ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

Automatic Singular Spectrum Analysis for Time-Series Decomposition

A.M. Álvarez-Meza and C.D. Acosta-Medina and G. Castellanos-Domínguez *

Universidad Nacional de Colombia, Signal Processing and Recognition Group
Campus La Nubia, km 7 vía al Magdalena, Manizales-Colombia

Abstract. An automatic Singular Spectrum Analysis based methodology is proposed to decompose and reconstruct time-series. We suggest a clustering based procedure to identify the main dynamics of the input signal, by computing a subset of orthogonal basis using a power spectrum criterion. The subset of basis are represented by the Discrete Fourier Transform to infer basis vectors encoding similar data structures. Thus, it is possible to highlight hidden components into the signal. Our approach is tested over some synthetic and real-world datasets, showing that our algorithm is a good tool to decompose time-series.

1 Introduction

In many real-world problems related to signal processing it is necessary to identify hidden structures (components) from the given input, in order to improve the performance of denoising, feature selection/extraction, and classification stages. In this sense, projective techniques appear as a tool to generate an alternative representation of the data, where such structures could be easily identified and interpreted. Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) are the most common multivariate data projective techniques, which are widely used as multivariate statistical analysis for feature extraction. Those methods analyze a covariance function over the data to find the alternative data representation. Projective techniques usually comprise three main steps: projection of the data, selection of relevant components, and the reconstruction.

However, traditional projective techniques can not be directly applied to one dimensional signal, which is the case of time-series [1]. Recently, Singular Spectrum Analysis - SSA has been developed as a projective technique that can be applied to time-series [1, 2, 3]. SSA decomposes the original signal into a sum of small numbers of interpretable components, such as, slowly varying trends, oscillatory components, and noise. Commonly, SSA is applied as a denoising stage, without interpreting different components inside the signal, which could be used to build an automatic time-series decomposition approach.

Here, we proposed an automatic SSA based methodology to decompose and reconstruct time-series. Therefore, a subset of orthogonal basis computed from the input are selected using a power spectrum criterion [4]. Moreover, we present a clustering based procedure to highlight the main dynamics of the time-series.

*Research carried out under grants provided by a PhD scholarship and by the project 111045426008 funded by Colciencias.

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

Thus, the subset of basis are represented by the Discrete Fourier Transform (DFT), to identify basis encoding similar data structures. Our aim is to reconstruct each component of the input signal by gathering orthogonal basis that share similar power spectrum properties. The original signal is reconstructed by a linear combination of the estimated components. Our approach is tested over some synthetic and real-world datasets. The remainder of this paper is organized as follow. Section 2 describes the main ideas behind SSA. Section 3 presents the proposed methodology to decompose and to reconstruct time-series based on SSA. Section 4 shows the experimental set-up and results. Finally, in sections 5 and 6 we discuss and conclude about the attained results.

2 Singular Spectrum Analysis - SSA

Traditional projective techniques, such as SVD/PCA, can not be directly applied to one dimensional time-series, being necessary the embedding of the one-dimensional signal into a high-dimensional space of time delayed coordinates [1]. Let $\boldsymbol{y} = \{y_t : t = 1, \ldots, T\}$ be a real-valued time-series with $\boldsymbol{y} \in \mathbb{R}^T$, which is mapped into the multidimensional set $\boldsymbol{H} = \{\boldsymbol{y}_l^\top : l = 1, \ldots, L\}$, $\boldsymbol{H} \in \mathbb{R}^{K \times L}$, termed Hankel matrix, comprising L-lagged vectors, and with $\boldsymbol{y}_l \in \Re^K$. In this regard, $\boldsymbol{H}$ can be calculated as

$$\boldsymbol{H} = \begin{bmatrix} y_1 & y_2 & \cdots & y_{L-1} & y_L \\ y_2 & y_3 & \cdots & y_L & y_{L+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ y_{K-1} & y_K & \cdots & y_{L-3+K} & y_{L-2+K} \\ y_K & y_{K+1} & \cdots & y_{L-2+K} & y_{L-1+K} \end{bmatrix}, \tag{1}$$

where $K = T - L + 1$ is the employed window size to embed the original time-series. Now, the Hankel-matrix $\boldsymbol{H}$ in (1) can be considered as a multivariate representation of $\boldsymbol{y}$, which is used by SSA to perform a SVD analysis. Thus, an alternative representation of $\boldsymbol{H}$ can be written as $\boldsymbol{H} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, being $\boldsymbol{U} \in \mathbb{R}^{K \times K}$ and $\boldsymbol{V} \in \mathbb{R}^{L \times L}$ the left and right singular vectors of $\boldsymbol{H}$, respectively, and $\boldsymbol{\Sigma} \in \mathbb{R}^{K \times L}$ is a rectangular matrix containing the $\boldsymbol{H}$ singular values on its diagonal. A reconstructed version $\hat{\boldsymbol{H}}$ of the original Hankel-matrix can be obtained by $\hat{\boldsymbol{H}} = (\hat{\boldsymbol{U}}^\top)^\dagger \boldsymbol{Z}$, being $\boldsymbol{Z} \in \mathbb{R}^{m \times L}$ the projection of $\boldsymbol{H}$ by the matrix $\hat{\boldsymbol{U}} \in \mathbb{R}^{K \times m}$, which is conformed by the m most relevant basis vectors of $\boldsymbol{U}$ (e.g. analyzing the singular values contained in $\boldsymbol{\Sigma}$). Finally, to estimate the reconstructed version $\hat{\boldsymbol{y}}$ of $\boldsymbol{y}$, a diagonal averaging is computed over $\hat{\boldsymbol{H}}$, as $\hat{\boldsymbol{y}} = \xi(\hat{\boldsymbol{H}})$, with $\xi(\cdot) : \mathbb{R}^{K \times L} \mapsto \mathbb{R}^T$ [1]. Thus, $\hat{\boldsymbol{y}}$ highlights the unfolded main dynamics of $\boldsymbol{y}$ according to the selected basis $\hat{\boldsymbol{U}}$.

3 Automatic Signal Decomposition based on SSA

The aim of SSA is to achieve a decomposition of the original time-series as $\boldsymbol{y} = \sum_{j=1}^{C} \boldsymbol{y}_j + \boldsymbol{\eta}$, being $\boldsymbol{y}_j \in \mathbb{R}^T$ an interpretable component (dynamic) of $\boldsymbol{y}$, C is the

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

number of considered components, and $\boldsymbol{\eta} \in \mathbb{R}^T$ is a noise component perturbing the signal. Hence, in order to attain a suitable representation of the original time-series, it is necessary to fix two main free parameters: the embedding dimension K and the number of basis vectors m. It is important to note that SSA considers $\boldsymbol{y}$ as a finite rank series [2], therefore, looking for a suitable window size value, K is incremented according to the set $\boldsymbol{\theta} = \{2, \ldots, T-2\}$, until that

$$rank(\boldsymbol{H}^{\theta_{r+1}}) < rank(\boldsymbol{H}^{\theta_r}), \tag{2}$$

where $rank(\cdot)$ is the rank function, $r = 1, \ldots, T-3$; $\boldsymbol{H}^{\theta_r} \in \mathbb{R}^{\theta_r \times L_r}$, and $L_r = T - \theta_r + 1$. When (2) is accomplished, the SSA window size is fixed as $K = \theta_r$.

SSA is able to find out different basis vectors containing the main dynamics of $\boldsymbol{y}$, however, such dynamics can be properly described by mixing basis with similar properties. Here, we proposed a clustering based approach to infer the main components of $\boldsymbol{y}$ from SSA. Namely, given the SVD of $\boldsymbol{H}$, the singular values of $\boldsymbol{H}$ are stored in $\boldsymbol{\lambda} = \{\lambda_a : a = 1, \ldots, rank(\boldsymbol{H})\}$, $\boldsymbol{\lambda} \in \mathbb{R}^{rank(\boldsymbol{H})}$, with $\lambda_1 > \lambda_2 > \cdots > \lambda_{rank(\boldsymbol{H})}$ [4]. To fix m, we look for the first λ_m value such that $\lambda_m / \sum_{a=1}^{rank(\boldsymbol{H})} \lambda_a < \tau_\lambda$. Then, from the subset of singular vectors $\{\lambda_1 > \lambda_2 > \cdots > \lambda_m\}$, their corresponding left singular vectors are stored in the matrix $\hat{\boldsymbol{U}}$. Consequently, the main components of the original time-series are represented by the selected basis. Note that in this approach we assume that the power of each time-series component $\boldsymbol{y}_j$ is higher than the power of the noise component $\boldsymbol{\eta}$ (see [4] for details). Due to each $\boldsymbol{y}_j$ can be represented by one or a mixture of basis vectors, it is necessary to properly identify how many components C are hidden into the signal. Therefore, we proposed to characterize each column vector $\hat{\boldsymbol{u}}_i \in \mathbb{R}^K$ in $\hat{\boldsymbol{U}}$, with $i = 1, \ldots, m$; by a function $\vartheta(\cdot) : \mathbb{R}^K \mapsto \Re^p$. Our goal is to cluster basis with a similar behavior according to $\vartheta(\cdot)$. For such purpose, we employs the Discrete Fourier Transform - DFT as $\hat{\boldsymbol{u}}_i$ descriptor, to obtain a feature representation matrix $\boldsymbol{X} \in \mathbb{R}^{p \times m}$, being p the number of considered frequencies in DFT. In this approach the DFT is employed to reveal the main power spectrum features of each $\boldsymbol{y}_j$, however, other kind of representations could be employed. Now, to infer the number of components C that are hidden into $\boldsymbol{y}$, a singular value analysis is performed over $\boldsymbol{X}$. Then, the columns of $\boldsymbol{X}$ are clustered into C groups by a distance based method (e.g. Euclidean distance). From the obtained labels, the $\hat{\boldsymbol{U}}^{(j)} \in \mathbb{R}^{K \times n_j}$ matrices are conformed, being n_j the number of elements (basis) in the j-th group. Besides, for each $\hat{\boldsymbol{U}}^{(j)}$, a projection $\boldsymbol{Z}^{(j)} \in \mathbb{R}^{n_j \times L}$ is computed as in traditional SSA, which encodes the data structure properties of $\boldsymbol{y}_j$. Finally, the reconstructed time-series $\hat{\boldsymbol{y}}$ is estimated as in (3)

$$\hat{\boldsymbol{y}} = \sum_{j=1}^{C} \hat{\boldsymbol{y}}_j = \sum_{j=1}^{C} \xi\left(\hat{\boldsymbol{U}}^{(j)} \boldsymbol{Z}^{(j)}\right). \tag{3}$$

4 Experiments

In order to test the capability of the proposed approach to automatically identify different dynamics over time-series, and to properly reconstruct the input

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

signal, even against noise conditions, two synthetic and one real-world dataset are tested. For all the provided experiments, τ_λ is fixed as 0.025, and the Euclidean distance is used together with a cutoff based clustering algorithm to identify the components of the signal. The first synthetic dataset is called the Three-Sin, which is composed by three different sinusoidal functions as $y(t) = sin(2\pi ft) + \frac{1}{2}sin(4\pi ft) + \frac{1}{3}sin(6\pi ft)$, where $f = 60Hz$. Thus, 300 samples are generated uniformly from 0 to $\frac{1}{3}f$ seconds. The second dataset is named Sin-Sinc, which is created as $y(t) = 2sinc(t) + \frac{1}{4}sin(2\pi ft)$. Again, 300 samples are generated uniformly from 0 to 0.2 seconds. For both synthetic datasets, we test perturbing the input time-series with additive white Gaussian noise against different signal to noise ratio, SNR[dB]=$\{5, 10, 15\}$, to verify the algorithm robustness. The average relative error - AE is computed as in (4)

$$\mathrm{AE}\left(\hat{\boldsymbol{y}}\right) = 100\frac{1}{10}\sum\nolimits_{b=1}^{10}\frac{\left\|\hat{\boldsymbol{y}}^{(b)} - \boldsymbol{y}\right\|^2}{\|\boldsymbol{y}\|^2}[\%], \tag{4}$$

where $\hat{\boldsymbol{y}}^{(b)}$ is the reconstructed signal according to (3) at the b-th simulation. Table 1 describes the AE results and the number of employed basis m for each synthetic dataset. All the provided results are presented with the obtained standard deviation for 10 simulations. Furthermore, in Fig. 1(a) and 1(c) the estimated dynamics (SNR=5[dB]) are presented for the Three-Sin and the Sin-Sinc datasets, respectively. Alike, in Fig. 1(b) and 1(d) the reconstructed time-series are shown. Finally, the European Climate Assessment-ECA real-world dataset is tested [5]. This database is a weather daily summary of Berlin, Germany between 2001 to 2004. Nine variables are measured: could cover, mean relative humidity, mean barometric pressure, snow depth, precipitation amount, sunshine, amount of rain, minimum air temperature, maximum air temperature and mean air temperature. For concrete testing, the mean air daily temperature is studied as input signal. In Fig. 1(e) and 1(f) the obtained components and the reconstruction of the ECA mean air temperature are presented.

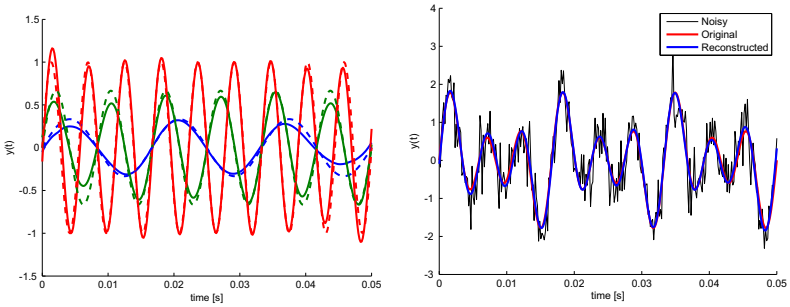
Table 1: Synthetic data reconstruction results against different noise levels.

Data	SNR = 2[dB]		SNR = 5[dB]		SNR = 10[dB]	
	AE[%]	m	AE[%]	m	AE[%]	m
Three-Sin	14.9 ± 3.6	6.9 ± 0.3	8.5 ± 1.1	7 ± 0	4.4 ± 0.7	7 ± 0.0
Sin-Sinc	13.8 ± 2.3	6.8 ± 0.4	9.1 ± 1.5	6.6 ± 0.5	4.6 ± 1.2	7.1 ± 0.6

5 Discussion

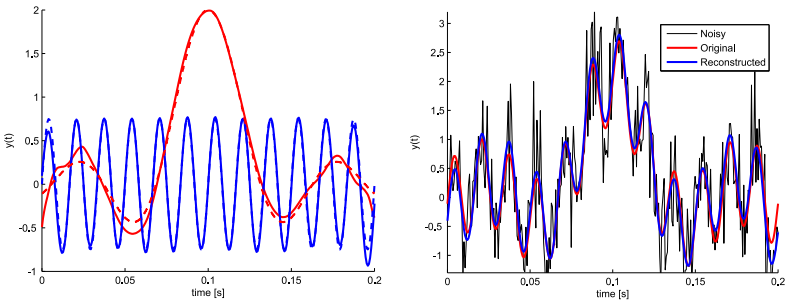
According to the attained results shown in Table 1, it is possible to notice that our approach exhibits suitable AE reconstruction performances with low standard deviation, even against low SNR conditions. Hence, our approach seems to be stable to reconstruct the input time-series, and it is useful as a denoising

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.



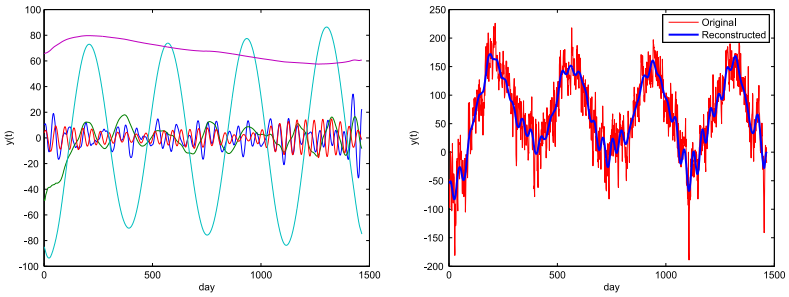
(a) Three-Sin decomposition-SNR=5[dB]. Dashed lines: original components. Continuous lines: estimated components.

(b) Three-Sin reconstruction-SNR=5[dB]



(c) Sin-Sinc decomposition-SNR=5[dB]. Dashed lines: original components. Continuous lines: estimated components.

(d) Sin-Sinc reconstruction-SNR=5[dB]



(e) ECA decomposition $m = 15$

(f) ECA reconstruction $m = 15$

Fig. 1: Some visualization results

tool. Besides, it is possible to notice how our proposed methodology is able to identify the main data structure, fixing the number of required basis m stably, and clustering them to estimate each hidden component. Additionally, as can be seen from Fig. 1(a) and 1(c) the main components of each synthetic dataset are recovered perturbing the input signal with SNR=5[dB]. From the Three-Sin

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

dataset it can be seen how even when each component presents a similar distribution, our approach is able to separate them. The above performance could be explained by the fact that the proposed clustering scheme takes into account the frequency properties of each basis, using the DFT as feature representation criterion. Moreover, the above statement is verified in Fig. 1(b), which presents the reconstruction of the input signal based on the computed components. Now, regarding to the real-world dataset ECA, from Fig. 1(e) it can be noted how the input signal is divided in 5 different components. The component with highest power reveals the cycle structure of the signal, while the remain ones could be related with high frequency changes of the time-series. Finally, based on the selected components, the mean daily temperature is reconstructed as shown in Fig. 1(f), which can be viewed as a free of noise version of the input.

6 Conclusions

An automatic SSA based methodology was proposed to decompose and reconstruct time-series. In this sense, we suggested a clustering based procedure to decompose the main dynamics of the input signal, using a DFT representation to characterize the properties of the SSA orthogonal basis. Moreover, a singular value analysis is employed to conserve the basis related to the main components of the signal and to discard noisy ones. In this approach we assume that the power of each component is enough different in comparison with the others, and higher that the power of the noise. Besides, a rank based method to select the window size of the Hankel matrix in SSA is presented. We tested our approach in two synthetic and one real-world datasets. Attained results showed that our approach is able to identify basis encoding similar data structures, which are employed to infer hidden components of the signal, discarding noise elements. As future work we are interested in decompose time-series in stationary and non-stationary parts using the proposed approach, to support regression and pattern recognition tasks (e.g. biosignal analysis). Furthermore, it would be interesting to incorporate different kind of basis representation, besides the DFT, to deal with more complex hidden components into the signal.

References

[1] Ana Rita Teixeira, Ana Maria Tomé, M. Boehm, Carlos Puntonet, and Elmar Lang. How to apply nonlinear subspace techniques to univariate biomedical time series. *IEEE Transactions on Instrumentation and Measurement*, 58(8):2433–2443, 2009.

[2] N. Golyandina and E. Osipov. The Caterpillar-SSA method for analysis of time series with missing values. *Journal of Statistical Planning and Inference*, 137(8):2642 – 2653, 2007.

[3] Hassani H. and Thomakos D. A review on singular spectrum analysis for economic and financial time series. *Statistics and its Interface*, 3:377–397, 2010.

[4] C. Hansen, J. Nagy, and D. Oleary. *Deblurring Images: Matrices, Spectra, and Filtering*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2006.

[5] A. M. Tank et. al. Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. *International Journal of Climatology*, 22(12):1441–1453, 2002.