medRxiv preprint doi: https://doi.org/10.1101/2020.05.24.20111872; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Predicting COVID-19 Pandemic in Saudi Arabia Using Modified Singular Spectrum Analysis

Nader Alharbi

King Saud bin Abdulaziz University For Health Sciences, Riyadh, Saudi Arabia

King Abdullah International Medical Centre, Riyadh, Saudi Arabia

Abstract

This research presents a modified Singular Spectrum Analysis (SSA) approach for the analysis of COVID-19 in Saudi Arabia. We have proposed this approach and developed it in [1–3] for separability and grouping step in SSA, which plays an important role for reconstruction and forecasting in the SSA. The modified SSA mainly enables us to identify the number of the interpretable components required for separability, signal extraction and noise reduction. The approach was examined using different number of simulated and real data with different structures and signal to noise ratio. In this study we examine its capability in analysing COVID-19 data. Then, we use Vector SSA for predicting new data points and the peak of this pandemic. The results shows that the approach can be used as a promising one in decomposing and forecasting the daily cases of COVID-19 in Saudi Arabia.

Keywords: COVID-19; Prediction; Singular Spectrum Analysis, Separability; Eigenvalues.

1 Introduction

One of the main issues that threats our health around the glop are infectious diseases. Nowadays, the outbreak of 2019 virus disease (COVID-19) is a global pandemic [4, 5]. The first case of this virus was recognized and reported on 31-12-2019 in the city of Wuhan, the capital of Hubei in China [6]. Then, the virus has spread rapidly around the world and affected more than 200 countries [7].

The number of cases and deaths of this virus are globally considered as serious problems. The number of confirmed cases were more than 4 million and around 200 thousand deaths by 12-05-2020. Although the outbreak seems to have decreased in China, the virus and its impacts are still going global, and those numbers are increasing. This leads to our concerns about variation in the affected cases and the mortality rate of the COVID-19 pandemic. Furthermore, there are a lot of concerns about economic global impact of this crises. It is now understood that the devastating influence of the virus on economy and world health is incomparable [8].

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.24.20111872; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

The primary objective of this manuscript is construction of a reliable, robust and interpretable model describing, decomposing, forecasting the number of confirmed cases, and predict the peak of this pandemic in Saudi Arabia. The rate of mortality in Saudi Arabia is low, less than 1% till writing this paper. Thus, we are only interested in the new daily cases affected by the virus and try to detect its peak. The number of cumulative cases is more than 40000 by 12-05-2020.

There are many standard epidemiological models for modelling epidemics such SIR, see e.g. [9–11]. However, since our aim is to analyse the daily data series of COVID-19, we seek to use a promising, reliable, and capable method for analysing time series. There is a number of various methods for analysing time series, but several of these methods requiring, for example, linearity or non linearity of a particular form as they are parametric methods.

An alternative method uses non-parametric approaches that are neutral with respect to problematic areas of specification, such as linearity, stationarity and normality [13]. Thus, such approaches can show a reliable and better means of decomposing time series data. Singular Spectrum Analysis (SSA) is a relatively new non-parametric technique that has shown and proved its capabible use in several applications of time series in different disciplines, such as genetics and biology [14, 15], medicine [16, 17], engineering [18, 19], economics and finance [20, 21], and other areas. For its history, see [22, 23]. For more details on the theory of SSA and its applications, refer to [13, 24, 25]. A comprehensive review of the method and description of its extensions and modifications can be found in [26].

Although the signals can be affected by an internal or external noise, which often have unknown characteristics, they can be identified if the signal and noise subspaces are accurately separated. It is known that removing noises from any signal is necessary for analysing any kind of time series, and is helpful in decomposing the signal in a proper manner [27].

The main idea of SSA is to analyse the main series into different components, then reconstruct the noise free series for further analysis. It depends upon two main choices; namely, the window length L and the number of required eigenvalues, denoted by r, for reconstruction. Thus, an appropriate selection of L and r leads to a perfect analysis and separability between the time series components. It was discussed in [28] that for a series of length N, selecting $L = N/4$ is common practice.It also should be mentioned that L should be large enough, but not larger than half of the series [24]. In [29], it was shown that for a series of length N and the optimal selection of the number of eigenvalues r for reconstructing the signal, the appropriate value of the window length is $median\{1, \ldots, N\}$. Despite various attempts that have been applied, there is no universal rule for obtaining optimal selections of L and r.

We have proposed an approach in [2, 3] for the selection of the value of r for noise reduction, filtering, and signal extraction in SSA. This has also been applied to the distinction of noise from chaos in time series analysis [30], and for the correction of noise in gene expression data [31]. In [3], we have developed the approach and introduced new criteria to the discrimination between epileptic seizure and normal EEG signals, the filtering of the EEG signal segments, and elimination of the noise included in the signal. The approach is mainly used to identify the required number of eigen-

medRxiv preprint doi: https://doi.org/10.1101/2020.05.24.20111872; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

values/singular values corresponding to the signal component, which depends on the distribution of the eigenvalues of a scaled Hankel matrix. The correlation between eigenvalues, the coefficients of skewness, kurtosis and variation of the eigenvalues distributions were proposed and proved to be new criteria for the separability between signal and noise components as they can split the eigenvalues into two groups [2]. Different simulated and real signals were used considering different signal to noise ratio in [2, 3], and evaluated to show the ability of the approach in the selection of r.

The remainder of this paper is structured as follows: the following section gives a short description of the modified SSA approach and its algorithm. In Section 3, we show that this approach can decompose a synthetic data into two main distinct subspaces. Section 4 presents the implementation of the approach in decomposing and reconstructing COVID-19 daily cases series. The section also presents the prediction of COVID-19 in Saudi Arabia using Vector SSA for the extracted signal by the modified SSA. Section 5 draws the conclusion of this paper with some ideas for future work.

2 The Modified SSA method

2.1 Review

This section presents a short description of the modified SSA used in this manuscript (for more details refer to [2]). A time series is decomposed by the SSA technique into a sum of components, allowing the identification of each one as either a main or noise component. The goal here is to consider the signal as a whole so that we can identify the appropriate value of r related to the whole signal component. In other words, we are not interested in each signal component, so the selection of L rational to the periodicity of the signal components becomes less important [25]. Therefore, the modified SSA focuses on the selection of r to identify the signal subspace.

Consider a one-dimensional series $Y_N = (y_1, \ldots, y_N)$ of length N. Transferring this series into a multi-dimensional series $X_1, \ldots, X_K$ where $X_i = (y_i, \ldots, y_{i+L-1})^T \in \mathbf{R}^L$ provides $\mathbf{X} = (x_{i,j})_{i,j=1}^{L,K}$, where L is an integer $(2 \leqslant L \leqslant N/2)$ and $K = N - L + 1$. A matrix $\mathbf{X}$ is a Hankel matrix, all the elements along the diagonal $i + j = const$ are equal. Set $\mathbf{B} = \mathbf{X}\mathbf{X}^T$ and denote by λ_i $(i = 1, \ldots, L)$ the eigenvalues of $\mathbf{B}$ taken in decreasing order of magnitude ($\lambda_1 \geq \ldots \geq \lambda_L \geq 0$) and by $U_1, \ldots, U_L$ the orthonormal system of the eigenvectors of matrix $\mathbf{B}$ corresponding to these eigenvalues.

The SVD of matrix $\mathbf{X}$ can be written as follows:

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

where $\mathbf{X}_i = \sqrt{\lambda_i} U_i V_i^T$. The elementary matrices $\mathbf{X}_i$ having rank 1, U_i and V_i are the left and right eigenvectors of matrix $\mathbf{X}$. Note that the collection $(\sqrt{\lambda_i}, U_i, V_i)$ is called the ith eigentriple of the SVD. Note also that $||\mathbf{X}||_F^2 = tr(\mathbf{X}\mathbf{X}^T) = \sum_{i=1}^{L} \lambda_i$ and $||\mathbf{X}_i||_F^2 = \lambda_i$, where $|| \; ||_F$ denotes the Frobenius norm.

Fundamental to the question of the eigenvalues behaviour, λ_i, is that if the series size increases, there is a corresponding increase in the eigenvalues. This problem can be

medRxiv preprint doi: https://doi.org/10.1101/2020.05.24.20111872; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

overcome if $\mathbf{B}$ is dividing by its trace, $\mathbf{A} = \mathbf{B}/tr(\mathbf{B})$, which provides several important properties [1]. Let $\zeta_1, \ldots, \zeta_L$ denote the matrix $\mathbf{A}$ eigenvalues in decreasing order of magnitude ($1 \geq \zeta_1 \geq \ldots \geq \zeta_L \geq 0$). The simulation technique is performed to obtain the distribution of ζ_i and to understand the behaviour of each eigenvalue. This helps to identify the value of r. Here, the goal is to establish the distribution and related forms of ζ_i, which will be used to select the appropriate value of r for removing noise from COVID-19 series.

It was proved in our work [2] that the largest eigenvalue has a positive skewed distribution for a white noise process. Therefore, if $skew(\zeta_c)$ ($c \in \{1, \ldots, L\}$) is the maximum, and the pattern for $skew(\zeta_c)$ to $skew(\zeta_L)$ has the same pattern, the same as emerged for the white noise, then the first $r = c-1$ eigenvalues correspond to the signal and the rest to the noise. A similar procedure can be done using the the coefficients of kurtosis and variation of ζ_i. Furthermore, if $\rho_s(\zeta_{c-1}, \zeta_c)$ is the minimum, and the pattern for the set $\{\rho_s(\zeta_i, \zeta_{i+1})\}_{i=c}^{L-1}$ is similar to what was observed for the white noise, then we select the first $r = c-1$ eigenvalues for the signal and the rest for the noise component (for more information see [2]).

In this research, we use the third and fourth central measures moments of the distribution, which are the skewness ($Skew$) and kurtosis ($Kurt$). Skewness is a measure of asymmetry of the data distribution, whilst kurtosis describes the distribution of observed data in terms of shape or peak. We use these measures as criteria for choosing the value of r, which can be calculated for m simulation as follows:

$$Skew(\zeta_i) = \frac{\frac{1}{m}\sum_{n=1}^{m}\left(\zeta_{i,n} - \overline{\zeta}_i\right)^3}{\left[\frac{1}{m-1}\sum_{n=1}^{m}\left(\zeta_{i,n} - \overline{\zeta}_i\right)^2\right]^{3/2}}, \tag{2}$$

$$Kurt(\zeta_i) = \frac{\frac{1}{m}\sum_{n=1}^{m}\left(\zeta_{i,n} - \overline{\zeta}_i\right)^4}{\left[\frac{1}{m}\sum_{n=1}^{m}\left(\zeta_{i,n} - \overline{\zeta}_i\right)^2\right]^2} - 3. \tag{3}$$

Moreover, the coefficient of variation, (CV), which is defined as the ratio of the standard deviation $\sigma(\zeta_i)$ and $\overline{\zeta}_i$ can be calculated mathematically from the following formula:

$$CV(\zeta_i) = \frac{\sigma(\zeta_i)}{\overline{\zeta}_i}. \tag{4}$$

In addition, the Spearman correlation ρ_s between the eigenvalues ζ_i and ζ_j ($i, j = 1, \ldots, L$) is also calculated to enhance the results obtained by those measures:

$$\rho_s = cor(\zeta_i, \zeta_j) = 1 - \frac{6\sum d_n^2}{m(m^2-1)}, \tag{5}$$

where $d_n = x_n - y_n$ ($n = 1, \ldots, m$) is the difference between x_n and y_n which are the ranks of $\zeta_{i,n}$ and $\zeta_{j,n}$ respectively, and $\zeta_{i,n}$ is the n-th observation for the i-th eigenvalue (ζ_i), $\overline{\zeta}_i = \left(\sum_{n=1}^{m}\zeta_{i,n}\right)/m$.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.24.20111872; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

These measures of difference between the eigenvalues related to the signal and noise components can specify the cut-off point of separability; the number of leading SVD components that are separated from the residual. Thus, the last cut-off point of separability between the signal and noise components obtained by the suggested measures, corresponds to the rank estimation.

The eigenvalues can be split into two groups by using the above criteria; the first corresponds to the signal and the second to the noise component. Furthermore, the Spearman correlation ρ between ζ_i and ζ_j is also calculated to support the outcomes obtained by those measures. The absolute value of the correlation coefficient is considered; 1 shows that ζ_i and ζ_j have perfect positive correlation, whilst 0 indicates there is no correlation between them. The matrix of the absolute values of the Spearman correlation gives a full analysis of the trajectory matrix, and in this analysis each eigenvalue corresponds to an elementary matrix of the SVD. Note that if the absolute value of ρ is close to zero, then the corresponding components are almost orthogonal; however, if it is close to one, then the two components are far from being orthogonal and so it is difficult to separate them. Thus, if $\rho = 0$ between two reconstructed components, this shows that these two reconstructed series are separable. The results of ρ between the eigenvalues for the white noise are quite large (see [2]), which helps in the discrimination of the noise part.

Once r is identified, then the matrices $\mathbf{X}_i$ can be split into two groups. Therefore, equation (1) can be written as follows:

$$\mathbf{X} = \mathbf{S} + \mathbf{E}, \tag{6}$$

where $\mathbf{S} = \sum_{i=1}^{r} \mathbf{X}_i$ is the signal matrix and $\mathbf{E} = \sum_{i=r+1}^{L} \mathbf{X}_i$ is the noise one. We then use diagonal averaging to transform matrix $\mathbf{S}$ into a new series of size N (see [24]).

2.2 Algorithm

The algorithm consists of two main stages. The steps of the first stage using the coefficients of skewness, kurtosis, variation and correlation can help us to obtain the optimal value of r for the separability between signal and noise as they split the eigenvalues into two groups. The steps of the second stage are used to reconstruct the free noise series

2.2.1 Stage 1:

1. Map a one-dimensional time series $Y_N = (y_1, \ldots, y_N)$ into multi-dimensional series $X_1, \ldots, X_K$ with vectors $X_i = (y_i, \ldots, y_{i+L-1})^T \in \mathbf{R}^L$, where the window length L is an integer; $2 \leq L \leq N/2$, and $K = N - L + 1$. This step gives us the Hankel matrix $\mathbf{X} = [X_1, \ldots, X_K] = (x_{ij})_{i,j=1}^{L,K}$.

2. Compute the matrix $\mathbf{A} = \mathbf{X}\mathbf{X}^T / tr(\mathbf{X}\mathbf{X}^T)$.

3. Decompose matrix $\mathbf{A}$ as $\mathbf{A} = \mathbf{P}\mathbf{\Gamma}\mathbf{P}^T$, where $\mathbf{\Gamma} = diag(\zeta_1, \ldots, \zeta_L)$ is the diagonal matrix of the eigenvalues of $\mathbf{A}$ that has the order $(1 \geq \zeta_1 \geq \zeta_2, \ldots, \zeta_L \geq 0)$ and

medRxiv preprint doi: https://doi.org/10.1101/2020.05.24.20111872; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

$\mathbf{P} = (P_1, P_2, \ldots, P_L)$ is an orthogonal matrix whose columns are the corresponding eigenvectors.

4. Simulate the original series m times and calculate the eigenvalues for each series. We simulate y_i from a uniform distribution with boundaries $y_i - a$ and $y_i + b$, where $a = \mid y_{i-1} - y_i \mid$ and $b = \mid y_i - y_{i+1} \mid$.

5. Compute the skewness coefficient for each eigenvalue, $skew(\zeta_i)$. If $skew(\zeta_c)$ is the maximum, and the pattern for $skew(\zeta_c)$ to $skew(\zeta_L)$ has a similar pattern to the white noise, select $r = c - 1$.

6. Compute the coefficient of kurtosis for each eigenvalue, $kurt(\zeta_i)$. If $kurt(\zeta_c)$ is the maximum, select $r = c - 1$.

7. Compute the coefficient of variation, $CV(\zeta_i)$. The result of CV splits the eigenvalues in two groups, from ζ_1 to ζ_{c-1} which correspond to the signal, and the remainder, which have an almost U shape, correspond to the noise.

8. Compute the absolute values of the correlation matrix between the eigenvalues, and represent them in a 20-grade grey scale from white to black corresponding to the values of the correlations from 0 to 1. This matrix also splits the eigenvalues into two groups, from ζ_1 to ζ_r which correspond to the signal, and the remainder, which correspond to the noise.

2.2.2 Stage 2

1. Calculate the approximated signal matrix $\widetilde{\mathbf{S}}$, that is $\widetilde{\mathbf{S}} = \sum_{i=1}^{r} \mathbf{X}_i$, where r is obtained from the first stage, $\mathbf{X}_i = \sqrt{\lambda_i} U_i V_i^T$, U_i and V_i stands for the left and right eigenvectors of the trajectory matrix.

2. By averaging over the diagonals of matrix $\widetilde{\mathbf{S}}$, this gives a one dimensional series, which is the approximate signal $\widetilde{S}$.

The capability of the modified SSA using different synthetic data, including series generated from chaotic map systems with different Signal to Noise ratio (SNR), were presented in [2]. This result confirms that the approach works promisingly for any series that is mixed with a low or high noise level.

Each eigenvalue or singular value contributes to the trajectory matrix decomposition. We can consider the ratio $\overline{\zeta}_i \times 100$ to be the characteristic of matrix $\mathbf{H}_i$ to Eq. (1). Thus, $100 \times \sum_{i=1}^{r} \overline{\zeta}_i$ is considered as the characteristic of the optimal approximation of $\mathbf{H}$ by matrices of rank r.

3 Separability in Synthetic data

We should mention that using the standard criteria in the basic SSA, the weighted correlation or w-correlation for separability and grouping (for more information see

medRxiv preprint doi: https://doi.org/10.1101/2020.05.24.20111872; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

[25]), does not always provide a good separability and correct selection of r, specially for real data.

It was shown in [2] that the results based on $Skew$, $Kurt$, CV, and ρ_s are more accurate than those obtained by the w-correlations for small window length, particularly for a data where a linear trend is included in the series.

Thus, we use the modified SSA, in particular, we use some of those proved criteria on the distribution of ζ_i, which given in the previous sections to identify r. The results are plausible and reliable.

We will provide here one synthetic example to show the capability of the approach before applying it to COVID-19 data, for more examples considering different types of series and evaluations with different criteria refer to [2].

Simulated data: In the following example, a white noise process ϵ_t was added to an exponential trend series.

$$y_t = \alpha_1 + \alpha_1 exp(\alpha_2 t) + \epsilon_t,$$

where $t = (1, \ldots, N)$, $N = 42$, $\alpha_1 = 10$, $\alpha_2 = 0.09$, and ϵ_t is a Gaussian white noise process with variance 1 (see 1). It is obvious that the number of eigenvalues required to reconstruct the signal for this series is 2, as we have a constant adding to exponential curve, which corresponds to the rank estimation (see [24].

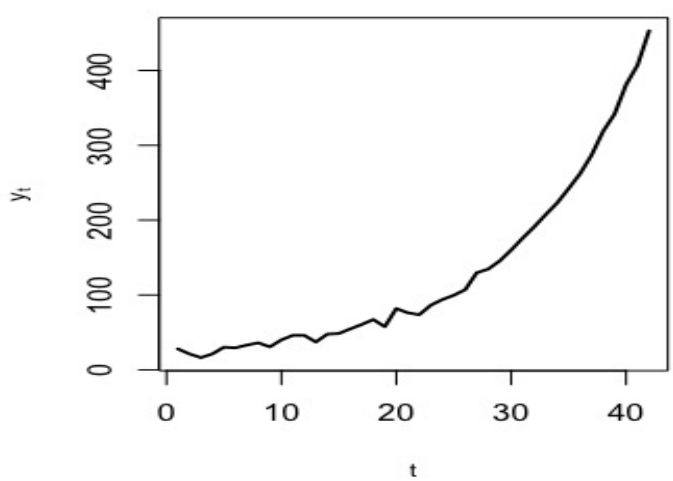

Fig. 1: A realization of the simulated series.

By looking at the w-correlations, and the logarithm of the eigenvalues, we may use only the first component to extract the signal (see Fig. 2).

However, using the suggested measures and criteria, this gives us the correct value of r. Fig. 3 represents the kurtosis coefficient of ζ_i $(i = 1, \ldots, L)$. The maximum value of the kurtosis coefficient is considered as one of the rules and indicators we use for the start of the noise. It is clear that the maximum kurtosis coefficients of ζ_i is obtained for

medRxiv preprint doi: https://doi.org/10.1101/2020.05.24.20111872; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

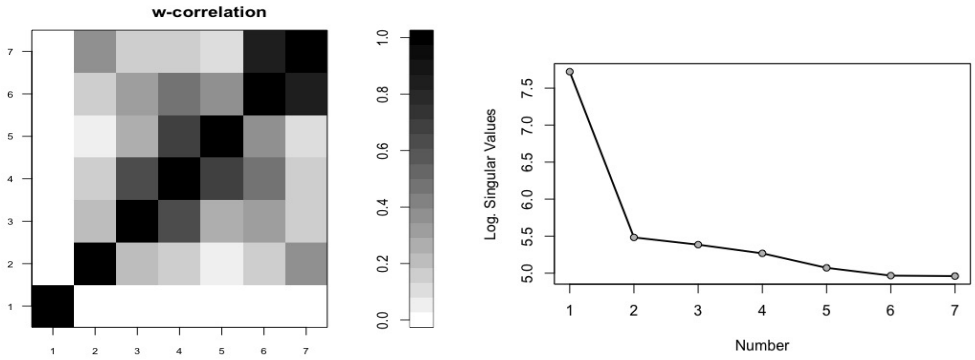

Fig. 2: w-correlations matrix (left) for the 7 reconstructed components of the simulated series, and logarithms of the 7 simulated series eigenvalues (right).

$\zeta_{c=3}$. Thus, the number of eigenvalues required to extract the signal is $r = c - 1 = 2$. Similar results emerged by using the values of *skew* and CV (see Fig 4).

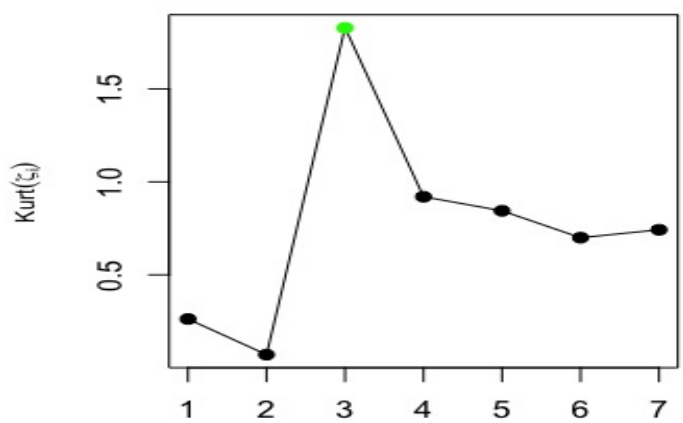

Fig. 3: Kurtosis of ζ_i for the simulated series.

In addition, the Spearman correlation coefficient between ζ_i and ζ_{i+1} is also calculated. Fig. 5 (left) shows the correlation between ζ_i and ζ_{i+1}. For the correlation coefficient, the minimum value of ρ_s between ζ_{c-1} and ζ_c is used as an another indicator for the cut-off point. The results are similar to what emerged by other criteria, and confirm that the approach works properly. Different criteria were used in [2] to evaluate the approach, for example, RMSE and MAE, which confirm that the modified approach can be used as a promising one.

The correlation matrix also enables us to distinguish and separate the different components from each other. Thus, the correlation matrix of ζ_i is identify the separability between the components. If the absolute value of the correlation coefficient between ζ_i

medRxiv preprint doi: https://doi.org/10.1101/2020.05.24.20111872; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

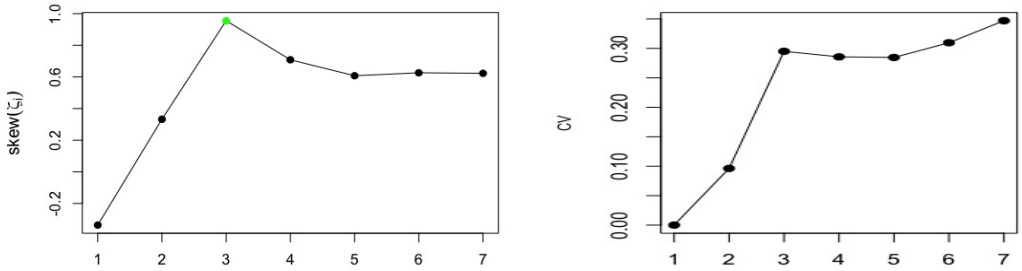

Fig. 4: Skewness (left) and coefficient of variation (right) of ζ_i for the simulated series.

and ζ_j is small, then the corresponding components are almost orthogonal; however, if the value is large, then the corresponding series are far from being orthogonal and thus they are not neatly separable. It is clear that the signal can be separated from the noise since the top right pattern from the correlation matrix is related to the white noise process (see Fig. 5 (right)).

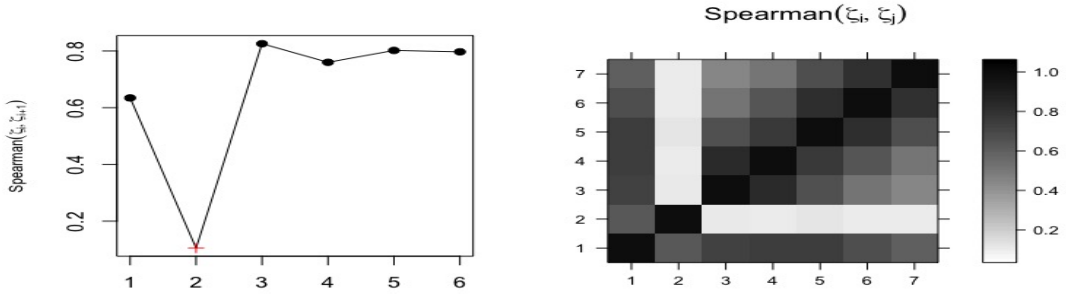

Fig. 5: Spearman correlation of $(\zeta_i,\ \zeta_{i+1})$ (left) and matrix of Spearman correlation between $(\zeta_i,\ \zeta_j)$.

4 COVID-19 data analysis

The daily confirmed cases of COVID-19 in Saudi Arabia [32] is used in this research. First, We have used the first 42 days data; from 02-03-2020 to 12-04-2020. The aim is to analyse the data and make prediction from 13-04-2020 to end of June 2020, and detect the peak. The number of daily cases series is depicted in Fig. 6. Second, we have updated our data on 12-05-2020, and included values from 13-04-2020 to 12-05-2020, so the total became 71 values. This does not affect the required number of eigenvalues for the reconstruction stage, this will be discussed in the following part.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.24.20111872; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

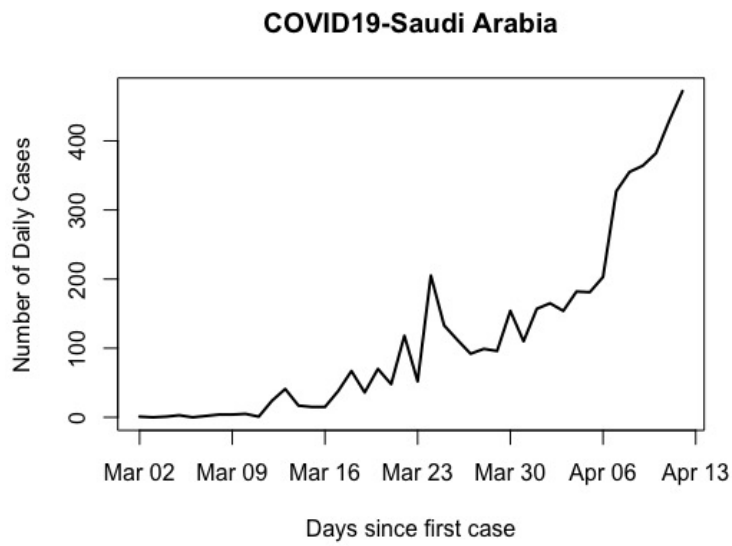

Fig. 6: COVID-19 daily confirmed cases time series in Saudi Arabia (02-03-2020 to 12-04-20202)

4.1 Separability and selection of the components

let us now start with the first data. As we mentioned earlier, since our aim is to extract the signal as a whole, we can choose any value for L, and the goal to find the best choice of r. Furthermore, based on our research [2], we showed that one can use a small window length when analysing exponential series, like the one of COVID-19 series. The selection of $L = 7$, provide the best and reasonable results with the required r that will be obtained by the proposed approach.

Th results based on those measure in extraction the signal for forecasting, give a curve with likely peak. However, the prediction using many other choices of L and r do not give any end or peak for this pandemic and go up exponentially, and this is impossible as this pandemic will not stay forever. This also support the obtained results. Therefore, the important task now is the selection of the number of eigenvalues r that required for reconstruction and build the model for forecasting.

Fig. 7 illustrates the results of the coefficients of skewness and kurtosis for each eigenvalue, and the results of the matrix correlations and correlation between ζ_i and ζ_{i+1} for $L = 7$. As shown by the results, for the COVID-19 daily series, the maximum values of $Skew$, $Kurt$, are observed for $\zeta_{c=3}$, and the minimum value of ρ_s is obtained between $\zeta_{c-1=2}$ and $\zeta_{c=3}$. In addition, the matrix of Spearman correlation for ζ_i and ζ_j $(i, j = 1, \ldots, 7)$ split the eigenvalues or the components into two groups; this indicates that the value of $r = 2$.

Fig. 8 depicts the the result of the reconstructed series, which is obtained by using $L = 7$ and eigentriples $r = 2$. The red and the black lines correspond to the reconstructed series and the original series, respectively. It seems that the reconstructed series has been obtained well. However, we will see later that the the reconstructed

medRxiv preprint doi: https://doi.org/10.1101/2020.05.24.20111872; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

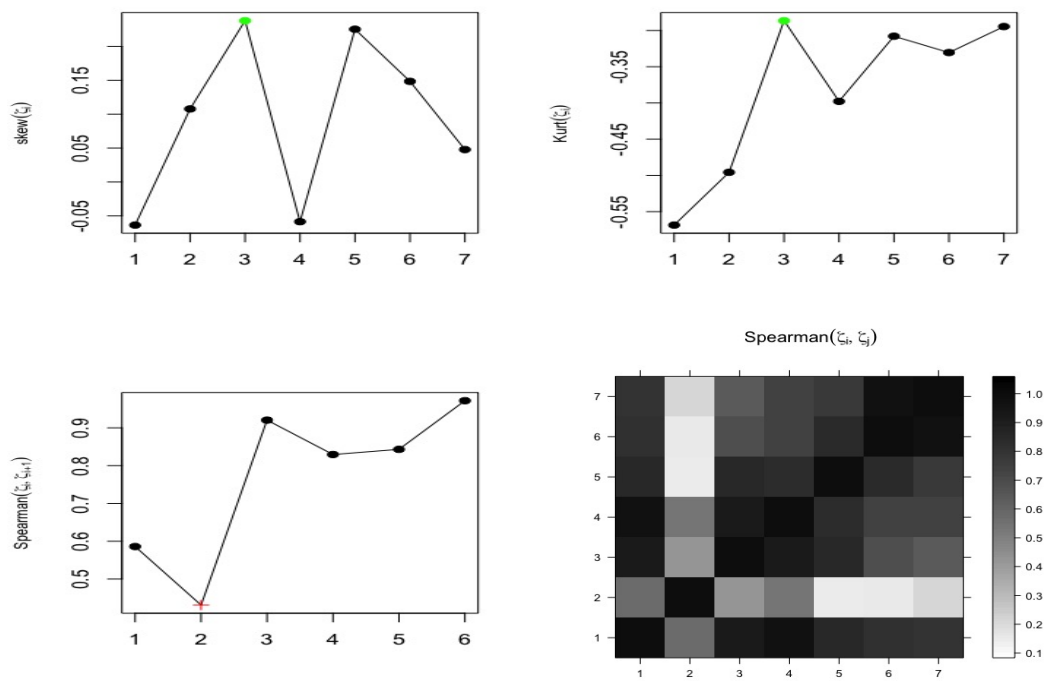

Fig. 7: All measures results for ζ_i.

series using the whole data is better than this fitted series.

4.2 Prediction daily cases of COVID-19 using VSSA

After obtaining the reconstructed series, the next aim is to predict new data, we will predict values from 13-04-2020 to the end of June 2020. There are two main forecasting methods in SSA, Vector SSA (VSSA) and Recurrent SSA (RSSA). The VSSA forecasting algorithms is the most widely used in SSA [24]. Generally, this method works more robustly than RSSA especially when a series contains outliers or when faced big shocks in the series [12]. Therefore, we use the VSSA algorithm for forecasting in this research as recommended in [13].

Vector forecasting algorithm: For performing SSA forecasting, the basic requirement is that the series satisfies a linear recurrent formula (LRF). The series $Y_N = [y_1, \ldots, y_N]$ satisfies a LRF of order $L-1$ if:

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \cdots + a_{L-1} y_{t-L+1}, \quad t = L+1, \ldots, N \tag{7}$$

The coefficient vector $A = a_1, \ldots, a_{L-1}$ is defined as follows:

$$A \equiv \tfrac{1}{1-\nu^2} \sum_{j=1}^{r} \pi_j U_j^{\nabla},$$

where $\nu^2 = \sum_{j=1}^{r} \pi_j^2$, and U_j^{∇} is the vector of the first $L-1$ components of the

medRxiv preprint doi: https://doi.org/10.1101/2020.05.24.20111872; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

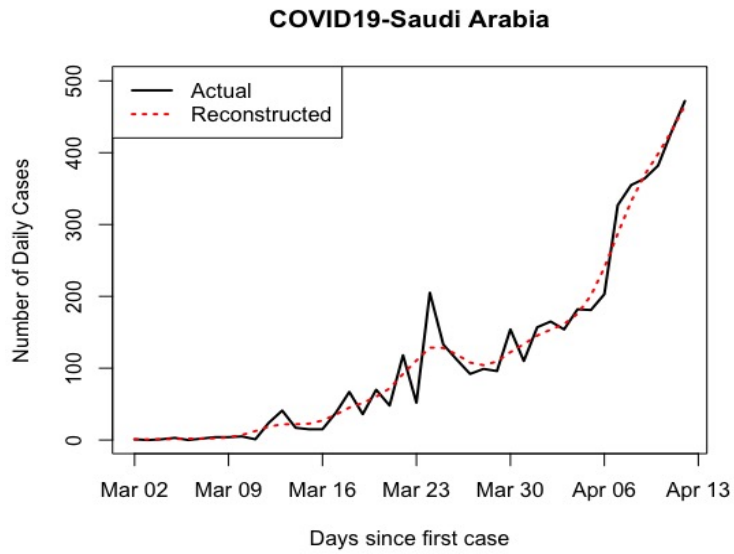

Fig. 8: Plot of the daily Covid-19 series in Saudi Arabia and fitted curve.

eigen-vector U_j, and π_j the last component of U_j $(j = 1, \ldots, r)$.

Consider the following matrix

$$\mathbf{\Pi} = \mathbf{U}^{\nabla}\mathbf{U}^{\nabla\mathbf{T}} + (\mathbf{1} - \nu^{\mathbf{2}})\mathbf{A}\mathbf{A}^{\mathbf{T}} \tag{8}$$

let us now define the linear operator:

$$\mathfrak{f}^{\nu} : \mathfrak{L}_r \to \mathbb{R}^L, \tag{9}$$

where $\mathfrak{L}_r = \text{span}\{U_1, \ldots, U_r\}$ and

$$\mathfrak{f}^{\nu}Y = \begin{pmatrix} \mathbf{\Pi}Y_{\triangle} \\ A^T Y_{\triangle} \end{pmatrix}, \; Y \in \mathfrak{L}_r, \tag{10}$$

where $Y_{\triangle}$ is the vector of the last $L-1$ elements of Y_N. The vector Z_j is defined as follows:

$$Z_i = \begin{cases} \widetilde{X_i} & \text{for} \quad j = 1, \ldots, K, \\ \mathfrak{f}^{\nu} Z_{j-1} & \text{for} \quad j = K+1, \ldots, K+h+L-1 \end{cases}$$

where the $\widetilde{X_i}$ are the reconstructed columns of the trajectory matrix of the ith series after grouping and leaving out noise components. Now, by constructing matrix $Z = [Z_1, \ldots, Z_{K+h+L-1}]$ and performing diagonal averaging, a new series $\hat{y}_1, \ldots, \hat{y}_{K+h+L-1}$ is obtained, where $\hat{y}_{N+1}, \ldots, \hat{y}_{N+h}$ from the h terms of the VSSA forecast.

As we discussed above, the best values for reconstruction and forecasting are $L = 7$ and $r = 2$. Similar procedures have been done for the new data that updated from 02-03-2020 to 12-05-2020. Same values of L and r were used in analysing the new data,

medRxiv preprint doi: https://doi.org/10.1101/2020.05.24.20111872; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

and used for prediction new data points. Fig. 9 presents the updated data and the reconstructed series by the first 2 eigentriples. It is obvious that the reconstructed series is obtained precisely. Fig. 10 shows two curves predictions and the whole actual data, the red one is the prediction using the first data, and the blue one is the predictions using the updated data. It is obvious that there is no big difference, as the peak by the red curve around May 20 where around beginning of June by the blue curve that used the updated data. In addition, and the end of this pandemic will be between mid of June and mid of July.

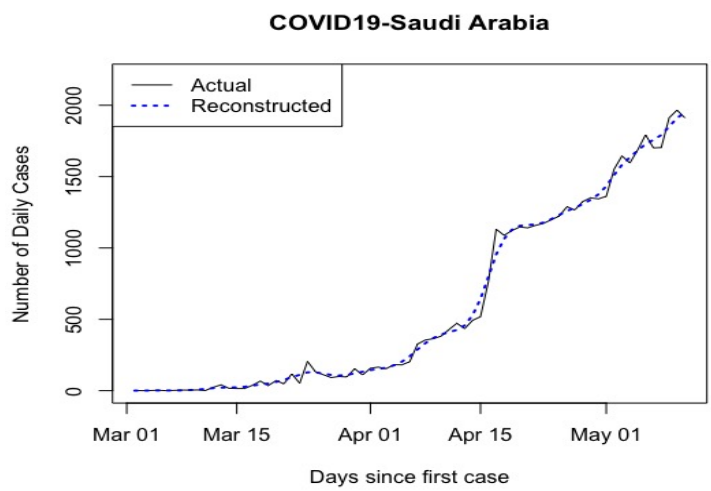

Fig. 9: Plot of the daily COVID-19 series in Saudi Arabia and fitted curve for the whole data.

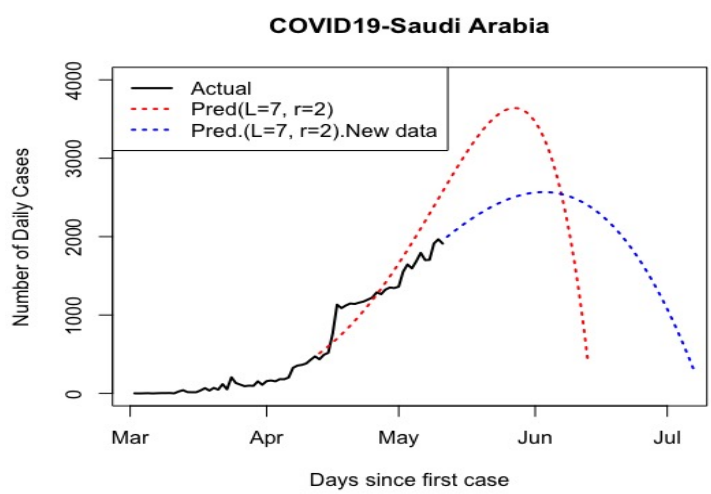

Fig. 10: Comparison of two forecasting scenarios with actual observations.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.24.20111872; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

5 Conclusion

A modified Singular Spectrum Analysis approach were used in this research for the decomposing and forecasting COVID-19 data in Saudi Arabia. The approach was examined in our previous research, and here in analysing COVID-19 data.

In the first stage, the first 42 confirmed daily values (02-03 to 12-04-2020) were used and analysed to identify the value of r for separability between noise and the signal. After obtaining the value of r, which was 2, and extracting the signals, the Vector SSA were used for prediction and determine the pandemic peak. In the second stage, we updated the data and included 71 daily values. We have used the same window length and number of eigenvalues for reconstruction and forecasting. The results of both forecasting scenarios have indicated that the peak will be around end of May and mid of June, and the end of this crises will be between end of June and mid of July.

All our results confirm the impressive performance of the modified SSA in analysing COVID-19 data and selecting the value of r for identifying the signal subspace from a noisy time series, and then make a good prediction using Vector SSA method. Note that we have not examined all possible values of window length in this research, and for forecasting we have used only the basic Vector SSA.

For future research, we will include more data and considered different window length L that may give a better forecasting. In addition, chaotic behaviour in COVID-19 data will be examined as we have some results that show strange patterns, which can be found in chaotic systems.

References

[1] Hassani, H., Alharbi, Nader., and Ghodsi, M. (2015). A study on the empirical distribution of the scaled Hankel matrix eigenvalues. *Journal of Advanced Research* 6 (6) (2015) 925–929.

[2] N. Alharbi, H. Hassani. A New approach for selecting the number of the eigenvalues in singular spectrum analysis. *Journal of the Franklin Institute*, 353 (1) (2015) 1–16.

[3] N. Alharbi. A novel approach for removing noise from EEG signal. *Biomedical Signal Processing and Control*, 39 (2018) 23–33.

[4] https://www.who.int/emergencies/diseases/novel-coronavirus-2019

[5] Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet*, 395 (2020) 565–574.

[6] Zhu, N., Zhang, D., Wang, W., et al; China Novel Coronavirus Investigating and Research Team. A novel coronavirus from patients with pneumonia in China 2019. *N Engl J Med*. doi:10.1056/NEJMoa2001017

medRxiv preprint doi: https://doi.org/10.1101/2020.05.24.20111872; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

[7] Zhou, G., Chen, S. and Chen, Z. Back to the spring of Wuhan: facts and hope of COVID-19 outbreak. Front. Med. (2020). https://doi.org/10.1007/s11684-020-0758-9

[8] https://www.worldometers.info/coronavirus/.

[9] Daley D.J., Gani J. *Epidemic modelling: an introduction*. Cambridge University Press, 2001.

[10] Hethcote H. W. The mathematics of infectious diseases. SIAM review. 42 (4) (2000) 599–653.

[11] Ferguson, M.N. et al. Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-9-impact-of-npis-on-covid-19/

[12] Hassani, H., Rahim, M., Hardi, O., Silva, E. A Preliminary Investigation into the Effect of Outlier(s) on Singular Spectrum Analysis. *Fluctuation and Noise Letters*. 13 (4) (2014) 1–14.

[13] Hassani, H., Mahmoudvand, R. *Singular Spectrum Analysis Using R*; Palgrave Macmillan: Basingstoke, UK, 2018.

[14] Movahedifar, M., Yarmohammadi, M., Hassani, H. Bicoid signal extraction: Another powerful approach. *Math. Biosci*. 303 (2018) 52–61. [CrossRef] [PubMed]

[15] Ghodsi, Z., Silva, E.S., Hassani, H. Bicoid Signal Extraction with a Selection of Parametric and Nonparametric Signal Processing Techniques. *Genom. Proteom. Bioinform*. 13 (2015) 183–191. [CrossRef] [PubMed]

[16] S anei, S., Hassani, H. *Singular Spectrum Analysis of Biomedical Signals*; Taylor and Francis/CRC: Boca Raton, FL, USA, 2016.

[17] Safi, S.M.M., Pooyan, M., Nasrabadi, A.M. Improving the performance of the SSVEP-based BCI system using optimized singular spectrum analysis (OSSA). *Biomed. Signal Proces. Control*, 46 (2018) 46–58.

[18] Muruganatham, B., Sanjith, M.A., Krishnakumar, B., Satya Murty, S.A.V. Roller element bearing fault diagnosis using singular spectrum analysis. *Mech. Syst. Signal Process*, 35 (2013) 150–166.

[19] Liu, K.; Law, S.S., Xia, Y., Zhu, X.Q. Singular spectrum analysis for enhancing the sensitivity in structural damage detection. *J. Sound Vib*, 233 (2014) 392—417.

[20] Hassani, H., Rua, A., Silva, E.S., Thomakos, D. Monthly forecasting of GDP with mixed-frequency multivariate singular spectrum analysis. *Int. J. Forecast*, 35 (2019) 1263–1272.

[21] Carvalho, M., Rodrigues, P.C., Rua, A. Tracking the US business cycle with a singular spectrum analysis. *Econ. Lett*, 114 (2012) 32–35.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.24.20111872; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

[22] Broomhead, D., King, G. Extracting qualitative dynamics from experimental data. *Physica D* ,20 (1986) 217–236.

[23] Broomhead, D., King, G. On the qualitative analysis of experimental dynamical systems. In Nonlinear Phenomena and Chaos; Sarkar, S., Ed.; Adam Hilger: Bristol, UK, (1986) 113–144.

[24] N. Golyandina, V. Nekrutkin, A. Zhigljavsky. *Analysis of time series structure: SSA and related techniques*. Chapman and Hall/CRC, New York, 2001.

[25] N. Golyandina, A. Zhigljavsky. *Singular spectrum analysis for time series*. Springer briefs in statistics. Springer, Verlag Berlin Heidelberg, 2013.

[26] Golyandina, N. Particularities and commonalities of singular spectrum analysis as a method of time series analysis and signal processing. arXiv 2019, arXiv:1907.02579v1.

[27] Kalantari, M., Hassani, H. Automatic Grouping in Singular Spectrum Analysis.*forcasting*, 1 (2019) 189–204.

[28] Elsner, J. B., and Tsonis, A. A. (1996).*Singular spectrum analysis: A new tool in time series analysis*. New York: Plenum Press.

[29] H. Hassani, M. Mahmoudvand, M. Zokaei, M. Ghodsi. On the separability between signal and noise in singular spectrum analysis. *Fluctuation and Noise Letters*. 11 (2) (2012) 14–25.

[30] H. Hassani, N. Alharbi, M. Ghodsi. Distinguishing chaos from noise: A new approach. *International Journal of Energy and statistics*. 2 (2) (2014) 137–150.

[31] N. Alharbi, Z. Ghodsi, H. Hassani. Noise correction in gene expression data: A new approach based on subspace method. *Mathematical Methods in the Applied Sciences*. 39 (13) (2016) 3750–3757.

[32] https://covid19.moh.gov.sa.