Statistics and Its Interface Volume 16 (2023) 217–225

Modified recurrent forecasting in singular spectrum analysis using Kalman filter and its application for Bicoid signal extraction

Reza Zabihi Moghadam, Masoud Yarmohammadi, and Hossein Hassani*

One of the important topics in Drosophila melanogaster is statistical analysis of bicoid protein gradient. The bicoid protein gradient plays an important role in the segmentation stage of embryo development in the head and thorax and also has considerable noise. Therefore, it has been considered by many researchers. In this paper the state space model and Kalman filter algorithms are used for noise elimination and smoothing bicoid gene expression. The state-space allows the unobserved variables, each with a specific interpretation, to be included in the estimate with the observed model and can be analyzed using the Kalman filter algorithm. Then, the less noise bicoid gene expression are used for forecast by singular spectrum analysis (SSA) method. The results with strong evidence indicate that the proposed method can be considered as a powerful technique in the analysis and prediction of gene expression measurements.

Keywords and phrases: Forecasting, Kalman filter, Singular spectrum analysis, State space form, Recurrent forecasting, Bicoid, Drosophila melanogaster.

1. INTRODUCTION

This belief has been widely accepted that the pattern of morphogen products plays a very important role in the process of developing a simple fertilized cell to a complex multicellular organism. One of the most important of morphogens is the bicoid, first identified by Nsslein-Volhard in 1988 [1]. During the ovogenesis bicoid mRNA is localised at the anterior end of the egg. After fertilization, bicoid translation begins, and as a result, the bicoid protein, which is distributed in the anterior-posterior (A-P) axis of the egg, form a concentration gradient that determines most aspects of head and thorax development [2]. Extensive studies show that bicoid morphogen plays an important role in developing of the anterior structure of Drosophila. For example, Frohnhfer et al. [3] showed that embryos receiving various doses of bicoid protein have differently sized anterior structures, Figure 1, and in the absence of bicoid, anterior structures of body are replaced with posterior regions.

*Corresponding author.

Newadays, there exist many datasets in the field of genetics and expression measurement and also there are various parametric and non-parametric methods and techniques for analyzing this dataset [4, 5, 6]. Historically, these datasets have been analyzed using parametric methods. [7, 8] But most of these parametric methods require a limiting condition such as the stationary assumption, which has led to more use of non-parametric methods. One of these non-parametric methods that has recently been considered by many researchers in the field of genetics is the SSA method [9, 10]. For example, Zara Ghodsi et al. [11] used six different parametric and non-parametric methods autoregressive integrated moving average (ARIMA), autoregressive fractionally integrated moving average (ARFIMA), exponential smoothing (ETS), neural networks (NN), synthesis diffusion degradation (SDD) and SSA to determine the most efficient method in the analysis of bicoid genetic data. Their results show the superiority of the SSA method over other methods using the Root Mean Squared Error criteria (RMSE). Also Hassani et al. [12] introduced a modified version of the single spectrum analysis to filter and extract the expression signal of the bicoid gene and showed that it is more efficient than the original SSA. Another comprehensive description of the theoretical and practical aspects of SSA for bicoid signal extraction with several examples can be found [13, 14].

The SSA method is a powerful nonparametric technique which has both filtering and forecasting capabilities, and it is useful for univariate or multivariate time series data. Unlike standard methods such as ARIMA methods, the SSA method does not require the assumption of stationary and linearity, and since most data are non-stationary, the SSA method is more general and may be more relevant to a particular situation. The introduction of SSA dates back to work by Broomhead and King [15, 16] in 1986. Since then, there have been different attempts to improve and using this method in time series analysis including meteorology, marine science, medicine, signal processing, and econometrics. While reviewing all the work done in the SSA method is beyond the scope of this paper, those interested can refer to the cases [17, 18, 19, 20, 21, 22, 23, 24, 25]. A complete explanation of the SSA technique can be found in the books by [26, 27, 28].

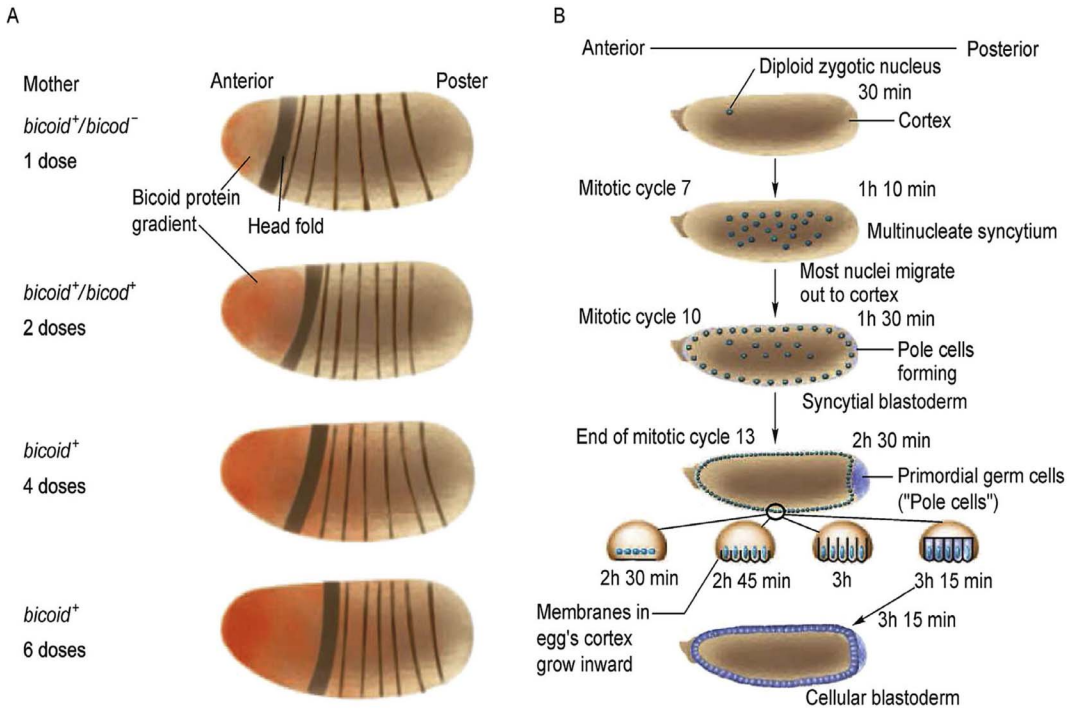

Figure 1. The BCD protein is a morphogen. (A) various doses of bcd affects the size of the head in Drosophila embryo. (B) Nuclear divisions in 14 cleavage cycles produce a syncytial blastoderm. Figures adapted from [3] with permission.

This study extends the application of SSA method into the field of Biology and attempts to predict Bicoid gene expression using the SSA method. The SSA method first breaks down the observations of bicoid gene expression into two components of noise and signal (noise-free). Then it calculates the forecast by reconstructing the original series using the signal component and a recursive linear relationship from the original series. Since bicoid gene expression data are often associated with significant noise, the forecasts are also contaminated with noise which subsequently lead to a reduction in forecast accuracy. Therefore, to improve the performance of recursive prediction for bicoid gene expression, this paper intends to first generate a noise less series using the space state equations and the Kalman filter algorithm, and then to advance new observations using the SSA recursive prediction method.

The rest of this paper organized as follows. Section 2 presents a review of the state-space model and the Kalman filter algorithm. Section 3 presents a short description of the SSA and SSA-R algorithm. In section 4 we introduce the newly SSA-R forecasting method based on the Kalman filter algorithm. In Section 5, we will present empirical results and the efficiency of new technique is compared with the original SSA-R method via simulation studies. We will present the conclusions in Section 6.

2. A SHORT DESCRIPTION OF THE STATE-SPACE AND KALMAN FILTER ALGORITHM

State-space models are broad models that include many linear and nonlinear models, and first introduced by Kalman [29] and Kalman and Bucy [30]. The goal of the state space is to infer information about the states, given the observations, as new data arrives. A well-known algorithm for performing this method is the Kalman filter, which does not require stationary and inversions. More detailed information on the theory of the state-space model and Kalman filter can be found in [31, 32].The state-space model and kalman filter algorithm are concisely presented below, and in doing so we mainly follow [33].

In brief, a state-space model for time series observation $\{y_t : t = 1, \ldots, N\}$ includes a measurement equation and a transition equation relating in the measurement equation of the observed data to a state vector, and this state vector in the transfer equation is obtained as a markovian process. The state-space equations for $t = 1, \ldots, N$ have the following form:

$$y_t = Z_t \alpha_t + \epsilon_t, \qquad \epsilon_t \sim N(0, \sigma_\epsilon^2) \tag{1}$$

$$\alpha_t = \mathbf{T}_t \alpha_{t-1} + \eta_t, \qquad \eta_t \sim N(0, \mathbf{Q}_t) \tag{2}$$

where Z_t is an $1 \times m$ matrix, $\mathbf{T}_t$ is an $m \times m$ transition matrix, α_t is an $m \times 1$ state vector, $\alpha_0 \sim N(a_0, P_0)$ is the

initial state. Note that ϵ_t and η_t represent disturbances in measurement equation and transition equation, respectively and they are mutually uncorrelated variables and are also uncorrelated with the initial state. In state space equations if matrices Z_t, $\mathbf{T}_t$ and $\mathbf{Q}_t$ are constant with respect to time, the model is called time invariant. In this models, the parameters are usually easily estimated, so in many applications the time invariant model is used.

To provide a simple illustration of the results of this section, consider model

$$y_t = \mu_t + \epsilon_t, \qquad \epsilon_t \sim N(0, \sigma_\epsilon^2)$$

where μ_t is the trend component and ϵ_t is the disturbance component. The component μ_t is taken to be locally linear

$$\begin{aligned} \mu_t &= \mu_{t-1} + \beta_{t-1} + \eta_t, \quad & \eta_t &\sim N(0, \sigma_\eta^2) \\ \beta_t &= \beta_{t-1} + \zeta_t, & \zeta_t &\sim N(0, \sigma_\zeta^2) \end{aligned}$$

where β_t is slope, η_t and ζ_t are disturbances that are mutually uncorrelated and are also uncorrelated with the disturbance ϵ_t. In the state space form of this model, for $t = 1, \ldots, N$

$$\alpha_t = \begin{pmatrix} \mu_t \\ \beta_t \end{pmatrix}, \qquad Z_t = \begin{pmatrix} 1 & 0 \end{pmatrix}$$
$$\mathbf{T} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{Q} = \begin{pmatrix} \sigma_\eta^2 & 0 \\ 0 & \sigma_\zeta^2 \end{pmatrix}$$

This model is referred to as a local linear trend.

Kalman filter

Kalman filter is the most famous algorithm for analyzing state-space equations. In the Gaussian state-space model, the Kalman filter is a set of recursive equations to obtain the estimates of the vector α_t, filtering and predicting time series data based on observations $Y_t = \{y_1, ..., y_t\}$. The filter includes two sets of predicting and updating equations.

In order to perform Kalman filter algorithm, first let $Y_t = \{y_1, \ldots, y_t\}$, where y_j are the observed variable, $a_{t-1} = E[\alpha_{t-1}|Y_{t-1}]$ is optimal estimator of α_{t-1} and $\mathbf{p}_{t-1} = E[(\alpha_{t-1} - a_{t-1})(\alpha_{t-1} - a_{t-1})|Y_{t-1}]$ is MSE matrix of a_{t-1}. Therefore by having a_{t-1} and $\mathbf{p}_{t-1}$, the Kalman filter is, for $t = 1, \ldots, N$,

$$a_{t|t-1} = \mathbf{T}_t a_{t-1|t-1}, \tag{3}$$

$$\mathbf{p}_{t|t-1} = \mathbf{T}_t \mathbf{p}_{t-1|t-1} \mathbf{T}_t' + \mathbf{Q}_t, \tag{4}$$

$$\hat{y}_{t|t-1} = Z_t a_{t|t-1}, \tag{5}$$

$$\begin{aligned} \upsilon_t &= Z_t(\alpha_t - a_{t|t-1}) + \epsilon_t, \\ \mathbf{F}_t &= Var(\upsilon_t|Y_{t-1}) \\ &= Z_t \mathbf{p}_{t|t-1} Z_t' + \sigma_\epsilon^2, \end{aligned} \tag{6, 7}$$

$$a_{t|t} = a_{t|t-1} + \mathbf{p}_{t|t-1} Z_t' \mathbf{F}_t^{-1} \upsilon_t, \tag{8}$$

$$\mathbf{p}_{t|t} = \mathbf{p}_{t|t-1} - \mathbf{p}_{t|t-1} Z_t' \mathbf{F}_t'^{-1} Z_t \mathbf{p}_{t|t-1}' \tag{9}$$

Where equations (3) and (4) are prediction equations, equation (5) is the optimal predictor of y_t (the value of y_t is obtained free-noise), equations (6) and (7) are the prediction error and its associated MSE matrix, respectively and equations (8) and (9) are updating equations.

3. A SHORT DESCRIPTION OF THE SSA

The SSA is a nonparametric method for data analysis that can break down series into several components and predict. This method involves decomposition and reconstruction stages, each of which includes two separate steps. The Basic SSA method is briefly presented below, and in doing so, we mainly follow [34, 26].

Stage 1. Decomposition

Decomposition stage includes two steps: embedding and Singular Value Decomposition (SVD).

Step 1. Embedding

At the first step, first the time series $Y_N = \{y_1, ..., y_N\}$ is organized into the matrix $\mathbf{X} = \{X_1, ..., X_K\}$, where $X_i = (y_i, ..., y_{i+L-1})' \in \mathbb{R}^L$, L is the window length and $2 \leq L \leq N/2$ and $K = N - L + 1$. In this procedure, matrix $\mathbf{X}$ is called trajectory matrix, which is also a Hankel matrix.

Step 2. Singular Value Decomposition (SVD)

In the next step of decomposition stage, the trajectory matrix $\mathbf{X}$ is broken into the sum of rank-one matrices. If $\lambda_1, ..., \lambda_L$, $U_1, ..., U_L$ be the eigenvalues, eigenvectors of the matrix $\mathbf{X}\mathbf{X}'$, respectively and $d = max\{i, \; such \; that \; \lambda_i > 0\} = rank(\mathbf{X})$, then the SVD of the trajectory matrix can be written as $\mathbf{X} = \mathbf{X}_1 + \ldots + \mathbf{X}_d$, where $\mathbf{X}_i = \sqrt{\lambda_i} U_i V_i'$ and $V_i = \mathbf{X}' U_i / \sqrt{\lambda_i} (i = 1, \ldots, d)$.

Stage 2. Reconstruction

Reconstruction stage includes two steps: grouping and diagonal averaging.

Step 1. Grouping

In this step, the elementary matrices $\mathbf{X}_i$ are separated into several groups and sums the matrices within each group. In this step the aim is the signal and noise components to be distinguished. If signal group of indices $i_1, \ldots, i_r$ is denoted by $I_r = \{i_1, \ldots, i_r\}$, then the matrix $\mathbf{X}_{I_r}$ corresponding to the group I_r is defined as $\mathbf{X}_{I_r} = \mathbf{X}_{i_1} + \ldots + \mathbf{X}_{i_r}$.

Step 2. Diagonal Averaging

In the diagonal averaging step, first the matrices obtained from the grouping step are converted into a Hankel matrix and then they are transformed to a time series.

3.1 SSA method parameters

There are two important parameters to the SSA method. The first parameter is the window length (L), which is required in the embedding step to form the trajectory matrix. Improper selection of this parameter will lead to improper time series decomposition and grouping. Unfortunately, there is no single way to determine L, but there is a set of general principles and rules that have good theoretical and practical support and can help you choose the

right L parameter. The second parameter that plays a key role in signal reconstruction over a time series is the number of eigentriples that explain the signal and are used in its reconstruction. This parameter is usually denoted by r and is called the reconstruction parameter. The information contained in eigenvalues and eigenvectors can be used to determine r. For more information on choosing the L and r parameters, can refer to [26, 28].

The SSA method has the capability of generate predictions after performing decomposition and reconstruction steps using the two forecasting methods Recurrent (SSA-R) and Vector (SSA-V) [27]. In what follows, forecasting with SSA is done with a greater focusing on SSA-R.

3.2 Recurrent SSA forecasting (SSA-R)

In order to obtion SSA-R forecasts for the time series $Y_N = \{y_1, \ldots, y_N\}$, suppose that $I_r = \{i_1, \ldots, i_r\}$ be the selected set of eigentriples from the signal group, $U_i \in \mathbb{R}^L$ and $i \in I_r$ be the corresponding eigenvectors, $\underline{\mathrm{U}}_i \in \mathbb{R}^{L-1}$ be the vector including of the first $L-1$ components of the vector U_i, π_i be the last component of the vector U_i, $\nu^2 = \sum_{i\in I_r} \pi_i^2$, and $\tilde{Y}_N = \{\tilde{y}_1, \ldots, \tilde{y}_N\}$ be the series reconstructed by I_r. Therefore SSA-R forecasts can be presented as follows:

$$z_i = \begin{cases} \tilde{y}_i, & i = 1, \ldots, N \\ \sum_{j=1}^{L-1} \phi_j z_{i-j}, & i = N+1, \ldots, N+h. \end{cases}$$

where $z_{N+1}, ..., z_{N+h}$ are the h step ahead the SSA-R method forecasts and $\phi_1, \ldots, \phi_{L-1}$ are called linear recurrent relation (LRR) coeffcients and be computed as follows:

$$R = (\phi_{L-1}, \ldots, \phi_1)' = \frac{1}{1-\nu^2} \sum_{i\in I_r} \pi_i \underline{\mathrm{U}}_i$$

Next, we introduce the newly suggested SSA-R forecasting method.

4. NEW SSA-R FORCASTING USING KALMAN FILTER ALGORITHM

Suppose the initial model for observations $\{y_t : t = 1, \ldots, N\}$ is

$$y_t = s_t + \epsilon_t \tag{10}$$

where s_t and ϵ_t are the signal and the noise components, respectively. Therefore if $\mathbf{X}$ is $L \times K$ trajectory matrix for observations $\{y_t : t = 1, \ldots, N\}$, it is clear that:

$$\mathbf{X} = \mathbf{S} + \boldsymbol{\epsilon} \tag{11}$$

where $\mathbf{S}$ and $\boldsymbol{\epsilon}$ represent $L \times K$ trajectory matrices for these components.

For any time series with constant window length L, there are $L-1$, SSA-R coefficient $\phi_1, \ldots, \phi_{L-1}$, which are obtained from eigenvectors of $\mathbf{X}\mathbf{X}'$. Therefore if the observational series includes noise, it is evident that the eigenvectors obtained from the trajectory matrices are also contaminated, and the estimated coefficients $\phi_1, \ldots, \phi_{L-1}$ may not be very accurate. Using these incorrect coefficients play an important role in forecasting and will reduce the accuracy of the prediction. We will use the state-space equation and Kalman filter algorithms to obtain data with less noise and improve our prediction using the SSA-R method. With this idea, we will define a new prediction method of SSA based on the Kalman filter (KF-SSA-R).

Let $\{\hat{y}_t : t = 1, \ldots, N\}$ be a less-noise time series generated by the Kalman filter equation (5), $\tilde{\mathbf{X}}$ be the trajectory matrix of less-noise series and $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_L$ and $\tilde{U}_1, \ldots, \tilde{U}_L$ are the eigenvalues and eigenvectors of $\tilde{\mathbf{X}}\tilde{\mathbf{X}}'$, respectively. Let I_r be the selected set of eigentriples, then the coefficients of the KF-SSA-R method are:

$$R = (\tilde{\phi}_{L-1}, \ldots, \tilde{\phi}_1)' = \frac{1}{1-\tilde{\nu}^2} \sum_{i\in I_r} \tilde{\pi}_i \underline{\tilde{\mathrm{U}}}_i \tag{12}$$

where $\underline{\tilde{\mathrm{U}}}_i$ is the vector including of the first $L-1$ components of the vector $\tilde{U}_i$, $\tilde{\pi}_i$ is the last component of the vector $\tilde{U}_i$ and $\tilde{\nu}^2 = \sum_{i\in I_r} \tilde{\pi}_i^2$. The h step ahead forecast for the KF-SSA-R method is:

$$z_i = \begin{cases} \tilde{\tilde{y}}_i, & i = 1, \ldots, N \\ \sum_{j=1}^{L-1} \tilde{\phi}_j z_{i-j}, & i = N+1, \ldots, N+h. \end{cases} \tag{13}$$

$\tilde{\tilde{y}}_i, i = 1, ..., N$ are the series reconstructed by I_r.

In the next section, in order to evaluate the results of the KF-SSA-R forecasting method, we compare the forecast accuracy of orginal SSA-R and KF-SSA-R methods using the Root Mean Squared error criteria (RMSE).

5. EMPIRICAL RESULTS

In this section, we compare the performance of original SSA-R and KF-SSA-R forecasting methods using simulated time series and real data based on the RMSE criteria. The time series $Y_N = \{y_1, \ldots, y_N\}$ is divided into two parts: a training set and a test set. To compare the SSA-R and KF-SSA-R forecasting methods, we will find the ratio of RMSE given by:

$$\begin{aligned} RRMSE_h &= \frac{RMSE_h(\text{KF-SSA-R})}{RMSE_h(\text{original SSA-R})} \\ &= \frac{(\sum_{t=M}^{N-h}(y_{t+h} - \hat{y}_{t+h|t})^2)^{1/2}}{(\sum_{t=M}^{N-h}(y_{t+h} - \hat{\hat{y}}_{t+h|t})^2)^{1/2}}, \end{aligned} \tag{14}$$

Where N is the length of time series data, M is the number of observations in the training set and h is the forecast horizon. On the other hand $\hat{y}_{t+h|t}$ and $\hat{\hat{y}}_{t+h|t}$ are the h-step ahead forecast generated by KF-SSA-R and SSA-R, respectively. If the $RRMSE_h < 1$, then the KF-SSA-R procedure

performs better than SSA-R. It should be noted that to perform calculations related to SSA method and Kalman filter, R software programming environment and Rssa and stats packages have been used.

5.1 Simulation studies

In the following simulation studies, 150 data points were generated using different models and normally distributed noise, then we consider 100 observations as the training sample, i.e $M = 100$, to obtain less noisy data by using State space equations and Kalman filter alghorithm in the Kf-SSA-R method. In both methods, the number of eigenvalues for reconstruction and forecasting (r) were obtained based on the rank of the trajectory matrix. The simulation for each of the models were repeated 1000 times and the average of RRMSEs were computed. In order to evaluate the effect of noise levels on forecasting results, different levels of signal to noise variance ratios (SNR) for various forecast horizons were used as $SNR = 0.125, 0.5, 1, 1.5, 10$.

Example 5.1. **Sin series**
Consider the Sin series for $t = 1, 2, \ldots, 150$:

$$y_t = sin(\frac{\pi t}{6}) + \epsilon_t, \qquad \epsilon_t \sim N(0, \sigma_\epsilon^2),$$

where $sin(\frac{\pi t}{6})$ is the signal component and ϵ_t is the noise component. In both methods SSA-R and KF-SSA-R, based on the rank of the trajectory matrix, the first two eigenvalues were chosen for reconstruction and forecasting ($r = 2$). In addition, structural models using the StructTS option in R software were used to obtain low-noise data in the Kf-SSA-R method. Figure 2 shows the RRMSE for different values of SNR for different lengths of forecast horizons $h = 1, 3, 6, 12$ of the Sin series. Based on the RRMSE results obtained from this figure, it can be concluded that the KF-SSA-R method forecasting performs better than the SSA-R method, especially when the values of the SNR are low and also the values of window length (L) are low. It can also be seen that there are no differences between RRMSEs for various forecast horizons h. In Table 1, the RRMSE values are presented the ratio of the minimum RMSE of the KF-SSA-R method over all possible window lengths to the minimum RMSE of the SSA-R method over all possible window lengths for various values of forecast horizons h. Based on the results obtained from this table, it can be concluded that the SSA-R method forecasting using Kalman filter performs better than the performance of SSA-R method.

Example 5.2. **Deterministic linear trend model**
As a second example, consider the local linear trend model for $t = 1, 2, \ldots, 150$:

$$\begin{aligned} y_t &= \mu_t + \epsilon_t, & \epsilon_t &\sim N(0, \sigma_\epsilon^2) \\ \mu_t &= \mu_{t-1} + \beta_{t-1} + \eta_t, & \eta_t &\sim N(0, \sigma_\eta^2) \\ \beta_t &= \beta_{t-1} + \zeta_t, & \zeta_t &\sim N(0, \sigma_\zeta^2) \end{aligned}$$

Table 1. RRMSE for the ratio of the minimum RMSE of the KF-SSA-R method over all possible window lengths to the minimum RMSE of the SSA-R method over all possible window lengths

h	SNR	RRMSE
1	10	0.83
	1.5	0.79
	1	0.67
	0.5	0.62
	0.125	0.59
3	10	0.83
	1.5	0.78
	1	0.68
	0.5	0.62
	0.125	0.60
6	10	0.84
	1.5	0.79
	1	0.68
	0.5	0.63
	0.125	0.59
12	10	0.82
	1.5	0.79
	1	0.68
	0.5	0.63
	0.125	0.60

where $\mu_0 = 2.5$, $\beta_0 = 0.005$, ϵ_t is the disturbance component, η_t and ζ_t are disturbances that $\sigma_\eta^2 = \sigma_\zeta^2 = 0$. This model are called deterministic linear trend model. For both methods, based on the rank of the trajectory matrix, the first two eigenvalues were chosen for reconstruction and forecasting ($r = 2$). In addition, structural models using the StructTS option in R software were used to obtain low-noise data in the Kf-SSA-R method. Figure 3 shows the RRMSE for different values of SNR for different lengths of forecast horizons $h = 1, 3, 6, 12$ of the deterministic linear trend models. Similar to sin model, Based on the RRMSE results obtained from this figure, it can be concluded that the KF-SSA-R method forecasting performs better than the SSA-R method, especially when the values of the SNR are low and also the values of window length (L) are low. It can also be seen that there are no differences between RRMSEs for various forecast horizons h. In Table 2, the RRMSE values are presented the ratio of the minimum RMSE of the KF-SSA-R method over all possible window lengths to the minimum RMSE of the SSA-R method over all possible window lengths for various values of forecast horizons h. Based on the results obtained from this table, it can be concluded that the SSA-R method forecasting using Kalman filter performs better than the performance of SSA-R method.

5.2 Real data

In this section, to appraise the efficiency of our presented algorithm, the efficiency of orginal SSA-R forecasting algorithm and SSA-R forecasting algorithm using Kalman filter

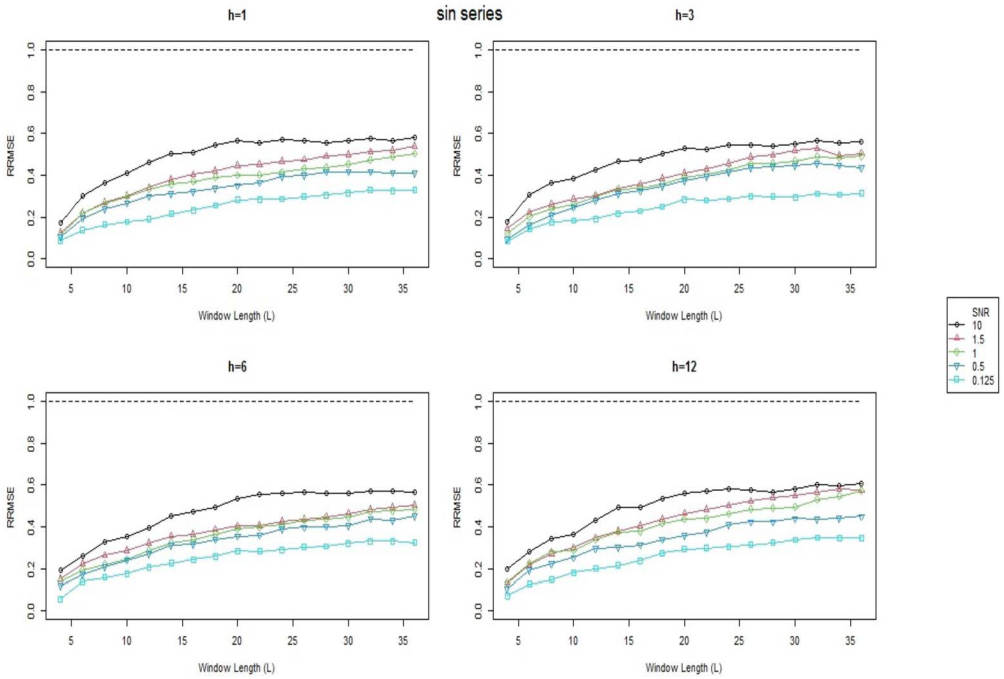

Figure 2. RRMSE for Sine series.

Table 2. RRMSE for the ratio of the minimum RMSE of the KF-SSA-R method over all possible window lengths to the minimum RMSE of the SSA-R method over all possible window lengths

h	SNR	RRMSE
1	10	0.85
	1.5	0.81
	1	0.73
	0.5	0.69
	0.125	0.63
3	10	0.85
	1.5	0.80
	1	0.73
	0.5	0.68
	0.125	0.62
6	10	0.84
	1.5	0.80
	1	0.72
	0.5	0.68
	0.125	0.64
12	10	0.86
	1.5	0.81
	1	0.74
	0.5	0.69
	0.125	0.62

are compared for real data. Five Drosophila melanogaster embryos introduced by Alexandrov et al. [14], which was originally obtained from FlyEx database [35, 36] are considered. The biological characteristics of these series and the statistical methods performed on them could be found in [11, 12, 36, 37, 38, 39, 40]. Figures 4 show the time series plots of these data sets.

In Table 3, the value of RRMSE are presented for various values of forecast horizons h. At each of forecast horizons h, different values of window length (L) and r have been employed. For both methods, the number of eigenvalues for reconstruction and forecasting (r) were obtained based on the rank of the trajectory matrix. In addition, structural models using the StructTS option in R software were used to obtain low-noise data in the Kf-SSA-R method. Based on the results obtained from this table, it can be concluded that the SSA-R method forecasting using Kalman filter performs better than the performance of SSA-R method, especially when the values of window length (L) are low. This is consistent with previous results in simulated series. In Table 4, the RRMSE values are presented the ratio of the minimum RMSE of the KF-SSA-R method over all possible window lengths to the minimum RMSE of the SSA-R method over all possible window lengths and the ratio of the minimum RMSE of the KF-SSA-R method over all possible window lengths to RMSE from forecasting equation in state space model for various values of forecast horizons h. Based on the results obtained from this table, it can be concluded that the SSA-R method forecasting using Kalman filter performs better than the performance of SSA-R method and state space model.

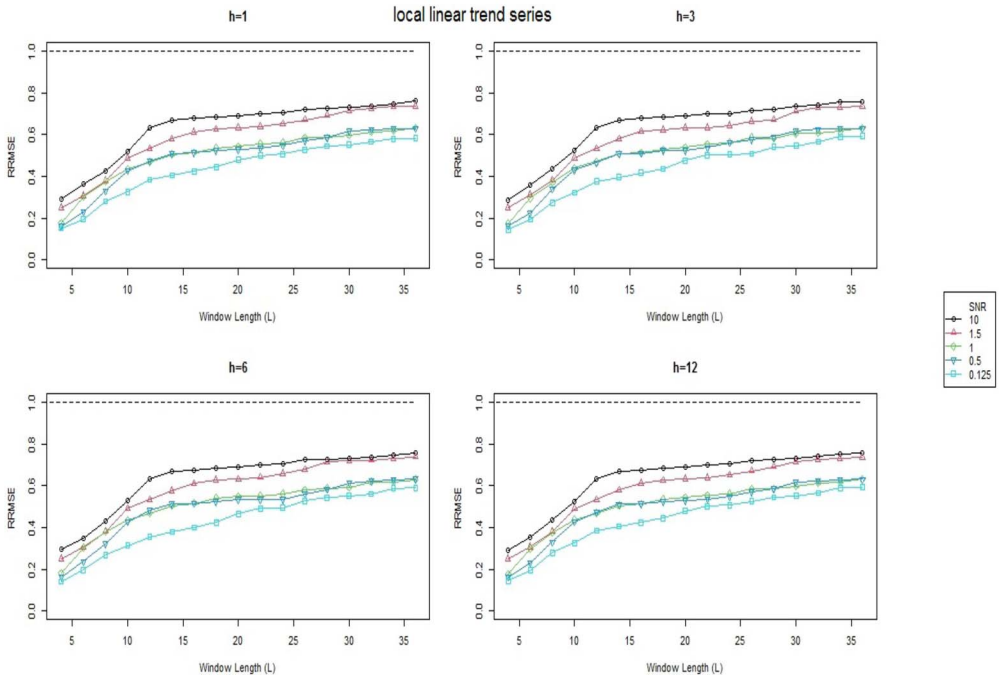

Figure 3. RRMSE for Deterministic linear trend series.

Table 3. RRMSE for real data of the Drosophila melanogaster embryos

Embryo	**h**	**RRMSE (L,r)**				
ab2	1	0.86 (14,3)	0.90 (16,3)	0.92 (36,2)	0.79 (12,2)	0.87 (32,5)
	3	0.79 (13,3)	0.87 (26,5)	0.92 (24,3)	0.95 (28,2)	0.75 (20,4)
	6	0.93 (30,5)	0.98 (33,4)	0.92 (27,2)	0.74 (13,3)	0.87 (35,3)
	12	0.77 (16,3)	0.80 (14,2)	0.91 (27,4)	0.94 (25,3)	0.91 (20,2)
ab7	1	0.96 (36,2)	0.98 (36,1)	0.96 (25,1)	0.87 (32,3)	0.89 (25,2)
	3	0.98 (26,1)	0.91 (30,2)	0.99 (32,1)	0.93 (36,2)	0.86 (26,3)
	6	0.98 (34,1)	0.95 (36,2)	0.90 (22,2)	0.86 (32,3)	0.91 (27,2)
	12	0.81 (27,2)	0.99 (34,1)	0.82 (34,2)	0.86 (31,3)	0.99 (36,1)
ab11	1	0.96 (36,2)	0.95 (36,3)	0.91 (16,2)	0.96 (32,2)	0.95 (30,4)
	3	0.97 (36,3)	0.88 (10,3)	0.97 (32,2)	0.97 (30,3)	0.96 (36,4)
	6	0.98 (36,2)	0.98 (36,3)	0.98 (29,2)	0.98 (33,3)	0.96 (34,4)
	12	0.94 (28,2)	0.93 (28,3)	0.96 (35,3)	0.95 (34,4)	0.95 (34,3)
ac30	1	0.98 (36,1)	0.98 (30,1)	0.97 (25,1)	0.98 (25,2)	0.98 (30,2)
	3	0.98 (29,1)	0.98 (36,1)	0.97 (36,2)	0.97 (27,2)	0.97 (32,2)
	6	0.97 (28,2)	0.98 (36,3)	0.98 (27,1)	0.97 (34,2)	0.99 (34,1)
	12	0.95 (36,1)	0.93 (35,3)	0.95 (33,2)	0.97 (28,1)	0.94 (36,2)
ad4	1	0.71 (30,2)	0.98 (86,2)	0.67 (24,2)	0.98 (78,2)	0.97 (70,2)
	3	0.67 (32,3)	0.98 (86,2)	0.98 (90,3)	0.98 (67,2)	0.70 (32,2)
	6	0.98 (67,2)	0.98 (77,2)	0.99 (80,1)	0.73 (43,3)	0.99 (77,1)
	12	0.97 (70,3)	0.98 (80,2)	0.98 (72,2)	0.97 (66,2)	0.97 (66,1)

6. CONCLUSION

In this paper, to predict the expression of bicoid gene using SSA method, a new approach based on the state-space equations and Kalman filter algorithms is suggested. The KF-SSA-R forecasting coefficients are then obtained from the less noise time series after filtering the original data. In a Monte Carlo simulation study, we investigated the impact of noise on the accuracy of forecast. We assessed the forecasting performance of the proposed denoised SSA method for different levels of signal to noise ratio (SNR) and for various forecast horizons based on the RMSE criteria. The results obtained from the simulation studies illustrate that the suggested method forecasting is better than the SSA-R, especially for lower values of the SNR and the window length

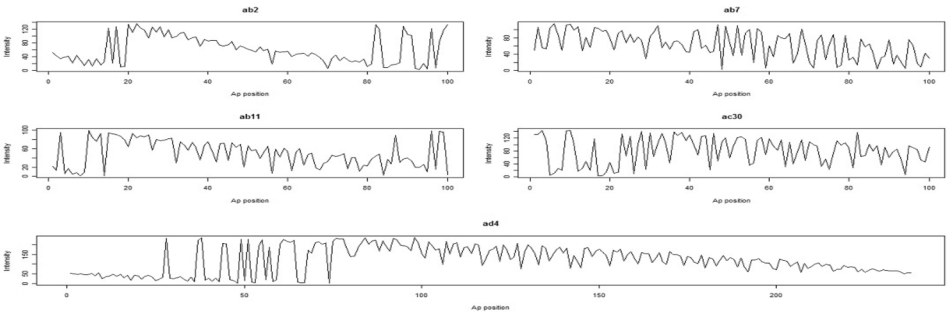

Figure 4. Time series plot of Drosophila melanogaster embryos data.

Table 4. RRMSE for ratio of the minimum RMSE of the KF-SSA-R method over all possible window lengths to the minimum RMSE of the SSA-R method over all possible window lengths

Embryo	**h**	**RRMSE (KF-SSA-R, SSA-R)(r)**	**RRMSE (KF-SSA-R, State space)**
ab2	1	0.89 (3)	0.87 (3)
	3	0.90 (2)	0.87 (2)
	6	0.90 (3)	0.87 (3)
	12	0.93 (4)	0.88(4)
ab7	1	0.89 (3)	0.88 (3)
	3	0.92 (2)	0.90 (2)
	6	0.89 (3)	0.88 (3)
	12	0.98 (1)	0.92(1)
ab11	1	0.95 (2)	0.93 (2)
	3	0.96 (3)	0.94 (3)
	6	0.96 (2)	0.94 (2)
	12	0.96 (3)	0.94(3)
ac30	1	0.97(1)	0.95 (1)
	3	0.96 (2)	0.94 (2)
	6	0.98 (1)	0.95 (1)
	12	0.96 (2)	0.94(2)
ad4	1	0.89 (3)	0.88 (3)
	3	0.89 (2)	0.88 (2)
	6	0.92 (1)	0.90 (1)
	12	0.89 (2)	0.88(2)

(L). Finally, the results show that the KF-SSA-R technique mentioned performs better than the SSA technique used to predict noisy bicoid and can be considered as a powerful method for analyzing the expression of bicoid gene data.

Received 23 May 2021

REFERENCES

[1] W. Driever, C. Nusslein-Volhard, The bicoid protein determines position in the drosophila embryo in a concentration dependent manner.(1988) 95–104.

[2] D. M. Holloway, L. G. Harrison, D Kosman, C.E. Vanario Alonso, Spirov AV. Analysis of pattern precision shows that Drosophila segmentation develops substantial independence from gradients of maternal gene products. Dev Dyn **235** (2006) 2949–60.

[3] http//highered.mheducation.com.

[4] P. G. S. M. Allison DB, Cui X, Microarray data analysis: from disarray to consolidation and consensus. Nat. Rev. Genet. (2006) 55–65.

[5] B.P. Goeman JJ, Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics (2007) 7–980.

[6] F. W. VanGuilder HD, Vrana KE, Twenty-five years of quantitative pcr for gene expression analysis. Biotechniques (2008) 26–619, https://doi.org/10.2144/000112776.

[7] Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES. Parametric and nonpara-metric linkage analysis: a unified multipoint approach. Am J Hum Genet **58**(6)(1996)1347.

[8] Grimm O, Coppey M, Wieschaus E. Modelling the bicoid gradient. Development(2010);**137**(14):2253–64.

[9] L. A. S. D. Y. H. Du LP, Wu SH, Spectral analysis of microarraygene expression time series data of Plasmodium falciparum, IJBRA**4** (2008) 49–337.

[10] Y. H. Tang VT, Noise reduction in microarray gene expression data based onspectral analysis, IJMLC **3** (2012) 7–51.

[11] Z. Ghodsi, E. S. Silva and H. Hassani, Bicoid Signal Extraction with a Selection of Parametric and Nonparametric Signal Processing Techniques, Genomics Proteomics Bioinformatics, **13**(3) (2015) 183–191.

[12] H. Hassani and Z. Ghodsi, Pattern recognition of gene expression with singular spectrum analysis, Med. Sci. (3) (2014) 127–139,

https://doi.org/10.3390/medsci2030127.
[13] H. Hassani, E.S. Silva, Z. Ghodsi, Optimizing bicoid signal extraction, Math. Biosci. 294 (2017) 46–56, https://doi.org/10.1016/j.mbs.2017.09.008. MR3732322
[14] T. Alexandrov, N. Golyandina and A. Spirov, Singular spectrum analysis of gene expression profiles of early Drosophila embryo: exponential-in-distance patterns, Res Lett Signal Process, (2008) 1–5.
[15] D.S. Broomhead, G.P. King, Extracting qualitative dynamics from experimental data, Phys. D, Nonlinear Phenom. **20** (1986) 217–236. MR0859354
[16] D.S. Broomhead, G.P. King, On the qualitative analysis of experimental dynamical systems, in: S. Sarkar (Ed.), Nonlinear Phenomena and Chaos, Adam Hilger, Bristol, (1986), pp. 113–144. MR0854700
[17] S. Aydin, H. M. Saraoglu and S. Kara, Singular Spectrum Analysis of Sleep EEG in Insomnia, Journal of Medical Systems, **35**(4) (2011) 457–461.
[18] K. L. Bail, J. M. Gipson and D. S. MacMillan, Quantifying the Correlation Between the MEI and LOD Variations by Decomposing LOD with Singular Spectrum Analysis, Earth on the Edge: Science for a Sustainable Planet International Association of Geodesy Symposia, **139** (2014) 473–477.
[19] M. Carvalho and A. Rua, Real-time nowcasting the US output gap: Singular spectrum analysis at work, International Journal of Forecasting, **33**(1) (2017) 185–198.
[20] H. S. Chao and C. H Loh, Application of singular spectrum analysis to structural monitoring and damage diagnosis of bridges, Structure and Infrastructure Engineering: Maintenance, Management, Life-Cycle Design and Performance, **10**(6) (2014) 708–727.
[21] H. Hassani, Z. Xu and A. Zhigljavsky, Singular spectrum analysis based on the perturbation theory, Nonlinear Analysis: Real World Applications, **12** (2011) 2752–2766. MR2813219
[22] K. Liu, S. S. Law, Y. Xia and X. Q. Zhu, Singular spectrum analysis for enhancing the sensitivity in structural damage detection, Journal of Sound and Vibration, **333** (2) (2014) 392–417.
[23] R. Mahmoudvand and P. C. Rodrigues, Missing value imputation in time series using singular spectrum analysis, International Journal of Energy and Statistics, **4**(1) (2016) 1650005.
[24] B. Muruganatham, M. A. Sanjith, B. Krishnakumar and S. A. V Satya Murty, Roller element bearing fault diagnosis using singular spectrum analysis, Mechanical Systems and Signal Processing, **35**(1-2) (2013) 150–166.
[25] R. Wang, H. G. Ma, G. Q. Liu and D. G. Zuo, Selection of window length for singular spectrum analysis, Journal of the Franklin Institute, **352** (2015) 1541–1560. MR3325504
[26] S. Sanei and H. Hassani, Singular Spectrum Analysis of Biomedical Signals, Taylor & Francis/CRC, (2016).
[27] N. Golyandina, V. Nekrutkin and A. Zhigljavsky, Analysis of Time Series Structure: SSA and Related Techniques, Chapman & Hall/CRC, (2001). MR1823012
[28] N. Golyandina and A. Zhigljavsky, Singular Spectrum Analysis for Time Series, Springer Briefs in Statistics, Springer, (2013). MR3024734
[29] R. E. Kalman, A new approach to linear filtering and prediction problems, Trans. ASME, J. of Basic Engineering, **83** (1960) 35–45. MR3931993
[30] R. E. Kalman and R. S. Bucy, New results in linear filtering and prediction theory, J. of Basic Engineering, Transactions ASME, D. **83** (1961) 95–108. MR0234760
[31] A. C. Harvey, Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press (1989).
[32] R. H. Shumway and D. S. Stoffer, Time Series Analysis and Application, 3rd ed, Springer, New York,(2011). MR2721825
[33] E. Zivot and J. Wang, Modeling Financial Time Series with S-PLUS, Springer-Verlag, New York (2003). MR2000944
[34] H. Hassani, Singular spectrum analysis: methodology and comparison, J. Data Sci. **5**(2) (2007) 239–257.
[35] M. Kozlov and E. Myasnikova, Method for spatial registration of the expression patterns of drosophila segmentation genes using wavelets, Comput. Technol. **5** (2000) 112–119.
[36] A. Pisarev, E. Poustelnikova, The quantitative atlas on segmentation gene expression at cellular resolution, Nucleic Acids Res.**37** (2009): D560-6.
[37] Z. Ghodsi, H. Hassani, M. Kalantari, and E. S. Silva, Estimation of protein diffusion parameters, Journal for the Rapid, (2018). https://doi.org/10.1002/sta4.192. MR3905849
[38] S. Surkova, D. Kosman, K. Kozlov, E. Myasnikova, A. A. Samsonova, A. Spirov and et al, Characterization of the Drosophila segment determination morphome. Dev Biol, **313** (2007):844–62.
[39] M. Movahedifara, M. Yarmohammadia and H. Hassani, Bicoid signal extraction: Another powerful approach, Math Bioscie, **303** (2018) 52–61. MR3836632
[40] E. Poustelnikova, A. Pisarev, M. Blagov, M. Samsonova and J. Reinitz, A database for management of gene expression data in situ. Bioinformatics, **20** (2004) 2212–21.

Reza Zabihi Moghadam
Department of Statistics, Payame Noor University, 19395-4697, Tehran, Iran
E-mail address: Rezazm63@gmail.com

Masoud Yarmohammadi
Department of Statistics, Payame Noor University, 19395-4697, Tehran, Iran
E-mail address: Yarmohammadi.mas@gmail.com

Hossein Hassani
Research Institute for Energy Management and Planning (RIEMP), University of Tehran, No. 9 Qods St, Tehran, Iran
E-mail address: hassani.stat@gmail.com