Universidade de Aveiro
2023

Alberto
Oliveira da Silva

Contribuições Multivariadas na Decomposição de uma Série Temporal

Multivariate Contributions in the Decomposition of a Time Series

Universidade de Aveiro
2023

Alberto
Oliveira da Silva

Contribuições Multivariadas na Decomposição de uma Série Temporal

Multivariate Contributions in the Decomposition of a Time Series

Tese apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Doutor em Matemática, realizada sob a orientação científica da Doutora Adelaide de Fátima Baptista Valente Freitas, Professora Associada do Departamento de Matemática da Universidade de Aveiro.

To my wife, Danielli.

o júri / the jury

presidente / president

Doutora Susana Isabel Barreto de Miranda Sargento
Professora Catedrática da Universidade de Aveiro

vogais / examiners committee

Doutor Paulo Canas Rodrigues
Professor and Head of the Department of Statistics of the Federal University of Bahia, Brazil

Doutora Adelaide de Fátima Baptista Valente Freitas
Professora Associada da Universidade de Aveiro (orientadora)

Doutora Isabel Maria Marques da Silva Magalhães
Professora Auxiliar da Universidade do Porto

Doutor Pedro Filipe Pessoa Macedo
Professor Auxiliar da Universidade de Aveiro

Doutora Susana Luísa da Custódia Machado Mendes
Professora Adjunta da Escola Superior de Turismo e Tecnologia do Mar – Politécnico de Leiria

agradecimentos / acknowledgements

I want to thank all my Professors in the Department of Mathematics at the University of Aveiro, especially my supervisor and advisor, Professor Adelaide Freitas, who granted me the honor of being her student. I also thank the Center for Research and Development in Mathematics and Applications (CIDMA) for supporting this research through the Portuguese Foundation for Science and Technology (FCT Fundação para a Ciência e a Tecnologia (FCT), within project UID/MAT/04106/2020 (CIDMA).

Palavras Chave

Séries Temporais, Análise Espectral Singular, *Biplots*, Algoritmo NIPALS, Decomposição em Valores Singulares, Análise de Componentes Principais.

Resumo

Um dos objetivos da análise de séries temporais é extrair características essenciais da série para fins exploratórios ou preditivos. A Análise Espectral Singular (SSA) é um método utilizado para esse fim, transformando a série original em uma matriz de Hankel, também chamada de matriz trajetória. O seu único parâmetro é o chamado comprimento da janela. A decomposição em valores singulares da matriz trajetória permite a separação das componentes da série, uma vez que a estrutura em termos de valores e vetores singulares está de alguma forma associada à tendência, componente oscilatória e ruído. Por sua vez, a visualização das etapas daquele método é pouco explorada ou carece de interpretabilidade. Neste trabalho, aproveitamos os resultados de uma particular decomposição em valores singulares através do algoritmo NIPALS para implementar uma exibição gráfica das componentes principais usando HJ-biplots, nomeando-o método SSA-HJ-biplot. Trata-se de uma ferramenta de natureza exploratória e cujo principal objetivo é aumentar a interpretabilidade visual da SSA, facilitando o passo de agrupamento e, consequentemente, identificar características da série temporal. Ao explorar as propriedades dos HJ-biplots e ajustar o comprimento da janela para a metade do comprimento série, linhas e colunas da matriz trajetória podem ser representadas em um mesmo SSA-HJ-biplot simultaneamente e de maneira ótima. Para contornar o potencial problema de mudanças estruturais na série temporal, que podem dificultar a visualização da separação das componentes, propomos uma metodologia para a detecção de *change points* e a aplicação do SSA-HJ-biplot em intervalos homogéneos, ou seja, entre *change points*. Essa abordagem de detecção é baseada em mudanças bruscas na direção das componentes principais, que são avaliadas por uma métrica de distância criada para esse fim. Por fim, desenvolvemos um outro método de visualização baseado na SSA para estimar as periodicidades dominantes de uma série temporal por meio de padrões geométricos, ao que chamamos SSA Área biplot. Nesta parte da investigação, implementámos em R um pacote chamado *areabiplot*, disponível na Comprehensive R Archive Network (CRAN).

Keywords Time Series, Singular Spectrum Analysis, Biplots, NIPALS Algorithm, Singular Value Decomposition, Principal Component Analysis

Abstract

One of the goals of time series analysis is to extract essential features from the series for exploratory or predictive purposes. The SSA is a method used for this intent, transforming the original series into a Hankel matrix, also called a trajectory matrix. Its only parameter is the so-called window length. The decomposition into singular values of the trajectory matrix allows the separation of the series components since the structure in terms of singular values and vectors is somehow associated with the trend, oscillatory component, and noise. In turn, the visualization of the steps of that method is little explored or lacks interpretability. In this work, we take advantage of the results of a particular decomposition into singular values using the NIPALS algorithm to implement a graphical display of the principal components using HJ-biplots, naming the method SSA-HJ-biplot. It is an exploratory tool whose main objective is to increase the visual interpretability of the SSA, facilitating the grouping step and, consequently, identifying characteristics of the time series. By exploring the properties of the HJ-biplots and adjusting the window length to half the series length, rows and columns of the trajectory matrix can be represented in the same SSA-HJ-biplot simultaneously and optimally. To circumvent the potential problem of structural changes in the time series, which can make it challenging to visualize the separation of the components, we propose a methodology for the detection of *change points* and the application of the SSA-HJ-biplot in homogeneous intervals, that is, between *change points*. This detection approach is based on sudden changes in the direction of the principal components, which are evaluated by a distance metric created for this purpose. Finally, we developed another visualization method based on SSA to estimate the dominant periodicities of a time series through geometric patterns, which we call the SSA Biplot Area. In this part of the research, we implemented a package in R called *areabiplot*, available on the Comprehensive R Archive Network (CRAN).

Contents

List of Figures

Part I

Introduction

Chapter 1

Overview

The linear methods used in Multivariate Statistical Analysis are typically applied to i) summarize the data; ii) explore data structures and patterns; iii) examine and explain the relationship between parts of the data; iv) make decisions and make inferences based on data. When it comes to investigating the structure of datasets, mainly those with high dimensionality, Principal Component Analysis (PCA) stands out for being a dimension reduction tool and, at the same time, having an exploratory nature [47].

In short, PCA applies to datasets in which rows and columns represent individuals and quantitative variables, respectively. One can consider the data matrix as either a set of rows or columns, which allows studying individuals or variables according to object choice. PCA considers the correlation matrix of a set of variables and computes the eigenvalues representing each principal component's variance when performing the data dimension reduction. Hence, the method yields a smaller group of relevant and uncorrelated components.

PCA applications have a broad spectrum, comprising biology and medicine to climate and computer science fields. For example, data related to cell gene expression patterns [108], magnetic resonance imaging [6] and computed tomography imaging [28], climate change indicators based on temperature and precipitation [96], and face recognition [81]. In all of them, the space of variables is usually high-dimensional and favors the use of the PCA method.

Although several algorithms are available to calculate the principal components, the singular value decomposition (SVD) method is the most used approach for this purpose. In this context, the SVD decomposes a high-dimensional dataset into factors that most explain the variability of the data in terms of singular vectors and singular values. An alternative and iterative way to extract the principal components is the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm. NIPALS can be seen as a simple ordinary least squares (OLS) regression sequence and can be applied even in case of missing data [22]. Besides, it is faster than SVD [71] when it's used for large datasets.

After obtaining and selecting the principal components of interest, visualizing and exploring the reduced variable space is crucial. Visualizing the principal components with the highest percentages of explained variability allows us to perceive the essential constitution of the original data, as they preserve the maximum amount of information in their reduced form. The graphic display of this underlying structure intensifies the researcher's perception in a subjective rather than a quantitative way. Furthermore, the visualization can facilitate the assimilation of what is represented, working as a complement to the coldness of the numbers [56]. In this aspect, Biplot methods reinforce the results of the PCA analysis. The PCA biplot, henceforward called just biplot, is a multivariate exploratory visualization technique that simultaneously represents individuals (by points) and variables (by arrows) at the same graphic, providing an approximation for the data matrix elements through the projection of the points onto the arrows [24].

Multivariate Analysis is sometimes combined with other techniques to create new methods. An example of this combination is the Singular Spectrum Analysis (SSA), a methodology used in Time Series Analysis for many different purposes, such as exploratory inspection and forecasting. The SSA objective is to decompose an original time series (TS) into a summation of some interpretable components, e.g., a slowly varying trend, oscillatory components, and an irregular component, or noise [39]. The method transforms a univariate TS into a trajectory matrix with a Hankel format, specifying a window length as the only parameter. Unlike the classical multivariate data matrix, the rows and columns of the trajectory matrix (lagged vectors) represent subseries of the original TS instead of individuals and variables. Aiming to estimate the spectral structure of the TS, the trajectory matrix is then factorized using the SVD, resulting in a summation of elementary matrices of rank one [38]. A relevant issue of the SSA method is how to group and add these rank-one elementary matrices to separate the components of the TS.

1.1 Motivating reasons

When it comes to multivariate analysis, extracting meaningful information from raw data or a method's direct result can take some effort and experience. In turn, multivariate data visualization techniques harness human visual perception capabilities to facilitate users to identify hidden dependencies and correlations. A practical effect of the graphical representation of data is that it produces a different and brand-new informational reality in the observer's mind [109].

On the other hand, it is necessary to guarantee some interpretability to the graphic representation so that the recipient can extract some knowledge, incorporating the visual structure into the thought itself [111]. When the problem involves reducing the dimensionality of the data, as in PCA, the retention of all factors extracted makes interpretation either impractical or unproductive since many of them are irrelevant to the method used [100]. In this sense, the choice of the visualization method in the circumstances of a multivariate scenario requires some care to explore the results obtained both comprehensively and efficiently.

In the context of the SSA method, after the SVD decomposition of the trajectory matrix, the visualization techniques employed are mainly used to identify the groups of rank-one matrices that will separate the additive components of the TS. The approaches proposed in the literature are limited to scatter plots of pairs of eigenvectors, the scree plot of eigenvalues, and a correlation matrix based on reconstructed series. Although these procedures work in practice, they lack interpretability and a formal basis [45]. The use of biplot methods is proposed in the present study to fill this gap.

Biplots are a valuable visualization tool for exploratory analysis since allowing a structural evaluation of a high-dimensional data matrix. When applied to the results of PCA, the method permits an assessment in an interpretable way of i) the similarity between individuals; ii) the variance of the variables; iii) the correlation between two variables [24]. The geometric properties of the scalar product between rows and columns of lower-rank matrices resulting from the SVD decomposition are the background for interpreting classical biplots [77]. The biplot methods have a consolidated theoretical basis with several variations, notably those proposed by Gabriel [24], Gower & Hand [42], Galindo [25], and Gower, Groenen & van de Velden [41].

The results of the trajectory matrix factorization carried out in the SSA decomposition stage favor the application of biplot methods. However, the challenge is to build the bases for the graphic interpretation of what is shown since the eigenstructure no longer represents individuals and variables but intervals of the original series.

1.2 Research aims and novel contributions

The main objective of this thesis is to develop new exploratory methods to assist in decomposing and analyzing a TS employing multivariate data visualization techniques. The work is based on the eigenstructure revealed in the SSA decomposition stage.

Besides, this investigation explores the existing correspondence between the components of a TS and the principal components extracted from the respective trajectory matrix, seeking to demonstrate this association using the biplot approach elaborated throughout the research.

In the context of SSA, one can choose the window length as half the size of the series to capture as many of its characteristics as possible. Then the corresponding trajectory matrix will reveal a Hankel structure in which the rows and columns are equal or nearly equal. Therefore, to represent the rows and columns simultaneously and with maximum quality, an appropriate biplot must be chosen. The so-called HJ-biplots [25] meet this criterion, allowing the interpretation of the lagged vectors under the perspective of both rows and columns in the same graph. Following this premise, the first innovative contribution presented in this thesis refers to constructing the SSA-HJ-biplot (Chapter 4). This new multivariate visualization approach assists in the grouping step while providing interpretability to the principal components in terms of the TS analysis. Subsequently, we sought to expand the meaning of what can be visually extracted from the SSA-HJ-biplot, clarifying the details that help in the interpretation (Chapter 5).

On the other hand, TS complexity can become the visual representation of the trajectory matrix eigenstructure challenging. In this particular, identifying structural breaks in the TS allows its segmentation and the construction of biplots between change points, making them easier to interpret. Some of these disturbances are trend reversals, shifts in the TS levels, variability increase, and frequency alteration. Hence, another contribution of this investigation is developing an approach that explores sudden steering corrections of the principal components to identify points of structural changes (Chapter 6). This original technique improves the capabilities of the SSA-HJ-biplot, allowing the TS decomposition over homogeneous intervals, and guaranteeing more visual interpretability.

Regarding a more specific purpose, a third novel contribution refers to creating a graphical tool to explore the oscillatory components, taking advantage of the autocorrelation between eigenvectors extracted from the trajectory matrix (Chapter 7). This visualization procedure intends to determine essential characteristics related to periodicity based on the idea that the perception of structural similarities suits the human ability to identify what is comparable [101]. Thus, we build a type of biplot closely related to the area biplot method [41], in which lagged vectors are connected to form polygons. These geometric figures establish groups of almost similar triangles that can reveal the periodicities of a series. Chapters 8 and 9 bring the most embryonic phase of the investigation, in which the first questions about the conceived theme are raised.

Finally, it is known that the lack of software related to biplot methods prevents their widespread use [77]. With that in mind, our last contribution concerns filling this gap with the implementation of a package in R [91] that automates the approach developed in this investigation to estimate the periodicity of a TS (Chapter 10). The package has dual applicability, being able to function both as a periodicity estimation tool and in the broader multivariate context proposed by Gower [41].

1.3 Structure and organization

This thesis consists of a coherent and relevant set of research works and is divided into three parts, namely:

- Part I - covers Chapters 1 to 3 and is dedicated to providing essential knowledge about the issues under study. Chapter 1 provides an overview of the relevant aspects of the research, pointing out the motivating reasons, the established objectives to be reached, and elements of innovation. Chapter 2 introduces the concepts and mathematical foundations of the methods used to develop and implement the techniques, tools, and approaches throughout the investigation. Chapter 3 outlines the issues dealt with in the study, indicating the relevance of the contributions and the appropriate framework concerning the present state of the art.
- Part II - comprises the Chapters 4 to 10, that bring seven works developed during the investigation, as follows: i) four articles submitted to scientific journals indexed in Scopus (Article I - published; Article II, Article III, and Article IV - in review); ii) two articles submitted to international conferences (Article V - published in the conference proceedings; Article VI - published as a book chapter); iii) an open source Software (R package), published and available at the Comprehensive R Archive Network (CRAN).
- Part III - contains the Chapters 11 and 12, committed to the articulated discussions and the endings. Chapter 11 seeks to close the study by presenting some points of intersection between the articles and other topics not previously addressed but crucial for understanding what is proposed. Some open questions that emerged from the work and perspectives for future research are also handled. Finally, Chapter 12 concludes the work.

Relevant scientific communications resulting from this research are listed in the Appendix.

Chapter 2

Background

2.1 Principal Components Analysis

2.1.1 Initial notes

Let S be a vector space over a field $\mathbb{F}$ and $T : S \to S$ a linear transformation. A nonzero vector $\mathbf{x} \in S$ is called an eigenvector of T if exists an scalar $\lambda \in \mathbb{F}$ such that

$$T(\mathbf{x}) = \lambda\mathbf{x}. \tag{2.1}$$

Furthermore, the scalar λ is named the eigenvalue of T associated to the eigenvector $\mathbf{x}$ [3, 94]. When S is a finite-dimensional vector space, then the expression in (2.1) is equivalent to write

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}, \tag{2.2}$$

where $\mathbf{A}$ is the matrix representation of the linear transformation T relative to a fixed base at S. Hence, the p eigenvalues of $\mathbf{A}_{p\times p}$ are the roots of the characteristic polynomial $det(\mathbf{A} - \lambda\mathbf{I}_p)$. Besides, if $\mathbf{A}_{p\times p}$ is a real, symmetric matrix, then its eigenvalues are also real. If, in addition, all eigenvalues are distinct, then the respective eigenvectors are orthogonal [19].

Theorem 1 *(Spectral Decomposition): Let $\mathbf{A} \in \mathbb{M}_{p\times p}$. Then $\mathbf{A}$ is symmetric if and only if there exists an orthogonal matrix $\mathbf{Q} \in \mathbb{M}_{p\times p}$ and a diagonal matrix $\mathbf{\Lambda} \in \mathbb{M}_{p\times p}$, such that*

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{'}, \tag{2.3}$$

that is equivalent to

$$\mathbf{A} = \sum_{j=1}^{p} \lambda_j \mathbf{q}_j \mathbf{q}_j^{'}, \tag{2.4}$$

where λ_j, for $j = 1, \ldots, p$, are the eigenvalues of $\mathbf{A}$ and constitutes the diagonal elements of $\boldsymbol{\Lambda}$. In turn, $\mathbf{q}_j$ are the correspondent eigenvectors that form the columns of $\mathbf{Q}$, constituting an orthonormal set. In case all λ_j are distinct, then the associated $\mathbf{q}_j$ are unique [87].

2.1.2 PCA model construction

Let's consider a set of m random variables $\mathscr{X}_1$, ..., $\mathscr{X}_m$ related to a sample of n objects or individuals, organized as a random vector $\mathbf{x} = (\mathscr{X}_1, \ldots, \mathscr{X}_m)$. Next, it will be summarized as an $(n \times m)$ data matrix $\mathbf{X} = [x_{ij}]_{i=1,\cdots,n;j=1,\cdots,m}$, in which the columns represent the variables. In turn, each row vector belonging to the space $\mathbb{R}^m$ expresses an individual. Thus:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \ldots & x_{1m} \\ x_{21} & x_{22} & x_{23} & \ldots & x_{2m} \\ x_{31} & x_{32} & x_{33} & \ldots & x_{3m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \ldots & x_{nm} \end{bmatrix}, \tag{2.5}$$

and since m is usually large, projecting the m-dimensional row vectors into a reduced space allows us to explore the data visually [32]. It is often essential to level the m random quantities under study by the same reference point. Then a statistical transformation is implemented consisting of centering the columns $[\mathbf{X}_1 \cdots \mathbf{X}_m]$ on the mean by computing (for $j = 1, \ldots, m$):

$$\hat{\mathbf{X}}_j = \mathbf{X}_j - \bar{\mathbf{X}}_j, \tag{2.6}$$

where $\mathbf{X}_j = (x_{1j}, x_{2j}, \cdots, x_{nj})'$, and the j^{th} column mean is calculated as

$$\bar{\mathbf{X}}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}. \tag{2.7}$$

Eventually, an extra transformation called scaling is applied, in which all centered $\hat{x}_{ij}$ elements are divided by the standard deviation of the corresponding column such that, for $j = 1, \ldots, m$, we have

$$\tilde{\mathbf{X}}_j = \frac{\hat{\mathbf{X}}_j}{s_j}, \tag{2.8}$$

being that the standard deviation of the column $\mathbf{X}_j$ is given by

$$s_j = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{\mathbf{X}}_j)^2}, \tag{2.9}$$

resulting that the respective $\tilde{\mathbf{x}}_j$ will be a unit vector. Note that no scaling is needed when all the variables are measured in the same units and have similar dispersion [31]. Hereafter, we will assume that the matrix $\mathbf{X}$ refers to variables centered on the mean and will define the sample covariance matrix $\mathbf{S}$ as

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}^{'}\mathbf{X}. \tag{2.10}$$

However, now and then, it can be convenient to use $\mathbf{X}^{'}\mathbf{X}$ instead of $\mathbf{S}$. That is because the eigenvectors of both matrices are the same, and the eigenvalues of the sample covariance matrix are $(\frac{1}{n-1})$ of the eigenvalues of $\mathbf{X}^{'}\mathbf{X}$ [52, 53], i.e.,

$$(n-1)\mathbf{S} = \mathbf{X}^{'}\mathbf{X}. \tag{2.11}$$

That said, the PCA approach seeks to establish new uncorrelated variables through linear combinations of the $\mathbf{X}$ columns that successively maximize the variance [53], as below

$$\mathbf{t} = \sum_{j=1}^{m} p_j \mathbf{X}_j, \tag{2.12}$$

or in its matrix form

$$\mathbf{t} = \mathbf{X}\mathbf{p}, \tag{2.13}$$

in which $\mathbf{p}$ is a vector of constants $p_1, \ldots, p_m$. The variance of the linear combination in (2.13) is given by [53]

$$Var(\mathbf{X}\mathbf{p}) = \mathbf{p}^{'}\mathbf{S}\mathbf{p}. \tag{2.14}$$

Thus, the m-dimensional vector $\mathbf{p}$ that maximizes the quadratic form $\mathbf{p}^{'}\mathbf{S}\mathbf{p}$, subject to the normalization constraint $\mathbf{p}^{'}\mathbf{p} = 1$ [52], will also be the unit solution that maximizes the variance of $\mathbf{t}$

$$\begin{aligned} \max \quad & \phi = \mathbf{p}^{'}\mathbf{S}\mathbf{p} \\ \text{s.t.} \quad & \mathbf{p}^{'}\mathbf{p} = 1. \end{aligned} \tag{2.15}$$

It is equivalent to moving the constraint into the objective function and writing

$$\max \phi = \mathbf{p}^{'}\mathbf{S}\mathbf{p} - \lambda(\mathbf{p}^{'}\mathbf{p} - 1), \tag{2.16}$$

where λ is a Lagrange multiplier. Following, taking the partial derivatives of ϕ with respect to $\mathbf{p}$ and equating them to the null vector, results in

$$\begin{aligned} \frac{\partial \phi}{\partial \mathbf{p}} = 0 \quad & \iff 2\mathbf{S}\mathbf{p} - 2\lambda\mathbf{p} = 0 \\ & \iff \mathbf{S}\mathbf{p} = \lambda\mathbf{p}. \end{aligned} \tag{2.17}$$

Therefore, it shows that the λ in (2.16) is the largest eigenvalue of $\mathbf{S}$ and $\mathbf{p}$ is the corresponding eigenvector. Besides, λ represents the variance of $\mathbf{t}$ since

$$\mathbf{p}^{'}\mathbf{S}\mathbf{p} = \lambda\mathbf{p}^{'}\mathbf{p} = \lambda. \tag{2.18}$$

According to Theorem 1, and given that $\mathbf{S}$ is a real and symmetric matrix, then it has an eigenstructure with precisely m real eigenvalues, whose associated eigenvectors form an orthonormal set of vectors [19]. Hence, (2.15) must be adjusted to a sequence of optimization problems indexed by $k = 1, \ldots, m$, and adding the orthogonality constraint between the eigenvectors so that

$$\begin{aligned} \max \quad & \phi = \mathbf{p}_k^{'} \mathbf{S} \mathbf{p}_k \\ \text{s.t.} \quad & \mathbf{p}_k^{'} \mathbf{p}_k = 1 \\ & \mathbf{p}_j^{'} \mathbf{p}_k = 0, \quad {}_{j \in \{1,\ldots,m\} \setminus \{k\}}. \end{aligned} \tag{2.19}$$

The solutions to the problem in (2.19) are the entire set of eigenvectors of $\mathbf{S}$, and that will compose the m new variables described as linear combinations of the $\mathbf{X}$ columns as below:

$$\mathbf{t}_k = \sum_{j=1}^{m} p_{jk} \mathbf{X}_j, \tag{2.20}$$

for $k = 1, \ldots, m$, what is equivalent to

$$\mathbf{t}_k = \mathbf{X} \mathbf{p}_k. \tag{2.21}$$

To summarize, it suffices to say that the linear combinations $\mathbf{X}\mathbf{p}_k$ express the PCs algebraically [53], meaning that they point to the directions of maximum data variability in decreasing order. Moreover, the elements of the eigenvectors $\mathbf{p}_k$ are called PC *unit-loadings*, and the elements of $\mathbf{t}_k$ are called PC *scores*, which stand for the data cloud projections onto the PC directions [31].

2.1.3 Singular Value Decomposition

An equivalent way to obtain the spectral decomposition of $\mathbf{S}$ is computing the Singular Value Decomposition (SVD) of $\mathbf{X}$. The SVD can be seen as a generalization of the spectral decomposition theorem regarding non-symmetric matrices [61].

Theorem 2 *(Singular Value Decomposition): Let $\mathbf{X} \in \mathbb{M}_{n \times m}$ and have rank r. Then, one can write:*

$$\mathbf{X} = \mathbf{U} \boldsymbol{\Sigma} \mathbf{V}^{'}, \tag{2.22}$$

that is equivalent to

$$\mathbf{X} = \sum_{j=1}^{r} \sigma_j \mathbf{u}_j \mathbf{v}_j^{'}, \tag{2.23}$$

where the matrices $\mathbf{U}_{n \times r}$ and $\mathbf{V}_{m \times r}$ have orthonormal columns $\mathbf{u}_1, \ldots, \mathbf{u}_r$ and $\mathbf{v}_1, \ldots, \mathbf{v}_r$, designated as left and right singular vectors of $\mathbf{X}$, respectively. In turn, $\boldsymbol{\Sigma}_{r \times r}$ is a diagonal matrix in which the σ_j are called singular values, satisfying $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0$.

To extract the singular vectors of the matrix $\mathbf{X}_{n\times m}$, let us consider:

1. its rows $\mathbf{x}_1^{'}, \ldots, \mathbf{x}_n^{'}$ as points in the m-dimensional space;
2. the problem of minimizing the sum of the squares of the perpendicular distances of $\mathbf{x}_1^{'}, \ldots, \mathbf{x}_n^{'}$ to a line that passes through the origin, which is equivalent to maximizing the sum of the squares of the lengths of the points projections onto the line;
3. a unit vector $\mathbf{v}$ that points in the same direction of the line.

In these circumstances, each intended length is given by the absolute value of the scalar projection of a row along the unit vector, i.e., $|\mathbf{x}_i^{'}.\mathbf{v}|$, for $i = 1, \ldots, n$. It means that

$$\|\mathbf{X}\mathbf{v}\|^2 = \mathbf{v}^{'}\mathbf{X}^{'}\mathbf{X}\mathbf{v} \tag{2.24}$$

provides the sum of the squares of the scalar projections. Now, defining $\mathbf{v}_1$ as the first singular vector of $\mathbf{X}$, then the best-fit line in the least-squares sense is given by simply doing

$$\mathbf{v}_1 = \underset{|\mathbf{v}|=1}{\operatorname{argmax}} \|\mathbf{X}\mathbf{v}\|. \tag{2.25}$$

The first singular value that is associated with $\mathbf{v}_1$ is designated as σ_1 and can be computed as follows:

$$\sigma_1 = \|\mathbf{X}\mathbf{v}_1\|. \tag{2.26}$$

The subsequent singular vectors are calculated successively, imposing the orthogonality constraint between them such that

$$\mathbf{v}_2 = \underset{\mathbf{v}\perp\mathbf{v}_1, |\mathbf{v}|=1}{\operatorname{argmax}} \|\mathbf{X}\mathbf{v}\|, \tag{2.27}$$

and so on until

$$\mathbf{v}_r = \underset{\mathbf{v}\perp(\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_{r-1}), |\mathbf{v}|=1}{\operatorname{argmax}} \|\mathbf{X}\mathbf{v}\|, \tag{2.28}$$

with the respective singular values computed as

$$\sigma_j = \|\mathbf{X}\mathbf{v}_j\|. \tag{2.29}$$

Ultimately, naming the left singular vectors of $\mathbf{X}$ as $\mathbf{u}_j$, for $j = 1, \ldots, r$, and defining them as the normalized vectors $\frac{1}{\sigma_j}\mathbf{X}\mathbf{v}_j$, then they will also establish a set of orthogonal vectors [93]. Besides, from (2.18) and (2.24), it is clear that the $\mathbf{v}_j$'s are the eigenvectors of the matrix $\mathbf{X}^{'}\mathbf{X}$, and the σ_j^2's are the respective eigenvalues.

Matrix approximation

The next theorem establishes that SVD provides the optimal low-rank approximation to a matrix $\mathbf{X}$.

Theorem 3 *(Eckart-Young): The optimal rank-k approximation to* $\mathbf{X}$*, in an* L_2 *sense, is given by the rank-k SVD truncation* $\mathbf{X}_k$ [18]*:*

$$\underset{\mathbf{X}_k}{\operatorname{argmax}} \| \boldsymbol{X} - \mathbf{X}_k \|_2 = \mathbf{U}_k \boldsymbol{\Sigma}_k \mathbf{V}_k^{'}, \tag{2.30}$$

where the matrices $\mathbf{U}_k$ *and* $\mathbf{V}_k$ *are the leading* k *columns of* $\mathbf{U}$ *and* $\mathbf{V}$*, and* $\boldsymbol{\Sigma}_k$ *contains the leading* $k \times k$ *sub-block of* $\boldsymbol{\Sigma}$*.*

It means that for $k < r$, one can approximate the matrix $\mathbf{X}$ by doing:

$$\mathbf{X} \approx \sum_{j=1}^{k} \sigma_j \mathbf{u}_j \mathbf{v}_j^{'}, \tag{2.31}$$

where each σ_j is a singular value, being $\mathbf{u}_j$ and $\mathbf{v}_j$ the corresponding left and right singular vectors, respectively.

2.1.4 NIPALS algorithm

The Partial Least Squares (PLS) method is a family of iterative least squares-based algorithms that solve problems related to multivariate analysis, such as multiple regression and path modeling. The nonlinear iterative partial least squares (NIPALS) algorithm is the PLS approach used to compute the principal components in (2.21) when it comes to PCA. In its turn, the PCA method allows us to write a rank-r matrix $\mathbf{X}$ as a summation of r rank-*one* matrices

$$\mathbf{X} = \mathbf{M}_1 + \mathbf{M}_2 + \cdots + \mathbf{M}_r, \tag{2.32}$$

where each rank-*one* matrix is computed as

$$\mathbf{M}_j = \mathbf{t}_j \mathbf{p}_j^{'}, \tag{2.33}$$

being $\mathbf{t}_j$, for $j = 1, \ldots, r$, the vectors containing the projections of the sample points on the j^{th} PC direction and called score vectors. As for $\mathbf{p}_j$, those are the vectors containing the angle cosines of the direction vector corresponding to the j^{th} PC and are named loading vectors.
The NIPALS algorithm calculates one PC at a time so that it first computes $\mathbf{t}_1$ and $\mathbf{p}_1$ from the $\mathbf{X}$ matrix. Next, $\mathbf{X}$ is deflated by doing

$$\mathbf{E}_1 = \mathbf{X} - \mathbf{t}_1 \mathbf{p}_1^{'}, \tag{2.34}$$

where $\mathbf{E}_1$ is the first residual matrix. Then, $\mathbf{t}_2$ and $\mathbf{p}_2$ are calculated from the first residual matrix. After, the second residual matrix $\mathbf{E}_2$ is obtained deflating $\mathbf{E}_1$ as follows

$$\mathbf{E}_2 = \mathbf{E}_1 - \mathbf{t}_2 \mathbf{p}_2^{'}. \tag{2.35}$$

The procedure is repeated until $\mathbf{E}_r$ is achieved, such that

$$\mathbf{E}_{r+k} = 0, \forall k > 0. \tag{2.36}$$

The steps of the NIPALS algorithm are described as follows:

Algorithm 1: NIPALS algorithm

```
**Input:** $\mathbf{E}_0 = \mathbf{X}$
**Output:** $\mathbf{P} = [\mathbf{p}_1, \dots, \mathbf{p}_r], \mathbf{T} = [\mathbf{t}_1, \dots, \mathbf{t}_r]$
**for** $j = 1$ **to** $r$ **do**
    Step 0: Initialize $\mathbf{t}_j$
    Step 1:
    **repeat**
        Step 1.1: $\mathbf{p}_j = \mathbf{E}'_{j-1}\mathbf{t}_j/(\mathbf{t}'_j\mathbf{t}_j)$;
        Step 1.2: $\mathbf{p}_j = \mathbf{p}_j/\|\mathbf{p}_j\|$ ;
        Step 1.3: $\mathbf{t}_j = \mathbf{E}_{j-1}\mathbf{p}_j$;
    **until** convergence of $\mathbf{t}_j$;
    Step 2: $\mathbf{E}_j = \mathbf{E}_{j-1} - \mathbf{t}_j\mathbf{p}'_j$
**end**
```

Observe that, from the relations established inside Step 1 of the NIPALS algorithm, one can obtain

$$\mathbf{E}'_{j-1}\mathbf{E}_{j-1}\mathbf{p}_j = \mathbf{t}'_j\mathbf{t}_j\mathbf{p}_j, \tag{2.37}$$

and

$$\mathbf{E}_{j-1}\mathbf{E}'_{j-1}\mathbf{t}^*_j = \mathbf{t}'_j\mathbf{t}_j\mathbf{t}^*_j, \tag{2.38}$$

where $\mathbf{t}^*_j$ is the j^{th} unit score vector, which is equivalent to write

$$\mathbf{t}_j = \sqrt{\mathbf{t}'_j\mathbf{t}_j}\mathbf{t}^*_j.$$

It means that the scalar $\mathbf{t}'_j\mathbf{t}_j$ is the largest eigenvalue of both $\mathbf{E}'_{j-1}\mathbf{E}_{j-1}$ and $\mathbf{E}_{j-1}\mathbf{E}'_{j-1}$ matrices, being $\mathbf{p}_j$ and $\mathbf{t}_j$ the corresponding eigenvectors. Moreover, when $j = 1$ in the algorithm, the equations in (2.37) and (2.38) can be write as

$$\begin{aligned} \mathbf{X}'\mathbf{X}\,\mathbf{p}_1 &= \mathbf{t}'_1\mathbf{t}_1\,\mathbf{p}_1 \\ \mathbf{X}\mathbf{X}'\,\mathbf{t}^*_1 &= \mathbf{t}'_1\mathbf{t}_1\,\mathbf{t}^*_1, \end{aligned}$$

implying that the first unit score vector and the first loading vector are exactly the first left and right singular vectors of $\mathbf{X}$, respectively, as well as $\sqrt{\mathbf{t}'_1\mathbf{t}_1}$ returns the first singular value of $\mathbf{X}$. For $j > 1$, and considering the column space of $\mathbf{X}$, the NIPALS computes the j-th principal component over the orthogonal complement of the subspace $\mathbf{t}_1\mathbf{p}'_1 + \cdots + \mathbf{t}_{j-1}\mathbf{p}'_{j-1}$, which is equivalent to the SVD approach of imposing the orthogonality restriction among singular vectors when maximizing its objective function. Thus, $\forall\ j = 1, \ldots, d$, $\mathbf{t}_j^*$ and $\mathbf{p}_j$ are equal to the left and right singular vectors of the SVD of $\mathbf{X}$, respectively, and each $\sqrt{\mathbf{t}_j'\mathbf{t}_j}$ is the corresponding singular value. The NIPALS decomposition of $\mathbf{X}$ is then given by

$$\mathbf{X} = \sqrt{\mathbf{t}_1'\mathbf{t}_1}\mathbf{t}_1^*\mathbf{p}_1' + \cdots + \sqrt{\mathbf{t}_d'\mathbf{t}_d}\mathbf{t}_d^*\mathbf{p}_d'. \tag{2.39}$$

Defining the matrix $\mathbf{\Sigma}$ as a diagonal matrix containing the singular values $\sqrt{\mathbf{t}_j'\mathbf{t}_j}$ arranged in decreasing order, one can write the matrix form of the expansion (2.39) as

$$\mathbf{X} = \mathbf{T}^*\mathbf{\Sigma}\mathbf{P}', \tag{2.40}$$

where $\mathbf{T}^*$ is the (unit-)scores matrix whose column vectors $\mathbf{t}_j^*$ are orthonormal, and $\mathbf{P}$ is the (unit-)loadings matrix whose column vectors $\mathbf{p}_j$ are also orthonormal.

2.2 Singular Spectrum Analysis

Given that, in the case of TS analysis, the main objective of the SSA is the decomposition of the original series into additive and interpretable components (e.g., trend, seasonality, and noise), this section begins with a brief discussion about them. Then, the basic SSA scheme is presented in detail, from transforming a univariate real series into a trajectory matrix to reconstructing the so-called elementary series through the diagonal averaging procedure.

2.2.1 Time series decomposition

Simply put, a TS can be defined as observations of a given phenomenon ordered over time. But not only that, it is crucial that these observations are time-sensitive. On the other hand, the time intervals at which measurements take place are usually regular, for example, annually, quarterly, monthly, weekly, or daily [14]. First, let's define a TS as a set of time-indexed random variables

$$\{\mathscr{Y}_1, \ldots, \mathscr{Y}_T\}, \tag{2.41}$$

whereas an observed TS is represented as a set of realizations of the random variables as

$$\{y_1, \ldots, y_T\}. \tag{2.42}$$

Decomposing a TS means, at any time t, separating unobservable components that are (or are not) associated with distinct variations in time. Traditionally, they are classified as i) long-term trend (T_t); ii) cyclical movements super-imposed upon the long-term trend (C_t) [15]; iii) seasonal movements within each year (S_t); iv) residual variations (R_t). Then, an additive decomposition model is written as

$$\mathscr{Y}_t = T_t + C_t + S_t + R_t. \tag{2.43}$$

When decomposing a TS, it is common to consider the trend and the cycle as a single component, designating it only as *trend* for simplicity [49]. Hence, for this investigation, the TS components are:

1. a global trend represented by a low-degree polynomial function, such as

   $$f(t) = a_0 + a_1 t + \cdots + a_m t^m, \tag{2.44}$$

   where t is the time, a_i are real constants, and m denotes the degree of the polynomial function;
2. a regular oscillatory component, with a period p, that can be synthesized as a function $g(t)$ and defined by a linear combination of sines and cosines with constant coefficients, i.e., expressed as a Fourier series such that

   $$g(t) = \sum_{m=0}^{p} (\alpha_m cos(\omega_m t) + \beta_m sin(\omega_m t)), \tag{2.45}$$

   where $\omega_m = 2\pi m/p$;
3. an irregular component ε_t, or noise at instant t.

Thus, the additive decomposition model to be considered becomes [49]:

$$\mathscr{Y}_t = f(t) + g(t) + \varepsilon_t. \tag{2.46}$$

Although other TS decomposition methods have some effectiveness, such as X11 [15, p. 79] and Seasonal Extraction in ARIMA Time Series (SEATS) [15, p. 121], the most used is the well-known Seasonal and Trend Decomposition using Loess (STL). That is a filtering procedure for decomposing a TS into the three previously mentioned components: trend, seasonal, and remainder [13]. Below is an example of applying the method using the *stl* function of the *stat* package in R [83], represented in Figure 2.1 and coded as:

```
> plot(stl(data, t.window=13, s.window=17)
```

The data are the records of CO_2 concentration in the Earth's atmosphere, measured monthly from January 1965 to December 1980 at an observing station on Mauna Loa in Hawaii [58]. The parameter *t.window* provides the number of consecutive observations that can be used to estimate the trend. In turn, *s.window* represents the number of consecutive years used to estimate a value in the seasonal component.

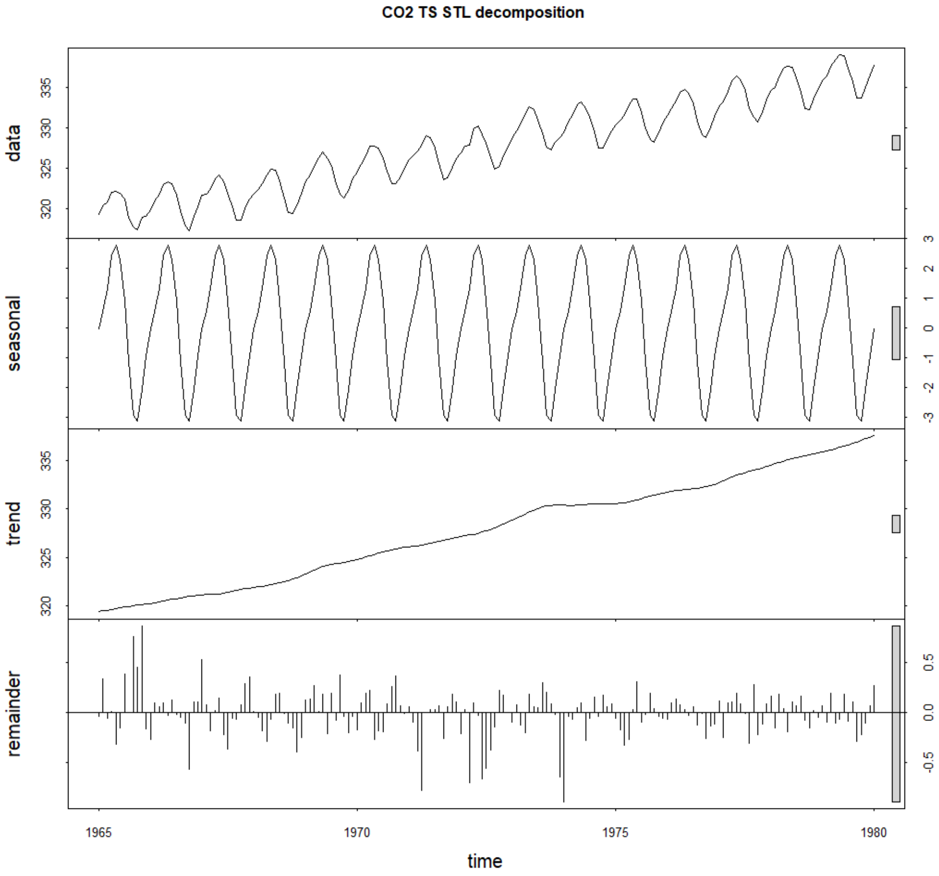

Figure 2.1: CO_2 TS decomposition through STL method.

2.2.2 Basic SSA

To deconstruct a TS into some interpretable components, SSA draws to some extent on the theory behind PCA [45]. As we will see below, the trajectory matrix's SVD decomposition backs the method and provides the desired interpretability. The basic scheme of the SSA method consists of two stages, decomposition and reconstruction. The decomposition stage is split into embedding and SVD steps, while the reconstruction is into grouping and diagonal averaging. The details regarding each one of them come next.

Stage 1 (1^{st} Step): Embedding

Consider a univariate real-valued TS $Y = (y_0, \ldots, y_{n-1})$ of length n, and let ℓ be the integer representing the so-called window length, in which $0 < \ell < n$. The embedding step maps Y to a sequence of multidimensional ℓ-lagged vectors

$$\mathbf{x}_i = (y_{i-1}, \ldots, y_{i+\ell-2})', \;\; 1 \leq i \leq \kappa, \tag{2.47}$$

where $\kappa = n - \ell + 1$. From there, the trajectory matrix is constructed so that

$$\mathbf{X} = [\mathbf{x}_1 \ldots \mathbf{x}_\kappa], \tag{2.48}$$

a Hankel matrix in which the columns are the ℓ-lagged vectors. It means that $\mathbf{X}$ consists of the transformation of a TS into a Hankel matrix using an embedding operator $\mathfrak{T}$, such that

$$\mathfrak{T}(Y) = \mathbf{X} = \begin{bmatrix} y_0 & y_1 & y_2 & \cdots & y_{\kappa-1} \\ y_1 & y_2 & y_3 & \cdots & y_\kappa \\ y_2 & y_3 & y_4 & \cdots & y_{\kappa+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_{\ell-1} & y_\ell & y_{\ell+1} & \cdots & y_{n-1} \end{bmatrix}, \tag{2.49}$$

where $x_{ij} = y_{i+j-2}$. As a final note, the value of ℓ must be large enough for the lagged vectors to capture an essential part of the behavior of the series.

Stage 1 (2^{nd} Step): SVD

The result of this step is the decomposition of the trajectory matrix as a sum of rank-one matrices so that

$$\mathbf{X} = \sum_{i=1}^{d} \sqrt{\lambda_i} \mathbf{u}_i \mathbf{v}_i' \tag{2.50}$$

where $d = rank(\mathbf{X})$. Besides, $\lambda_i, i = 1, \ldots, d$, are the eigenvalues of the matrix $\mathbf{X}\mathbf{X}'$ arranged in decreasing order of magnitudes ($\lambda_i > 0$), and associated to the orthonormal system of the eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_d$, being

$$\mathbf{v}_i = \mathbf{X}' \mathbf{u}_i / \sqrt{\lambda_i}. \tag{2.51}$$

In the SVD context, the elements of the collection $(\sqrt{\lambda_i}, \mathbf{u}_i, \mathbf{v}_i)$ are the singular values, left and right singular vectors of $\mathbf{X}$, respectively. Moreover, one can write (2.48) in matrix form if we define

$$\mathbf{X}_i = \sqrt{\lambda_i} \mathbf{u}_i \mathbf{v}_i \tag{2.52}$$

to get

$$\mathbf{X} = \mathbf{X}_1 + \cdots + \mathbf{X}_d. \tag{2.53}$$

Stage 2 (3^{rd} Step): Grouping

The grouping step performs the separation of the additive components of the series [40], identifying the sets of rank-one matrices in (2.51) associated with the trend, the oscillatory component, and the noise. Then, given the index set $I = \{1, \ldots, d\}$, the step starts with the partitioning of I into disjoint subsets $I_k, k = 1, \ldots, p$, yielding in the following decomposition

$$\mathbf{X} = \mathbf{X}_{I_1} + \cdots + \mathbf{X}_{I_p}, \tag{2.54}$$

where $\mathbf{X}_I = \sum_{i \in I} \mathbf{X}_i$ [39].

Stage 2 (4^{th} Step): Diagonal Averaging

The diagonal averaging transfers a matrix resulting from the grouping step to a corresponding TS [39]. Concretely, each $\ell \times \kappa$ matrix $X_{I_k}, k = 1, ..., p$, in (2.54) provides a reconstruction of an associated series $\tilde{Y}^{(k)} = (\tilde{y}_0^{(k)}, \ldots, \tilde{y}_{n-1}^{(k)})$, whose elements correspond to the means of the anti-diagonals in X_{I_k}. Hence, each $\tilde{y}_\iota^{(k)}$, $\iota = 0, \ldots, (n-1)$, is computed as:

$$\tilde{y}_\iota^{(k)} = \begin{cases} \frac{1}{\iota+1} \sum_{m=1}^{\iota+1} \tilde{x}_{m,\iota-m+2}, & \text{if } 0 \leq \iota < \ell - 1, \\ \frac{1}{\ell} \sum_{m=1}^{\ell} \tilde{x}_{m,\iota-m+2}, & \text{if } \ell - 1 \leq \iota < \kappa, \\ \frac{1}{n-\iota} \sum_{m=\iota-\kappa+2}^{n-\kappa+1} \tilde{x}_{m,\iota-m+2}, & \text{if } \kappa \leq \iota < n. \end{cases} \tag{2.55}$$

Just for the sake of argument, let's consider a toy example to show the procedure of this step schematically, such that $Y = (y_0, \ldots, y_5)$ and where $\ell = 3$. Next, suppose that the matrix $\tilde{\mathbf{X}} = \mathbf{X}_{I_k}$ was determined in the grouping step, and it is associated with some component of Y. Thus, the corresponding TS $\tilde{Y}$ will be computed by averaging the skewed diagonals ($i + j = constant$) as in Figure 2.2 and established in (2.55). Essential to note yet that the final result provides the decomposition of the original TS Y into a sum of reconstructed series as below

$$Y = \sum_{k=1}^{p} \tilde{Y}^{(k)}. \tag{2.56}$$

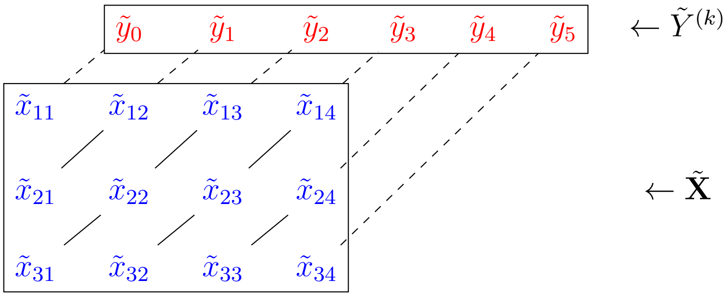

Figure 2.2: The diagonal averaging scheme through a toy example.

2.2.3 Separability

The concept of separability plays a crucial role in the series decomposition, especially in the grouping step. Considering the TS Y and a fixed ℓ, let us assume just two identified groups associated to the matrices $\mathbf{X}_{I_1}$ and $\mathbf{X}_{I_2}$, where $I_1 = \{i_1, \ldots, i_r\}$ and $I_2 = I \setminus I_1$, for $1 \leq i_1 < \cdots < i_r \leq d$. Furthermore, consider the corresponding reconstructed series (2.55) $\tilde{Y}^{(1)}$ and $\tilde{Y}^{(2)}$ and their respective trajectory matrices $\tilde{\mathbf{X}}^{(1)}$ and $\tilde{\mathbf{X}}^{(2)}$. Thus, if $\mathbf{X}_{I_1} = \tilde{\mathbf{X}}^{(1)}$ and $\mathbf{X}_{I_2} = \tilde{\mathbf{X}}^{(2)}$, then we say that $\tilde{Y}^{(1)}$ and $\tilde{Y}^{(2)}$ are separable.

Denoting the linear spaces spanned by the columns of $\tilde{\mathbf{X}}^{(1)}$ and $\tilde{\mathbf{X}}^{(2)}$ by $L^{(\ell,1)}$ and $L^{(\ell,2)}$, and those spanned by the columns of $(\tilde{\mathbf{X}}^{(1)})'$ and $(\tilde{\mathbf{X}}^{(2)})'$ by $L^{(\kappa,1)}$ and $L^{(\kappa,2)}$, the separability stated before means that $L^{(\ell,1)} \perp L^{(\ell,2)}$ and $L^{(\kappa,1)} \perp L^{(\kappa,2)}$, being named *weak*. In this case, $\tilde{\mathbf{X}}^{(1)}(\tilde{\mathbf{X}}^{(2)})' = \mathbf{0}_{\ell\times\ell}$, and also $(\tilde{\mathbf{X}}^{(1)})'\tilde{\mathbf{X}}^{(2)} = \mathbf{0}_{\kappa\times\kappa}$. If, in addition to the spaces orthogonality, the set of singular values of $\tilde{\mathbf{X}}^{(1)}$ and $\tilde{\mathbf{X}}^{(2)}$ are disjoint, then $\tilde{Y}^{(1)}$ and $\tilde{Y}^{(2)}$ are said to be strongly separable [9].

On the other hand, if $\mathbf{X}_{I_1}$ and $\mathbf{X}_{I_2}$ are just close to being Hankel matrices, we will say that $\tilde{Y}^{(1)}$ and $\tilde{Y}^{(2)}$ are approximately separable. Therefore, the main objective of the grouping step is to find $I_1, \ldots, I_p$ so that $\mathbf{X}_{I_1}, \ldots, \mathbf{X}_{I_p}$ are close to being Hankel matrices and satisfies 2.54, ensuring the successful decomposition of Y [9, 39].

2.2.4 Visualization tools available in SSA

In the vocabulary of the SSA method, each collection $(\sqrt{\lambda_i}, \mathbf{u}_i, \mathbf{v}_i), i = 1, \ldots, d$, is called *eigentriple* [39]. Their elements are used to build additional graphics that help to separate the components in the grouping step. The scree plot of $\sqrt{\lambda_i}$ against i is helpful to identify the eigentriples associated with both trend and seasonal components. Dominant singular values indicate the corresponding eigentriples are related to the trend. In their turn, pairs of singular values very close to each other point out that the correspondent eigentriples are related to a harmonic component.

For example, take the trajectory matrix $\mathbf{X}$ of the CO_2 series mentioned above, using a window length ℓ equal to $n/2$. From the SVD of $\mathbf{X}$, the scree plot of the singular values looks as shown in Figure 2.3. The first singular value stands out from the others, indicating that the first eigentriple and the trend are related. Another way to separate the trend group is through a scatter plot of the elements of an eigenvector. A slowly varying of them implies an association of the corresponding eigentriple with the trend [36]. Figure 2.4 brings about an example related to the TS CO_2, in which the first left singular vector presents such behavior.

The periodicity extraction with period $P = 1/w$ is performed through scatterplots of pairs of eigenvectors, where w is the frequency of a harmonic component. When patterns appear in quasi-regular polygons, the number of vertices must coincide with P [45]. In Figure 2.5, one can see that the second graph formed by the eigenvectors $\mathbf{u}_2$ and $\mathbf{u}_3$ establishes a pattern where $P = 12$, meaning that the correspondent eigentriples are related to the seasonal component. The plateau formed by the second and third singular values in Figure 2.3 corroborates this understanding. The same goes for the fourth graph in Figure 2.5, with the eigenvectors $\mathbf{u}_4$ and $\mathbf{u}_5$, where $P = 6$.

1^{st} visualization tool available in SSA

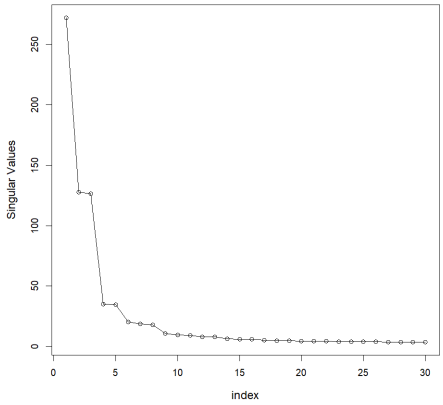

Figure 2.3: The scree plot of the singular values of the trajectory matrix $\mathbf{X}$ corresponding to the TS CO_2.

2^{nd} visualization tool available in SSA

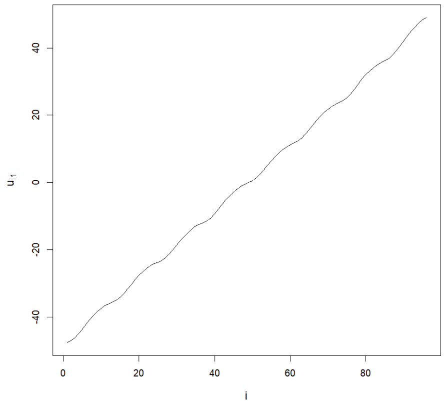

Figure 2.4: Slowly varying left singular vector $\mathbf{u}_1$ (TS CO_2).

3^{rd} visualization tool available in SSA

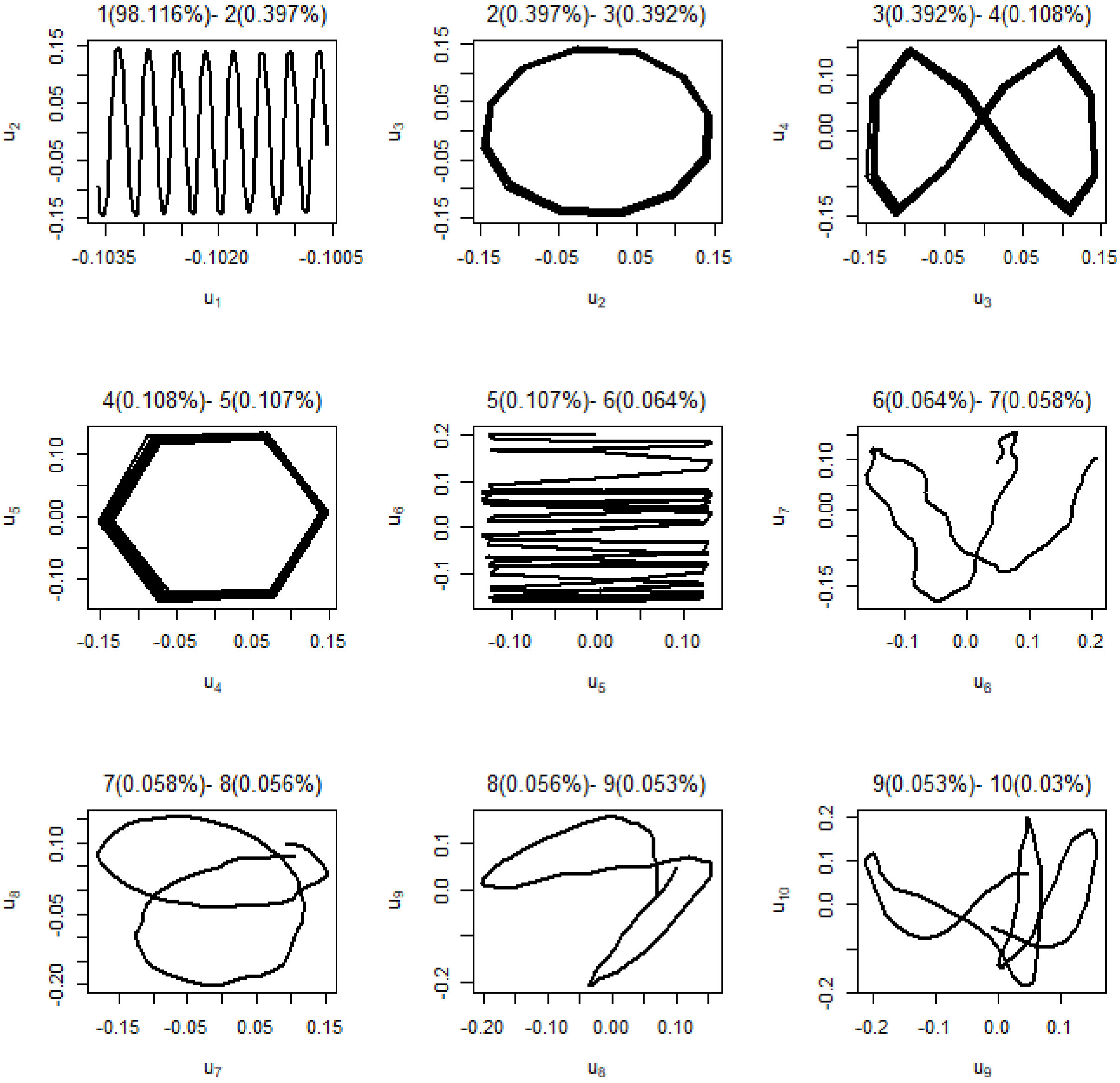

Figure 2.5: Paired eigenvectors 1–10 for the TS CO_2.

2.3 Biplot methods

The biplot is a multivariate visualization technique, usually constructed as a 2- or 3-dimensional graph, allowing simultaneous representation of both objects and variables of data sets. Biplots can reveal essential characteristics of multivariate data structure, e.g., patterns of correlations between variables or similarities between individuals [43]. Generally, any $(\ell \times \kappa)$ matrix $\mathbf{X}$ of rank d can be factorized as

$$\mathbf{X} = \mathbf{A}\mathbf{B}', \tag{2.57}$$

where $\mathbf{A}$ and $\mathbf{B}$ are $(\ell \times d)$ and $(\kappa \times d)$ matrices each of rank d [84]. For example, Gabriel [24] conceived the Biplot method by adopting the following classical notation for this type of factorization:

$$\mathbf{X} = \mathbf{G}\mathbf{H}', \tag{2.58}$$

in which $\mathbf{G}$ is a $(\ell \times q)$ matrix and $\mathbf{H}$ has a dimension of $(\kappa \times q)$, with $q \leq d$ and such that their rows create two sets of q-dimensional points, as follows

$$\mathbf{G} = \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1q} \\ g_{21} & g_{22} & \cdots & g_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ g_{\ell 1} & g_{\ell 2} & \cdots & g_{\ell q} \end{bmatrix}, \quad \text{and} \quad \mathbf{H} = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1q} \\ h_{21} & h_{22} & \cdots & h_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ h_{\kappa 1} & h_{\kappa 2} & \cdots & h_{\kappa q} \end{bmatrix}. \tag{2.59}$$

Taking $q = 2$, one can simultaneously represent $\mathbf{X}$'s rows and columns on the same graph, 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). Even when $q > 2$, it is possible to construct a biplot after obtaining the best rank two approximation of $\mathbf{X}$, in the sense of least square. A 2-dimensional biplot displays both row markers $\mathbf{g}_1, \cdots, \mathbf{g}_\ell$ and column markers $\mathbf{h}_1, \cdots, \mathbf{h}_\kappa$ of $\mathbf{X}$, so that the inner product $\mathbf{g}_i'\mathbf{h}_j$ approximates the element x_{ij} of $\mathbf{X}$ [77], such that

$$[x]_{i,j=1}^{\ell,\kappa} \approx \begin{bmatrix} \mathbf{g}_1' \\ \mathbf{g}_2' \\ \vdots \\ \mathbf{g}_\ell' \end{bmatrix} \begin{bmatrix} \mathbf{h}_1 & \mathbf{h}_2 & \cdots & \mathbf{h}_\kappa \end{bmatrix}, \tag{2.60}$$

for $i = 1, \cdots, \ell$ and $j = 1, \cdots, \kappa$. Taking into account the singular values decomposition performed by the NIPALS algorithm (2.40), the factorization of $\mathbf{X}$ can assume different configurations (Figure 2.6), resulting in different types of biplots [43].

When the factorization considers $\mathbf{G} = \mathbf{T}^*$ and $\mathbf{H} = \mathbf{P}\boldsymbol{\Sigma}$, the result is said to preserve the column metrics of $\mathbf{X}$, and the corresponding biplot receives different denominations, such as *i*) Gabriel biplot; *ii*) classic biplot; *iii*) covariance biplot; and *iv*) GH-biplot. Assuming $\mathbf{X} columns have centered$, the GH-biplot satisfies the following properties [77]:

1. The norm of a column marker $\mathbf{h}_j$ is proportional to the standard deviation of the respective variable;
2. The cosine of the angle formed by column markers approximates the correlation between the related variables;
3. The columns are better represented than the rows in terms of quality.

Another possibility is to define $\mathbf{J} = \mathbf{T}^*\mathbf{\Sigma}$ and $\mathbf{K} = \mathbf{P}$. In this respect, the factorization will preserve the metric of the rows in the so-called form biplot or JK-biplot. Ergo, the Euclidean distances between the row markers approximate the Euclidean distances between the respective individuals in the full space. The representation of the rows is better than the columns.

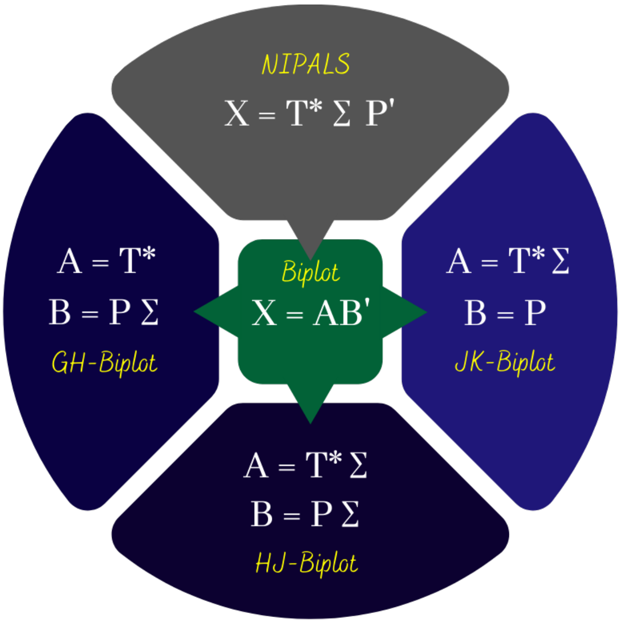

Figure 2.6: Biplot types according to the matrix factorization configuration.

2.3.1 HJ-Biplot

Based on SVD, a different type of biplot, called HJ-biplot, was proposed in 1986 by Galindo [25] in which optimal quality representation of the ℓ rows and the κ columns of $\mathbf{X}$ is ensured in the same Euclidean space. An HJ-biplot version based on NIPALS instead of the SVD can be constructed taking the rows of the matrix $\mathbf{J} = \mathbf{T}^*\mathbf{\Sigma}$ as row markers, and the rows of the matrix $\mathbf{H} = \mathbf{P}\mathbf{\Sigma}$ as column markers of $\mathbf{X}$. Indeed, the NIPALS decomposition in (2.40) results in that

$$\mathbf{X}\mathbf{P} = \mathbf{T}^*\mathbf{\Sigma}\mathbf{P}'\mathbf{P} = \mathbf{T}^*\mathbf{\Sigma}.$$

So, the ℓ rows of the matrix $\mathbf{J} = \mathbf{T}^*\boldsymbol{\Sigma}$ correspond to the projections of the ℓ points represented by the rows (individuals) of $\mathbf{X}$ onto the subspace spanned by the loading vectors $\mathbf{p}_1$ and $\mathbf{p}_2$, that is, the best-fit two-dimensional subspace for $\mathbf{X}$. Likewise, the κ rows of the matrix $\mathbf{H} = \mathbf{P}\boldsymbol{\Sigma}$ coincide with the projections of the κ points expressed by the columns (variables) of $\mathbf{X}$ onto the subspace spanned by the normalized score vectors $\mathbf{t}_1^*$ and $\mathbf{t}_2^*$, as seen below:

$$(\mathbf{T}^*)'\mathbf{X} = (\mathbf{T}^*)'\mathbf{T}^*\boldsymbol{\Sigma}\mathbf{P}' \Longleftrightarrow \mathbf{X}'\mathbf{T}^* = \mathbf{P}\boldsymbol{\Sigma}.$$

From (2.40) follows that both row and column representations of $\mathbf{X}$, $\mathbf{XP}$ and $\mathbf{X}'\mathbf{T}^*$, respectively, are related since

$$\mathbf{A} = \mathbf{X}'\mathbf{B}\boldsymbol{\Sigma}^{-1} \tag{2.61}$$

and

$$\mathbf{B} = \mathbf{X}\mathbf{A}\boldsymbol{\Sigma}^{-1}, \tag{2.62}$$

where $\mathbf{A} = \mathbf{XP}$ and $\mathbf{B} = \mathbf{X}'\mathbf{T}^*$.

It means that the coordinates of the variables can be expressed as a weighted average of the coordinates of the individuals and vice-versa. Consequently, it allows the representation of the rows and columns in the same Cartesian coordinates system with optimal quality of representation [25, 77]. However, in the HJ simultaneous representation, the inner product $\mathbf{j}_\ell'\mathbf{h}_\kappa$ does not provide an approximation to the element $x_{\ell\kappa}$ of $\mathbf{X}$ anymore, and consequently

$$\mathbf{X} \neq \mathbf{J}\mathbf{H}'.$$

Accordingly, the interpretation of the HJ-biplot representation can be performed as follows:

- The distance between points corresponds to how different the associated individuals are (dissimilarities), just like in JK-biplots;
- As it occurs in GH-biplot, the size of an arrow (variable) is proportional to the standard deviation of the associated variable, i.e., the longer the arrow, the greater the correspondent standard deviation;
- The cosine of the angle between arrows approximates the correlation between the variables they represent. Thus, if the angle is next to 90 degrees it indicates a poor correlation, while an angle close to 0 degrees or 180 degrees suggests a strong correlation, being positive in the first case and negative in the other.

2.3.2 Area Biplot

The Area Biplot is a visualization technique used to estimate data values through the areas spanned by triangles constructed from the results of the SVD of $\mathbf{X}$. To guarantee that the row and column markers exhibit a similar spread, another type of target matrix factorization is proposed in [41]. Standardizing both matrices $\mathbf{A}$ and $\mathbf{B}$ facilitate the visual inspection of the biplots as follows:

$$\mathbf{A} = (\frac{\ell}{\kappa})^{\frac{1}{4}}\mathbf{T}_2^*\mathbf{\Sigma}_2^{\frac{1}{2}}, \tag{2.63}$$

and

$$\mathbf{B} = (\frac{\kappa}{\ell})^{\frac{1}{4}}\mathbf{P}_2\mathbf{\Sigma}_2^{\frac{1}{2}}, \tag{2.64}$$

where $\mathbf{T}_2^*$ and $\mathbf{P}_2$ denote two consecutive columns of $\mathbf{T}^*$ and $\mathbf{P}$, and $\mathbf{\Sigma}_2$ the diagonal matrix with the two corresponding singular values. The inner product matrix $\mathbf{AB}^{'}$ provides an approximation for $\mathbf{X}$. Briefly, the procedure to construct an area biplot starts rotating the row markers by 90°, i.e., doing

$$\mathbf{a}_i^{[r]} = \mathbf{Ra}_i, \tag{2.65}$$

in which $\mathbf{R}$ is the (2×2) 90° counterclockwise rotation matrix. If we consider θ_{ij} as the angle between the vectors $\mathbf{a}_i$ and $\mathbf{b}_j$, then

$$cos(\theta_{ij}) = sin(\theta_{ij} + \pi/2) = sin(\phi_{ij}). \tag{2.66}$$

Hence, instead of writing the inner product between $\mathbf{a}_i$ and $\mathbf{b}_j$ as

$$\mathbf{a}_i^{'}\mathbf{b}_j = \|\mathbf{a}_i\|\,\|\mathbf{b}_j\|\,cos(\theta_{ij}), \tag{2.67}$$

one can consider

$$\mathbf{a}_i^{'}\mathbf{b}_j = \|\mathbf{a}_i\|\,\|\mathbf{b}_j\|\,sin(\phi_{ij}), \tag{2.68}$$

where ϕ_{ij} is the angle between the 90°-rotated biplot point $\mathbf{a}_i^{[r]}$ and the biplot vector $\mathbf{b}_j$. Besides, that is what justifies the choice for the $90° = \pi/2$ counterclockwise rotation. So, the expression (2.68) provides twice the area of the triangle formed by the origin and the endpoints of the vectors $\mathbf{a}_i^{[r]}$ and $\mathbf{b}_j$ (Figure 2.7). Therefore, the element x_{ij} may be estimated by the double of the triangle area.

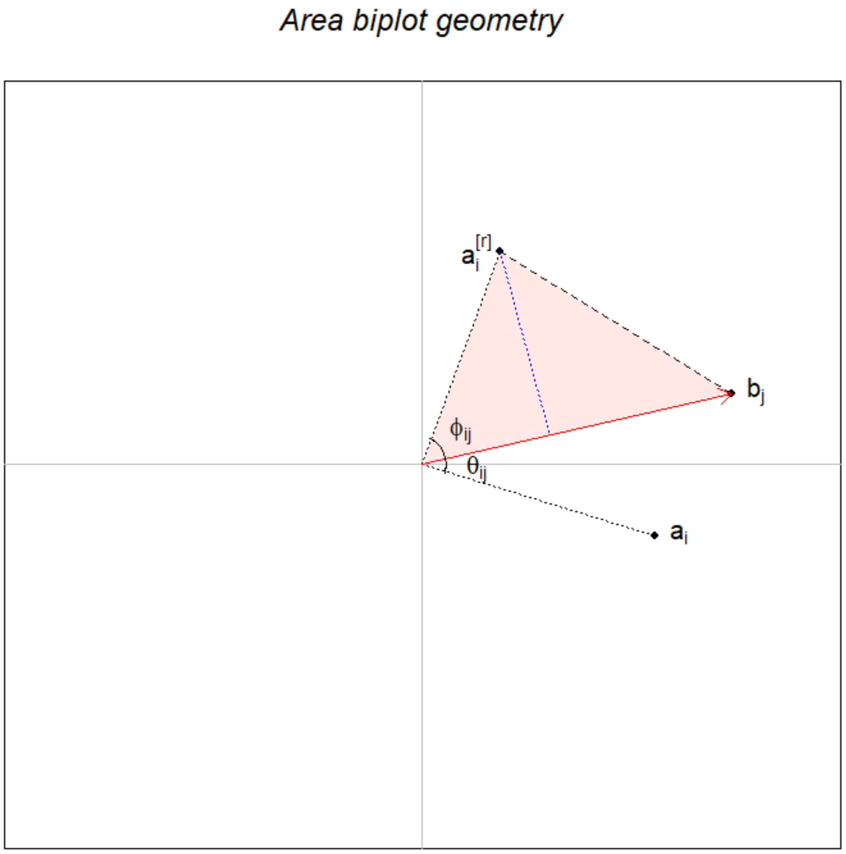

Figure 2.7: Area biplot construction: after rotating the point $\mathbf{a}_i$, twice the highlighted triangle area provides an estimation for element x_{ij}.

Chapter 3

Research progression

In the context of Statistics, matrix decomposition and matrix factorization (MF) are interchangeable ways of referring to techniques used for data dimensionality reduction while preserving as much information as possible. Two prominent areas in which MF techniques are widely applied are Omics technology [78, 92] and recommender systems [59, 70]. In the first case, the high-dimensional biological data is represented through a matrix containing in its rows expression counts, methylation levels, protein concentrations, etc., and individual samples in its columns [92]. The most commonly used MF techniques in Omics data are independent component analysis (ICA), non-negative matrix factorization (NMF), and PCA [30, 78, 92, 102]. Unlike PCA, which imposes the orthogonality constraint between components, ICA uses statistical independence between them when projecting the data into a lower-dimensional space [78]. The objective is to extract components that are maximally independent and non-Gaussian [50, 54]. NMF methods place non-negativity constraints on the data model, imposing all elements of the factor matrices to be greater than or equal to zero [102]. As for recommender systems, these engines deal with data sparseness while seeking to predict the rating a user would give to an item [70]. The primary MF technique in recommendation platforms is SVD or variations such as Improved Regularized SVD (RSVD2) [80].

In this thesis, the NIPALS [105] was the MF method chosen to decompose the trajectory matrix $\mathbf{X}$ (in 2.49) due to its ability to deal with missing data [22, 106]. The versatility of NIPALS allows its application in several fields, such as data science [48], where it was used to improve the performance of classification algorithms in high-dimensional data, and geophysics [33], in a study carried out to determine temporal variations of geoid heights. The algorithm performs a singular value decomposition calculating the scores and loadings iteratively. The NIPALS yields a sequence of orthogonal vector linear combinations of the data points. It is difficult to interpret these raw results in vector terms other than through visualization [68]. Therefore, we sought to apply biplot methods to understand the underlying data structure and its generation process. With this, we intend to establish the minimum criteria of visual interpretation to associate groups of eigentriples to the components of a TS. Also, we expect to circumvent eventual issues related to changes in the TS structure by applying the segmentation procedure and creating a way to identify the breakpoints. We expect other insights will emerge regarding the series' characteristics due to the biplots' interpretability.

3.1 Trajectory matrix factorization

3.1.1 State of the art

In the original basic SSA, the Hankel trajectory matrix $\mathbf{X}$ factorization using the SVD method [38, 39, 45] poses some problems. A well-known setback is that the SVD of $\mathbf{X}$ is the most time-consuming step of the procedure [60]. The SVD computation of a given matrix evolved from methods based on planes rotation [23, 46, 57] to the most usual Golub and Reinsch algorithm (GR) [34]. Even so, GR algorithm performs $\mathscr{O}(\ell^2\kappa+\ell\kappa^2+\kappa^3)$ multiplications, reaching $\mathscr{O}(n^3)$ when $\ell \approx N/2$, its worst computational complexity [60].

It is natural in an SSA-based application to expect to get just a few leading eigentriples. Since they carry a lot of signal information, the cost of a complete decomposition is wasteful. In multivariate SSA (MSSA) cases, performing a low-rank factorization via the Lanczos bidiagonalization algorithm [11, 29] or via randomized SVD [79] are a thrifty (regarding memory) and agile way to reduce the number of operations required for SSA purposes. In the *basic* context, a randomized SSA (rSSA) is proposed in [86], in which the SVD step is carried out by the randomized SVD proposed by Halko et al. [44]. This is done by obtaining an approximation of the trajectory matrix $\mathbf{X}$ using random sampling methods, followed by decomposing the resulting matrix.

Another possible drawback is the decomposition of a trajectory matrix when the TS presents missing data (MD). An imputation method for TS of finite rank is proposed in [35] and called "Caterpillar"-SSA (CSSA). Assuming that the TS is governed by some linear recurring formula (LRF), the CSSA relies on the continuity of the computed component structure to fill the gaps. An improved CSSA (ICSSA) is proposed in [55], modifying the original to reject outliers and make the model parameters more robust to the different temporal patterns. Other methods based on different approaches have been applied in geophysics [88, 90], climatology [65], and finance [16].

3.1.2 Addressed issues

To strike a balance between solving the problem of time-consuming computation associated with SVD and an eventual need for imputation when dealing with MD, we raise the following points at issue:

(P1) A comprehensive solution : *Is it feasible to build or adapt an existing algorithm to simultaneously solve both problems (speed and MD)?*

Throughout the investigation, we proposed applying the NIPALS algorithm as a direct way to get around the gap-filling problem. We made only a few modifications to express the decomposition results in terms of singular values and singular vectors, adapting them to the biplot scheme. The NIPALS algorithm is powerful because of its speed and simplicity [75]. Instead of computing the PCs simultaneously as in SVD, NIPALS performs an iterative regression procedure when factorizing $\mathbf{X}$ and calculating the PCs. NIPALS handles missing values without any imputation [71, 106] since it ignores the blanks when executing the regressions, which is equivalent to setting all missing points to zero in the least-squares objective function.

Besides, it is faster than SVD when applied to large matrices [71]. To compute the first q PCs, the complexity for NIPALS is $\mathscr{O}(\ell\kappa q i)$, in which i is the number of iterations until convergence [95].

(P2) Instability and other weaknesses : *Is the convergence of the chosen algorithm always guaranteed?*

The NIPALS algorithm is very similar to the Power method [22, 67], except that the latter applies to square matrices. Besides, the convergence in the Power method is not always guaranteed if the given matrix is non-diagonalizable [103]. The NIPALS is applied directly to centered and scaled data matrices. According to Geladi and Kowalski [31], NIPALS usually converges in practical situations. But eventually, it does not happen when there are two or more very similar eigenvalues.

Some criticisms are made of the NIPALS algorithm, but not only concerning convergence. Miyashita *et al.* [75] hold that, in some instances, the first principal component may not be obtained. They suggest a modification to the original algorithm to work around this problem. Seasholtz and Gates [89] reinforce that if the algorithm converges in just one step, then the eigenvalue is ambiguous, i.e., it is impossible to say which eigenvalue was computed if the first or any other. They propose to keep using the SVD to decompose a matrix.

3.2 PCs visualization

3.2.1 State of the art

Originally formulated by Gabriel [24], the biplot method was detailed by Gower and Hand [42]. In the context of PCA, biplots are a simultaneous graphical representation of the rows and columns of a multivariate data matrix in reduced-dimensional subspaces [43]. When applying the method, it is essential to consider which configuration to preserve in the subspace, if of the rows (individuals) or the columns (variables). Each choice produces a different biplot with distinct interpretations.

In the GH-biplot construction (or column-metric-preserving biplot), the singular values are fully assigned to the right singular vectors. On the other hand, when the matrix $\mathbf{\Sigma}$ is allocated to the $\mathbf{T}^*$ matrix in (2.40), then it characterizes the row-metric-preserving case. Recently, Balcerowska-Czerniak *et al.* [4] used this concept in the chemometrics context to compound identification using the X-ray diffraction technique; Bassani *et al.* [5] applied the JK-biplots to genomics to discover common patterns of expression; Torres-Salinas *et al.* [97] used this factorization configuration to represent bibliometric indicators through biplots in the information systems field.

Galindo [25] elaborated an alternative possibility called HJ-biplot. On it, $\mathbf{\Sigma}$ is assigned to both $\mathbf{T}^*$ and $\mathbf{P}$ matrices, obtaining a simultaneous representation of both rows and columns of $\mathbf{X}$ with maximum quality. Several scientific fields have newly applied this approach, including hydrology [10], biotechnology [74], environmental science [27], engineering [98], health [20], and sustainable development [26, 69].

Concerning applying biplot methods to time series, mention should be made of Alvarez & Galindo [2]. The authors used the HJ-biplot to study the traffic in communication networks, suggesting a better biplot method performance than the PCA alone, being more informative about time series behavior. In the same direction, Yang *et al.* [107] proposed an algorithm to visualize and analyze large time series using PCA and biplots. In turn, Ivanov & Evtimov [51] used biplots to examine the seasonality of a univariate time series containing temperature anomalies of the Northern Hemisphere.

Nieto *et al.* [77] pointed out the lack of software has been a deterrent to popularizing biplot utilization as a multivariate visualization technique and presented an inferential version of a biplot in R [83], based on bootstrap confidence intervals for the parameters defined by the row and column markers [64]. Still, in R, the GGEBiplotGUI package [7] is another attempt to expand the availability of software based on the method.

3.2.2 Addressed issues

One of the main issues raised throughout this work is visually capturing the series' essence as much as possible and how it behaves over time. The other is to present our contribution to the absence of software concerning biplots.

(P3) Capturing the essence of the TS : *What parameters and adjustments are needed to explore the trajectory matrix in the representation through biplots optimally?*

Consider the conformation of the trajectory matrix $\mathbf{X}$, containing in its rows ℓ κ-lagged vectors and, in its columns, κ ℓ-lagged vectors. Also, note that both cases are representations of subseries of the TS. The ideal situation is to get some balance in the κ-lagged and ℓ-lagged vectors' lengths. The equilibrium is achieved by setting the window length as $\ell = n/2$. This procedure follows the SSA theory, which establishes an ℓ large enough so that the ℓ-lagged vectors (and, in this case, also the κ-lagged vectors) incorporate an essential part of the behavior of the TS. Likewise, the choice of the HJ-biplot to illustrate the $\mathbf{X}$ decomposition is a natural consequence. The method created by Galindo [25] guarantees optimal quality representation of the ℓ rows and the κ columns of $\mathbf{X}$ in the same Euclidean space.

(P4) Revealing the TS periodicity : *How can the biplot method be used to emphasize the visualization of features of the TS studied?*

It is possible to check a periodicity of a TS when using a pair of eigentriples associated with an oscillatory component to construct the HJ-biplot. Nevertheless, this characteristic is not visually highlighted. A possible solution to help bring more contrast to the periodicity is using geometric patterns to detach it. Area biplots approximate the values of elements of a data matrix through the area of specific triangles constructed from the eigenvectors [41]. In this investigation, we propose a different construction to the polygons, in which their areas will help to evaluate the autocorrelation among ℓ-lagged vectors and from which the periodicity will emerge more clearly.

The original area biplot is based on building triangles from the counterclockwise rotation of row markers. On the other hand, our suggested version rotates the column markers to construct the triangles.

(P5) Automating the visualization : *How to automate (and distribute) the suggested method's visualization process to reveal the TS periodicity?*

An R package is a suitable way to collect and distribute codes to be reusable. Submitting it to the Comprehensive R Archive Network (CRAN) is an intelligent decision to get traction with the community. The *devtools* R package [104] provides all the functions necessary to build a package aimed at automating tasks in R.

The approach proposed in this investigation to reveal the periodicity of a TS is based on the method proposed by Gower [41]. However, no computational tool has been implemented so far for the automatic construction of an area biplot. We set out to resolve this by building an R package and submitting it to CRAN.

3.3 TS structural changes

3.3.1 State of the art

A broad spectrum of problems can be addressed through SSA, ranging from exploratory analysis and forecasting in the field of time series analysis to parameter estimation in signal processing [37]. As a result, the scope of applications of SSA is also extensive, including quality control [12, 21, 82], renewable energy [72, 73, 110], and health [99]. Lately, SSA methods have been created to assess structural damage control [8, 66]. They are pretty much based on the idea of finding out the points where the eigenstructure presents some important modification.

This brings us to the problem of detecting heterogeneities in a TS. A series is homogeneous if it is governed by some linear recurrence relation (LRR) [39]. Heterogeneity occurs when the TS goes from one homogeneous state to another due to its exposure to a local perturbation. A disturbance can result in spectrum dispersion, causing a variation over time. One strategy to get around this inconvenience is to segment the original series with some overlaps. Then, for example, the SSA can be applied to each subseries and the results analyzed accordingly. Leles *et al.* [62, 63] proposed a method called overlap-SSA (ov-SSA) using this segmentation approach to analyze non-stationary TS.

Finding the short interval within the TS where the transition occurs characterizes a change-point detection (CPD) problem. Golyandina *et al.* [39] suggested a more comprehensive solution, calling it a structural change detection problem. Their proposal is based on heterogeneity detection and uses a sequential SSA application. 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 series intervals. Moskvina & Zhigljavsky [76] developed a similar algorithm but restricted to a CPD perspective. More recently, Alanqari *et al.* [1] presented another CPD algorithm based on the underlying dynamics of multivariate time series observations through a Spatio-temporal model.

3.3.2 Addressed issues

Complex series in which change points appear can make it challenging to identify the dominant structures visually. We will work to implement a sequential application of the SSA to determine the point where an interruption of the LRR arises, combining it with a metric for CPD through robust trajectory matrix factorization strategies.

(P6) A simple segmentation approach : *How to take advantage of the SVD sensitivity to data contamination in order to build a CPD strategy using SSA?*

In previous work [39, 76], the segmentation scheme determines two disjunct intervals (base and test), taken sequentially from the original series, which means constructing two trajectory matrices in each iteration. Then, the occurrence of a perturbation is evaluated in terms of Euclidean distances related to the (base and test) trajectory matrices. In this investigation, we have in view a particular procedure. Instead of applying a single decomposition method to two different trajectory matrices (base and test) iteratively throughout the series, we intend to apply two different decomposition methods (one robust and the other ordinary) on the same trajectory matrix.

Using a procedure similar to the one used by Rodrigues *et al.* in [85], it seems possible to capture an interruption of the LRR by measuring some distances created by adopting a robust and regular approach to decompose the same matrix of trajectory.

(P7) Creating metrics : *How to handling the distances?*

Moskvina & Zhigljavsky [76] associate the arising of a disturb with an increase in the Euclidean distance between the lagged vectors of the TM (base) and the subspace generated by the eigenvectors of the lag-covariance matrix (test). Differently, we intend to use the differences caused by an eventual change of direction of some PCs (eigenvectors) to compute a metric in the identification of interruption of the LRR.

3.4 The research conduction piece by piece

This research project was designed to publish a coherent and relevant set of scientific articles in journals with recognized international merit selection committees. In addition, the proposal also intended to offer complimentary contributions in the form of posters and oral communications in scientific meetings and proceedings of international conferences. The following publications list seeks to answer the questions formulated and explained in Sections 3.1 to 3.3. One can find the entire content of these works in Part II.

3.4.1 The main point

The first research paper (Article I) brings the main concepts behind a new SSA-HJ-biplot method to support the SSA grouping stage. It is a visualization method based on biplots that allow the graphical representation of the spectral structure of a TS. Its interpretability helps analyze the series through eigenvectors' representation characteristics. Article I explores and formalizes the general ideas presented in Article V, going beyond by considering a series containing missing data in applying the method. In sequence, Article II treats the details concerning the interpretability enhancement of the method, caring to extract information as much as possible about the SSA-HJ-biplots. Both deal with issues P1, P2, and P3. In addition, Article I and Article II partially address question P4, as the visual perception of the TS periodicity requires some effort when evaluated directly from a HJ-biplot.

3.4.2 Improving the method

The third work (Article III) sought to strengthen the visualization of the separation of TS components by applying the SSA-HJ-biplot method in series where structural changes occur. First, the TS is segmented concerning identified change points. Next, one can apply the SSA-HJ-biplot over each interval, making visual interpretation easier. A new approach for change point detection is suggested based on sudden shifts in the direction of the PCs. It is evaluated by computing the difference by applying two decomposition methods (robust and ordinary) on the same trajectory matrix. These differences will be more accentuated when there is an eventual change in the direction of some PCs (eigenvectors) in case of interruption of the LRF. Article III focuses on bringing solutions to issues P6, and P7.

3.4.3 A different approach

Article IV discusses a multivariate visualization technique to estimate the dominant periodicities of a time series. It consists of a version of the area biplot method built from PCs associated with pairs of singular values close to each other. After pinning a biplot vector of interest (i.e., some loading vector related to a lagged vector), the remaining are rotated at 90 degrees. Triangles are created connecting the origin of the factorial axes to the endpoints of these vectors. These polygons establish patterns that provide visual information regarding the autocorrelation between the corresponding lagged vectors. Periodicity emerges from these patterns. The method gave rise to a Software (package) called areabiplot, implemented in R [83]. Article IV addresses the issue raised in point P4.

3.4.4 The starting point

Article V is a guide to validate the initial ideas, serving as a compass for subsequent work. Article I and Article II elaborate and give depth to the concepts presented in that initial essay. Article III and Article IV intend to solve the difficulties faced in the first experiments conducted while writing that embryonic monograph. In this way, this content is somehow related to all the questions raised from P1 to P7.

3.4.5 Preparation for the graphical tool

Article VI's content is related to applying biplot methods from factoring matrices using PLS regression algorithms. However, this article was a first attempt at developing the *areabiplot* R package [91]. As mentioned earlier, the lack of software prevents the widespread use of biplot methods. In this sense, the code implemented for Article VI served as the basis for developing the function built into the *areabiplot* package. Since the main application of this software is directed to the results of the trajectory matrix decomposition of the SSA method, the work developed in Article VI helped in answering question P5.

3.4.6 Graphical tool implementation

The last piece consists of a Software implementation in R Programming Environment [83], called *areabiplot* [91]. The R package areabiplot was conceived under MIT License, giving users express permission to reuse code for any purpose. The tool can be used both to reveal an estimate for the periodicity of a TS (context of this research) and to estimate the elements of a data matrix (original context suggested by Gower). The Software is intrinsically connected to the issue P5.

From the publication date, the areabiplot package has more than 6000 downloads from the cloud.r-project.org CRAN mirror (3.1), as data collected through the following code:

```r
devtools::install_github("metacran/cranlogs")

library(cranlogs)

X = cran_downloads(packages = "areabiplot", from = "2021-03-10",
               to = "2022-09-30")

inds <- seq(as.Date("2021-03-10"), as.Date("2022-09-30"), by = "day")

Y <- ts(X[,2], start = c(2021, as.numeric(format(inds[1], "%j"))),
        frequency = 365)
```

```


plot(Y,type="l",col="blue",ylab="Number of downloads per day", main="R package areabiplot",
     font.main = 3, cex.main = 1.2)
```

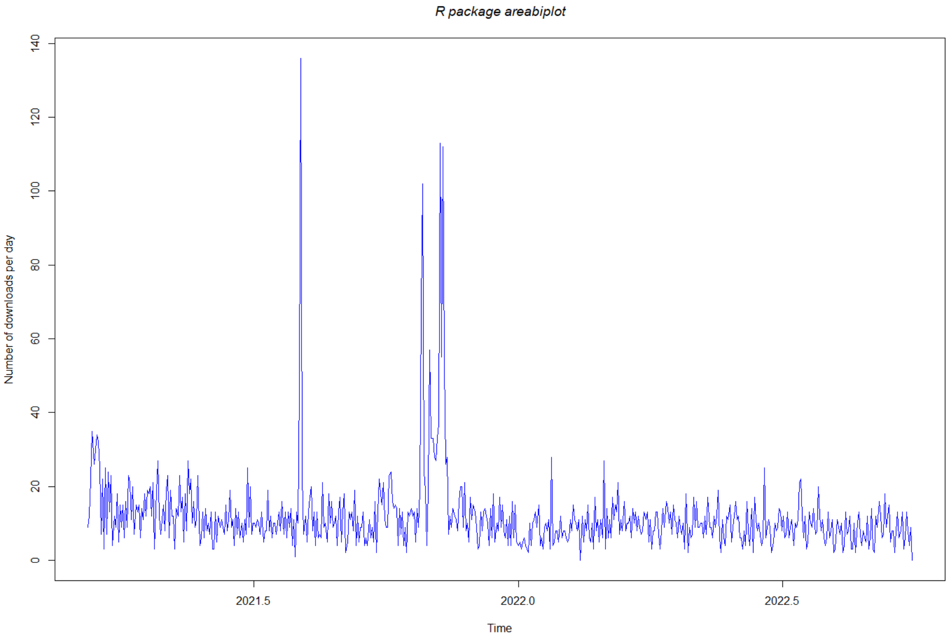

Figure 3.1: Number of downloads per day of the *areabiplot* package from the cloud.r-project.org CRAN mirror.

Part II

Research

Chapter 4

Article I

Time Series components separation based on Singular Spectral Analysis visualization: an HJ-biplot method application

Published:

Silva, A., Freitas, A. (2020). Time Series components separation based on Singular Spectral Analysis visualization: an HJ-biplot method application. Statistics, Optimization and Information Computing, 8(2):346:358.

DOI https://doi.org/10.19139/soic-2310-5070-897

STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **8**, June 2020, pp 346–358.
Published online in International Academic Press (www.IAPress.org)

Time Series components separation based on Singular Spectral Analysis visualization: an HJ-biplot method application

Alberto Oliveira da Silva [1,2,*], Adelaide Freitas [1,2]

[1] *Department of Mathematics, University of Aveiro, Portugal*
[2] *Center for Research & Development in Mathematics and Applications (CIDMA), University of Aveiro, Portugal*

Abstract The extraction of essential features of any real-valued time series is crucial for exploring, modeling and producing, for example, forecasts. Taking advantage of the representation of a time series data by its trajectory matrix of Hankel constructed using Singular Spectrum Analysis, as well as of its decomposition through Principal Component Analysis via Partial Least Squares, we implement a graphical display employing the biplot methodology. A diversity of types of biplots can be constructed depending on the two matrices considered in the factorization of the trajectory matrix. In this work, we discuss the called HJ-biplot which yields a simultaneous representation of both rows and columns of the matrix with maximum quality. Interpretation of this type of biplot on Hankel related trajectory matrices is discussed from a real-world data set.

Keywords Singular Spectrum Analysis, NIPALS algorithm, Biplots

AMS 2010 subject classifications 62-09, 62M15

DOI: 10.19139/soic-2310-5070-897

1. Introduction

Time series (TS) data emerge in many scientific fields. The analysis of this type of data can be based on the time or the frequency domain. When its evaluation is based on time-domain, parametric models are proposed. When the study is in terms of frequency-domain, approaches using non-parametric models, such as the Spectral Analysis, are usually developed. A TS can be classified as stationary if properties of the underlying generation process do not change over time, e.g., the mean and the variance are constants. Otherwise, the TS is said to be nonstationary and, in this case, can reveal a trend, i.e., a smooth and slowly varying part of the series. Any TS can be decomposed into a variety of components, for instance, (1) global trend (only for nonstationary TS), (2) oscillatory shape, e.g., a seasonal variation, and (3) irregular component, or noise. Generally, it is possible to model the trend by mean of a low degree polynomial function

$$f(t) = a_0 + a_1 t + \cdots + a_m t^m$$

where t is the time, a_i are real constants and m denoting here the degree of the polynomial function. Likewise, the regular oscillatory, with a period p, can be synthesized as a function $g(t)$ defined by a linear combination of sines and cosines with constant coefficients, i.e., expressed as a Fourier series

$$g(t) = \sum_{m=0}^{p} (\alpha_m \cos(\omega_m t) + \beta_m \sin(\omega_m t))$$

[*] Correspondence to: Alberto Oliveira da Silva (Email: albertos@ua.pt). Center for Research & Development in Mathematics and Applications (CIDMA), University of Aveiro. 3810-193 Aveiro, Portugal.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

where $\omega_m = 2\pi m/p$. Thus, a classic model of a TS may be obtained by a systematic component, given by the addition (or multiplication) of $f(t)$ and $g(t)$, plus a random component defined by a white noise, $\{\varepsilon_t\}$, which is independent of the signal.

Singular Spectrum Analysis (SSA) is a methodology used in Time Series Analysis for many different purposes, such as exploratory inspection and forecasting [6]. Differently from other techniques, the SSA does not take into consideration whether the model is additive or multiplicative. Besides, no trend model or previous knowledge about the periodicity of the series is required to apply SSA [5]. In the basic version of SSA, the object is a one-dimensional real-valued TS and consists of two successive stages. In the first one, the decomposition stage, the TS is transformed into a Hankel matrix, named as trajectory matrix, on which the Singular Value Decomposition (SVD) is applied, resulting in the summation of rank-one matrices. Next, in the reconstruction stage, some of these rank-one matrices are grouped appropriately (grouping step). From those groups, in the so-called diagonal averaging step, an approximation for the original object or components of the TS, like trend, oscillatory shape, and noise, separately, can be obtained. The form of a TS and the eigenvectors resulting from the decomposition of its trajectory matrix are related. Thus, the graphical representation of these eigenvectors is a proper way to visualize the components of a TS. For example, a plot of the first eigenvector is suitable to evaluate the existence of a trend. In turn, a scatterplot of two eigenvectors close to each other can determine a geometric pattern in some cases, being useful for assessing the existence of a seasonal component ([6] for more details). This work presents a graphical tool based on pairs of eigenvectors that, in the same plot, combines more information and can lead to the identification of relevant features of a TS.

For any matrix $\mathbf{Z}$ whose rank is r, the SVD factorization provides the best approximation matrix $\tilde{\mathbf{Z}}$, in the least-squares sense, whose rank is less than r. Further, if $\tilde{\mathbf{Z}}$ has rank 2, then the SVD allows practical graphical representations of both rows and columns of the approximation matrix employing the Biplot method [2, 3]. Biplots provide easier interpretations and are much more informative than the traditional scatterplots, beyond that might facilitate the work in the grouping step in SSA. Several types of biplots can be constructed depending on how the three factors identified by SVD are aggregated to obtain only two factors. Herein, the option is the HJ-biplot method proposed by Galindo (1986), which yields a simultaneous representation of both rows and columns of $\tilde{\mathbf{Z}}$ with maximum quality [3].

Taking the Hankel-related trajectory matrix arisen from Basic SSA, an HJ-biplot type exploratory tool is constructed to visualize and identify patterns in nonstationary TS, and its interpretation is discussed. This work suggests the factorization of the trajectory matrix using the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm [14] since it is capable of dealing with missing values, commonly in TS, without employing any imputation method [13, 15]. For complete matrices, it is essential to point out that the NIPALS algorithm provides equivalent results to their factorization via SVD concerning the singular vectors and the singular values.

The paper is organized as follows. In Section 2, a short overview of theoretical background related to methods involved in this work is provided. In Section 3, the proposed biplot approach to the SSA method and its interpretation are discussed. In Section 4, the suggested technique is performed on a real-world TS using the statistical software R [12]. Some R-code fragments are indicated. Conclusions are presented in Section 5.

2. Methods

2.1. *Basic Singular Spectrum Analysis*

Consider a real-valued TS $Y = (y_1, , y_n)$ of length n. Basic SSA is a model-free tool used to recognize and identify the structure of Y [6]. As aforementioned above, the SSA consists of two complementary stages: decomposition and reconstruction. Each stage in this algorithm includes two steps.

First Stage: Decomposition.

Let ℓ ($1 < \ell < n$) an integer value representing the so-called window length, as well as $\kappa = n - \ell + 1$. Hereupon, the embedding procedure consists in representing Y in κ lagged vectors $\mathbf{x}_1, \cdots, \mathbf{x}_\kappa$, each one of size ℓ (ℓ-lagged vectors), i.e., $\mathbf{x}_j = [y_j, , y_{j+\ell-1}]'$, $1 \leq j \leq \kappa$. This sequence of κ vectors forms the trajectory matrix $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_\kappa]$, a Hankel matrix that has in its columns the ℓ-lagged vectors. Thus, the trajectory matrix consists of the transformation of a time series into a Hankel matrix by means of an embedding operator $\mathcal{T}$, such that

$$\mathcal{T}(Y) = \mathbf{X} = \begin{bmatrix} y_1 & y_2 & y_3 & \cdots & y_\kappa \\ y_2 & y_3 & y_4 & \cdots & y_{\kappa+1} \\ y_3 & y_4 & y_5 & \cdots & y_{\kappa+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_\ell & y_{\ell+1} & y_{\ell+2} & \cdots & y_n \end{bmatrix}$$

Next, SVD is executed for the trajectory matrix $\mathbf{X}$ resulting

$$\mathbf{X} = \sum_{i=1}^{d} \sqrt{\lambda_i}\mathbf{u}_i\mathbf{v}_i' \,,$$

where $d = rank(\mathbf{X})$, λ_i, $i = 1, \cdots, d$, are the eigenvalues of the matrix $\mathbf{X}'\mathbf{X}$ arranged in decreasing order of magnitudes ($\lambda_i > 0$), and associated to the orthonormal system of the eigenvectors $\{\mathbf{v}_1, \cdots, \mathbf{v}_d\}$ of $\mathbf{X}'\mathbf{X}$, and

$$\mathbf{u}_i = \mathbf{X}\mathbf{v}_i / \sqrt{\lambda_i} \,.$$

The elements of the triple $(\sqrt{\lambda_i}, \mathbf{u}_i, \mathbf{v}_i)$ are also known as singular values, left and right singular vectors of $\mathbf{X}$, respectively. Besides, defining

$$\mathbf{X}_i = \sqrt{\lambda_i}\mathbf{u}_i\mathbf{v}_i',$$

$\mathbf{X}$ can be represented by a sum of d 1-rank matrices, i.e.,

$$\mathbf{X} = \mathbf{X}_1 + \cdots + \mathbf{X}_d. \tag{1}$$

Second Stage: Reconstruction.
Once the expansion (1) has been determined, the second stage of SSA starts with the partitioning of the index set $\{1, \cdots, d\}$ into disjoints subsets I_j, $j = 1, \cdots, p$, leading to the decomposition given as follows

$$\mathbf{X}_I = \mathbf{X}_{I_1} + \cdots + \cdots \mathbf{X}_{I_p} \,,$$

where $\mathbf{X}_I = \sum_{i \in I} \mathbf{X}_i$. The intention of the grouping procedure is the separation of the additive components of the TS [7]. The objective of the next phase, the diagonal averaging step, is to take each matrix $\mathbf{X}_{I_j}$ of the grouping step and transform it into a Hankel matrix $\tilde{\mathbf{X}}_{I_j}$, converting the result into a TS [6] by means of

$$\tilde{Y}^{I_j} = \mathcal{T}^{-1}\left(\tilde{\mathbf{X}}_{I_j}\right).$$

2.2. *PCA through NIPALS*

The NIPALS algorithm belongs to the Partial Least Squares (PLS) family, a set of iterative algorithms that implement a wide range of multivariate explanatory and exploratory techniques. The NIPALS is designed as an iterative estimation method for Principal Component Analysis (PCA), that computes the principal components through an iterative sequence of simple ordinary least squares regressions [13, 14]. NIPALS on $\mathbf{X}$ produces a decomposition of the matrix so that the principal components are computed one-by-one [13], providing equivalent results to the SVD concerning the singular vectors and the singular values. A particular feature of the NIPALS algorithm is that, in each iteration, only present data are considered in the regressions performed, ignoring the missing elements. It is equivalent to defining all missing points as zero in the least-squares objective function. Therefore, in the case of missing data, no imputation method is necessary when applying NIPALS, which can be evaluated as an advantage over the SVD.

Considering a d-rank trajectory matrix, the algorithm decomposes $\mathbf{X}$ as a sum of d 1-rank matrices in terms of the outer product of two vectors, a score $\mathbf{t}_i$ and a loading $\mathbf{p}_i$, so that

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1' + \cdots + \mathbf{t}_d\mathbf{p}_d'.$$

The elements of the scores vector $\mathbf{t}_i$ are the projections of the sample points on the principal component (PC) direction, while each loading in $\mathbf{p}_i$ is the cosine of the angle between the component direction vector and a variable axis [4]. The NIPALS first computes $\mathbf{t}_1$ and $\mathbf{p}_1$ from $\mathbf{X}$ and, then, the outer product $\mathbf{t}_1\mathbf{p}_1'$ is subtracted from $\mathbf{X}$ to calculate the residual matrix $\mathbf{E}_1$. After, $\mathbf{E}_1$ is used to compute $\mathbf{t}_2$ and $\mathbf{p}_2$, and the residual $\mathbf{E}_2$ is calculated subtracting $\mathbf{t}_2\mathbf{p}_2'$ from $\mathbf{E}_1$, and so on until to obtain $\mathbf{t}_d$ and $\mathbf{p}_d$. The NIPALS algorithm is shown in Algorithm 1.

Algorithm 1 (NIPALS)

```
Input: $\mathbf{E}_0 = \mathbf{X}$
Output: $\mathbf{P} = [\mathbf{p}_1 \cdots \mathbf{p}_d],\ \mathbf{T} = [\mathbf{t}_1 \cdots \mathbf{t}_d]$
  For all $i = 1, \cdots d$ Do
     1) initialize $\mathbf{t}_i$
     2) Repeat
        $\mathbf{p}_i = \mathbf{E}_{i-1}'\mathbf{t}_i/\mathbf{t}_i'\mathbf{t}_i$
        $\mathbf{p}_i = \mathbf{p}_i/||\mathbf{p}_i||$
        $\mathbf{t}_i = \mathbf{E}_{i-1}\mathbf{p}_i$
        until convergence of $\mathbf{p}_i$
     3) $\mathbf{E}_i = \mathbf{E}_{i-1} - \mathbf{t}_i\mathbf{p}_i'$
  End For
```

From the internal relations in each iteration of the NIPALS algorithm, and after normalizing $\mathbf{t}_i$, such that

$$\mathbf{t}_i^* = \mathbf{t}_i/||\mathbf{t}_i|| \iff \mathbf{t}_i = \sqrt{\mathbf{t}_i'\mathbf{t}_i}\,\mathbf{t}_i^*,$$

the following equations can be verified [13]:

$$\mathbf{E}_{i-1}'\mathbf{E}_{i-1}\mathbf{p}_i = \lambda_i\mathbf{p}_i \tag{2}$$

$$\mathbf{E}_{i-1}\mathbf{E}_{i-1}'\mathbf{t}_i^* = \lambda_i\mathbf{t}_i^*, \tag{3}$$

where $\lambda_i = \mathbf{t}_i'\mathbf{t}_i$ is the eigenvalue of both matrices $\mathbf{E}_{i-1}'\mathbf{E}_{i-1}$ and $\mathbf{E}_{i-1}\mathbf{E}_{i-1}'$, as well as $\mathbf{p}_i$ and $\mathbf{t}_i^*$ are their corresponding eigenvectors with unit norm. In the first iteration of the algorithm, i.e., for $i = 1$, the equations in (2) and (3) are reduced to

$$\begin{aligned} \mathbf{X}'\mathbf{X}\,\mathbf{p}_1 &= \lambda_1\,\mathbf{p}_1 \\ \mathbf{X}\mathbf{X}'\,\mathbf{t}_1^* &= \lambda_1\,\mathbf{t}_1^* \end{aligned}$$

and it is clear that the first normalized score vector and the first loading vector are exactly the first left and right singular vectors of $\mathbf{X}$, respectively, as well as $\sqrt{\mathbf{t}_1'\mathbf{t}_1}$ returns the first singular value of $\mathbf{X}$. Moreover, for $i > 1$, and considering the column space of $\mathbf{X}$, the NIPALS computes the i-th principal component over the orthogonal complement of the subspace $\mathbf{t}_1\mathbf{p}_1' + \cdots + \mathbf{t}_{i-1}\mathbf{p}_{i-1}'$, which is equivalent to the SVD approach of imposing the orthogonality restriction among singular vectors when maximizing its objective function. It implies yet that, $\forall$ $i = 1, \cdots, d$, $\mathbf{t}_i^*$ and $\mathbf{p}_i$ are equal to the left and right singular vectors of the SVD of $\mathbf{X}$, respectively, and each $\sqrt{\mathbf{t}_i'\mathbf{t}_i}$ is the corresponding singular value. The NIPALS decomposition of $\mathbf{X}$ is then given by

$$\mathbf{X} = \sqrt{\mathbf{t}_1'\mathbf{t}_1}\mathbf{t}_1^*\mathbf{p}_1' + \cdots + \sqrt{\mathbf{t}_d'\mathbf{t}_d}\mathbf{t}_d^*\mathbf{p}_d'. \tag{4}$$

Defining the matrix $\mathbf{\Sigma}$ as a diagonal matrix containing the singular values $\sqrt{\mathbf{t}_i'\mathbf{t}_i}$ arranged in decreasing order, one can write the matrix form of the expansion (4) as

$$\mathbf{X} = \mathbf{T}^*\mathbf{\Sigma}\mathbf{P}', \tag{5}$$

where $\mathbf{T}^*$ is the (unit-)scores matrix whose column vectors $\mathbf{t}_i^*$ are orthonormal, and $\mathbf{P}$ is the (unit-)loadings matrix whose column vectors $\mathbf{p}_i$ are also orthonormal.

2.3. Biplots

Biplot is a 2- or 3-dimensional graph that allows the joint plotting of both objects and variables of multivariate data sets [2]. Biplots can reveal essential characteristics of multivariate data structure, e.g., patterns of correlations between variables or similarities between observations [8]. Consider a d-rank target $(\ell \times \kappa)$ data matrix $\mathbf{X}$ with factorization in the form

$$\mathbf{X} = \mathbf{G}\mathbf{H}', \tag{6}$$

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. Thus, a (2-dimensional) biplot displays both row markers $\mathbf{g}_1, \cdots, \mathbf{g}_\ell$ and column markers $\mathbf{h}_1, \cdots, \mathbf{h}_\kappa$ of $\mathbf{X}$, such that the inner product $\mathbf{g}_i'\mathbf{h}_j$ provides an approximation to the element x_{ij} of $\mathbf{X}$ [9]. Based on SVD of $\mathbf{X}$, many factorization form of $\mathbf{X}$ can be taken into account, and hence different choices of $\mathbf{G}$ and $\mathbf{H}$, and types of biplots [8]. From the equation (5), and considering $\mathbf{G} = \mathbf{T}^*$ and $\mathbf{H} = \mathbf{P}\mathbf{\Sigma}$, the resultant factorization is characterized by preserving the column metrics of $\mathbf{X}$, and the associated biplot is called Gabriel biplot, or classic biplot or covariance biplot or GH-biplot. In this case, if $\mathbf{X}$ has been centered by columns, this type of biplot satisfies the following properties [9]:

- The norm of a column marker $\mathbf{h}_j$ is proportional to the standard deviation of the respective variable;
- The cosine of the angle formed by column markers approximates the correlation between the related variables;
- The columns are better represented than the rows in terms of quality.

On the other hand, by defining $\mathbf{G} = \mathbf{T}^*\mathbf{\Sigma}$ and $\mathbf{H} = \mathbf{P}$, this factorization will preserve the metric of the rows in the so-called form biplot or JK-biplot, in which the Euclidean distances between the row markers approximate the Euclidian distances between the respective individuals in the full space, and the quality of representation of the rows is better than the columns.

Based on SVD, a different type of biplot, called HJ-biplot, was proposed in 1986 by Galindo [3] in which optimal quality representation of the ℓ rows and the κ columns of $\mathbf{X}$ is ensured in the same Euclidean space. An HJ-biplot version based on NIPALS instead of the SVD can be constructed taking the rows of the matrix $\mathbf{J} = \mathbf{T}^*\mathbf{\Sigma}$ as row markers, and the rows of the matrix $\mathbf{H} = \mathbf{P}\mathbf{\Sigma}$ as column markers of $\mathbf{X}$. Indeed, from the NIPALS decomposition in (5), results that

$$\mathbf{X}\mathbf{P} = \mathbf{T}^*\mathbf{\Sigma}\mathbf{P}'\mathbf{P} = \mathbf{T}^*\mathbf{\Sigma}.$$

So, the ℓ rows of the matrix $\mathbf{J} = \mathbf{T}^*\mathbf{\Sigma}$ correspond to the projections of the ℓ points represented by the rows (individuals) of $\mathbf{X}$ onto the subspace spanned by the loading vectors $\mathbf{p}_1$ and $\mathbf{p}_2$, that is, the best-fit two-dimensional subspace for $\mathbf{X}$. Likewise, the κ rows of the matrix $\mathbf{H} = \mathbf{P}\mathbf{\Sigma}$ coincide with the projections of the κ points expressed by the columns (variables) of $\mathbf{X}$ onto the subspace spanned by the normalized score vectors $\mathbf{t}_1^*$ and $\mathbf{t}_2^*$, as seen below:

$$(\mathbf{T}^*)'\mathbf{X} = (\mathbf{T}^*)'\mathbf{T}^*\mathbf{\Sigma}\mathbf{P}' \Longleftrightarrow \mathbf{X}'\mathbf{T}^* = \mathbf{P}\mathbf{\Sigma}$$

From (5) follows that both row and column representations of $\mathbf{X}$, $\mathbf{X}\mathbf{P}$ and $\mathbf{X}'\mathbf{T}^*$, respectively, are related since

$$\mathbf{B} = \mathbf{X}'\mathbf{A}\mathbf{\Sigma}^{-1} \quad \text{and} \quad \mathbf{A} = \mathbf{X}\mathbf{B}\mathbf{\Sigma}^{-1},$$

where $\mathbf{B} = \mathbf{X}'\mathbf{T}^*$ and $\mathbf{A} = \mathbf{X}\mathbf{P}$. It means that the coordinates of the variables can be expressed as a weighted average of the coordinates of the individuals, and vice-versa. As a consequence, it allows the representation of the rows and columns in the same cartesian coordinates system with optimal quality of representation [3, 9]. However, in the HJ simultaneous representation, the inner product $\mathbf{j}_i'\mathbf{h}_j$ does not provide an approximation to the element x_{ij}, $i = 1, 2, \cdots, \ell$ and $j = 1, 2, \cdots, \kappa$, of $\mathbf{X}$ anymore, and consequently

$$\mathbf{X} \neq \mathbf{J}\mathbf{H}'$$

Accordingly, the interpretation of the HJ-biplot representation can be performed as follows:

- The distance between points corresponds to how different the associated individuals are (dissimilarities), just like in JK-biplots;
- As it occurs in GH-biplot, the size of an arrow (variable) is proportional to the standard deviation of the associated variable, i.e., the longer the arrow, the greater the correspondent standard deviation;
- The cosine of the angle between arrows approximates the correlation between the variables they represent. Thus, if the angle is next to 90 degree it indicates a poor correlation, while an angle close to 0 degree or 180 degree suggests a strong correlation, being positive in the first case and negative in the other.

3. The SSA-HJ-Biplot

In the basic version of the SSA, the trajectory matrix that will be decomposed by the NIPALS algorithm has some peculiarities in relation to the usual multivariate data matrix. Instead of individuals and variables, the rows and columns of the Hankel trajectory matrix represent κ-lagged and ℓ-lagged vectors of a univariate time series, respectively. A row marker, determined by the rows of $\mathbf{T}^*\mathbf{\Sigma}$, i.e., $\mathbf{j}_i' = \mathbf{t}_i'$, $i = 1, \cdots, \ell$, is depicted as a point in the SSA-HJ-biplot and corresponds to a κ-lagged vector. Each HJ-biplot point may receive a label from identifying the period, e.g., month or year, in which that κ-lagged vector begins, and thus improve the graph interpretation regarding the data. It means that the points in the SSA-HJ-biplot can represent not only the κ-lagged vectors that start in a given period but also the month itself. In turn, an arrow represents the column marker associated with a ℓ-lagged vector. An SSA-HJ-biplot uses any 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. In its turn, each PC is associated with a TS component, e.g., trend, seasonality, and noise, explaining a proportion of the variability of the data, given by

$$PC_{\%}(i) = \frac{\mathbf{t}_i'\mathbf{t}_i}{\sum_{j=1}^{d} \mathbf{t}_j'\mathbf{t}_j}. \tag{7}$$

Some auxiliary graphs can reveal this relationship between a PC and a TS component. For instance, in a scree plot of $\sqrt{\mathbf{t}_i'\mathbf{t}_i}$ [5], most of the time the first principal components are related to highlighted singular values, indicating an association with the trend. Once these PCs are identified, one can visualize the trend plotting each one of these PCs against an index $j = 1, \cdots, \kappa$. Some precautions are necessary to obtain the best results when constructing the SSA-HJ-biplot. For example, the window length ℓ has to be large enough so that each ℓ-lagged vector captures a substantial part of the behavior of the TS [6]. Still, but at the same time, it should permit the interpretability of the graphics display. A window length equals to $n/2$ provides both capabilities because it allows for a most detailed decomposition [6]. Beyond that, it is worth keeping in mind, the higher the percentage of variability explained, the better the quality of the adjustment of the SSA-HJ-biplot [3].

The interpretation of the SSA-HJ-biplot for a Hankel trajectory matrix is performed as follows:

- Proximity of points. Biplot points whose euclidean distances are small imply similarity in the behavior of the associated κ-lagged vectors; hence, if there is a natural number π such that

  $$||\mathbf{g}_t - \mathbf{g}_{t+\pi}|| \approx 0, \quad \forall t\,,$$

  then,

  $$||(y_t, \cdots, y_{t+\kappa-1}) - (y_{t+\pi}, \cdots, y_{t+\pi+\kappa-1})|| \approx 0, \quad \forall t\,.$$

  This fact leads to suspect that $(y_t, \cdots, y_{t+\kappa-1}) = (y_{t+\pi}, \cdots, y_{t+\pi+\kappa-1})$, for all t, which means that the TS might have periodic fluctuations (seasonality) with period π;
- Length of biplot vectors. If some arrows have roughly the same size, this indicates that the correspondent ℓ-lagged vectors have standard deviation also close; hence, if there is a natural number $\tau < \kappa$ such that

  $$||\mathbf{h}_\tau|| \approx ||\mathbf{h}_{\tau+j}||, \quad \forall j = 1, 2, \cdots, \kappa - \tau\,,$$

then,

$$var(\mathbf{x}_\tau) \approx var(\mathbf{x}_{\tau+j}), \quad \forall j = 1, 2, \cdots, \kappa - \tau\,,$$

which suggests that the TS $(y_1, \cdots, y_t, \cdots, y_n)$ might be generated by a variance stationary process for $t > \tau$;

- Angle between two biplot vectors. If the angle between two arrows is close to 90 degrees, a correlation near to zero is expected which means no similar behaviors of the respective ℓ-lagged vectors. If the angle between two arrows is next to 0 (180, resp.) degree, then a strong and positive (negative, resp.) correlation between the two ℓ-lagged vectors associated is suspected. Thus, if there exists a natural number $\eta < \kappa$ such that

  $$|\cos\left(\angle(\mathbf{h}_j, \mathbf{h}_{j+\eta})\right)| \approx 1, \quad \forall j = 1, \cdots, \tau,$$

  for some $\tau \in \{1, \cdots, \kappa - \eta\}$, then

  $$|cor(\mathbf{x}_j, \mathbf{x}_{j+\eta})| \approx 1, \quad \forall j = 1, \cdots, \tau.$$

  Consequently, there are real constants a and $b \neq 0$ such that $y_{j+\eta} \approx a + by_j$, for $j = 1, \cdots, \tau$. This means that $(y_1, \cdots, y_\kappa)$ might be generated by a τ-th order stationary process.

As a rule, a singular value represents the contribution of the corresponding PC in the form of the TS. As the trend generally characterizes the shape of a TS, its singular values are higher than the others, that is, they are the first eigenvalues [1]. It means that the directions of the highest variability of a time series are related to the trend, and as mentioned before, it can be modeled employing a low degree polynomial function, such as in (1). Still, when two singular values are close enough, i.e.,

$$\sqrt{\mathbf{t}_i'\mathbf{t}_i} \approx \sqrt{\mathbf{t}_h'\mathbf{t}_h},$$

this is an evidence of the formation of plateaus in the scree plot and indicates that the associated PC is informative about the oscillatory components of the TS [6], as long as the principal component explains high variability of the data. It occurs because the periodical shape of a TS can be expressed as a Fourier series, as in (1). Consequently, for each m, the sine and cosine of $(\omega_m t)$ will determine orthogonal directions of a pair of PC, and the associated singular values will be close to each other. Apart from the interpretation of similarities using Euclidean distances, the projections of the biplot points into a PC axis helps in the identification of the TS components. If the projections evolve in time in the same principal component growth direction, it means this PC is associated with the trend, as well as the trend is crescent. Otherwise, if the evolution in time occurs in the opposite direction, the trend is decreasing. Moreover, this procedure allows detecting a trend change direction quickly. In turn, a pattern in terms of proximity between the projections can occur. In this case, it indicates the correspondence between the PC and the periodicity of the TS.

4. Example

In this section, a real-world TS is used to demonstrate the capabilities of the SSA-HJ-biplot. Two R-libraries accompany this study: `imputeTS` [10] which contains a function for obtaining graphical representations of TS with missing data, and `nipals` [16] which is aimed to perform PCA using NIPALS algorithm. This TS (`data` in the R-code below) contains the records of the carbon dioxide concentration in the Earth's atmosphere, measured monthly from January of 1965 to December of 1980 ($n = 192$) at an observing station on Mauna Loa in Hawaii [11]. This time series is referred to as TS CO2 in this work and presents missing data, as can be verified in Figure 1. For constructing this plot, the following R-code was used:

```
> Y <- ts(data, start=1965, end=1980, frequency = 12)
> library(imputeTS)
> plotNA.distribution(Y, main="Monthly Dioxide Carbon Concentration
   + (ts CO2)",xlab="", ylab="CO2 concentration")
```

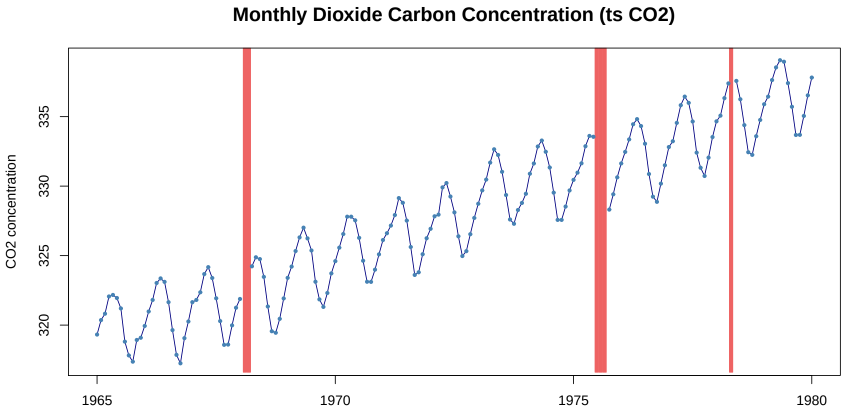

Figure 1. Records of carbon dioxide (CO2) concentration in the Earth's atmosphere measured monthly from January of 1965 to December of 1980 at an observing station on Mauna Loa in Hawaii. Missing data are represented in red.

As mentioned before, the NIPALS algorithm handles the missing values conveniently without the need to complete the data. For the execution of the NIPALS algorithm on the $(\ell \times \kappa)$ trajectory matrix (X), the following R-code was used:

```
> n = 192; L = ceiling(n/2); K = n - L + 1
> X = outer((1:L),(1:K),function(x,y) data[(x+y-1)])
> library(nipals)
> res = nipals(X, center = TRUE, scale = FALSE)
```

In the embedding step of the SSA, the window length used was $\ell = n/2 = 96$ observations, resulting in $\kappa = 97$. The R-object res above contains, among others, the (unit-)scores matrix $\mathbf{T}^*$, the diagonal matrix $\mathbf{\Sigma}$, and the (unit-)loadings matrix $\mathbf{P}$ related to the NIPALS decomposition (5) of the trajectory matrix, and are computed as:

```
> Tstar = res$scores
> Sigma = diag(res$eig)
> P = res$loadings
```

respectively. Table 1 shows the proportion of variability explained by the ten first PC and calculated according to (7). Therefore, it turns out that the five first PC explain about 98% of the data variability, with less than 2% remaining from the 6th PC onwards. In its turn, Figure 2 brings the scree plot, in which the dominant singular value $\sqrt{\mathbf{t}_1'\mathbf{t}_1}$ represents the 1st PC and explains about 67% of the data variability, being associated with the trend. Figure 3 shows the first SSA-HJ-biplot, in which the biplot points are labeled with the month and year when the κ-lagged vector starts, ranging from January of the first year (J1) to December of the eighth year (D8). The biplot markers displayed in Figure 3 were obtained using the following R-code:

```
> J = Tstar %*% Sigma
> H = P %*% Sigma
> points(J[,1],J[,2],col="blue")
> arrows(0,0,H[,1],H[,2],length=0.1,lwd=1, col="grey")
```

Table 1. Proportion of variability explained by the ten first principal components

PC (i)	Variance (%)	PC (i)	Variance (%)
1	66.830	6	0.376
2	14.707	7	0.316
3	14.444	8	0.291
4	1.096	9	0.106
5	1.070	10	0.080

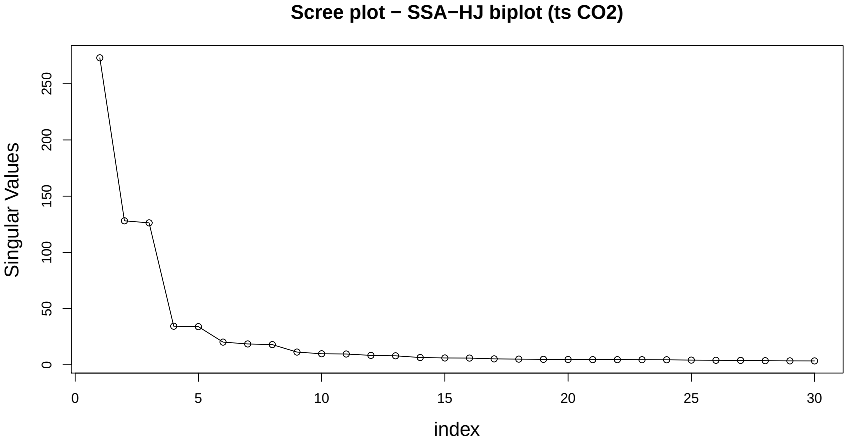

Figure 2. Scree plot of the singular values of the trajectory matrix after NIPALS decomposition.

```
> arrows(0,0,H[1:6,1],H[1:6,2],length=0.1,lwd=1, col="black")
```

It can be verified in Figure 3 that as the 1st PC increases, the projection of the points representing the κ-lagged vectors also progress over time, indicating a crescent trend. On the other hand, the projections into the 2nd PC determine a pattern in terms of proximity regarding the months, i.e., projections of points with the same tag, e.g., J1, $\cdots$,J8, always falls close to the same coordinate. This pattern repeats for all months and indicates, then, a periodicity of twelve months. Therefore, the first SSA-HJ-biplot combines different structural components of the TS CO2, since the 1st PC is related to trend and the 2nd PC to seasonality.

Still in Figure 3, for any year, points near to each other indicate similarity in the behavior of the κ-lagged vectors, e.g., the set of points $\{$A,Y,U$\}$ or $\{$O,N,D$\}$. It means that the κ-lagged vectors starting in April, May, and June, or the κ-lagged vectors beginning in October, November, and December resemble each other in terms of the object of interest. Also, the labeling strategy proved to be useful to capture the series behavior in the month itself, since April, May, and June correspond precisely to the periods in which the highest concentration of carbon dioxide occurs in the atmosphere. Besides, October, November, and December are the months with the lowest measured level of CO2.

In turn, the column markers (ℓ-lagged vectors) are represented as black arrows up to the sixth ℓ-lagged vector (tagged as L1, ..., L6 in Figure 3), ordered from top to bottom. From the seventh ℓ-lagged vector onwards, the pattern repeats itself, and so they were plotted in gray. It means that the first group of arrows, which is at the top, refer to the ℓ-lagged vectors beginning in January and July, just below those as starting in February and August, and so on. The angle between two consecutive arrows Li and Lj, such that $i = 1, \cdots, 5$ and $j = i + 1$, indicates a strong autocorrelation between the respective ℓ-lagged vectors since Li and Lj form very sharp angles. Comparing the angles between L1 and the others up to L6, they vary from a value close to 0o to a value close to 90o, which suggests a fading of the autocorrelations. Figure 4 shows the SSA-HJ-biplot formed by the 2nd and 3rd PCs, while Figure 5 exhibits the SSA-HJ-biplot constructed from the 4th and 5th PCs. Along with the first SSA-HJ-biplot, these are

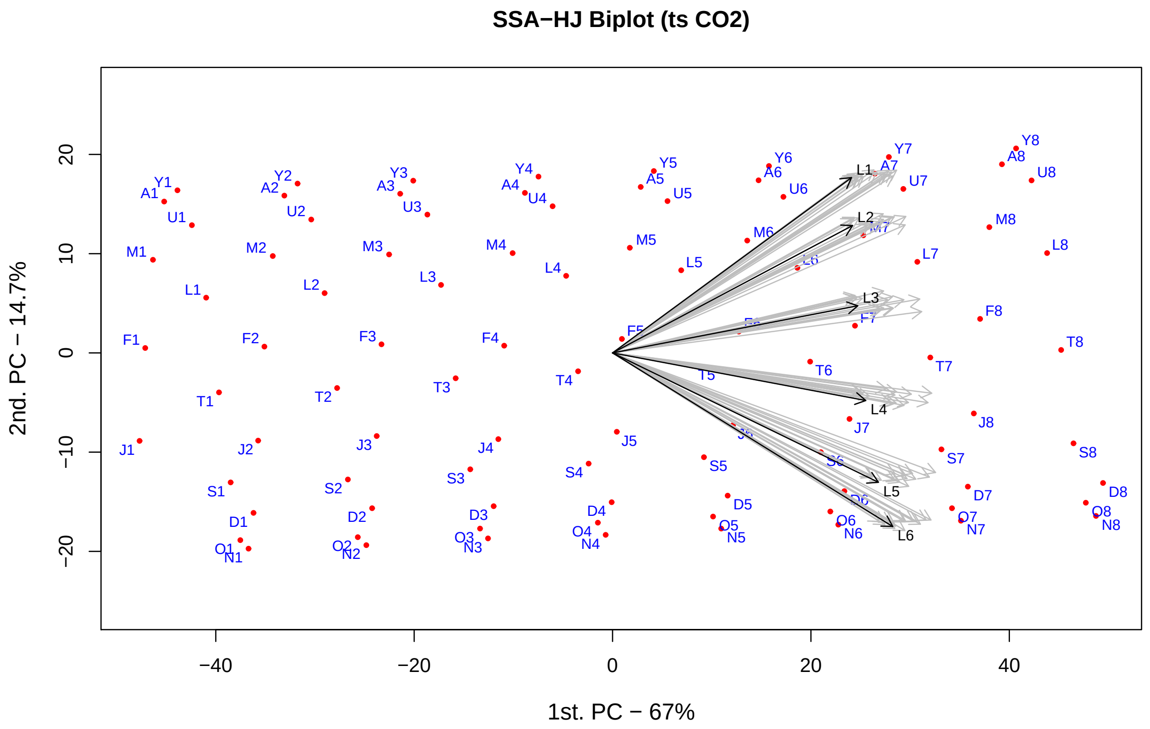

Figure 3. First SSA-HJ-biplot, constructed with the 1st and 2nd PC.

the only ones that produce interpretable results or evidence some pattern in the time series, being that these results are in agreement with the one verified in the scree plot of the singular values in Figure 2, where the pair of points related to $\sqrt{\mathbf{t}_2'\mathbf{t}_2}$ and $\sqrt{\mathbf{t}_3'\mathbf{t}_3}$ are around at the same level, the same concerning $\sqrt{\mathbf{t}_4'\mathbf{t}_4}$ and $\sqrt{\mathbf{t}_5'\mathbf{t}_5}$. In Figure 4, there are 12 distinct groups of row markers, each one of them referring to a κ-lagged vector starting in a specific month. Also, the column markers associated with each one of these groups show strong autocorrelation between the ℓ-lagged vectors. All of this indicates a seasonal pattern, with peaks and valleys separated by 12 months. In turn, the SSA-HJ-biplot in Figure 5 groups the lagged vectors two by two, e.g., January and July, February and August, and so on. Interpreting this together with the biplot in Figure 4, where these same groups occur but in the opposite directions, one can conclude that the valleys tend to be six months behind the peaks.

Therefore, the result of the grouping step for the decomposition of the TS CO2 should be $\mathbf{X}_1$ and $\mathbf{X}_2$, the first corresponding to the trend component, and the second describing the seasonal component, in which

$$\mathbf{X}_1 = \sqrt{\mathbf{t}_1'\mathbf{t}_1}\mathbf{t}_1^*\mathbf{p}_1',$$

and

$$\mathbf{X}_2 = \sum_{i=2}^{5} \sqrt{\mathbf{t}_i'\mathbf{t}_i}\mathbf{t}_i^*\mathbf{p}_i',$$

with the remaining being related to the noise component. The application of the diagonal averaging procedure over $\mathbf{X}_1$ and $\mathbf{X}_2$ produces the reconstructed series $\tilde{Y}^{(1)}$ and $\tilde{Y}^{(2)}$, whose graphical representations are shown in Figure 6.

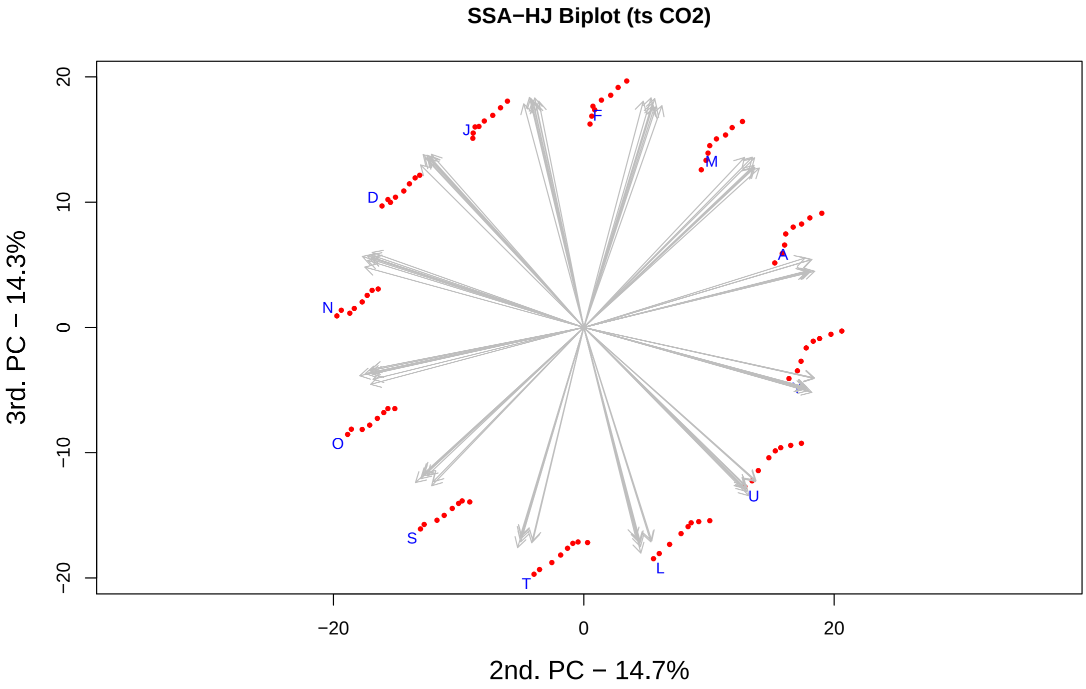

Figure 4. Second SSA-HJ-biplot of the 2nd and 3rd PCs describing an oscillatory component of TS CO2 of period 12

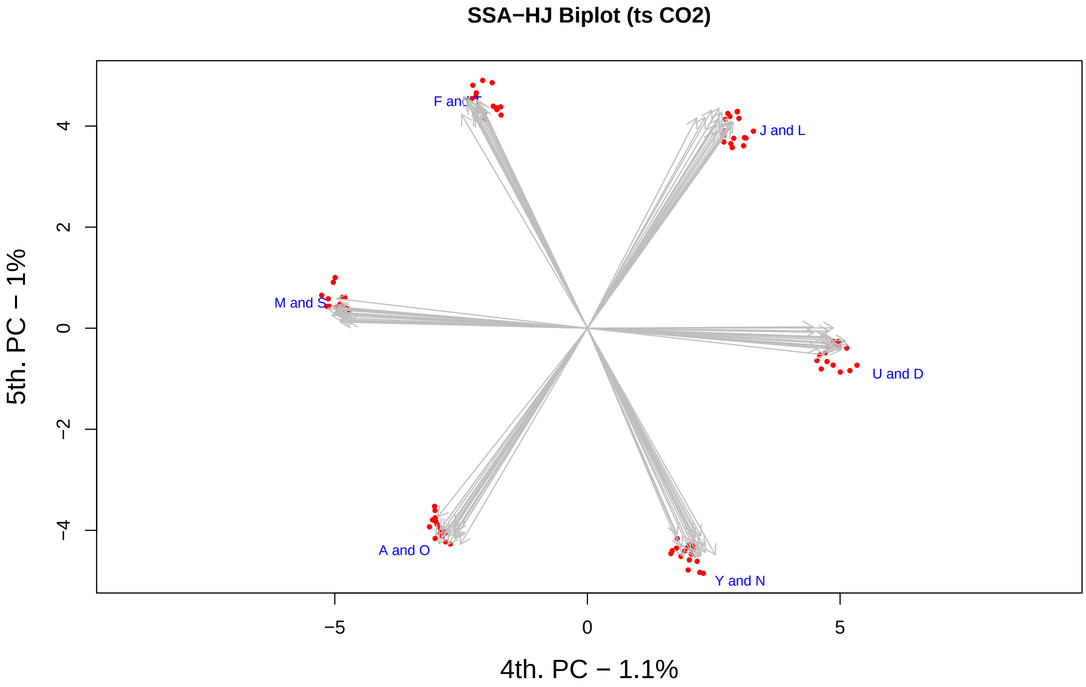

Figure 5. Third SSA-HJ-biplot of the 4th and 5th PCs also describing an oscillatory component of TS CO2.

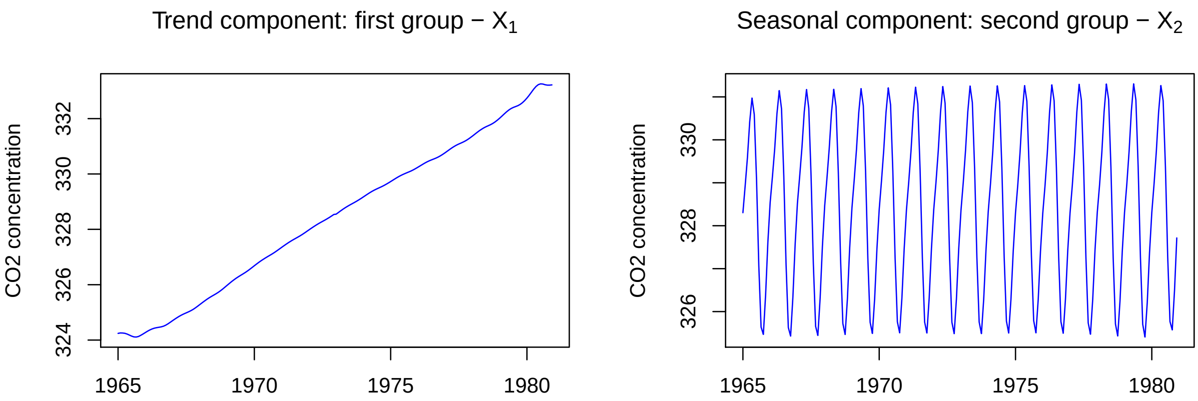

Figure 6. Separation of the trend and seasonal components of TS CO2 using the SSA-HJ-biplot technique.

5. Conclusions

This paper attempts to provide an integrative graphical tool to visualize and understand the underlying structure of the trajectory matrix, which is the result of the embedding step of the SSA. The SSA-HJ-biplot visualization method appears to be a promising exploratory technique, as it provides interpretability for the results of the SSA decomposition step, as illustrated by an example in this work. The SSA-HJ-biplots and auxiliary graphics provided a visual solution for the decomposition of the analyzed time series, properly separating the trend and the oscillatory component, using biplot axes up to the 5th PC. Also, it allowed the identification of all relevant eigentriple, composed by the singular values $\sqrt{\mathbf{t}_i'\mathbf{t}_i}$, by the left singular vectors $\mathbf{t}_i^*$, and by the right singular vectors $\mathbf{p}_i$, $i = 1, , 5$, to perform the grouping step. The study also revealed that the SSA-HJ-biplot points, representative of the row markers $(\mathbf{j}_i')$, can also depict the period itself in terms of dissimilarities, being possible to visually verify the months with the highest and lowest levels of CO2 concentration in the atmosphere throughout the years. The first SSA-HJ-biplot, built using the 1st and 2nd PCs, proved yet to be useful in dealing with autocorrelations between the column markers, which are drawn as arrows and represent the ℓ-lagged vectors. This study is promising in the sense that the SSA-HJ-biplot has great potential as an exploratory tool to analyze the structure of a univariate TS due to its visual appeal in such a complex issue. Nevertheless, TS may present complicated characteristics that make their analysis more challenging. For instance, prior detection of change-points in the TS is essential to highlight vital features, and consequently, to provide a better interpretation of SSA-HJ-biplots in complex TS data.

Acknowledgement

This work was 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), references UIDB/04106/2020 and UIDP/04106/2020.

REFERENCES

1. T. Alexandrov, *A method of trend extraction using Singular Spectrum Analysis*, REVSTAT, Statistical Journal, vol. 7, n. 1, pp. 1-22, 2009.
2. K. Gabriel, *The biplot graphic display of matrices with application to principal component analysis*, Biometrika, vol. 58, n. 3, pp. 453-467, 1971
3. M.P. Galindo, *An alternative of simultaneous representation: HJ-biplot*, Questii, vol. 10, 1, pp. 13-23, 1986.
4. P. Geladi and B.R. Kowalsky, *Partial Least Squares regression: a tutorial*, Analytica Chimica Acta, vol. 185, pp. 1-17, 1986.
5. N. Golyandina, V. Korobeynikov and A. Zhigljavsky, *Singular Spectrum Analysis with R*, 1st ed., Springer, Berlin, 2018.
6. N. Golyandina, V. Nekrutkin and A. Zhigljavsky, *Analysis of Time Series Structure: SSA and Related Techniques*, 1st ed. Chapman & Hall/CRC, Boca Raton, Florida, 2001.
7. N. Golyandina and A. Shlemov, *Variations of Singular Spectrum Analysis for separability improvement: non-orthogonal decompositions of time series*, Statistics and its Interface, vol. 8, n. 3, pp. 277-294, 2015.
8. M. Greenacre, *Biplots in Practice*, FBBVA, Bilbao, Biscay, 2010.
9. A.B. Nieto, M.P. Galindo, V. Leiva and P.V. Galindo, *A methodology for biplots based on bootstrapping with R*, Colombian Journal of Statistics, vol. 37, n. 2, pp. 367-397, 2014.
10. S. Moritz, T. Bartz-Beielstein, *imputeTS: Time Series Missing Value Imputation in R*, The R Journal, vol. 9, n. 1, pp. 207-218, 2017. https://doi.org/10.32614/RJ-2017-009.
11. NOAA Homepage, https://www.esrl.noaa.gov/gmd/ccgg/trends/, last accessed 24/05/2019.
12. R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2019 https://www.R-project.org/.
13. V.E. Vinzi and G. Russolillo, *Partial Least Squares algorithms and methods*, WIREs Comput Stat, vol. 5, pp. 1-19, 2013.
14. H. Wold, *Estimation of principal components and related models by iterative least squares*, in Multivariate Analysis, edited by P.R. Krishnaiah, Academic Press, New York, pp. 391–420, 1966.
15. S. Wold, C. Albano, W.J. Dunn, K. Esbensen, S. Hellberg, E. Johansson and M. Sjostrom, *Pattern recognition: finding and using regularities in multivariate data*, in Food Research and Data Analysis, edited by H. Martens and H. Russwurm, Applied Science Publishers, London, pp. 147-189, 1983.
16. K. Wright, *nipals: Principal Components Analysis using NIPALS or Weighted EMPCA, with Gram-Schmidt Orthogonalization*. R package version 0.7, 2020. https://CRAN.R-project.org/package=nipals

Chapter 5

Article II

SSA-HJ-biplot method: going further in the interpretation

Preprint:

Silva, A., Nieto-Librero, A.B., Freitas, A.. SSA-HJ-biplot method: going further in the interpretation. *Preprint submitted.*

SSA-HJ-biplot method: going further in the interpretation

Alberto Silva [*a,b], Ana B. Nieto-Librero[c], and Adelaide Freitas[a,b]

[a] *Department of Mathematics, University of Aveiro, Portugal*
[b] *Center for Research & Development in Mathematics and Applications (CIDMA), University of Aveiro, Portugal*
[c] *Department of Statistics, University of Salamanca, Spain*

October 25, 2022

An SSA-HJ-biplot is a tool designed to visualize the characteristics of a time series, having an exploratory nature. The graphical display is based on the HJ-biplot methodology and the NIPALS decomposition of Hankel ($\ell \times \kappa$) trajectory matrices. The approach aims to increase the visual interpretability of the Singular Spectrum Analysis method, making the grouping step easier. In this work, we detail the interpretation of this type of biplot when $\ell \approx \kappa$ and for columns-centered trajectory matrices, associating the eigenstructure to the components of the time series.

keywords: Singular Spectrum Analysis; HJ-biplot; Hankel matrix; NIPALS algorithm.

1 Introduction

The Singular Spectrum Analysis (SSA) is a non-parametric procedure based on principal components (PC) and used for signal extraction in time series (TS) analysis. The SSA can be used to decompose the original TS into a sum of a small number of interpretable components, like a slowly varying trend, different oscillatory components, and a noise (Golyandina et al., 2001). Briefly, in the basic version of SSA, a one-dimensional real-valued TS is transformed into a Hankel matrix, the so-called trajectory matrix ($\mathbf{X}$). The singular value decomposition is applied on , resulting in a summation of rank-one matrices. Some of these rank-one matrices are grouped appropriately in the grouping

[*] Corresponding author: albertos@ua.pt

step, and one can reconstruct the TS components by applying the diagonal averaging step over each obtained group.

SSA has been applied to several fields, e.g., *i*) for forecasting and exploratory purposes in economic data (Hassani and Zhigljavsky, 2009; de Carvalho et al., 2012); *ii*) for forecasting the number of cases, deaths, and recoveries of disease in public health data (Kalantari, 2021); and *iii*) understanding temperature and precipitation behavior in climate records (Benzi et al., 1997). Despite being a powerful technique, SSA presents some frailty in identifying the eigenvalues associated with the oscillatory components in the grouping step (Hassani and Mahmoudvand, 2018). Because of this problem, SSA requires external intervention to identify the extracted components' harmonic frequencies (Bógalo et al., 2017). Several studies have addressed the issue, emphasizing Ghil and Mo (1991), who associated the oscillatory component with two eigenvalues close to each other; the use of a periodogram to identify the association of pairs of eigenvectors to a harmonic by Vautard et al. (1992); the application of cluster analysis proposed by Alonso and Salgado (2008); the use of the asymptotic properties of a Toeplitz matrices' eigenvalues to relate SSA and the Fourier analysis suggested by Bozzo et al. (2010); and applying a versioned SSA based on circulant matrices Bógalo et al. (2017).

The SSA-HJ-biplot (da Silva and Freitas, 2020) was developed to improve the visual interpretability of the SSA and can also be an alternative method to identify the PCs associated with the harmonics. The tool takes advantage of the trajectory matrix decomposition results and consists of a version of the biplot proposed by Galindo (1986). On it, the HJ-biplot simultaneously represents both rows and columns of $\mathbf{X}$ through markers computed from the left and right singular vectors.

Usually, the row and column markers express variables and individuals in the multivariate analysis context. On the other hand, they refer to lagged vectors in the SSA-HJ biplot, i.e., they designate subseries of the original TS. Even so, the biplot interpretation remains valid regarding Euclidean distances, angles, and projections, with their due specificities. In addition, other meanings and insights emerge from the SSA-HJ-biplot, bringing different points of view regarding the relations between singular vectors and singular values and the components of the TS. This investigation aims to look into possible interpretations made from the visual inspection of an SSA-HJ-biplot, going beyond those elaborated according to the more general characteristics and properties of an ordinary HJ-biplot. That includes a suggestion to identify the eigenstructure related to the oscillatory components of a TS.

The paper is organized as follows. In **Section 2**, a short overview of the theoretical background of the SSA and the SSA-HJ-biplot technique is provided. In **Section 3**, the general bases for interpreting the SSA-HJ-biplot are first presented and then some specific insights from the graphical results are shown. In **Section 4**, after applying the tool on two real-world TS using the statistical software R (R Core Team, 2021), the suggested interpretation is performed on the graphs produced. Discussions and conclusions are presented in **Section 5**.

2 Methodology

2.1 The basic SSA

Consider a real-valued time series $Y = (y_1, \cdots, y_n)$ of length n. To construct a graphical tools to exploit features of Y, we consider the embedding procedure of the Basic SSA in representing Y in κ lagged vectors $\mathbf{x}_1, \cdots, \mathbf{x}_\kappa$, each one of size ℓ (ℓ-lagged vectors), i.e., $\mathbf{x}_j = [y_j, \cdots, y_{j+\ell-1}]'$, $1 \leq j \leq \kappa$, where ℓ $(1 < \ell < n)$ is an integer value representing the so-called window length and, consequently, $\kappa = n - \ell + 1$. This sequence of κ vectors forms the trajectory matrix $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_\kappa]$ of Y which is defined by a Hankel matrix with ℓ-lagged vectors by columns and given by

$$\mathbf{X} = \begin{bmatrix} y_1 & y_2 & y_3 & \cdots & y_\kappa \\ y_2 & y_3 & y_4 & \cdots & y_{\kappa+1} \\ y_3 & y_4 & y_5 & \cdots & y_{\kappa+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_\ell & y_{\ell+1} & y_{\ell+2} & \cdots & y_n \end{bmatrix} \tag{1}$$

The next step is to decompose the trajectory matrix into singular values, centering $\mathbf{X}$ or not. For example, centering $\mathbf{X}$ by the columns is suitable when the TS presents a trend since it extracts linear-like signals (Golyandina et al., 2001). The procedure starts from computing the matrix

$$\mathbf{C} = \frac{1}{\ell}\mathbf{1}_\ell \mathbf{1}_\ell' \mathbf{X}, \tag{2}$$

in which $\mathbf{1}_\ell = (1, \cdots, 1)'$, and in which $\mathbf{C}$ has ℓ identical rows. Then, the columns-centered trajectory matrix is

$$\tilde{\mathbf{X}} = \mathbf{X} - \mathbf{C}, \tag{3}$$

where each column of $\mathbf{X} - \mathbf{C}$ results from the comparison, in an all-to-one way, of ℓ terms (all) of the TS Y with a single element (one) of the moving average series of order ℓ, as illustrated in Figure 1.

Concretely, $\tilde{\mathbf{X}}$ is defined by κ vectors $\mathbf{v}_j$ so that

$$\mathbf{X} - \mathbf{C} = [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \cdots \quad \mathbf{v}_\kappa], \tag{4}$$

and, for $j = 1, \cdots, \kappa$,

$$\mathbf{v}_j = [(y_j - \bar{y}_j) \quad (y_{j+1} - \bar{y}_j) \quad \cdots \quad (y_{j+\ell-1} - \bar{y}_j)]', \tag{5}$$

where

$$\bar{y}_j = \frac{1}{\ell}\sum_{i=j}^{j+\ell-1} y_i. \tag{6}$$

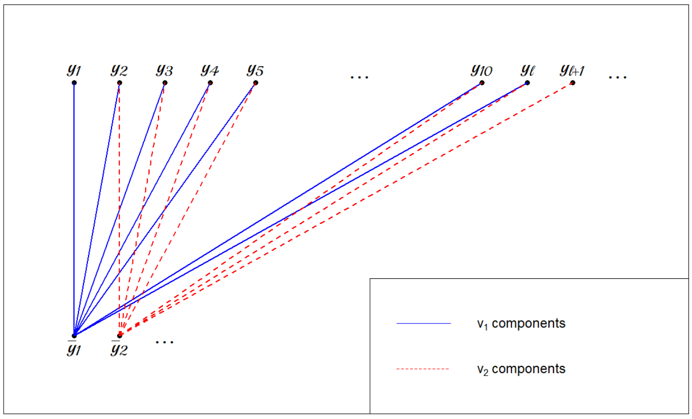

Figure 1: Columns centering all-to-one scheme (lines' length conceived just for the sake of illustration).

In terms of the rows, centering $\mathbf{X}$ by the columns corresponds to comparing, in a one-to-one way, each of the κ observations in Y with κ moving averages of order ℓ. Figure 2 schematically illustrates the terms of the comparison for a hypothetical series. Specifically,

$$\mathbf{X} - \mathbf{C} = [\mathbf{u}'_1 \quad \mathbf{u}'_2 \quad \cdots \quad \mathbf{u}'_\ell]', \tag{7}$$

such that

$$\mathbf{u}'_i = [(y_i - \bar{y}_1) \quad (y_{i+1} - \bar{y}_2) \quad \cdots \quad (y_{i+\kappa-1} - \bar{y}_\kappa)]. \tag{8}$$

The second step of the SSA consists of the decomposition of $\tilde{\mathbf{X}}$ using the NIPALS algorithm, such that

$$\tilde{\mathbf{X}} = \mathbf{T}^* \boldsymbol{\Sigma} \mathbf{P}', \tag{9}$$

in which i) $\mathbf{T}^*$ is the normalized scores matrix whose column vectors $\mathbf{t}^*_i$ are orthonormal; ii) $\mathbf{P}$ is the loadings matrix whose column vectors $\mathbf{p}_i$ are also orthonormal; and iii) $\boldsymbol{\Sigma}$ is a diagonal matrix containing the singular values $\sqrt{\mathbf{t}'_i \mathbf{t}_i}$ arranged in decreasing order, and $\mathbf{t}_i$ is the i^{th} score vector of $\tilde{\mathbf{X}}$. Another way to express (9) is by writing it as a summation of d 1-rank matrices, as below:

$$\tilde{\mathbf{X}} = \sum_{i=1}^{d} \tilde{\mathbf{X}}_i = \sqrt{\mathbf{t}'_1 \mathbf{t}_1} \mathbf{t}^*_1 \mathbf{p}'_1 + \cdots + \sqrt{\mathbf{t}'_d \mathbf{t}_d} \mathbf{t}^*_d \mathbf{p}'_d, \tag{10}$$

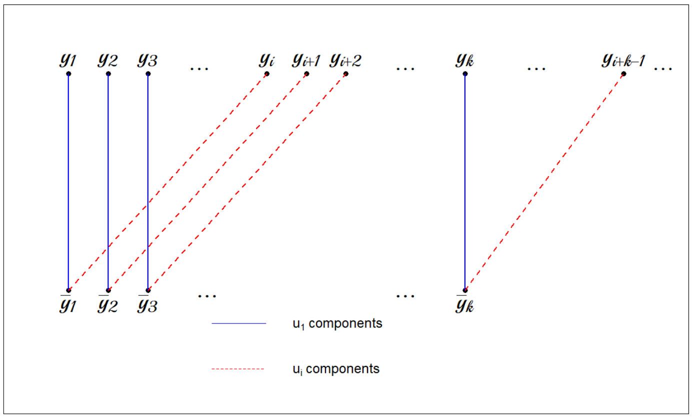

Figure 2: The row vectors $\mathbf{u}_1'$ and $\mathbf{u}_i'$ of the columns-centered trajectory matrix according to the one-to-one scheme (lines' length and parallelism conceived just for the sake of illustration).

where d is de rank of $\tilde{\mathbf{X}}$. The decomposition will reflect the columns $\mathbf{v}_1, \cdots, \mathbf{v}_\kappa$ and rows $\mathbf{u}_1', \cdots, \mathbf{u}_\ell'$ in terms of the orthogonal vectors $\mathbf{p}_1, \cdots, \mathbf{p}_d$ and $\mathbf{t}_1^*, \cdots, \mathbf{t}_d^*$, respectively.

The following step consists of grouping the elementary matrices $\tilde{\mathbf{X}}_i$ into $m < d$ disjoint groups, summing them within each group. Let $I_k = \{k_1, \cdots, k_p\}$, $k = 1, \cdots, m$, be each disjoint group of indices corresponding to the respective eigenvectors. Doing so, the matrix $\tilde{\mathbf{X}}_{I_k} = \tilde{\mathbf{X}}_{k_1} + \cdots + \tilde{\mathbf{X}}_{k_p}$ corresponds to I_k group. Consequently,

$$\tilde{\mathbf{X}} = \tilde{\mathbf{X}}_{I_1} + \cdots + \tilde{\mathbf{X}}_{I_m}. \tag{11}$$

The last step (diagonal averaging) transforms each matrix $\tilde{\mathbf{X}}_{I_k}$ into a Hankel matrix $\mathbf{X}_{I_k}$, converting the result into a TS, such that

$$\tilde{Y}_{I_k} = \mathcal{T}^{-1}\left(\mathbf{X}_{I_k}\right), \tag{12}$$

in which $\mathcal{T}$ represents the embedding operator. Additionally, since

$$\mathbf{X} = \sum_{i=1}^{d} \tilde{\mathbf{X}}_i + \mathbf{C}, \tag{13}$$

the matrix $\mathbf{C}$ yields an apart TS component afterward applying the diagonal averaging step.

2.2 The SSA-HJ-biplot construction

The primary purpose of the SSA-HJ-biplot is to be an auxiliary visualization tool in the decomposition of a TS. Therefore, it applies between the trajectory matrix decomposition step and the 1-rank matrix grouping step. For the two first PC, the SSA-HJ-biplot considers the rows of the matrix $\mathbf{J} = \mathbf{T}_2^*\mathbf{\Sigma}_2$ as row markers, and the rows of the matrix $\mathbf{H} = \mathbf{P}_2\mathbf{\Sigma}_2$ as column markers of $\tilde{\mathbf{X}}$, being that $\mathbf{T}_2^*$ and $\mathbf{P}_2$ denote the first two columns of $\mathbf{T}^*$ and $\mathbf{P}$, and $\mathbf{\Sigma}_2$ the diagonal matrix containing the two largest singular values in decreasing order. Consequently,

$$\tilde{\mathbf{X}}\mathbf{P}_2 = \mathbf{T}_2^*\mathbf{\Sigma}_2\mathbf{P}_2'\mathbf{P}_2 = \mathbf{T}_2^*\mathbf{\Sigma}_2. \tag{14}$$

It means that the ℓ rows of the matrix $\mathbf{J}$ correspond to the projections of the ℓ points representing the rows of $\tilde{\mathbf{X}}$ onto the subspace spanned by the loading vectors $\mathbf{p}_1$ and $\mathbf{p}_2$, i.e., the best-fit two-dimensional subspace for $\tilde{\mathbf{X}}$. Correspondingly, the κ rows of the matrix $\mathbf{H}$ coincide with the projections of the κ points expressing the columns of $\tilde{\mathbf{X}}$ onto the subspace spanned by the normalized score vectors $\mathbf{t}_1^*$ and $\mathbf{t}_2^*$, as below:

$$(\mathbf{T}_2^*)'\tilde{\mathbf{X}} = (\mathbf{T}_2^*)'\mathbf{T}_2^*\mathbf{\Sigma}_2\mathbf{P}_2' \Longleftrightarrow \tilde{\mathbf{X}}'\mathbf{T}_2^* = \mathbf{P}_2\mathbf{\Sigma}_2. \tag{15}$$

In addition, $\tilde{\mathbf{X}}\mathbf{P}_2$ and $\tilde{\mathbf{X}}'\mathbf{T}_2^*$ are related since calling $\mathbf{A} = \tilde{\mathbf{X}}\mathbf{P}_2$ and $\mathbf{B} = \tilde{\mathbf{X}}'\mathbf{T}_2^*$ one can obtain

$$\mathbf{A} = \tilde{\mathbf{X}}'\mathbf{B}\mathbf{\Sigma}_2^{-1} \tag{16}$$

and

$$\mathbf{B} = \tilde{\mathbf{X}}\mathbf{A}\mathbf{\Sigma}_2^{-1}. \tag{17}$$

In other words, the coordinates of the rows of $\tilde{\mathbf{X}}$ can be expressed as a weighted average of the coordinates of the columns and vice-versa. Consequently, it allows the representation of the rows and columns in the same Cartesian coordinates system with optimal quality of representation (Galindo, 1986; Nieto et al., 2014). The same reasoning goes for other pairs of PCs, with the proper adjustments. At last, in the SSA-HJ-biplot construction, the all-to-one relations (Figure 1) of the columns of $\tilde{\mathbf{X}}$ will be depicted as arrows, while the one-to-one relations (Figure 2) of the rows of $\tilde{\mathbf{X}}$ will be represented by points.

3 SSA-HJ-biplot interpretation

3.1 Noticeable properties

1. Consider the matrix $\mathbf{\Lambda} = diag(\lambda_1, \cdots, \lambda_d)$, where λ_i, $i = 1, \cdots, d$, are the eigenvalues of $\tilde{\mathbf{X}}\tilde{\mathbf{X}}'$, and d is its rank. In addition, keep in view that each $\lambda_i = \mathbf{t}_i'\mathbf{t}_i$, i.e., the respective squared singular value of $\tilde{\mathbf{X}}$. Then, given that $\mathbf{J} = \mathbf{T}^*\mathbf{\Sigma}$, the following relation is established:

$$\mathbf{J}\mathbf{J}' = \mathbf{T}^*\boldsymbol{\Sigma}(\mathbf{T}^*\boldsymbol{\Sigma})' = \mathbf{T}^*\boldsymbol{\Sigma}\boldsymbol{\Sigma}'(\mathbf{T}^*)' = \mathbf{T}^*\boldsymbol{\Lambda}(\mathbf{T}^*)' = \tilde{\mathbf{X}}\tilde{\mathbf{X}}', \tag{18}$$

implying that the scalar product $\mathbf{u}_i'\mathbf{u}_r$ is equal to $\mathbf{j}_i'\mathbf{j}_r$.

2. The Euclidean distances between two row vectors of $\tilde{\mathbf{X}}$ and the distance between the corresponding row markers in the full space are the same, so that

$$d^2(\mathbf{u}_i', \mathbf{u}_r') = (\mathbf{u}_i' - \mathbf{u}_r')'(\mathbf{u}_i' - \mathbf{u}_r') = (\mathbf{j}_i' - \mathbf{j}_r')'(\mathbf{j}_i' - \mathbf{j}_r') = d^2(\mathbf{j}_i', \mathbf{j}_r'), \tag{19}$$

and where, from (8), if the difference between the row vectors is approximately the zero vector, such that

$$(\mathbf{u}_i' - \mathbf{u}_r') = [(y_i - y_r) \quad (y_{i+1} - y_{r+1}) \quad \cdots \quad (y_{i+\kappa-1} - y_{r+\kappa-1})]' \approx \mathbf{0}, \tag{20}$$

then it is expected to observe the periodic behavior of the TS on the SSA-HJ-biplot, indicating a period $p = \frac{1}{k}(r - i)$, for $k = 1, \cdots, \lfloor \frac{n}{p} \rfloor$;

3. Regarding the columns of $\tilde{\mathbf{X}}$, it is known that $\mathbf{H} = \mathbf{P}\boldsymbol{\Sigma}$ in the HJ-biplot scheme, and then

$$\mathbf{H}\mathbf{H}' = \mathbf{P}\boldsymbol{\Sigma}(\mathbf{P}\boldsymbol{\Sigma})' = \mathbf{P}\boldsymbol{\Sigma}\boldsymbol{\Sigma}'\mathbf{P}' = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}' = \tilde{\mathbf{X}}'\tilde{\mathbf{X}}, \tag{21}$$

resulting in the following properties:

- Since, for $q = 1, \cdots, \kappa$, $||\mathbf{h}_q'||^2 = Var(\tilde{\mathbf{X}}_q)$, the biplot arrows length approximate the standard deviation of the corresponding $\mathbf{v}_q$. In addition, as stated in (da Silva and Freitas, 2020), if there is a $\tau \in \mathbb{N}, \tau < \kappa$, such that

$$||\mathbf{h}_\tau'|| \approx ||\mathbf{h}_{\tau+\iota}'|| \quad \Rightarrow \quad ||\mathbf{v}_\tau|| \approx ||\mathbf{v}_{\tau+\iota}||, \quad \forall \iota = 1, 2, \cdots, \kappa - \tau, \tag{22}$$

then it suggests that the ts $(y_1, \cdots, y_t, \cdots, y_n)$ might be generated by a variance stationary process for $t > \tau$. Moreover, if $\bar{y}_\tau \neq \bar{y}_{\tau+\iota}$, then one can suspect there is a trend component in the TS.

- Given that $\theta_{q,r}$ is the angle between the column markers $\mathbf{h}_q'$ and $\mathbf{h}_r'$, then $cos(\theta_{q,r})$ approximates $cor(\mathbf{v}_q, \mathbf{v}_r) = cor(\mathbf{x}_q, \mathbf{x}_r)$. Therefore, when $cos(\theta_{q,r}) \approx 1$, this indicates that the TS might be generated by a stationary process (da Silva and Freitas, 2020).

4. The proximity of two singular values ($\sqrt{\mathbf{t}_i'\mathbf{t}_i} \approx \sqrt{\mathbf{t}_h'\mathbf{t}_h}$) indicates that the associated PCs are informative about the oscillatory components of the TS (Golyandina et al., 2001).

3.2 Initial considerations

The only parameter for the SSA is the window length. When it comes to the SSA-HJ-biplot, choosing ℓ allows one to emphasize the representation of points or arrows. The arrows will more faithfully reflect the series for narrower windows, i.e., for $\kappa \gg \ell$. That is because the longer the κ-lagged vectors, the better the trajectory matrix columns capture the TS behavior. Otherwise, if $\ell \gg \kappa$, a long window length privileges the representation of the points in the sense of detaining the shape of the TS. An alternative choice is to opt for equilibrium, making $\ell \approx \kappa$. Note that this is a method that depends on the proportion of variability explained by the components, calculated as

$$PC_{\%}(i) = \frac{\mathbf{t}_i'\mathbf{t}_i}{\sum_{j=1}^{d} \mathbf{t}_j'\mathbf{t}_j}. \tag{23}$$

In the SSA-HJ-biplot, the points and arrows are labeled according to the period when the lagged vector starts, e.g., the month or year. Consider a TS of period p, so that each observation refers to the terms $\psi_1, \cdots, \psi_p$, periodically. Thus, a biplot point $\mathbf{j}_k'$, $(k = 1, \cdots, \ell)$, will be tagged according to the ψ_f, $(f = 1, \cdots, p)$, of the corresponding lagged vector's first observation, such that $f = i - \upsilon p$, and where $(\upsilon = \lfloor i/p \rfloor)$. With this, it is possible to visually capture the behavior of the phenomenon under study in terms of projections and distances between points (da Silva and Freitas, 2020).

From this moment on, let us assume an additive decomposition, in which we can write the TS as

$$Y = T + S + R, \tag{24}$$

where $Y = (y_1, \cdots, y_n)$ is the data, T is the trend-cycle, S is the seasonal component, and R is the remainder component. First, we will examine the SSA-HJ-biplots constructed from the PC associated with the dominant eigenvalue (trend) and from each of the components with the highest similar eigenvalues. These last two correspond to the dominant periodicity of the TS, with one referring to the cosine and the other to the sine since we can express S as a Fourier series. By doing so, one can synthesize the regular oscillatory with period p as a function $s(t)$ defined by a linear combination of sines and cosines such that

$$s(t) = \sum_{m=0}^{p} \left(\alpha_m \cos(\omega_m t) + \beta_m \sin(\omega_m t)\right), \tag{25}$$

where $\omega_m = 2\pi m/p$. Given a pure harmonic in which ω is the frequency, and ℓ and κ are multiples of the period $p = 1/w$, then the $\mathbf{t}_i^*$ and the principal components have the shape of sine and cosine sequences with the same p and the same phase. Hence, to identify the PCs generated by a harmonic it is enough to determine the pairs of left singular vectors presenting such shapes (Hassani, 2007).

As for the trend, we will approximate T by representing it through $\tilde{y}_t$, i.e., the moving average of order η (η-MA) such that, for $\eta = p$ and $t \geq 1$,

$$\tilde{y}_t = \frac{1}{\eta} \sum_{i=-\nu}^{\nu} y_{t+i}, \tag{26}$$

where $\eta = 2\nu + 1$. Somehow the PCA over the trajectory matrix is just another way to approximate these terms, with the 1^{st} PC representing the dominant trend (if any), pairs of PCs with similar singular values reflect the sine and cosine for different frequencies, and those PCs with negligible singular values are associated with noise.

3.3 Going into details

In the circumstances outlined above, the following is observed:

1. First, let's examine the SSA-HJ-biplot constructed from the 1^{st} PC (trend) and the PC corresponding to the cosine direction in the pair of the largest similar eigenvalues (2^{nd} or 3^{rd} PC). The biplot points $\mathbf{j}'_k = (j_{k,1}, j_{k,2})$ near to the 1^{st} PC axis explain the behavior of the p-MA. Also, when an observation for a determined ψ_f, $(f = 1, \cdots, p)$, is located above the p-MA, it is expected that the respective $\mathbf{j}'_k$ will appear above the 1^{st} PC axis. On the other hand, it will appear below, i.e., for $k = 1, \dots \ell$,
$$y_k - \tilde{y_k} > 0 \Rightarrow (\pm j_{k,1}, +j_{k,2}), \tag{27}$$
and
$$y_k - \tilde{y_k} < 0 \Rightarrow (\pm j_{k,1}, -j_{k,2}). \tag{28}$$
In this case, the biplot points tend to replicate the behavior of Y concerning the increasing or decreasing order of the observations. Hence, given that y_k corresponds to some ψ_f, and for $k = 1, \cdots, (\ell - 1)$,
$$y_k > y_{k+1} \Rightarrow j_{k,2} > j_{(k+1),2}, \tag{29}$$
where $f = i - \upsilon p$.

2. Next, considering the pair with the highest similar eigenvalues again, we will deal with the SSA-HJ-biplot formed by the 1^{st} PC and the one related to the sine direction. Due to the orthogonality of the sine and cosine functions, the projections of the trajectory matrix rows change positions in the plane formed by the factorial axes. Thus, as long as the data are well represented, the biplot points labeled with the ψ_f corresponding to tops and valleys in the TS place now near to the 1^{st} PC, and the same goes for the opposite case, i.e.,
$$\max(|y_k - \tilde{y_k}|) \Rightarrow \min(|j_{k,2}|), \tag{30}$$
and
$$\min(|y_k - \tilde{y_k}|) \Rightarrow \max(|j_{k,2}|). \tag{31}$$

3. In addition to the angle between the arrows, another way of interpreting the representation of the ℓ-lagged vectors is through the cosine of the angle formed between the arrows and a factor axis. The more acute the angle, the more the ℓ-lagged vector is related to the axis. This relationship is called relative contribution (RC) and represents the proportion of the variability of each ℓ-lagged vector explained

by the PC of interest (Nieto et al., 2014). Hence, considering θ is the angle between $\mathbf{h}'_q$ and the axis corresponding to the i^{th} PC, the greater the $|cos(\theta)|$, the more the variability of this associated column vector will have been affected by this component.

4. The projection of a biplot point onto an arrow corresponds to the level of agreement between the κ-lagged vector that determines the point and the ℓ-lagged vector that induces the arrow.

4 Examples

To demonstrate the capabilities of the SSA-HJ-biplot in terms of interpretability, we applied the SSA-HJ-biplot to two datasets. The first is a TS containing records of the concentration of carbon dioxide in the Earth's atmosphere, measured monthly at the Mauna Loa observation station in Hawaii (TS CO_2). The other is a dataset containing the records of the average monthly wildfire statistics provided by the U. S. National Interagency Fire Center (NIFC), available from January 2013 to December 2020 and called here as TS Wildfire.

4.1 TS $\mathbf{CO_2}$

The TS CO_2 consists of 192 observations from January 1965 to December 1980. Figure 3 shows the series with the respective 12-MA overlap.

The following procedures were applied for the construction of each SSA-HJ-biplot: *i*) First, we defined the length of the window $\ell = 96$, i.e., as $n/2$, and constructed the trajectory matrix ($\mathbf{X}$) using the embedding approach; *ii*) Next, we performed the NIPALS algorithm to the centered $\tilde{\mathbf{X}}$ to obtain the singular vectors ($\mathbf{t}_i^*$ and $\mathbf{p}'_i$) and the singular values ($\sqrt{\mathbf{t}'_i\mathbf{t}_i}$), $i = 1, \cdots, d$, where d is the rank of $\tilde{\mathbf{X}}$; and *iii*) Lastly, depending on the PCs that will function as factorial axes, we set the $(\ell \times 2)$ matrix $\mathbf{J}$ and the $(\kappa \times 2)$ matrix $\mathbf{H}$ choosing the adequate pairs of singular vectors along with the corresponding singular values. Table 1 shows the proportion of variability explained by the first ten PCs and computed according to (23). From Table 1, one can verify that the eigenvalue $\mathbf{t}'_1\mathbf{t}_1$ is dominant, and then the 1^{st} PC is associated with the trend. Further, the proximity of $\mathbf{t}'_2\mathbf{t}_2$ and $\mathbf{t}'_3\mathbf{t}_3$ indicates that 2^{nd} and 3^{rd} PCs are related to the periodicity of the TS. The same interpretation goes for the 4^{th} and 5^{th} PCs, whose corresponding eigenvalues are close.

The set containing the labels is $\psi = \{J, F, \cdots, D\}$, whose each element represents, respectively, the months of January, February, and so on, until December. Hence, Figure 4 shows the SSA-HJ-biplot built from the 1^{st} and 2^{nd} PCs, which explain more than 80% of the data variability. The $\mathbf{j}'_i$ points' projections onto the horizontal axis evolve in the same growth direction of the 1^{st} PC, meaning a crescent trend. Differently, the projections of points of the same label onto the vertical axis fall always close to the same coordinate, indicating an association of the 2^{nd} PC with the seasonality. Furthermore, considering the shape of the the biplot points' contour referring to the origin of the factor

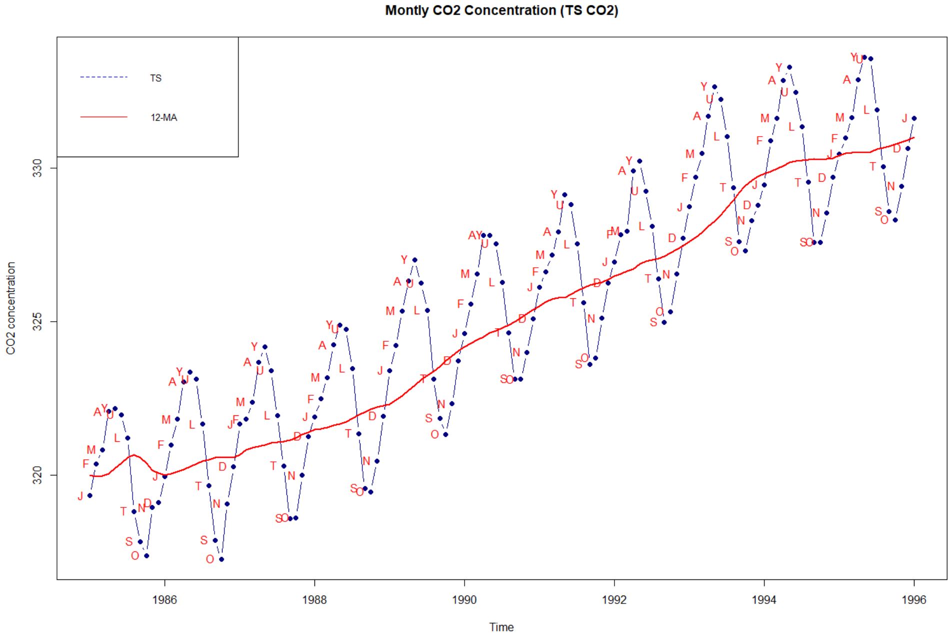

Figure 3: Time series CO_2 representation and the respective moving average of order 12.

Table 1: Proportion of variability explained by the ten first principal components

PC (i)	Variance (%)	PC (i)	Variance (%)
1	66.830	6	0.376
2	14.707	7	0.316
3	14.444	8	0.291
4	1.096	9	0.106
5	1.070	10	0.080

axes, we can infer that the 2^{nd} PC is associated with the cosine direction. It can be seen that February and August describe the behavior of the 12-MA as the points with F and T labels are positioned close to the 1^{st} PC along the entire axis. Within each period, a biplot point $\mathbf{j}'_i$ marked as Y always has the highest value compared to the 2^{nd} PC, which means we expect peaks to occur in May in the TS. Likewise, the $\mathbf{j}'_i$ points tagged as N always have the lowest values concerning the 2^{nd} PC, so one can presume that the valleys will appear in the series in November.

We can assume that the CO_2 concentration decreases throughout the year from June to November, starting to increase again from December to May. Besides, from March to July, the CO_2 concentration around the Hawaii station places above the 12-MA since the projections of the corresponding $\mathbf{j}'_i$ points onto the 2nd PC axis are always positive $(+j_{i,2})$. Correspondingly, from September to January, the accumulation of carbon diox-

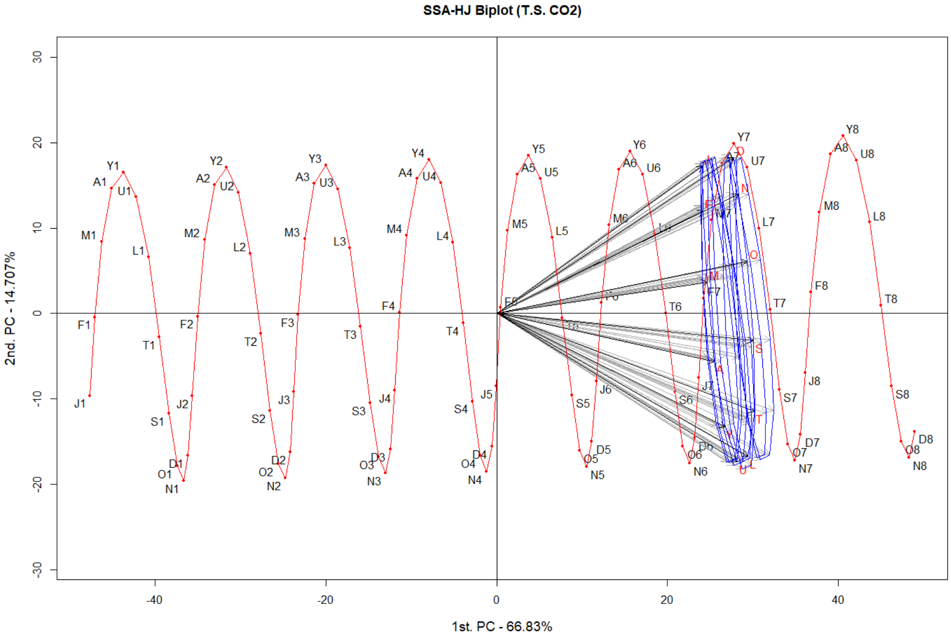

Figure 4: TS CO2 SSA-HJ-biplot regarding the 1^{st} and 2^{nd} PCs.

ide is below the moving average. The proximity of the points labeled A, Y and U, as well as O, N and D indicates that the CO2 concentration levels in the corresponding months are also close.

In the case of the SSA-HJ-biplot constructed using the 1^{st} and 3^{rd} PCs, the behavior of the biplot points on the plane formed by the factor axes seems to confirm that the 1^{st} component is related to the sine direction (Figure 5). Due to the sine and cosine orthogonality, the projections represented by $\mathbf{j}_i'$ points and $\mathbf{h}_k$ arrows appear in inverted positions in this plane. The months that appeared in the first SSA-HJ-biplot describing the behavior of the 12-MA now form the tops and valleys of the contour points. To better understand what happens here, it is necessary to remember that these elements (dots and arrows) represent projections of the original observations in a reduced dimension.

As for the arrows, they represent the ℓ-lagged vectors (columns of $\tilde{\mathbf{X}}$). The contour of the row vectors $\mathbf{h}_q'$ consist of a representation similar to that of the $\mathbf{j}_i'$ points but compressed. It occurs because the trend is still present in the columns of $\tilde{\mathbf{X}}$, unlike the rows. In addition, compared to the points, there is also an inversion of both the position of the months (upside down) and the direction of growth (counterclockwise). It happens because the $\mathbf{j}_i'$ points are eigenvectors of the matrix $\tilde{\mathbf{X}}\tilde{\mathbf{X}}'$, while $\mathbf{h}_q'$ are eigenvectors of $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$.

It is necessary to analyze the SSA-HJ-biplots jointly to interpret the autocorrelation through the $\mathbf{h}_q'$ vectors. In this example, for the same ψ_f label, all the $\mathbf{h}_q'$ form an angle close to 0° with each other, regardless of the PCs used in the construction of the

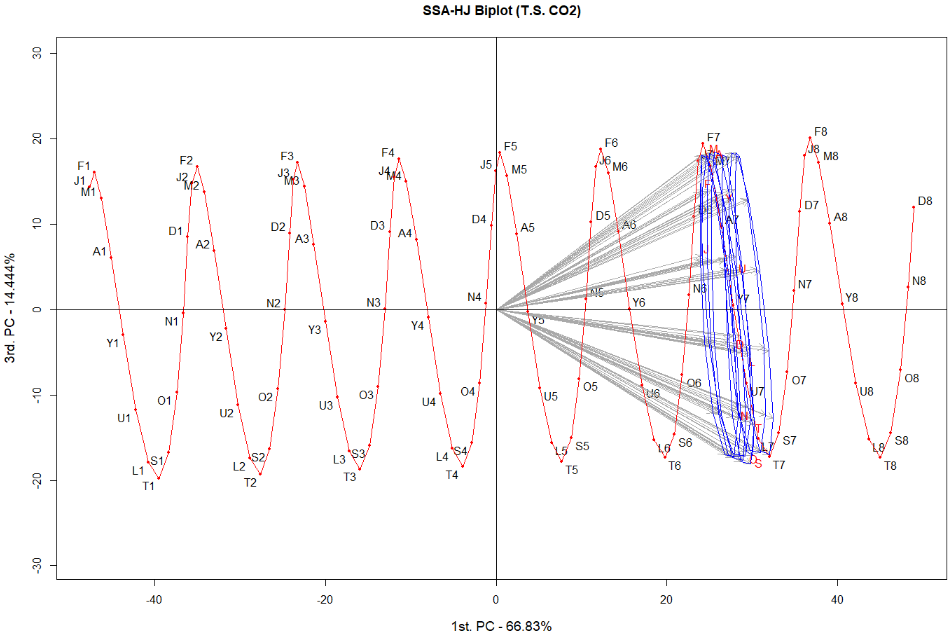

Figure 5: TS CO2 SSA-HJ-biplot regarding the 1st and 3rd PCs.

graph, meaning that the corresponding ℓ-lagged vectors are strongly correlated. On the other hand, for different labels, the orthogonality of sine and cosine can lead to wrong conclusions if we look only at a specific SSA-HJ-biplot. For example, the angles among the arrows labeled with ψ_2 = February and ψ_{11} = November are close to 0° in Figure 4, in which the 2^{nd} PC is related with the cosine direction.

However, something quite different occurs in the graph constructed using the PC associated with the sine direction (Figure 5), where those angles are much closer to 90° than 0°. The SSA-HJ-biplot of the 2^{nd} and 3^{rd} settles the issue by showing a right angle between the two sets of arrows. All of it indicates the SSA-HJ-biplot that uses the 1^{st} and 2^{nd} PCs and is associated with the cosine direction better represents the biplot points $\mathbf{j}'_i$ than the arrows $\mathbf{h}'_q$.

On the other hand, the SSA-HJ-biplot with the 1^{st} and 3^{rd} PCs as factorial axes provides more accurate information regarding the arrows $\mathbf{h}'_q$. Based on this, in relation to the 3rd PC, it can be seen in Figure 5 that the smallest angles are those formed by the factor axis and the arrows labeled with M, A, S or O. Thus, we conclude that the variability of the corresponding ℓ-lagged vectors is strongly affected by 3^{rd} PC. Likewise, regarding the 2^{nd} PC, the ℓ-lagged vectors that start in January, June, July, and December are those whose proportion of variability is more explained by that component, i.e., the greater RC. Proceeding to a joint interpretation, comparing the SSA-HJ-biplots formed by the 2nd and 3rd PCs (Figure 6) and the 4^{th} and 5^{th} PCs (Figure 7) with those of Figures 4 and 5, we will see the same information as before, but

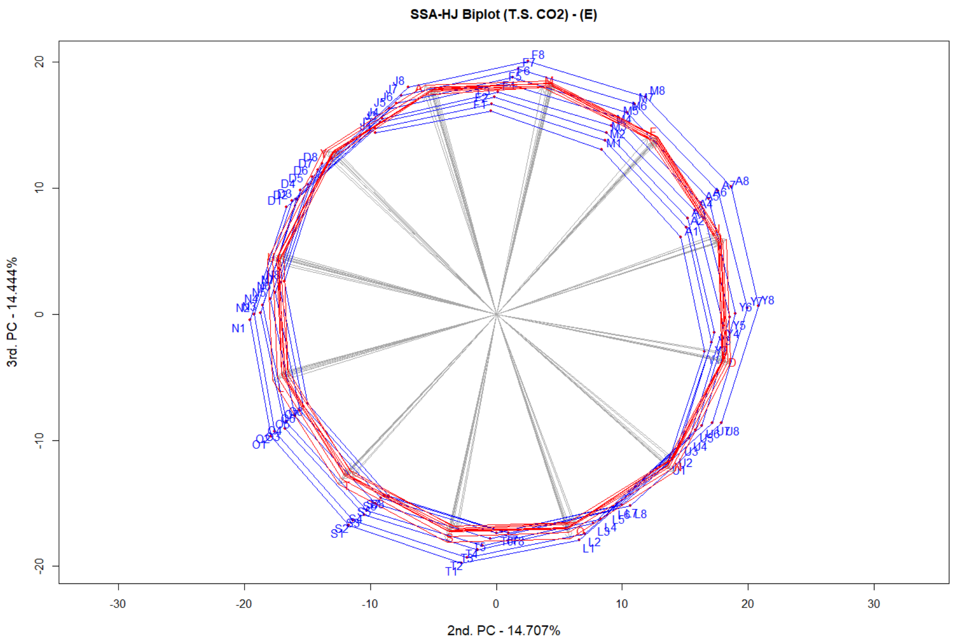

Figure 6: TS CO2 SSA-HJ-biplot regarding the 2^{nd} and 3^{rd} PCs.

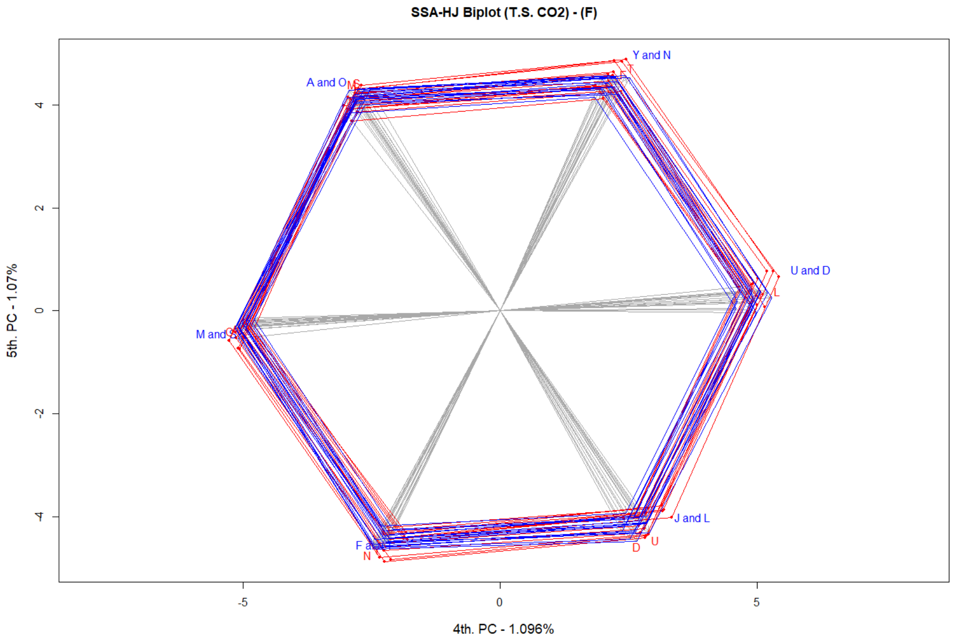

Figure 7: TS CO2 SSA-HJ-biplot regarding the 4^{th} and 5^{th} PCs.

now without any interference from the trend direction. For example, as for Figure 6 and regarding the 2^{nd} PC, the $\mathbf{j}'_i$ points tagged from M to L are located on the positive side of the 2^{nd} component, while those labeled from S to J are on the opposing part.

We can find the points related to the 12-MA around zero. The position of the arrows also repeats the same positioning pattern concerning the labels on the first SSA-HJ-biplot. In Figure 6, it is even more evident which months are most related to each PC, i.e., the largest RC of each component as a function of the cosine of the angles formed by the arrows and the axes. In addition, the periodicity appears in the formation of 12 groups of both points and arrows.

When it comes to the Figure 7, it reinforces the previous conclusions but now shows the elements $\mathbf{j}'_i$ and $\mathbf{h}'_q$ labeled with the same ψ_f in pairs, instead of mirrored regarding some axis. For example, $\psi_6 = J$ and $\psi_{12} = D$ appear in a symmetrical position in the SSA-HJ-biplots of the 2^{nd} and 3^{rd} PCs, but appear together in the SSA-HJ-biplot of the 4th and 5th PCs. It suggests that the valleys tend to be six months behind the peaks.

4.2 TS Wildfire

Regarding the second example, the sample size of the TS Wildfire series is $n = 96$, spanning January 2013 to December 2020. As usual, the window length is $\ell = n/2 = 48$ and $\kappa = n - \ell + 1 = 49$. Figure 8 shows the shape of the TS Wildfire, and the overlaid 12-MA suggests there is no trend in the series. Table 2 confirms the absence of a slowly varying trend since the two first eigenvalues $\mathbf{t}'_1\mathbf{t}_1$ and $\mathbf{t}'_2\mathbf{t}_2$ are next to each other, indicating the corresponding 1^{st} and 2^{nd} PCs are associated with the TS oscillatory component. And so do the 3^{rd} and 4^{th} PCs, given that the eigenvalues $\mathbf{t}'_3\mathbf{t}_3$ and $\mathbf{t}'_4\mathbf{t}_4$ are similar. From the eigenvalue $\mathbf{t}'_5\mathbf{t}_5$ on the absence of a geometric pattern on the SSA-HJ-biplots suggests the corresponding PCs are related with the noise.

Table 2: Proportion of explained variance by the ten first principal components (TS Wildfire).

PC$_\mathrm{i}$	Variance (%)	PC$_\mathrm{i}$	Variance (%)
1	14	6	4
2	13	7	3
3	11	8	3
4	10	9	3
5	4	10	3

The TS Wildfire is a challenging case to analyze, as the behavior of the observations appears more erratic than in the previous example. Even so, the SSA-HJ-biplots of the 1^{st} and 2^{nd} PCs (Figure 9) and the 3^{rd} and 4^{th} PCs (Figure 10) suggest a periodicity of 4 and 12. In the first case, the reason stems from i) four clusters inside which the points are close to each other; ii) four bundles of arrows where the angle between them is close to zero within each group.

The SSA-HJ-biplot (Figure 9) indicates that the κ-lagged vectors associated with the

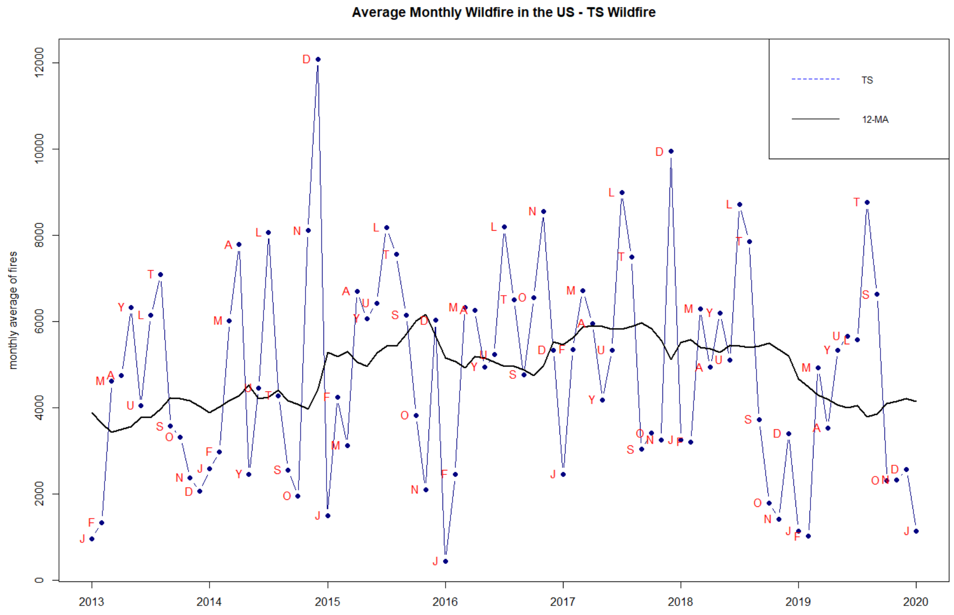

Figure 8: Time series Wildfire and the corresponding 12-MA.

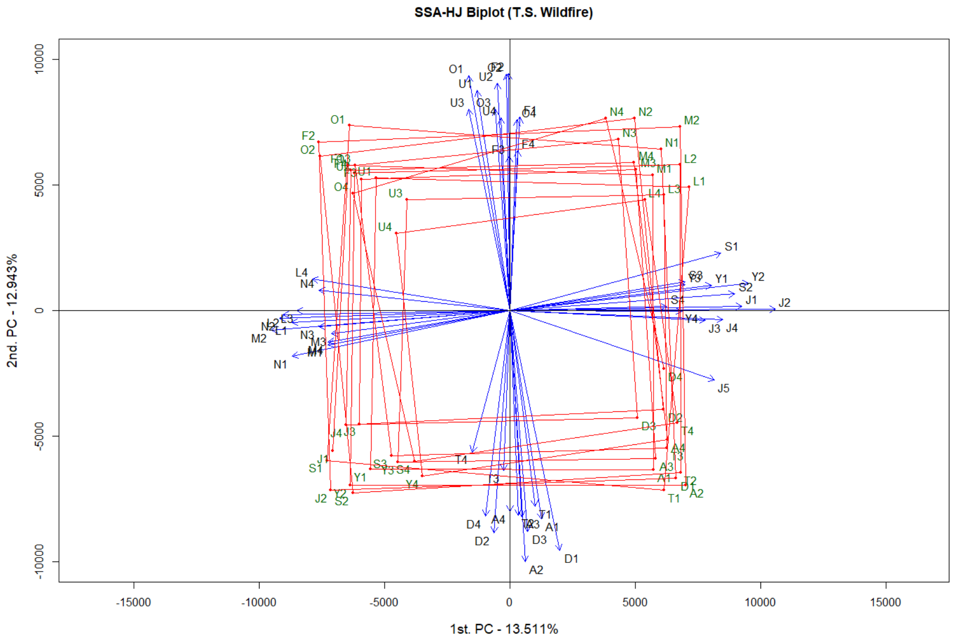

Figure 9: TS Wildfire SSA-HJ-biplot of the 1st and 2nd PCs.

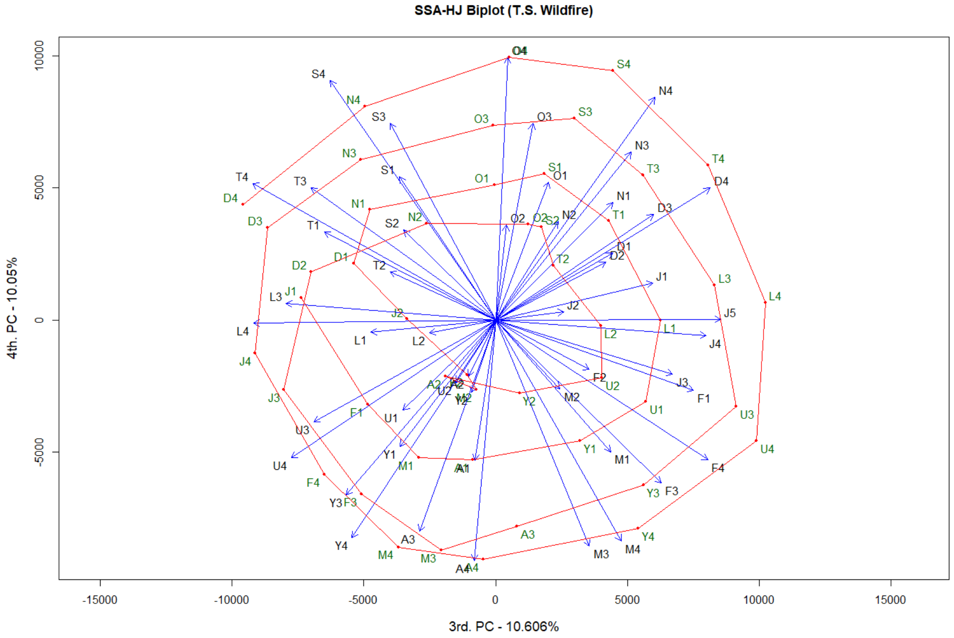

Figure 10: TS Wildfire SSA-HJ-biplot of the 3^{rd} and 4^{th} PCs.

$\mathbf{j}_i'$ points tagged as J, Y, and S are those where some movement starts, finishing in A, T, and D. The κ-lagged vectors starting in other months represent intermediate situations. The angle between the $\mathbf{h}_q'$ arrows and the axes indicates the ℓ-lagged vectors starting at J, Y, and S are more related to the 1^{st} PC, while those starting at A, T, and D are more associated with the 2^{nd} PC. About the underlying phenomenon, one can expect the occurrence of peaks of wildfires in April (A), August (T), and December (D) relative to the previous three months.

Finally, the SSA-HJ-biplot of Figure 10 shows some visual deterioration, but it is still possible to verify that the amplitude of the series does not show a uniform pattern of variation over time since the size of the $\mathbf{h}_q'$ arrows goes back and forth. Therefore, one can expect that the associated TS is non-stationary. Biplot points labeled with a specific month (e.g., T) are close to each other and form twelve groups around the origin of the axes but present an irregular contour. Regarding the RC, the ℓ-lagged vectors associated with J and L are more related to the 3^{rd} PC than the 4^{th} PC. In contrast, the 4^{th} PC explains a greater proportion of the variability of ℓ-lagged vectors starting at A and O.

5 Conclusion

This article seeks to develop a general and comprehensive interpretation of the SSA-HJ-biplot, providing a visual understanding of the linkage between the trajectory matrix eigenstructure and the components of the corresponding TS. It is natural to look first for PCs that explain a considerable proportion of data variability in the construction of each SSA-HJ-biplot. But notice that some harmonics may be associated with pairs of eigenvectors whose PC explains less than 1% of the variability. The two examples in Section 4 show i) the importance of performing joint analysis of SSA-HJ-biplots; ii) the more components the TS has, the more informative the SSA-HJ-biplots will be; iii) PCs associated with noise can explain a considerable amount of data variability; iv) sometimes, due to the existence of a dominant eigenvalue (associated with the trend) and two close eigenvalues (associated with the oscillatory component), it is crucial to building two SSA-HJ-biplots for the 1^{st} PC, one of them using the PC associated with the sine and the other with the cosine. It is because biplot points may be better represented in one biplot while arrows in another.

The SSA-HJ-biplots' properties proved convenient in identifying the periodicity of the studied TS. Regarding the phenomenon recorded in the first example (CO_2), the suggested interpretation highlighted the months with the highest and lowest concentration of carbon dioxide. In addition, it showed the months whose records most resemble the moving average of the series. As for the second example (wildfires), the SSA-HJ-biplots helped to raise suspicions about the months that tend to have the highest incidence of fires within every four months. The proposed interpretation strengthens the SSA-HJ-biplot method, expanding its capacity as a visual exploratory technique. In more complex data, we suggest the segmentation of the TS and, after, applying the approach in more or less homogeneous intervals to maintain its interpretability.

Acknowledgement

This research is part of first author's Ph.D. project, which is carried out at University of Aveiro. The authors were supported by Fundação para a Ciência e a Tecnologia (FCT), within project UID/MAT/04106/2019 (CIDMA).

References

Alonso, F. and Salgado, D. (2008). Analysis of the structure of vibration signals for tool wear detection. *Mechanical systems and signal processing*, 22(3):735–748.

Benzi, R., Deidda, R., and Marrocu, M. (1997). Characterization of temperature and precipitation fields over sardinia with principal component analysis and singular spectrum analysis. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 17(11):1231–1262.

Bógalo, J., Poncela, P., and Senra, E. (2017). Automatic signal extraction for stationary and non-stationary time series by circulant ssa.

Bozzo, E., Carniel, R., and Fasino, D. (2010). Relationship between singular spectrum analysis and fourier analysis: Theory and application to the monitoring of volcanic activity. *Computers & Mathematics with Applications*, 60(3):812–820.

da Silva, A. O. and Freitas, A. (2020). Time series components separation based on singular spectral analysis visualization: an hj-biplot method application. *Statistics, Optimization & Information Computing*, 8(2):346–358.

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

Galindo, M. P. (1986). An alternative for simultaneous representation: Hj-biplot. *Questíio*, 10:12–23.

Ghil, M. and Mo, K. (1991). Intraseasonal oscillations in the global atmosphere. part i: Northern hemisphere and tropics. *Journal of Atmospheric Sciences*, 48(5):752–779.

Golyandina, N., Nekrutkin, V., and Zhigljavsky, A. A. (2001). *Analysis of time series structure: SSA and related techniques*. CRC press.

Hassani, H. (2007). Singular spectrum analysis: methodology and comparison.

Hassani, H. and Mahmoudvand, R. (2018). *Singular spectrum analysis: Using R*. Springer.

Hassani, H. and Zhigljavsky, A. (2009). Singular spectrum analysis: methodology and application to economics data. *Journal of Systems Science and Complexity*, 22(3):372–394.

Kalantari, M. (2021). Forecasting covid-19 pandemic using optimal singular spectrum analysis. *Chaos, Solitons & Fractals*, 142:110547.

Nieto, A. B., Galindo, M. P., Leiva, V., and Vicente-Galindo, P. (2014). A methodology for biplots based on bootstrapping with r. *Revista colombiana de estadística*, 37(2):367–397.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Vautard, R., Yiou, P., and Ghil, M. (1992). Singular-spectrum analysis: A toolkit for short, noisy chaotic signals. *Physica D: Nonlinear Phenomena*, 58(1-4):95–126.

Chapter 6

Article III

An enhanced version of the SSA-HJ-biplot for Time Series with complex structure

Preprint:

Silva, A., Freitas, A.. An enhanced version of the SSA-HJ-biplot for Time Series with complex structure. *Preprint submitted.*

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

Alberto Silva[1,2*] and Adelaide Freitas[1,2†]

[1*]Department of Mathematics, University of Aveiro, Campus de Santiago, Aveiro, 3810-193, Aveiro, Portugal.
[2]Center for Research & Development in Mathematics and Applications (CIDMA), University of Aveiro, Campus de Santiago, Aveiro, 3810-193, Aveiro, Portugal.

*Corresponding author(s). E-mail(s): albertos@ua.pt;
Contributing authors: adelaide@ua.pt;
†This author contributed equally to this work.

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

MSC 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 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 $Y = (y_1, \cdots, y_n)$ in which the stochastic structure related to Y is said to be strictly stationary. In this case, given $t_1, \cdots, t_k \in \{1, \cdots, n\}$, the joint distribution functions of the random vectors $(y_{t_1}, \cdots, y_{t_k})$ and $(y_{t_1+\tau}, \cdots, y_{t_k+\tau})$ 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 $a_1, \cdots, a_r$ 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 Section 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 Section 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 Section 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 $Y = (y_1, \cdots, y_n)$ 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 $\mathbf{x}_1, \cdots, \mathbf{x}_\kappa$, each one of size ℓ, in which

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

These κ lagged vectors form a Hankel matrix $\mathbf{X}$ called trajectory matrix, i.e., $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_\kappa]$. 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

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

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 $(\sqrt{\mathbf{t}_i'\mathbf{t}_i}, \mathbf{t}_i^*, \mathbf{p}_i), i = 1, \cdots, d$, 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 $\sqrt{\mathbf{t}_i'\mathbf{t}_i}, i = 1, \cdots, d$, 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 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. The columns of $\mathbf{H}'$ are depicted as vectors connected to the origin (arrows). When $q > 2$, the best 2-rank approximation of $\mathbf{X}$ is considered in the sense of least square. Assuming $\mathbf{G} = \mathbf{T}^*$ and $\mathbf{H} = \mathbf{P}\mathbf{\Sigma}$, 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 $\mathbf{G} = \mathbf{T}^*\mathbf{\Sigma}$ 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. The representation of the rows is better than the columns. From this point, consider the matrix $\mathbf{J} = \mathbf{T}^*\mathbf{\Sigma}$, and the matrix $\mathbf{H} = \mathbf{P}\mathbf{\Sigma}$. 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.

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 $Y = (y_1, \cdots, y_n)$ be a univariate and real-valued TS, and $y_{h+1}, \cdots, y_{h+m}$ be a subseries so that $m < n$ and $h = 0, \cdots, n-m$ (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

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

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

noticing that

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

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

The parameter m is crucial to adequately capture the change of direction of the principal component in the NIPALS decomposition 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 $\mathcal{D}_\phi, \phi = 1, 2, \cdots, d$, 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 $Y(y_{h+1}, \cdots, y_{h+m})$, 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

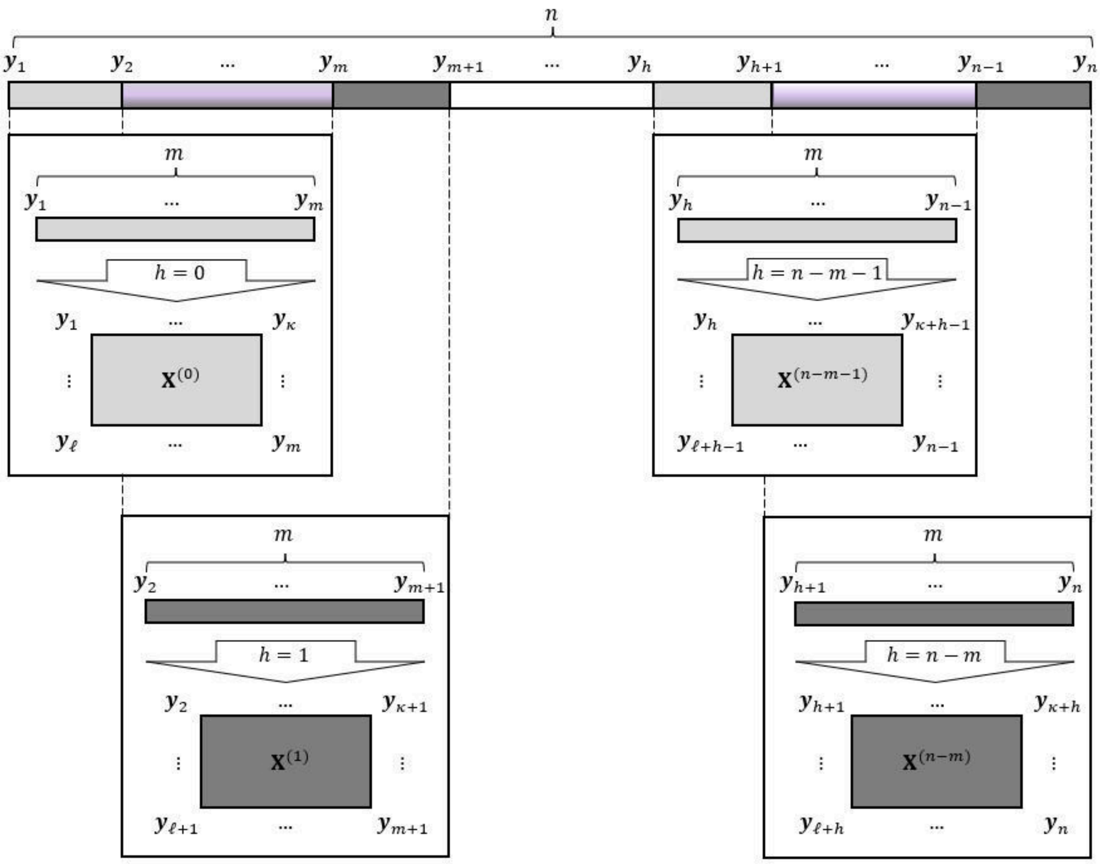

Fig. 1 Segmentation of a TS of length n, and the embedding step of the SSA applied to each subseries of length m.

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

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

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

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

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 1^{st} to the d^{th} 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 $\mathcal{D}_\phi, \phi = 1, \cdots, d$, correspond to the Frobenius norm of the matrices

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

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 $i = 0, 1, 2, \cdots, n-m$ 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 $\mathcal{D}_1, \cdots, \mathcal{D}_{d-1}$, we vertically look for the most remarkable differences among their curves, i.e., higher difference values among $\mathcal{D}_1(k), \cdots, \mathcal{D}_{d-1}(k)$. 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):

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)
```

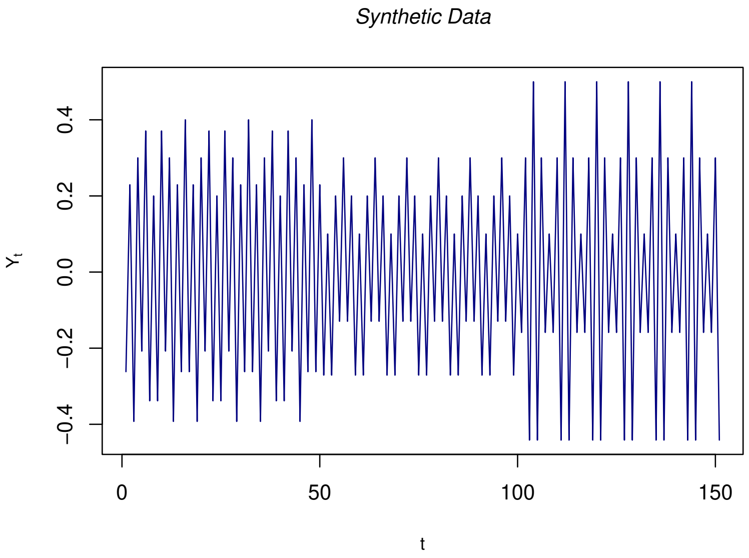

Fig. 2 Signal of the Synthetic data in which there are two structural changes (t_{51} and t_{101}).

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. 3 and are following as awaited. Also, there are three lines in the graph because, for $h = 0, \cdots, n - m$, $rank(\mathbf{X}^{(h)}) = 3$. 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.

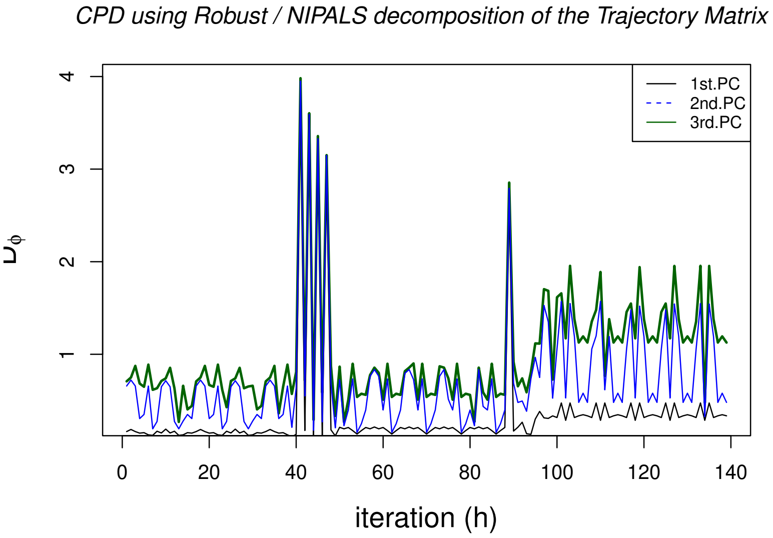

Fig. 3 The spikes in the curves representing $\mathcal{D}_\phi$ suggest two structural changes at observations y_{51} and y_{101}.

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.

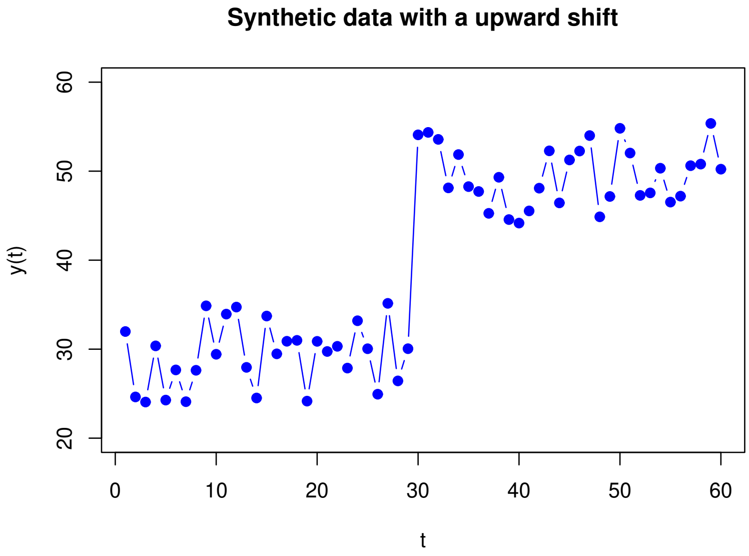

Fig. 4 Synthetic data presents an upward shift disturbance structural at t_{30}.

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

```
n = 60; Y = numeric(n)
m = 30; s = 2
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 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 $h = 0, \cdots, n - m$.

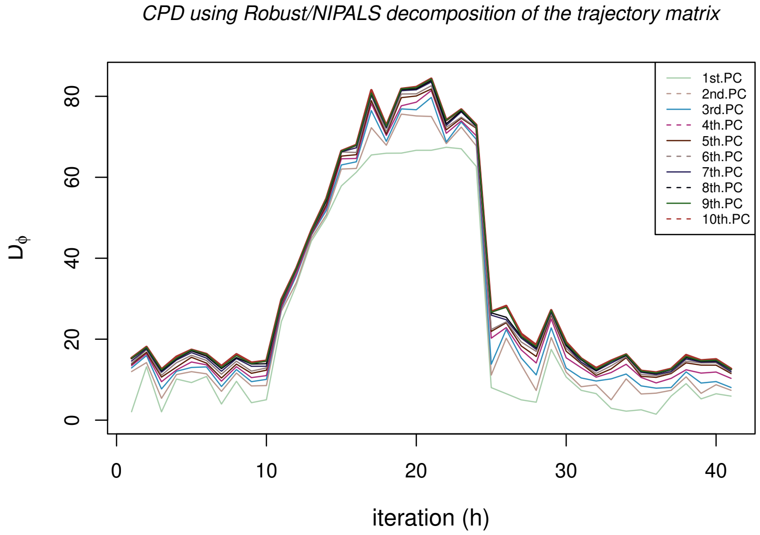

Fig. 5 A sharp increment in the curves representing $\mathcal{D}_\phi$ at iteration 11 (i.e., $h = 10$) suggests a structural change in the series from observation y_{30}. Since $m = 20$, this agrees with the proposed method, as $m + h = 30$.

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. Fig. 7 represents the proposed structural change detection method using the optimal value of m^* equals 24 and, therefore, a window length $\ell = 12$.

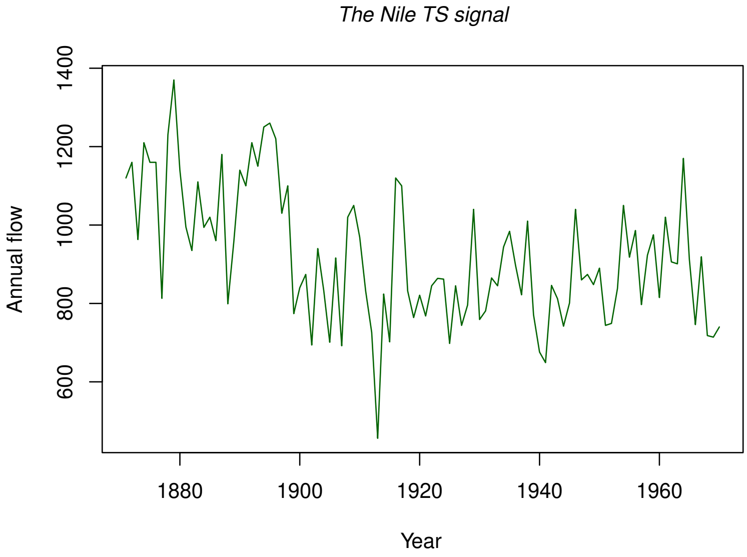

Fig. 6 Signal of the Nile TS, that represents the annual flow of the River Nile from 1871 to 1970.

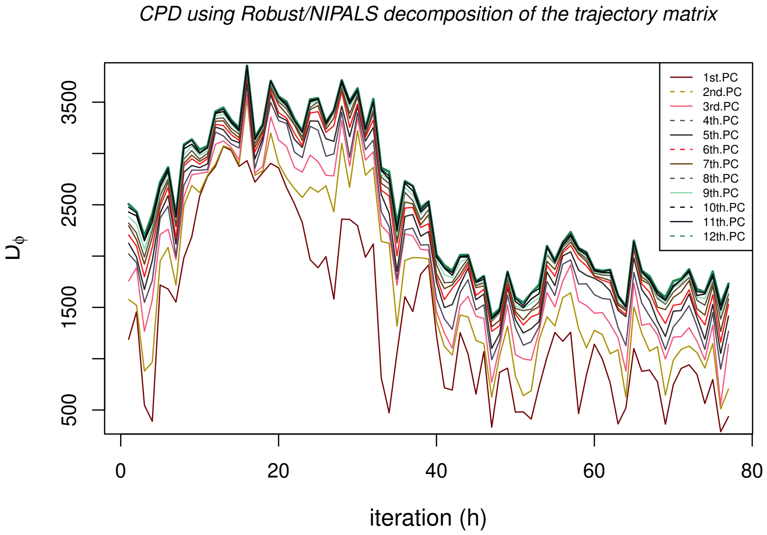

Fig. 7 A change point was detected at iteration 6 (i.e., $h = 5$) after a consistent increase in the values of $\mathcal{D}_\phi$. The subseries' length used was $m = 24$ and, therefore, the change point occurs in the observation $y_{h+m} = y_{29}$.

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 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 Subsection 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 Subsection 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.

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 20^{th}-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.

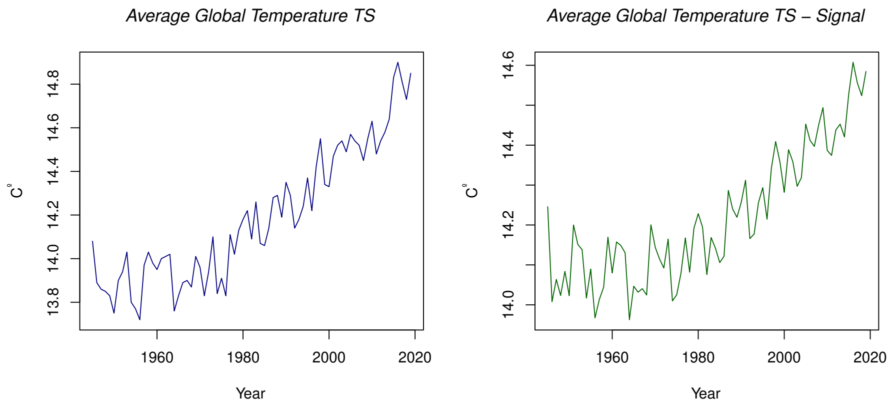

Fig. 8 TS of the average global surface temperatures plus anomalies, from 1945 to 2019. The original data is represented on the left and the series signal on the right.

Next, taking advantage of the SSA's grouping step, the signal was filtered using the first three eigentriples. The corresponding TS was reconstructed 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 2^{nd} 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 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). Fig. 11 shows the SSA-HJ-biplot of the TS for the 1^{st} and 2^{nd} principal components concerning the interval immediately before the estimated structural change, which comprises $1945 - 1965$. The first two principal components 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.

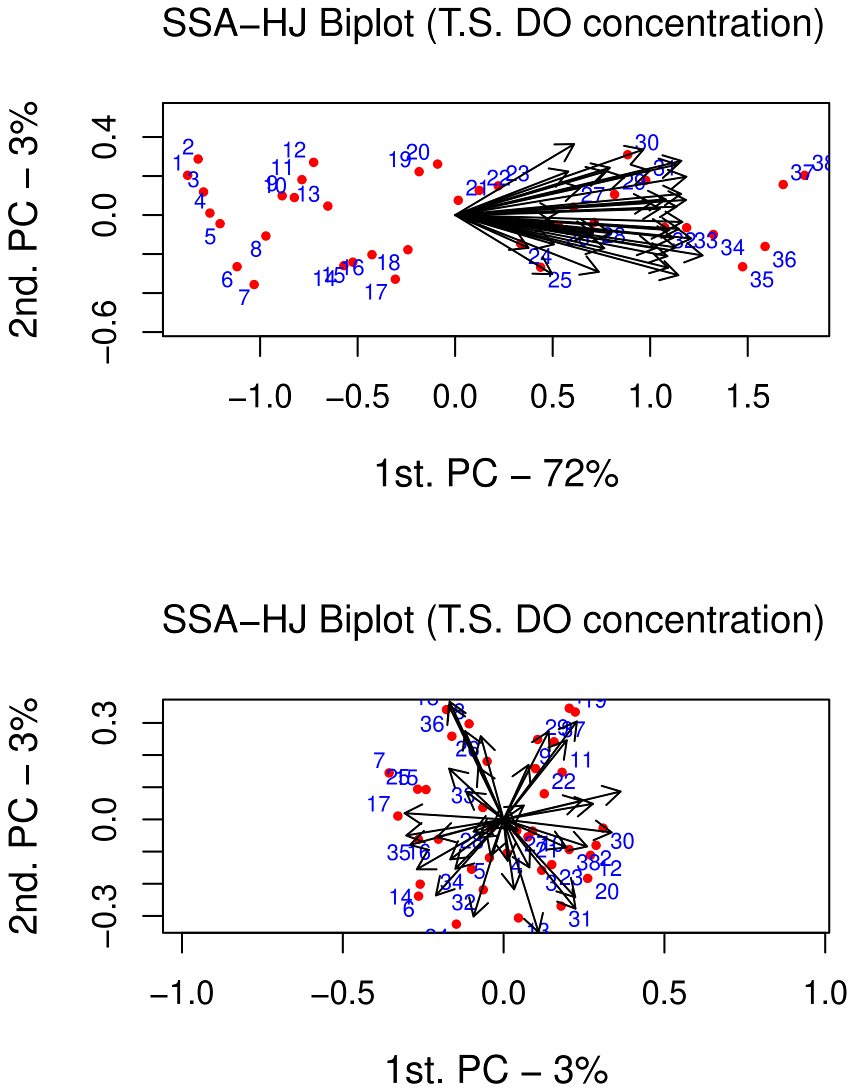

Fig. 9 The original approach of the SSA-HJ-biplot concerning the TS of average temperatures on the globe's surface from 1945 to 2019. The graph at the top refers to the 1^{st} and 2^{nd} principal components, followed by the SSA-HJ-biplot of the 2^{nd} and 3^{rd} principal components.

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

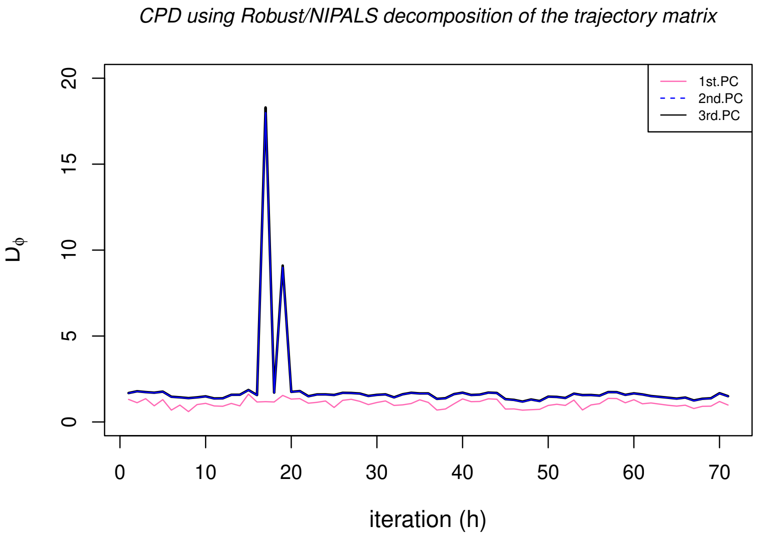

Fig. 10 The graph points out the values of $\mathcal{D}_\phi$ in each iteration according to each principal component's contribution. In this case, there is a substantial increase in iteration 17 (or $h = 16$). This means the occurrence of a change point in the observation y_{21}, since $m = 5$.

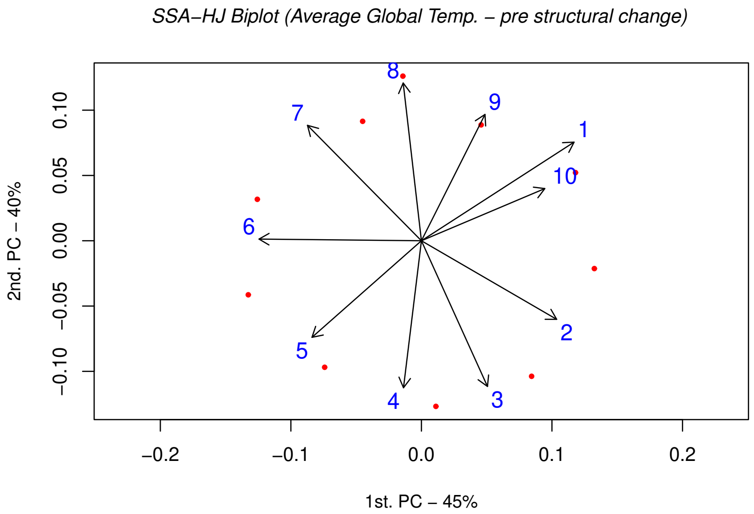

Fig. 11 SSA-HJ-biplot (1^{st} and 2^{nd} principal components) of the Average Global Temperature TS regarding the interval before the estimated structural modification (1945-1965).

In turn, the biplot points' projection on the 2^{nd} 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 2^{nd} 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 2^{nd} and 3^{rd} principal components is represented in Fig. 13. Since the 2^{nd} and 3^{rd} 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.

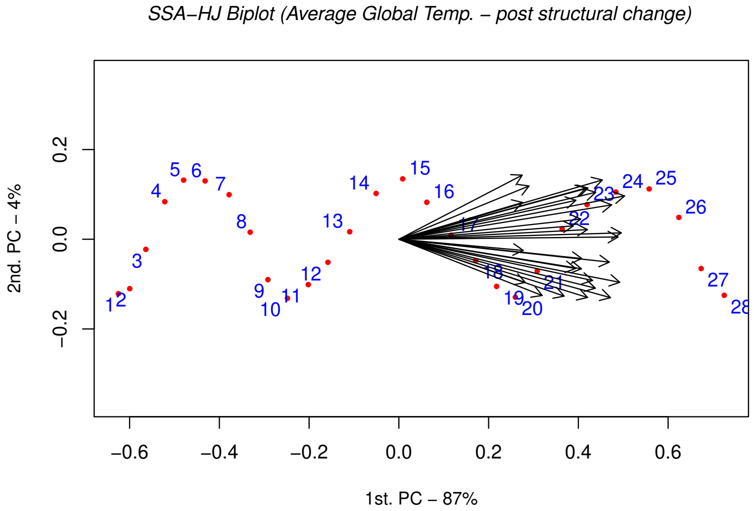

Fig. 12 SSA-HJ-biplot (1^{st} and 2^{nd} principal components) of the Average Global Temperature TS regarding the interval after the estimated structural modification (1966-2019).

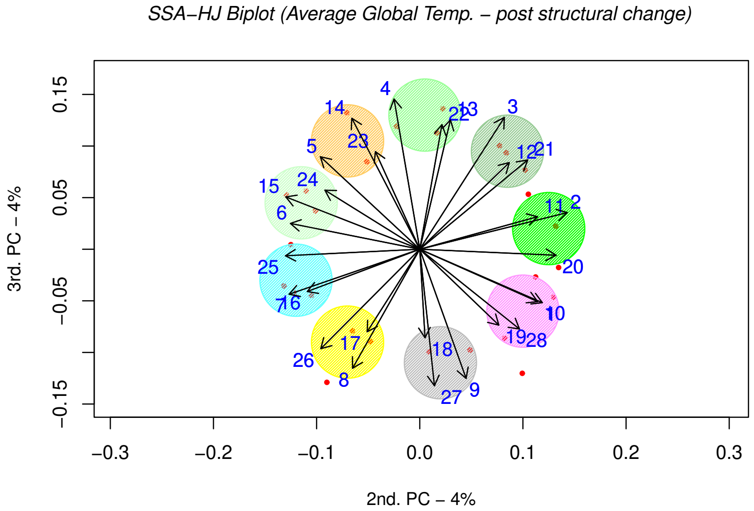

Fig. 13 SSA-HJ-biplot (2^{nd} and 3^{rd} principal components) of the Average Global Temperature TS regarding the interval after the estimated structural modification (1966-2019).

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. Fig. 14 shows the TS filtered after retaining the components explaining more than 1% of the data variability. Fig. 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.

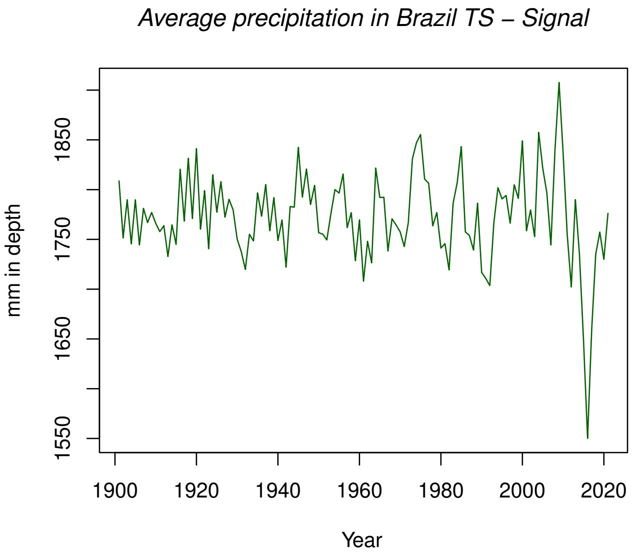

Fig. 14 TS of the average annual precipitation in Brazil, from 1901 to 2021. The data represents the series' signal.

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 $\mathcal{D}_\phi$.

Fig. 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} (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.

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

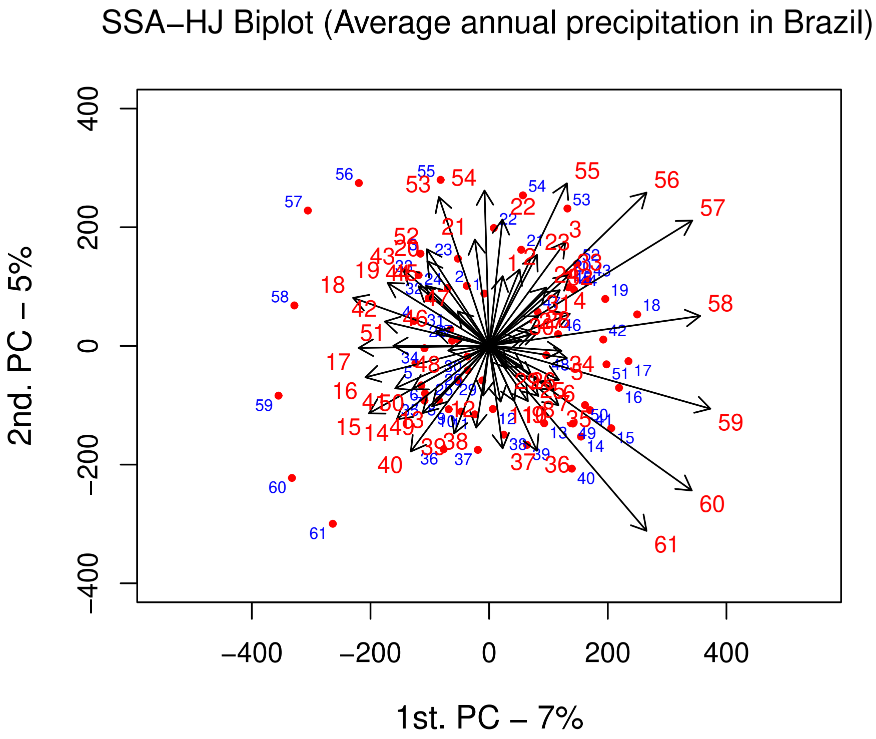

Fig. 15 The original approach of the SSA-HJ-biplot regarding the TS of average annual precipitation in Brazil from 1901 to 2021.

In Fig. 17 and 18, the SSA-HJ-biplot for the 1st and 2nd principal components for both intervals are presented. The subseries detected no trend 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 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.

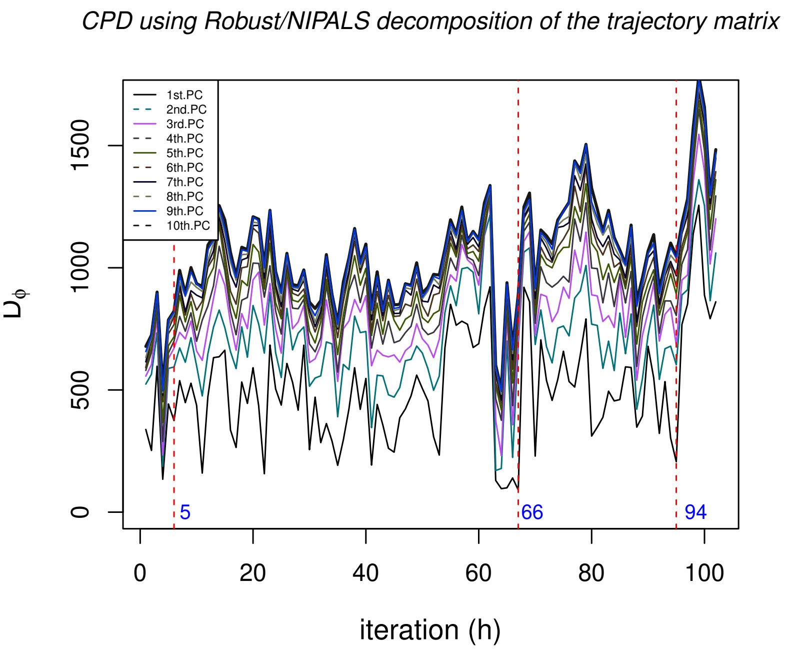

Fig. 16 For $m = 20$, the graph suggests the existence of three change points since $\mathcal{D}_d$ grows fast after the iterations 6 ($h = 5$) , 67 ($h = 66$), and 95 ($h = 94$).

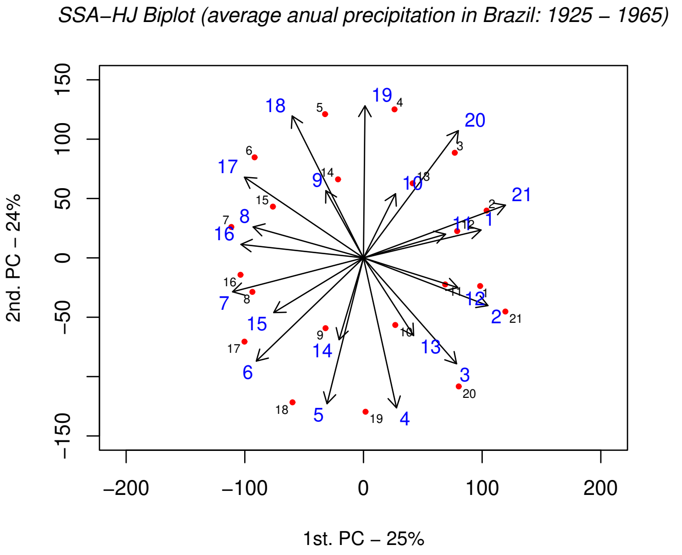

Fig. 17 SSA-HJ-biplot (1^{st} and 2^{nd} principal components) of the Average Annual Precipitation in Brazil TS regarding the interval between the two first change points (1925-1965).

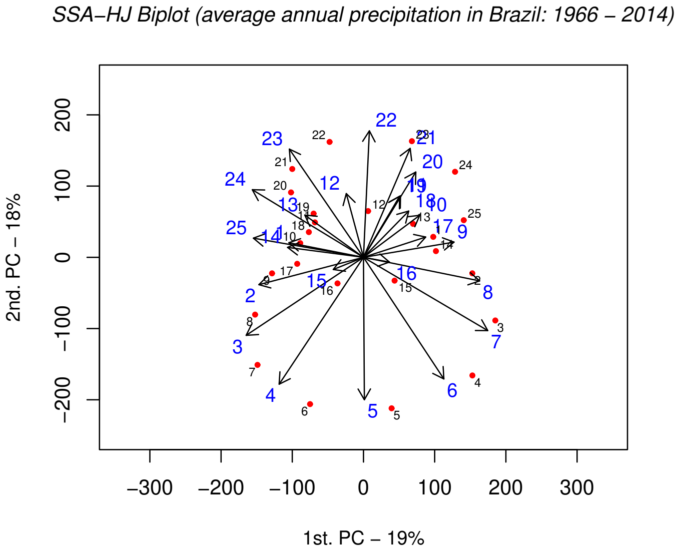

Fig. 18 SSA-HJ-biplot (1^{st} and 2^{nd} principal components) of the Average Annual Precipitation in Brazil TS regarding the interval between the second and third detected change points (1966-2014).

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 trajectory matrices since it allows dealing with missing data without needing any imputation. This substitution is possible because the matrices $\mathbf{T}$ (scores matrix) and $\mathbf{UD}$ (the product of the left singular vectors matrix times the singular values matrix) are equivalent, as well as $\mathbf{P}$ (loadings matrix) and $\mathbf{V}$ (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 Section 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.

Acknowledgments. 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 material. 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.

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 change-point problem. Biometrika 65(2):243–251

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

Esposito Vinzi V, Russolillo G (2013) Partial least squares algorithms and methods. Wiley Interdisciplinary Reviews: Computational Statistics 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. Analytica chimica acta 185:1–17

Golyandina N, Nekrutkin V, Zhigljavsky AA (2001) Analysis of time series structure: SSA and related techniques. CRC press

Hassani H, Mahmoudvand R (2018) Singular spectrum analysis: Using R. Springer

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 International Publishing, 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. Journal of chemometrics 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, Austria, URL 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

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

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: Food research and data analysis: proceedings from the IUFoST Symposium, September 20-23, 1982, Oslo, Norway/edited by H. Martens and H. Russwurm, Jr, London: Applied Science Publishers, 1983.

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

Chapter 7

Article IV

Time series periodicity detection using area biplots

Preprint:

Silva, A., Freitas, A.. Time series periodicity detection using area biplots. *Preprint submitted.*

Preprint submitted article: NOV2021

Time series periodicity detection using area biplots

Alberto Silva *[a,b] and Adelaide Freitas[a,b]

[a] *Department of Mathematics, University of Aveiro, Portugal*
[b] *Center for Research & Development in Mathematics and Applications (CIDMA), University of Aveiro, Portugal*

October 25, 2022

An exploratory approach is proposed based on multivariate visualization techniques to estimate the dominant periodicities of a time series. The application of the NIPALS algorithm to the trajectory matrix of the Singular Spectral Analysis (SSA) is presented, resulting in (*i*) a diagonal matrix containing the norms of the score vectors (singular values); (*ii*) a matrix formed by the normalized score vectors (left single vectors); (*iii*) another one formed by the loading vectors (right singular vectors). Pairs of singular values close to each other suggest the respective principal components (PCs) are associated with the periodicity of the time series. The proposed method consists of constructing the biplot of these PCs, pinning a biplot vector of interest (i.e., some loading vector associated with a lagged vector), and the 90° rotation of the others. Depending on the variability explained by the PCs, the areas of the triangles formed by (*i*) the origin of the factorial axes, (*ii*) the endpoints of the pinned vector, and (*iii*) each of the rotated vectors will provide visual information regarding the magnitude of the autocorrelation between the corresponding lagged vectors. In addition, the periodicity will emerge from the appearance of groups of similar triangles because of the strong autocorrelation between groups of lagged vectors. The periodogram should confirm the analysis if the data are not well represented in the biplot. In addition to the method, the authors developed the R package *areabiplot*, available for use in the Comprehensive R Archive Network (CRAN).

keywords: Area Biplots, NIPALS algorithm, Periodicity Detection, Singular Spectrum Analysis.

*Corresponding author: albertos@ua.pt

1 Introduction

Understanding a time series (TS) behavior can be an essential advantage for understanding the associated phenomenon and making predictions about it. In this context, periodicity is a crucial feature of an oscillatory TS. If its additive components are separable or at least approximately separable, the basic Singular Spectrum Analysis (SSA) is an effective tool for extracting the periodic component.

The SSA is a powerful non-parametric method used to analyze a TS, both for exploratory purposes and for making predictions (Elsner and Tsonis, 1996). In contrast to other methods, SSA is indifferent to (*i*) the model's specification; (*ii*) restrictive assumptions (e.g., stationarity); (*iii*) the length of the TS for forecasting purposes. Basic SSA is intended for one-dimensional, real-value TS and consists of two complementary stages: *decomposition* and *reconstruction*. First, the TS is transformed into a Hankel trajectory matrix ($\mathbf{X}$) in the so-called embedding step of the decomposition stage. Then, the singular value decomposition (SVD step) of $\mathbf{X}$ is calculated, resulting in a sum of rank-one matrices such that

$$\mathbf{X} = \mathbf{X}_1 + \mathbf{X}_2 + \cdots + \mathbf{X}_d, \tag{1}$$

where d is the rank of $\mathbf{X}$. Each $\mathbf{X}_i$ is computed as the product of the respective singular value ($\sqrt{\lambda_i}$), the left singular vector ($\mathbf{u}_i$), and the transposed right singular vector ($\mathbf{v}_i^{'}$). Note that $\mathbf{u}_i$ is the ith eigenvector of $\mathbf{X}\mathbf{X}^{'}$, $\mathbf{v}_i$ is the ith eigenvector of $\mathbf{X}^{'}\mathbf{X}$, and λ_i is the ith eigenvalue of both $\mathbf{X}\mathbf{X}^{'}$ and $\mathbf{X}^{'}\mathbf{X}$. In the literature of the SSA, the collection ($\sqrt{\lambda_i}, \mathbf{u}_i, \mathbf{v}_i$) is called the ith *eigentriple* of the SVD.

In the second stage, some of those rank-one matrices are appropriately grouped. The diagonal averaging step obtains an approximation to the original TS (or each component separately). The eigentriples and the shape of the TS are related. Applying data visualization techniques to these eigentriples helps select the proper groups and extract the corresponding components. For instance, a scree plot can reveal seasonal components showing the formation of plateaus in the eigenvalue spectra or by mean of geometric patterns in a pairwise scatter plot of the singular vectors (see Golyandina et al. (2001) and Hassani and Mahmoudvand (2018) for more details).

Another way to use eigenvectors to reveal the components of a TS graphically is through Biplot methods (Silva and Freitas, 2020). A biplot is usually a 2-dimensional graph that allows the joint plotting of both objects (rows) and variables (columns) of multivariate data matrices (Gabriel, 1971). Biplots can reveal essential characteristics of a multivariate data structure, e.g., similarities between observations, correlations between variables, and data variability. In the biplot theory, points represent the objects, and vectors connected to the origin (arrows) describe the variables. An element x_{ij} of the data matrix is visually estimated by projecting the point associated with the row i into the vector related to column j, then multiplying the result by the length of this vector. The so-called Area Biplot (Gower et al., 2010) is a particular method case. In this, the area covered by a triangle formed by (*i*) a point rotated by 90°, (*ii*) the origin of the factorial axes, (*iii*) and the end of an arrow is used to estimate an element of the data matrix. In this technique, the graphical representation of the correlation structure can also be made through the areas of the triangles (Gower et al., 2010; Graffelman, 2013).

In any case, when using biplots to visualize the eigentriples of the trajectory matrix, one should observe that the rows and columns of **X** are lagged vectors (subseries) of the original TS. Besides, the interpretability of biplots strongly depends on (*i*) the percentage of variability explained by the principal components extracted in the **X** decomposition; (*ii*) characteristics of the TS; (*iii*) and the separability of its components. When interpretation makes interpretation difficult, the periodogram can help identify the components in the SSA decomposition through biplots (Golyandina et al., 2001).

The periodogram analysis is an important auxiliary tool of the SSA to identify the eigentriples associated with periodicity. The periodogram is an estimator for the spectral density f and can be computed from the Discrete Fourier Transform (DFT) of a real-valued sequence y_n, $n = 0, \ldots, N-1$. Note that the normalized DFT of y_n is a sequence of complex numbers $Y(f)$ given by

$$Y(f_{k/N}) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} y(n) e^{-\frac{i2\pi kn}{N}}, \ for\ k = 0, \ldots, (N-1), \tag{2}$$

where k is the index in the frequency domain of the DFT and k/N indicates the frequency that each coefficient registers. The periodogram is then computed as the squared norm of each Fourier coefficient associated with the frequency k/N:

$$\mathcal{P}(f_{k/N}) = \|Y(f_{k/N})\|^2, \ for\ k = 0, \ldots, \lceil \frac{N-1}{2} \rceil. \tag{3}$$

Within the scope of SSA, one should search for singular vectors whose frequencies coincide with those of the original TS. Considering the periodogram of an eigentriples pair, existing a peak at a particular frequency, one should expect they are related to the signal (Hassani and Mahmoudvand, 2018).

This study suggests a visualization approach for identifying the periodicity of a TS through a variation of the Area Biplot method. Pairs of singular values close to each other suggest the respective principal components (PCs) are associated with the periodicity of the TS. The proposed method consists of constructing the biplot of these PCs, pinning a biplot vector of interest (i.e., some loading vector associated with a lagged vector), and the 90° rotation of the others. Depending on the percentage of variability explained by the PCs involved, the areas of the triangles formed by the origin of the factorial axes and the endpoints of the pinned vector and each of the rotated vectors will provide visual information regarding the magnitude of the autocorrelation between the corresponding lagged vectors. In addition, the periodicity will emerge from the appearance of groups of similar triangles because of the strong autocorrelation between groups of lagged vectors. If the data are not well represented in the biplot, the periodogram should confirm the analysis.

The paper is organized as follows. In **Section 2**, a short overview of the theoretical background related to the SSA-HJ-biplot and Biplot methods is provided. In **Section**

3, the proposed approach for TS periodicity detection is presented. In **Section 4**, the suggested technique is performed on two real-world TS using the statistical software R (R Core Team, 2021) and the R package *areabiplot*. Conclusions are presented in **Section 5**.

2 Background

2.1 Basics about SSA-HJ-Biplot

Using the results of both SSA and Biplot methods, SSA-HJ-Biplot is an exploratory graphic tool based on pairs of eigenvectors that, in the same plot, gather relevant information about a TS and can lead to the identification of its main characteristics (Silva and Freitas, 2020). It operates mainly after the SSA decomposition stage, i.e., following the *embedding* and *SVD* steps. In turn, its results are a helpful feature in conducting the *grouping* step of the reconstruction stage.

2.1.1 SSA decomposition stage

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 representing Y as κ lagged vectors $\mathbf{x}_1, \ldots, \mathbf{x}_\kappa$, each one of size ℓ, in which

$$\mathbf{x}_j = [y_j, \ldots, y_{(j+\ell-1)}]^{'}, \; 1 \leq j \leq \kappa. \tag{4}$$

These κ lagged vectors form the Henkel matrix $\mathbf{X} = [\mathbf{x}_1 \ldots \mathbf{x}_\kappa]$, which is called the trajectory matrix, having the following aspect:

$$\mathbf{X} = \begin{bmatrix} y_1 & y_2 & y_3 & \cdots & y_\kappa \\ y_2 & y_3 & y_4 & \cdots & y_{\kappa+1} \\ y_3 & y_4 & y_5 & \cdots & y_{\kappa+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_\ell & y_{\ell+1} & y_{\ell+2} & \cdots & y_n \end{bmatrix}. \tag{5}$$

Next, different from the classic approach of the basic SSA, in which X is factored through the SVD method, the SSA-HJ-biplot uses the NIPALS algorithm to decompose the trajectory matrix, and that results in

$$\mathbf{X} = \sum_{j=1}^{d} \mathbf{t}_j \mathbf{p}_j^{'}, \tag{6}$$

where $d = rank(\mathbf{X})$, each $\mathbf{t}_j$ is a score vector, and $\mathbf{p}_j^{'}$ is a loading vector. In matrix terms, one can rewrite (6) as follows

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

in which the columns of the matrix $\mathbf{T}$ are the score vectors $\mathbf{t}_j$, and the matrix $\mathbf{P}$ columns are the loading vectors $\mathbf{p}_j$. At each iteration, the NIPALS algorithm performs a linear regression of the $\mathbf{X}$ columns on a score vector $\mathbf{t}$, resulting in a loading vector $\mathbf{p}$. Then, the algorithm runs a linear regression of the $\mathbf{X}$ rows on the loading vector to get a new estimate for $\mathbf{t}$. The cycle repeats until it converges according to some criterion (Wold, 1966). The option for the NIPALS algorithm in this procedure is justified because it 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). Representing the indices of missing values by s in the estimation of the jth principal component, each iteration of NIPALS is such that

$$p_{rj} = \sum_{\substack{i=1\\ i\neq s}}^{\ell} x_{ir}t_{ij} \bigg/ \sum_{\substack{i=1\\ i\neq s}}^{\ell} t_{ij}^2 \ , \ for \ r = 1, \ldots, \kappa \tag{8}$$

and

$$t_{ij} = \sum_{\substack{r=1\\ r\neq s}}^{\kappa} x_{ir}p_{rj} \bigg/ \sum_{\substack{r=1\\ r\neq s}}^{\kappa} p_{rj}^2 \ , \ for \ i = 1, \ldots, \ell. \tag{9}$$

It means that, in the case of missing data, no imputation method is required when applying the NIPALS algorithm. To get the results of the NIPALS decomposition into factors equivalent to those of the SVD, one can normalize the score vectors so that

$$\mathbf{t}_i^* = \frac{\mathbf{t}_i}{\|\mathbf{t}_i\|} \Leftrightarrow \mathbf{t}_i = \sqrt{\mathbf{t}_i'\mathbf{t}_i}\mathbf{t}_i^* \ , \tag{10}$$

meaning that the decomposition of $\mathbf{X}$ can be write in terms of its left singular vectors $\mathbf{t}_i^*$, right singular vectors $\mathbf{p}_i$, and singular values $\sqrt{\mathbf{t}_i'\mathbf{t}_i}$. Each NIPALS eigentriple $(\sqrt{\mathbf{t}_i'\mathbf{t}_i}, \mathbf{t}_i^*, \mathbf{p}_i^{'})$, $i = 1, \ldots, d$, establishes an elementary matrix in which

$$\mathbf{X}_i = \sqrt{\mathbf{t}_i'\mathbf{t}_i}\mathbf{t}_i^*\mathbf{p}_i^{'} \ , \tag{11}$$

and consequently

$$\mathbf{X} = \sqrt{\mathbf{t}_1'\mathbf{t}_1}\mathbf{t}_1^*\mathbf{p}_1^{'} + \cdots + \sqrt{\mathbf{t}_d'\mathbf{t}_d}\mathbf{t}_d^*\mathbf{p}_d^{'}. \tag{12}$$

Defining $\mathbf{\Sigma}$ as a diagonal matrix containing the singular values $\sqrt{\mathbf{t}_i'\mathbf{t}_i}$ arranged in decreasing order, one can write the matrix form of the decomposition in (12) as

$$\mathbf{X} = \mathbf{T}^*\mathbf{\Sigma}\mathbf{P}^{'}, \tag{13}$$

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 $\mathbf{p}_i$.

2.1.2 Biplot method

A biplot is a graphical representation of a multivariate structure, generally used to reveal essential data characteristics, such as correlations between variables and similarities between observations (Greenacre, 2010), among others. The basic scheme starts from, e.g., factoring a d-rank $(\ell \times \kappa)$ trajectory matrix $\mathbf{X}$ in the form

$$\mathbf{X} \approx \mathbf{A}\mathbf{B}^{'}, \tag{14}$$

where $\mathbf{A}$ is an $(\ell \times p)$ matrix and $\mathbf{B}$ is an $(\kappa \times p)$ matrix, with $p \leq d$. These two matrices create two sets of p-dimensional points. Set $p = 2$, and then rows and columns of $\mathbf{X}$ can be simultaneously represented as a biplot. The biplot points reproduce the rows of $\mathbf{A}$, the so-called row markers $\mathbf{a}_1^{'}, \ldots, \mathbf{a}_\ell^{'}$. On the other hand, the biplot vectors connected to the origin (arrows) depict the columns of $\mathbf{B}^{'}$, i.e., the column markers $\mathbf{b}_1, \ldots, \mathbf{b}_\kappa$. Projecting a point onto an arrow followed by a multiplication by the length of the biplot vector is equivalent to computing the inner product $\mathbf{a}_i^{'}\mathbf{b}_j$, which provides an approximation to the corresponding element x_{ij}.

Regarding the decomposition in (13), one could consider more than one factorization form of $\mathbf{X}$ by making different choices for $\mathbf{A}$ and $\mathbf{B}$ and, consequently, getting back distinct biplot results. For example, opting for $\mathbf{A} = \mathbf{T}^*$ and $\mathbf{B} = \mathbf{P\Sigma}$, the factorization will preserve the metric of the columns of $\mathbf{X}$. That is the *classic* biplot (Gabriel, 1971), also called GH-biplot by Galindo (1986). Besides, if the matrix is centered by columns, this type of biplot satisfies the following properties: i) the norm of a column marker $\mathbf{b}_1$ is proportional to the standard deviation of the respective variable; ii) the cosine of the angle formed by column markers approximates the correlation between the related variables; iii) the columns are better represented than the rows in terms of quality. In turn, choosing $\mathbf{A} = \mathbf{T}^*\mathbf{\Sigma}$ and $\mathbf{B} = \mathbf{P}$, this factorization will preserve the metric of the rows so that the Euclidean distances between the row markers approximate the Euclidean distances between the respective individuals in the full space. In addition, the representation quality of the rows is better than the columns. It is called *form* biplot (Gabriel, 1971), or JK-biplot (Galindo, 1986).

Another possibility is the selection that occurs in the HJ-biplot method, where $\mathbf{A} = \mathbf{T}^*\mathbf{\Sigma}$ and $\mathbf{B} = \mathbf{P\Sigma}$. In this case, one can obtain an optimal representation of the ℓ rows and the κ columns of $\mathbf{X}$ in the same Euclidean space (Galindo, 1986; Nieto et al., 2014; Silva and Freitas, 2020). Conditioned on the quality of data representation in two-dimensional space, some of the possible interpretations of an HJ-biplot are i) the distance between points is expected to correspond to the dissimilarity between the associated individuals, just as in JK-biplots; ii) as it occurs in GH-biplots, approximately, the longer the arrow, the greater the correspondent standard deviation of the associated variable; iii) the cosine of the angle between arrows approximates the correlation between the variables they represent. A 90° angle indicates a weak correlation, while an angle close to 0 degrees or 180° suggests a strong correlation, positive in the first case and negative in the other. In the HJ-biplot, the inner product $\mathbf{a}_i^{'}\mathbf{b}_j$ does not approximate the element $x_{ij}, i = 1, \ldots, \ell$, and $j = 1, \ldots, \kappa$, but it is not a problem when it comes to the SSA-HJ-biplot. On it, the main objective is to visually identify the components of a TS (trend, seasonality, and noise) and associate the corresponding eigentriple to each of these. An SSA-HJ-biplot uses any 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 (Galindo, 1986; Silva and Freitas, 2020). Each PC associated with a TS component explains a proportion of the variability of the data, given by

$$PC_i(\%) = \frac{\mathbf{t}_i^{'}\mathbf{t}_i}{\sum_{j=1}^{d}\mathbf{t}_j^{'}\mathbf{t}_j}, \tag{15}$$

being that the higher the percentage of variability explained, the better the quality of the adjustment of the SSA-HJ-biplot. In this context, (i) biplot points whose Euclidean distances are small imply similarity in the behavior of the associated κ-lagged vectors; (ii) arrows with roughly the same size, indicating that the correspondent ℓ-lagged vectors have standard deviation also close; (iii) the angle between two arrows pinpoints the autocorrelation between the two ℓ-lagged vectors associated.

2.2 Area Biplot

The Area Biplot is a visualization technique used to estimate data values through the areas spanned by a triangle constructed from the results of the SVD factorization of a data matrix. In the original approach, Gower et al. (2010) propose another type of target matrix factorization to guarantee that the row and column markers exhibit a similar spread, facilitating the visual inspection of the produced biplots. It is done by standardizing the matrices $\mathbf{A}$ and $\mathbf{B}$, such that:

$$\mathbf{A} = (\frac{\ell}{\kappa})^{\frac{1}{4}}\mathbf{T}_2^{*}\mathbf{\Sigma}_2^{\frac{1}{2}}, \tag{16}$$

and

$$\mathbf{B} = (\frac{\kappa}{\ell})^{\frac{1}{4}}\mathbf{P}_2\mathbf{\Sigma}_2^{\frac{1}{2}}, \tag{17}$$

where $\mathbf{T}_2^{*}$ and $\mathbf{P}_2$ denote the first two columns of $\mathbf{T}^{*}$ and $\mathbf{P}$, and $\mathbf{\Sigma}_2$ the diagonal matrix with the two largest singular values. Doing so, the inner product matrix $\mathbf{AB}^{'}$ still approximates the matrix $\mathbf{X}$. The procedure starts with the rotation of the row markers by 90°, i.e.,

$$\mathbf{a}_i^{[r]} = \mathbf{R}\mathbf{a}_i, \tag{18}$$

in which $\mathbf{R}$ is the (2×2) 90° counterclockwise rotation matrix. Considering θ_{ij} the angle between the vectors $\mathbf{a}_i$ and $\mathbf{b}_j$, as well as that the inner product can be written as

$$\mathbf{a}_i^{'}\mathbf{b}_j = \|\mathbf{a}_i\|\,\|\mathbf{b}_j\|\,cos(\theta_{ij}), \tag{19}$$

the option for the $90° = \pi/2$ counterclockwise rotation is justified because

$$cos(\theta_{ij}) = sin(\theta_{ij} + \pi/2) = sin(\phi_{ij}), \tag{20}$$

and then

$$\mathbf{a}_i^{'}\mathbf{b}_j = \|\mathbf{a}_i\| \, \|\mathbf{b}_j\| \, sin(\phi_{ij}), \tag{21}$$

where ϕ_{ij} is the angle between the 90°-rotated biplot point $\mathbf{a}_i^{[r]}$ and the biplot vector $\mathbf{b}_j$. Besides, the expression (21) provides the area of the triangle formed by the origin and the endpoints of the vectors $\mathbf{a}_i^{[r]}$ and $\mathbf{b}_j$ multiplied by two (**Figure 1**). Therefore, the element x_{ij} may be estimated by the double of the triangle area.

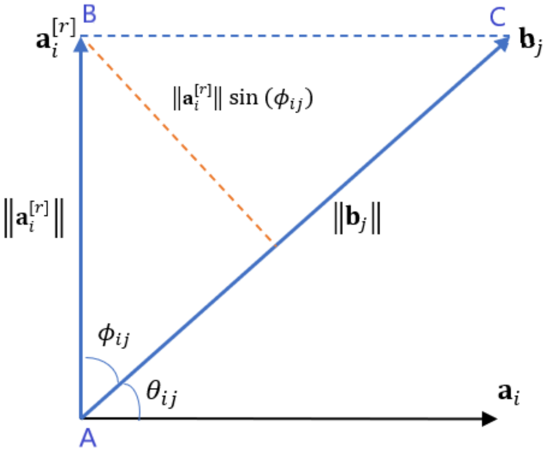

Figure 1: The inner product $\mathbf{a}_i^{'}\mathbf{b}_j$ is twice the area of the triangle ABC.

3 Our approach for periodicity detection

This section shows the foundations of an exploratory visualization technique used to estimate the periodicity of a TS called SSA Area Biplot. Through triangles built from sets of lagged vectors and whose autocorrelations are very close to each other, an estimate for the TS periodicity is obtained by computing the number of groups of almost similar triangles on the biplot graph. Similarly to in the area biplot method, the triangles here are built with the singular vectors resulting from the trajectory matrix decomposition. However, this approach uses the HJ-biplot factorization instead of the one proposed by Gower et al. (2010). As $\mathbf{X}$ is a Hankel matrix and the ℓ-lagged and κ-lagged vectors of the same order represent the same subseries (or almost the same), we are interested in representing rows and columns simultaneously with maximum quality. Let us initiate considering the trajectory matrix row markers as

$$\mathbf{A} = \mathbf{T}_2^*\Sigma_2 = \begin{bmatrix} \mathbf{a}_1^{'} \\ \mathbf{a}_2^{'} \\ \vdots \\ \mathbf{a}_\ell^{'} \end{bmatrix}, \tag{22}$$

and the column markers as

$$\mathbf{B} = \mathbf{P}_2\Sigma_2 = \begin{bmatrix} \mathbf{b}_1^{'} \\ \mathbf{b}_2^{'} \\ \vdots \\ \mathbf{b}_\kappa^{'} \end{bmatrix}. \tag{23}$$

In this context, $\mathbf{\Sigma}_2$ is a diagonal matrix with two consecutive singular values associated with some harmonic component of the series, and $\mathbf{T}_2^*$ and $\mathbf{P}_2$ denote the two corresponding columns of $\mathbf{T}^*$ and $\mathbf{P}$.

3.1 Triangles construction

The procedure starts with identifying a pair of singular values of $\mathbf{X}$ close to each other, such that

$$\sqrt{\mathbf{t}_i'\mathbf{t}_i} \approx \sqrt{\mathbf{t}_{i+1}'\mathbf{t}_{i+1}}. \tag{24}$$

The proximity between them means that the associated principal components (PC) are related to the periodicity of the TS (Golyandina et al., 2001; Silva and Freitas, 2020). Then, an adapted area biplot is built for the two selected PC, fixing one of the biplot vectors as a reference ($\mathbf{b}_f^{'}$). The vector $\mathbf{b}_f^{'}$ will serve as the base for the triangles, and hence it will be referred to as the base vector.

Next, all others biplot vectors are counterclockwise rotated by 90° ($\mathbf{b}_j^{[r]}$, $j \in \{1, \ldots, \kappa\} \setminus \{f\}$), unlike what occurs in the original area biplot method, in which the points are the objects that undergo the rotation. Triangles are then formed by connecting the endpoints of the base vector and each rotated vector (**Figure 2**). According to the SSA-HJ-biplot interpretation, the cosine of the angle formed by two biplot vectors approximates the autocorrelation between the corresponding lagged vectors. Thus, naming the angle between $\mathbf{b}_f^{'}$ and $\mathbf{b}_j^{'}$ as θ_{fj},

$$cos(\theta_{fj}) \approx corr(\mathbf{x}_f; \mathbf{x}_j), \tag{25}$$

and also

$$cos(\theta_{fj}) = sin(90^\circ \pm \theta_{fj}). \tag{26}$$

In this way, taking into account the direction of $\mathbf{b}_f^{'}$, and considering $\mathbf{b}_j^{[r]}$ is on its left, then the correlation will have a positive sign. It occurs because, if $0 \leq (90^\circ \pm \theta_{fj}) \leq$

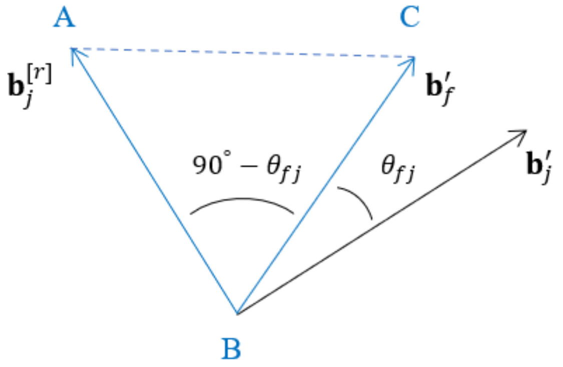

Figure 2: Triangle ABC formed by the reference and rotated vectors.

180°, then $sin(90^\circ \pm \theta_{fj}) > 0$. Otherwise, the correlation will have a negative sign. Furthermore, concerning the angle whose vertex coincides with the origin of the factorial axes, when triangle ABC (**Figure 2**) is almost a right triangle, then the $corr(\mathbf{x}_f; \mathbf{x}_j)$ will be close to 1.

Assuming that the data is well represented in the biplot in terms of the percentage of variability explained by the PCs, it is expected that:

1. For a fixed biplot vector $\mathbf{b}_f^{'}$, and if π is the dominant period of the TS, then there exists an integer η such that

$$\measuredangle(\mathbf{b}_f^{'}, \mathbf{b}_{\eta+\mathrm{k}\pi}^{[r]}) \approx 90^\circ, \;\; \mathrm{k} \in \{0, 1, \ldots, \lfloor(\kappa - \eta)/\pi\rfloor\}. \tag{27}$$

   In other words, the corresponding images will be close to being right-angled triangles and, for stationary series in the variance, among those with the largest area. Considering the direction of the base vector, if the vectors $\mathbf{b}_{\eta+\mathrm{k}\pi}^{[r]}$ are to its left, then the autocorrelation between the associated ℓ-lagged vectors will be close to 1 and positive. Otherwise, it will be strongly negative.

2. Likewise, there exists an integer τ for which the norm of the difference between the base vector $\mathbf{b}_f^{'}$ and some rotated biplot vectors is close to zero, such that

$$||\mathbf{b}_f^{'} - \mathbf{b}_{\tau+\mathrm{k}\pi}^{[r]}|| \approx 0, \;\; \mathrm{k} \in \{0, 1, \ldots, \lfloor(\kappa - \tau)/\pi\rfloor\}, \tag{28}$$

   and then the related triangles will have minimal areas and imply a weak autocorrelation between the corresponding ℓ-lagged vectors.

3. In intermediate situations, one or more cohesive groups of triangles may appear (depending on the periodicity). Inside each group, The associated ℓ-lagged vectors will be strongly correlated with each other but less intensely with the ℓ-lagged vector corresponding to the base vector.

4. In any case, considering the direction of the base vector, the rule is that groups of triangles with similar shapes to the left of $\mathbf{b}_f^{'}$ will indicate a positive correlation. When to the right, the correlation will be negative.

5. The periodicity of the TS is estimated by the number of groups of almost similar triangles.

6. If some biplot vectors have approximately the same size, this indicates that the corresponding ℓ-lagged vectors also have close standard deviations. So, if there is a natural number $\nu < \kappa$ such that

$$\|\mathbf{b}_\nu^{'}\| \approx \|\mathbf{b}_{\nu+j}\|, \ \ \forall j = 1, 2, \ldots, \kappa - \nu, \tag{29}$$

then,

$$var(\mathbf{x}_\nu) \approx var(\mathbf{x}_{\nu+j}), \ \ \forall j = 1, 2, \ldots, \kappa - \nu. \tag{30}$$

This leads to a greater similarity between triangles within the same group and suggests that a stationary process in the variance may have generated the series.

As stated before, the SSA does not require rigid assumptions or model specifications, and the separability of the components sought is usually only approximate (Golyandina et al., 2001; Hassani and Mahmoudvand, 2018). Thus, these setbacks can lead to misrepresenting data in the biplot. To work around this, one could use the periodogram of the series to compare with the SSA area biplot results. By providing a different perspective on the spectral structure of the TS, the periodogram analysis can be an important confirmation tool on what is extracted from the analysis of biplot triangles.

4 Experimental results

We implemented an R package entitled x to test the technique, applying it over two datasets provided by the North-American agency National Oceanic and Atmospheric Administration (NOAA) and involving climate change-related issues, i.e., carbon dioxide (CO_2) concentration in the atmosphere and Wildfires. The first TS contains the records of the CO_2 concentration in the Earth's atmosphere, measured monthly from January 1965 to December 1980 at an observing station on Mauna Loa in Hawaii. This is referred to in this work as TS CO2. The graphical representation of the TS CO2 is shown in **Figure 3**. The second dataset records the average monthly wildfire statistics provided by the National Interagency Fire Center (NIFC), available from January 2013 to December 2020 (called here as TS Wildfire).

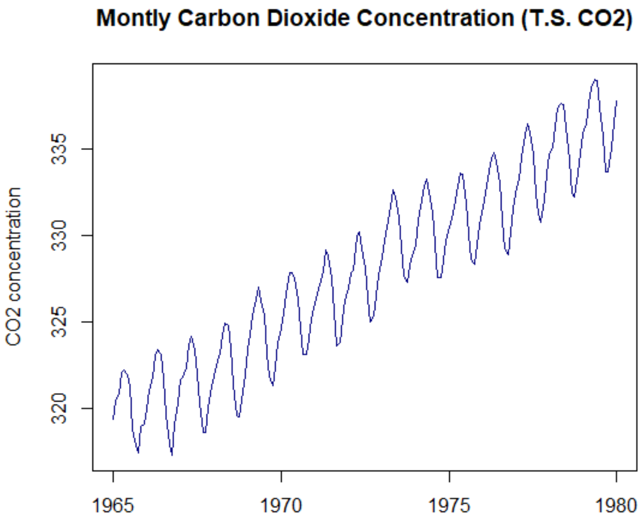

Figure 3: Records of CO_2 concentration in the Earth's atmosphere measured monthly from January of 1965 to December of 1980 at an observing station on Mauna Loa in Hawaii.

4.1 Time Series CO2 - Hawaii

In this case, the length of the series is $n = 192$ and the window length is $\ell = n/2 = 96$, resulting in $\kappa = n - \ell + 1 = 97$.

Figure 4 brings the scree plot of the singular values resulting from the NIPALS decomposition of $\mathbf{X}$ related to the TS CO2. This criterion suggests the first five principal components are associated with the series signal. As the first singular value is dominant, the first eigentriple is related to the trend (Silva and Freitas, 2020). Following, $\sqrt{\mathbf{t}_2'\mathbf{t}_2}$ and $\sqrt{\mathbf{t}_3'\mathbf{t}_3}$ are close to each other and form a plateau in the scree plot. So, it indicates the eigentriples associated with them are related to the oscillatory component. The same occurs with the 4^{th} and 5^{th} singular values, which corresponding eigentriples will also be associated with the periodicity of the series. The first five components explain 98% of the data variability (**Table 1**). From the 6th principal component onwards, the proportion of explained variance drops off and can be considered noise-related.

To detect periodicities in the TS CO2, two area biplots were built, one using the 2^{nd} and 3^{rd} PCs (**Figure 5**), and another related to the 4^{th} and 5^{th} PCs (**Figure 6**). In both, we labeled the vertices of the triangles according to the start month of the respective ℓ-lagged vector. Besides, we choose the first ℓ-lagged vector (1st column of $\mathbf{X}$) to be the base vector. Picking out a different base vector, one can get similar results. Regarding the SSA area biplot in **Figure 5**, the proportion of explained variance by the factorial axes is 29%, with the 2^{nd} PC explaining 15% and the 3^{rd} PC explaining

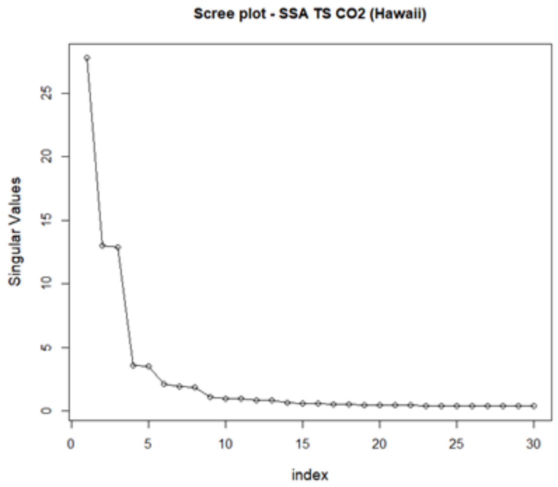

Figure 4: Scree plot of the singular values resulting from the NIPALS decomposition of the TS CO2 trajectory matrix. By this criterion, the first five components must be considered when separating the signal from the series.

Table 1: Proportion of explained variance by the ten first PCs of the **X** (TS CO2).

PC_i	Variance (%)	PC_i	Variance (%)
1	67	6	0.4
2	15	7	0.3
3	14	8	0.3
4	1	9	0.1
5	1	10	0.1

14% of the data variability. There are twelve distinct groups of nearly similar triangles, and since the size of the biplot vectors is very close to each other, this also suggests the stationarity of the series in the variance. Furthermore, each group of triangles is composed of biplot vectors corresponding to ℓ-lagged vectors that start in a specific month of the year. Therefore, each of the twelve groups presents biplot vectors whose proximity of the angles suggests a strong autocorrelation between the related ℓ-lagged vectors. And this points to a periodicity 12. Thus, the harmonic component of period 12 will be associated with the eigentriples $(\mathbf{t}_2^*, \sqrt{\mathbf{t}_2'\mathbf{t}_2}, \mathbf{p}_2)$ and $(\mathbf{t}_3^*, \sqrt{\mathbf{t}_3'\mathbf{t}_3}, \mathbf{p}_3)$.

The subseries corresponding to the 1st ℓ-lagged vector associated with the base vector $(\mathbf{b}_1^{'})$ starts in January. The triangles formed by $\mathbf{b}_1^{'}$ and each rotated biplot vector labeled J (January) establish a figure very close to a right triangle. Since they are to the left of the base vector, it indicates a strong and positive autocorrelation between the corresponding ℓ-lagged vectors.

On the other hand, triangles labeled L (July) establish patterns close to the right

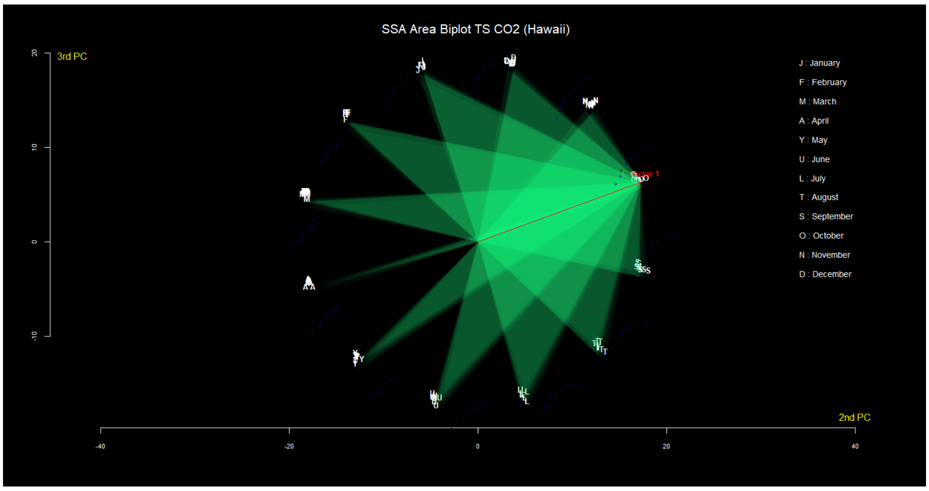

Figure 5: SSA Area Biplot of the TS CO2 related to the 2nd PC (15%) and 3rd PC (14%).

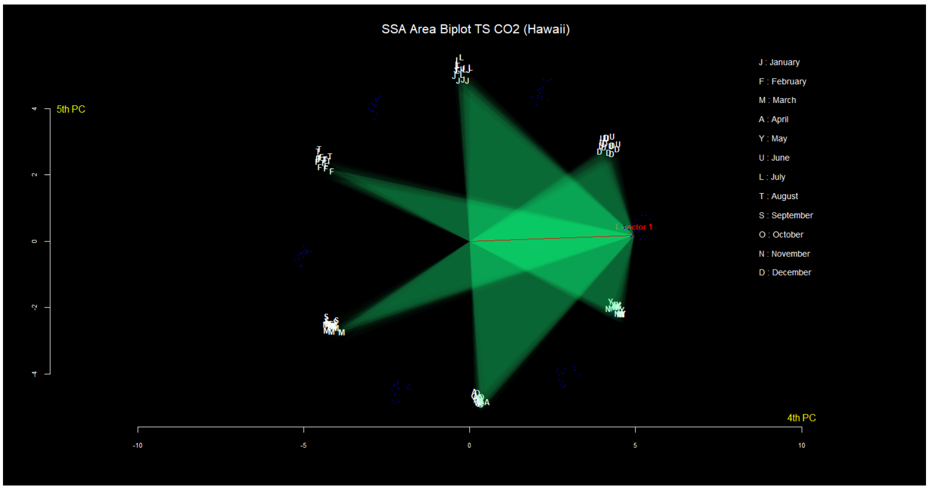

Figure 6: SSA Area Biplot of the TS CO2 related to the 4th PC (1%) and 5th PC (1%).

triangles but on the right side of the base vector. Therefore, it suggests a strong and negative autocorrelation between the 1st ℓ-lagged vector and each of the rotated ℓ-lagged vectors starting in July. In both cases, these are the triangles with the largest area, which will be proportional to the magnitude of the autocorrelations. In addition, triangles labeled N (November), D (December), J (January), F (February), M (March), and A (April) are on the left of the base vector, while that labeled Y (May), U (June), L (July), T (August), S (September), and O (October) are on the right. This suggests that the values of the series increase from November to April and decrease from May to October.

Concerning the SSA area biplot formed by the 4th and 5th PCs (**Figure 6**), each group of almost similar triangles is established by biplot vectors associated with pairs of months, i.e., J and L, F and T, and so on. Thus, the six groups lead us to conclude that the eigentriples $(\mathbf{t}_4^*, \sqrt{\mathbf{t}_4'\mathbf{t}_4}, \mathbf{p}_4)$ and $(\mathbf{t}_5^*, \sqrt{\mathbf{t}_5'\mathbf{t}_5}, \mathbf{p}_5)$ are related to the harmonic component of period 6. Considering the biplot vectors related to the months of these pairs appear in opposite directions in the biplot in **Figure 5**, one can conclude that the valleys tend to be six months behind the peaks.

To conclude, we built the TS CO2 periodogram (**Figure 7**), in which we can observe two peaks at frequencies 0.08(3) and 0.1(6), which leads us to infer that the two dominant periods of the series are 12 and 6 and confirms the SSA area biplot analysis.

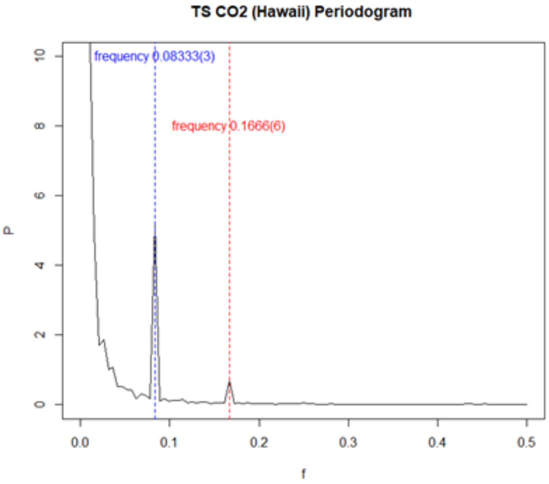

Figure 7: Periodogram of the TS CO2 showing two peaks at frequencies 0.08(3) and 0.1(6) and indicating 12 and 6 as the dominant periods.

4.2 Time Series Wildfire – USA

The sample size of the TS Wildfire series is $n = 96$, spanning January 2013 to December 2020. As usual, the window length is $\ell = n/2 = 48$ and $\kappa = n - \ell + 1 = 49$. **Figure 8** shows the shape of the TS Wildfire, and **Figure 9** depicts the scree plot of the singular values of its trajectory matrix $\mathbf{X}$.

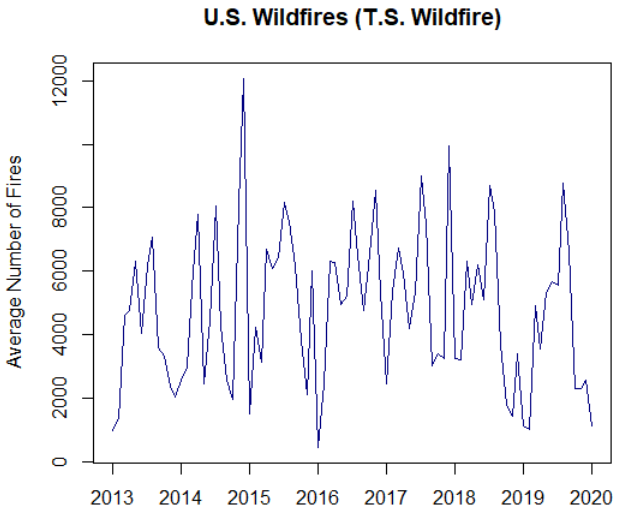

Figure 8: The monthly average number of wildfires in the USA.

The scree plot does not detach any dominant singular value, indicating no trend in the TS Wildfire. When it comes to the oscillatory component, it is much more difficult to identify a pair of eigentriples associated with periodicity in the TS Wildfire since the proportion of explained variance is not so high even for the first four PCs (**Table 2**), which totals just 48%.

Table 2: Proportion of explained variance by the ten first PCs of $\mathbf{X}$ (TS Wildfire).

PC$_i$	Variance (%)	PC$_i$	Variance (%)
1	14	6	4
2	13	7	3
3	11	8	3
4	10	9	3
5	4	10	3

Nevertheless, through the scree plot criterion, the first two singular values are close enough to suggest a correspondence between $(\mathbf{t}_1^*, \sqrt{\mathbf{t}_1'\mathbf{t}_1}, \mathbf{p}_1)$ and $(\mathbf{t}_2^*, \sqrt{\mathbf{t}_2'\mathbf{t}_2}, \mathbf{p}_2)$ with a periodic component. We can apply an analogous reasoning regarding the proximity of the next two singular values, $\sqrt{\mathbf{t}_3'\mathbf{t}_3}$ and $\sqrt{\mathbf{t}_4'\mathbf{t}_4}$, to deduce a relation between the eigentriples $(\mathbf{t}_3^*, \sqrt{\mathbf{t}_3'\mathbf{t}_3}, \mathbf{p}_3)$ and $(\mathbf{t}_4^*, \sqrt{\mathbf{t}_4'\mathbf{t}_4}, \mathbf{p}_4)$ with another harmonic component. From there is verified a sharp drop in the singular values, denoting the remaining ones are related to noise. Accordingly, **Figure 10** and **Figure 11** bring the two SSA area biplots corresponding to the 1st and 2nd PCs and the 3rd and 4th PCs.

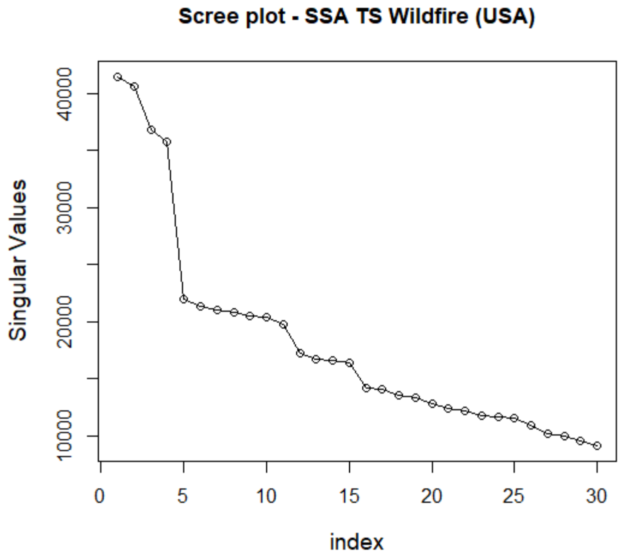

Figure 9: Scree plot of singular values resulting from the NIPALS decomposition of the TS Wildfire trajectory matrix. By this criterion, the first four components must be considered when separating the signal from the series.

Regarding the SSA area biplot related to the first two PCs (**Figure 10**), it is possible to identify four groups of triangles according to the start month of the corresponding ℓ-lagged vector, i.e., (J,Y,S),(F,U,O),(M,L,N), and (A,T,D). They are a little farther from the desired quasi-similarity, but the groups are sufficiently distinct to visually recognize a strong autocorrelation between the associated ℓ-lagged vectors. Thus, it looks like the first pair of eigentriples is connected to the harmonic component of period 4, but a confirmation through the periodogram is strongly recommended because it seems the data are not well represented in the biplot.

As for the biplot of the SSA area using the 3rd and 4th PCs (**Figure 11**), it is still possible to distinguish twelve groups of triangles regarding the start month of the ℓ-lagged vectors. But the heights concerning the base vector vary significantly, discarding the expectation of obtaining quasi-similar triangles. One could suspect that the 3rd and 4th PCs are related to the oscillatory component of period 12, but again, we need the periodogram to confirm these results.

Comparing what has been extracted from the SSA area biplot with what is shown in **Figure 12** allows us to conclude that the results of the initial visual analysis are consistent with the periodogram results, which present two peaks at frequencies 0.25 and 0.08(3), meaning that the two dominant periods of the TS Wildfire should be 4 and 12.

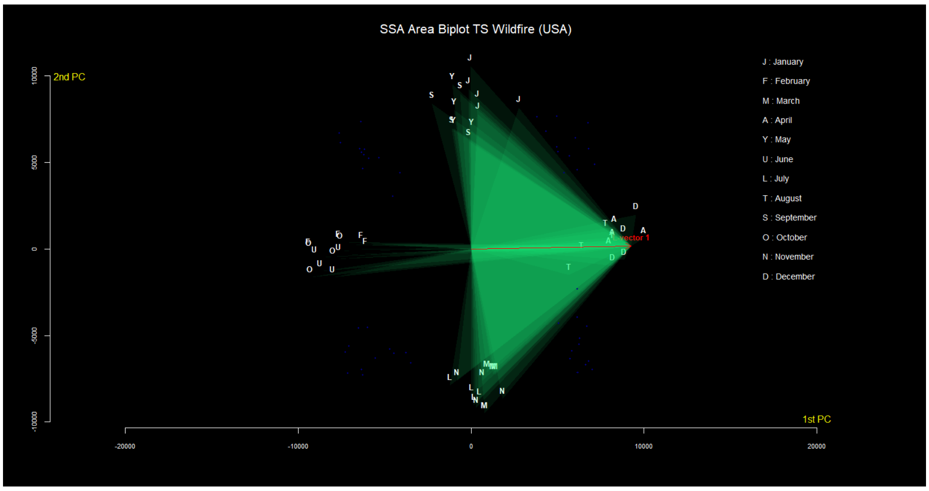

Figure 10: SSA Area Biplot of the TS Wildfire related to the 1st PC (14%) and 2nd PC (13%).

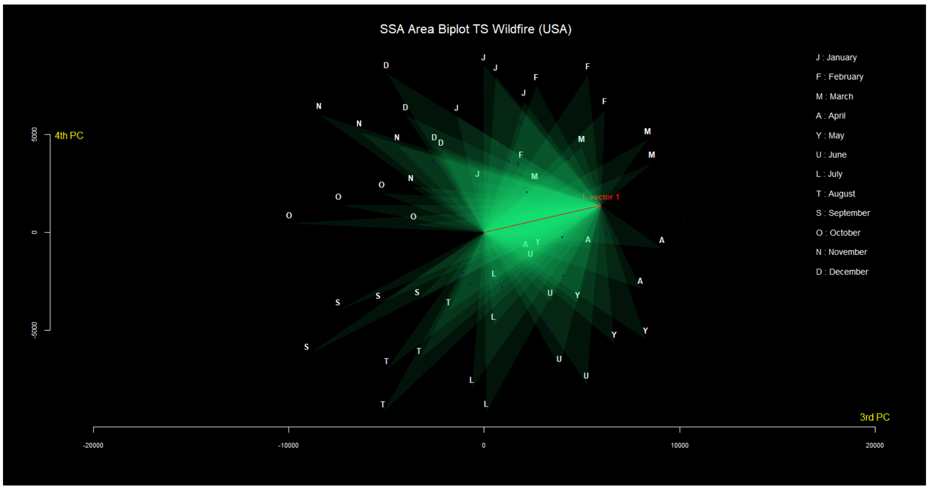

Figure 11: SSA Area Biplot of the TS Wildfire related to the 3rd PC (11%) and 4th PC (10%).

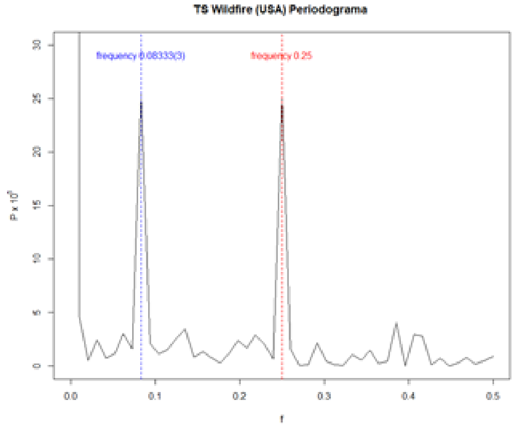

Figure 12: Periodogram of the TS Wildfire showing two peaks at frequencies 0.25 and 0.08(3) and indicating 4 and 12 as the dominant periods.

5 Discussions and Conclusion

The SSA method offers few resources in terms of visualization, mainly regarding the results of the trajectory matrix decomposition. In this paper, combining the SSA and the Area Biplot methods, we proposed a straightforward exploratory approach to visually estimate the periodicity of a TS through groups of quasi-similar triangles, which we call the SSA area biplot. Due to the lack of R packages capable of building area biplots, we have implemented the *areabiplot* package in R, which serves both the purposes stated in this article and more general multivariate data. The technique presented also provides interpretability for the ℓ-lagged vectors (which constitute the columns of the trajectory matrix) in terms of their autocorrelations. For cases in which the data are poorly represented in the biplot, either by the low proportion of variability explained or by the complexity of the series, we proposed using the periodogram as a confirmatory measure. The results obtained in the two applications of the procedure to climate TS data (TS CO2 and TS Wildfire) were consistent with those given by the respective periodograms regarding the periodicity of the series, even in the case of data badly described in the biplot (TS Wildfire). This study is promising in two senses: i) it has the potential to provide reliable starting points for the application of other more sophisticated methods; ii) for being an easily perceived visualization technique for any user.

Acknowledgement

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 Fundação para a Ciência e a Tecnologia (FCT) within project UID/MAT/04106/2019 (CIDMA).

References

Elsner, J. B. and Tsonis, A. A. (1996). *Singular spectrum analysis: a new tool in time series analysis*. Springer Science & Business Media.

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

Galindo, M. (1986). An alternative for simultaneous representation: Hj-biplot. *Questíio*, 10:12–23.

Golyandina, N., Nekrutkin, V., and Zhigljavsky, A. A. (2001). *Analysis of time series structure: SSA and related techniques*. CRC press.

Gower, J., Groenen, P., and van de Velden, M. (2010). Area biplots. *Journal of Computational and Graphical Statistics*, 19(1):46–61.

Graffelman, J. (2013). Linear-angle correlation plots: new graphs for revealing correlation structure. *Journal of Computational and Graphical Statistics*, 22(1):92–106.

Greenacre, M. J. (2010). *Biplots in practice*. Fundacion BBVA.

Hassani, H. and Mahmoudvand, R. (2018). *Singular spectrum analysis: Using R*. Springer.

Nieto, A. B., Galindo, M. P., Leiva, V., and Vicente-Galindo, P. (2014). A methodology for biplots based on bootstrapping with r. *Revista colombiana de estadística*, 37(2):367–397.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Silva, A. and Freitas, A. (2020). Time series components separation based on singular spectral analysis visualization: an hj-biplot method application. *Statistics, Optimization & Information Computing*, 8(2):346–358.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. *Multivariate analysis*, pages 391–420.

Wold, S., Albano, C., Dunn III, W., Esbensen, K., and Hellberg, S. (1983). Pattern recognition: finding and using regularities in multivariate data food research, how to relate sets of measurements or observations to each other. In *Food research and data analysis: proceedings from the IUFoST Symposium, September 20-23, 1982, Oslo, Norway/edited by H. Martens and H. Russwurm, Jr.* London: Applied Science Publishers, 1983.

Chapter 8

Article V

The HJ-Biplot Visualization of the Singular Spectrum Analysis Method

Published:

Silva, A., Freitas, A. (2020). The HJ-Biplot Visualization of the Singular Spectrum Analysis Method. in ITISE 2019. Proceedings of papers. Vol 2 (2019).

The HJ-Biplot Visualization of the Singular Spectrum Analysis Method

Alberto Oliveira da Silva (✉) [0000-0002-3496-6802] , Adelaide Freitas[0000-0002-4685-1615]

Department of Mathematics - University of Aveiro
Aveiro, 3810-193, Portugal
albertos@ua.pt

Abstract. Time series data usually emerge in many scientific domains. The extraction of essential characteristics of this type of data is crucial to characterize the time series and produce, for example, forecasts. In this work, we take advantage of the trajectory matrix constructed in the Singular Spectrum Analysis, as well as of its decomposition through the Principal Component Analysis via Partial Least Squares, to implement a graphical display employing the Biplot method. In these graphs, one can visualize and identify patterns in time series from the simultaneous representation of both rows and columns of such decomposed matrices. The interpretation of various features of the proposed biplot is discussed from a real-world data set.

Keywords: Singular Spectrum Analysis, NIPALS algorithm, Biplots.

1 Overview

Singular Spectrum Analysis (SSA) is a non-parametric method and a suitable tool to perform exploratory analysis on time series [6]. The Basic SSA schema is the version that deals with the description and identification of the structure of a one-dimensional real-valued time series. Basic SSA can be described as two successive stages: *decomposition* and *reconstruction*. The first one is subdivided into step 1, the *embedding*, and step 2, the *Singular Value Decomposition* (SVD), while the second consists of two other phases, the *grouping* and the *diagonal averaging*. The primary purpose is to decompose the original time series into the sum of a few interpretable components, such as trend, oscillatory shape (e.g., seasonality) which should be separated from a noise component [5].

For any matrix, the factorization given by SVD allows practical graphical representations of both rows and columns of the matrix employing biplots methods [2, 3]. Biplots provide easier interpretations, are much more informative than the traditional scatterplots, and might facilitate the work in the grouping step in SSA. Several types of biplots can be constructed depending on how the three factors identified by SVD are aggregated to obtain only two factors. Herein, the option is the biplot method proposed by Galindo [3], called HJ-biplot, which yields a simultaneous representation of both rows and columns of a matrix of interest with maximum quality [3].

The main objective of this paper is to propose a new exploratory procedure to visualize and identify patterns in the time series through the construction of an HJ-biplot from the results of the SVD step on the Basic SSA. Moreover, this work suggests an alternative approach to obtain the factorization referred in step 2 (first stage) based on the *Nonlinear Iterative Partial Least Squares* (NIPALS) algorithm [11] instead of the usual SVD method. Although it provides equivalent results concerning the singular vectors and the singular values, it empowers the SSA to deal with missing values in the data, without employing any imputation method, since NIPALS is a suitable tool to treat this problem [10, 12]. That occurs because, in each iteration of the NIPALS algorithm, only present data are considered in the regressions performed, ignoring the missing elements. This is equivalent to defining all missing points in the least squares objective function as zero.

The paper is organized as follows. In Section 2, we provide a short description of the theoretical background related to the methods involved in this work. In Section 3, we propose a biplot approach to the SSA method and some possible interpretations of it. In Section 4, we perform an application of the proposed technique by using real-world data set. Final conclusions are contained in Section 5.

2 Methods

2.1 Basic Singular Spectrum Analysis

The Basic SSA is a model-free tool used to recognize and identify the structure of a time series [5]. As before mentioned, it is composed of two complementary stages, as follows.

First Stage: Decomposition.
Consider a real-valued time series $\mathrm{Y} = (y_1, \ldots, y_N)$ of length N. Let the integer value L $(1 < L < N)$ be the so-called *window length*, as well as $K = N - L + 1$. Hereupon, the *embedding* procedure, that is the first step of the Basic SSA, consists in representing Y in K lagged vectors, $\mathbf{x}_1, \ldots, \mathbf{x}_K$, each one of size L (L-lagged vectors), i.e., $\mathbf{x}_j = (y_j, \ldots, y_{j+L-1})$, $1 \leq j \leq K$. This sequence of K vectors forms the trajectory matrix $\mathbf{X} = [\mathbf{x}_1 : \ldots : \mathbf{x}_K]$, that has as its columns the L-lagged vectors. Step 2, the SVD step, results in the singular value decomposition of the trajectory matrix. Consider that rank($\mathbf{X}$) is equal to d, and the matrix $\mathbf{S}$ is defined as the product $\mathbf{X}'\mathbf{X}$. So, the SVD of $\mathbf{X}$ is the decomposition in the form

$$\mathbf{X} = \textstyle\sum_{i=1}^{d} \sqrt{\lambda_i}\mathbf{u}_i\mathbf{v}_i', \tag{1}$$

where λ_i, $i = 1, \ldots, d$, are the eigenvalues of the matrix $\mathbf{S}$ arranged in decreasing order of magnitudes ($\lambda_i > 0$), $\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$ is the orthonormal system of the eigenvectors of $\mathbf{S}$ associated with the eigenvalues $\lambda_1, \ldots, \lambda_d$, and

$$\mathbf{u}_i = \mathbf{X}\mathbf{v}_\mathbf{i}/\sqrt{\lambda_i}. \tag{2}$$

The elements of the triple $\sqrt{\lambda_i}, \mathbf{u}_i, \mathbf{v}_i$ are also known as *singular values, left* and *right singular vectors* of **X**, respectively. Besides, defining

$$\mathbf{X}_\mathrm{i} = \sqrt{\lambda_i}\mathbf{u}_i\mathbf{v}_i', \tag{3}$$

one can represent **X** as a sum of d 1-rank matrices, i.e.,

$$\mathbf{X} = \mathbf{X}_1 + \cdots + \mathbf{X}_d. \tag{4}$$

Second Stage: Reconstruction.

Once the expansion (4) has been determined, the third step of the SSA starts with the partitioning of the index set $\{1, \dots, d\}$ into disjoints subsets $I_j, j = 1, \dots, p$. Let

$$\mathbf{X}_I = \sum_{i \in I} \mathbf{X}_i \tag{5}$$

and the decomposition can be written as

$$\mathbf{X} = \mathbf{X}_{I_1} + \cdots + \mathbf{X}_{I_p}. \tag{6}$$

The intention of the grouping procedure is the separation of the additive components of the time series [6]. The objective of the next phase, the *diagonal averaging* step, is to transform each matrix of the *grouping* decomposition into a new time series [5]. At this point, as in [6], it is convenient to define: $\mathbb{M}_{L,K}$ as the space of the matrices of dimension $(L \times K)$; $\mathbb{M}_{L,K}^{(H)}$ the space of Hankel matrices of dimension $(L \times K)$; the embedding operator $\mathcal{T}: \mathbb{R}^N \mapsto \mathbb{M}_{L,K}$ as $\mathcal{T}(Y) = \mathbf{X}$; and the projector $\mathcal{H}$ of $\mathbb{M}_{L,K}$ to $\mathbb{M}_{L,K}^{(H)}$, that carries out the projection by changing entries on auxiliary diagonals (where $i + j$ is a constant) to their averages along the diagonal. So, the diagonal averaging procedure corresponds to obtaining

$$\tilde{Y}^{(k)} = \mathcal{T}^{-1}[\mathcal{H}(\mathbf{X}_{I_k})] \tag{7}$$

and, then

$$Y = \sum_{k=1}^{p} \tilde{Y}^{(k)}. \tag{8}$$

2.2 PCA through NIPALS

The NIPALS algorithm belongs to the Partial Least Squares family, a set of iterative algorithms that implement a wide range of multivariate explanatory and exploratory techniques. The NIPALS is designed as an iterative estimation method for Principal Component Analysis (PCA), that computes the principal components through an iterative sequence of simple ordinary least squares regressions [10, 11]. It produces a singular value decomposition (SVD) of a matrix regardless of its dimensions and the presence of missing data [10]. Again, considering that the trajectory matrix has rank d, the method decomposes **X** as a sum of d 1-rank matrices in terms of the outer product of two vectors, a score $\mathbf{t}_i$ and a loading $\mathbf{p}_i$, so that

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1' + \cdots + \mathbf{t}_d\mathbf{p}_d'. \tag{9}$$

The elements of the scores vector $\mathbf{t}_i$ are the projections of the sample points on the principal component direction, while each loading in $\mathbf{p}_i$ is the cosine of the angle between the component direction vector and a variable axis [4]. The NIPALS first computes $\mathbf{t}_1$ and $\mathbf{p}_1$ from $\mathbf{X}$ and, then, the outer product $\mathbf{t}_1\mathbf{p}_1'$ is subtracted from $\mathbf{X}$ to calculate the residual matrix $\mathbf{E}_1$. After, $\mathbf{E}_1$ is used to compute $\mathbf{t}_2$ and $\mathbf{p}_2$, and the residual $\mathbf{E}_2$ is calculated subtracting $\mathbf{t}_2\mathbf{p}_2'$ from $\mathbf{E}_1$, and so on until to obtain $\mathbf{t}_d$ and $\mathbf{p}_d$. The NIPALS algorithm is shown in Algorithm 1.

Algorithm 1. NIPALS internal relations.

```
NIPALS
Input: $\mathbf{E_0} = \mathbf{X}$
Output: $\mathbf{P} = |\mathbf{p_1}: \ldots: \mathbf{p}_d|, \mathbf{T} = |\mathbf{t_1}: \ldots: \mathbf{t}_d|$
    for all $i = 1, \ldots, d$ do
        step 0: initialize $\mathbf{t}_i$
        step 1:
        repeat
            step 1.1: $\mathbf{p}_i = \mathbf{E}_{i-1}'\mathbf{t}_i/\mathbf{t}_i'\mathbf{t}_i$
            step 1.2: $\mathbf{p}_i = \mathbf{p}_i/\|\mathbf{p}_i\|$
            step 1.3: $\mathbf{t}_i = \mathbf{E}_{i-1}\mathbf{p}_i$
        until convergence of $\mathbf{p}_i$
        step 2: $\mathbf{E}_i = \mathbf{E}_{i-1} - \mathbf{t}_i\mathbf{p}_i'$
    end for
```

From the internal relations in each iteration of the NIPALS algorithm, and after normalizing $\mathbf{t}_i$, such that

$$\mathbf{t}_i^* = \mathbf{t}_i/\|\mathbf{t}_i\| \Leftrightarrow \mathbf{t}_i = \sqrt{\mathbf{t}_i'\mathbf{t}_i}\ \mathbf{t}_i^*\ , \tag{10}$$

the following equations can be verified [10]:

$$\mathbf{E}_{i-1}'\mathbf{E}_{i-1}\mathbf{p}_i = \lambda_i\mathbf{p}_i \tag{11}$$

$$\mathbf{E}_{i-1}\mathbf{E}_{i-1}'\mathbf{t}_i^* = \lambda_i\mathbf{t}_i^*, \tag{12}$$

where $\lambda_i = \mathbf{t}_i'\mathbf{t}_i$ is the eigenvalue of both matrices $\mathbf{E}_{i-1}'\mathbf{E}_{i-1}$ and $\mathbf{E}_{i-1}\mathbf{E}_{i-1}'$, as well as $\mathbf{p}_i$ and $\mathbf{t}_i^*$ are their corresponding eigenvectors. Thus, the NIPALS decomposition of $\mathbf{X}$ can be written as

$$\mathbf{X} = \sqrt{\mathbf{t}_1'\mathbf{t}_1}\ \mathbf{t}_1^*\mathbf{p}_1' + \cdots + \sqrt{\mathbf{t}_d'\mathbf{t}_d}\ \mathbf{t}_d^*\mathbf{p}_d'. \tag{13}$$

Now, define the matrix $\mathbf{\Sigma}$ as a diagonal matrix containing the singular values $\sqrt{\mathbf{t}_i'\mathbf{t}_i}$ arranged in decreasing order. So, one can write the matrix form of the expansion (13) as

$$\mathbf{X} = \mathbf{T}^*\mathbf{\Sigma}\mathbf{P}', \tag{14}$$

where $\mathbf{T}^*$ is the scores matrix whose column vectors $\mathbf{t}_i^*$ are orthonormal, and $\mathbf{P}$ is the loadings matrix whose column vectors $\mathbf{p}_i$ are also orthonormal.

2.3 HJ-Biplot

The term biplot is due to Gabriel [2] and is associated to a graphical representation that reveals essential characteristics of multivariate data structure, e.g., patterns of correlations between variables or similarities between observations [7]. Consider a target data matrix $\mathbf{Z}$ of dimension $(I \times J)$, and its decomposition in the form

$$\mathbf{Z} = \mathbf{A}\mathbf{B}', \tag{15}$$

where $\mathbf{A}$ is a matrix of dimension $(I \times Q)$, and $\mathbf{B}$ is a matrix of dimension $(J \times Q)$. The matrices $\mathbf{A}$ and $\mathbf{B}$ create two sets of points, and if $Q = 2$, then the rows and columns of $\mathbf{Z}$ can be simultaneously represented into a two-dimensional graph called biplot, in which the rows of $\mathbf{A}$ are reproduced by points and the columns of $\mathbf{B}'$ are expressed as vectors connected to the origin (arrows). Thus, the biplot displays the row markers $\mathbf{a}_1, \ldots, \mathbf{a}_I$ of $\mathbf{Z}$, as well as its column markers $\mathbf{b}_1, \ldots, \mathbf{b}_J$, so that the inner product $\mathbf{a}_i'\mathbf{b}_j$ is the element z_{ij} of $\mathbf{Z}$ [8]. Very briefly, the interpretation of the biplot representation can be performed as follows:

1. The distance between points corresponds to how different the associated individuals are (dissimilarities), mainly if they are well represented;
2. The size of the arrow is proportional to the standard deviation of the associated variable. The longer the arrow, the greater the standard deviation;
3. The cosine of the angle between arrows approximates the correlation between the variables they represent. Thus, if the angle is next to 90° it indicates a poor correlation, while an angle close to 0° or 180° suggests a strong correlation, being positive in the first case and negative in the other.

The most popular biplot is the classic one [2], in which the metric of the columns is preserved. This version is also designated by GH-biplot [8]. An essential property of the GH-biplot is that the biplot vectors have the same configuration of the data matrix columns and the quality of representation of columns is maximum. By choosing row and column markers properly, the HJ-biplot allows representing the rows and columns simultaneously in the same Euclidean space with optimal quality for both [3].

To construct an HJ-biplot version based on NIPALS instead of SVD as proposed in [3], it’s enough to demonstrate the relationship between $\mathbf{t}_i^*$ and $\mathbf{p}_i$, as will be done next.

From the equation (12), multiplying it to the left by $\mathbf{E}_{i-1}'$, it becomes

$$\mathbf{E}_{i-1}'\mathbf{E}_{i-1}(\mathbf{E}_{i-1}'\mathbf{t}_i^*) = \lambda_i(\mathbf{E}_{i-1}'\mathbf{t}_i^*) \tag{16}$$

Next, the vector normalization of $(\mathbf{E}_{i-1}'\mathbf{t}_i^*)$ results in $\mathbf{E}_{i-1}'\mathbf{t}_i^*/\sqrt{\mathbf{t}_i'\mathbf{t}_i}$, i.e., the vector $\mathbf{p}_i$. Proceeding in the same way with respect to equation (11), and multiplying it to the left by $\mathbf{E}_{i-1}$ we have

$$\mathbf{E}_{i-1}\mathbf{E}_{i-1}'(\mathbf{E}_{i-1}\mathbf{p}_i) = \lambda_i(\mathbf{E}_{i-1}\mathbf{p}_i). \tag{17}$$

After, $(\mathbf{E}_{i-1}\mathbf{p}_i)$ is normalized, which produces $\mathbf{E}_{i-1}\mathbf{p}_i/\sqrt{\mathbf{t}_i'\mathbf{t}_i}$, i.e., the vector $\mathbf{t}_i^*$. Hence,

$$\sqrt{\mathbf{t}_i'\mathbf{t}_i}\mathbf{p}_i = \mathbf{E}_{i-1}'\mathbf{t}_i^*, \tag{18}$$

and

$$\sqrt{\mathbf{t}_i'\mathbf{t}_i}\mathbf{t}_i^* = \mathbf{E}_{i-1}\mathbf{p}_i. \tag{19}$$

To unify the biplot axes scales similarly to what is done in [3], the following designation is done

$$\mathbf{a}_i = \mathbf{E}_{i-1}\mathbf{p}_i = \sqrt{\mathbf{t}_i'\mathbf{t}_i}\mathbf{t}_i^* \tag{20}$$

$$\mathbf{b}_i = \mathbf{E}_{i-1}'\mathbf{t}_i^* = \sqrt{\mathbf{t}_i'\mathbf{t}_i}\mathbf{p}_i. \tag{21}$$

Substituting (18) into (20), it follows that

$$\mathbf{a}_i = \mathbf{E}_{i-1}\mathbf{b}_i/\sqrt{\mathbf{t}_i'\mathbf{t}_i}, \tag{22}$$

and plugging (20) in (21) we get

$$\mathbf{b}_i = \mathbf{E}_{i-1}'\,\mathbf{a}_i/\sqrt{\mathbf{t}_i'\mathbf{t}_i}. \tag{23}$$

Thus, from (22) and (23), the coordinates of the i-th column are expressed as a function of the coordinates of the i-th row and vice versa. As a consequence, it allows the representation of the rows and columns in the same Cartesian coordinates system. Moreover, these expressions of the column and row coordinates lead to the maximum quality of the representation for rows and columns in the same system [3]. Considering the matrix form of the NIPALS decomposition in (14), it is worth to mention that for the configuration of the HJ-biplot, we have

$$\mathbf{A} = \mathbf{T}^*\mathbf{\Sigma}, \tag{24}$$

$$\mathbf{B} = \mathbf{P\Sigma}, \tag{25}$$

and so,

$$\mathbf{X} \neq \mathbf{AB}'. \tag{26}$$

3 The SSA-HJ-Biplot

The trajectory matrix that will be decomposed by the NIPALS algorithm at the second step of the first stage of the SSA has some peculiarities in relation to the usual multivariate data matrix. Instead of individuals and variables, the rows and columns of the trajectory matrix represent L-lagged and K-lagged vectors of a time series, respectively. That said, after the decomposition of **X**, a row marker in the HJ-biplot denotes a K-lagged vector and is depicted in the graph as a point. In turn, a column marker represents a L-lagged vector, being that an arrow symbolizes it. One of the goals of SSA-HJ-biplot is to assist in the grouping step and for this, building more than one SSA-HJ-biplot may be needed. The first SSA-HJ-biplot uses the 1st and the 2nd principal components (PC), the next one uses the 2nd PC and the 3rd PC, and so on as long as the remain components can explain the variability of the data, which is given by

$$PC_{\%}^{(i)} = \mathbf{t}_i'\mathbf{t}_i / \sum_{j=1}^{d} \mathbf{t}_j'\mathbf{t}_j, \tag{27}$$

or visually through the scree plot of the singular values ($\sqrt{\mathbf{t}_i'\mathbf{t}_i}$) [5].

The window length L has to be large enough so that each L-lagged vector captures a substantial part of the behavior of the time series [5], but at the same time, it permits the interpretability of the graphics display. A window length equals to $N/2$ provides both capabilities because it allows for the most detailed decomposition [5]. The interpretation of the first SSA-HJ-biplot is performed in terms of:

1. The proximity of points. Biplot points whose Euclidean distances are small imply similarity in the behavior of the associated K-lagged vectors;
2. The length of the biplot vectors. If the arrows are roughly the same size, this indicates that the L-lagged vectors have standard deviation also close, which suggests that the process is stationary in the variance;
3. The angle formed between biplot vectors. If the angle between the two arrows is next to 0°, it hints a strong and positive autocorrelation between the two L-lagged vectors associated (negative if next to 180°). If the angle is close to 90°, it is expectable an autocorrelation near to zero.

It is worth to take in mind the percentage of explained variability represented by the first two components, since the higher the percentage, the better the quality of the adjust of the SSA-HJ-biplot [3].

As a rule, a singular value represents the contribution of the corresponding PC in the form of the time series. As the tendency generally characterizes the shape of a time series, its singular values are higher than the others, that is, they are the first eigenvalues [1]. On the other hand, when two singular values are close enough, i.e.,

$$\sqrt{\mathbf{t}_i'\mathbf{t}_i} \approx \sqrt{\mathbf{t}_h'\mathbf{t}_h},$$

this is an evidence of the formation of plateaus in the scree plot and indicates that the associated SSA-HJ-biplot is informative about the oscillatory components of the time series [5], as long as the PC explain high variability of the data.

4 Example

In this Section, an SSA-HJ-biplot is constructed to a time series that contains the records the carbon dioxide concentration in the Earth's atmosphere, measured monthly from January of 1965 to December of 1980 at an observing station on Mauna Loa in Hawaii [9], referred as T.S. CO2 in this work and that is represented in **Fig. 1**. Two auxiliary plots in **Fig. 2** provide some hints for what to expect in an SSA-HJ-biplot analysis in the data. In **Fig. 2** (b), where the 1st PC is plotted against an index $j =$

$1, \ldots, K$, the presence of a trend component in T.S. CO2 is manifest, and this should emerge somehow in the first SSA-HJ-biplot, i.e., in the biplot where the axes are the 1st and 2nd PCs. **Fig. 3** brings the first SSA-HJ-biplot, where one can verify that the 1st PC explains 67% of the data variability, i.e., the trend direction. A channel formed by two dotted lines helps in the perception of the presence of the trend, although the 2nd PC contributes to attenuate the slope if compared with the plot in **Fig. 2** (b). Each one of the biplot points (in red) represents a K-lagged vector, and its corresponding label indicates the month in which the lagged vector starts. In this sense, accordingly to the graph legend, a point labeled as "O" means a K-lagged vector starting in October of some year, and a label "D" symbolizes that the respective K-lagged vectors begins in December, and so on. These points are the row markers, determined by the rows of $\mathbf{T}^{*}\mathbf{\Sigma}$, that is $\mathbf{a}_i' = \mathbf{t}_i'$, $i = 1, \ldots, L$.

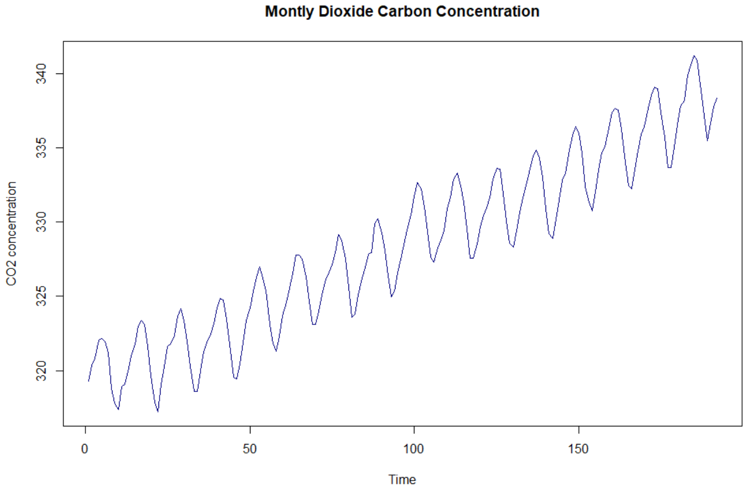

Fig. 1. Carbon dioxide concentration in the Earth's atmosphere measured monthly from January of 1965 to December of 1980 at an observing station on Mauna Loa in Hawaii.

According to the biplot theory, near points indicate similarity in the behavior of the lagged vectors, e.g., the points tagged as A, Y, and U in **Fig. 3**, i.e., the K-lagged vectors starting in April, May, and June. But not only that. Considering the labeling procedure before mentioned, the SSA-HJ-biplot is also capable of capturing the behavior of the months, since April, May, and June correspond precisely to the periods in which the highest concentration of carbon dioxide occurs in the atmosphere. It means that the points in the first SSA-HJ-biplot can represent not only the K-lagged vectors that start in a given month but also the month itself.

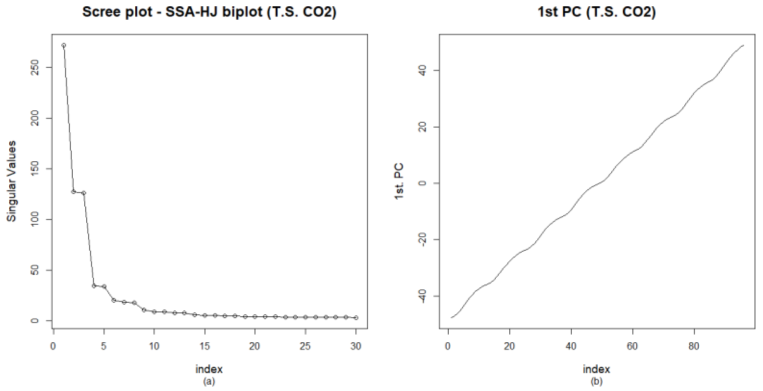

Fig. 2. Auxiliary plots in the SSA-HJ-biplot analysis.

In **Fig. 3**, the SSA-HJ-biplot represents the column markers (the L-lagged vectors) as black arrows up to the sixth L-lagged vector (tagged as $L1$ until $L6$), ordered from top to bottom. From the seventh L-lagged vector onwards the pattern repeats itself, and so they were plotted in gray. It means that the first group of arrows, which is at the top, refer to the L-lagged vectors beginning in January and July, just below those as starting in February and August, and so on. The angle between two consecutive arrows Li and Lj, such that $i = 1, \ldots, 5$ and $j = i + 1$, indicates a strong autocorrelation between the respective L-lagged vectors since Li and Lj form very sharp angles. As for $L1$ and the others up to $L6$, the angles range from something close to 0 to something close to 90 degrees, which suggests a fading of the autocorrelations. And this cycle repeats from $L7$ periodically, which suggests the non-stationarity also in the seasonality.

Fig. 4 shows the SSA-HJ-biplot formed by the 2nd and 3rd PCs, while **Fig. 5** exhibit the SSA-HJ-biplot constructed from the 4th and 5th PCs. Along with the first SSA-HJ-biplot, these are the only ones that produce interpretable results or evidence some pattern in the time series, being that these results are in agreement with the one verified in the scree plot of the singular values in **Fig. 2** (a), where the pair of points related to $\sqrt{\mathbf{t}'_2\mathbf{t}_2}$ and $\sqrt{\mathbf{t}'_3\mathbf{t}_3}$ are around at the same level, the same with respect to $\sqrt{\mathbf{t}'_4\mathbf{t}_4}$ and $\sqrt{\mathbf{t}'_5\mathbf{t}_5}$. In the SSA-HJ-biplot of **Fig. 4**, there are well defined 12 groups of row markers, where each one of these groups refers to a K-lagged vector that starts for a specific month. Also, the column markers associated with each one of these groups show strong autocorrelation between the L-lagged vectors. All of this indicates a seasonal pattern, with peaks and valleys separated by 12 months. In turn, the SSA-HJ-biplot of **Fig. 5** groups the lagged vectors two by two, e.g., January and July, February and August, and so on. Interpreting this together with the biplot of **Fig. 4**, where these same groups occur but in the opposite directions, one can conclude that the valleys tend to be six months behind the peaks.

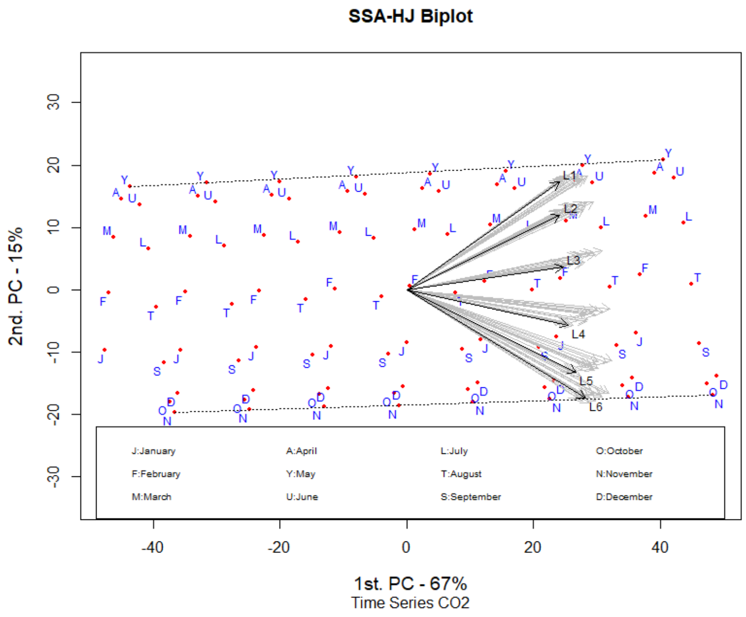

Fig. 3. First SSA-HJ-biplot of the T.S.CO2 trajectory matrix decomposition.

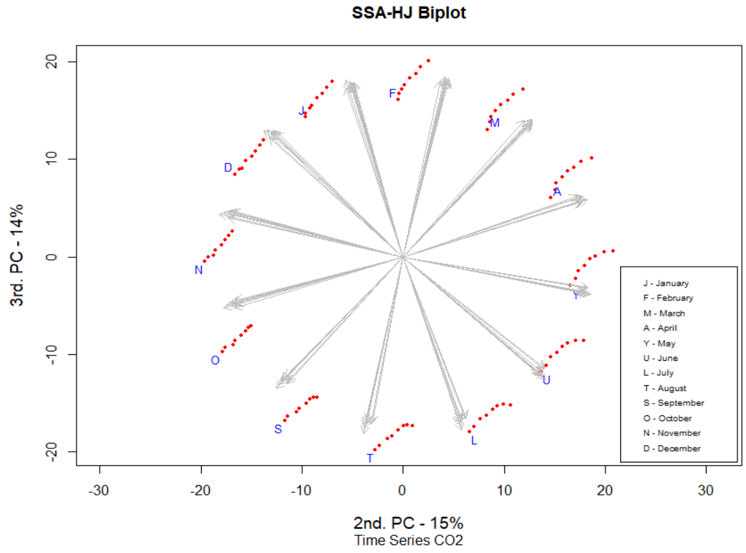

Fig. 4. The second SSA-HJ-biplot whose axes are the 2nd and 3rd PCs.

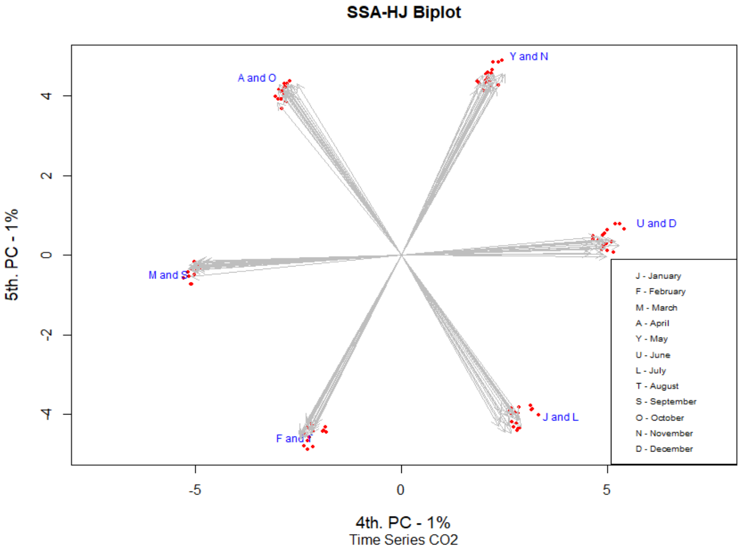

Fig. 5. The third SSA-HJ-biplot whose axes are the 4th and 5th PCs.

Therefore, the result of the grouping step for the decomposition of the T.S. CO2 should be $\mathbf{X}_1$ and $\mathbf{X}_2$, the first corresponding to the trend component, and the second describing the seasonal component, in which

$$\mathbf{X}_1 = \sqrt{\mathbf{t}_1'\mathbf{t}_1}\ \mathbf{t}_1^*\mathbf{p}_1', \tag{28}$$

and

$$\mathbf{X}_2 = \sum_{i=2}^{5}\sqrt{\mathbf{t}_\mathrm{i}'\mathbf{t}_\mathrm{i}}\ \mathbf{t}_\mathrm{i}^*\mathbf{p}_i', \tag{29}$$

with the rest being related to the noise component.

5 Conclusions

This paper attempts to provide an alternative way to visualize and understand the underlying structure of the trajectory matrix, that is the result of the embedding step of the SSA. The HJ biplot visualization method appears to be a promisor exploratory technique adequate to the purposes of this work since it provides interpretability to the results of the SVD step as was illustrated by an application. The SSA-HJ-biplots and auxiliary graphics provided a visual solution for the decomposition of the analyzed time series, properly separating the trend and the oscillatory component, using biplot axes up to the fifth PC. Also, allowed the identification of all relevant eigentriple, composed by the singular values $\sqrt{\mathbf{t}_\mathrm{i}'\mathbf{t}_\mathrm{i}}$, by the left singular vectors $\mathbf{t}_\mathrm{i}^*$, and by the right singular vectors $\mathbf{p}_\mathrm{i}$, $i = 1, \dots, 5$, to perform the grouping step. The study also revealed that the SSA-HJ-biplot points, representative of the row markers ($\mathbf{a}_\mathrm{i}'$) and symbol of the K-lagged vectors that begin in a given period of the series (months in this specific case) could also depict the period itself in terms of dissimilarities, being possible to visually verify the months with the highest and lowest levels of CO_2 concentration in the atmosphere throughout the years. The SSA-HJ-biplot built with the 1st and 2nd PCs proved yet to be useful in dealing with autocorrelations between the column markers, which are drawn as arrows and represent the L-lagged vectors. This study is promising in the sense that the SSA-HJ-biplot has a great potential as an exploratory tool to analyze the structure of a univariate time series due to its visual appeal in such a complex issue.

Acknowledgments.

The authors were supported by Fundação para a Ciência e a Tecnologia (FCT), within project UID/MAT/04106/2019 (CIDMA).

References

1. Alexandrov, T.: A method of trend extraction using Singular Spectrum Analysis. REVSTAT, Statistical Journal, **7**(1), 1-22 (2009)
2. Gabriel, K.: The biplot graphic display of matrices with application to principal component analysis. Biometrika, **58**(3), 453-467 (1971)
3. Galindo, M.P.: An alternative of simultaneous representation: HJ-biplot. Questiió, **10**(1), 13-23 (1986)
4. Geladi, P., Kowalsky, B.R.: Partial Least Squares regression: a tutorial. Analytica Chimica Acta, **185**, 1–17 (1986)
5. Golyandina, N., Nekrutkin, V., Zhigljavsky, A.: Analysis of Time Series Structure: SSA and Related Techniques. 1st ed. Chapman & Hall/CRC, Boca Raton, Florida (2001)
6. Golyandina, N., Shlemov, A.: Variations of Singular Spectrum Analysis for separability improvement: non-orthogonal decompositions of time series. Statistics and its Interface, **8**(3), 277–294 (2015)
7. Greenacre, M.: Biplots in Practice. FBBVA, Bilbao, Biscay (2010)
8. Nieto, A.B., Galindo, M.P., Leiva, V., Galindo, P.V.: A methodology for biplots based on bootstrapping with R. Colombian Journal of Statistics, **37**(2), 367–397 (2014)
9. NOAA Homepage, https://www.esrl.noaa.gov/gmd/ccgg/trends/, last accessed 2019/05/24
10. Vinzi, V.E., Russolillo, G.: Partial Least Squares algorithms and methods. WIREs Comput Stat, **5**, 1–19 (2013)
11. Wold, H.: Estimation of principal components and related models by iterative least squares. In: Krishnaiah, P.R. (ed.), Multivariate Analysis, New York: Academic Press, 391–420 (1966)
12. Wold, S., Albano, C., Dunn, W.J. III, Esbensen, K., Hellberg, S., Johansson, E., Sjostrom, M.: Pattern recognition: finding and using regularities in multivariate data. In: Martens, H., Russwurm, H. (eds.), Food Research and Data Analysis, London: Applied Science Publishers, 147–189 (1983)

Chapter 9

Article VI

PLS visualization using biplots: An application to team effectiveness

Published:

Silva, A., Dimas, I. D., Lourenço, P. R., Rebelo, T. & Freitas, A. (2020) PLS visualization using biplots: An application to team effectiveness. In: International Conference on Computational Science and Its Applications – ICCSA 2020. Lecture Notes in Computer Science(), vol 12251, pp. 214-230. Springer, Cham.

DOI https://doi.org/10.1007/978-3-030-58808-3_17

PLS Visualization Using Biplots: An Application to Team Effectiveness

Alberto Silva[1,2], Isabel Dórdio Dimas[3,4](✉), Paulo Renato Lourenço[3,5], Teresa Rebelo[3,5], and Adelaide Freitas[1,2]

[1] Departament of Mathematics, University of Aveiro, 3810-193 Aveiro, Portugal
{albertos,adelaide}@ua.pt

[2] CIDMA, Center for Research & Development in Mathematics and Applications, University of Aveiro, 3810-193 Aveiro, Portugal

[3] CeBER, Centre for Business and Economics Research, University of Coimbra, 3004-512 Coimbra, Portugal

[4] FEUC, University of Coimbra, 3004-512 Coimbra, Portugal
idimas@fe.uc.pt

[5] FPCEUC, University of Coimbra, 3000-115 Coimbra, Portugal
{prenato,terebelo}@fpce.uc.pt

Abstract. Based on a factorization provided by the Partial Least Square (PLS) methodology, the construction of a biplot for both exploratory and predictive purposes was shown to visually identify patterns among response and explanatory variables in the same graph. An application on a team effectiveness research, collected from 82 teams from 57 Portuguese companies and their respective leaders, containing two effectiveness criteria (team performance and the quality of the group experience as response variables), was considered and interpretation of the biplot was analyzed in detail. Team effectiveness was considered as the result of the role played by thirteen variables: team trust (two dimensions), team psychological capital (four dimensions), collective behavior, transformational leadership, intragroup conflict (two dimensions), team psychological safety, and team cohesion (two dimensions). Results revealed that the biplot approach proposed was able to capture the most critical variables for the model and correctly assigned the signals and the strength of the regression coefficients. Regarding the response variable team performance, the most significant variables to the model were team efficacy, team optimism, and team psychological safety. Concerning the response variable quality of the group experience, intragroup conflict, team-trust, and team cohesion emerged as the most relevant predictors. Overall, the results found are convergent with the literature on team effectiveness.

Keywords: Partial least square · Biplot · Organizational teams · Team effectiveness

O. Gervasi et al. (Eds.): ICCSA 2020, LNCS 12251, pp. 214–230, 2020.
https://doi.org/10.1007/978-3-030-58808-3_17

1 Introduction

Frequently, multivariate data analysis seeks to perceive the existing underlying structure and to understand the relationships established within data. Visual information via graphic displays can be a useful tool to explore the dataset since it summarizes the data more directly and improves its understanding (Koch 2014). Likewise, a graph of the results of a specific statistical method, e.g., the Principal Component Analysis (PCA) biplot, enhances data familiarity. The biplot method permits visual evaluation of the structure of large data matrices through the approximation of a high-rank matrix by one of rank two. The PCA biplot represents observations with points and variables with arrows. Small distances between units can indicate the existence of clusters, while the size of an arrow depicts the standard deviation of the associated variable. Further, the angle between two vectors approximates the linear correlation of the related variables (Gabriel 1971).

When it comes to multivariate regression problems, sometimes one must fix some problems before applying any methodology to estimate parameters and thinking about the graphical representation of its results. This is the case of an ill-posed problem, in which the predictors are many and quasi-collinear, leading to an unstable Ordinary Least Squares (OLS) solution, i.e., the OLS estimates have high variance (Belsley et al. 2004). Under this condition, the Partial Least Squares (PLS) regression gives better results, since it eliminates the quasi-collinearity issue. The PLS method extracts factors that maximize the covariance between the predictors and response variables, and then regresses the response on these latent factors. Based on the outputs of the PLS (scores, loadings, and weights vectors), the variances and correlations of the variables can be revealed by employing an *exploratory PLS biplot*. On the other hand, the PLS biplot can be adapted to provide a visual approximation of the PLS coefficient estimates, hence the reason for naming it *predictive PLS biplot*.

The primary purpose of this article is to provide a straightforward interpretation for the PLS biplot applicable to both exploratory and predictive purposes, illustrating its application in team effectiveness research data. Interest in understanding complex relationships between variables of team effectiveness datasets has been growing in recent years (Mathieu et al. 2019; Ringle et al. 2018) and the PLS biplot method can play a crucial role in the analysis of this kind of data.

In order to achieve the main aim of the present work, the paper is structured in the following sections: Sect. 2 gives a brief overview of how PLS works, describing mathematical details; Sect. 3 presents an application of these methods on a subset of variables of real work teams, exploring the relationships between a set of team effectiveness predictors (team trust, team psychological capital, collective behavior, transformational leadership, intragroup conflict, team psychological safety and team cohesion) and two team effectiveness criteria (team performance and quality of group experience). All the statistical analysis was executed using R software; finally, Sect. 4 includes the discussion of the results found, as well as conclusions and future perspectives.

2 Methods

2.1 Partial Least Squares

Assume a multivariate regression model $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$, in which $\mathbf{Y}$ is an (n × q) response matrix and $\mathbf{X}$ is a (n × m) predictor matrix, and both are column centered, with m and q being respectively the number of predictors and response variables, and n the number of observations. Also, $\mathbf{B}$ is a (m × q) coefficients matrix, and $\mathbf{E}$ is a (n × q) error matrix, such that m > n or the m explanatory variables are highly correlated. In this case, one might use the PLS to estimate the regression coefficients. The PLS model consists of three other models, two external and one internal, as a result of the application of a suitable algorithm, usually the Nonlinear Iterative Partial Least Squares (NIPALS). The method seeks to estimate some underlying factors that decompose $\mathbf{X}$ and $\mathbf{Y}$ simultaneously, maximizing the covariance between them, establishing the so-called outer relations for $\mathbf{X}$ and $\mathbf{Y}$ individually (Geladi and Kowalsky 1986). Considering the extraction of all possible factors, the PLS decomposition results in

$$\mathbf{X} \ = \mathbf{TP}' \text{ and } \mathbf{Y} \ = \mathbf{UQ}',$$

where $\mathbf{T}$ contains the scores of the predictors' matrix, $\mathbf{P}$ holds the loadings of $\mathbf{X}$. In turn, $\mathbf{U}$ and $\mathbf{Q}$ are the matrices of scores and loadings relative to the response matrix $\mathbf{Y}$. Additionally, an inner relation links the $\mathbf{X}$-scores and $\mathbf{Y}$-scores matrices as follows:

$$\widehat{\mathbf{u}}_\mathbf{i} = \mathrm{a}_\mathbf{i}\mathbf{t}_\mathbf{i},$$

where

$$\mathrm{a}_\mathrm{i} = \frac{\mathbf{u}_\mathrm{i}'\mathbf{t}_\mathrm{i}}{\mathbf{t}_\mathrm{i}'\mathbf{t}_\mathrm{i}}$$

are the regression coefficients for a given factor. In order to ensure maximum covariance between $\mathbf{Y}$ and $\mathbf{X}$ when extracting PLS components, it is necessary to find two sets of weights $\mathbf{w}$ and $\mathbf{q}$, which allow the vectors $\mathbf{t} = \mathbf{Xw}$ and $\mathbf{u} = \mathbf{Yq}$ to be obtained. It can be done making $\mathbf{t'u}$ maximum and solving the optimization problem

$$\begin{gathered}\mathrm{argmax}_{w,q}\{\mathbf{w}'\mathbf{X}'\mathbf{Yq}\},\\ \text{subject to: COR } (t_i, t_j) = 0, \ \forall i \neq j;\\ \boldsymbol{w}'\boldsymbol{w} = 1.\end{gathered}$$

2.2 Partial Least Squares Regression

Concisely, the NIPALS algorithm[1] performs the following steps (Abdi 2010):

[1] In this context, the symbol $\propto$ means 'to normalize the result of the operation'.

- Step 1. $\mathbf{w} \propto \mathbf{X'u}$ (X-weights).
- Step 2. $\mathbf{t} \propto \mathbf{Xw}$ (X-factor scores).
- Step 3. $\mathbf{q} \propto \mathbf{Y't}$ (Y-weights).
- Step 4. $\mathbf{u} = \mathbf{Yq}$ (Y-scores).

At the i-th iteration of the algorithm, the PLS method estimates a single column $\mathbf{t}_i$ of the matrix $\mathbf{T}$ as a linear combination of the variables X with coefficients $\mathbf{w}$. This vector of weights $\mathbf{w}$ will compose the i-th column of the matrix of weights $\mathbf{W}$. Since in each iteration the matrix $\mathbf{X}$ is deflated, the columns of $\mathbf{W}$ are non-comparable and, hence, $\mathbf{T} \neq \mathbf{XW}$. In contrast, there exists a matrix $\mathbf{R} = \mathbf{W}(\mathbf{P}'\mathbf{W})^{-1}$ which allows direct computation of $\mathbf{T}$ by doing $\mathbf{T} = \mathbf{XR}$ (Wold et al. 2004).

The estimated PLS regression equation is:

$$\widehat{\mathbf{Y}} = \mathbf{T}\widehat{\mathbf{B}}, \text{where } \widehat{\mathbf{B}} = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{Y}.$$

Moreover, $\widehat{\mathbf{Y}} = \mathbf{XR}\widehat{\mathbf{B}}$ and, thus, $\widehat{\mathbf{B}}_{\mathrm{PLS}} = \mathbf{R}\widehat{\mathbf{B}} = \mathbf{R}(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{Y} = \mathbf{RT}'\mathbf{Y} = \mathbf{RQ}'$. Notice that $\mathbf{Q}'$ is the Y-weights matrix composed of the $\mathbf{q}$ vectors estimated in Step 3 of the NIPALS algorithm. Lastly, we can write the predictive model as

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\mathbf{B}}_{\mathrm{PLS}}, \text{where } \widehat{\mathbf{B}}_{\mathrm{PLS}} = \mathbf{RQ}'.$$

2.3 The Biplot

The term biplot was introduced by Gabriel (1971) and consists of a graphical representation that reveals important characteristics of data structure, e.g., patterns of correlations between variables or similarities between the observations (Greenacre 2010). To achieve this, it uses the decomposition of a $(n \times m)$ target matrix $\mathbf{D}$ into the product of two matrices, such that $\mathbf{D} = \mathbf{GH'}$. The dimension of the $\mathbf{G}$ matrix is $(n \times k)$, and the size of $\mathbf{H}$ matrix is $(m \times k)$. Therefore, each element $\mathrm{d_{ij}}$ of the matrix $\mathbf{D}$ can be written as the scalar product of the *i-th* row of the left matrix $\mathbf{G}$ and the *j-th* column of the right matrix $\mathbf{H'}$, as follows:

$$\mathbf{D} = \mathbf{GH}' = \begin{pmatrix} \mathbf{g}_1' \\ \vdots \\ \mathbf{g}_n' \end{pmatrix} (\mathbf{h}_1 \quad \ldots \quad \mathbf{h}_m) = \begin{pmatrix} \mathbf{g}_1'\mathbf{h}_1 & \cdots & \mathbf{g}_1'\mathbf{h}_m \\ \vdots & \ddots & \vdots \\ \mathbf{g}_n'\mathbf{h}_1 & \cdots & \mathbf{g}_n'\mathbf{h}_m \end{pmatrix}.$$

The matrices $\mathbf{G}$ and $\mathbf{H}$ that arise from the decomposition of $\mathbf{D}$ create two sets of points. If these points are two-dimensional (i.e., $k = 2$), then the rows and columns of $\mathbf{D}$ can be represented employing a two-dimensional graph, with the n rows of $\mathbf{G}$ represented by points, and the m columns of $\mathbf{H'}$ reproduced in the form of vectors connected to the origin. In the graph, projecting $\mathbf{g}_\mathrm{i}'$ onto the axis determined by $\mathbf{h}_\mathrm{j}$ and then multiplying the norm of that projection by the norm of $\mathbf{h}_\mathrm{j}$, the result will be equivalent to the geometric definition of the scalar product, which can also be used to represent the element d_{ij} of the target matrix $\mathbf{D}$, that is:

$$\mathrm{d_{ij}} = \mathbf{g}_{\mathrm{i}}'\mathbf{h}_{\mathrm{j}} = \|\mathbf{g}_{\mathrm{i}}\|\|\mathbf{h}_{\mathrm{j}}\|\cos\theta,$$

where θ is the angle formed by the vectors $\mathbf{g}_{\mathrm{i}}$ and $\mathbf{h}_{\mathrm{j}}$. Furthermore, each set of coordinates formed by a row of **G** (i.e., $\mathbf{g}_{\mathrm{i}}'$) is represented as a *biplot point*, and each column of the transpose of **H** (i.e., $\mathbf{h}_{\mathrm{j}}$) is plotted as a *biplot vector*.

2.4 The Exploratory PLS Biplot

Given a rank r data matrix, the PLS allows another matrix to be obtained with rank s that is an approximation of the former, in which $s < r$. The PLS dataset is composed of two centered matrices **X** and **Y**, wherein the matrix of predictors **X** has the dimension ($n \times m$), and the matrix of responses **Y** has the size ($n \times q$). Representing a target matrix **D** as a juxtaposition of **X** and **Y**, then it will be ($n \times (m + q)$) and denoted as **D** = [**X Y**]. Considering that the number of PLS components extracted is lower than the rank of **X**, i.e., $k < r$, thus the matrix product **TP′** provides an approximation of **X**. Similarly, the matrix product **TQ′** gives an approximation for **Y**, instead of **UQ′** (Oyedele and Lubbe 2015). As a consequence, $\tilde{\mathbf{D}}$ provides an approximation for **D** such that

$$\tilde{\mathbf{D}} = \begin{bmatrix}\tilde{\mathbf{X}} & \tilde{\mathbf{Y}}\end{bmatrix} = [\mathbf{TP}' \quad \mathbf{TQ}'] = \mathbf{T}[\mathbf{P} \quad \mathbf{Q}]'.$$

Extracting just two components, the dimension of **T** is ($n \times 2$) and the size of the block matrix [**P Q**]′ is ($2 \times (m + q)$). So, the rows of **T** represent the biplot points in the exploratory PLS biplot, expressing the observations of the sample, while the columns of the block matrix [**P Q**]′ indicate the biplot vectors and denote the variables, wherein those from column 1 to m refer to the predictors and from column ($m + 1$) to ($m + q$) are associated with the responses. Considering each set of biplot vectors separately (predictors and responses), the angle formed by two vectors provides an approximation for the sample correlation coefficient related to the associated variables (Graffelman 2012). Therefore, if $\angle\left(\mathbf{p}_{\mathrm{i}}', \mathbf{p}_{\mathrm{j}}'\right) \cong 0^{\circ}$, it means that the associated variables are strongly correlated because the cosine of the angle between the biplot vectors is close to one. On the other hand, when $\angle\left(\mathbf{p}_{\mathrm{i}}', \mathbf{p}_{\mathrm{j}}'\right) \cong 180^{\circ}$ and the biplot vectors point to almost opposite directions, then it indicates a negative but substantial correlation. Lastly, a right angle suggests a weak correlation between the related variables. However, the accuracy of this approximation will depend on how much the variables contribute to each of the underlying components estimated (Bassani et al. 2010), as well as the biplot explained variance (Greenacre 2012).

2.5 The Predictive PLS Biplot

As previously seen in Sect. 2.2, the ($m \times q$) matrix $\widehat{\mathbf{B}}_{\mathrm{PLS}} = \mathbf{RQ}'$ contains the PLS coefficient estimates, in which the **R** columns are the transformed PLS X-weights, and **Q** is the matrix of Y-weights. In the predictive PLS biplot, the rows of the matrix **R** denote the biplot points instead of the rows of **T**. Further, the columns of **Q′**

symbolize the responses through biplot vectors. Each response can also define a calibrated axis, on which one can project the set of points ($\mathbf{r}_i'$) to get an approximation of the coefficients. Considering a specific response Y_j and a fixed predictor X_i, each element of the matrix $\widehat{\mathbf{B}}_{PLS}$ is computed as

$$\widehat{b}_{PLS_{ij}} = \mathbf{r}_i'\mathbf{q}_j = \|\mathbf{r}_i\|\|\mathbf{q}_j\|\cos\theta_{\mathbf{r}_i,\mathbf{q}_j}.$$

Therefore, there are two ways to evaluate an approximation for these estimates in the biplot visually. The first manner consists of the *calibration* of biplot axes (Greenacre 2010; Oyedele and Lubbe 2015) and mentally reading the projection of the biplot point on the biplot axis. In the second mode, the *area biplot* method is applied (Gower et al., 2010; Oyedele and Lubbe 2015), in which the approximation of $\widehat{b}_{PLS_{ij}}$ is obtained from the area determined by the origin, the rotated biplot point $\mathbf{r}_i'$, and the endpoint of $\mathbf{q}_j$. The area and position of the triangles furnish other relevant information about the PLS regression coefficients, such as the signal and the importance of each predictor to the model.

3 Application

Teams of individuals working together to achieve a common goal are a central part of daily life in modern organizations (Mathieu et al. 2014). By bringing together individuals with different skills and knowledge, teams emerge as a competitive asset in the ever-changing organizational environment. When teams are created, the ultimate goal is to generate value for the organization. Accordingly, studying team effectiveness and the conditions that enable the team to be effective has been a central concern for both research and practice (Kozlowski and Ilgen 2006).

3.1 Variables

In the present research, in line with previous studies (e.g., Hackman 1987), we consider team effectiveness as a multidimensional construct. Thus, in this study, team effectiveness is evaluated through two criteria: team performance and the quality of group experience. *Team performance* (Y_1) refers to the extent to which team outcomes respect the standards set by the organization, in terms of quantity, quality, delivery time and costs (Rousseau and Aubé 2010). The *quality of the group experience* (Y_2) is related to the quality of the social climate within the team (Aubé and Rousseau 2005).

Team effectiveness will be considered, in the present study, as the result of the role played by thirteen variables: team trust (2 dimensions), team psychological capital (4 dimensions), collective behavior, transformational leadership, intragroup conflict (2 dimensions), team psychological safety, and team cohesion (2 dimensions). Each variable will be briefly described as follows.

Team trust refers to the aggregate levels of trust that team members have in their fellow teammates (Langfred 2004) and has been conceptualized as a bidimensional construct: the *affective dimension of team trust* (X_1) is related to the perception of the presence of shared ideas, feelings, and concerns within the team; the *task dimension of team trust* (X_2) has been associated with the recognition by team members of the levels of professionalism and competence of their teammates and on their ability to appropriately perform the tasks (McAllister 1995).

Team psychological capital (PsyCap) can be defined as a team positive psychological state characterized by: having confidence (efficacy) to succeed in challenging tasks; making a positive attribution (optimism) about succeeding now and in the future; persevering, and when necessary, redirecting paths to goals (hope) in order to be effective; and having the ability to bounce back from challenges and setbacks (resilience) (Luthans et al. 2007; Luthans and Youssef-Morgan 2017; Walumbwa et al. 2011). In summary, team PsyCap includes four team psychological resources: *team efficacy* (X_3), *team optimism* (X_4), *team hope* (X_5), and *team resilience* (X_6).

Collective behavior (X_7) refers to the members' tendency to coordinate, evaluate, and utilize task inputs from other team members when performing a group task (Driskell et al. 2010).

Transformational leadership (X_8) can be defined as a leadership style that encourages followers to do more than they originally expected, broadening and changing their interests and leading to conscientiousness and acceptance of the team's purposes (Bass 1990). Carless et al. 2000) described transformational leaders as those who exhibit the following seven behaviors: they 1) communicate a vision; 2) develop staff; 3) provide support for them to work towards their objectives through coordinated team work; 4) empower staff; 5) are innovative by using non-conventional strategies to achieve their goals; 6) lead by example; 7) are charismatic.

Intragroup conflict can be defined as a disagreement that is perceived as creating tension at least by one of the parties involved in an interaction (De Dreu and Weingart 2003). Conflicts in teams may emerge as a result of the presence of different ideas about the tasks performed (X_9) – *task conflict* – or may be related to differences between team members in terms of values or personalities (X_{10}) – *affective conflict* (Jehn 1994).

Team psychological safety (X_{11}) relates to team members' perceptions about what the consequences will be of taking interpersonal risks at the work environment. It means taking beliefs for granted about how others will react when one speaks up or participates. It is a confidence climate that comes from mutual respect and trust between members (Edmondson 1999).

Team cohesion can be defined as the team members' inclination to create social bonds, resulting in the group sticking together, remaining united, and wanting to work together (Carron 1982; Salas et al. 2015). It can be related to the task or the affective system of the team. *Task cohesion* (X_{12}) refers to the shared commitment among members towards achieving a goal that requires the collective efforts of the group. *Social cohesion* (X_{13}) refers to shared liking or attraction to the group and to the nature and quality of the emotional bonds of friendship, liking, caring, and closeness among group members (Chang and Bordia 2001).

3.2 Sample and Data Collection Procedure

Organizations were selected by convenience, using the personal and professional contacts network of the research team. To collect data, key stakeholders in each organization (CEOs or human resources managers) were contacted to explain the purpose and requirements of the study. When the organization agreed to participate, the selection of teams for the survey was based on the following criteria (Cohen and Bailey 1997): teams must be composed of at least three members; should be perceived by themselves and others as a team; they have to regularly interact, interdependently, to accomplish a common goal; and they must have a formal supervisor who is responsible for the actions of the team.

Data was collected following two strategies. In most organizations, questionnaires were filled in during team meetings, in the presence of a member of the research team. When it was not possible to implement this strategy, they were filled in online via an electronic platform. Data was obtained from 104 teams and their respective leaders. After eliminating from the sample teams with a team members' response rate below 50% and participants with more than 10% of missing values, the remaining sample had a total of 82 teams. In this remaining sample, missing values in the questionnaires were replaced by the item average (in case of a random distribution) or by expectation-maximization (EM) method (in case of a non-random distribution).

The 82 teams of the sample are from 57 Portuguese companies. Forty-two per cent of these organizations are small, and the most representative sector is the services sector (73%). Team size ranged from 3 to 18 members, with an average of approximately 6 members (SD = 3.55). Of the team members (N = 353), 67% were female, 63.3% had secondary education or less, with the remaining 36.7% having a higher education background. The mean age was approximately 38 years old (SD = 12.33). The average team tenure was approximately 6 years (SD = 7.25). Regarding team leaders (N = 82), 57% were male, the mean age was about 42 years old (SD = 10.86) and 55.7% had a higher education background. Leaders had, on average, 5 years of experience as leader of the current team (SD = 4.87).

3.3 Measures

Apart from team performance that was assessed by team leaders, all variables were measured by team members. The measures used are identified as follows: *team performance* was measured with a scale developed by Rousseau and Aubé (2010), which has five items; *quality of the group experience* was assessed with the scale developed by Aubé and Rousseau (2005), which is composed of three items; team trust was evaluated with the scale developed by McAllister (1995), which is constituted by 10 items; team psyCap was measured with the scale developed by Luthans et al. (2007), which is composed of 24 items; collective behavior was measured with the scale developed by Driskell et al. (2010), which has 10 items; transformational leadership was measured with the scale developed by Carless et al. (2000), which is composed of seven items; intragroup conflict was evaluated with the scale developed by Dimas and Lourenço (2015), which is composed of nine items; team psychological safety was assessed with the scale developed by Edmonson (1999), which is composed of seven items; team cohesion was measured with the scale developed by Chang and Bordia (2001), which is constituted by eight items. Team trust and team psycap were assessed using 6-point scales, intragroup conflict and team psychological safety were evaluated on 7-point scales and the remaining variables were measured on 5-point scales.

3.4 PLS Biplot Results

In order to reveal a linear relation between the variables describing team effectiveness and the explanatory variables, the PLS was used to construct the external and internal models. First, the predictor matrix $\mathbf{X}_{82\times 13}$ and the response matrix $\mathbf{Y}_{82\times 4}$ were centered and scaled. Next, the NIPALS algorithm was used to decompose the data matrices and to extract two PLS components, yielding the matrices $\mathbf{T}_{82\times 2} = [\mathrm{T}_1\mathrm{T}_2]$, $\mathbf{P}_{13\times 2}$, $\mathbf{U}_{82\times 2}$, $\mathbf{Q}_{4\times 2}$, $\mathbf{W}_{13\times 2}$, $\mathbf{R}_{13\times 2}$, and $\mathbf{B}_{13\times 2}$. The latter contains the estimates of the PLS regression coefficients, according to Table 1.

The first PLS component T_1 explains 56.5% of the data variability, while the proportion of variance explained by T_2 is 9.5%. Figure 1 shows the exploratory PLS biplot, in which the (black) biplot points (X-scores $\mathbf{t}_i'$) represent the 82 teams, the blue biplot vectors depict the responses (Y-loadings $\mathbf{q}_i'$), and the red biplot vectors symbolize the predictors (X-loadings $\mathbf{p}_i'$).

Figure 1 provides an approximation of the correlation structures of the data, but it must be taken into account that the total proportion of variance explained by the two components T_1 and T_2 is 66%. Table 2 shows some significantly correlated variables evidenced by the biplot (X_2 and X_{12}, X_7 and X_{12}, X_3 and X_{11}, and X_9 and X_{10}), a pair of variables that displayed negative correlation (X_2 and X_9), and others that manifested a weak correlation visually (X_5 and X_{13}, X_6 and X_{13}), all of them flanked by the exact sample correlation coefficients. Table 2 is not exhaustive, and it is possible to pinpoint other exciting examples regarding the correlation structure in Fig. 1, e.g., the weak correlation between the two responses (the correct sample correlation is $\cong 0.28$). Moreover, all of the variables are positively associated with the first PLS component T_1, except *Task conflict* (X_9) and *Affective conflict* (X_{10}), which are negatively associated. Regarding T_2, the predictor *Team trust-affective* (X_1) is negligibly correlated, and the predictors *Team trust-task* (X_2), *Collective behavior* (X_7), *Task cohesion* (X_{12}), and *Social cohesion* (X_{13}) are negatively correlated.

For comparison purposes only, Fig. 2 shows the results of the area biplot method. With respect to response *Team performance* (Y_1) – left biplot, the predictors *Team efficacy* (X_3), *Team optimism* (X_4), and *Team psychological safety* (X_{11}) stand out as the most influential variables to the model, since the triangle related to the regression coefficients b_3, b_4, and b_{11} show the most significant area. On the other side, the variables with the least positive impact on the model are *Team trust-task* (X_2) and *Team task conflict* (X_9), because they are related to the smallest areas. Further, the predictor *Social cohesion* (X_{13}) affects *Team performance* negatively, given that the triangle position is on the right side of the biplot axis. In its turn, regarding the response *Quality of the group experience* (Y_2), the most important predictors are *Task conflict* (X_9), *Affective conflict* (X_{10}), *Team trust-task* (X_2), *Task cohesion* (X_{12}), and *Social cohesion* (X_{13}), with the first two in a negative way. All these findings are following the

Table 1. Punctual estimates of the PLS regression coefficients.

Predictor name	Predictor identification	$\widehat{\beta}_1$ related to Team performance ($\boldsymbol{Y}_1$)	$\widehat{\beta}_2$ related to Quality of the group experience ($\boldsymbol{Y}_2$)
Team trust (affective)	X_1	0.076	0.079
Team trust (task)	X_2	0.005	0.141
Team efficacy	X_3	0.105	0.033
Team optimism	X_4	0.144	−0.027
Team hope	X_5	0.076	0.068
Team resilience	X_6	0.023	0.075
Collective behavior	X_7	0.047	0.095
Transformational leadership	X_8	0.055	0.043
Task conflict	X_9	0.017	−0.124
Affective conflict	X_{10}	0.051	−0.154
Team psychological safety	X_{11}	0.090	0.053
Task cohesion	X_{12}	0.059	0.103
Social cohesion	X_{13}	−0.035	0.109

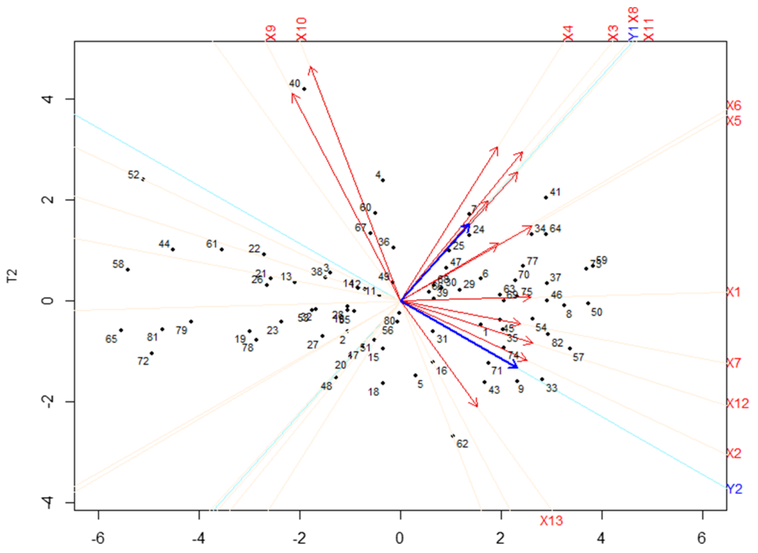

Fig. 1. Exploratory PLS biplot – sample and variables representation. (Color figure online)

PLS results shown in Table 2, but one can easily reach the same conclusions through Fig. 1 (exploratory PLS biplot).

Figure 3 brings a modified version of the exploratory PLS biplot, in which all of the biplot vectors are projected onto the calibrated biplot axis Y_1. One more time, the most significant vector projections refer to *Team efficacy* (X_3), *Team optimism* (X_4), and

Table 2. Correlation approximation by biplot vectors and sample correlation coefficients.

Variables	Correct Correlation Coefficient (r)
X_2 (Team trust - task) and X_{12} (Task cohesion)	0.71
X_7 (Team efficacy) and X_{12} (Task cohesion)	0.70
X_3 (Team efficacy) and X_{11} (Team psychological safety)	0.73
X_9 (Task conflict) and X_{10} (Affective conflict)	0.85
X_2 (Team trust - task) and X_9 (Task conflict)	−0.60
X_5 (Team hope) and X_{13} (Social cohesion)	0.35
X_6 (Team resilience) and X_{13} (Social cohesion)	0.25

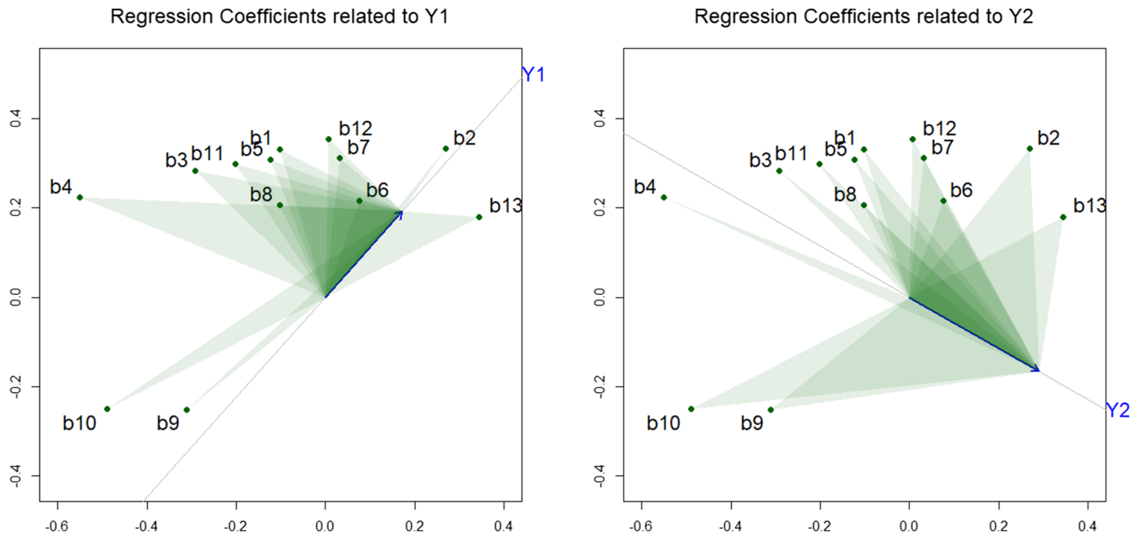

Fig. 2. Area biplot Method applied to the team effectiveness dataset.

Team psychological safety (X_{11}), as well as the less significant projection referring to *Team trust-task* (X_2). Beyond that, only the projection related to the variable *Social cohesion* (X_{13}) falls on the negative part of the biplot axis. The approximation of the regression coefficients related to the dependent variable *Quality of the group experience* is represented in Fig. 4, where the biplot vectors are projected onto the biplot axis Y_2. In this case, similarly to the results of the area biplot method, the largest projections indicate the more relevant variables. On the negative side, the predictors *Task conflict* (X_9) and *Affective conflict* (X_{10}) are the most influential in the model, while the explanatory variables *Team trust-task* (X_2), *Task cohesion* (X_{12}), and *Social cohesion* (X_{13}) have the most significant and positive impact concerning Y_2.

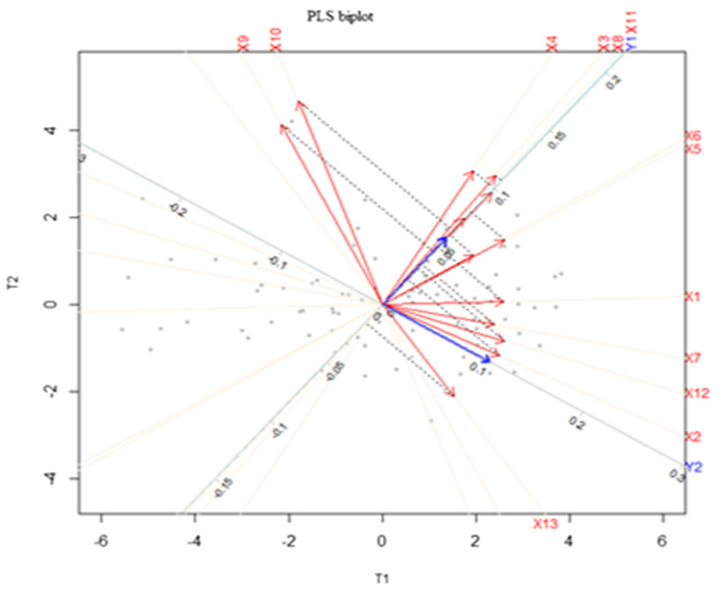

Fig. 3. Visual approximation of the regression coefficients (response Y_1).

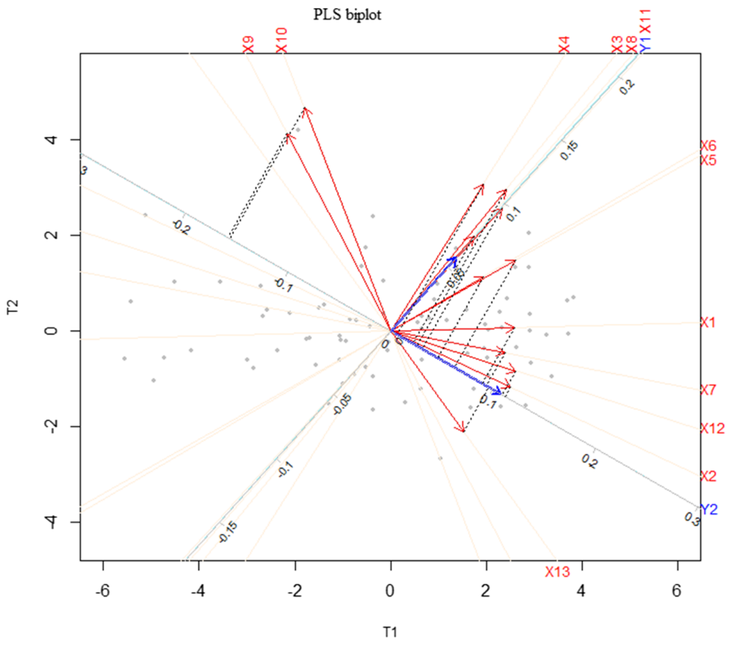

Fig. 4. Visual approximation of the regression coefficients (response Y_2).

4 Discussion and Conclusions

Regarding the application of the method we use in this work, the results point to the "validity" of such an application concerning the relationships found between the group processes and the team output variables considered. In fact, overall, the most significant results found in our study suggest relationships between the predictors and the criteria that are convergent with the literature.

One of the results points out the relevant role of cohesion as a predictor of team outcomes but the different behavior of each one of the team cohesion dimensions.

Indeed, task cohesion (X_{12}) showed a positive relationship with both team effectiveness criteria (although with higher magnitude regarding the quality of the group experience); however, social cohesion (X_{13}), although it emerged as one of the most relevant positive predictors of the quality of group experience, revealed a negative influence on team performance. These results are in line with the literature. Firstly, team cohesion is recognized by researchers as one of the most influential factors on group behavior and, consequently on group outcomes (Carron and Brawley 2000; Dionne et al. 2004). Secondly, and despite that, the literature, namely a meta-analysis conducted by Mullen and Cooper (1994), also suggests that the link between social cohesion and task cohesion with team outcomes can be different. Task cohesion tends to be positively associated with team outcomes, but social cohesion can have a more complex relationship with team outcomes due the fact that social cohesion, although it increases the willingness to help each other and to cooperate, can also lead to uncritical acceptance of solutions and to groupthink (Janis 1972). Thus, social cohesion can both increase the sense of belonging to a group, contributing to a positive perception of the group experience (quality of group experience), and decrease team performance, as suggested by our study.

Another interesting result to highlight is related to the negative relationship of both conflict types – task conflict (X_9) and affective conflict (X_{10}) – with the quality of group experience and the less clear role of task conflict in team performance. Indeed, task conflict revealed a negative relation with the quality of the group experience and a positive (albeit low-level) relation with team performance. These results tend to converge with the literature. On the one hand, the literature points out that conflict is always experienced as a negative experience (e.g., Jehn et al. 2008) and, as a result tends to have a negative influence on the attitudes of team members towards the group. However, on the other hand, studies are not totally consensual with respect to its effects on team performance, especially in what concerns to task conflict (De Wit et al. 2012; Dimas and Lourenço 2015). In fact, most studies found either negative associations between task conflict and team performance (e.g., Janssen et al. 1999) or a nonsignificant relation (e.g., Jordan and Troth 2004), and a meta-analysis conducted by De Dreu and Weingart (2003) supported those findings. However, more recently De Wit et al. (2012) conducted a new meta-analysis and concluded that the effects of task conflict on team outcomes are less negative (or even positive) as compared to affective conflict. Overall, the studies tend to suggest that, in certain circumstances, task conflict may be positively related to group outcomes (e.g., De Wit et al. 2012) emphasizing the role of moderators, such as the conflict-handling strategies used in the team.

It is also interesting to mention the positive role of team trust (X_1 and X_2) in team results and the more significant role of team task trust (X_2) compared to team social trust (X_1). Again, our results are supported by the literature which indicates that trust represents an important determinant of team effectiveness. In this regard, Dirks and Ferrin (2001) pointed out in their meta-analysis that team trust is positively related to performance and team satisfaction (an indicator of the quality of group experience). The fact that, in the present study, task trust has showed to be a more important predictor of performance than social trust can be explained by the fact that our sample is composed of work teams in productive organizations, where trust in the members' skills and their professionalism for the accomplishment of the tasks is more critical.

Finally, it is important to mention the role of team psychological safety (X_{11}), team self-efficacy (X_3) and team optimism (X_4) as positive predictors of team performance. Like the variables that we addressed above, the results obtained are supported by the literature. Regarding the relationship between team psychological safety and team performance, previous studies suggest that team performance can be facilitated, directly or indirectly, by the presence of a psychological security climate (e.g., Edmondson 1999). According to these studies, team performance is increased by a group climate in which team members are encouraged to express themselves without fear of the evaluations of the rest of the group. Regarding team efficacy and team optimism (dimensions of PsyCap), several studies show that collective PsyCap is positively related to team performance (e.g., Norman et al. 2010; Walumbwa et al. 2011). Additionally, previous research suggests that when team members have a collective belief in their ability to be effective, they explore and share knowledge and are more prepared to implement new ways of achieving results, because they believe these behaviors will lead to higher levels of performance (Bandura 1977). Similarly, a team with optimistic beliefs has positive expectations, is usually more actively involved in tasks than a team with a low level of optimism and use more adaptive coping skills when obstacles occur (Avey et al. 2011).

In general, the interpretation method proposed in this work provided excellent results in the application performed in Sect. 3, since it was able to capture the most critical variables for the model, correctly assigned the signals of the regression coefficients and gave an approximation to their values directly in the exploratory PLS biplot. Nevertheless, some inconsistencies were detected. For example, regarding the response Y_1, one can see in Table 2 that $\widehat{\beta}_{41} > \widehat{\beta}_{31}$, but the projections of the biplot vectors corresponding to X_4 and X_3 over the biplot axis yield the opposite result (Fig. 3). In the same sense, the projections of the vectors related to X_6 and X_9 seem to be overestimated considering the associated values ($\widehat{\beta}_{61}$ and $\widehat{\beta}_{91}$) in Table 2. Although with less intensity, the same occurs in Fig. 4, where the projections are made over the biplot axis Y_2.

However, we should keep in mind that biplot is a visualization method whose purpose is to provide a general idea of latent structures in the data, not to mention that the interpretation technique suggested in this paper provides only an approximation of the coefficients, which will be closer to the real values of the estimates, the higher the PLS components' ability to explain the variance.

Acknowledgments. This work was partially supported by The Center for Research and Development in Mathematics and Applications (CIDMA) and by The Center for Business and Economics Research (CeBER) through the Portuguese Foundation for Science and Technology (FCT - Fundação para a Ciência e a Tecnologia), references UIDB/04106/2020, UIDP/04106/2020 and UIDB/05037/2020.

References

Abdi, H.: Partial least squares regression and projection on latent structure regression (PLS regression). WIREs Comput. Stat. **2**, 97–106 (2010)

Avey, J.B., Reichard, R.J., Luthans, F., Mhatre, K.H.: Meta-analysis of the impact of positive psychological capital on employee attitudes, behaviors, and performance. Human Resource Dev. Quarterly **22**(2), 127–152 (2011)

Aubé, C., Rousseau, V.: Team goal commitment and team effectiveness: the role of task interdependence and supportive behaviors. Group Dynamics Theory Res. Practice **9**, 189–204 (2005)

Bandura, A.: Self-efficacy: Toward a unifying theory of behavioral change. Psychol. Rev. **84**(2), 191–215 (1977)

Bass, B.M.: From transactional to transformational leadership: Learning to share the vision. Org. Dyn. **18**(3), 19–31 (1990)

Bassani, N., Ambrogi, F., Coradini, D., Biganzoli, E.: Use of biplots and partial least squares regression in microarray data analysis for assessing association between genes involved in different biological pathways. In: Rizzo, R., Lisboa, Paulo J.G. (eds.) CIBB 2010. LNCS, vol. 6685, pp. 123–134. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-21946-7_10

Belsley, D.A., Kuh, E., Welsch, R.E.: Regression Diagnostics – Identifying Influential Data and Sources of Collinearity. Wiley-Interscience, New Jersey (2004)

Carless, S., Wearing, L., Mann, L.A.: Short measure of transformational leadership. J. Bus. Psychol. **14**(3), 389–405 (2000)

Carron, A.V.: Cohesiveness in Sport Groups: Interpretations and Considerations. J. Sport Psychol. **4**, 123–138 (1982)

Carron, A.V., Brawley, L.R.: Cohesion: Conceptual and measurement issues. Small Group Res. **31**, 89–106 (2000)

Chang, A., Bordia, P.: A multidimensional approach to the group cohesion group performance relationship. Small Group Res. **32**(4), 379–405 (2001)

Cohen, S.G., Bailey, D.E.: What makes teams work: Group effectiveness research from the shop floor to the executive suite. J. Manag. **23**(3), 239–290 (1997)

De Dreu, C.K.W., Weingart, L.R.: Task versus relationship conflict, team performance, and team member satisfaction: A meta-analysis. J. Appl. Psychol. **88**(4), 741–749 (2003)

De Wit, F.R., Greer, L.L., Jehn, K.A.: The paradox of intragroup conflict: A meta-analysis. J. Appl. Psychol. **97**(2), 360–390 (2012)

Dimas, I.D., Lourenço, P.R.: Intragroup conflict and conflict management approaches as determinants of team performance and satisfaction: Two field studies. Negot. Confl. Manage. Res. **8**(3), 174–193 (2015)

Dionne, S.D., Yammarino, F.J., Atwater, L.E., Spangler, W.D.: Transformational leadership and team performance. J. Organ. Change Manage. **17**(2), 177–193 (2004)

Dirks, K.T., Ferrin, D.L.: The role of trust in organizational settings. Organ. Sci. **12**(4), 450–467 (2001)

Driskell, J.E., Salas, E., Hughes, S.: Collective orientation and team performance: Development of an individual differences measure. Hum. Factors **52**(2), 316–328 (2010)

Edmondson, A.: Psychological safety and learning behavior in work teams. Adm. Sci. Q. **44**(2), 350–383 (1999)

Gabriel, K.R.: The biplot graphic display of matrices with application to principal component analysis. Biometrika **58**, 453–467 (1971)

Geladi, P., Kowalsky, B.R.: Partial least squares regression: A tutorial. Anal. Chim. Acta **186**, 1–17 (1986)

Gower, J.C., Groenen, P.J.F., Van de Velden, M.: Area biplots. J. Comput. Graphical Stat. **19**, 46–61 (2010)

Graffelman, J.: Linear-angle correlation plots: new graphs for revealing correlation structure. J. Comput. Graphical Stat. **22**(1), 92–106 (2012)

Greenacre, M.: *Biplots in Practice*. FBBVA, (2010)

Greenacre, M.: Contribution Biplots. J. Comput. Graphical Stat. **22**(1), 107–122 (2012)

Hackman, J.R.: The design of work teams. In: Lorsch, J. (ed.) Handbook of Organizational Behavior, pp. 315–342. Prentice-Hall, Englewood Cliffs (1987)

Janis, I.L.: Victims of groupthink: A psychological study of foreign-policy decisions and fiascoes. Houghton Mifflin (1972)

Janssen, O., Van de Vliert, E., Veenstra, C.: How task and person conflict shape the role of positive interdependence in management teams. J. Manag. **25**(2), 117–141 (1999)

Jehn, K.A.: Enhancing effectiveness: An investigation of advantages and disadvantages of value-based intragroup conflict. Int. J. Conflict Manage. **5**, 223–238 (1994)

Jehn, K.A., Greer, L., Levine, S., Szulanski, G.: The effects of conflict types, dimensions, and emergent states on group outcomes. Group Decis. Negot. **17**(6), 465–495 (2008)

Jordan, P.J., Troth, A.C.: Managing emotions during team problem solving: Emotional intelligence and conflict resolution. Hum. Perform. **17**(2), 195–218 (2004)

Kozlowski, S.W., Ilgen, D.R.: Enhancing the effectiveness of work groups and teams. Psychol. Sci. Public Interest **7**(3), 77–124 (2006)

Koch, I.: Analysis of Multivariate and High-Dimensional Data. Cambridge University Press, New York (2014)

Langfred, C.W.: Too much of a good thing? Negative effects of high trust and individual autonomy in self-managing teams. Acad. Manage. J. **47**(3), 385–399 (2004)

Luthans, F., Avolio, B.J., Avey, J.B., Norman, S.M.: Positive psychological capital: Measurement and relationship with performance and satisfaction. Pers. Psychol. **60**(3), 541–572 (2007)

Luthans, F., Youssef-Morgan, C.M.: Psychological capital: An evidence-based positive approach. Annu. Rev. Organ. Psychol. Organ. Behav. **4**(1), 339–366 (2017)

Mathieu, J.E., Gallagher, P.T., Domingo, M.A., Klock, E.A.: Embracing complexity: Reviewing the past decade of team effectiveness research. Annu. Rev. Organ. Psychol. Organ. Behav. **6**, 17–46 (2019)

Mathieu, J.E., Tannenbaum, S.I., Donsbach, J.S., Alliger, G.M.: A review and integration of team composition models: Moving toward a dynamic and temporal framework. J. Manage. **40** (1), 130–160 (2014)

McAllister, D.: Affect and cognition-based trust as foundations for interpersonal cooperation in organizations. Acad. Manage. J. **38**(1), 24–59 (1995)

Mullen, B., Cooper, C.: The relation between group cohesiveness and performance: An integration. Psychol. Bull. **115**, 210–227 (1994)

Norman, S.M., Avey, J.B., Nimnicht, J.L., Pigeon, N.: The interactive effects of psychological capital and organizational identity on employee organizational citizenship and deviance behaviors. J. Leadership Organ. Stud. **17**(4), 380–391 (2010)

Oyedele, O.F., Lubbe, S.: The construction of a partial least squares biplot. J. Applied Stat. **42** (11), 2449–2460 (2015)

Ringle, C.M., Sarstedt, M., Mitchell, R., Gudergan, S.P.: Partial least squares structural equation modeling in human resource management research. Int. J. Hum. Resour. Manage. **31**, 1617–1643 (2018)

Rousseau, V., Aubé, C.: Team self-managing behaviors and team effectiveness: The moderating effect of task routineness. Group Organ. Manage. **35**(6), 751–781 (2010)

Salas, E., Grossman, R., Hughes, A.M., Coultas, C.W.: Measuring team cohesion: Observations from the science. Hum. Factors **57**(3), 365–374 (2015)

Walumbwa, F.O., Luthans, F., Avey, J.B., Oke, A.: Authentically leading groups: The mediating role of collective psychological capital and trust. J. Organ. Behav. **32**(1), 4–24 (2011)

Wold, S., Eriksson, L., Trygg, J., Kettaneh, N.: The PLS Method - Partial Least Squares Projections to Latent Structures - and Its Applications in Industrial RDP (Research, Development, and Production). Umea University, Umea (2004)

Chapter 10

Software

R package *areabiplot*

Published:

Silva, A. & Freitas, A. areabiplot: Area Biplot R package version 1.0.0 (2021).

https://CRAN.R-project.org/package=areabiplot

10.1 R package *areabiplot*: documentation

Packages are essential to the R language, bringing together documentation, reusable functions, and data. They are community developed and easy to share with other users. The Comprehensive R Archive Network, or CRAN, is the public repository for R packages, consisting of web servers around the world that store identical and up-to-date versions of code and documentation for R. The *areabiplot* package is our contribution to this environment, implemented to meet multivariate generic purposes, and also for the specific intents of this investigation. After loading the package evoking in R the instruction $library(areabiplot)$, one can construct SSA-HJ-biplots executing the *areabiplot* function.

Package 'areabiplot'

March 10, 2021

Title Area Biplot

Version 1.0.0

Description Considering an (n x m) data matrix X, this package is based on the method proposed by Gower, Groener, and Velden (2010) <doi:10.1198/jcgs.2010.07134>, and utilize the resulting matrices from the extended version of the NIPALS decomposition to determine n triangles whose areas are used to visually estimate the elements of a specific column of X. After a 90-degree rotation of the sample points, the triangles are drawn regarding the following points: 1.the origin of the axes; 2.the sample points; 3. the vector endpoint representing some variable.

License MIT + file LICENSE

Encoding UTF-8

LazyData true

RoxygenNote 7.1.1

Suggests testthat

Imports grDevices, graphics, nipals

NeedsCompilation no

Author Alberto Silva [aut, cre] (<https://orcid.org/0000-0002-3496-6802>), Adelaide Freitas [aut] (<https://orcid.org/0000-0002-4685-1615>)

Maintainer Alberto Silva <albertos@ua.pt>

Repository CRAN

Date/Publication 2021-03-10 19:00:02 UTC

R topics documented:

areabiplot *Area Biplot*

Description

Consider an (n x m) centered data matrix X and let $rank(X) = r$. Alternatively to the ordinary NIPALS decomposition of X, where $X = TP'$, this package uses the resulting matrices from the extended version of the NIPALS decomposition ($X = GHP'$) to determine n triangles whose areas are used to visually estimate the n elements of a specific column of X (a variable of interest). After a 90-degree rotation of the sample points, the triangles are drawn regarding the following points:

1. the origin of the axes.
2. the sample points.
3. the vector endpoint representing the selected variable.

Just keep in mind that The extended NIPALS decomposition, $X = GHP'$, is equivalent to the SVD decomposition, $X = UDV'$, being that:

1. G is the matrix containing in its columns the normalized score vectors of X, i.e., the normalized columns of T. If t is the i-th score vector of the matrix T, then the i-th column of G will be $g = t/||t||$, which will correspond to the i-th left singular vector u.
2. If t is the i-th column of T, then $||t|| = \sqrt(t't)$ gives the i-th singular value of X. In addition, H is the diagonal matrix containing these singular values in decreasing order, i.e., $H = D$.
3. P is the loadings matrix, which is equivalent to the V matrix that contains the right singular vectors of X.

Usage

```
areabiplot(
  L,
  S,
  R,
  ord.row,
  mode = NULL,
  tri.rgb = NULL,
  bg.col = NULL,
  plot.title = NULL,
  plot.title.col = NULL,
  plot.title.font = NULL,
  plot.title.cex = NULL,
  plot.sub = NULL,
  plot.sub.col = NULL,
  plot.sub.font = NULL,
  plot.sub.cex = NULL,
  plot.cex = NULL,
  plot.col = NULL,
  plot.pch = NULL,
  plot.xlab = NULL,
  plot.ylab = NULL,
  plot.xlim = NULL,
  plot.ylim = NULL,
  points.lab = NULL,
  var.lab = NULL,
  text.col.var = NULL,
  text.cex = NULL,
  text.font = NULL,
  text.col = NULL,
  text.pos = NULL,
  axis.col = NULL,
  axis.cex = NULL,
  axis.font = NULL,
  axis.asp = NULL,
  arrow.lwd = NULL,
  arrow.len = NULL,
  arrow.col = NULL
)
```

Arguments

`L` A (n x 2) matrix containing normalized score vectors g (or left singular vectors u).

`S` An appropriate (2 x 2) diagonal matrix containing the corresponding singular values in decreasing order.

`R` A (m x 2) matrix containing the corresponding loading vectors (or right singular vectors).

`ord.row` The row of R used as the base of the triangle, e.g., if 1 is provided, then the first row of R will be taken.

`mode` a string providing the way the singular values will be allocated. The default is "SS", i.e., the similar spread proposed by Gower et al.. Alternatively, one can choose the "HJ" method (see more in Details).

`tri.rgb` The hexadecimal color and alpha transparency code for the triangle. The default is #19FF811A (green and 90% of transparency).

`bg.col` A string providing the color of the background. The default is #001F3D (blue).

`plot.title` A string providing the main title. The default is NONE.

`plot.title.col` A string specifying the color of the main title text. The default is "FFFFFF" (white).

`plot.title.font`
An integer providing the style of the main title text. The default is 1 (normal text).

`plot.title.cex` A number indicating the amount by which the main title text should be scaled relative to the default. 1 = default, 1.5 is 50% larger, and so on.

`plot.sub` A string providing a sub-title. The default is NONE.

`plot.sub.col` A string specifying the color of the sub-title text. The default is "FFFFFF" (white).

`plot.sub.font` An integer providing the style of the main title text. The default is 1 (normal text).

`plot.sub.cex` A number indicating the amount by which the sun-title text should be scaled relative to the default. 1 = default, 1.5 is 50% larger, and so on.

`plot.cex` A number indicating the expansion or contraction factor used to specify the point size. The default is 0.6 (40% smaller).

`plot.col` A string specifying the color of the points. The default is "FFFFFF" (white).

`plot.pch` An integer specifying the shape of the points. The default is 21 (circle)

`plot.xlab` A string specifying a label to the horizontal axis. The default is NONE.

`plot.ylab` A string specifying a label to the vertical axis. The default is NONE.

`plot.xlim` The limits for the x axis.

`plot.ylim` The limits for the y axis.

`points.lab` A vector of characters containing the names of the data matrix rows.

`var.lab` A string providing the variable name used as triangle base.

`text.col.var` A string specifying the color of the variable label text. The default is "FFFFFF" (white).

`text.cex` A number indicating the expansion or contraction factor used to specify the point labels. The default is 0.5.

`text.font` An integer providing the style of the point labels. The default is 2 (bold).

`text.col` A string specifying the color of the point labels text. The default is "FFFFFF" (white).

`text.pos` An integer providing the position of the point labels. The default is 3 (above).

`axis.col` A string specifying the color of the axis. The default is #FFFFFF (white).

`axis.cex` A number indicating the expansion or contraction factor used to specify the tick label. The default is 0.7

`axis.font` An integer providing the style of the tick label. The default is 1 (normal text).

`axis.asp` A number specifying the aspect ratio of the axes. The default is 1.

`arrow.lwd` A number specifying the line width of the arrow. The default is 1.

`arrow.len` The length of the edges of the arrow head (in inches). The default is 0.1.

`arrow.col` A string specifying the color of the arrow. The default is "FFFFFF" (white).

Details

1. If the variables (the columns of X) are measured in different units or their variability differs considerably, one could perform a variance scaling to get better visual results on the graph (see Examples). In this case, the percentage of variance explained by the first principal components might decrease.
2. The "HJ" mode is reserved for an application under implementation.

Value

An area biplot is produced on the current graphics device.

Author(s)

Alberto Silva albertos@ua.pt, Adelaide Freitas adelaide@ua.pt

References

J.C. Gower, P.J.F. Groenen, M. van de Velden (2010). Area Biplots. Journal of Computational and Graphical Statistics, v.19 (1), pp. 46-61. doi: 10.1198/jcgs.2010.07134

Examples

```
library(nipals)
data(uscrime)
Y = uscrime[, -1]

# first case: scale is false
nip = nipals(Y, ncomp = 2, center = TRUE, scale = FALSE, force.na = TRUE)
L = nip$scores
R = nip$loadings
S = diag(nip$eig[1:2])
areabiplot(L, S, R, 5, points.lab = c(uscrime[, 1]),var.lab= "burglary")

# second case: scale is true
nip = nipals(Y, ncomp = 2, center = TRUE, scale = TRUE, force.na = TRUE)
L = nip$scores
R = nip$loadings
S = diag(nip$eig[1:2])
areabiplot(L, S, R, 4, points.lab = c(uscrime[, 1]),var.lab= "assault")
```

Index

10.2 Code

```
#' Area Biplot
#'
#' @description
#' Consider a (n x m) centered data matrix \eqn{X} and let \eqn{rank(X) = r}.
#' Alternatively to the ordinary NIPALS decomposition of \eqn{X}, where
#' \eqn{X = T P'}, this package uses the resulting matrices from the extended
#' version of the NIPALS decomposition (\eqn{X = G H P'}) to determine \eqn{n}
#' triangles whose areas are used to visually estimate the \eqn{n} elements of
#' a specific column of \eqn{X} (a variable of interest). After a 90-degree
#' rotation of the sample points, the triangles are drawn regarding the
#' following points:

#'  1. the origin of the axes.
#'  2. the sample points.
#'  3. the vector endpoint representing the selected variable.

#' @description
#' Just keep in mind that the extended NIPALS decomposition, \eqn{X = G H P'}, is
#' equivalent to the SVD decomposition, \eqn{X = U D V'}, being that:
#' 1. \eqn{G} is the matrix containing in its columns the normalized score vectors
#' of \eqn{X}, i.e., the normalized columns of \eqn{T}. If \eqn{t} is the i-th score
#' vector of the matrix \eqn{T}, then the i-th column of  \eqn{G} will be
#' \eqn{g = t / || t || }, which will correspond to the i-th left singular vector \eqn{u}.
#' 2. If \eqn{t} is the i-th column of \eqn{T}, then \eqn{|| t || = \sqrt(t' t)} gives
#' the i-th singular value of \eqn{X}. In addition, \eqn{H} is the diagonal matrix
#' containing these singular values in decreasing order, i.e., \eqn{H = D}.
#' 3. \eqn{P} is the loadings matrix, which is equivalent to the \eqn{V} matrix that
#' contains the right singular vectors of \eqn{X}.
#'
#'
#' @param L                 A (n x 2) matrix containing normalized score vectors \eqn{g} (or left singular
#'                          vectors \eqn{u}).
#' @param S                 An appropriate (2 x 2) diagonal matrix containing the corresponding singular
#'                          values in decreasing order.
#' @param R                 A (m x 2) matrix containing the corresponding loading vectors (or right singular
#'                          vectors).
#' @param ord.row           The row  of \eqn{R} used as the base of the  triangle, e.g.,
#'                          if 1 is provided, then the first row of \eqn{R} will be taken.
#' @param mode              A string providing the way the singular values will be allocated. The default
#'                          is "SS", i.e., the similar spread proposed by Gower et al.. Alternatively,
#'                          one can choose the "HJ" method (see more in  Details).
#' @param tri.rgb           The hexadecimal color and alpha transparency code for the triangle. The
#'                          default is #19FF811A (green and 90% of transparency).
#' @param bg.col            A string providing the color of the background. The default is #001F3D (blue).
#' @param plot.title        A string providing the main title. The default is NONE.
#' @param plot.title.col    A string specifying the color of the main title text. The default is "FFFFFF"
#'                          (white).
#' @param plot.title.font An integer providing the style of the main title text. The default is 1 (normal
#'                          text).
#' @param plot.title.cex    A number indicating the amount by which the  main title text should be scaled
#'                          relative to the default. 1 = default, 1.5 is 50% larger, and so on.
#' @param plot.sub          A string providing a sub-title. The default  is NONE.
#' @param plot.sub.col      A string specifying the color of the sub-title text. The default is "FFFFFF"
#'                          (white).
#' @param plot.sub.font   An integer providing the style of the main title text. The default is 1 (normal
#'                        text).
#' @param plot.sub.cex    A number indicating the amount by which the sun-title text should be scaled
#'                        relative to the default. 1 = default, 1.5 is 50% larger, and so on.
#' @param plot.cex        A number indicating the expansion or contraction factor used to specify the
#'                        point size. The default is 0.6 (40% smaller).
#' @param plot.col        A string specifying the color of the points. The default is "FFFFFF" (white).
#' @param plot.pch        An integer specifying the shape of the points. The default is 21 (circle)
#' @param plot.xlab       A string specifying a label to the horizontal axis. The default is NONE.
#' @param plot.ylab       A string specifying a label to the vertical axis. The default is NONE.
#' @param plot.xlim       The limits for the x axis.
#' @param plot.ylim       The limits for the y axis.
#' @param axis.col        A string specifying the color of the axis. The default is #FFFFFF (white).
#' @param axis.cex        A number indicating the expansion or contraction factor used to specify
#'                        the tick label. The default is 0.7
#' @param axis.font       An integer providing the style of the tick label. The default is 1 (normal
#'                        text).
#' @param axis.asp        A number specifying the aspect ratio of the axes. The default is 1.
#' @param points.lab      A vector of characters containing the names of the data matrix rows.
#' @param var.lab         A string providing the variable name used as triangle base.
#' @param text.col.var    A string specifying the color of the variable label text. The default is
#'                        "FFFFFF" (white).
#' @param text.cex        A number indicating the expansion or contraction factor used to specify
#'                        the point labels. The default is 0.5.
#' @param text.font       An integer providing the style of the point labels. The default is 2 (bold).
#' @param text.col        A string specifying the color of the point labels text. The default is
#'                         "FFFFFF" (white).
#' @param text.pos         An integer providing the position of the point labels. The default is 3
#'                         (above).
#' @param arrow.lwd        A number specifying the line width of the arrow. The default is 1.
#' @param arrow.len        The length of the edges of the arrow head (in inches). The default is 0.1.
#' @param arrow.col        A string specifying the color of the arrow. The default is "FFFFFF"
#'                         (white).
#'
#'
#' @return An area biplot is produced on the current graphics device.
#'
#'
#' @author
#' Alberto Silva <albertos@ua.pt>, Adelaide Freitas <adelaide@ua.pt>
#'
#'
#' @references
#' J.C. Gower, P.J.F. Groenen, M. van de Velden (2010). Area Biplots. Journal of Computational and
#' Graphical Statistics, v.19 (1), pp. 46-61. \doi{10.1198/jcgs.2010.07134}
#'
#'
#' @examples
#' library(nipals)
#' data(uscrime)
#' Y = uscrime[, -1]
#'
#' # first case: scale is false
#' nip = nipals(Y, ncomp = 2, center = TRUE, scale = FALSE, force.na = TRUE)
#' L = nip["scores"][[1]]
#' R = nip["loadings"][[1]]
#' S = diag(nip["eig"][[1]][1:2])
#' areabiplot(L, S, R, 5, points.lab = c(uscrime[, 1]),var.lab= "burglary")
#'
#' # second case: scale is true
#' nip = nipals(Y, ncomp = 2, center = TRUE, scale = TRUE, force.na = TRUE)
#' L = nip["scores"][[1]]
#' R = nip["loadings"][[1]]
#' S = diag(nip["eig"][[1]][1:2])
#' areabiplot(L, S, R, 4, points.lab = c(uscrime[, 1]),var.lab= "assault")
#'
#'
#' @details
#' 1. If the variables (the columns of X) are measured in different units or
#' their variability differs considerably, one could perform a variance scaling
#' to get better visual results on the graph (see Examples). In this case, the
#' percentage of variance explained by the first principal components might decrease.
#' 2. The "HJ" mode is reserved for an application under implementation.
#'
#'
#' @export
#' @importFrom            grDevices rgb
#' @importFrom            graphics arrows par polygon axis text
#' @import                nipals
#'
#'


## area biplot function


areabiplot <-
  function(L, S, R, ord.row, mode = NULL, tri.rgb = NULL, bg.col = NULL, plot.title = NULL,
  plot.title.col = NULL, plot.title.font = NULL, plot.title.cex = NULL, plot.sub = NULL,
  plot.sub.col = NULL, plot.sub.font = NULL, plot.sub.cex = NULL, plot.cex = NULL,
  plot.col = NULL, plot.pch = NULL, plot.xlab = NULL, plot.ylab = NULL, plot.xlim = NULL,
  plot.ylim = NULL, points.lab = NULL, var.lab = NULL, text.col.var = NULL, text.cex = NULL,
  text.font = NULL, text.col = NULL, text.pos = NULL, axis.col = NULL, axis.cex = NULL,
  axis.font = NULL, axis.asp = NULL, arrow.lwd = NULL, arrow.len = NULL, arrow.col = NULL)
```

```
 {

## dimensions

  n <- nrow(L)
  m <- nrow(R)

## scaling mode

   if (is.null( mode )) mode <- "SS"
   if ( mode == "HJ" ) {

   ## HJ scaling mode

      A <- L %*% S
      B <- R %*% S

   } else if ( mode == "SS" ) {

   ## similar spread mode

      q    <-  (n / m)^(1 / 4)
      sig <-  S^(1 / 2)
      A    <-  q * L %*% sig
      B    <-  (1 / q) * R %*% sig

   } else { warning ( "the scaling mode provided was not recognized. Try 'SS' or 'HJ'!" )}

   ## rotate sample points

      rotate  <- matrix(c(0, -1, 1, 0), 2)
      M <- A %*% rotate
      b <- c(B[ord.row, c(1, 2)])

   ## other default parameters

   if (is.null( tri.rgb ))          tri.rgb         <- "#19FF811A"
   if (is.null( bg.col ))           bg.col          <- "#001F3D"
   if (is.null( plot.title.col ))   plot.title.col  <- "#FFFFFF"
   if (is.null( plot.title.font ))  plot.title.font <- 1
   if (is.null( plot.title.cex ))   plot.title.cex  <- 1
   if (is.null( plot.sub.col ))     plot.sub.col    <- "#FFFFFF"
   if (is.null( plot.sub.font ))    plot.sub.font   <- 1
   if (is.null( plot.sub.cex ))     plot.sub.cex    <- 0.8
```

```r
    if (is.null( plot.cex ))         plot.cex        <- 0.6
    if (is.null( plot.col ))         plot.col        <- "#FFFFFF"
    if (is.null( plot.pch ))         plot.pch        <- 21
    if (is.null( plot.xlab ))        plot.xlab       <- ""
    if (is.null( plot.ylab ))        plot.ylab       <- ""
    if (is.null( plot.xlim ))        plot.xlim       <- c(min( min( M[, 1]), b[1]),  max( max( M[, 1] ), b[1] ))
    if (is.null( plot.ylim ))        plot.ylim       <- c(min( min( M[, 2]), b[2]), max( max( M[, 2] ), b[2] ))
    if (is.null( text.cex ))         text.cex        <- 0.5
    if (is.null( text.font ))        text.font       <- 2
    if (is.null( text.col ))         text.col        <- "#FFFFFF"
    if (is.null( text.pos ))         text.pos        <- 3
    if (is.null( axis.col ))         axis.col        <- "#FFFFFF"
    if (is.null( axis.cex ))         axis.cex        <- 0.7
    if (is.null( axis.font ))        axis.font       <- 1
    if (is.null( axis.asp ))         axis.asp        <- 1
    if (is.null( arrow.lwd ))        arrow.lwd       <- 1
    if (is.null( arrow.len ))        arrow.len       <- 0.1
    if (is.null( arrow.col ))        arrow.col       <- "#FFFFFF"
    if (is.null( text.col.var ))     text.col.var    <- "#FF0000"

## graphical parameter

    temppar <- par(bg = bg.col)
    on.exit(par(temppar), add = TRUE)

## plot sample points

    plot(M[, 1], M[, 2], pch = plot.pch, cex = plot.cex, col  = plot.col, xlim = plot.xlim,
    ylim = plot.ylim, main = plot.title, col.main = plot.title.col, font.main = plot.title.font,
    cex.main = plot.title.cex, sub = plot.sub, col.sub = plot.sub.col, font.sub = plot.sub.font,
    cex.sub = plot.sub.cex, axes = FALSE, xlab = plot.xlab, ylab = plot.ylab, asp = axis.asp)
    text(M, labels = points.lab, cex = text.cex, font = text.font, col = text.col, pos = text.pos)
    text(x = b[1], y = b[2], labels = var.lab, cex = text.cex, font = text.font, col = text.col.var, pos = text.pos)
    axis(1, col = axis.col, col.axis = axis.col, col.ticks = axis.col, cex.axis = axis.cex, font = axis.font)
    axis(2, col = axis.col, col.axis = axis.col, col.ticks = axis.col, cex.axis = axis.cex, font = axis.font)
```

```

## draw the triangles

   for (j in 1:n) {
      v1 = c(0, M[j, 1], b[1])
      v2 = c(0, M[j, 2], b[2])
      polygon(x = v1, y = v2, col = tri.rgb, border = NA)
    }


## draw the base of the triangle

   arrows(0, 0, x1 = b[1], y1 = b[2], lwd = arrow.lwd, length = arrow
   .len, col = arrow.col)

}
```

Part III

Discussion and Conclusions

Chapter 11

Discussion

The visualization methods and alternative multivariate approaches presented throughout this investigation focused on creating additional tools for decomposing a TS through the SSA. The proposed biplots allow a deeper visual inspection of the trajectory matrix eigenstructure. This chapter is dedicated to examining the promising points verified along the research, but in an integrated and articulated way according to the sequence of articles contained in Part II. However, it does not neglect to present and emphasize the weaknesses of the suggested methods, drawing attention to possible mitigating solutions.

The discussion begins with consolidating the SSA-HJ-biplot interpretation rules in light of aspects not yet addressed, namely, the use of different window lengths (ℓ) and the issue of centering the trajectory matrix. The objective is to comprehensively understand the trajectory matrix structure associated with the TS components. The reason is because this perception is crucial to both evidencing the components separation (Article I and Article II) and to recognize the period through geometric patterns (Article IV). Next, we reveal a setback in developing a new method for detecting change points (Article III) using the NIPALS algorithm to decompose the trajectory matrix. The aim is to examine the pros and cons of using NIPALS and determine whether one can strike a balance between speed and instability. Finally, a summary of the development and small solutions adopted in implementing the R package *areabiplot* (Software) is presented.

11.1 Consolidation of SSA-HJ-biplots interpretability

In the broader context, the properties of row and column markers in an HJ-biplot are the same as in JK-biplot and GH-biplot, meaning that [77]:

- The smaller the distance between row markers, the greater the similarity between the respective individuals.
- The column vectors' lengths can approximate the variables' standard deviations.
- The correlations of the variables are approximated by the cosines of the angles of the column vectors. The more positively correlated the variables, the closer the angle between the arrows will be to 0°. Likewise, given two strongly negatively correlated variables, the angle between the respective arrows appears close to 180°. Also, the closer to 90° the angle between two or more arrows is, the less correlated the corresponding variables will be. Finally, correlations between column markers and PCs are also approximated by the cosines of the angles formed by them.

As stated before, the SSA-HJ-biplot [17] starts with a univariate real-valued TS $Y = (y_0, \cdots, y_{n-1})$. Given the trajectory matrix $\mathbf{X} = [x_{ij}]_{i,j=1}^{\ell,\kappa}$, and where $x_{ij} = y_{i+j-2}$, we set the window length $\ell = n/2$. Consequently, the κ-lagged vector $(y_{i-1}, \cdots, y_{i+\kappa-2})$ at the i^{th} row and the ℓ-lagged vector $(y_{j-1}, \cdots, y_{j+\ell-2})$ at the j^{th} column represent almost the same subseries for some i and j. As shown in Article II, depending on the characteristics of Y, one can center the trajectory matrix on the columns for better visualization. The graphical interpretation is performed based on the row markers $\mathbf{j}_i^{'}$, $i = 1, \ldots, \ell$, and the column markers $\mathbf{h}_q^{'}$, $q = 1, \ldots, \kappa$. Each $\mathbf{j}_i^{'}$ is displayed as a biplot point and corresponds to a κ-lagged vector, while each $\mathbf{h}_q^{'}$ is depicted as an arrow and relates to an ℓ-lagged vector.

Briefly, Article I handles the SSA-HJ-biplot interpretation in the following terms:

(R1) The proximity of points: *Short Euclidean distances indicate similarity in the behavior of the associated κ-lagged vectors.*

$$\forall i, \exists \pi \in \mathbb{N} : ||\mathbf{j}_i^{'} - \mathbf{j}_{i+\pi}^{'}|| \approx 0 \Rightarrow ||(y_i, \cdots, y_{i+\kappa-1}) - (y_{i+\pi}, \cdots, y_{i+\pi+\kappa-1})|| \approx 0;$$

(R2) The arrows length: *Biplot vectors having roughly the same length indicate the corresponding ℓ-lagged vectors have a standard deviation also close.*

$$\exists \tau \in \mathbb{N}, \tau < \kappa : ||\mathbf{h}_\tau^{'}|| \approx ||\mathbf{h}_{\tau+q}^{'}|| \Rightarrow var(\mathbf{x}_\tau) \approx var(\mathbf{x}_{\tau+q}), \quad \forall q = 1, \cdots, \kappa - \tau;$$

(R3) The angle between arrows: *The cosine of the angle formed by arrows approximates the autocorrelation between the corresponding ℓ-lagged vectors.*

$$\exists \eta \in \mathbb{N}, \eta < \kappa : |\cos\left(\angle(\mathbf{h}_q^{'}, \mathbf{h}_{q+\eta}^{'})\right)| \approx 1 \Rightarrow |cor(\mathbf{x}_q, \mathbf{x}_{q+\eta})| \approx 1 \quad \forall q = 1, \cdots, \kappa - \eta.$$

(R4) The trend and directions of PCs: *If there is a trend, then the singular value associated with the 1^{st} PC will be dominant.*

(R5) The seasonality and singular values: *If $\exists i \in \mathbb{N} : \sqrt{\mathbf{t}_i' \mathbf{t}_i} \approx \sqrt{\mathbf{t}_{i+1}' \mathbf{t}_{i+1}}$, then the associated PCs are informative about the oscillatory components as long as they explain* *a substantial proportion of variability.*

In turn, Article II brought more contributions to analyzing the elements in the SSA-HJ-biplot. Hence, given an additive model represented by $Y = T + S + R$, where T is the trend-cycle, S is the seasonal component, and R is the remainder component, the interpretation provided by that study is summarized below:

(R6) The singular vectors and the sine/cosine shape: *If the 1^{st} PC represents the trend, one should consider identifying in the next pair with similar singular values which one refers to the sine and the other one refers to the cosine, taking into account this distinction in building the SSA-HJ-biplots.*

(R7) The moving average of order η (η-MA) as a reference: *Choosing η properly, one can approximate the trend by the η-MA. Thus, the position of the points in the SSA-HJ-biplot concerning the 1^{st} PC axis allows us to assess the location of the observations regarding the trend in the TS.*

(R8) The biplot points and sine/cosine directions: *Biplots constructed from the 1^{st} PC, and each of the directions determined by sine and cosine will better evidence the time units that summarize the trend behavior and those that determine the peaks/valleys in the TS separately.*

(R9) The angle θ between an arrow and a factor axis: *The greater the $|cos(\theta)|$, the more the variability of the column vector associated with $\mathbf{h}'_q$ is affected by the corresponding i^{th} PC.*

(R10) The projection of a point onto an arrow: *It provides the level of agreement between the κ-lagged vector that determines the point and the ℓ-lagged vector that induces the arrow.*

As for the HJ-biplot elaborated in Article IV, it focuses on identifying the periodicity of a TS through the formation of geometrical patterns. Let $\mathbf{h}'_f$ be the fixed arrow that will serve as the basis for all triangles, and $\mathbf{h}^{[r]}_q$, $q \in \{1, \ldots, \kappa\} \setminus \{f\}$, the corresponding vectors representing the arrows counterclockwise rotated by 90°. Then, the possible insights that can emerge from the SSA area biplots are listed below:

(R11) The autocorrelation and right-angled triangles: *Considering p is the dominant period of the TS, the most expressive autocorrelations between the $\mathbf{h}'_f$ and some $\mathbf{h}^{[r]}_q$ are indicated by right triangles, i.e.,*

$$\exists \eta \in \mathbb{Z} : \measuredangle(\mathbf{h}'_f, \mathbf{h}^{[r]}_{\eta+\mathrm{kp}}) \approx 90^\circ, \ \ \mathrm{k} \in \{0, 1, \ldots, \lfloor(\kappa - \eta)/p\rfloor\}.$$

(R12) Weak autocorrelation and the triangles' area: *Considering p is the dominant period of the TS, almost zero-area triangles indicate the weaker autocorrelations between the $\mathbf{h}'_f$ and some $\mathbf{h}^{[r]}_q$, i.e.,*

$$\exists \tau \in \mathbb{Z} : ||\mathbf{h}'_f - \mathbf{h}^{[r]}_{\tau+\mathrm{kp}}|| \approx 0, \ \ \mathrm{k} \in \{0, 1, \ldots, \lfloor(\kappa - \tau)/p\rfloor\}.$$

(R13) Cohesive groups of triangles: *Within intermediate groups of almost similar triangles, the associated ℓ-lagged vectors and also the correspondent subseries are strongly* *correlated with each other.*

(R14) The triangles' position and the autocorrelation signal: *A group of almost similar triangles to the left of the base vector indicate that the correlation between the ℓ-lagged vector corresponding to each $\mathbf{h}_q^{[r]}$ and the ℓ-lagged vector associated with $\mathbf{h}_f^{'}$ will be positive. When on the right, the autocorrelation will be negative. In both situations, the same reasoning applies to the corresponding subseries.*

(R15) The TS periodicity: *The periodicity of the TS is estimated by the number of groups of almost similar triangles.*

Given the exploratory nature of the SSA-HJ-biplot, ideally, the graphical results should be interpreted together better to understand the underlying phenomenon and characteristics of the TS. The overlaps related to some interpretation rules and method properties bring consistency to the approach, providing the interpreter with different points of view regarding the details of the TS structure. Property R1 in Article I and the rules R7 and R8 in Article II are somehow related. The similarity in the behavior of the κ-lagged vectors evidenced by the proximity of the biplot points will also reflect in the determination of the time units that best show the peaks, valleys, and the trend. In turn, criterion R3 in Article I and the properties from R11 to R13 in Article IV should lead to the same understanding, as they are just particular ways of describing the same information in two different ways about two different biplots. The properties and rules of interpretation from R1 to R15 corroborate the solution to the problems previously raised in questions P3 and P4.

11.1.1 Window length choices

The strategy of adjusting the window length in $\ell/2$ and employing the HJ-biplot scheme proved efficient in decomposing and visualizing the characteristics of the series studied through biplots. However, one may wish to privilege the representation of the points, making the window length ℓ narrower, that is, making $\ell << \kappa$. Ideally, for a known period p, each row of the trajectory matrix will capture the entire behavior of the TS, being represented in the SSA-HJ-biplot as p points. Notice that one is giving up on representing the trend in this case. For example, compare the SSA-HJ-biplot of the monthly TS CO_2 in Figure 11.1, in which $\ell = 12$, with the respective biplots at Article I and Article II.

On the other hand, to give more importance to the representation of the arrows, one can set the window size so that $\kappa << \ell$. Again, if there is a good guess for p, taking $\kappa = p$ is equivalent to reproducing the κ columns of the trajectory matrix as p arrows in the SSA-HJ-biplot but better represented in terms of interpretability. Figure 11.2 shows the SSA-HJ-biplot of the monthly TS CO_2 in which $\kappa = 12$. The corresponding

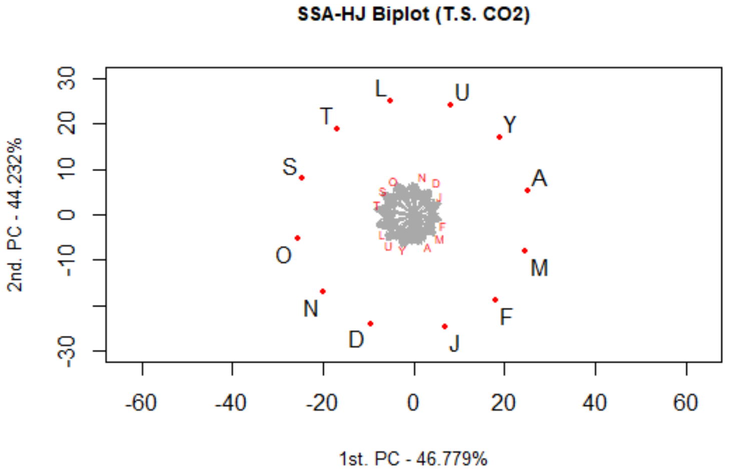

Figure 11.1: Representation of the TS CO2 SSA-HJ-biplot (1^{st} and 2^{nd} PCs) when the window length $\ell = 12$ is narrower than κ.

biplots at Article I and Article II do not illustrate the arrows so well in comparison with the one in Figure 11.2, despite the latter having lost the quality of representing the points.

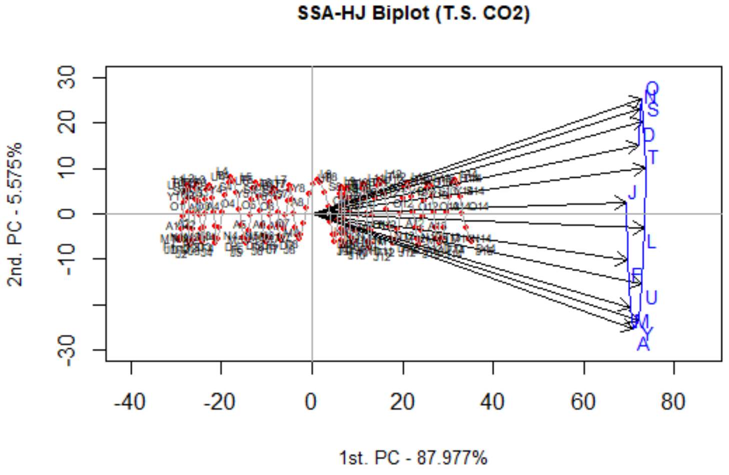

Figure 11.2: Representation of the TS CO2 SSA-HJ-biplot (1^{st} and 2^{nd} PCs) when the window length ℓ is greater than $\kappa = 12$.

11.1.2 Centering the trajectory matrix

Article II also explains why one might eventually want to center the trajectory matrix on the columns. In this case, it is shown that the ℓ-lagged vectors and the κ-lagged vectors are not the same subseries, even when $\ell = \kappa$. When $\mathbf{X}$ is centered, all elements of a specific ℓ-lagged vector ($\mathbf{X}$ column) are subtracted from the same average (of the corresponding column) in what we call an all-to-one scheme. In turn, each element of a given κ-lagged vector ($\mathbf{X}$ row) is subtracted from a different average (of each column), called an one-to-one schema. Figure 11.3 represents the all-to-one scheme related to the TS CO_2 trajectory matrix, and where the column averages appear shifted downwards just for the sake of illustration.

The segment bundle connects the first ℓ observations of the TS CO_2 to the average of the first column of $\mathbf{X}$, in terms of what is explained in Article II. The purpose of the graph is to somehow represent the differences between the observations and $\bar{y}_1$. With that, the periodicity emerges when verifying the same pattern of segment bundle for different averages. Besides, if this pattern occurs even when the involved mean are expressive different, i.e., when for some k, $\bar{y}_i << \bar{y}_{i+k}$ or $\bar{y}_i >> \bar{y}_{i+k}$, then it is probable that a trend is present. Likewise, Figure 11.4 shows the all-to-one scheme for the κ^{th} column of the centered $\mathbf{X}$, in which the segments are an illustration of the differences between the last ℓ observations of the series and the shifted $\bar{y}_\kappa$.

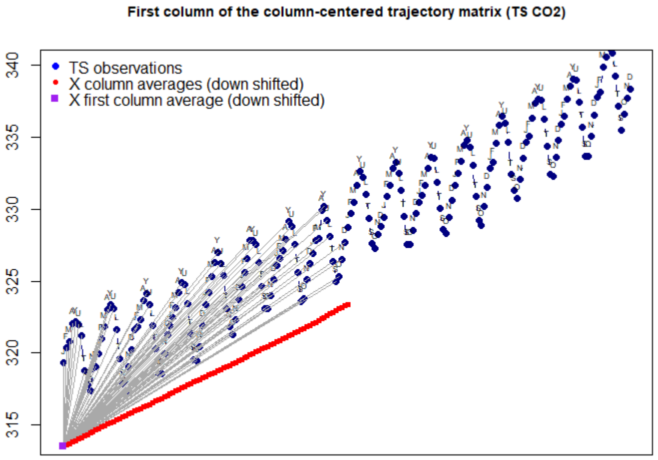

Figure 11.3: First column of the columns-centered trajectory matrix (TS CO2, 1^{st} and 2^{nd} PCs) in the all-to-one scheme.

In turn, the one-to-one scheme is represented by Figures 11.5 and 11.6, which aim to provide a visualization of what happens with the rows of the trajectory matrix when centered by the columns. In the first case, the segments link the first κ observations to all the column means $(\bar{y}_j, j = 1, \cdots, \kappa)$ shifted downwards. On the other, the procedure

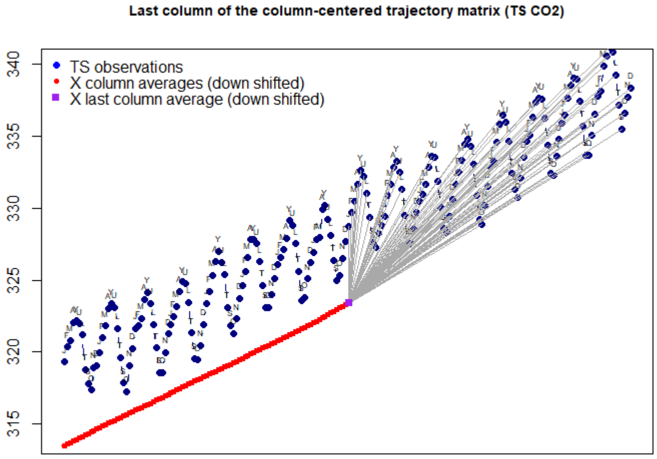

Figure 11.4: κ^{th} column of the columns-centered trajectory matrix (TS CO2, 1^{st} and 2^{nd} PCs) in the all-to-one scheme.

is repeated for the last observations. These graphics show what happens with the rows of **X** regarding differences in relation to distinct means. Ultimately, this is one way to understand why the contours of points and arrows are shown in different formats in the SSA-HJ-biplot.

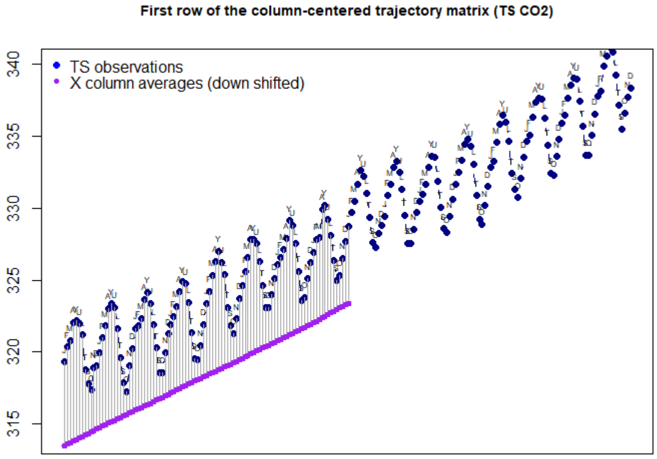

Figure 11.5: First row of the columns-centered trajectory matrix (TS CO2, 1^{st} and 2^{nd} PCs) in the one-to-one scheme.

Regarding computational efficiency, the problem P1 was partially addressed since

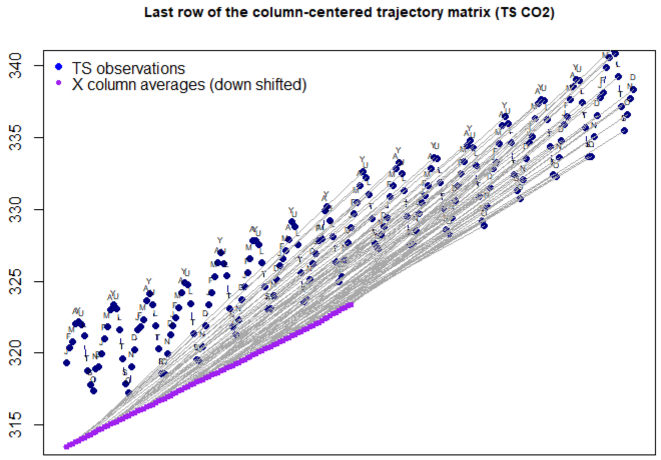

Figure 11.6: Last row of the columns-centered trajectory matrix (TS CO2, 1^{st} and 2^{nd} PCs) in the one-to-one scheme.

the choice made by the NIPALS algorithm did not prove to be advantageous over the SVD method in terms of speed, given that the trajectory matrices used were not of high dimension. Nevertheless, NIPALS worked well when asked to handle missing data, dispensing with imputation methods. In cases where the ST shows some structural change, visualization becomes more problematic. Thus, complementary measures were developed in Article III to guarantee the interpretability of the SSA-HJ-biplot.

11.2 The structural change detection and NIPALS instability

A structural change in the context of TS analysis can be characterized by the interruption of the LRR that governs the process over some time interval [39]. This type of heterogeneity brings undesired complexity for methods of visual representation based on determining directions of maximum variability of the data, as is the case of the SSA-HJ-biplot. Nevertheless, the intervals between two subsequent interruption points are suitable for applying the SSA-HJ-biplot. Article III handles the problem answering the questions P6 and P7.

During the implementation of the detection method, we used massive sequences of iterations of the SSA method to discover these breakpoints. At each iteration, a trajectory matrix constructed from subseries of the original series was decomposed by NIPALS twice, once in a robust way and once in a usual way. For the first case, we developed a robust version of the NIPALS based on the L1 rather than the L2 least-squares norm. In this phase, the instability of the NIPALS algorithm emerged so that, in many cases, it did not reach convergence. On the other hand, for the same data, the robust SVD method decomposed the corresponding trajectory matrix without effort. Therefore, it is worth mentioning that in the case of the CPD method proposed in Article III, some care must be taken when replacing the SVD method with the NIPALS algorithm due to the instability of the latter.

11.3 Development summary of the *areabiplot* package

Hosting a package in the CRAN repository guarantees code authenticity and more discoverability. But verification criteria can make processing time-consuming and cumbersome. Attention to small details is necessary even with simple code (e.g., the *areabiplot*). We use the *devtools* package in code development and the *Roxygen2* in the documentation phase to minimize errors. Below, as examples, are some specific solutions adopted:

- As it is a graphical application, the *areabiplot* allows the user to change graphical parameters, such as the background color. As a good practice, the function *on.exit* (exit code) for resetting graphic parameters was used.
- To ensure a unity ratio to factor axes, the *asp* parameter was set to 1 by default. The *asp* defines the aspect ratio, i.e., the proportional relationship between the width and height of the plot axes.
- The *scaling mode* solution. At this point, the code allows the user to assign which way he intends to allocate the diagonal matrix of singular values, whether in the HJ scheme or the one proposed by the creator of the area biplot.

An example of using the package in the multivariate context is shown at the end of the documentation. The result is shown in Figure 11.7.

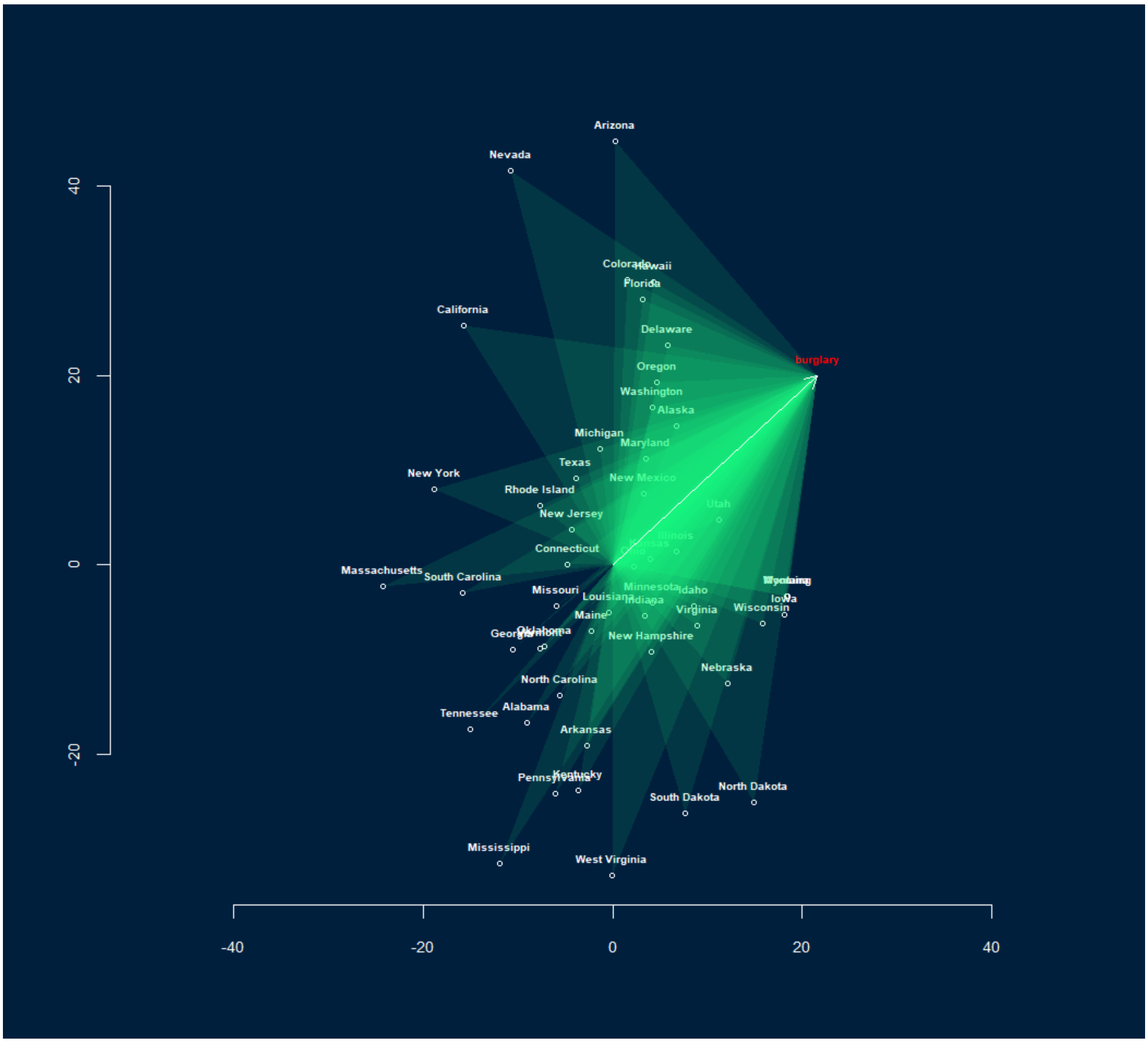

Figure 11.7: Example of the area biplot provided in the *areabiplot* package documentation.

Chapter 12

Conclusions

The research work carried out in this investigation resulted in a new way of visualizing the decomposition of a TS called SSA-HJ-biplot. The closer the series is to being classified as separable, the greater the ability of this approach to provide interpretable graphs. The properties of the biplot methods, especially the HJ-biplot, assure the interpretability of the tool. The exploratory nature of this new method also allows evaluating the spectral structure of the trajectory matrix of a univariate TS, providing indications of other characteristics, for example, regarding the stationarity or not of the series. One of the research questions aligned with this aspect is related to expanding the capabilities of the method created to emphasize the visualization of the approach. And the SSA Area biplot method is one of the achievements that help answer this issue, given its capacity to visually estimate the dominant periodicities of a TS. To further increase the range of possibilities of both tools, we created an alternative procedure for their application between points of structural changes in the analyzed series.

During the investigation, some of the objectives initially established were achieved, and other setbacks that arose in the course of the development of the work were also circumvented, highlighting:

- The definition of at least fifteen properties and interpretation rules of graphical tools in the context of SSA-HJ-biplot and SSA Area biplot.
- The implementation of a Software (R package) that automated the SSA Area biplot method but also works in the more general context of the Area biplot method.

- The definition of a distance measure to evaluate sudden changes in the direction of a PC to detect structural changes in a TS.
- The development of a CPD method based on a sudden PC's change of direction and using SSA.
- The verification of an optimal value for the window length, permitting representing the ℓ-lagged vectors and κ-lagged vectors simultaneously in the same SSA-HJ-biplot with optimal quality.
- The assignment of an alternative matrix decomposition method (NIPALS) within the SSA allowing the application of the method to TS in which missing data is checked.

This new method contributes to the visual interpretability of SSA and provides another perspective on time series analysis. As expected, it has its limitations, but the answers to the research questions raised here showed its versatility in applying it to more complex data. Thus, the following research steps will focus on improving the method and expanding its applicability. In the first case, we will seek to establish more robust connections between the SSA-HJ-biplot and the separability of the components of a TS, creating new graphical tools and improving those proposed here. The other consists of possible adaptations of the SSA-HJ-biplot for use in TS forecasting.

References

1. Alanqary, A., Alomar, A. & Shah, D. Change Point Detection via Multivariate Singular Spectrum Analysis. *Advances in Neural Information Processing Systems* **34,** 23218–23230 (2021).

2. Álvarez, F. J. D. & Villardon, P. G. *A proposal for spatio-temporal analysis of traffic matrices using HJ-biplot* in *2015 IEEE International Workshop on Measurements & Networking (M&N)* (2015), 1–6.

3. Apostol, T. *Calculus, vol. 1 e vol. 2* 1969.

4. Balcerowska-Czerniak, G., Wronkowski, A., Antończak, A., Wronkowska, A., *et al.* The potential of multivariate analysis to phase identification based on X-ray diffraction patterns. *Chemometrics and Intelligent Laboratory Systems* **135,** 126–132 (2014).

5. Bassani, N., Ambrogi, F., Coradini, D., Boracchi, P. & Biganzoli, E. Validation of Gene Expression Profiles in Genomic Data through Complementary Use of Cluster Analysis and PCA-Related Biplots. *International Journal of Statistics in Medical Research* **1,** 162–173 (2012).

6. Baumgartner, R. *et al.* Comparison of two exploratory data analysis methods for fMRI: fuzzy clustering vs. principal component analysis. *Magnetic Resonance Imaging* **18,** 89–94 (2000).

7. Bernal, E. F., Villardon, P. G., Bernal, M. E. F. & BWidget, S. Package 'GGEBiplotGUI'. *Available in: https://cran. r-project. org/web/packages/GGEBiplotGUI* (2016).

8. Bhowmik, B., Krishnan, M., Hazra, B. & Pakrashi, V. Real-time unified single-and multi-channel structural damage detection using recursive singular spectrum analysis. *Structural Health Monitoring* **18,** 563–589 (2019).

9. Bógalo, J., Poncela, P. & Senra, E. *Strong Separability in Circulant SSA* in *Conference of the International Society for Non-Parametric Statistics* (2016), 295–309.

10. Carrasco, G. *et al.* Water quality evaluation through a multivariate statistical HJ-Biplot approach. *Journal of Hydrology* **577,** 123993 (2019).

11. Chen, J. & Saad, Y. Lanczos vectors versus singular vectors for effective dimension reduction. *IEEE Transactions on Knowledge and Data Engineering* **21,** 1091–1103 (2008).

12. Chiu, J.-E. & Tsai, C.-H. On-line concurrent control chart pattern recognition using singular spectrum analysis and random forest. *Computers & Industrial Engineering* **159,** 107538 (2021).

13. Cleveland, R. B., Cleveland, W. S., McRae, J. E. & Terpenning, I. STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *Journal of Official Statistics* **6,** 3–73 (1990).

14. Dagum, E. B. Time series modeling and decomposition. *Statistica* **70,** 433–457 (2010).

15. Dagum, E. B. & Bianconcini, S. *Seasonal adjustment methods and real time trend-cycle estimation* (Springer, 2016).

16. Dash, J. & Zhang, Y. Cleaning financial data using SSA and MSSA. *Available at SSRN 2808156* (2016).

17. Da Silva, A. O. & Freitas, A. Time Series components separation based on Singular Spectral Analysis visualization: an HJ-biplot method application. *Statistics, Optimization & Information Computing* **8,** 346–358 (2020).

18. Eckart, C. & Young, G. The approximation of one matrix by another of lower rank. *Psychometrika* **1,** 211–218 (1936).

19. Elsner, J. B. & Tsonis, A. A. *Singular spectrum analysis: a new tool in time series analysis* (Springer Science & Business Media, 1996).

20. Escobar, K. M. *et al.* Frequency of Neuroendocrine Tumor Studies: Using Latent Dirichlet Allocation and HJ-Biplot Statistical Methods. *Mathematics* **9,** 2281 (2021).

21. Espinosa, F., Bartolomé, A. B., Hernández, P. V. & Rodriguez-Sanchez, M. Contribution of Singular Spectral Analysis to Forecasting and Anomalies Detection of Indoors Air Quality. *Sensors* **22,** 3054 (2022).

22. Esposito Vinzi, V. & Russolillo, G. Partial least squares algorithms and methods. *Wiley Interdisciplinary Reviews: Computational Statistics* **5,** 1–19 (2013).

23. Forsythe, G. E. & Henrici, P. The cyclic Jacobi method for computing the principal values of a complex matrix. *Transactions of the American Mathematical Society* **94,** 1–23 (1960).

24. Gabriel, K. R. The biplot graphic display of matrices with application to principal component analysis. *Biometrika* **58,** 453–467 (1971).

25. Galindo, M. An alternative for simultaneous representation: HJ-Biplot. *Questıio* **10,** 12–23 (1986).

26. Gallego-Álvarez, I., Galindo-Villardón, M. & Rodrıguez-Rosa, M. Analysis of the sustainable society index worldwide: A study from the biplot perspective. *Social Indicators Research* **120,** 29–65 (2015).

27. Gallego-Álvarez, I., Rodrıguez-Domınguez, L. & Garcıa-Rubio, R. Analysis of environmental issues worldwide: a study from the biplot perspective. *Journal of Cleaner Production* **42,** 19–30 (2013).

28. Gao, H., Cai, J.-F., Shen, Z. & Zhao, H. Robust principal component analysis-based four-dimensional computed tomography. *Physics in Medicine & Biology* **56,** 3181 (2011).

29. Gao, J., Sacchi, M. D. & Chen, X. A fast reduced-rank interpolation method for prestack seismic volumes that depend on four spatial dimensions. *Geophysics* **78,** V21–V30 (2013).

30. Gastinel, L. N. Principal component analysis in the era of «Omics» data. *Principal component analysis–multidisciplinary applications,* 21–42 (2012).

31. Geladi, P. & Kowalski, B. R. Partial least-squares regression: a tutorial. *Analytica chimica acta* **185,** 1–17 (1986).

32. Gewers, F. L. *et al.* Principal component analysis: A natural approach to data exploration. *ACM Computing Surveys (CSUR)* **54,** 1–34 (2021).

33. Godah, W., Szelachowska, M. & Krynski, J. Application of the PCA/EOF method for the analysis and modelling of temporal variations of geoid heights over Poland. *Acta Geodaetica et Geophysica* **53,** 93–105 (2018).

34. Golub, G. H. & Reinsch, C. in *Linear algebra* 134–151 (Springer, 1971).

35. Golyandina, N. & Osipov, E. The "Caterpillar"-SSA method for analysis of time series with missing values. *Journal of Statistical planning and Inference* **137,** 2642–2653 (2007).

36. Golyandina, N. Particularities and commonalities of singular spectrum analysis as a method of time series analysis and signal processing. *Wiley Interdisciplinary Reviews: Computational Statistics* **12,** e1487 (2020).

37. Golyandina, N. & Korobeynikov, A. Basic singular spectrum analysis and forecasting with R. *Computational Statistics & Data Analysis* **71,** 934–954 (2014).

38. Golyandina, N., Korobeynikov, A., Shlemov, A. & Usevich, K. Multivariate and 2D extensions of singular spectrum analysis with the Rssa package. *arXiv preprint arXiv:1309.5050* (2013).

39. Golyandina, N., Nekrutkin, V. & Zhigljavsky, A. A. *Analysis of time series structure: SSA and related techniques* (CRC press, 2001).

40. Golyandina, N. & Shlemov, A. Variations of singular spectrum analysis for separability improvement: non-orthogonal decompositions of time series. *arXiv preprint arXiv:1308.4022* (2013).

41. Gower, J., Groenen, P. & van de Velden, M. Area biplots. *Journal of Computational and Graphical Statistics* **19,** 46–61 (2010).

42. Gower, J. C. & Hand, D. J. *Biplots* (CRC Press, 1995).

43. Greenacre, M. J. *Biplots in practice* (Fundacion BBVA, 2010).

44. Halko, N., Martinsson, P.-G. & Tropp, J. A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review* **53,** 217–288 (2011).

45. Hassani, H. & Mahmoudvand, R. *Singular spectrum analysis: Using R* (Springer, 2018).

46. Hestenes, M. R. Inversion of matrices by biorthogonalization and related results. *Journal of the Society for Industrial and Applied Mathematics* **6,** 51–90 (1958).

47. Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology* **24,** 417 (1933).

48. Howley, T., Madden, M. G., O'Connell, M.-L. & Ryder, A. G. *The effect of principal component analysis on machine learning accuracy with high dimensional spectral data* in *International Conference on Innovative Techniques and Applications of Artificial Intelligence* (2005), 209–222.

49. Hyndman, R. J. & Athanasopoulos, G. *Forecasting: principles and practice* (OTexts, 2018).

50. Hyvärinen, A. Independent component analysis: recent advances. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **371,** 20110534 (2013).

51. Ivanov, M. A. & Evtimov, S. N. Seasonality in the biplot of Northern Hemisphere temperature anomalies. *Quarterly Journal of the Royal Meteorological Society* **140,** 2650–2657 (2014).

52. Jolliffe, I. T. Springer series in statistics. *Principal component analysis* **29** (2002).

53. Jolliffe, I. T. & Cadima, J. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **374,** 20150202 (2016).

54. Jutten, C. & Herault, J. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal processing* **24,** 1–10 (1991).

55. Kandasamy, S. *et al.* Improving the consistency and continuity of MODIS 8 day leaf area index products. *International Journal of Electronics and Telecommunications* **58,** 141–146 (2012).

56. Koch, I. *Analysis of multivariate and high-dimensional data* (Cambridge University Press, 2013).

57. Kogbetliantz, E. Solution of linear equations by diagonalization of coefficients matrix. *Quarterly of Applied Mathematics* **13,** 123–132 (1955).

58. Komhyr, W. *et al.* Global atmospheric CO2 distribution and variations from 1968–1982 NOAA/GMCC CO2 flask sample data. *Journal of Geophysical Research: Atmospheres* **90,** 5567–5596 (1985).

59. Koren, Y., Bell, R. & Volinsky, C. Matrix factorization techniques for recommender systems. *Computer* **42,** 30–37 (2009).

60. Korobeynikov, A. Computation-and space-efficient implementation of SSA. *arXiv preprint arXiv:0911.4498* (2009).

61. Lange, K., Chambers, J. & Eddy, W. *Numerical analysis for statisticians* (Springer, 1999).

62. Leles, M. C., Sansão, J. P. H., Mozelli, L. A. & Guimarães, H. N. A new algorithm in singular spectrum analysis framework: The Overlap-SSA (ov-SSA). *SoftwareX* **8,** 26–32 (2018).

63. Leles, M. C., Sansão, J. P. H., Mozelli, L. A. & Guimarães, H. N. Improving reconstruction of time-series based in Singular Spectrum Analysis: A segmentation approach. *Digital Signal Processing* **77,** 63–76 (2018).

64. Librero, A. B. N., Villardon, P. G. & Freitas, A. *biplotbootGUI: Bootstrap on Classical Biplots and Clustering Disjoint Biplot* R package version 1.2 (2019). https://CRAN.R-project.org/package=biplotbootGUI.

65. Lima, G. *et al. Gap filling of precipitation data by SSA-singular spectrum analysis* in *Journal of Physics: Conference Series* **759** (2016), 012085.

66. Liu, K., Law, S.-S., Xia, Y. & Zhu, X. Singular spectrum analysis for enhancing the sensitivity in structural damage detection. *Journal of Sound and Vibration* **333,** 392–417 (2014).

67. Lorber, A., Wangen, L. E. & Kowalski, B. R. A theoretical foundation for the PLS algorithm. *Journal of Chemometrics* **1,** 19–31 (1987).

68. Mahoney, M. W. & Drineas, P. CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences* **106,** 697–702 (2009).

69. Martınez-Regalado, J. A., Murillo-Avalos, C. L., Vicente-Galindo, P., Jiménez-Hernández, M. & Vicente-Villardón, J. L. Using HJ-Biplot and External Logistic Biplot as Machine Learning Methods for Corporate Social Responsibility Practices for Sustainable Development. *Mathematics* **9,** 2572 (2021).

70. Mehta, R. & Rana, K. *A review on matrix factorization techniques in recommender systems* in *2017 2nd International Conference on Communication Systems, Computing and IT Applications (CSCITA)* (2017), 269–274.

71. Meng, C. *et al.* Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in bioinformatics* **17,** 628–641 (2016).

72. Mi, X., Liu, H. & Li, Y. Wind speed prediction model using singular spectrum analysis, empirical mode decomposition and convolutional support vector machine. *Energy conversion and management* **180,** 196–205 (2019).

73. Mi, X. & Zhao, S. Wind speed prediction based on singular spectrum analysis and neural network structural learning. *Energy Conversion and Management* **216,** 112956 (2020).

74. Miranda, A. R. *et al.* Association of Dietary Intake of Polyphenols with an Adequate Nutritional Profile in Postpartum Women from Argentina. *Preventive Nutrition and Food Science* **27,** 20 (2022).

75. Miyashita, Y., Itozawa, T., Katsumi, H. & Sasaki, S.-I. Comments on the NIPALS algorithm. *Journal of chemometrics* **4,** 97–100 (1990).

76. Moskvina, V. & Zhigljavsky, A. An algorithm based on singular spectrum analysis for change-point detection. *Communications in Statistics-Simulation and Computation* **32,** 319–352 (2003).

77. Nieto, A. B., Galindo, M. P., Leiva, V. & Vicente-Galindo, P. A methodology for biplots based on bootstrapping with R. *Revista colombiana de estadıstica* **37,** 367–397 (2014).

78. Ochs, M. F. & Fertig, E. J. *Matrix factorization for transcriptional regulatory network inference* in *2012 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)* (2012), 387–396.

79. Oropeza, V. & Sacchi, M. Simultaneous seismic data denoising and reconstruction via multichannel singular spectrum analysis. *Geophysics* **76,** V25–V32 (2011).

80. Paterek, A. *Improving regularized singular value decomposition for collaborative filtering* in *Proceedings of KDD cup and workshop* **2007** (2007), 5–8.

81. Paul, L. C. & Al Sumam, A. Face recognition using principal component analysis method. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)* **1,** 135–139 (2012).

82. Plaza, E. G. & López, P. N. Surface roughness monitoring by singular spectrum analysis of vibration signals. *Mechanical systems and signal processing* **84,** 516–530 (2017).

83. R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2021). https://www.R-project.org/.

84. Rao, C. *Linear Statistical Inference and Its Applications (Paper back ed)* 2002.

85. Rodrigues, P. C., Lourenço, V. & Mahmoudvand, R. A robust approach to singular spectrum analysis. *Quality And Reliability Engineering International* **34,** 1437–1447 (2018).

86. Rodrigues, P. C., Tuy, P. G. & Mahmoudvand, R. Randomized singular spectrum analysis for long time series. *Journal of Statistical Computation and Simulation* **88,** 1921–1935 (2018).

87. Roman, S., Axler, S. & Gehring, F. *Advanced linear algebra* (Springer, 2005).

88. Schoellhamer, D. H. Singular spectrum analysis for time series with missing data. *Geophysical research letters* **28,** 3187–3190 (2001).

89. Seasholtz, M. B., Pell, R. J. & Gates, K. E. Comments on the power method. *Journal of chemometrics* **4,** 331–334 (1990).

90. Shen, Y., Peng, F. & Li, B. Improved singular spectrum analysis for time series with missing data. *Nonlinear Processes in Geophysics* **22,** 371–376 (2015).

91. Silva, A. & Freitas, A. *areabiplot: Area Biplot* R package version 1.0.0 (2021). https://CRAN.R-project.org/package=areabiplot.

92. Stein-O'Brien, G. L. *et al.* Enter the matrix: factorization uncovers knowledge from omics. *Trends in Genetics* **34,** 790–805 (2018).

93. Stewart, G. W. On the early history of the singular value decomposition. *SIAM review* **35,** 551–566 (1993).

94. Strang, G. *Linear algebra and its applications.* (Belmont, CA: Thomson, Brooks/Cole, 2006).

95. Strauch, M. & Galizia, C. G. *Fast PCA for processing calcium-imaging data from the brain of Drosophila melanogaster* in *BMC Medical Informatics and Decision Making* **12** (2012), 1–10.

96. Tadić, L., Bonacci, O. & Brleković, T. An example of principal component analysis application on climate change assessment. *Theoretical and Applied Climatology* **138,** 1049–1062 (2019).

97. Torres-Salinas, D., Robinson-Garcıa, N., Jiménez-Contreras, E., Herrera, F. & López-Cózar, E. D. On the use of biplot analysis for multivariate bibliometric and scientific indicators. *Journal of the American Society for Information Science and Technology* **64,** 1468–1479 (2013).

98. Vairinhos, V., Parreira, R., Lampreia, S., Lobo, V. & Galindo, P. Vibration analysis based on HJ-biplots. *International Journal of Prognostics and Health Management* **9** (2018).

99. Vaisman, L., Zariffa, J. & Popovic, M. R. Application of singular spectrum-based change-point analysis to EMG-onset detection. *Journal of Electromyography and Kinesiology* **20,** 750–760 (2010).

100. Vellido, A., Martın-Guerrero, J. D. & Lisboa, P. J. *Making machine learning models interpretable* in *ESANN* **12** (2012), 163–172.

101. Vlachos, M., Yu, P. & Castelli, V. *On periodicity detection and structural periodic similarity* in *Proceedings of the 2005 SIAM international conference on data mining* (2005), 449–460.

102. Wang, Y.-X. & Zhang, Y.-J. Nonnegative matrix factorization: A comprehensive review. *IEEE Transactions on knowledge and data engineering* **25,** 1336–1353 (2012).

103. Watkins, D. S. Understanding the QR algorithm, part II. *URL http://www. researchgate. net/publication/228966308_Understanding_the_QR_algorithm_Part_II/file/79e4150e24b0e6b291. pdf* (2008).

104. Wickham, H., Hester, J., Chang, W. & Bryan, J. *devtools: Tools to Make Developing R Packages Easier* R package version 2.4.3 (2021). https://CRAN.R-project.org/package=devtools.

105. Wold, H. Estimation of principal components and related models by iterative least squares. *Multivariate analysis,* 391–420 (1966).

106. Wold, S., Albano, C., Dunn III, W., Esbensen, K. & Hellberg, S. *Pattern recognition: finding and using regularities in multivariate data Food research, how to relate sets of measurements or observations to each other* in *Food research and data analysis: proceedings from the IUFoST Symposium, September 20-23, 1982, Oslo, Norway/edited by H. Martens and H. Russwurm, Jr* (1983).

107. Yang, D., Dong, Z., Lim, L. H. I. & Liu, L. Analyzing big time series data in solar engineering using features and PCA. *Solar Energy* **153,** 317–328 (2017).

108. Yeung, K. Y. & Ruzzo, W. L. Principal component analysis for clustering gene expression data. *Bioinformatics* **17,** 763–774 (2001).

109. Zakharova, A. *et al.* Quantitative assessment of cognitive interpretability of visualization. *Scientific Visualization* **10,** 145–153 (2018).

110. Zhang, Y., Le, J., Liao, X., Zheng, F. & Li, Y. A novel combination forecasting model for wind power integrating least square support vector machine, deep belief network, singular spectrum analysis and locality-sensitive hashing. *Energy* **168,** 558–572 (2019).

111. Ziemkiewicz, C. & Kosara, R. *Preconceptions and individual differences in understanding visual metaphors* in *Computer Graphics Forum* **28** (2009), 911–918.

Appendix

List of communications

III Conference on Statistics and Data Science. *Area biplot for time series feature extraction.* Alberto Silva, and Adelaide Freitas. Salvador (BA), Brazil, 2021

XXV Congress of the Portuguese Statistical Society. *Time Series Periodicity Detection using Area Biplots.* Alberto Silva, and Adelaide Freitas. Évora, Portugal, 2021.

The HJ-Biplot Visualization of the Singular Spectrum Analysis Method. VIII Workshop of Probability and Statistics group of the Center for Research Development in Mathematics and Applications (CIDMA). July 8, 2020. Department of Mathematics. University of Aveiro. Portugal.

III Encontro Luso-Galaico de Biometria, EBio2018. *Entropia normalizada e outros métodos de seleção de variáveis: um estudo comparativo com dados simulados.* Alberto Silva, Rodney Sousa, and Pedro Macedo. Aveiro, Portugal. June, 2018.