) boils down to obtaining the vector $\boldsymbol{\alpha}$, which can be retrieved from the symbolic singular value decomposition via Golyandina et al. (2001, Proposition1). Following Rodrigues and de Carvalho (2013) the vector $\boldsymbol{\alpha}$ can be computed as follows

$$\frac{\boldsymbol{\alpha} = \boldsymbol{M}[\boldsymbol{\Pi}_1 \bigodot (\boldsymbol{\pi}_1 \bigotimes \mathbf{1}_{l-1})]\mathbf{1}_m}{1 - ||\boldsymbol{\pi}_1||^2,} \tag{7}$$

where $\|\cdot\|$ is the Euclidean norm, $\bigodot$ and $\bigotimes$ are the Hadamard and tensor Kronecker products, respectively, and $\boldsymbol{M} \in \mathbb{R}^{(l-1)\times(l-1)}$ is an antidiagonal matrix with ones in the main antidiagonal. In addition, $\boldsymbol{\Pi}_1 \in \mathbb{R}^{(l-1)\times m}$ is a matrix composed by the first $(l-1)$ components of the m eigenvectors associated to signal, whereas $\boldsymbol{\pi}_1 \in \mathbb{R}^{1\times m}$ contains the last components of those eigenvectors.

The next section will consider multivariate extensions of IVSSA.

2.4 | Multivariate extensions

Suppose now that we observe D interval time series, $\{\mathbf{y}_i = [a_{ij},\ b_{ij}]\}_{j=1}^{N_i}$, where N_i denotes the series ith length, for $i = 1, \ldots, D$. Multivariate IVSSA (MIVSSA) entails a similar course of action as that described in Section 2.2. To streamline the discussion we assume that $N_1 = \cdots = N_D \equiv N$ and $l_1 = \cdots = l_D \equiv l$, but all steps below can be easily adapted otherwise.

EMBEDDING AND SINGULAR VALUE DECOMPOSITION. Let $\boldsymbol{Y}_i \in \mathcal{H}_{l,k}$ be the interval trajectory matrix corresponding to the ith series; we consider either of the two stacked trajectory matrix:

$$\boldsymbol{Y}_V = \begin{bmatrix} \boldsymbol{Y}_1 \\ \vdots \\ \boldsymbol{Y}_D \end{bmatrix}, \text{ or } \boldsymbol{Y}_H = [\boldsymbol{Y}_1 \cdots \boldsymbol{Y}_D], \tag{8}$$

where $\boldsymbol{Y}_V \in \mathcal{H}_{lD,k}$ stand for vertical stack and $\boldsymbol{Y}_H \in \mathcal{H}_{l,kD}$ for horizontal stack. In regard to the stacking strategy, we consider the eigen–pairs corresponding to matrices:

$$\boldsymbol{S}_V = \begin{bmatrix} \boldsymbol{S}_{11} & \cdots & \boldsymbol{S}_{1D} \\ \vdots & \ddots & \vdots \\ \boldsymbol{S}_{D1} & \cdots & \boldsymbol{S}_{DD} \end{bmatrix}, \quad \boldsymbol{S}_H = \sum_{i=1}^{D} \boldsymbol{S}_{ii}, \tag{9}$$

where $[\boldsymbol{S}_{ii'}]_{jj'} = \sum_{q=1}^{k} \{2a_{i,j+q-1}a_{i',j'+q-1} + a_{i,j+q-1}b_{i',j'+q-1} + b_{i,j+q-1}a_{i',j'+q-1} + 2b_{i,j+q-1}b_{i',j'+q-1}\}/6$, for $j,j' = 1,\ldots,l$ and $i,i' = 1,\ldots,D$. For vertical staking, the elements in the diagonal of $\boldsymbol{S}_V$ correspond to different interval covariance matrices obtained when applying IVSSA on each interval time series separately. Considering vertical stacking and denoting as $\{(\lambda_1^{\mathrm{v}}, \mathbf{u}_1^{\mathrm{v}}), \ldots, (\lambda_{lD}^{\mathrm{v}}, \mathbf{u}_{lD}^{\mathrm{v}})\}$ the eigen–pairs of $\boldsymbol{S}_V$, then $\boldsymbol{Y}_V = \sum_{i=1}^{d} \boldsymbol{Y}_i$, where $\boldsymbol{Y}_i = \sqrt{\lambda_i^{\mathrm{v}}}\mathbf{u}_i^{\mathrm{v}}(\mathbf{v}_i^{\mathrm{v}})^T$ and $\mathbf{v}_i^{\mathrm{v}} = \boldsymbol{Y}_V^T\mathbf{u}_i^{\mathrm{v}}/\sqrt{\lambda_i^{\mathrm{v}}}$, for $i = 1,\ldots,d$ where $d = \max\{i \in \{1,\ldots,lD\} : \lambda_i^{\mathrm{v}} > 0\}$.

GROUPING AND DIAGONAL AVERAGING. Not all terms in the decomposition of $\boldsymbol{Y}_V$ contain information about the signal, hence we retain a subset $I \subset \{1,\ldots,d\}$ so to compute $\boldsymbol{Y}_I = \sum_{i \in I} \boldsymbol{Y}_i$. Averaging over all the elements of the (anti)diagonal $i+j = const$ of $\boldsymbol{Y}_I$ yields an interval Hankel matrix, from where our interval trendline indicators results. We allow for each trendline to be constructed from a different number of components, that is

$$[\widetilde{\mathbf{y}}_1 \cdots \widetilde{\mathbf{y}}_D] = [\mathbb{D}(\boldsymbol{Y}_{I_1}^1) \cdots \mathbb{D}(\boldsymbol{Y}_{I_D}^D)]. \tag{10}$$

where

$$\begin{bmatrix} \boldsymbol{Y}_{I_1}^1 \\ \vdots \\ \boldsymbol{Y}_{I_1}^D \end{bmatrix} = \sum_{i \in I_1} \boldsymbol{Y}_i, \ldots, \begin{bmatrix} \boldsymbol{Y}_{I_D}^1 \\ \vdots \\ \boldsymbol{Y}_{I_D}^D \end{bmatrix} = \sum_{i \in I_D} \boldsymbol{Y}_i.$$

Note that (10) is thus a multivariate version of (5). Next, we assess the finite sample performance of the proposed methods in a simulation study.

3 | SIMULATION STUDY

3.1 | Data-generating scenarios and preliminary experiments

In this section we assess the performance of the proposed methods via a simulation study. A Monte Carlo study will be presented in Section 3.2; for now we concentrate on discussing the data generating processes from which the data are simulated, and on illustrating a fit from the proposed methods on a single run experiment with $n = 100$. We consider the following interval-valued data generating processes:

$$\begin{aligned} x_t &= [\mu_t^x + \varepsilon_t^x, \mu_t^x + 2 + \varepsilon_t^x], \\ y_t &= [\mu_t^y + \varepsilon_t^y, 2(\mu_t^y + 1) + \varepsilon_t^y], \end{aligned} \tag{11}$$

where $t \in T = \{2i\pi/n\}_{i=1}^{n}$, $\mu_t^x = 8 + t + \sin(\pi t)$, $\mu_t^y = \sqrt{t} + \cos(\pi t/2)$. Here, ε_t^x and ε_t^y are zero mean normally distributed errors with covariance function given by

$$\begin{aligned} \boldsymbol{\Sigma}(s,t) &\equiv \begin{bmatrix} \mathrm{Cov}(\varepsilon_t^x, \varepsilon_s^x) & \mathrm{Cov}(\varepsilon_t^x, \varepsilon_s^y) \\ \mathrm{Cov}(\varepsilon_t^y, \varepsilon_s^x) & \mathrm{Cov}(\varepsilon_t^y, \varepsilon_s^y) \end{bmatrix} \\ &= \delta(t-s)\begin{bmatrix} \sigma^2 & \rho \\ \rho & \sigma^2 \end{bmatrix}, \end{aligned}$$

where $\delta(0) = 1$ and $\delta(t-s) = 0$ for $t \neq s$. In Scenario A we set $\rho = 0$ and $\sigma^2 = 1$, thus $\{x_t\}$ and $\{y_t\}$ are independent interval-valued processes, and in Scenario B we set $\rho = 1/2$ and $\sigma^2 = 1$, leading to dependent interval-valued processes.

The processes $\{x_t\}$ and $\{y_t\}$ will be used to illustrate all versions of the proposed method, namely: Interval-Valued Singular Spectrum Analysis (IVSSA; Section 2.2) as well as its multivariate extensions (h-MIVSSA and v-MIVSSA; Section 2.4). In Figure 1 we present one instance of an interval trendline estimate yield using our methods corresponding to a one shot experiment for Scenarios A–B. As it can be seen from Figure 1 our methods closely track the true interval means of both processes for Scenarios A–B; of course such finding should be regarded as tentative, as this is the outcome of a single run experiment, but the same inquiry will be revisited in Section 3.2 through the lenses of a Monte Carlo simulation study.

Some comments on the selection of (m, l) over our numerical experiments are in order. As anticipated in Section 2, to learn about m we adapt the periodogram-based approach in de Carvalho and Martos (2020) to an interval-valued setting; details on the latter are available from Appendix A.2. Keeping in mind theoretical results on the window length achieving maximum rank (Hassani & Mahmoudvand, 2013, Section 4.1), we consider $l = \lceil (n+1)/(D+1) \rceil$ in the case of vertical stacking and $l = \lceil D(n+1)/(D+1) \rceil$ in the case of horizontal stacking, where $\lceil \cdot \rceil$ is the ceiling function.

As mentioned by a reviewer, the data generating scenarios from (11) imply a deterministic relationship between the lower and upper limits of the interval-valued process. Thus, in the supplementary materials we examine the performance of the proposed methods in the case where that relationship is random; performance is similar to that presented here. In addition, since in some applied

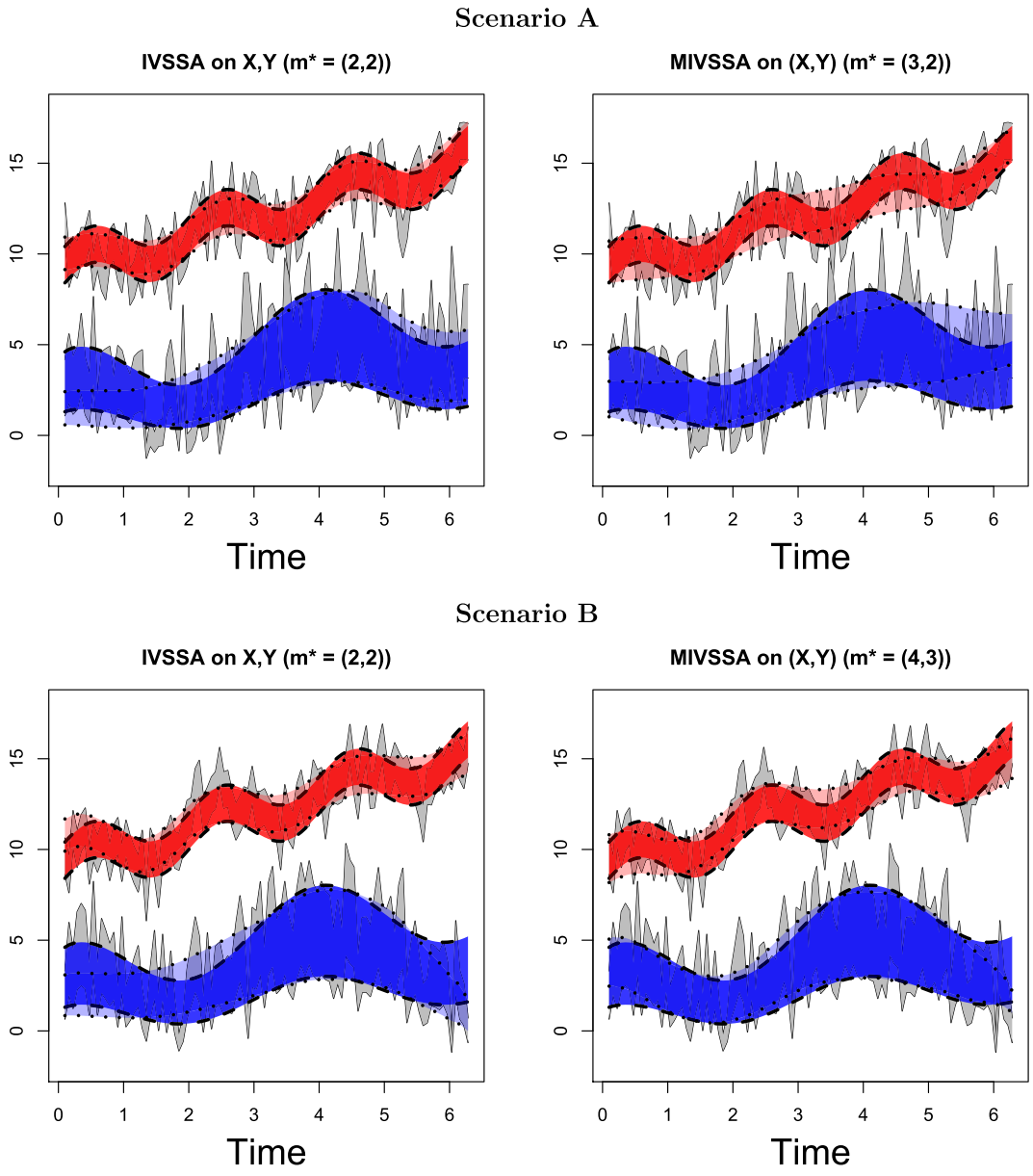

FIGURE 1 One shot experiments for Scenarios A and B. The solid gray areas corresponds to raw interval data for $\{x_t\}$ and $\{y_t\}$. Conditional means $[\mu_t^X, \mu_t^X + 2]$ and $[\mu_t^Y, 2(\mu_t^Y + 1)]$ and trendline estimators are depicted in solid and transparent red and blue, respectively

settings the interest is on modeling intervals defined by the minima and maxima of a process that is sampled on a higher frequency, we present further numerical results illustrating the proposed method on that setting.

3.2 | Monte Carlo simulation study

We now report the main findings of a Monte Carlo simulation experiment based on the data generating processes described in Section 3.1; here, we consider $S = 1000$ Monte Carlo simulations. For Scenarios A and B we consider the sample sizes $n = 100, 250, 1000$, and allow for the number of ERC to be retained to be $m = 1, \ldots, 6$. Since the processes under study are set-valued, we assess performance using the Hausdorff distance between the mean set-valued process ($\mathrm{E}(x_t)$) and the estimated interval trendline ($\widetilde{x}_t$), that is

$$D_{\mathrm{H}}(\mathrm{E}(x_t), \widetilde{x}_t) = \max\{|\mathrm{E}(a_t) - \widetilde{a}_t|, |\mathrm{E}(b_t) - \widetilde{b}_t|\},$$

with $\{x_t \equiv [a_t, b_t]\}_t$ and $\{\widetilde{x}_t \equiv [\widetilde{a}_t, \widetilde{b}_t]\}_t$. More specifically, we compute the average Hausdorff residuals (HR) here defined as

$$\mathrm{HR} = \frac{1}{n}\sum_{t} D_{\mathrm{H}}(\mathrm{E}(x_t), \widetilde{x}_t).$$

Figure 2 depicts side-by-side boxplots of HRs for Scenarios A and B; in the Supplementary Material, we also report the Monte Carlo mean HR for all the sample sizes and ERCs in this study.

As it can be seen from Figure 2, as the sample size increases the HR tends to decrease, regardless of the number of ERC. This thus indicates a better performance, from an Hausdorff residual perspective, of the proposed methods as the number of observations increases. We now switch gears and examine the periodogram-based criterion used for learning about the number of ERCs (see Appendix A.2). Figure 3 displays the distribution of the number of ERCs selected with our automatic criterion on the Monte Carlo experiment. The joint analysis of Figures 2 and 3 suggests that our periodogram-based approach does a sensible job at learning about the number of ERCs as it tends to select a number of components

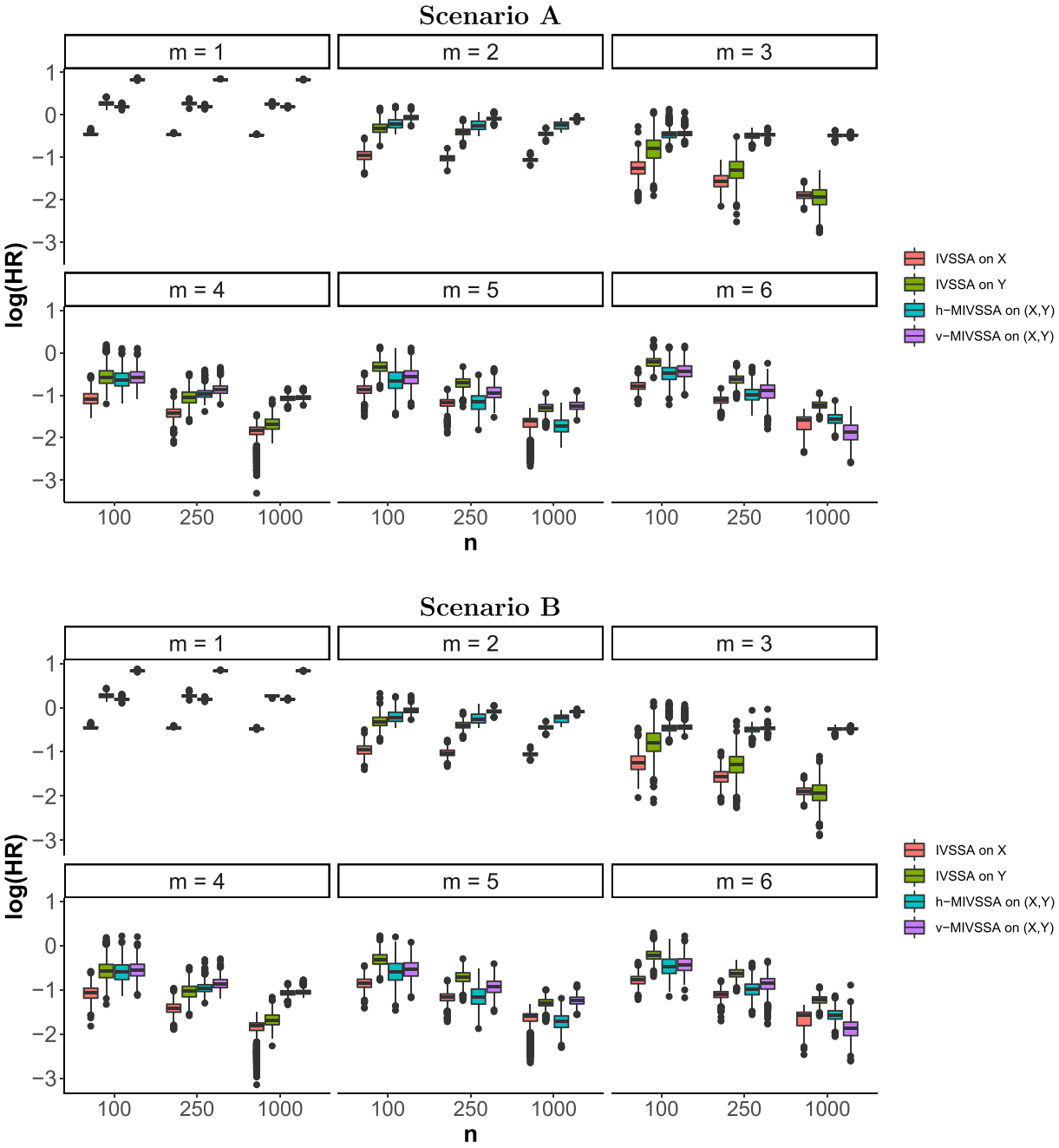

FIGURE 2 Monte Carlo simulation study Hausdorff residuals (HR). Boxplots of HR for Scenarios A and B over different values of m and n

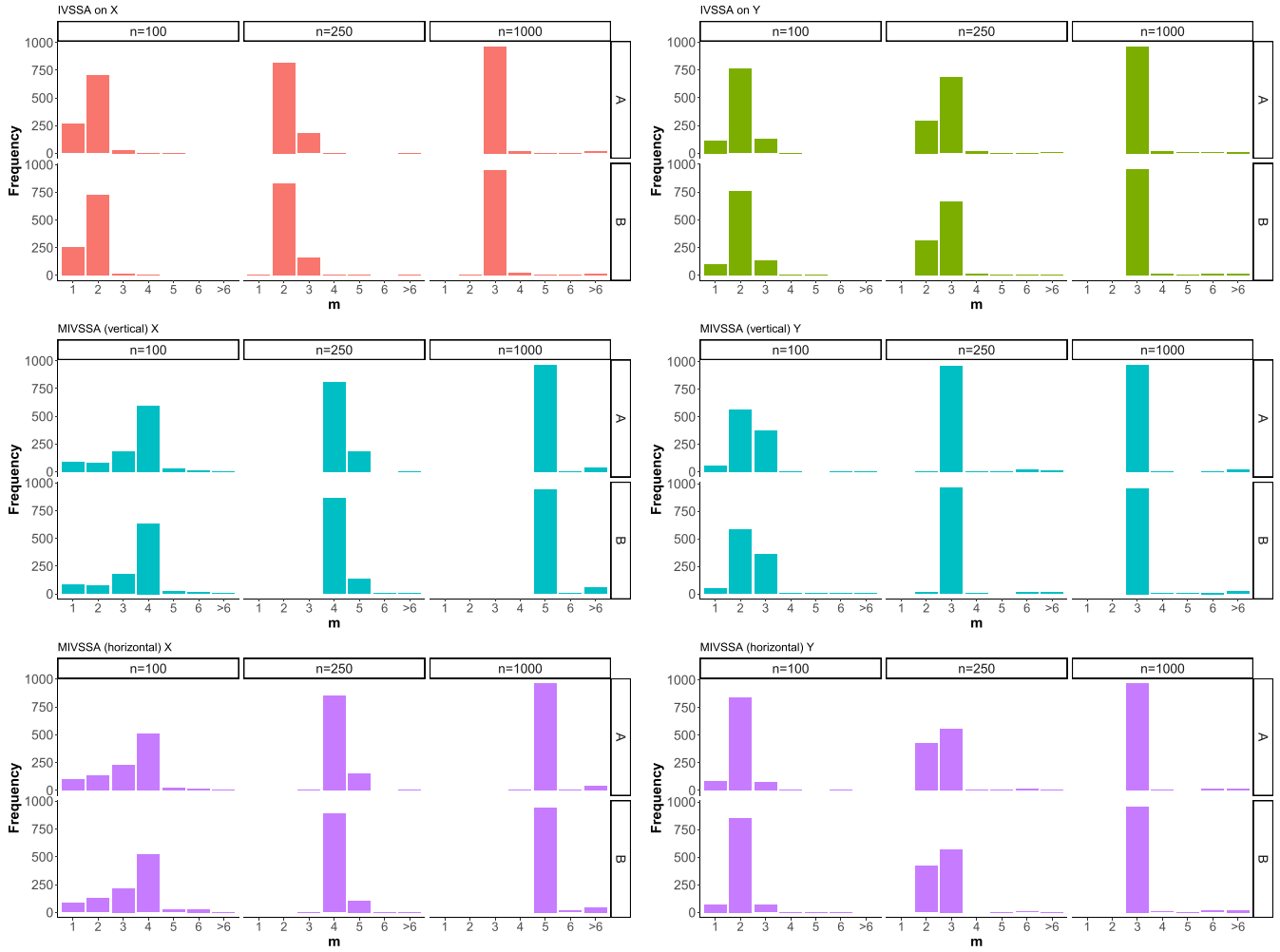

FIGURE 3 Number of ERC selected according to the periodogram-based of de Carvalho and Martos (2020) adapted for the context of interval-valued time series

(see Figure 3) that closely follows the number of components achieving the lowest HR in the Monte Carlo simulation study (see Figure 2).

4 | INTERVAL TRENDLINES FOR ARGENTINA STOCK MARKET

4.1 | Data description and motivation for the analysis

We now apply our methods so to learn about interval trendlines for the MERVAL index—the principal index of Argentina stock market. In Figure 5 we depict the raw interval data series from Yahoo Finance corresponding to weekly minimum and maximum values of MERVAL ranging from January 1, 2016 to September 30, 2020. During the period of interest the economy of Argentina was impacted by several episodes of financial interest, and we will aim to examine how the trendlines of MERVAL reacted to those. Examples include (a) currency crisis started in (Q1 2018) (Sturzenegger, 2019, Section 4.1), that involved the IMF (International Monetary Fund) intervention with a three-year lending program of USD 50bn (Sturzenegger, 2019, Section 4.2) approved in the end of (Q2 2018) (IMF Press release NO.18/245) and later increased by USD 7bn (IMF Press release NO.18/362) on (Q3 2018); (b) the so-called PASO (primary elections in Argentina) whose surprising outcome (Q3 2019) has led to the imposition of foreign exchange controls (BBC, press note) and also a virtual sovereign debt default (Q4 2019) (DNU 49/2019); (c) a sovereign debt restructuring process between (Q1 2020) and (Q3 2020) (Bloomberg, press note); and (d) COVID-19 lockdown over (Q1 2020–onwards) (Bloomberg, press note). The next section will employ the proposed methods and will assess how have the MERVAL trendlines reacted when those episodes took place.

4.2 | Modeling, nowcasting, and forecasting interval trendlines

Figure 4 depicts the first 12 ERCs of MERVAL obtained via the proposed decomposition methods. As it can be seen from the latter figure, the first components seem to correspond to movements associated with an interval drift, whereas the last few components seem to represent a cycle or noise; to draw a distinction between what ERCs that actually correspond to a drift, and which ones represent noise, we resort to an interval-valued version of the periodogram-based method of de Carvalho and Martos (2020)—which is discussed in Appendix A.2, and whose performance has been examined in Section 3. In Figure 5 we depict the MERVAL trendline corresponding to the IVSSA version of the proposed methods. To learn about the interval trendlines, we consider $l = \lceil (n+1)/2 \rceil = 125$ (as discussed on p. 10) and to learn about the number of components (m) we resort to our periodogram-based criterion. From a visualization

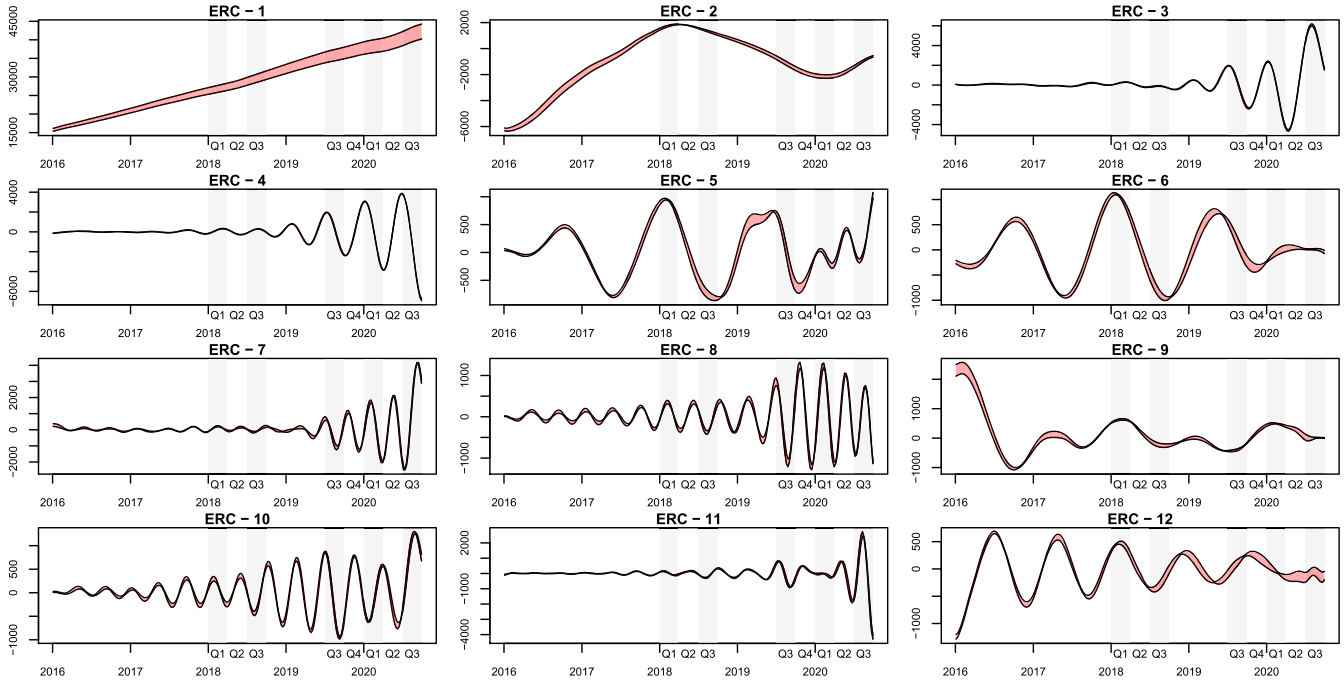

FIGURE 4 First 12 ERCs of MERVA index obtained via IVSSA. The shaded areas represent episodes a–d from Section 4.1

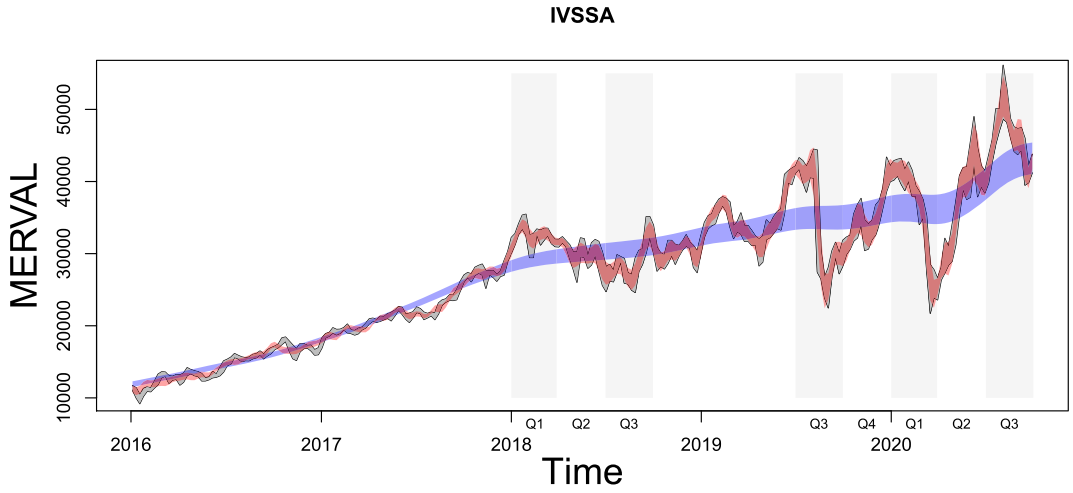

FIGURE 5 "Post-mortem" interval trendlines: MERVAL interval-valued data defined by weekly minimum and maximum values of the index (■), IVSSA interval trendline obtained with periodogram-based approach (■) and with Haussdorff residual-based approach (■). The shaded areas represent episodes a–d from Section 4.1

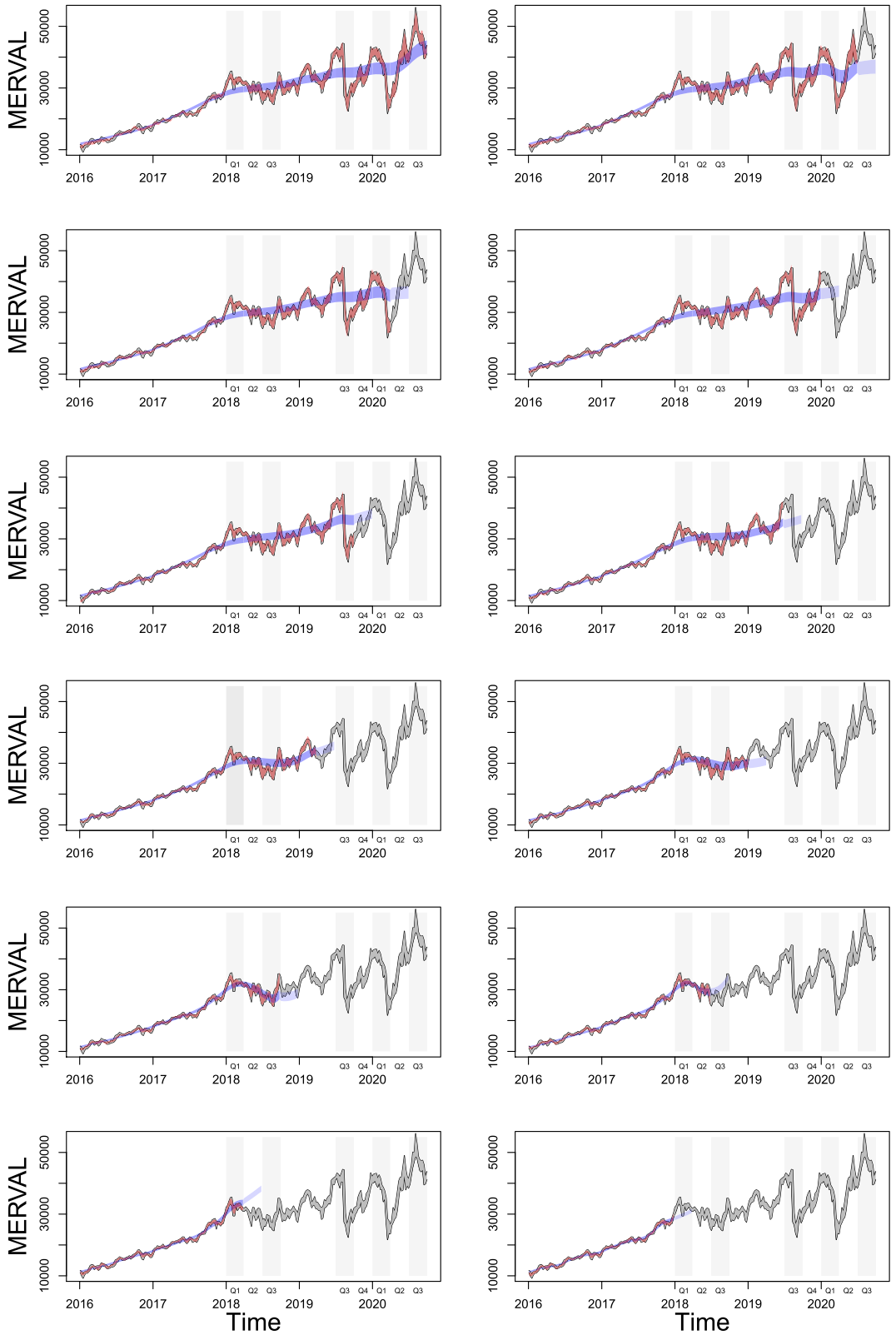

FIGURE 6 Real-time analysis ranging from (Q3 2020) to (Q1 2018) along with MERVAL interval data (■) and IVSSA interval trendline obtained with periodogram-based approach (■) and with Hausdorff residual-based approach (■) along with the corresponding interval forecasts. The gray shaded areas represent episodes a–d from Section 4.1

viewpoint, perhaps a number components smaller than $m^* = 31$ is preferable, despite the overall good performance suggested by Section 3 of our criterion for selecting m. Keeping in mind this, and the fact that for forecasting the latter choice of m may not be the most appropriate—as the interval-valued signal may follow the data too closely—we also use out-of-sample evaluations for selecting the values of l and m. Roughly speaking, this is achieved by minimizing the 1Q (12 weeks) out-of-sample Haussdorff residual corresponding to models fitted over an expanding window, and it yields $(l^\star, m^\star) = (80, 2)$; see Appendix A.3 for details. Regardless of the value of m, as it can be seen from Figure 5, the interval trendlines produced by our method have clear links with the episodes a–d mentioned in Section 4.1.

The interval trendlines depicted in Figure 5 are a post-mortem in the sense that they are based on the entire sample period. To assess how much the trendlines produced by our method would be revised when they are produced in real time, we conduct a real-time analysis. In fields such as economics, the ability of a method to be coherent over real-time—in the sense of not revising estimates once new data arrives—is key, and it has been a subject of wide interest (see, for instance, Orphanides & Van Norden, 2002, and references therein). We conduct a real-time analysis by sequentially removing the most recent quarters of data from the whole data set so to compute interval trendlines at the end of every quarter. The sequence of real-time interval trendlines is depicted in Figure 6. As it can be seen from the latter figure, our method does not revise substantially the produced interval trendlines; that is, the real-time interval trendlines (Figure 6) resemble those obtained from the post-mortem (Figure 5), thus suggesting a sensible real-time performance of our method. Such sturdy real-time performance is in line with what has been found for SSA for time series (de Carvalho et al. 2012; de Carvalho & Martos, 2020)—rather than for interval time series as examined here. The real-time analysis from Figure 6 also presents a sequence of out-of-sample forecasts that were obtained via the methods from Section 2.3. As it can be noticed from the latter figure, the out-of-sample forecasts obtained by the proposed method are reasonably in line with the true targets, thus suggesting a good forecast accuracy of the proposed methods in this real-time exercise.

5 | CLOSING REMARKS

From a methodological outlook a main goal of this article was on extending SSA-based methods to an interval time series context. The proposed extension is tailored for modeling a range of values over time so to learn about interval-valued signals and to yield out-of-sample forecasts of the said range of values. To our knowledge this paper pioneers the development of statistical decomposition methods for set-valued stochastic processes, and the proposed method can be used for decomposing an interval-valued time series into a string of components that can be interpreted as an interval-valued signal, cycle, or noise. The proposed method coincides with standard SSA when the data of interest are standard time series—rather than interval time series—and the multivariate extension of our method allows for combining both time series and interval time series. Naively, one could think of applying standard multivariate SSA to interval-valued data by treating each of the limits of an interval as a vector as an alternative to the methodology proposed herein; yet such naive multivariate SSA-based approach would not have in mind the interval-valued nature of the data.

Some remarks on future research are in order. It would seem natural to use the spectral features learned via symbolic SSA so to cluster or classify interval time series, or even to cluster according to these features; given the recent applied relevance of clustering interval time series (Maharaj et al. 2019), we believe this could be a natural methodological target for future investigation. Another potential follow-up within the remit of this paper is the development of decomposition methods for functional set-valued data, that would aim to extend the methods proposed here to a continuous time setting. While Functional Data Analysis (Ferraty & Vieu, 2006; Horváth & Kokoszka, 2012; Ramsay & Silverman, 2002; Ramsay, 2006)—that is, the analysis of data in the form of a continuous time stochastic process—is a fast-evolving field, to our knowledge no developments have been made on the statistical analysis of a set-valued version of functional data (i.e., data in the form of a continuous time set-valued stochastic process), nor have been devised decomposition methods for set-valued functional. We leave such open problems for future analysis.

ACKNOWLEDGEMENTS

We thank the Editor, the Associate Editor, and two anonymous reviewers for insighful and constructive comments on a previous version of the paper. The research was partially funded by the project INTERSTATA (**Inter**disciplinary **Stat**istics in **A**ction) from the International Research and Partnership Fund (Developing Countries), and by FCT (Fundação para a Ciência e a Tecnologia, Portugal), through the projects PTDC/MAT-STA/28649/2017 and UID/MAT/00006/2019.

DATA AVAILABILITY STATEMENT

The data analyzed in this article are publicly available at Yahoo Finance (https://finance.yahoo.com/).

ORCID

Miguel de Carvalho https://orcid.org/0000-0003-3248-6984

REFERENCES

Billard, L. (2006). Symbolic data analysis: What is it? *Compstat 2006-Proceedings in Computational Statistics*: Springer, pp. 261–269.

Billard, L. (2007). Dependencies and variation components of symbolic interval-valued data, *Selected Contributions in Data Analysis and Classification*: Springer, pp. 3–12.

Billard, L., & Diday, E. (2003). From the statistics of data to the statistics of knowledge: symbolic data analysis. *Journal of the American Statistical Association*, *98*(462), 470–487.

Billard, L., & Le-Rademacher, J. (2012). Principal component analysis for interval data. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(6), 535–540.

Brockwell, P. J., & Davis, R. A. (2002). *Time Series: Theory and Methods*. New York: Springer.

de Carvalho, M., & Martos, G. (2018). ASSA: Applied singular spectrum analysis. R package version 1.0.

de Carvalho, M., & Martos, G. (2020). Brexit: Tracking and disentangling the sentiment towards leaving the EU. *International Journal of Forecasting*, *36*, 1128–1137.

de Carvalho, M., Rodrigues, P. C., & Rua, A. (2012). Tracking the US business cycle with a singular spectrum analysis. *Economics Letters*, *114*(1), 32–35.

Ferraty, F., & Vieu, P. (2006). *Nonparametric functional data analysis: Theory and practice*. New York: Springer.

Golyandina, N., Nekrutkin, V., & Zhigljavsky, A. A. (2001). *Analysis of time series structure: SSA and related techniques*. Boca Raton, FL: Chapman and Hall/CRC.

Golyandina, N., & Zhigljavsky, A. (2013). *Singular Spectrum Analysis for Time Series*. New York: Springer.

González-Rivera, G., & Arroyo, J. (2012). Time series modeling of histogram-valued data: The daily histogram time series of s&p500 intradaily returns. *International Journal of Forecasting*, *28*(1), 20–33.

González-Rivera, G., & Lin, W. (2013). Constrained regression for interval-valued data. *Journal of Business & Economic Statistics*, *31*(4), 473–490.

Hassani, H., & Mahmoudvand, R. (2013). Multivariate singular spectrum analysis: A general view and new vector forecasting approach. *International Journal of Energy and Statistics*, *1*(1), 55–83.

Horváth, L., & Kokoszka, P. (2012). *Inference for functional data with applications*, Vol. 200. New York: Springer.

Khan, M. A. R., & Poskitt, D. S. (2017). Forecasting stochastic processes using singular spectrum analysis: Aspects of the theory and application. *International Journal of Forecasting*, *33*(1), 199–213.

Kisielewicz, M. (2013). Set-valued stochastic processes, *Stochastic differential inclusions and applications* (pp. 67–102). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-6756-4_2

Le-Rademacher, J., & Billard, L. (2012). Symbolic covariance principal component analysis and visualization for interval-valued data. *Journal of Computational and Graphical Statistics*, *21*(2), 413–432.

Lin, W., & González-Rivera, G. (2016). Interval-valued time series models: Estimation based on order statistics exploring the agriculture marketing service data. *Computational Statistics & Data Analysis*, *100*, 694–711.

Maharaj, E. A., Teles, P., & Brito, P. (2019). Clustering of interval time series. *Statistics and Computing*, *29*(5), 1011–1034.

Mahmoudvand, R., & Rodrigues, P. C. (2018). A new parsimonious recurrent forecasting model in singular spectrum analysis. *Journal of Forecasting*, *37*(2), 191–200.

Orphanides, A., & Van Norden, S. (2002). The unreliability of output-gap estimates in real time. *Review of Economics and Statistics*, *84*(4), 569–583.

R Development Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ramsay, J. O. (2006). *Functional data analysis*. New York: Wiley.

Ramsay, J. O., & Silverman, B. W. (2002). *Applied Functional Data Analysis: Methods and Case Studies*, Vol. 77: Citeseer.

Rodrigues, P. C., & de Carvalho, M. (2013). Spectral modeling of time series with missing data. *Applied Mathematical Modelling*, *37*(7), 4676–4684.

Rodrigues, P. M. M., & Salish, N. (2015). Modeling and forecasting interval time series with threshold models. *Advances in Data Analysis and Classification*, *9*(1), 41–57.

Sturzenegger, F. (2019). Macri's macro: The elusive road to stability and growth. *Brookings Papers on Economic Activity*, *2019*(2), 339–436.

Sun, Y., Han, A., Hong, Y., & Wang, S. (2018). Threshold autoregressive models for interval-valued time series data. *Journal of Econometrics*, *206*(2), 414–446.

Wang, X., Zhang, Z., & Li, S. (2016). Set-valued and interval-valued stationary time series. *Journal of Multivariate Analysis*, *145*, 208–223.

AUTHOR BIOGRAPHIES

Miguel de Carvalho is a Reader in Statistics at the University of Edinburgh. He is currently the Director of the Centre for Statistics, University of Edinburgh, as well as President of the Portuguese Statistical Society (SPE). He is an Associate Editor for the *Journal of the American Statistical Association* (Case Studies), *Annals of Applied Statistics*, and *Computational Statistics and Data Analysis*.

Gabriel Martos is a statistician with a variety of interdisciplinary research interests, including Applied Statistics, Computational Statistics, and Machine Learning. He received his PhD from University Carlos

III de Madrid and was a Post-Doctoral Fellow at Pontificia Universidad Catolica de Chile as well as a Research Associate at the Universidad de Buenos Aires. He is currently an Assistant Professor at the Universidad Torcuato Di Tella, Argentina.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: de Carvalho, M., & Martos, G. (2022). Modeling interval trendlines: Symbolic singular spectrum analysis for interval time series. *Journal of Forecasting*, *41*(1), 167–180. https://doi.org/10.1002/for.2801

APPENDIX

A.1 | Technical Details

Proof of Proposition A1. Our strategy is similar to that of Golyandina et al. (2001, Proposition 6.3). Since $\mathbf{H}=\{[\alpha_{i,j},\beta_{i,j}]\}$ is an Hankel matrix of ordered pairs, it follows that it is constant across anti-diagonals, that is, $[\alpha_{i,j},\beta_{i,j}]=[g_{1,s},g_{2,s}]$, for $i+j=s$ and some pair of numbers $(g_{1,s},g_{2,s})$. Thus, it follows that

$$\begin{aligned}\|\mathbf{Y}-\mathbf{H}\|_{\mathrm{C}}^2 &= \sum_{i,j}\{(a_{i,j}-\alpha_{i,j})^2+(b_{i,j}-\beta_{i,j})^2\}\\ &= \sum_{i,j}(a_{i,j}-\alpha_{i,j})^2+\sum_{i,j}(b_{i,j}-\beta_{i,j})^2\\ &= \sum_{s=2}^{l+k}\sum_{i+j=s}(a_{i,j}-g_{1,s})^2+\sum_{s=2}^{l+k}\sum_{i+j=s}(b_{i,j}-g_{2,s})^2,\end{aligned}$$

which is minimized for $g_{1,s}=n_s{}^{-1}\sum_{i+j=s}a_{i,j}$ and $g_{2,s}=n_s{}^{-1}\sum_{i+j=s}b_{i,j}$.

A.2 | Automatic criterion to choose the number of ERC

In this section we extend the automatic criterion to choose the number of ERC of de Carvalho and Martos (2020) to an interval-valued time series context.

A.2.1 | Groundwork on spectral analysis for interval-value time series

Prior to introducing our criterion we need to lay the groundwork. The spectral density of an interval-valued time series $\{y_t\equiv[a_t,b_t]\}_{t=1}^n$ is here defined as

$$f(\omega)=\frac{1}{2\pi}\sum_{h=-\infty}^{\infty}\gamma(h)e^{-ih\omega},\ \omega\in(-\pi,\pi], \tag{A1}$$

where $\gamma(h)=\mathrm{cov}(y_t,y_{t-h})$ is the autocovariance function at lag $h\in\mathbb{N}$. The definition in (A1) is motivated from the well-known relation between the autocovariance function and the spectral density (Brockwell & Davis, 2002, Proposition 10.1.2). Further, we consider the interval residuals $\mathbf{e}=\phi(\mathbf{y}-\tilde{\mathbf{y}})=\{[e_1^L,e_1^U],\ldots,[e_n^L,e_n^U]\}$ and the corresponding spectral density plug-in estimator, which we will refer to as the periodogram for interval-valued time series, computed as follows:

$$\hat{f}(\omega_j)=\frac{1}{2\pi}\sum_{|h|\le n-1}\hat{\gamma}_e(h)e^{-ih\omega_j}, \tag{A2}$$

where $\omega_j=2\pi j/n$ are the so-called Fourier frequencies, for $j=1,\ldots,J=\lfloor(n-1)/2\rfloor$, with $\lfloor\cdot\rfloor$ denoting the floor function, and $\hat{\gamma}_e(h)$ is the empirical autocovariance function of interval residuals which readily follows by adapting Billard and Le-Rademacher (2012, Equation 3):

$$\hat{\gamma}_e(h)=\frac{1}{6n}\sum_{t=1}^{n-h}\{2e_t^Le_{t+h}^L+e_t^Le_{t+h}^U+e_t^Ue_{t+h}^L+2e_t^Ue_{t+h}^U\},$$

where $h=0,1,\ldots,n-1$. Next, we show how (A2) can be used for learning about the number of ERC.

A.2.2 | Periodogram-based criterion for learning about the number of ERC

Our periodogram-based criterion for learning about the number of ERC is tantamount to that of de Carvalho and Martos (2020), but based on the periodogram for interval-valued time series in (A2). Below, $\omega_j=2\pi j/n$ are Fourier frequencies, for $j=1,\ldots,J=\lfloor n/2\rfloor$, with $\lfloor\cdot\rfloor$ denoting the floor function. Formally, the method is as follows:

Targeted grouping based on the Kolmogorov–Smirnov statistic

Set $i = 1$ and execute the steps:

1. Compute the interval residual vector $\mathbf{e} = \phi(\mathbf{y} - \widetilde{\mathbf{y}})$, yield from the interval trendline based on $I = \{1, \ldots, i\}$.
2. Compute the cumulative periodogram of $\mathbf{e}$, and test the null hypothesis of white noise using the Kolmogorov–Smirnov test based on the statistic

$$\sqrt{J}\max\{|C(\omega_j) - j/J|\}_{j=1}^{J},$$

$$C(\omega_j) = \frac{\sum_{i=1}^{j}\mathbb{I}(\omega_i)}{\sum_{i=1}^{J}\mathbb{I}(\omega_i)}. \quad \text{(A3)}$$

If the null is rejected then increment i and repeat Steps 1 and 2. Otherwise stop.

In words, the method sequentially adds components until there is evidence from the cumulative periodogram of the interval residuals suggesting that the interval residuals constitutes white noise.

A.3 | Hausdorff residual-based criterion for learning about window length and number of ERC for forecasting

Let $\hat{y}_t \equiv \hat{y}_t(\mathbb{D}_w, l, m)$ be the forecast obtained with IVSSA using the method from Section 2.3 with the training data set $\mathbb{D}_w = (y_1, \ldots, y_w)$, for $1 < w \leq n$, and with parameters (l, m). To learn about the smoothing parameters (l, m) in the context of forecasting p-steps ahead of period w_0, we solve the minimization problem

$$(l^\star, m^\star) = \underset{(l,m)}{\operatorname{argmin}} \sum_{w=w_0}^{n-p} \sum_{t=w+1}^{w+p} D_{\mathrm{H}}(y_t, \hat{y}_t),$$

where D_{H} is the Hausdorff distance. (For example, for the forecasts depicted in Figure 6, we consider $w_0 = 104$—the weeks corresponding to years 2016 and 2017 in MERVAL data—and $p = 12$, that is, we choose (l, m) so to maximize 1Q forecasting accuracy.)