Advances in Data Analysis and Classification (2024) 18:409–430
https://doi.org/10.1007/s11634-023-00541-x

REGULAR ARTICLE

An enhanced version of the SSA-HJ-biplot for time series with complex structure

Alberto Silva[1,2] · **Adelaide Freitas**[1,2]

Received: 4 February 2022 / Revised: 16 March 2023 / Accepted: 4 April 2023 /
Published online: 18 April 2023

Abstract
HJ-biplots can be used with singular spectral analysis to visualize and identify patterns in univariate time series. Named SSA-HJ-biplots, these graphs guarantee the simultaneous representation of the trajectory matrix's rows and columns with maximum quality in the same factorial axes system and allow visualization of the separation of the time series components. Structural changes in the time series can make it challenging to visualize the components' separation and lead to erroneous conclusions. This paper discusses an improved version of the SSA-HJ-biplot capable of handling this type of complexity. After separating the series' signal and identifying points where structural changes occurred using multivariate techniques, the SSA-HJ-biplot is applied separately to the series' homogeneous intervals, which is why some improvement in the visualization of the components' separation is intended.

Keywords Structural change detection · Singular spectrum analysis · NIPALS algorithm · Biplots

Mathematics Subject Classification 62H99

1 Introduction

The Biplot method is a multivariate technique that can be useful to visualize some steps of the decomposition of univariate Time Series (TS) using the singular spectrum

✉ Alberto Silva
albertos@ua.pt

Adelaide Freitas
adelaide@ua.pt

1 Department of Mathematics, University of Aveiro, Campus de Santiago, 3810-193 Aveiro, Aveiro, Portugal

2 Center for Research and Development in Mathematics and Applications (CIDMA), University of Aveiro, Campus de Santiago, 3810-193 Aveiro, Aveiro, Portugal

analysis (SSA) method (da Silva and Freitas 2020). The SSA is a powerful technique involving several other methodologies, including classical TS analysis, signal processing, and multivariate statistics. Summarily, the basic version of the method maps the original TS into a Hankel trajectory matrix, whose columns are the so-called lagged vectors of size ℓ (the window length). After, the technique performs a Singular Value Decomposition (SVD) to factorize the trajectory matrix into a summation of 1-rank matrices. These elementary matrices are combined to capture a specific structure in the grouping step. Then, the diagonal averaging step reconstructs the TS from the resulting matrix. For more details, see Elsner and Tsonis (1996), Golyandina et al. (2001), and Hassani and Mahmoudvand (2018).

The resulting eigenvectors and eigenvalues from the SVD step of the SSA allow the graphical representation of relevant characteristics of the TS through HJ-biplots (Galindo-Villardón 1986), which we named the SSA-HJ-biplot method (da Silva and Freitas 2020). The points' position in the SSA-HJ-biplot, the arrows' size, and their location in the factorial axes system can reveal patterns leading to the identification of TS' features (Nieto et al. 2014). However, some care is needed to ensure proper representation through biplots when facing more complex data. For example, structural changes in a TS can make visualization more difficult and interpretation confusing. Thus, prior knowledge about the occurrence of a modification in the TS structure can facilitate the graphical exploratory analysis via SSA-HJ-biplot.

To state the problem, let us consider a univariate TS $Y = (y_1, \ldots, y_n)$ in which the stochastic structure related to Y is said to be strictly stationary. In this case, given $t_1, \ldots, t_k \in \{1, \ldots, n\}$, the joint distribution functions of the random vectors $(y_{t_1}, \ldots, y_{t_k})$ and $(y_{t_1+\tau}, \ldots, y_{t_k+\tau})$ are the same for all adequate integers τ and k. In turn, a weakly stationary structure occurs when the process's first and second-order moments do not depend on t, and the autocovariance between y_t and $y_{t+\tau}$ depends just on the lag τ. On the other hand, perturbations can occur in real data, bringing about modifications on either the mean, the variance, or the autocorrelation structure. Thus, it characterizes the process as nonstationary, and these disturbances provoke structural changes (Kleiber 2018).

Another way to approach the issue is characterizing the TS Y as homogeneous in the sense that, for all t, some linear recurrent formula drives the process such that (Golyandina et al. 2001)

$$y_t = a_1 y_{t-1} + \cdots + a_r y_{t-r}, \tag{1}$$

in which $a_1, \ldots, a_r$ are constant coefficients, and $r < n$ is the dimension of the linear recurrent formula. A TS is heterogeneous when a disturbance results in the linear recurrent formula interruption and, after a short transition period, another one begins to govern the series again. Thus, there are two ways to deal with the structural change detection problem: (i) regarding the heterogeneity or (ii) concerning the transition interval. The latter is also known as a change-point detection problem (Golyandina et al. 2001).

Golyandina et al. (2001) proposed solving the structural change detection problem based on heterogeneity detection and using the SSA method. They created a metric to evaluate the distances between lagged vectors and the trajectory space, i.e., the space

spanned by some eigenvectors of the lag-covariance matrix, determined in different intervals of the series. Moskvina and Zhigljavsky (2003) used a quite similar approach to suggest an application of the SSA to the detection of change points in TS. In both studies, two disjunct intervals (base and test) are taken sequentially from the original series, which initially follows a linear recurrent formula. Then, the associated trajectory matrices are constructed. In case of disturbance, it is expected an increase in the Euclidean distance between the lagged vectors of the trajectory matrix (base) and the subspace generated by the eigenvectors of the lag-covariance matrix (test).

Considering a TS with structural changes, two problems emerge for applying the SSA-HJ-biplot method. First, retaining more principal components to capture such essential characteristics of the series can be necessary. Second, visualizing these characteristics can be more challenging than when the TS is entirely homogeneous. Thus, our primary goal is to refine the exploratory capacity of the SSA-HJ-biplot in heterogeneous time series, applying the technique in its homogeneous intervals to improve its interpretability. To detect the points where the linear recurrent formula is interrupted, we have as a secondary objective the creation of a procedure based on the SSA method to evaluate the occurrence of disturbances.

The paper is organized as follows. Section 2 provides a brief overview of the theoretical background of the SSA-HJ-biplot method. In Sect. 3, a new structural change detection method is proposed to improve the performance of the SSA-HJ-biplot when applied to heterogeneous TS, followed by examples that use synthetic and real data. In Sect. 4, we establish the steps for the SSA-HJ-biplot strengthening. Section 5, the suggested procedure is performed on two real-world TS using the statistical software R (R Core Team 2019). Conclusions are presented in Sect. 6.

2 Brief overview

The SSA-HJ-biplot consists of an exploratory tool for visually inspecting the main characteristics of univariate TS, using the results of both SSA and Biplot methods. First, consider $Y = (y_1, \ldots, y_n)$ a univariate and real-valued TS, and let ℓ be the greatest integer less than or equal to $n/2$ representing the window length, as well as $\kappa = n - \ell + 1$. The SSA embedding step comprises defining Y as κ lagged vectors $\mathbf{x}_1, \ldots, \mathbf{x}_\kappa$, each one of size ℓ, in which

$$\mathbf{x}_j = [y_j \quad \ldots \quad y_{j+l-1}]', \quad 1 \leq j \leq \kappa. \tag{2}$$

These κ lagged vectors form a Hankel matrix $\mathbf{X}$ called trajectory matrix, i.e., $\mathbf{X} = [\mathbf{x}_1 \ldots \mathbf{x}_\kappa]$. Then, $\mathbf{X}$ is decomposed using the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm (Wold 1966). The NIPALS is the forerunner of the Partial Least Squares (PLS) method and was designed to iteratively estimate the principal components of a multivariate data matrix through a sequence of simple ordinary least squares regressions (Esposito Vinzi and Russolillo 2013; Wold et al. 1983). The algorithm decomposes the matrix computing the principal components one by one, with results equivalent to the SVD concerning singular vectors and values. The NIPALS decomposition of $\mathbf{X}$ results in a sum of d matrices of rank 1 in terms of the outer

product of a score vector $\mathbf{t}_i$ and a loading vector $\mathbf{p}_i$, so that

$$\mathbf{X} = \sum_{i=1}^{d} \mathbf{t}_i \mathbf{p}_i', \tag{3}$$

where $d = rank(\mathbf{X})$. The elements of the score vector $\mathbf{t}_i$ correspond to the projections of the sample points in the associated principal component direction. In contrast, each loading in $\mathbf{p}_i$ is the cosine of the angle between the component direction vector and the corresponding variable axis (Geladi and Kowalski 1986). At each iteration, the NIPALS algorithm performs a linear regression of the $\mathbf{X}$ columns on a score vector $\mathbf{t}_i$, resulting in a loading vector $\mathbf{p}_i$. Then, the algorithm runs a linear regression of the $\mathbf{X}$ rows on the loading vector to get a new estimate for $\mathbf{t}_i$. The cycle repeats until it converges according to some criterion (Wold 1966).

The NIPALS algorithm ignores any missing data when executing the regressions, which is equivalent to setting all missing points to zero in the least-squares objective function (Wold et al. 1983). Consequently, the proposed approach can be applied even when missing values are detected in the series without the need to use imputation methods. In addition, to get the results of the NIPALS decomposition equivalent to those of the SVD, one can normalize the score vectors as follows

$$\mathbf{t}_i^* = \frac{\mathbf{t}_i}{||\mathbf{t}_i||} \iff \mathbf{t}_i = \sqrt{\mathbf{t}_i' \mathbf{t}_i}\, \mathbf{t}_i^*. \tag{4}$$

Thus, the decomposition of $\mathbf{X}$ is obtained in terms of its left singular vectors $\mathbf{t}_i^*$, right singular vectors $\mathbf{p}_i$, and singular values $\sqrt{\mathbf{t}_i' \mathbf{t}_i}$ (Esposito Vinzi and Russolillo 2013; da Silva and Freitas 2020). Each one of these NIPALS eigentriple $\left(\sqrt{\mathbf{t}_i' \mathbf{t}_i}, \mathbf{t}_i^*, \mathbf{p}_i\right), i = 1, \ldots, d$, lays down an elementary matrix such that

$$\mathbf{X}_i = \sqrt{\mathbf{t}_i' \mathbf{t}_i}\, \mathbf{t}_i^* \mathbf{p}_i', \tag{5}$$

and

$$\mathbf{X} = \mathbf{X}_1 + \cdots + \mathbf{X}_d. \tag{6}$$

On the other hand, defining $\mathbf{\Sigma}$ as a diagonal matrix containing the singular values $\sqrt{\mathbf{t}_i' \mathbf{t}_i}, i = 1, \ldots, d$, arranged in decreasing order, the matrix form of the decomposition in (3) is

$$\mathbf{X} = \mathbf{T}^* \mathbf{\Sigma} \mathbf{P}', \tag{7}$$

where $\mathbf{T}^*$ is the matrix containing the orthonormal score vectors $\mathbf{t}_i^*$ in its columns, and $\mathbf{P}$ is the matrix whose columns are the orthonormal loading vectors $\mathbf{p}_i$.

The decomposition in (7) allows the assignment of the matrix $\mathbf{\Sigma}$ in different ways to obtain the biplot scheme. Any $\ell \times \kappa$ matrix $\mathbf{X}$ of rank d can be factorized as $\mathbf{X} = \mathbf{G}\mathbf{H}'$, where $\mathbf{G}$ is a $(\ell \times q)$ matrix and $\mathbf{H}$ is a $(\kappa \times q)$ matrix, with $q \leq d$. The matrices $\mathbf{G}$ and $\mathbf{H}$ create two sets of q-dimensional points. If $q = 2$, then the rows and columns of $\mathbf{X}$

can be simultaneously represented in the so-called biplot, in which the rows of $\mathbf{G}$ are reproduced by points, and the columns of $\mathbf{H}'$ are depicted as vectors connected to the origin (arrows). When $q > 2$, the best 2-rank approximation of $\mathbf{X}$, in the sense of least square, is considered. Assuming $\mathbf{G} = \mathbf{T}^*$ and $\mathbf{H} = \mathbf{P}\boldsymbol{\Sigma}$, the resultant factorization is characterized by preserving the column metrics of $\mathbf{X}$. The associated biplot is called Gabriel biplot (Gabriel 1971), later named $\mathbf{GH}'$-biplot in Galindo-Villardón (1986). In this case, the columns are better represented than the rows in terms of quality. On the other hand, by defining $\mathbf{G} = \mathbf{T}^*\boldsymbol{\Sigma}$ and $\mathbf{H} = \mathbf{P}$, this factorization will preserve the metric of the rows in the so-called form biplot, later designated as $\mathbf{JK}'$-biplot in Galindo-Villardón (1986). On it, the Euclidean distances between the row markers approximate the Euclidean distances between the respective individuals in the full space, and the representation of the rows is better than the columns. From this point, consider the matrix $\mathbf{J} = \mathbf{T}^*\boldsymbol{\Sigma}$, and the matrix $\mathbf{H} = \mathbf{P}\boldsymbol{\Sigma}$. Then, the rows and columns of $\mathbf{X}$ can be simultaneously represented with maximum quality through the so-called HJ-biplot (Galindo-Villardón 1986), a 2-dimensional biplot in which the points reproduce the rows of $\mathbf{J}$ (the row markers), and the rows of $\mathbf{H}$ (the column markers) are depicted as vectors connected to the origin.

In turn, the SSA-HJ-biplot provides useful visual information to separate the groups related to the TS components (trend, seasonality, and noise) (da Silva and Freitas 2020). An SSA-HJ-biplot uses two principal components to visualize information about a TS in an integrative way since the row and column markers are displayed simultaneously on the same graph, with maximum representation quality. Each principal component is associated with a TS component and explains a proportion of the variability of the data. A point corresponds to a κ-lagged vector (a row marker), while an arrow designates an ℓ-lagged vector (a column marker). The points are labeled to identify the specific time unit where the κ-lagged vector starts. This improves the interpretation of the SSA-HJ-biplot against the data, which means that the points can represent not only the time unit where the κ-lagged vector starts but also the time unit itself. When the evolution of the projection of the points onto a principal component coincides with its growth, it means the existence of a trend and the association between the latter and the principal component. The corresponding singular value will be dominant in this case. On the other hand, when projections of points with the same label tend to concentrate on small intervals of the factorial axes, it indicates an association between the periodical component of the TS and the related principal components. In this case, the two corresponding singular values will be very close.

To substantially capture the behavior of the TS through the rows and, simultaneously, the columns of $\mathbf{X}$, da Silva and Freitas (2020) proposed a window length $\ell = n/2$, which allows an enhancement in the interpretability of the graphics display. Considering the SSA-HJ-biplot interpretation is based on the proximity of points, the arrow length, and the angle between arrows, complex structures tend to blur the biplot, turning its visual understanding into a challenging task. Next, a segmentation of the TS is suggested as a solution to this problem.

3 Enhancing the SSA-HJ-biplot through structural change detection

3.1 Basics of the proposed structural change detection method

In previous works (Golyandina et al. 2001; Moskvina and Zhigljavsky 2003), the procedure adopted to detect eventual structural changes in a TS using SSA consists of applying a single decomposition method to two different trajectory matrices (base and test) iteratively throughout the series. In each iteration, the distances between some eigenvectors and an appropriate subspace are computed, creating a measure for later comparison. We propose to assess this difference using a distinct approach in this work. The comparison is based on the difference between applying two decomposition methods (one robust and the other ordinary) on the same trajectory matrix. These differences will be more accentuated when there is an eventual change in the direction of some principal components (eigenvectors) in case of interrupting the linear recurrent formula. The main advantage of this strategy over those suggested by Golyandina et al. (2001) and Moskvina and Zhigljavsky (2003) lies in the possibility of interpretation in terms of principal components that the visualization of the results provides. As a drawback, the NIPALS algorithm may eventually present instability in determining the principal components (Miyashita et al. 1990) and achieving convergence (Geladi and Kowalski 1986).

Let $Y = (y_1, \ldots, y_n)$ be a univariate and real-valued TS, and $y_{h+1}, \ldots, y_{h+m}$ be a subseries so that $m < n$ and $h = 0, \ldots, n - m$ (Fig. 1). Based on the SSA method, the following steps describe how to compute the proposed differences.

1. Iteratively, from $h = 0$ to $h = n-m$, for some m previously defined, the respective $\ell \times \kappa$ trajectory matrix $\mathbf{X}^{(h)}$ is constructed as follows, where $1 < \ell \leq m/2$ and $\kappa = m - \ell + 1$:

$$\mathbf{X}^{(h)} = \begin{bmatrix} y_{h+1} & y_{h+2} & \cdots & y_{h+\kappa} \\ y_{h+2} & y_{h+3} & \cdots & y_{h+\kappa+1} \\ y_{h+3} & y_{h+4} & \cdots & y_{h+\kappa+2} \\ \vdots & \vdots & \ddots & \vdots \\ y_{h+\ell} & y_{h+\ell+1} & \cdots & y_{h+m} \end{bmatrix}. \tag{8}$$

2. In each iteration, $\mathbf{X}^{(h)}$ is decomposed in singular values in two different ways. One uses a robust method (hereinafter, the subscript "*rob*"), and the other uses the NIPALS algorithm (hereinafter, the subscript "*nip*"). The robust decomposition method implemented in R in this work is a NIPALS-based adaptation of the one described in (Rodrigues et al. 2018), which is based on the $L1$ norm instead of the frequent least-squares $L2$ norm. The resulting factorization from the two mentioned methods are, respectively:

$$\mathbf{X}^{(h)}_{rob} = \mathbf{T}_{rob}\mathbf{P}'_{rob} = \mathbf{T}^{*}_{rob}\mathbf{\Sigma}_{rob}\mathbf{P}'_{rob}, \tag{9}$$

and

$$\mathbf{X}^{(h)}_{nip} = \mathbf{T}_{nip}\mathbf{P}'_{nip} = \mathbf{T}^{*}_{nip}\mathbf{\Sigma}_{nip}\mathbf{P}'_{nip}. \tag{10}$$

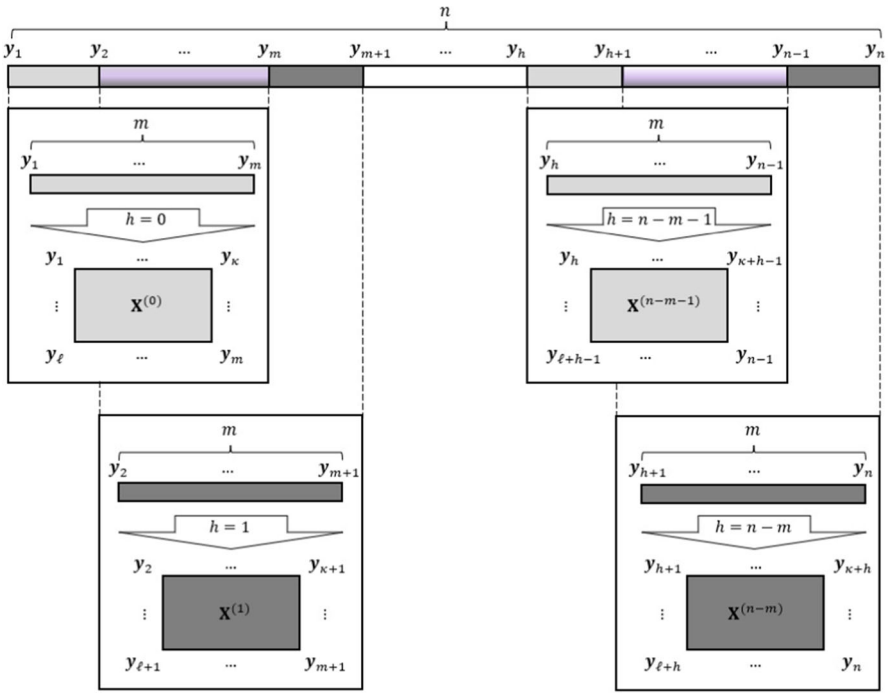

Fig. 1 Segmentation of a TS of length n, and the embedding step of the SSA applied to each subseries of length m

3. Next, for each h, it is computed a matrix formed by the difference between the *nip* and *rob* score matrices, such that

$$\mathbf{Z}^{(h)} = \mathbf{T}^*_{nip}\boldsymbol{\Sigma}_{nip} - \mathbf{T}^*_{rob}\boldsymbol{\Sigma}_{rob}. \tag{11}$$

   The purpose of the $\mathbf{Z}^{(h)}$ matrix is to figure out possible deviations between the homologous principal components provided by a sensitive method and a non-sensitive one concerning outliers.
4. Taking into account the decreasing variability of the 1st to the dth column of the score matrices that generated the $\mathbf{Z}^{(h)}$ matrix, d metrics that cumulatively add more information are introduced. The proposed d metrics $\mathcal{D}_\phi$, $\phi = 1, \ldots, d$, correspond to the Frobenius norm of the matrices

$$[\mathbf{z}_1^{(h)}], \quad [\mathbf{z}_1^{(h)} \quad \mathbf{z}_2^{(h)}], \quad \ldots \quad , \quad [\mathbf{z}_1^{(h)} \quad \mathbf{z}_2^{(h)} \quad \ldots \quad \mathbf{z}_d^{(h)}],$$

   respectively, in which $\mathbf{z}_j^{(h)}$ indicates the jth column of $\mathbf{Z}^{(h)}$, and $d = rank(\mathbf{Z}^{(h)})$. They are given by

$$\mathcal{D}_\phi(h) = \sqrt{\sum_{j=1}^{\phi}\sum_{i=1}^{\ell}(\mathbf{z}_{ij}^{(h)})^2}, \quad \text{for} \quad \phi = 1, \ldots, d, \tag{12}$$

noticing that

$$\mathcal{D}_\phi^2(h) = \mathcal{D}_{\phi-1}^2(h) + \sum_{i=1}^{\ell}(\mathbf{z}_{i\phi}^{(h)})^2. \tag{13}$$

Equation (13) holds for each ϕ as long as, by convention, we have $D_0(h)$ equal to the null function. Besides, the term $\sum_{i=1}^{\ell}(\mathbf{z}_{i\phi}^{(h)})^2$ in the second member of (13) provides information about the structure of the trajectory matrix in each iteration, helping to identify in which Principal Component in the NIPALS decomposition (PC_{nip}) the change of direction occurs.

The parameter m is crucial to adequately capture the change of direction of the PC_{nip} and, consequently, a structural change of the TS. For minimal values of m, the behavior of $\mathcal{D}_\phi$ tends to replicate the signal, while for higher values of m, the structural change can occur inside the first subseries and not be noticed. An optimal value of m would undoubtedly provide a graphical resolution of the $\mathcal{D}_\phi$ curves with the best visual perception of the principal components' direction shifts.

3.2 Graphical assessment of a TS structural change

To evaluate possible structural changes of a TS Y, we propose visually assessing the behavior of $\mathcal{D}_\phi$, $\phi = 1, 2, \ldots, d$, through a simultaneous graphical representation of these d functions. Concretely, let us consider the existence of a structural change in the TS Y at the time point $i = h + m$, i.e., occurring at observation y_{h+m}. In these conditions, we expect a sharp increase of some functions $\mathcal{D}_\phi$ more highlighted for the highest curve starting at the iteration $k = h+1$, that is, at $\mathcal{D}_d(k)$. It is because when the observation y_{h+m} first appears in one of the m-sized subseries of $Y(y_{h+1}, \ldots, y_{h+m})$, it will also be the last element of the trajectory matrix $\mathbf{X}^{(h)}$, causing changes of direction in some principal component when applying the NIPALS decomposition in $\mathbf{X}^{(h)}$, but not when applying the robust method. Consequently, for some positive integer H and some $k \in [h + 1, h + 1 + H]$, larger values of $\mathcal{D}_\phi(k)$ are expected relative to those obtained in previous iterations ($k \leq h$). These differences will be more pronounced when considering the cumulative differences contained in $\mathcal{D}_d$. Then, horizontally, the analysis of the functions $\mathcal{D}_\phi$ focuses on $\mathcal{D}_d$ because it contains the highest cumulative differences in different iterations. Thus, we look for some iteration k such that $\mathcal{D}_d(k)$ presents an elbow, evidencing a marked change in the slope of the curve. On the other hand, the calculation of $\mathcal{D}_d$ at the iteration point $h + m$, for different values of m, can also establish an estimate for m, as described below.

Since $d = d(m)$, i.e., d depends on the dimension of the trajectory matrix $\mathbf{X}^{(i)}$, for some $i = 0, 1, 2, \ldots, n - m$ associated with the series of size n, we first determine the iteration h^* that maximizes the function $\mathcal{D}_d^2$ normalized to $d(m)$ for each m, and

given by

$$\frac{\sum_{j=1}^{d(m)} \sum_{i=1}^{\ell} (z_{ij}^{(h+m)})^2}{d(m)}.$$

Hence,

$$h^* = \arg\max_h \frac{\mathcal{D}_d^2(h+m)}{d(m)}. \tag{14}$$

Thus, h^* corresponds to defining, for a given m, the iteration h where the average increments of $\mathcal{D}_d^2(h+m)$ in (13) are maximums. Since (14) is only dependent on m, the optimal m^* could be estimated by

$$m^* = \arg\max_m \left(\max_h \frac{\mathcal{D}_d^2(h+m)}{d(m)} \right). \tag{15}$$

After identifying the moment of occurrence of the structural change in the series Y, says $i = h + m$, it is essential to know the type of change that occurred at the observation y_{h+m}. One could expect that the principal component related to $\mathbf{X}^{(h)}$ identified as presenting the most major direction change will correspond to the homologous component of the TS Y (trend, periodicity, etc.). The first subseries containing the observation where structural change begins (y_{h+m}) is no longer homogeneous. The subseries will also carry a heterogeneous part of the TS until Y starts obeying a new linear recurrent formula. The increasing input of observations from the heterogeneous interval of the TS makes the decomposition of the trajectory matrices related to these subseries continue to show a structural change until $\mathcal{D}_\phi$ curves reach a peak. Consequently, evaluating the graphs of functions $\mathcal{D}_1, \ldots, \mathcal{D}_{d-1}$, we vertically look for the most remarkable differences among their curves, i.e., higher difference values among $\mathcal{D}_1(k), \ldots, \mathcal{D}_{d-1}(k)$. It is expected that more substantial differences will occur in the curves of the first functions as they reflect the first principal components and carry more information (higher eigenvalues).

3.3 Examples

This subsection evaluates the proposed structural change detection method through three examples. First, a synthetic dataset where occurs two structural changes regarding the periodicity. After, another synthetic dataset with an upward shift in the series. Finally, the Nile database, described in R Documentation (R Core Team 2019) as measurements of the annual flow of the River Nile at Aswan (formerly Assuan), 1871–1970, in $10^8 m^3$.

I—**Synthetic data (disturbance in periodicity)**: The constructed signal contains 151 observations, and there are two change points at the time t_{51} and t_{101}. Below is the R code (Listing 1) used to generate the signal and its graphical representation (Fig. 2):

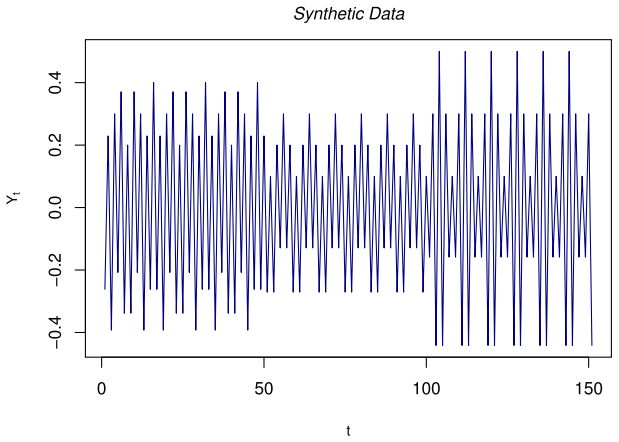

Fig. 2 Signal of the Synthetic data in which there are two structural changes (t_{51} and t_{101})

Listing 1 Code of the synthetic data presenting disturbances in periodicity.

```
Y = numeric()
for(t in 1:50){
   Y[t]   = 0.1*cos(3*pi*t/8) + 0.2*cos(pi*t) + 0.1*cos(7*pi*t)
   }
for(t in 51:100){
   Y[t] = 0.1*cos(pi*t/4) + 0.1*cos(5*pi*t) + 0.1*cos(3*pi*t)
   }
for(t in 101:151){
   Y[t]= 0.2*cos(3*pi*(t/4)) + 0.2*cos(5*pi*t) + 0.1*cos(5*pi*t)
   }
plot(Y, type = "l", col="navy", main = "Synthetic Data", xlab="t",
ylab = expression("Y"[t]), cex.lab = 1.3, cex.main=1.5,font.main=3)
```

In this case, the subseries size's optimal value obtained according to (15) is $m^* = 13$, resulting in a window length $\ell = 7$, and $\kappa = 7$. As the first interruption of the linear recurrent formula takes place in y_{51}, thus is expected an increase in $\mathcal{D}_\phi$ in iteration 39 and following, i.e., from $h = 38$ onwards. Since the second interruption occurs in y_{101}, then $\mathcal{D}_\phi$ should spike at iteration 89 (i.e., $h = 88$). The proposed structural change detection method results are shown in Fig. 3and are following as awaited. Also, there are three lines in the graph because, for $h = 0, \ldots, n - m$, $rank(\mathbf{X}^{(h)}) = 3$. Those graph's lines capture each principal component's contribution in the increase of $\mathcal{D}_\phi$, or in other words, which principal components vary more in direction when the singular value decomposition of the trajectory matrix is not robust.

II—**Synthetic data (upward shift disturbance)**: This example shows a sequence in which $n = 60$ and occurs an upward shift at the time t_{30}. The generated series was based on the patterns presented in (Alcock et al. 1999), with implementation in R summarized in the code below (Listing 2) and a graphical representation in Fig. 4.

Listing 2 Code of the synthetic data presenting an upward shift disturbance.

```
n = 60; Y = numeric(n)
m = 30; s = 2
```

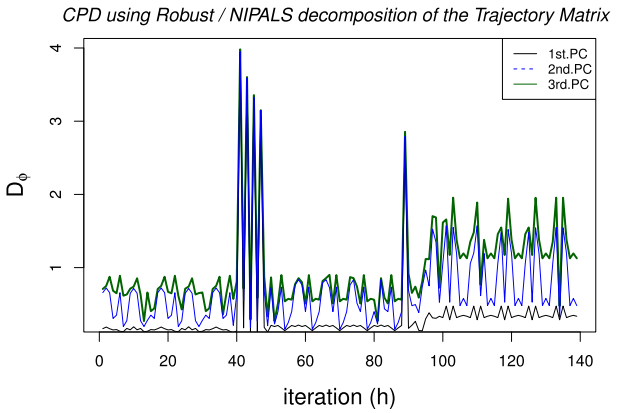

Fig. 3 The spikes in the curves representing $\mathcal{D}_\phi$ suggest two structural changes at observations y_{51} and y_{101}

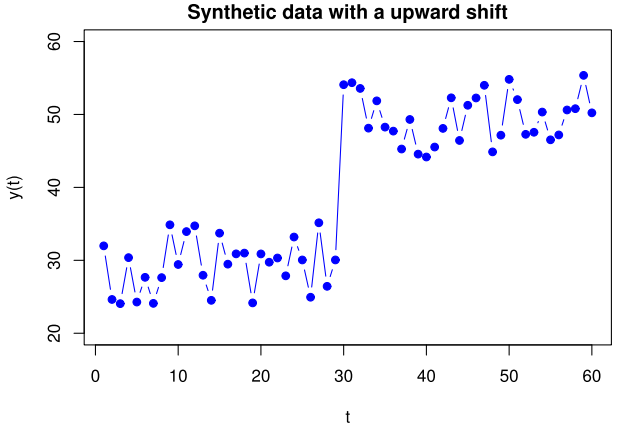

Fig. 4 Synthetic data presents an upward shift disturbance structural at t_{30}

```
r = runif(n, min = −1, max = 1)
x = 10; t3 = 30
for (t in 1:n){
   k = ifelse (t < t3, 0, 1)
   Y[t] = m + s∗r[t] + k∗x
   }
```

According to (15), the optimal subseries length in the second example is $m^* = 20$, following a window length $\ell = 10$ and a $\kappa = 11$. Since the series level moves up in y_{30}, one could await a sharp increment of $\mathcal{D}_\phi$ from iteration 11 onwards ($h = 10$). And that is precisely what Fig. 5 shows since it suggests a structural change in the series from observation y_{30}, as $m + h = 30$. Besides, the curves corresponding to the first two principal components seem to significantly contribute to the $\mathcal{D}_\phi$ increment. In this specific sample, $rank(\mathbf{X}^{(h)}) = \ell, \forall h$. On the other hand, due to the way of construction of the trajectory matrix in the SSA, eventually, $\mathbf{X}^{(h)}$ may not have full

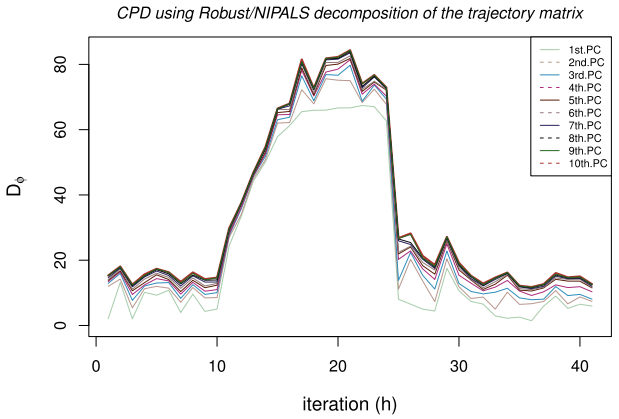

Fig. 5 A sharp increment in the curves representing $\mathcal{D}_\phi$ at iteration 11 (i.e., $h = 10$) suggests a structural change in the series from observation y_{30}. Since $m = 20$, this agrees with the proposed method, as $m + h = 30$

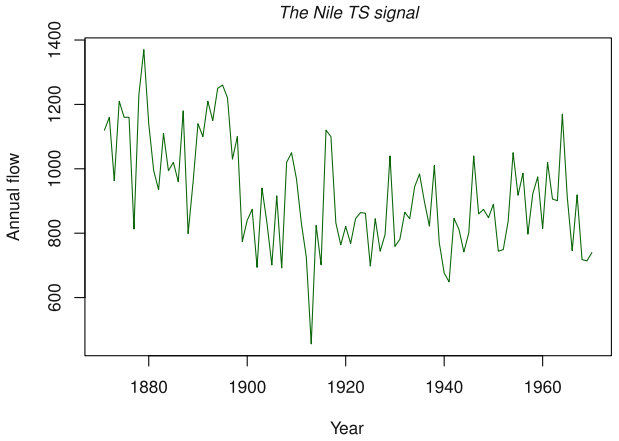

Fig. 6 Signal of the Nile TS, that represents the annual flow of the River Nile from 1871 to 1970

rank in every iteration. If that happens, the number of extracted principal components must be reduced to the lowest computed rank of $\mathbf{X}^{(h)}$, for $h = 0, \ldots, n - m$.

III—**The Nile data**: After separating the signal from the noise using the SSA method, the series looks like it appears in Fig. 6. The literature points out "an apparent change point near 1898" (Cobb 1978), i.e., from the observation y_{29} onwards, the initial linear recurrent formula is no longer in effect. Figure 7 represents the proposed structural change detection method using the optimal value of m^* equals 24 and, therefore, a window length $\ell = 12$.

In this case, all matrices $\mathbf{X}^{(h)}$ are full rank, with 12 rows. In the graph, one can verify that $\mathcal{D}_\phi$ starts to grow with $h = 5$ (iteration 6). Therefore, there is an eventual change point in the observation $y_{h+m} = y_{29}$, following previous literature results. Besides, the graph shows that the first two principal components are the most affected in terms of

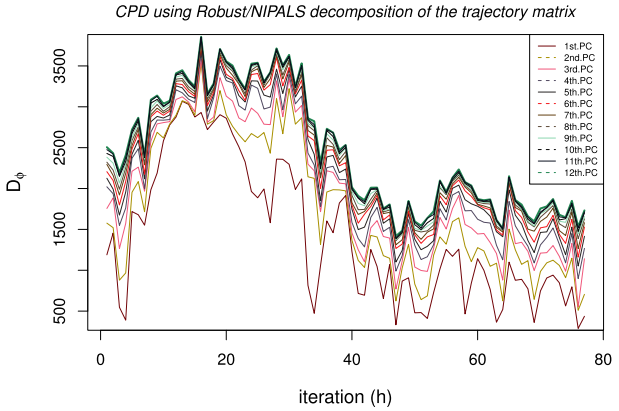

Fig. 7 A change point was detected at iteration 6 (i.e., $h = 5$) after a consistent increase in the values of $\mathcal{D}_\phi$. The subseries' length used was $m = 24$ and, therefore, the change point occurs in the observation $y_{h+m} = y_{29}$

change of direction and, thus, are the ones that most contribute to the increase in $\mathcal{D}_\phi$. Therefore, this is relevant information and should be considered when interpreting the SSA-HJ-biplot.

4 Strengthening the SSA-HJ-biplot

The SSA-HJ-biplot on any univariate TS (homogeneous or heterogeneous) will be helpful if the interpretability of its elements related to the decomposition of a TS is visually highlighted. This section brings an enhanced version of the technique application, adding extra steps for seeking structural change points at the TS. Therefore, the analysis of the global characteristics of the TS is based on the inspection of homogeneous subseries. In this sense, a method for detecting interruptions in the linear recurrent formula was presented in Sect. 3.1, seeking to improve the SSA-HJ-biplot approach. As stated before, the objective here is to increase the range of cases in which the SSA-HJ-biplot technique is suitable for separating TS components. Thus, the following steps are performed preliminarily in the case of a heterogeneous TS.

1. First, to increase the detection performance in the next step 2, a first round of the SSA-HJ-biplot is applied to the entire TS to separate the signal from the noise, followed by the series' reconstruction concerning the signal.
2. Then, the structural change detection method proposed in Sect. 3.1 is applied to the reconstructed time-series signal to identify the observations y_i in which an interruption of the linear recurrent formula is supposed to occur. This step separates the TS into homogeneous subseries between the change points.
3. Finally, the SSA-HJ-biplot is performed and interpreted in each homogeneous interval, that is, between change points.

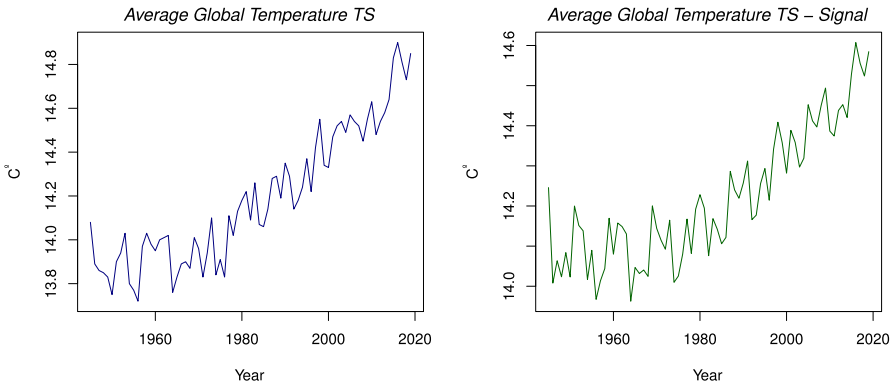

Fig. 8 TS of the average global surface temperatures plus anomalies, from 1945 to 2019. The original data is represented on the left and the series signal on the right

5 Applications

The enhanced SSA-HJ-biplot technique was applied to two real climate time series to assess the method comprehensively.

Case 1

The first TS (Fig. 8—left), referring to the period from 1945 to 2019, was obtained from the National Oceanic and Atmospheric Administration (NOAA 2020) website by adding the twentieth-century global mean temperature ($13.9°C$) to the Earth's surface temperature anomalies, defined as the difference between the measured Sea Surface Temperature (SST) and the average temperature for a certain period (Yang et al. 2018). When applied to the entire series, the SSA-HJ-biplot results in the biplots in Fig. 9. For now, note that using the SSA-HJ-biplot interpretation rules, one can only identify a growing global trend, and nothing can be concluded about the periodicity using the graph in Fig. 9.

Next, taking advantage of the SSA's grouping step, the signal was filtered using the first three eigentriples and reconstructed the corresponding TS through the diagonal averaging step, represented in Fig. 8—right. In the present case, the visual perception of the principal components' direction shifts occurs for an optimal value of $m^* = 5$, resulting in $\ell = \kappa = 3$. It means that the trajectory matrix in each iteration is a square matrix of order 3. Also, $\mathbf{X}^{(h)}$ is a full-rank matrix for all h and, therefore, there must exist three lines to represent each component's contribution to the increment of $\mathcal{D}_\phi$. In Fig. 10, it can be seen that $\mathcal{D}_\phi$ begins to grow rapidly in iteration 17, i.e., when $h = 16$. Therefore, all of that suggests that an eventual linear recurrent formula was interrupted around 1966 (observation y_{21}) since $h + m = 21$ in this case. Besides, Fig. 10 shows the most marked change of direction occurs in the extraction of the 2nd principal component.

Knowing that an eventual modification in the TS structure occurred in observation y_{21}, the series is then segmented in the intervals 1945–1965 and 1966–2019 to build

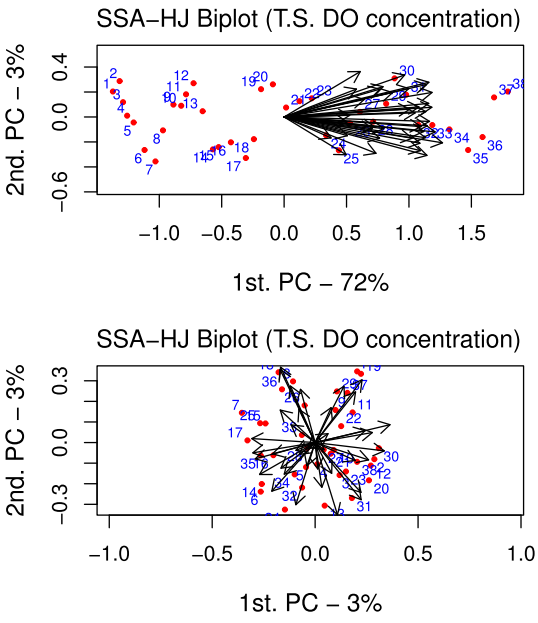

Fig. 9 The original approach of the SSA-HJ-biplot concerning the TS of average temperatures on the surface of the globe from 1945 to 2019. The graph at the top refers to the 1st and 2nd principal components, followed by the SSA-HJ-biplot of the 2nd and 3rd principal components

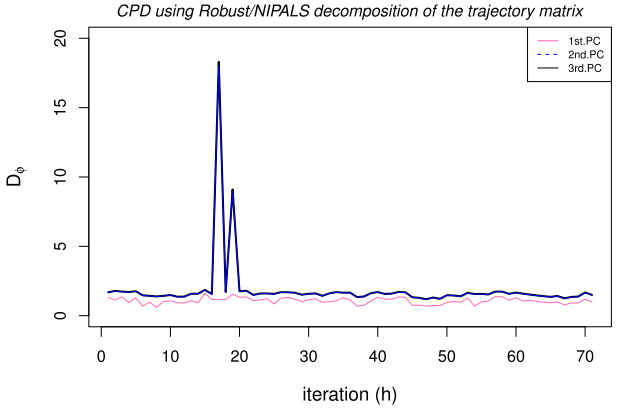

Fig. 10 The graph points out the values of $\mathcal{D}_\phi$ in each iteration according to each principal component's contribution. In this case, there is a substantial increase in iteration 17 (or $h = 16$). This means the occurrence of a change point in the observation y_{21}, since $m = 5$

the SSA-HJ-biplots, aiming to improve the visualization and facilitate the graphic interpretation. Following the SSA-HJ-biplot approach, one can set labels to the biplot points according to the year each of the κ-lagged vectors starts. Thus, tag "1" indicates that the first κ-lagged vector (first row of the trajectory matrix) begins in the year 1945 (or 1966), "2" indicates the year 1946 (or 1967) as the starting year of the second κ-lagged vector, and so on. Eventually, one could use this approach to label the biplot arrows instead of the points and make the tags indicate each ℓ-lagged vector (columns of the trajectory matrix). Figure 11 shows the SSA-HJ-biplot of the TS for the 1st and

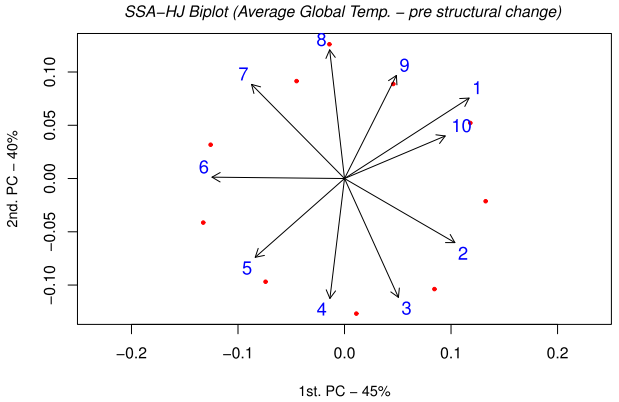

Fig. 11 SSA-HJ-biplot (1st and 2nd principal components) of the Average Global Temperature TS regarding the interval before the estimated structural modification (1945–1965)

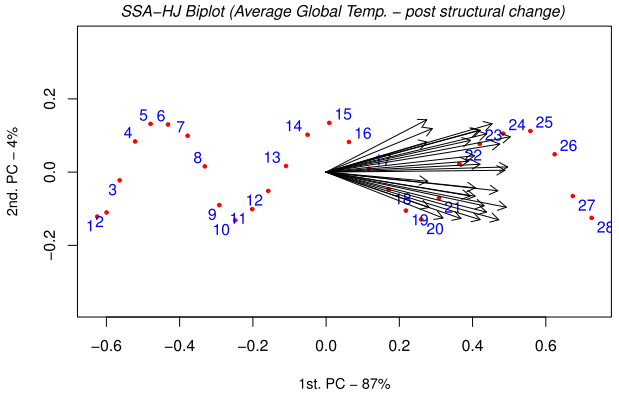

Fig. 12 SSA-HJ-biplot (1st and 2nd principal components) of the Average Global Temperature TS regarding the interval after the estimated structural modification (1966–2019)

2nd principal components concerning the interval immediately before the estimated structural change, which comprises 1945–1965. The first two principal components together explain about 85% of the data variability. According to the SSA-HJ-biplot interpretation (da Silva and Freitas 2020), the graph does not indicate a trend in this section since the points (red tags, from 1 to 10) do not grow in the same direction as any component. On the other hand, the circular pattern suggests some periodicity, but the lack of enough observations prevents a more comprehensive interpretation of this interval.

Figures 12 and 13 show the SSA-HJ-biplot concerning the interval after the structural change, which comprises the years 1966 to 2019 and $n = 55$, $\ell = \kappa = 28$. Figure 12 shows the SSA-HJ-biplot regarding the 1st and 2nd principal components and explains 91% of the data variability. The biplot points projections in the 1st principal component evolve in the same growth direction as that component, which means that the 1st principal component is associated with a crescent trend.

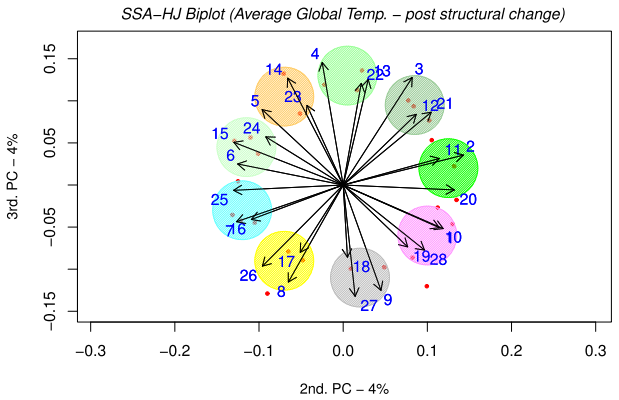

Fig. 13 SSA-HJ-biplot (2nd and 3rd principal components) of the Average Global Temperature TS regarding the interval after the estimated structural modification (1966–2019)

In turn, the biplot points' projection on the 2nd principal component reveals a pattern in terms of proximity. For example, the projections of the peak points are always within the same vicinity. It indicates the correspondence between the 2nd principal component and the periodicity of the TS. However, the periodicity is always associated with a pair of principal components (da Silva and Freitas 2020) and, therefore, another SSA-HJ-biplot is required. The SSA-HJ-biplot for the 2nd and 3rd principal components is represented in Fig. 13. Since the 2nd and 3rd principal components explain only 8% of the data variability, it is preferable to label the arrows to reveal the periodicity using the angles between them to search for the most positively correlated ℓ-lagged vectors. Even considering the low percentage of explained variability, the proximity pattern between the arrows suggests a periodicity of 9 years in this interval.

Case 2

The second TS refers to the average annual precipitation in Brazil from 1901 to 2021, obtained in the World Bank Group Climate Change Knowledge Portal. Figure 14 shows the TS filtered after retaining the components explaining more than 1% of the data variability. Figure 15 brings up the SSA-HJ-biplot of the entire TS, and Table 1 shows normalized $\mathcal{D}_d^2$ for several values of m.

Except for the evident increase in the variability of the data expressed by the variation in the size of the arrows, little information can be extracted from the biplot representation in Fig. 15. Then, we compute the normalized $\mathcal{D}_d^2$ to determine the m for which the function (14) is maximum. For convenience, Table 1 shows just a few values of them. Therefore, the optimal value for the size of the subseries will be $m^* = 20$, with which it will be possible to better perceive the changes in the direction of the components by plotting the curves $\mathcal{D}_\phi$.

Figure 16 suggests the existence of three change points from 1901 to 2021. As $m = 20$ and the iterations where the curve $\mathcal{D}_d$ start to grow rapidly are those regarding to $h = 5$, $h = 66$, and $h = 94$, then the observations of interest are y_{25} (1925), y_{66}

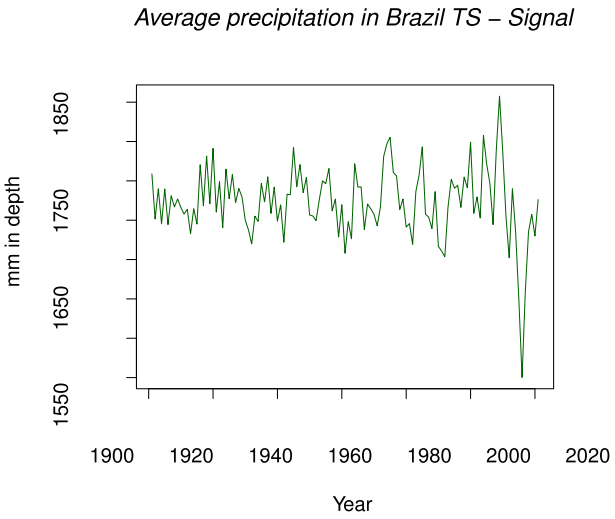

Fig. 14 TS of the average annual precipitation in Brazil, from 1901 to 2021. The data represents the series' signal

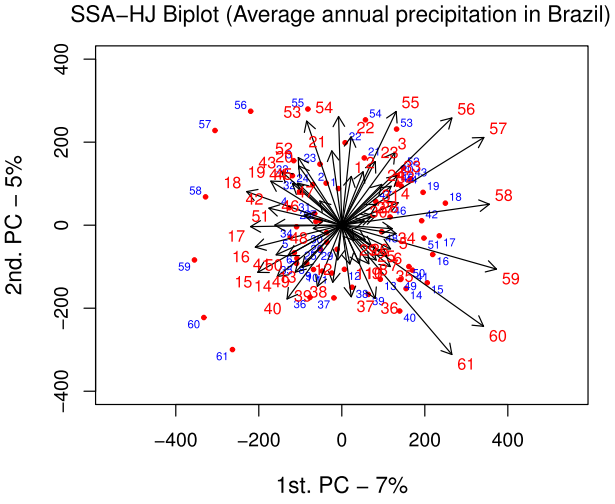

Fig. 15 The original approach of the SSA-HJ-biplot regarding the TS of average annual precipitation in Brazil from 1901 to 2021

(1966), and y_{114} (2014). As we have a few observations before the first change point and after the last one, we are interested in constructing the SSA-HJ-biplots for the intervals between 1925–1965 e 1966–2014.

In Figs. 17and 18, the SSA-HJ-biplot for the 1st and 2nd principal components for both intervals are presented. No trend was detected in the subseries from 1925 to 1965 (Fig. 17). In addition, the decrease and increase in the size of the arrows indicate some variability in the TS, with the pattern of proximity among arrows suggesting a periodicity of around ten years. Regarding Fig. 18, the circular pattern again indicates the

Table 1 The optimal value of m obtained from the maximum normalized $\mathcal{D}_d^2$

m	max $\mathcal{D}_d^2/d(m)$	m	max $\mathcal{D}_d^2/d(m)$
17	27685.6	27	23694.4
18	29022.5	28	25959.6
19	27915.7	29	22797.9
20	**31880.1**	30	22873.5
21	23725.2	31	19802.1
22	22915.9	32	21482.7
23	27838.9	33	18017.6
24	27939.1	34	18189.9
25	21397.8	35	18528.6
26	20981.5	36	20036.4

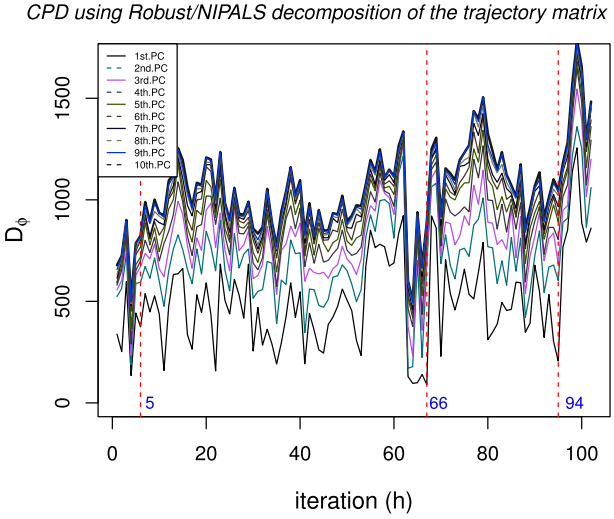

Fig. 16 For $m = 20$, the graph suggests the existence of three change points since $\mathcal{D}_d$ grows fast after the iterations 6 ($h = 5$), 67 ($h = 66$), and 95 ($h = 94$)

absence of a trend. The difference in the sizes of the arrows likewise shows variability in the data, but no periodicity is suggested in the subseries from 1966 to 2014.

6 Conclusion

This paper proposes an improved version of the SSA-HJ-biplot visualization method, intending to enlarge its applicability to univariate time series with more complex structures, especially when a structural change occurs. A simple approach based on multivariate techniques was performed to identify TS structural changes, preliminarily effective in the analyzed series. As usual in the SSA-HJ-biplot, the application of NIPALS ($\mathbf{X} = \mathbf{TP}'$) prevails over the SVD ($\mathbf{X} = \mathbf{UDV}'$) method to decompose the

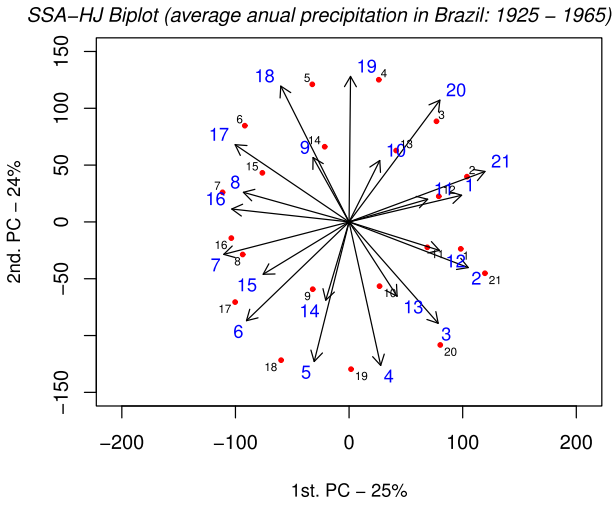

Fig. 17 SSA-HJ-biplot (1st and 2nd principal components) of the Average Annual Precipitation in Brazil TS regarding the interval between the two first change points (1925–1965)

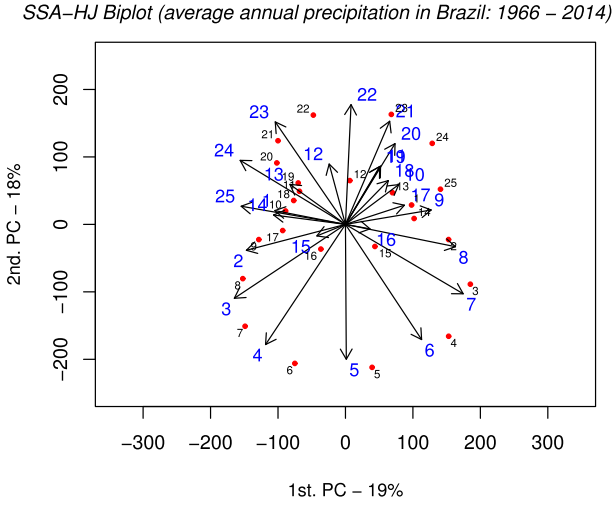

Fig. 18 SSA-HJ-biplot (1st and 2nd principal components) of the Average Annual Precipitation in Brazil TS regarding the interval between the second and third detected change points (1966–2014)

trajectory matrices since it allows dealing with missing data without needing any imputation. This substitution is possible because the matrices **T** (scores matrix) and **UD** (the product of the left singular vectors matrix times the singular values matrix) are equivalent, as well as **P** (loadings matrix) and **V** (right singular vectors matrix). Regarding the proposed structural change detection method, the procedure could recognize the boundaries of homogeneous intervals in three series analyzed. The method correctly pinpointed the moments when the linear recurrent formula interruptions occurred in

the two synthetic series containing previously established structural changes (Examples I and II). The same success was obtained using real data (Example III—The Nile River), where the change point is well-known in the literature. The applications in Sect. 5 showed the effectiveness of the proposed method. In Case 1, the suggested procedure allowed segmenting a real climate time-series data into two homogeneous subseries. The modified method proved useful in confirming the TS' structural change since it identified the absence of a trend in the first interval (1945–1965), in contrast to what occurred in the second (1966–2019). Besides, the second version of the SSA-HJ-biplot also captured a 9-year periodic component in the second interval (1966–2019). As the first interval is short, it was impossible to recognize the existence of periodicity by analyzing only the SSA-HJ-biplot in Fig. 11. However, the periodicity identified in the subsequent interval and the small angle formed by arrows 1 and 10 in Fig. 11 insinuates nine years for the entire TS, agreeing with what is stated in other studies (Keeling and Whorf 1997), which claim an approximately decadal periodicity in surface air temperature from 1945 onwards. In Case 2, we focused on showing how we can estimate the size of the subseries, the parameter (m), in the procedure for detecting structural changes in the analyzed series. In addition, the proposed approach handled a more complex series, segmenting the TS in four intervals. From the SSA-HJ-biplots analysis, the results suggest that the rainfall pattern in Brazil has changed on at least three occasions, becoming more irregular from 1966 onwards, accentuating after 2014. Therefore, these two cases illustrate an improvement in the modified method.

Acknowledgements This research is part of the first author's Ph.D. project, which is carried out at the University of Aveiro. The authors were supported by the Center for Research and Development in Mathematics and Applications (CIDMA) through the Portuguese Foundation for Science and Technology (FCT Fundação para a Ciência e a Tecnologia (FCT), within project UID/MAT/04106/2020 (CIDMA).

Availability of data and materials Data is publicly available with reference in the manuscript.

Code availability Code for the analysis in this manuscript can be found at: https://github.com/albertoosilva/ssa-hj-biplot.

Declarations

Conflict of interest Both authors declare that they have no conflicts of interest.

References

Alcock RJ, Manolopoulos Y, et al (1999) Time-series similarity queries employing a feature-based approach. In: 7th Hellenic conference on informatics, pp 27–29

Cobb GW (1978) The problem of the nile: conditional solution to a changepoint problem. Biometrika 65(2):243–251

da Silva AO, Freitas A (2020) Time series components separation based on singular spectral analysis visualization: an HJ-biplot method application. Stat Optim Inf Comput 8(2):346–358

Elsner JB, Tsonis AA (1996) Singular spectrum analysis: a new tool in time series analysis. Springer, Berlin

Esposito Vinzi V, Russolillo G (2013) Partial least squares algorithms and methods. Wiley Interdiscip Rev Comput Stat 5(1):1–19

Gabriel KR (1971) The biplot graphic display of matrices with application to principal component analysis. Biometrika 58(3):453–467

Galindo-Villardón MP (1986) Una alternativa de representación simultánea: Hj-biplot. Qüestiió pp 13–23

Geladi P, Kowalski BR (1986) Partial least-squares regression: a tutorial. Anal Chim Acta 185:1–17
Golyandina N, Nekrutkin V, Zhigljavsky AA (2001) Analysis of time series structure: SSA and related techniques. CRC Press, Cambridge
Hassani H, Mahmoudvand R (2018) Singular spectrum analysis: using R. Springer, Berlin
Keeling CD, Whorf TP (1997) Possible forcing of global temperature by the oceanic tides. Proc Natl Acad Sci 94(16):8321–8328
Kleiber C (2018) Structural change in (economic) time series. Springer, Cham, pp 275–286. https://doi.org/10.1007/978-3-319-64334-2_21
Miyashita Y, Itozawa T, Katsumi H et al (1990) Comments on the nipals algorithm. J Chemom 4(1):97–100
Moskvina V, Zhigljavsky A (2003) An algorithm based on singular spectrum analysis for change-point detection. Commun Stat Simul Comput 32(2):319–352
Nieto AB, Galindo MP, Leiva V et al (2014) A methodology for biplots based on bootstrapping with R. Rev Colomb Estad 37(2):367–397
R Core Team (2019) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/
Rodrigues PC, Lourenço V, Mahmoudvand R (2018) A robust approach to singular spectrum analysis. Qual Reliab Eng Int 34(7):1437–1447
Wold H (1966) Nonlinear estimation by iterative least squares procedures In: David FN (hrsg.) Festschrift for j. Neyman: Research Papers in Statistics, London
Wold S, Albano C, Dunn III W, et al (1983) Pattern recognition: finding and using regularities in multivariate data food research, how to relate sets of measurements or observations to each other. In: Martens H, Russwurm Jr H (eds) Food research and data analysis: proceedings from the IUFoST symposium, September 20–23, 1982, Oslo, Norway. Applied Science Publishers, London
Yang B, Emerson SR, Peña MA (2018) The effect of the 2013–2016 high temperature anomaly in the subarctic northeast pacific (the “blob”) on net community production. Biogeosciences 15(21):6747–6759