Statistics and Its Interface Volume 12 (2019) 527–536

Comparison between Basic and Toeplitz SSA applied to non-stationary time-series

Michel C. R. Leles*, Mariana G. Moreira, Adriano S. Vale-Cardoso, Cairo L. Nascimento Júnior, Elton F. Sbruzzi, and Homero N. Guimarães

A comparison between two approaches of Singular Spectrum Analysis (SSA) methodology is presented: the Basic and the Toeplitz SSA. These approaches differ in assumptions about some SSA properties. Toeplitz SSA assumes time-series stationarity, which means that the process needs to be mean-reverting. However, such assumption is not a necessary condition for the Basic SSA. Therefore, the applicability of the Toeplitz SSA to non-stationary signals is still an under discussion subject. In this paper both approaches are applied to this kind of signal. Similarities and differences between these techniques are addressed. The frequency domain interpretation of eigenvectors as well as forecasting performance are presented for both methodologies. Several computer simulations involving both synthetic and actual data time-series, using the same parameters, were executed in order to compare the studied SSA approaches. The obtained results suggest the Toeplitz SSA should not be used for non-stationary time-series before removing their trend component.

Keywords and phrases: Singular Spectrum Analysis (SSA), Non-Stationary Signals, Basic SSA, Toeplitz SSA.

1. INTRODUCTION

Singular Spectrum Analysis (SSA) is a nonparametric approach and, as so, there are no major restrictions on the data for its use. In this manner, SSA-based methods can be applied for decomposing a time-series into components of trends, oscillations and noise [1], each having meaningful interpretation. These capabilities have made it widely used in time-series analysis [2]. It has been applied to several knowledge fields, including earth sciences [3], biomedical [4] and financial [5] time-series.

Despite the fact that several approaches for the SSA have been presented in the literature [6, 7, 8, 9, 10, 11, 12, 13], two specific approaches of the SSA method are discussed in this paper: the "Basic" SSA [14, 15, 16]; and the "Toeplitz" SSA [3, 17, 18]. These two versions can also be called BK and VG, respectively, as they were first named by their original papers [19, 20].

The main differences between the Basic and Toeplitz approaches concern assumptions made in the study of the SSA properties [1]. For the Toeplitz, the stationarity of the analyzed time-series is assumed. Moreover, the analyzed series is represented according a "*signal plus noise*" model. Nevertheless, in the Basic approach, the main reasoning is oriented to the "*separability*" issue of time-series components, and no assumptions about its stationarity or its model are made. Golyandina [1] stated that using Toeplitz SSA in non-stationary time-series may yield to wrong results. Ghil et al. [17], on the other hand, argued that the difference between the Basic and Toeplitz approaches of SSA – when dealing with non-stationary time-series – is marginal. In this case, however, the author refers to climatic time-series, in which the adoption of a pre-processing stage for trend removing is a common practice for data analysis. Therefore, the applicability of the Toeplitz SSA to non-stationary signals is still an under discussion subject and it is the main focus of this paper. An in-depth investigation concerning trend estimation in financial time-series comparing these two methodologies were carried out in Leles et al. [21].

This paper addresses theoretical and application aspects of these two SSA approaches focused on non-stationary signals with any kind of pre-processing step before time-series analysis. Several computer simulations were conducted concerning synthetic and real data time-series. The results show that the dominant structure of non-stationary signals were lost when Toeplitz SSA was applied, and consequently, its usage is discouraged for such signals.

2. METHODOLOGY

A definition of stationarity is revised and the two concerned approaches of the SSA are briefly exposed: Basic and Toeplitz. The frequency domain interpretation of SSA eigenvectors is presented, followed by the forecasting algorithm.

2.1 Stationary signals

An infinite time-series $\mathbf{x}_\infty = x_1, x_2, \cdots, x_N, \cdots$, is stationary if, for all non-negative values of k and m, the limit

[*] Corresponding author.

is satisfied [16, Section 1.4.1]:

$$(1) \qquad \lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} x_{j+k}\, x_{j+m} = R(k-m)$$

where the function $R(.)$ represents the covariance function. Additionally, for any value of k, it is assumed that

$$(2) \qquad \lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} x_{j+k} = 0.$$

This means the time-series has zero-mean.

Stationarity can also be defined considering a finite time-series. A time-series $\mathbf{x_k}, \mathbf{k} \in \mathbb{Z}$, where $\mathbb{Z}$ is an integer set, is stationary if: its variation is finite, its first moment is constant and its second moment only depends on $(k-m)$ and not on k nor m [22].

2.2 Basic SSA methodology

A brief description of the Basic SSA methodology is presented, based on Golyandina and Zhigljavsky [14], Golyandina et al. [16].

2.2.1 Embedding

Let the time-series $\mathbf{x} = (x_0, x_1, \cdots, x_n, \cdots, x_{N-1})^T$, of length N, represent the analyzed signal. The mapping of this signal into a matrix $\mathbf{A}$, of dimension $L \times K$, assuming $L \leq K$, is called *immersion*, and can be defined as:

$$(3) \qquad \mathbf{A} = \begin{bmatrix} x_0 & x_1 & \cdots & x_{K-1} \\ x_1 & x_2 & \cdots & x_K \\ \vdots & \vdots & & \vdots \\ x_{L-1} & x_L & \cdots & x_{N-1} \end{bmatrix}.$$

L is the window length, or embedding dimension, and $K = N - L + 1$. $\mathbf{A}$ in equation 3 is called the trajectory matrix.

2.2.2 Singular value decomposition

The Singular Value Decomposition (SVD) of the trajectory matrix yields to:

$$(4) \qquad \mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \sum_{r=1}^{R} \sigma_r \mathbf{u}_r \mathbf{v}_r^T,$$

where $R = rank(\mathbf{A}) \leq L$. $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices. $\mathbf{\Sigma}$ is a diagonal matrix, whose diagonal elements $\{\sigma_r\}$ are the singular values of $\mathbf{A}$. The main components are obtained by:

$$(5) \qquad \mathbf{w}_r = \sigma_r \mathbf{v}_r = \mathbf{A}^T \mathbf{u}_r,$$

where $\mathbf{w}_r$ is a $K \times 1$ vector. The r-th elementary matrix, $\mathbf{A}_r$, an unitary rank $L \times K$ matrix, can be written as

$$(6) \qquad \mathbf{A}_r = \sigma_r \mathbf{u}_r \mathbf{v}_r^T = \mathbf{u}_r \mathbf{w}_r^T.$$

Equation 4 can be then rewritten as

$$(7) \qquad \mathbf{A} = \sum_{r=1}^{R} \mathbf{u}_r \mathbf{w}_r^T.$$

Additionally, $\sigma_r^2 = \|\mathbf{A}_r\|^2$ and $\|\mathbf{A}\|^2 = \sum_{r=1}^{R} \sigma_r^2$, where $\|.\|^2$ represents the Frobenius norm. A coefficient $\mathcal{C}_r$, defined as

$$(8) \qquad \mathcal{C}_r = \sigma_r^2 / \sum_{m=1}^{R} \sigma_m^2,$$

is called the contribution of the elementary matrix $\mathbf{A}_r$ to the trajectory matrix $\mathbf{A}$.

2.2.3 Grouping

The grouping step is the procedure of arranging the R elementary matrices into M disjoint subsets I_m. For a set of elementary matrices $\{\mathbf{A}_r \mid r \in I_m\}$, the resulting matrix from this grouping is:

$$\mathbf{A}_{I_m} = \sum_{r \in I_m} \mathbf{A}_r = \mathbf{A}_{\{.\}},$$

where $\{.\}$ designates the indexes of the I_m set. Each group is intended to represent an additive component of the original signal, such as a trend, an oscillatory component or noise.

2.2.4 Diagonal averaging

The purpose of this step is to recover a time-series, $\widetilde{\mathbf{x}}_r$, of length N, from an elementary matrix $\mathbf{A}_r$. The diagonal averaging in N anti-diagonals of $\mathbf{A}_r$ can be computed according to [18]:

$$(9) \qquad \widetilde{\mathbf{x}}_n^{(r)} = \begin{cases} \frac{1}{n+1} \sum_{i=0}^{n} u_i^{(r)} w_{n-i}^{(r)} & \text{for } 0 \leq n < L-1, \\ \frac{1}{L} \sum_{i=0}^{L-1} u_i^{(r)} w_{n-i}^{(r)} & \text{for } L-1 \leq n < K, \\ \frac{1}{N-n} \sum_{i=n-K+1}^{N-K} u_i^{(r)} w_{n-i}^{(r)} & \text{for } K \leq n < N. \end{cases}$$

This procedure can be easily extended to any matrix resulting from the grouping process [16, Section 1.2.4].

2.3 Toeplitz SSA methodology

It should be noticed that in this methodology, a mean-corrected or zero-centered data is used. In contrast, such correction is not necessary for the Basic SSA.

In the Toeplitz approach, the covariance matrix is estimated as a *Toeplitz* matrix, $\widetilde{\mathbf{C}}$, whose elements $\widetilde{c}_{ij}$, $1 \leq i, j \leq L$, can be defined as [18]:

$$(10) \qquad \widetilde{c}_{ij} = \frac{1}{N - |i-j|} \sum_{m=1}^{N-|i-j|} x_m\, x_{m+|i-j|}.$$

Although there are several ways of calculating the covariance matrix estimate, in this paper Equation (10) was used, which can be considered the standard approach [23, Section 8.2].

From the Toeplitz covariance matrix $\widetilde{\mathbf{C}}$, the orthonormal eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_L$ are computed (assuming that $\widetilde{\mathbf{C}}$ has full rank), and the trajectory matrix is decomposed as:

$$\mathbf{A} \;=\; \sum_{l=1}^{L} \mathbf{e}_l \mathbf{q}_l^T \tag{11}$$

It can be noticed that the equation (11) is very similar to equation (7). $\tau_l = \|\mathbf{q}_l\|^2$ is often referred to as "squared Toeplitz singular values" and they are usually different from the eigenvalues of $\widetilde{\mathbf{C}}$.

Equation (11) shows that the trajectory matrix $\mathbf{A}$ can be decomposed in the same way as for Basic SSA. However, it is important to emphasize that this approach does not benefit from SVD optimality properties, as the covariance matrix $\widetilde{\mathbf{C}}$ replaces $\mathbf{C} = \mathbf{A}\mathbf{A}^T$ at the decomposition stage. Grouping and diagonal averaging procedures are the same as in the Basic approach.

2.4 SSA in frequency domain

Tome et al. [24] showed that the SSA eigenvectors (closed) frequency-response, for $K \le n \le N - L + 1$, is:

$$H_r(f) = \sum_{l=-(L-1)}^{L-1} \upsilon_l^{(r)} e^{j2\pi l f} = \upsilon_0^{(r)} + \sum_{l=1}^{L-1} 2\upsilon_l^{(r)} cos(2\pi l f). \tag{12}$$

The coefficients $\upsilon_l^{(r)}$ are the result of the convolution sum of two polynomials ($\mathbf{u}_r$ for Basic SSA; $\mathbf{e}_r$, for Toeplitz SSA) with the same coefficients in inverse order of the powers, i.e., $\upsilon_l^{(r)} = \upsilon_{-l}^{(r)},\ l = 1, 2, \cdots, L-1$.

Equation (12) represents the analytical solution for the frequency response of SSA eigenfilters. This filter is known as the middle point filter, which produces a zero-phase response, which implies in a zero delay between input and resulting signal. Leles et al. [2] provided a review of frequency domain interpretation of SSA filters. For n outside the range $[L, N - L + 1]$, there is a wider set of filters, as can be seen in Golyandina et al. [16, Section 2.9].

2.5 SSA forecasting

SSA forecasting algorithm can only be applied to a time-series $\mathbf{y}$ which fits a linear recurrent model:

$$y_{N-n} = \sum_{l=1}^{L-1} a_l\, y_{N-n-l}, \quad 0 \le n \le N - L \tag{13}$$

According to Golyandina et al. [16, Chapter 2], there is a wide variety of systems that satisfy this condition. It can be proved that the coefficient z_L of a vector $\mathbf{z} = (z_1, z_2, \cdots, z_L)^T$ is a linear combination of the $L-1$ former coefficients: $z_L = a_1 z_{L-1} + a_2 z_{L-2} + \cdots + a_{L-1} z_1$. The coefficients $\mathbf{a} = (a_1, a_2, \cdots, a_{L-1})$ can be expressed as:

$$\mathbf{a} = \frac{1}{1-\nu^2} \sum_{r=1}^{R} \upsilon_r \mathbf{u}_r^{\nabla}. \tag{14}$$

where υ_r is the last component of vector $\mathbf{u}_r$, $\nu^2 = \upsilon_1^2 + \upsilon_2^2 + \cdots + \upsilon_r^2 + \cdots + \upsilon_R^2$, and $\mathbf{u}_r^{\nabla} \in \Re^{L-1}$ is the vector consisting of $L-1$ former terms of $\mathbf{u}_r$. The eigenvector adopted depends on the SSA version used. Therefore, $\mathbf{u}_r$ is concerned by the Basic, whereas $\mathbf{e}_r$ is concerned by the Toeplitz approach.

Let b be the number of forecast points. Set $\mathbf{p}_r$, of length $N+b$, as the SSA forecast series, whose elements $p_k^{(r)}$, $k = 1, 2, \cdots, N+b$, are given by [16, Section 2.1]:

$$p_k^{(r)} = \begin{cases} \widetilde{x}_k^{(r)}, & k = 1, \cdots, N \\ \sum_{l=1}^{L-1} a_l p_{k-l}^{(r)}, & k = N+1, \cdots, N+b \end{cases} \tag{15}$$

Equation (15) shows that the first N points of series $\mathbf{p}_r$ were obtained from time-series synthesis, $\widetilde{x}^{(r)}$, and the points $p_{N+1}^{(r)}, \cdots, p_{N+b}^{(r)}$ are the b SSA forecast points.

3. APPLICATION

In this section some computer simulation results are presented in order to compare the Basic and Toeplitz SSA behavior when treating non-stationary time-series. Both approaches are applied to synthetic and experimental datasets. In section 3.1.1, an exponential modulated harmonic time-series is used. An example of real data time-series is presented in section 3.2.

The reconstruction and forecasting performances can be quantified by standard error metrics, such as the Mean Absolute Error (MAE), according to:

$$\mathrm{MAE} = \frac{1}{N} \sum_{n=0}^{N-1} |x_n - \tilde{x}_n|, \tag{16}$$

where x_n is an original time-series' sample at instant n, $\tilde{x}_n$ is the corresponding reconstructed sample and N indicates the series length.

3.1 Synthetic time series

In this section two different analyses are conducted based upon synthetic time-series, as follows.

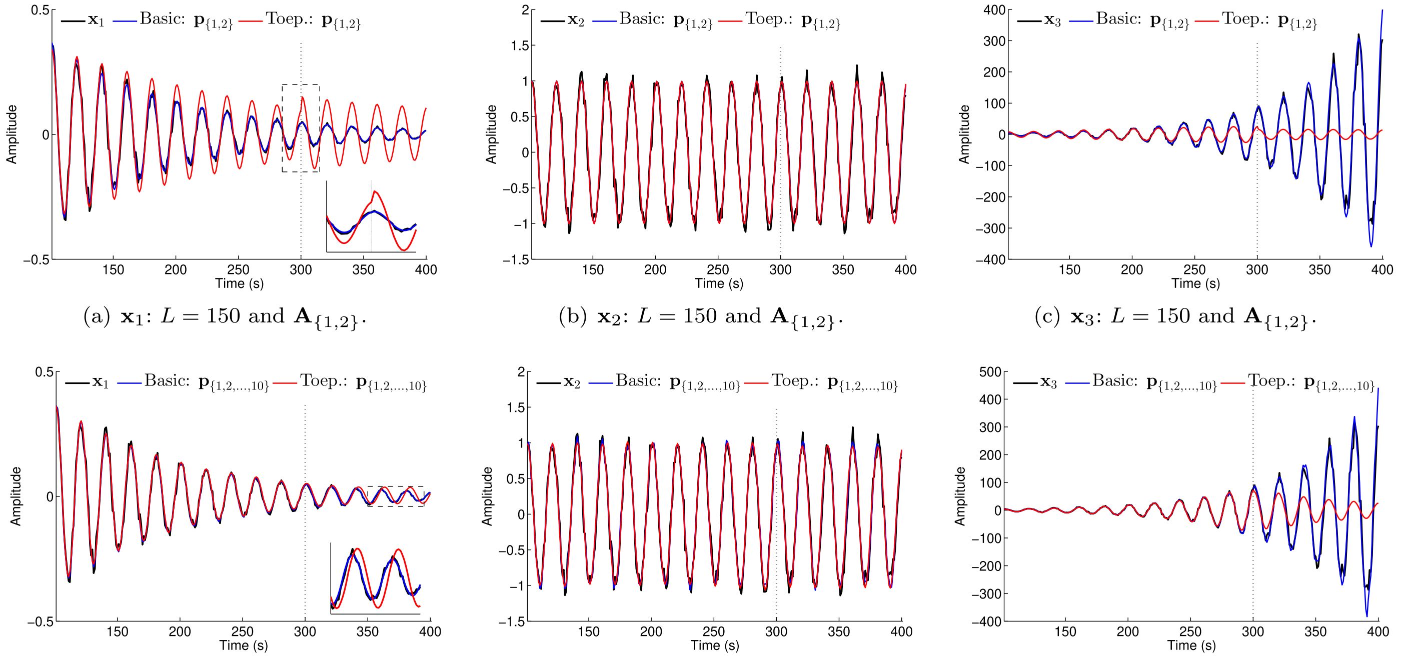
3.1.1 Case 1

Three different time-series are concerned in this section, described by their equations as follows:

$$\mathbf{x}_1 \;=\; 0.99^n \left(\cos(2\pi n/20) + \epsilon\right); \tag{17}$$

$$\mathbf{x}_2 \;=\; \left(\cos(2\pi n/20) + \epsilon\right); \tag{18}$$

$$\mathbf{x}_3 \;=\; 1.015^n \left(\cos(2\pi n/20) + \epsilon\right). \tag{19}$$



(a) $\mathbf{x}_1$: $L = 150$ and $\mathbf{A}_{\{1,2\}}$.

(b) $\mathbf{x}_2$: $L = 150$ and $\mathbf{A}_{\{1,2\}}$.

(c) $\mathbf{x}_3$: $L = 150$ and $\mathbf{A}_{\{1,2\}}$.

(d) $\mathbf{x}_1$: $L = 150$ and $\mathbf{A}_{\{1,2,\cdots,10\}}$.

(e) $\mathbf{x}_2$: $L = 150$ and $\mathbf{A}_{\{1,2,\cdots,10\}}$.

(f) $\mathbf{x}_3$: $L = 150$ and $\mathbf{A}_{\{1,2,\cdots,10\}}$.

Figure 1. Reconstruction and forecasting results for both SSA approaches, using $L = 150$ and two different groups of elementary matrices. The first 300 points generated for each series, $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$, were used for reconstruction. Then 100 more points were computed through the SSA forecasting method. In every situations, the results showed that Basic approach accomplished a much better time-series approximation in both synthesis and forecasting procedures. The Toeplitz approach, on the other hand, failed in the synthesis and forecasting procedures in (a), (e) and (f). Although in (b) the synthesized series were quite equivalent for both approaches, the Toeplitz presented a deviation from the original time-series for long-term forecasting.

Time-series $\mathbf{x}_1$ is a sinusoidal variation with a decreasing amplitude, $\mathbf{x}_2$ is a sinusoidal variation with constant amplitude and $\mathbf{x}_3$ is a sinusoidal variation with an increasing amplitude. All signals have the same frequency and are zero-mean, with $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$, and $\sigma_\epsilon^2 = 0.01$. $\mathbf{x}_1$ and $\mathbf{x}_3$ are actually amplitude modulated signals, with a known exponential modulating wave. For this reason, they do not satisfy the stationarity requirements for a time-series.

These time-series were presented in Golyandina [1, Section 8.2]. However, a deeper investigation is conducted here, especially concerning the frequency characteristics of eigenfilters and their pairwise scatterplots.

In Figure 1, the results obtained by both approaches for each synthetic time-series are depicted. The first 300 points generated for each series, $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$, were used for reconstruction. Then, 100 more points were computed through the SSA forecasting method. A vertical dashed line was inserted in order to discriminate these two parts of resulting time-series.

Table 1 exhibits the reconstruction and forecasting MAE for Basic and Toeplitz approaches, considering every synthetic signals, $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$.

Table 1. Mean Absolute Error – MAE. $\mathbf{A}_{\{a\}}$ means $\mathbf{A}_{\{1,2\}}$; $\mathbf{A}_{\{b\}}$ means $\mathbf{A}_{\{1,2,...,10\}}$

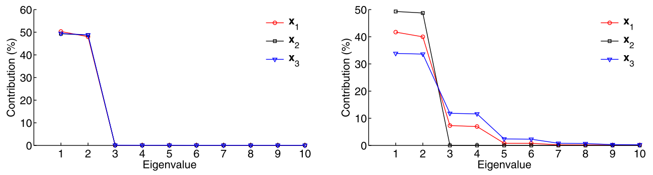
	Reconstruction Error				**Forecasting Error**			
	Basic		Toeplitz		Basic		Toeplitz	
	$\mathbf{A}_{\{a\}}$	$\mathbf{A}_{\{b\}}$	$\mathbf{A}_{\{a\}}$	$\mathbf{A}_{\{b\}}$	$\mathbf{A}_{\{a\}}$	$\mathbf{A}_{\{b\}}$	$\mathbf{A}_{\{a\}}$	$\mathbf{A}_{\{b\}}$
$\mathbf{x}_1$:	0.02	0.02	0.08	0.02	0.00	0.00	0.06	0.01
$\mathbf{x}_2$:	0.08	0.07	0.08	0.07	0.09	0.09	0.09	0.09
$\mathbf{x}_3$:	1.44	0.96	6.26	1.63	22.36	24.88	116.11	98.07

Figure 2 illustrates the eigenvalues contributions for the series variance – Equation[1] (8) – for every time-series and both methods.

Figure 3 illustrates the pairwise scatterplot of every analyzed time-series, synthesized by both Basic and Toeplitz SSA. Each point on the plot consists of a pair of corresponding elements of eigenvectors $\mathbf{x}_1$ and $\mathbf{x}_2$.

The synthetic time-series spectral estimation obtained from both SSA approaches, and also from the classic DFT method, are shown in Figure 4.

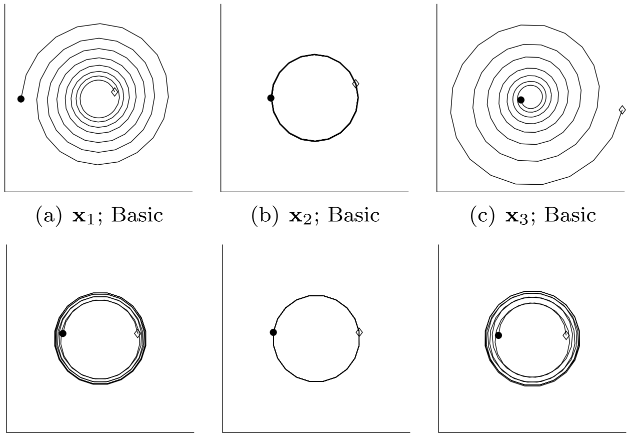
[1] To compute this contribution for the Toeplitz SSA σ^2 in equation (8) is replaced by τ.



(a) Basic approach.

(b) Toeplitz approach.

Figure 2. Eigenvalues percent contributions for the time-series variance according to both Basic and Toeplitz approaches. **(a)** *The graphs show close results for the contributions of eigenvectors obtained from the Basic approach, for all the time-series. Moreover, the biggest contribution is concentrated on eigenvectors $\mathbf{u}_1$ and $\mathbf{u}_2$.* **(b)** *The results obtained from Toeplitz approach are quite different from one time-series to another, and have spanned from eigenvectors $\mathbf{u}_1$ to $\mathbf{u}_4$ for the amplitude modulated time-series.*



(a) $\mathbf{x}_1$; Basic

(b) $\mathbf{x}_2$; Basic

(c) $\mathbf{x}_3$; Basic

(d) $\mathbf{x}_1$; Toeplitz

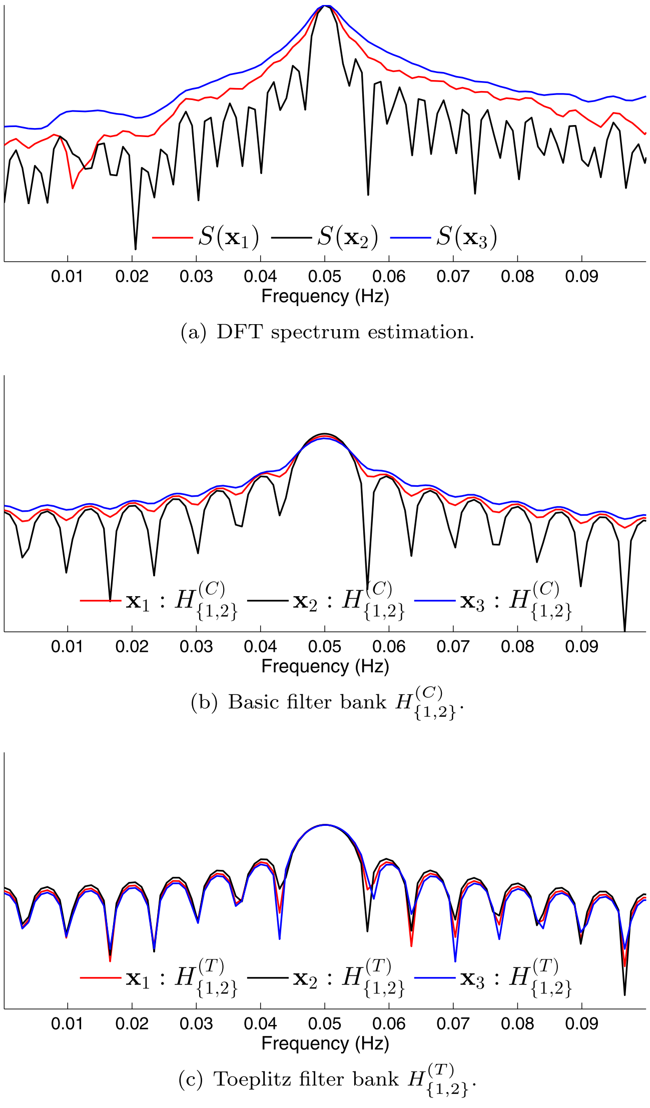
(e) $\mathbf{x}_2$; Toeplitz

(f) $\mathbf{x}_3$; Toeplitz

Figure 3. Pairwise scatterplots $\mathbf{u}_1 \times \mathbf{u}_2$. The black circle indicates the starting point and the white diamond the end point. The results displayed by the scatterplots obtained from Basic SSA made possible the time-series behavior identification. In (a), the decreasing amplitude of modulated time-series is portrayed by the shrinking spiral. The opposite happens for the increasing amplitude series in (c). In (b), a circle represents a constant amplitude sinusoidal time-series. In contrast, the scatterplots obtained from Toeplitz SSA were unable to discriminate the time-series amplitude behavior, as depicted in (d), (e) and (f). In particular the scatterplots (d) and (f) did not show a clear spiral.

3.1.2 Case 2

This section aims to extend the simulated example proposed in Golyandina et al. [25, Section 2.2.3.3], focused on the forecasting performance comparison. The MAE is computed for the forecasts based on both SSA approaches as a function of the window length (L) and the modulating signal damping factor (α).



(a) DFT spectrum estimation.

(b) Basic filter bank $H^{(C)}_{\{1,2\}}$.

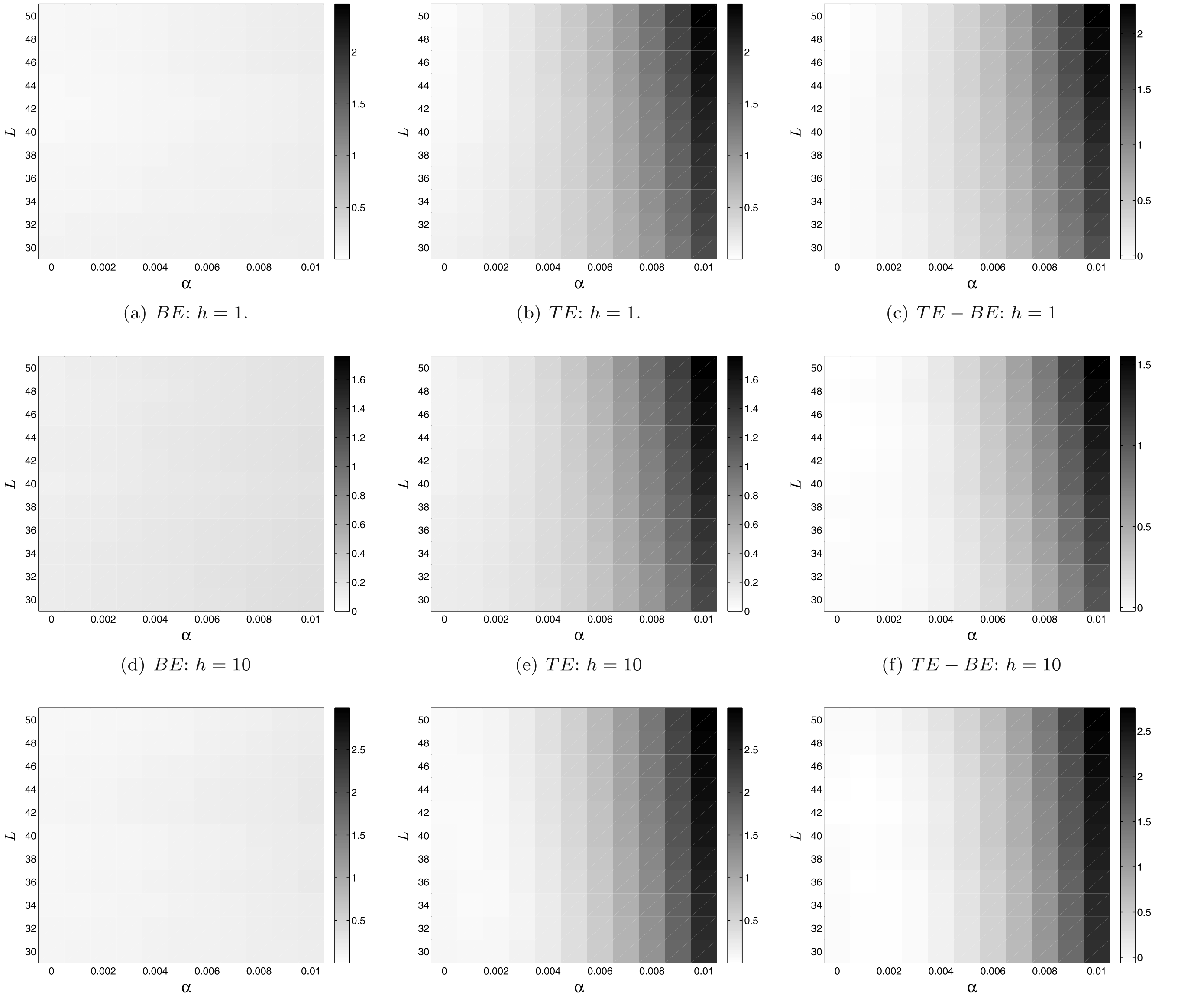
(c) Toeplitz filter bank $H^{(T)}_{\{1,2\}}$.

Figure 4. Log scale charts of time-series normalized spectra estimated by: (a) DFT; (b) Basic SSA eigenfilters frequency response; and (c) Toeplitz SSA eigenfilters frequency response. In both SSA approaches the eigenvectors ($\mathbf{u_1}$ and $\mathbf{u}_2$) were taken into account. The estimated spectra in (a) and (b) were quite similar and displayed a smooth spread of energy around the fundamental frequency for the amplitude modulated series, $\mathbf{x}_1$ and $\mathbf{x}_3$, and more concentrated response for the sinusoidal time-series. The estimated spectra in (c), however, showed very close frequency content for all the analyzed time-series, coincident with those obtained for the pure sinusoidal time-series on (a) and (b).

Let x_n be a new time-series described by:

$$x_n = e^{(\alpha n)}\sin(2\pi n/7), +\sigma\epsilon_n \tag{20}$$

where $\sigma = 0.5$ and $\epsilon \sim \mathcal{N}(0,1)$. For all $\alpha \neq 0$, x_n is a non-stationary time-series.



(a) BE: $h = 1$.

(b) TE: $h = 1$.

(c) $TE - BE$: $h = 1$

(d) BE: $h = 10$

(e) TE: $h = 10$

(f) $TE - BE$: $h = 10$

(g) BE: $h = 50$

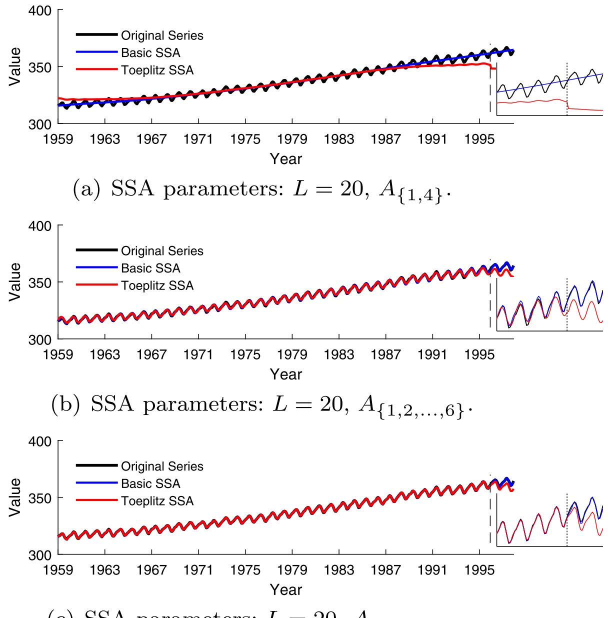
(h) TE: $h = 50$

(i) $TE - BE$: $h = 50$

Figure 5. MAE for short- and long-term forecasting results obtained from Basic and Toeplitz SSA. h means the number of steps ahead for the current forecast. The error behavior of SSA approaches are quite different, as can be observed by comparing (a) and (b), (d) and (e), (g) and (h). Basic SSA seems to be insensitive to the damping factor (α), which emphasizes its ability to deal with non-stationary time-series. Toeplitz, on the other hand, displays an increasing MAE as α increases. The window length (L) did not significantly influence the forecasting error, as can be observed in every images, which might be a particular characteristic of the simple analyzed time-series. In conclusion, the MAE differences corroborate with the possible insensitiveness to α previously pointed out. Actually, images (c), (f) and (i) are pretty similar to those in (b), (e) and (h), except for a light reduction of intensity, which can be explained by the almost constant error surfaces in (a), (d) and (g) produced by Basic SSA forecasts.

By varying α from 0 to 0.01, and L from 30 to 50, the h steps ahead forecasts are computed and then the MAEs for each pair (α, L) are calculated. The results are summarized by the images at Figure 5, which allow the easy comparison between error performance on forecasting.

Each row in figure displays the image, as shades of gray, corresponding to the MAEs obtained from Basic SSA (BE), figures 5(a), 5(d) and 5(g); followed by those resulting from Toeplitz SSA (TE), figures 5(b), 5(e) and 5(h); and then the difference between them ($TE - BE$), figures 5(c), 5(f) and 5(i).

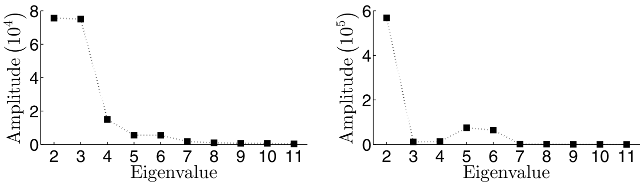


(a) SSA parameters: $L = 20$, $A_{\{1,4\}}$.

(b) SSA parameters: $L = 20$, $A_{\{1,2,\ldots,6\}}$.

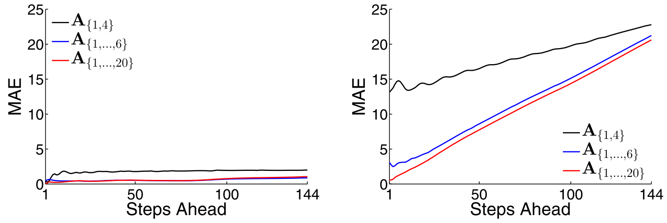
(c) SSA parameters: $L = 20$, $A_{\{1,2,\ldots,20\}}$.

Figure 6. CO_2 atmospheric concentration (ppm) data and analysis results. The original series was plotted together with the reconstructed ones using both Basic and Toeplitz SSA. A predicted interval was also plotted, after the vertical dashed line at 1997, equally using both approaches. The main observed difference among the distinct eigenvectors groups is related to the ability to describe short-period variations. In (a), both approaches described the long-term trend. However, a deviation from the series trend can be clearly observed for the Toeplitz synthesized data before 1967 and after 1987. In (b) and (c), both approaches were able to describe the short-period variations, but the deviation from original data is clearly greater for the Toeplitz SSA in the forecast stage. In conclusion, it is clear that the forecasts from Toeplitz approach have lost the series trend, in contrast to Basic, which showed a plausible evolution.



(a) Basic. (b) Toeplitz.

Figure 7. Eigenvalues spectra. In both graphs the eigenvalues of $\mathbf{u}_1$ were omitted to improve readability. ***(a)*** *At least two pairs of adjacent eigenvalues with approximately equal values could be identified: $(\mathbf{u}_2, \mathbf{u}_3)$ and $(\mathbf{u}_5, \mathbf{u}_6)$. They are clear indications of harmonic components in the series.* ***(b)*** *Only one pair of adjacent eigenvalues with close amplitude could be noticed: $(\mathbf{u}_5, \mathbf{u}_6)$. In Toeplitz SSA, the values of τ_l are equal to the contribution values. However, the sorting of eigenvectors is performed through the covariance matrix eigenvalues.*



(a) Basic. (b) Toeplitz.

Figure 8. Forecasting error (MAE) for 1 to 144 steps ahead prediction (h). The time-series was interrupted at December, 1986. Basic SSA forecasting showed a low and almost constant MAE as h increases. Toeplitz SSA forecasting, on the other hand, displayed an increasing MAE as h increases.

3.2 Real data time series

The experimental time-series shows the atmospheric CO_2 concentration in ppm from 1959 to 1997, collected by the Mauna Loa Observatory, Hawaii. This time-series was analyzed in [26]. In this paper, the simulations carried out compares the synthesis and forecasting results in a different perspective.

Figure 6 illustrates CO_2 concentration together with the reconstructed and forecast SSA time-series approaches, using three different groups and the same embedded dimensions. Figure 7 shows the eigenvalues contributions for the series variance according to both SSA approaches.

Table 2 shows the values of MAE, for six different experiments, which consisted of a whole year forecasting (12 months) following the periods:

– **Period I** – 1959 to December, 1996;
– **Period II** – 1959 to December, 1994;
– **Period III** – 1959 to December, 1992;
– **Period IV** – 1959 to December, 1990;
– **Period V** – 1959 to December, 1988;
– **Period VI** – 1959 to December, 1986.

In Figure 8 the obtained MAE for the experiment **VI** is depicted using $1 \leq h \leq 144$.

Although Basic and Toeplitz SSA have exhibited great performances on trend estimation, it is important to mention that the Toeplitz approach needs additional processing steps. In order to obtain the results in Figure 6 it is necessary to subtract the series mean before the covariance matrix calculation. Moreover, after SSA computations, it is necessary to add the previously subtracted mean to the reconstructed signal.

Table 2. MAE. Groups of elementary matrices $\mathbf{A}_{\{1,4\}}$, $\mathbf{A}_{\{1,2,\ldots,6\}}$ and $\mathbf{A}_{\{1,2,\ldots,20\}}$ are represented, respectively, as $\mathbf{A}_{\{a\}}$, $\mathbf{A}_{\{b\}}$ and $\mathbf{A}_{\{c\}}$

	Reconstruction Error						**Forecasting Error**					
	Basic			Toeplitz			Basic			Toeplitz		
	$\mathbf{A}_a$	$\mathbf{A}_b$	$\mathbf{A}_c$	$\mathbf{A}_a$	$\mathbf{A}_b$	$\mathbf{A}_c$	$\mathbf{A}_a$	$\mathbf{A}_b$	$\mathbf{A}_c$	$\mathbf{A}_a$	$\mathbf{A}_b$	$\mathbf{A}_c$
I:	1.75	0.28	0.16	2.75	0.58	0.18	1.79	0.40	0.31	13.42	3.48	2.05
II:	1.74	0.29	0.16	2.73	0.58	0.18	1.91	1.02	0.58	15.14	4.93	2.65
III:	1.76	0.30	0.17	2.93	0.58	0.18	1.88	0.27	0.22	15.07	3.70	2.00
IV:	1.77	0.30	0.16	2.91	0.56	0.19	2.33	1.26	0.74	14.46	2.88	1.91
V:	1.79	0.34	0.16	2.72	0.56	0.18	1.95	1.20	0.18	14.12	3.80	2.20
VI:	1.79	0.34	0.16	2.69	0.58	0.18	1.98	0.23	0.52	14.60	4.21	2.12

4. DISCUSSION

4.1 Synthetic data

The results from Figure 1 showed that the Basic SSA has successfully decomposed, synthesized and predicted, based on two eigenfilters, all the analyzed time-series, stationary and non-stationary. Toeplitz SSA, on the other hand, was unable to describe the non-stationary time-series. In Figure 1(a) and 1(c), eigenfilters $\mathbf{u}_1$ and $\mathbf{u}_2$ were selected. In such cases, both reconstruction and forecasting intervals deviated significantly from the original data. In 1(d) and 1(f), eigenfilters $\mathbf{u}_1$ to $\mathbf{u}_{10}$ were employed, and led to good reconstruction results, but one more time, deficient forecast achievements. Table 1 evidences that for $\mathbf{x}_1$ and $\mathbf{x}_3$, the MAEs obtained from Basic SSA are smaller than those produced by Toeplitz SSA. For $\mathbf{x}_2$, a stationary time-series, both SSA approaches display close MAE results.

According to SSA theory, a pair of eigenfilters is required for describing a harmonic oscillation. The results from Figure 2 not only confirm this statement for Basic SSA, but also show that they can represent an exponential amplitude modulation. In Figure 2(a), it can be noticed that eigenfilters $\mathbf{u}_1$ and $\mathbf{u}_2$ are actually the main responsible for the time-series variance. However, Toeplitz SSA results diverted one more time, attributing to eigenfilters $\mathbf{u}_3$ to $\mathbf{u}_6$ part of the series variability, as depicted in Figure 2(b).

The effect of exponential amplitude modulation becomes more clear by analyzing Figure 4. In 4(a) the frequency content of the analyzed time-series in a log-scale is displayed, obtained through DFT. As $\mathbf{x}_2$ is a harmonic oscillation, a single frequency is indicated by the main peak at 0.05 Hz surrounded by a spurious response due to the finite window-length and noise. The modulated series $\mathbf{x}_1$ and $\mathbf{x}_3$ display a continuous and smooth decreasing around the main component located at 0.05 Hz, resulting from the frequency convolution of harmonic and exponential Fourier Transforms[2].

Figure 4(b) depicts the Basic SSA eigenfilters $\mathbf{u}_1$ and $\mathbf{u}_2$ frequency response for each time-series, in log-scale. The results are very similar to those obtained from DFT. Figure 4(c) illustrates the same results reached by Toeplitz SSA. In this case, however, the frequency responses for every time-series are similar to those obtained for $\mathbf{x}_2$ by Basic SSA, which is related to a non-modulated harmonic variation. This is a clear indication that Toeplitz SSA was not capable of identifying the exponential amplitude modulation when the eigenvectors $\mathbf{u}_1$ and $\mathbf{u}_2$ was employed. Figure 3 confirms this statement. The spiral trajectory in 3(a) describes a decreasing amplitude while in 3(c) an increasing amplitude is depicted. In contrast, the corresponding trajectories obtained by the Toeplitz SSA, Figures 3(d) and 3(f) do not show a spiral trajectory, but several turns of a circular trajectory with some radius variation.

A set of experiments concerning the forecasting performance of both SSA methodologies was carried out, as described in Section 3.1.2. The exponential modulating signal makes possible the damping factor (α) variation, which can be useful for quantifying the effects of non-stationarity. Figure 5 summarizes these results. In summary, Basic SSA exhibited a superior performance compared to Toeplitz SSA, since it did not show significant error variations as a consequence of damping factor variation. Toeplitz SSA, on the other hand, showed an increasing error behavior as α increases, which can be understood as a fragility to non-stationary data. Neither Basic nor Toeplitz methods showed a significant error variation as a function of window length (L).

4.2 Real data

The results illustrated in Figure 6 showed that Toeplitz SSA failed in several aspects for both synthesizing and forecasting procedures. In Figure 6(a), the time-series trend was successfully retrieved by the Basic SSA. However, a noticeable deviation can be observed, for Toeplitz SSA, before 1967 and after 1990. In the forecasting section, the deviation is still worse.

Figures 6(b) and 6(c) showed that Basic SSA succeeded in the time-series reconstruction, trend and harmonic components, and also in forecasting section. The same did not happen for the Toeplitz SSA, which was capable of synthe-

[2] $\mathcal{F}\{u(t)\,e^{-\alpha t}\,\cos(\omega_0 t)\} = \frac{1}{\alpha+j\omega} * \pi[\delta(\omega+\omega_0)+\delta(\omega-\omega_0)] = \frac{\alpha+j\omega}{\omega_0^2+(\alpha+j\omega)^2}$. $\mathcal{F}\{.\}$: Fourier transform. $u(t)$: time-domain unit step function. $\delta(w)$: frequency-domain Dirac delta function.

sizing the series components, but showed a noticeable deviation from the original data in the forecasting section. By comparing Figures 6(b) and 6(c), it can be observed that increasing the eigenvectors set ($\mathbf{u}_1$ to $\mathbf{u}_{20}$) did not produce noticeable improvement to Toeplitz based forecasting. By inspecting Figure 8, one can recognize that Basic SSA forecasting showed a superior error performance.

The Basic SSA results observed in Figure 6 can be explained by the eigenvalues contributions illustrated in Figure 7(a). The series trend was attributed to eigenvector $\mathbf{u}_1$, which represents a great part of its variability. At least two harmonic components could be identified by the eigenvalues plateaus, $(\mathbf{u}_2, \mathbf{u}_3)$ and $(\mathbf{u}_5, \mathbf{u}_6)$.

The identification of trend and oscillatory components based on Toeplitz eigenvalues is not a straightforward procedure. In order to achieve good results in time-series synthesizing, a large number of eigenvectors must be taken into account.

5. CONCLUSION

This paper addresses theoretical and application aspects of Toeplitz and Basic SSA approaches focused on non-stationary signals with any kind of preprocessing step before time-series analysis.

Several aspects related to synthesized and forecast time-series were analyzed: the eigenfilters frequency responses; the pairwise scatterplots; and the error performance for synthesis and forecasting. Actual and synthetic data were employed.

In conclusion, the analyses results indicate that Toeplitz SSA should not be used for non-stationary time-series, which usually results in an increased number of elementary matrices necessary to achieve an accurate reconstruction. An analogous could be made to the problem of *overfitting* in some models identification procedures. Additionally, the dominant structure of non-stationary series was lost in forecasting section.

Although Toeplitz SSA is considered the mainstream SSA version, the results obtained in this paper recommends its usage only if the series under analysis is stationary or if a pre-processing stage (which includes trend removing procedure) is applied to a non-stationary signal. Thus, Toeplitz SSA may be seen as a particular case of the Basic approach, in which no assumptions are made about time-series stationarity. Finally, it should be noted that if the time-series under analysis is stationary, Golyandina [1] and Ghil et al. [17] agree that best choice is the Toeplitz SSA.

ACKNOWLEDGMENTS

Cairo L. Nascimento Jr. and Michel C. R. Leles gratefully acknowledge the support from FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) under grants 2016/04992-6 and 2017/20248-8, respectively.

Received 27 June 2018

REFERENCES

[1] Golyandina, N., On the choice of parameters in Singular Spectrum Analysis and related subspace-based methods, Statistics and its interface 3 (2010) 259–279. MR2720132

[2] M. C. R. Leles, Vale-Cardoso, A. S., Moreira, M. G., Guimarães, H. N., Silva, C. M., Pitsillides, A., Frequency-domain characterization of Singular Spectrum Analysis eigenvectors, in: Signal Processing and Information Technology (ISSPIT), 2016 IEEE International Symposium on, IEEE, 2016, pp. 22–27.

[3] Groth, A., Ghil, M., Monte Carlo Singular Spectrum Analysis (SSA) revisited: Detecting oscillator clusters in multivariate datasets, Journal of Climate 28 (2015) 7873–7893.

[4] Sanei, S., Ghodsi, M., Hassani, H., An adaptive Singular Spectrum Analysis approach to murmur detection from heart sounds, Medical Engineering & Physics 33 (2011) 362–367.

[5] Leles, M. C. R., Mozelli, L. A., Guimarães, H. N., New trend-following indicator: Using SSA to design trading rules, Fluctuation and Noise Letters 16 (2017) 1750016.

[6] Leles, M. C. R., Sansao, J. P., Mozelli, L. A., Guimarães, H. N., Improving Reconstruction of Time-Series With Singular Spectrum Analysis: A Segmentation Approach, 2017. To appear in *Digital Signal Processing*. MR3794306

[7] Leles, M. C. R., Sansao, J. P., Mozelli, L. A., Guimarães, H. N., A new algorithm in Singular Spectrum Analysis framework: the Overlap-SSA (ov-SSA), 2017. To appear in *Software X*. MR3794306

[8] Golyandina, N., Shlemov, A., Variations of Singular Spectrum Analysis for separability improvement: non-orthogonal decompositions of time series, Statistics and Its Interface 8 (2015) 277–294. MR3341327

[9] Zhang, J., Hassani, H., Xie, H., Zhang, X., Estimating multi-country prosperity index: A two-dimensional singular spectrum analysis approach, Journal of Systems Science and Complexity 27 (2014) 56–74. MR1833045

[10] Ghodsi, M., Hassani, H., Rahmani, D., Silva, E. S., Vector and recurrent Singular Spectrum Analysis: which is better at forecasting? Journal of Applied Statistics 45 (2018) 1872–1899. MR3811848

[11] Hassani, H., Heravi, S., Brown, G., Ayoubkhani, D., Forecasting before, during, and after recession with singular spectrum analysis, Journal of Applied Statistics 40 (2013) 2290–2302. MR3291164

[12] Hassani, H., Mahmoudvand, R., Zokaei, M., Ghodsi, M., On the separability between signal and noise in Singular Spectrum Analysis, Fluctuation and Noise Letters 11 (2012) 1250014.

[13] Rodriguez-Aragon, L. J., Zhigljavsky, A., Singular Spectrum Analysis for image processing, Statistics and Its Interface 3 (2010) 419–426. MR2720144

[14] Golyandina, N., Zhigljavsky, A. A., Singular Spectrum Analysis for Time Series, Springer Briefs in Statistics, 2013. MR3024734

[15] Hassani, H., Singular Spectrum Analysis: Methodology and comparison, Journal of Data Science 5 (2007) 259–267. MR2538869

[16] Golyandina, N., Nekrutkin, V., Zhigljavsky, A. A., Analysis of Time Series Structure: SSA and Related Techniques, Chapman & Hall/CRC, 2001. MR1823012

[17] Ghil, M., Allen, M. R., Dettinger, M. D., Ide, K., Kondrashov, D., Mann, M. E., Robertson, A. W., Saunders, A., Tian, Y., Yiou, P., Advanced spectral methods for climatic time series, Reviews of Geophysics 40 (2002) 1–41.

[18] Vautard, R., Yiou, P., Ghil, M., Singular-Spectrum Analysis: A toolkit for short, noisy chaotic signals, Physica D 58 (1992) 95–126.

[19] Broomhead, D. S., King, G. P., On the qualitative analysis of experimental dynamical systems, in: S. Sarkar (Ed.), Nonlinear Phenomena and Chaos, Adam Hilger, Bristol, England, 1986, pp. 113–144.

[20] Vautard, R., Ghil, M., Singular Spectrum Analysis in nonlinear

dynamics, with applications to paleoclimatic time series, Physica D 35 (1989) 395–424. MR1004204

[21] Leles, M. C. R., Cardoso, A. S. V., Moreira, M. G., Sbruzzi, E. F., Nascimento, C. L., Guimarães, H. N., A Singular Spectrum Analysis based trend-following trading system, in: 2018 Annual IEEE International Systems Conference (SysCon), 2018, pp. 1–5. doi:10.1109/SYSCON.2018.8369524.

[22] Hamilton, J. D., Time Series Analysis, Princeton, NJ, 1994. MR1278033

[23] Anderson, T. W., The statistical analysis of time series, volume 19, John Wiley & Sons, 2011. MR0283939

[24] Tome, A. M., Teixeira, A. R., Figueiredo, N., Santos, I. M., Georgieva, P., Lang, E., SSA of biomedical signals: A linear invariant systems approach, Statistics and Its Interface 3 (2010) 345–355. MR2720138

[25] Golyandina, N., Korobeynikov, A., Zhigljavsky, A., Singular Spectrum Analysis with R, Springer, 2018. MR3793637

[26] Golyandina, N., Korobeynikov, A., Basic Singular Spectrum Analysis and forecasting with r, Computational Statistics & Data Analysis 71 (2014) 934 – 954. MR3132018

Michel C. R. Leles
Department of Technology
Universidade Federal de São João del-Rei
Ouro Branco, MG
Brazil
E-mail address: mleles@ufsj.edu.br

Mariana G. Moreira
Department of Technology
Universidade Federal de São João del-Rei
Ouro Branco, MG
Brazil
E-mail address: marianageny@ufsj.edu.br

Adriano S. Vale-Cardoso
Department of Technology
Universidade Federal de São João del-Rei
Ouro Branco, MG
Brazil
E-mail address: adrianosvc@ufsj.edu.br

Cairo L. Nascimento Júnior
Electronics Engineering Division
Instituto Tecnológico de Aeronáutica
São José dos Campos, SP
Brazil
E-mail address: cairo@ita.br

Elton F. Sbruzzi
Division of Computer Science
Instituto Tecnológico de Aeronáutica
São José dos Campos, SP
Brazil
E-mail address: elton@ita.br

Homero N. Guimarães
Department of Electrical Engineering
Universidade Federal de Minas Gerais
Belo Horizonte, MG
Brazil
E-mail address: homero@cpdee.ufmg.br