Digital Signal Processing ••• (••••) •••–•••

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Improving reconstruction of time-series based in Singular Spectrum Analysis: A segmentation approach

M.C.R. Leles [a,b,*], J.P.H. Sansão [a,b], L.A. Mozelli [a,c], H.N. Guimarães [d]

[a] CELTA – Center for Studies in Electronics Engineering and Automation, UFSJ – Federal University of São João del-Rei, Rod. MG 443 km 7, 36420-000, Ouro Branco, MG, Brazil
[b] PPGEE – Graduate Program in Electrical Engineering, UFMG – Federal University of Minas Gerais, Av. Antônio Carlos 6627, 31270-901, Belo Horizonte, MG, Brazil
[c] Department of Electronics Engineering – UFMG, Brazil
[d] Department of Electrical Engineering – UFMG, Brazil

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Singular spectrum analysis
Non-stationary signals
Segmentation

ABSTRACT

Singular Spectrum Analysis (SSA) is a powerful non-parametric framework to analysis and enhancement of time-series. SSA may be capable of decomposing a time-series into its meaningful components: trends, oscillations and noise. However, if the signal under analysis is non-stationary, with its spectrum spreading and varying in time, the reliability of the reconstruction is guaranteed only when many elementary matrices are used. As a consequence, the capability to discriminate dominant structures from time-series may be impaired. To circumvent this issue, a new method, called overlap-SSA (ov-SSA), is proposed for segmentation, analysis and reconstruction of long-term and/or non-stationary signals. The raw time series is divided into smaller, consecutive and overlapping segments, and standard SSA procedures are applied to each segment with the resulting series being concatenated. This variation of SSA seeks to: improve reconstruction and component separability for non-stationary time-series; enable the analysis for large datasets, avoiding the issues of concatenation of many segments; and present some benefits of the segmentation in terms of better time–frequency characterization. These advantages are illustrated in several synthetic and experimental datasets.

1. Introduction

Singular Spectrum Analysis (SSA) is a non-parametric approach that can be used to decompose the original time-series into additive structures, promoting an efficient way to separate meaningful components, such as trends or oscillations, whilst discarding noise. It has been used into a plethora of applications and fields: time-series reconstruction [17], frequency estimation [27,16], trend extraction [29], time-series forecasting [18], trends and harmonic extraction [30], and many others [15].

The technique consist in four major steps. The *Immersion*, which maps the raw time-series into a Hankel matrix, called trajectory matrix. Parameter L, the embedding dimension, is used to assemble the trajectory matrix. The *Singular Value Decomposition* (SVD), which factorizes the trajectory matrix into a set of L elementary matrices, $A_1, A_2, \dots, A_L$. In *Grouping*, occurs the combination of a subset of elementary matrices, $A_{\{I\}} = \sum_{i \subset I} A_i$, that capture the desired structures. Finally, *Diagonal Averaging* transforms the resultant matrix from grouping into the reconstructed time-series.

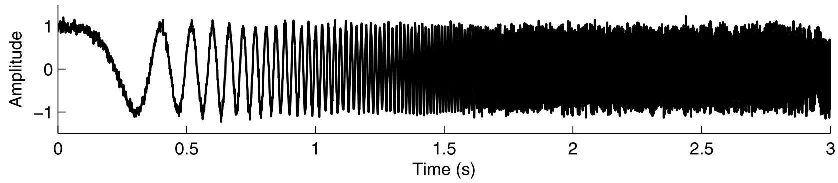
Classical implementations of SSA suffer from some drawbacks, one of which will be deepened in the sequence. Consider a synthetic signal depicted in Fig. 1(a), which is an exponential chirp signal[1] with addition of noise (SNR = 20 dB).

If a classical SSA approach is used, whose mathematical details will be provided latter in Section 3.1, one may attempt to capture the main features of this signal on decomposition stage. Lets consider two groupings, one resorting to the first 10 components and the other considering the first 20 components, and their respective reconstructed time-series, portrayed in Fig. 1. It is noticeable, as time increases, that reconstructed time-series are less trustworthy. Likewise, as more components are introduced into grouping stage, the reconstructed time-series becomes more reliable. Therefore, the only way to achieve a reasonable reconstruction, in this

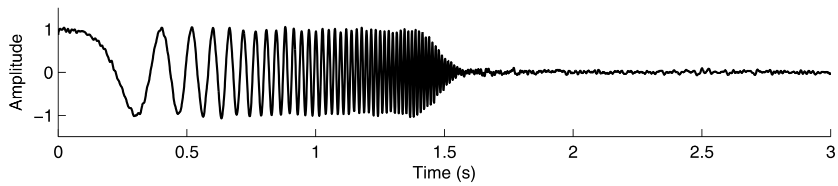
* Corresponding author at: CELTA – Center for Studies in Electronics Engineering and Automation, UFSJ – Federal University of São João del-Rei, Rod. MG 443 km 7, 36420-000, Ouro Branco, MG, Brazil.
E-mail addresses: mleles@ufsj.edu.br (M.C.R. Leles), joao@ufsj.edu.br (J.P.H. Sansão), mozelli@cpdee.ufmg.br (L.A. Mozelli), hnguimar@gmail.com (H.N. Guimarães).

[1] To generate this signal *Matlab*® command `chirp(t,0.5,2,150, 'q');` is used.

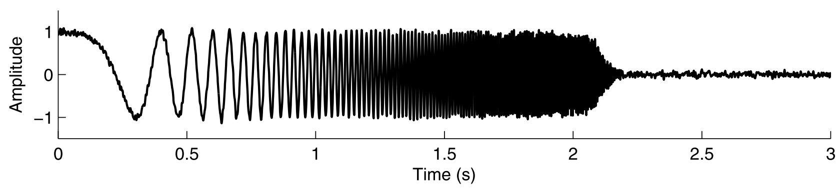
https://doi.org/10.1016/j.dsp.2017.10.025
1051-2004/



(a) Cosine signal with variable frequency.



(b) Reconstructed time-series using $L = 60$ and $\mathbf{A}_{\{1,2,\cdots,10\}}$.



(c) Reconstructed time-series using $L = 60$ and $\mathbf{A}_{\{1,2,\cdots,20\}}$.

Fig. 1. Original time-series and reconstructed ones by SSA using $L = 60$ and two groupings.

example, is by resorting to a large number of components. However, as the number of components increases, it becomes more difficult to distinguish dominant structures from each other.

This example shows a shortcoming of classical SSA: non-stationary signals. The concept of "non-stationarity" is often defined in different ways. In this paper, non-stationarity is invoked to refer to deterministic signals whose spectral properties are time-variant [33, Chapter 11].

Other drawback appears when applying SSA to large sequential datasets, typically associated with very long time-series, arising from telecommunications, finance, bioinformatics or web mining. Usage of SSA for the entire time-series may lead to cumbersome issues, due to computation of SVD of a high dimensional matrix, when basic (off-the-shelf) numeric packages are adopted. Alternative implementations were proposed by Korobeynikov [26] which are computationally more efficient and may reduce this impact. For the Multivariate Singular Spectral Analysis (M-SSA) a solution for this drawback was proposed by Pukenas [35].

An efficient approach to handle at once both these issues, non-stationarity and large data, can be segmenting. Yiou et al. [44] proposed a local approach of SSA, called MS-SSA (Multi-Scale SSA). It seeks to perform a joint time–frequency analysis, analogously to wavelet transform. However this capability is obtained at the cost of great computational effort. Rekapalli and Tiwari [36] proposed another modification of SSA algorithm called Windowed SSA (WSSA). The idea is to perform segmentation of the dataset and compute SSA for each segment. However, there is no treatment on how to join the segments properly.

In this paper, a new algorithm based on the SSA is proposed. Like many others variations of the standard SSA, as discussed latter in this paper, the objective is to improve the application of SSA to some common situations that may occur in various fields of signal processing. The algorithm is not intended to improve the SSA for every applications but is intended to be useful for:

- improve reconstruction and component separability for non-stationary time-series, like the aforementioned example;
- improve the analysis for large datasets, avoiding the issues of concatenation of many segments;
- discuss and present some advantages of the segmented analysis in terms of better time–frequency characterization.

The remaining of this paper is organized as follows: Section 2 provides the mathematical framework concerning standard SSA analysis; Section 3 epitomizes the main features about the proposed method[2]; Section 4 discuss several examples to illustrate the benefits provided by the ov-SSA, considering synthetic and experimental datasets; Section 5 presents some remarks about the proposed method; and finally, Section 6 lays down the conclusion and future perspectives.

2. SSA

In this section, a brief description of SSA methodology is given, according to Golyandina et al. [11].

2.1. Embedding

Time-series $\mathbf{x} = (x_0, x_1, \cdots, x_n, \cdots, x_{N-1})^T$, with length N, represents the signal under analysis. 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:

$$\mathbf{A} = \begin{bmatrix} x_0 & x_1 & \cdots & x_{K-1} \\ x_1 & x_2 & \cdots & x_K \\ \vdots & \vdots & & \vdots \\ x_{L-1} & x_L & \cdots & x_{N-1} \end{bmatrix}, \tag{1}$$

where L is the window length, or embedding dimension, and $K = N - L + 1$. $\mathbf{A}$, is a *Hankel* matrix, called the trajectory matrix [11].

[2] Complexity analysis and computational details of the proposed algorithm are provided in the unpublished paper by Leles et al. [28], which is companion for this Joint Special Issue.