Estimation of Protein Diffusion Parameters

Zara Ghodsi[a], Hossein Hassani[b], Mahdi Kalantari[c] and Emmanuel Sirimal Silva[d*]

[a] *Bournemouth University, Bournemouth, UK.*

[b] *Research Institute for Energy Management and Planning, University of Tehran, no. 13, Qods St, Tehran, Iran*

[c] *Department of StatisticsPayame Noor University, 19395–4697, Tehran, Iran.*

[d] *Fashion Business School, London College of Fashion, University of the Arts London. UK.*

Abstract

Protein diffusion offers an essential and elegant mechanism for morphogen gradient formation. Morphogens are signalling molecules that emanate from a particular region of the cell and create a gradient which has an impact on most biological processes, cell signalling and embryonic development. Using a method that is based on Singular Spectrum Analysis, we estimate parameters introduced in the Synthesis Diffusion Degradation model which is a commonly applied model for a transcription factor known as Bicoid. Our findings, consistent with simulation results, indicate that the proposed method can be practically applied as an enhanced parameter estimation technique with reduced sensitivity to various levels of noise.

Keywords: Bicoid, Diffusion, Morphogen, Parameter Estimation, Singular Spectrum Analysis, Simulation, Noise Reduction.

1 Introduction

Many characteristics of cells and tissues are specified during development by concentration gradients of morphogens. These gradients are essential for describing the formation of reproducible patterns during ontogenetic development. One of the best-studied examples of a morphogen gradient is formed by *bicoid* gene [1, 2]. In *Drosophila melanogaster*, the anterior-posterior (A-P) concentration gradient of Bicoid (Bcd) plays a fundamental role as a transcription factor in determining key features of anterior segments of the body [1–3].

Since the discovery of Bcd, several pattern formation mechanisms have been introduced to characterise its concentration gradient; from simple models with few parameters to more complex models involving many regulated quantities [4]. In this regard, the Synthesis Diffusion Degradation (SDD) model is a comparatively simple and most commonly used model for explaining the Bcd concentration gradient. The SDD model was initially described by Driever and Nusslein-Volhard [1], but its name was coined by [5].

According to [4, 6], by considering $c(x,t)$ as the Bcd concentration at embryonic A-P axis, $0 < x < L$, the evolution of Bcd along the egg can be characterised by:

$$\frac{\partial c(x,t)}{\partial t} = s(x,t) + D\nabla^2 c(x,t) - k_{deg}c(x,t), \tag{1}$$

*As all authors contributed equally towards this research, author names appear in alphabetical order.

where s is the source function, D is the Bcd diffusion coefficient, and k_{deg} is the Bcd degradation rate. As discussed in [4], the equilibrium solution of Equation (1) is well known to be a decaying exponential:

$$Ae^{-x/\lambda}, \tag{2}$$

where λ is a constant length such that $\lambda = \sqrt{D/k}$. Assuming the concentration of Bcd at $x = 0$ is A, λ is the distance to the source at which the Bcd concentration has dropped to $1/e$ of the maximal value, Figure 1 [4].

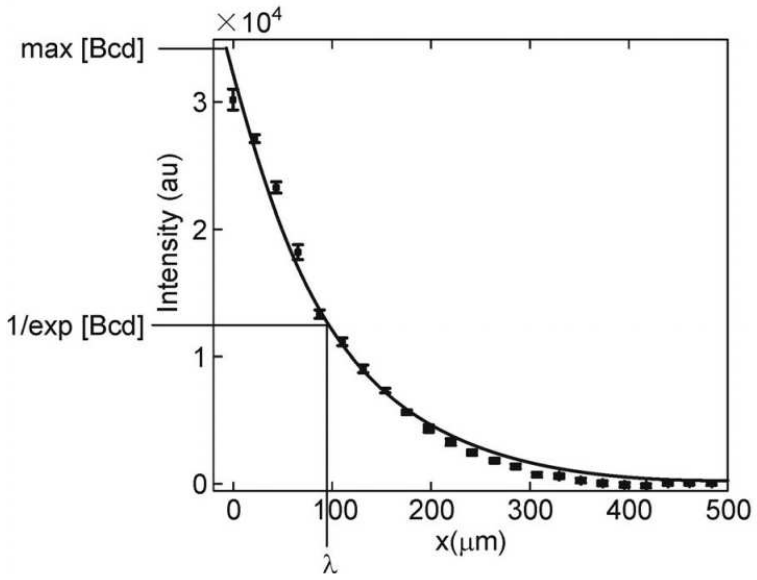

Figure 1: Calculating the length constant λ. Figure adopted from [12].

Although this model may be theoretically sound, there is considerable debate surrounding the ability of estimated parameters at accounting for much of the experimental data. It is well known that the estimation of A and λ are challenging tasks. However, it represents overly important tasks, simply because all future results from the SDD model are clearly based on these initial estimations, and therefore, if the initial estimations are inefficient, it implies that the future results too will be negatively affected. Thus, there is a strong need for research into measures which can lead to more efficient estimations of A and λ. Having identified this requirement, this study seeks to re-evaluate the estimation process of A and λ, which are the two most important parameters of the SDD model. Given the focus on parameter estimation, it is pertinent to note the promising signal processing approach which was developed and employed recently [8]. Here, the authors evaluated a variety of parametric and nonparametric signal processing techniques to identify the most efficient model for Bcd signal extraction. They found that an enhanced version of the nonparametric Singular Spectrum Analysis (SSA) approach produced the most efficient extraction. More importantly, they noted that denoising Bcd is a complex and arduous task. This is because, the data associates with both observational and biological noise, and the extracted residual is not normally distributed as required by parametric techniques [8]. Hence, from a practical point of view, it is relevant in such context to simultaneously filter the data and estimate the parameters of interest.

Accordingly, we propose a new method based on SSA which can generate more efficient estimations of Bcd parameters. The introduction of the SSA-based approach is motivated by several factors. First and foremost, Figure 1 illustrated how the SDD model follows an exponential pattern, and it is of great advantage to consider such a decaying exponential pattern. This leads to the second factor, which is that SSA is a general signal processing approach and

therefore a dedicated type of SSA that considers the exponential pattern alone would be ideal. Moreover, SSA has inherent filtering capabilities and is therefore less sensitive to noise. The theoretical aspects underlying the proposed SSA based approach are validated via a simulation study and through the evaluation over real data which considers all the cleavage cycles in which Bcd is present in the embryo.

Accordingly, we consider SDD as the benchmark given its well established nature within Bcd studies and also consider the SSA technique as a less noise sensitive alternative which has been proven to be useful with Bcd data. In particular, here we develop SSA for a typical exponential model in our attempts to improve the estimation of initial parameters and this is covered in detail in Section 2. Thereafter, in Section 3 we evaluate the accuracy and performance of the proposed SSA model against the SDD model via a series of simulations where the pattern and characterless of the data is considered. Then, real data is then used to demonstrate the viability of the proposed approach and to directly compare its performance with the well established SDD model. Finally, we present a concise summary in Section 4.

2 Estimating Values of A and λ

Let us now consider the new theoretical approach proposed to estimate A and λ. As mentioned above, we propose a tailored SSA approach. The SSA technique consists of two complementary stages called *Decomposition* and *Reconstruction*, each with two separate steps which are vital for signal extraction and noise filtering [13]. In brief, during the Decomposition stage, Bcd is decomposed into several interpretable components to distinguish between signal and noise. Thereafter, the Reconstruction stage is used to group the signal components together and construct a less noisy series. The basic SSA process is briefly explained below and in doing so we mainly follow [13].

The first step within the Decomposition stage is called embedding. Here, the time series $Y_N = \{y_1, \ldots, y_N\}$ is mapped into the vectors $X_1, \ldots, X_K$ where $X_i = (y_i, \ldots, y_{i+L-1})^T$, with L observations and $K = N - L + 1$. The single choice of this step is the *Window Length* L which is an integer, such that $2 \leq L \leq N - 1$. The output from the embedding step is the trajectory matrix $\mathbf{X} = [X_1 : \cdots : X_K]$ whose columns are the vectors X_i.

The second step in the Decomposition stage is called Singular Value Decomposition (SVD), and it is aimed at representing the trajectory matrix $\mathbf{X}$ as a sum of elementary matrices. The eigenvalues of $\mathbf{X}\mathbf{X}^T$ are denoted by $\lambda_1, \ldots, \lambda_L$ in decreasing order of magnitude ($\lambda_1 \geq \cdots \geq \lambda_L \geq 0$) and by $U_1, \ldots, U_L$, the eigenvectors of the matrix $\mathbf{X}\mathbf{X}^T$ corresponding to these eigenvalues. It is assumed that the eigenvectors have unit length, i.e., $\|U_i\| = 1$, where $\|\cdot\|$ is the Euclidean norm. If $d = \max\{i, \text{such that } \lambda_i > 0\} = rank\, \mathbf{X}$ then the SVD of the trajectory matrix can be written as $\mathbf{X} = \mathbf{X}_1 + \cdots + \mathbf{X}_d$, where $\mathbf{X}_i = \sqrt{\lambda_i} U_i V_i^T$ and $V_i = \mathbf{X}^T U_i / \sqrt{\lambda_i}$ $(i = 1, \ldots, d)$.

This is followed by the first step within the Reconstruction stage, i.e., the grouping step. This splits the elementary matrices $\mathbf{X}_i$ into several groups and sums the matrices within each group. Finally, the second step, diagonal averaging is used to transform each matrix of the grouped decomposition into a less noisy time series.

It is noteworthy that the eigenvectors of $\mathbf{X}\mathbf{X}^T$ play a very pivotal role in signal extraction. Let I be the chosen set of eigentriples attained via grouping within SSA and $U_i, i \in I$, be the corresponding eigenvectors. To extract the signal by set I, the matrix $\widehat{\mathbf{X}}$ is diagonally averaged where $\widehat{\mathbf{X}} = \sum_{j \in I} U_j U_j^T \mathbf{X}$. The matrix $\widehat{\mathbf{X}}$ consists of column vectors $\widehat{X}_i$ where $\widehat{X}_i = \sum_{j \in I} U_j U_j^T X_i$.

Given the crucial importance of eigenvectors $U_i, i \in I$, in signal extraction, extracting accurate values of these eigenvectors are of great importance for improving the accuracy of the signal extraction. It is noteworthy that in reality, the eigenvectors $U_i, i \in I$, are not entirely

noise free. This is due to the fact that bicoid gene expression profile is contaminated with highly volatile noise. If Bcd data at location t is denoted by y_t, it can be assumed that y_t is the sum of a noise free series (signal) and noise, such that:

$$y_t = s_t + n_t, \quad t = 1, \ldots, N, \tag{3}$$

where s_t and n_t represent the signal and noise components, respectively. Then, Equation (3) can also be expressed in the following matrix form:

$$\mathbf{X} = \mathbf{S} + \mathbf{N}, \tag{4}$$

where $\mathbf{S}$ and $\mathbf{N}$ represent $L \times K$ trajectory matrices of the signal and noise components, respectively. According to (4), the trajectory matrix $\mathbf{X}$ is not a noiseless matrix. Hence, the eigenvectors of $\mathbf{X}\mathbf{X}^T$ are contaminated with noise. Therefore, it appears that the accuracy of the Reconstruction stage in SSA is affected by the presence of some noise in eigenvectors $U_i, i \in I$, which result in reducing the accuracy of the signal extraction performance. Therefore, developing a more accurate signal extraction method for filtering the noisy Bcd profile is essential.

Following the removal of the noise levels as much as possible, one may assume that the reconstructed/filtered series s_t without noise can be represented by:

$$y_t = s_t. \tag{5}$$

Then, using the assumption that s_t follows a noise-free SDD model of length N, we have:

$$s_t = Ae^{-t/\lambda}, \quad t = 1, 2, \ldots, N. \tag{6}$$

We begin by probing the structure of the matrix $\mathbf{X}\mathbf{X}^T$ for this SDD model (6). If the ith row of the trajectory matrix $\mathbf{X}$ is denoted by H_i, i.e., $H_i = (y_i, \ldots, y_{i+K-1})$, then it can be concluded that the components of the ith row and jth column of matrix $\mathbf{X}\mathbf{X}^T$ is given as follows:

$$\begin{aligned} H_i H_j^T &= \sum_{l=0}^{K-1} y_{i+l} y_{j+l} = \sum_{l=0}^{K-1} Ae^{-(i+l)/\lambda} Ae^{-(j+l)/\lambda} = A^2 \sum_{l=0}^{K-1} e^{-(i+j+2l)/\lambda} \\ &= A^2 \sum_{l=0}^{K-1} e^{-(i+j-2+2+2l)/\lambda} = A^2 \sum_{l=0}^{K-1} e^{-(i+j-2)/\lambda} e^{-(2+2l)/\lambda} \\ &= A^2 e^{-(i+j-2)/\lambda} \sum_{l=0}^{K-1} e^{-2(1+l)/\lambda} = \gamma A^2 e^{-(i+j-2)/\lambda}. \end{aligned}$$

where $\gamma = \sum_{l=0}^{K-1} e^{-2(1+l)/\lambda}$. If the matrix $\mathbf{E}_L$ is defined as:

$$\mathbf{E}_L = (e^{-(i+j-2)/\lambda})_{i,j=1}^{L,L} = \begin{pmatrix} 1 & e^{-1/\lambda} & \ldots & e^{-(L-1)/\lambda} \\ e^{-1/\lambda} & e^{-2/\lambda} & \ldots & e^{-L/\lambda} \\ \vdots & \vdots & \ddots & \vdots \\ e^{-(L-1)/\lambda} & e^{-L/\lambda} & \ldots & e^{-2(L-1)/\lambda} \end{pmatrix}_{L \times L}, \tag{7}$$

it can be concluded that $\mathbf{X}\mathbf{X}^T = \gamma A^2 \mathbf{E}_L$. Since $\mathbf{X}\mathbf{X}^T$ is a multiple of matrix $\mathbf{E}_L$, the matrices $\mathbf{X}\mathbf{X}^T$ and $\mathbf{E}_L$ have similar eigenvectors (with different eigenvalues). Additionally, we have $X_j = e^{-1/\lambda} X_{j-1} = e^{-(j-1)/\lambda} X_1$ where X_j is the jth column of the trajectory matrix $\mathbf{X}$. Accordingly, $rank(\mathbf{X}) = rank(\mathbf{X}\mathbf{X}^T) = rank(\mathbf{E}_L) = 1$. As the rank of trajectory matrix for the SDD model

(6) is equal to one for different values of A and λ, we utilize the first eigentriple alone for signal extraction and consider the remainder as noise ($I = \{1\}$).

Then, to obtain the eigenvector of $\mathbf{X}\mathbf{X}^T$, it is sufficient to find the eigenvector of $\mathbf{E}_L$. Since $rank(\mathbf{E}_L) = 1$, the matrix $\mathbf{E}_L$ has only one eigenvector. It can be easily shown that $\mathbf{E}_L = \mathbf{e}_L\mathbf{e}_L^T$, where $\mathbf{e}_L = \left(1, e^{-1/\lambda}, \ldots, e^{-(L-1)/\lambda}\right)^T$. Thus, we obtain:

$$\mathbf{E}_L\mathbf{e}_L = \mathbf{e}_L\mathbf{e}_L^T\mathbf{e}_L = \|\mathbf{e}_L\|^2\,\mathbf{e}_L. \tag{8}$$

Relation (8) indicates that $\|\mathbf{e}_L\|^2$ and $\mathbf{e}_L$ are eigenvalue and eigenvector of matrix $\mathbf{E}_L$, respectively. Consequently, the eigenvalue and eigenvector of matrix $\mathbf{X}\mathbf{X}^T$ are $\gamma A^2\|\mathbf{e}_L\|^2$ and $\mathbf{e}_L$, respectively.

Now, we can exploit the noise-free eigenvector of $\mathbf{X}\mathbf{X}^T$, to extract the signal of the SDD model (6) based on $U_1 = \frac{\mathbf{e}_L}{\|\mathbf{e}_L\|}$. Since $\mathbf{E}_L = \mathbf{e}_L\mathbf{e}_L^T$, we obtain:

$$U_1U_1^T = \frac{\mathbf{e}_L}{\|\mathbf{e}_L\|}\frac{\mathbf{e}_L^T}{\|\mathbf{e}_L\|} = \frac{\mathbf{e}_L\mathbf{e}_L^T}{\|\mathbf{e}_L\|^2} = \frac{1}{\|\mathbf{e}_L\|^2}\mathbf{E}_L. \tag{9}$$

Therefore, diagonal averaging the matrix $\widehat{\mathbf{X}}$ provides the signal where $\widehat{\mathbf{X}} = \sum_{j\in I} U_jU_j^T\mathbf{X} = \frac{1}{\|\mathbf{e}_L\|^2}\mathbf{E}_L\mathbf{X}$. The matrix $\widehat{\mathbf{X}}$ consists of column vectors $\widehat{X}_i = \sum_{j\in I} U_jU_j^TX_i = \frac{1}{\|\mathbf{e}_L\|^2}\mathbf{E}_LX_i$.

It is noteworthy that the noise-free eigenvector $\mathbf{e}_L$ depends on the window length (L) and parameter λ of the SDD model (6) requires estimation. Note that, the length of Bcd in each profile (N) and parameter A of the SDD model have no effect on $\mathbf{e}_L$.

As mentioned previously, in reality the eigenvector U_1 is not noise free and therefore, we assume that $U_1 = \frac{\mathbf{e}_L}{\|\mathbf{e}_L\|} + \boldsymbol{\varepsilon}$. More precisely, the following non-linear model can be considered:

$$u_i = \frac{e^{-(i-1)/\lambda}}{\sqrt{1 + e^{-2/\lambda} + \cdots + e^{-2(L-1)/\lambda}}} + \varepsilon_i, \quad i = 1, 2, \ldots, L, \tag{10}$$

where u_i is the ith component of eigenvector U_1. For each fixed $L \geq 3$, the non-linear regression model (10) can be used to estimate the parameter λ by nonlinear least squares approach. However, there are many other approaches than can be used here to estimate the parameter lambda.

3 Empirical Findings

3.1 Simulation Study

In this subsection, the performance of the new signal extraction approach is compared with the SDD method by simulating an exponential curve drawn from the SDD model as the benchmark. As noted in [8–11], the Bcd profiles follow a highly volatile exponential trend. In order to obtain a noisy Bcd profile similar to the real one, the following multiplicative model was used:

$$y_t = Ae^{-t/\lambda}.\varepsilon_t, \quad t = 1, 2, \ldots, N,$$

where $A = 200$, $\lambda = 50$, y_t is Bcd data at location t and ε_t represents the random noise following the Log-Normal distribution with zero mean and standard deviation 0.2 in log scale. In total 100 observations were generated and the simulation was repeated 1000 times. It is noteworthy that changes to N do not result in any alterations to the main conclusions reported here.

The accuracy of signal extractions are evaluated using the frequently cited Root Mean Squared Error (RMSE) criterion. See for example, [16, 17] and references therein. The ratio of

the RMSE (RRMSE) is calculated as follows:

$$RRMSE = \frac{\left(\sum_{t=1}^{N}(y_t - \tilde{\tilde{y}}_t)^2\right)^{1/2}}{\left(\sum_{t=1}^{N}(y_t - \tilde{y}_t)^2\right)^{1/2}},$$

where $\tilde{\tilde{y}}_t$ and $\tilde{y}_t$ are extracted signals at location t obtained via the newly proposed and SDD methods, respectively. If $RRMSE < 1$, it can be concluded that the new technique outperforms the SDD method. Furthermore, to measure the approximate separability between signal and noise, metric *weighted correlation* or **w**-correlation is calculated:

$$\rho^{(w)} = \frac{\sum_{t=1}^{N} w_t s_t n_t}{\sqrt{\sum_{t=1}^{N} w_t s_t^2 \sum_{t=1}^{N} w_t n_t^2}},$$

where s_t and n_t represent the signal and noise components, respectively as defined in Equation (3). Also, $w_t = \min\{t, L, N - t + 1\}$ are the weights. Well separated components produce small correlations whereas poorly separated components have large correlations (in absolute values). Therefore, an absolute value of **w**-correlation close to zero ($\left|\rho^{(w)}\right| \simeq 0$) would indicate high separability between signal and noise [13]. The concept of separability plays a key role in SSA and the value of **w**-correlation indicates the quality of the decomposition by determining how well different components of a time series are separated from each other. More details on **w**-correlation can be found in [13].

To enable easier comparison, the ratio of the absolute values of **w**-correlations is computed as follows:

$$Ratio = \left| \frac{\rho^{(w)}_{new\ method}}{\rho^{(w)}_{SDD-NLS}} \right|,$$

If $Ratio < 1$, it can be concluded that the new method outperforms the SDD-NLS method in separating the signal from noise.

Here, we also study the optimum value of L which is very important for signal extraction or separability between extracted signal and noise. In Figure 2, the RRMSE values are depicted versus different values of window length (L). It is evident that all RRMSEs are less than one. Consequently, it can be concluded that the new method provides a more reliable signal extraction in comparison to that obtained via the SDD model across different values of L, with particular reference to the smaller values. It should be noted that when L is too large, the covariance matrix of the L variables is calculated with only a few observations. This, in turn, extends the imprecision of the result. Moreover, a considerably large value of L results in some parts of noise mixing with the signal. Overall, the simulation demonstrated that for all values of L, the newly proposed approach outperforms the basic approach, but that the outcomes are more visible and easily distinguished for smaller L.

Next, we consider the second criterion, separability between signal and noise. The results in Table 1 reports the ratio of **w**-correlation for simulated series. As can be seen in this table, all ratios are less than one which indicates that the new method proposed in this paper always extracts the signal better than the SDD model.

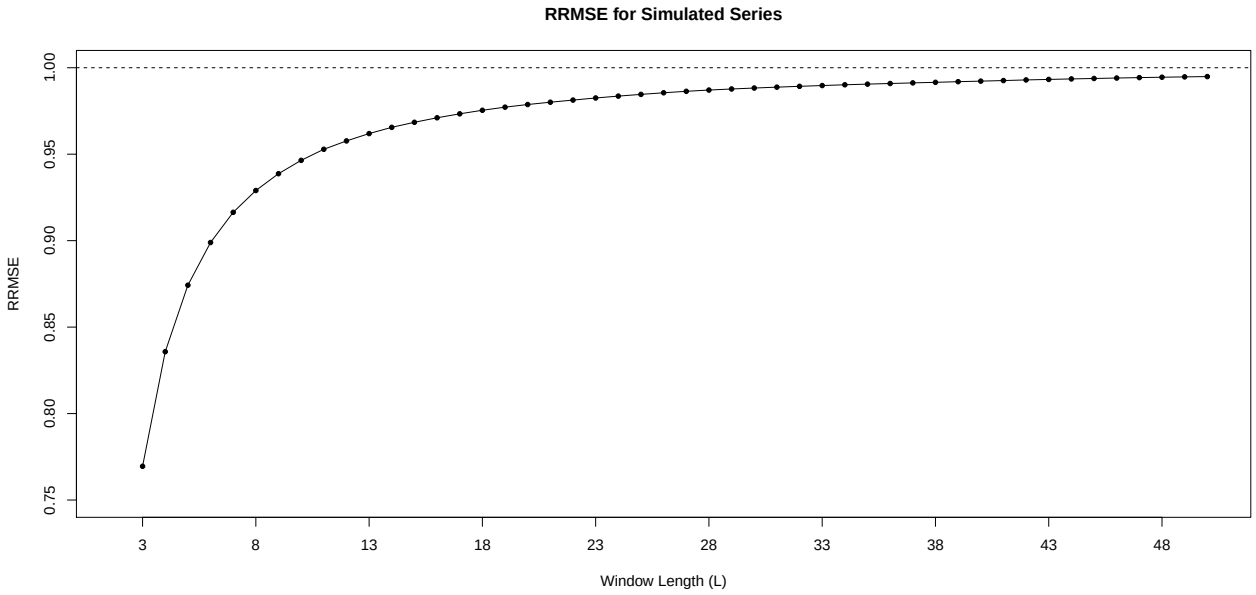

Figure 2: Comparison between SDD and new method for various value of L and based on RRMSE.

	L									
	5	10	15	20	25	30	35	40	45	50
Ratio	0.464	0.168	0.102	0.076	0.063	0.055	0.050	0.044	0.038	0.032

Table 1: Ratio of **w**-correlation for simulated series across different values of L.

Thus, the results of the simulation study confirms that the proposed approach works better than the SDD model based on the two criteria evaluated above, especially for smaller values of L which results in a more noisier signal. It is also noteworthy that the maximum separability between signal and noise occurred at different values of L and not always at the greater/max values of L (see Tables 2-4). This is a result of the structure of this particular data which has been explained in detail above.

As third criterion, we consider the ratio of RMSE of parameter estimators defined as follows:

$$RRMSE_A = \frac{\left(\sum_{i=1}^{1000}(A_i - \hat{\hat{A}}_i)^2\right)^{1/2}}{\left(\sum_{i=1}^{1000}(A_i - \hat{A}_i)^2\right)^{1/2}},$$

$$RRMSE_\lambda = \frac{\left(\sum_{i=1}^{1000}(\lambda_i - \hat{\hat{\lambda}}_i)^2\right)^{1/2}}{\left(\sum_{i=1}^{1000}(\lambda_i - \hat{\lambda}_i)^2\right)^{1/2}},$$

where $A_i = 200$ and $\lambda_i = 50$ for each of ith simulated series, $\hat{\hat{A}}_i$ and $\hat{A}_i$ are estimated value of parameter A obtained via SSA-NLS and SDD-NLS in the ith simulated series, respectively. Similarly, $\hat{\hat{\lambda}}_i$ and $\hat{\lambda}_i$ are estimated value of parameter λ given from SSA-NLS and SDD-NLS in the ith simulated series, respectively. If $RRMSE_A < 1$ ($RRMSE_\lambda < 1$), it would indicate that SSA-NLS parameter estimation method is more precise than SDD-NLS to estimate parameter A (λ). Figures 3 and 4 show the values of $RRMSE_A$ and $RRMSE_\lambda$, respectively. Again,

these results confirm that SSA-NLS approach estimates parameters A and λ more accurate than SSA-NLS method.

(Please add your comments on Figures 5 and 6)

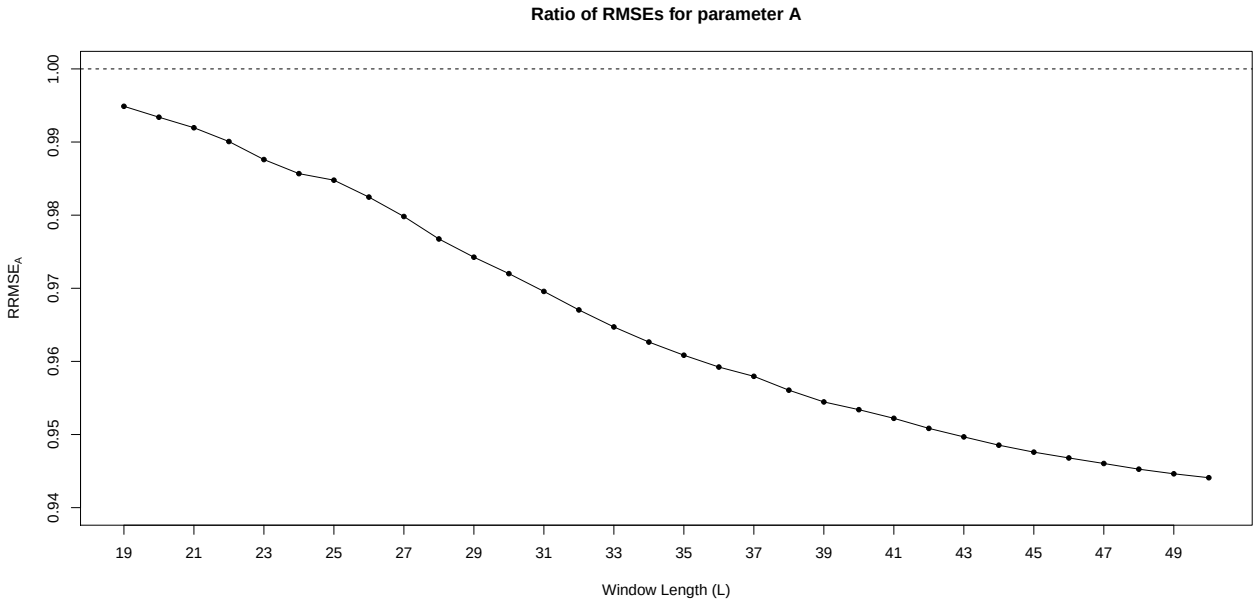

Figure 3: Ratio of RMSEs for parameter A

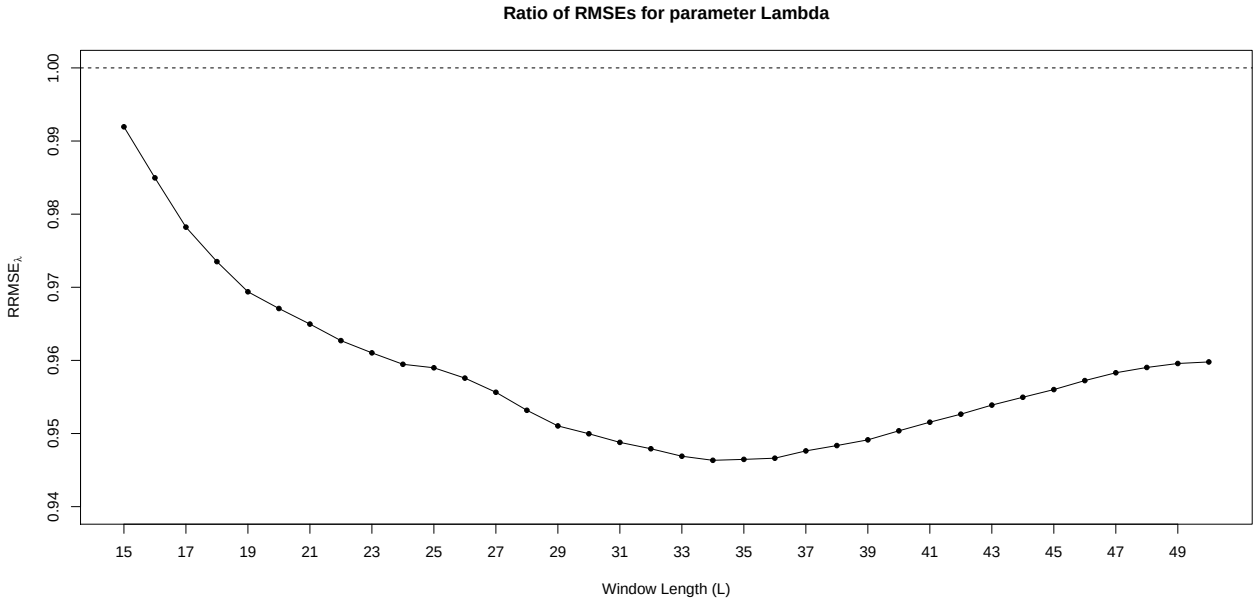

Figure 4: Ratio of RMSEs for parameter λ

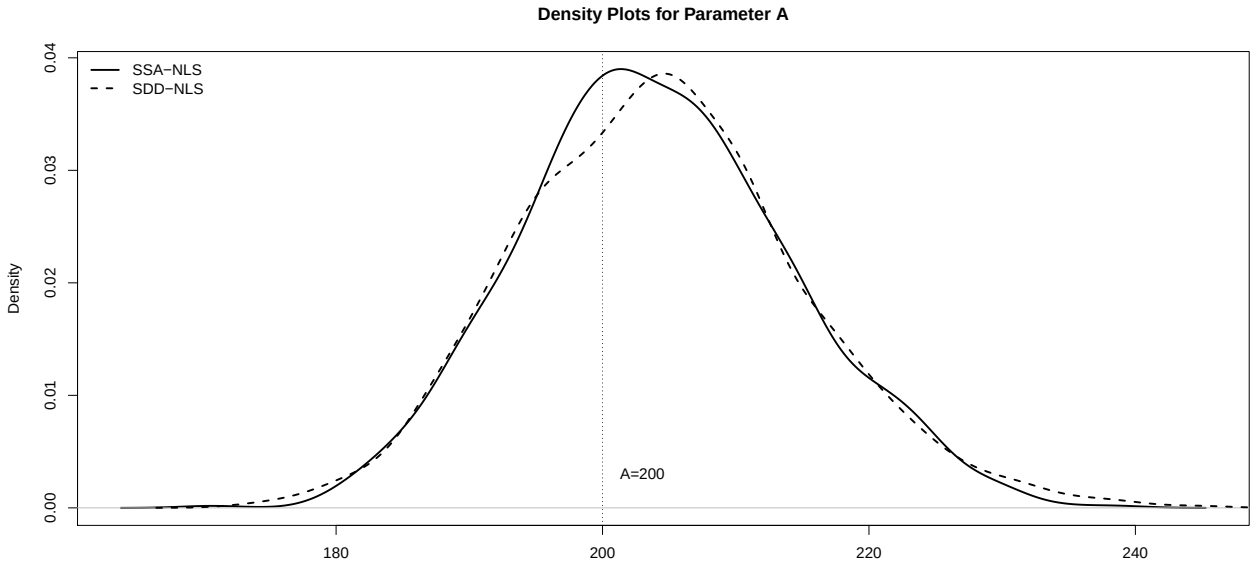

Figure 5: Density Plots for Parameter A

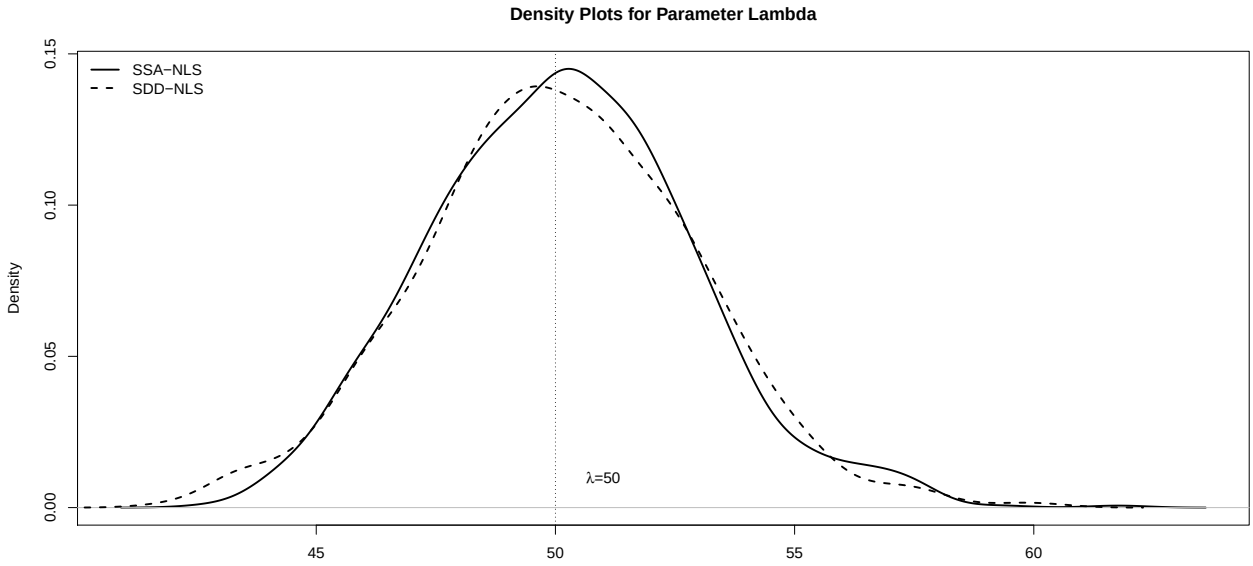

Figure 6: Density Plots for Parameter λ

3.2 Real Data:Bcd Data

Next, we go a step further and replicate the analysis on real Bcd profiles. The evaluation is performed on Bcd expression profiles introduced in the FlyEx database [7, 14]. FlyEx is a powerful database widely used for studying the dynamics of segment determination in early *Drosophila* development [15].

In FlyEx, the quantitative Bcd data was obtained using the confocal scanning microscopy of fixed embryos immunostained for segmentation proteins [7]. In FlyEx, the quantitative Bcd data was obtained using the confocal scanning microscopy of fixed embryos immunostained for segmentation proteins [7]. For analysis purposes, a 1024×1024 pixel confocal image with 8 bits of fluorescence data was created for each embryo and then transformed into an ASCII table.

The ASCII table contains the fluorescence intensity levels attributed to each nucleus of A-P axis. To present the data using a graph, the x-axis shows the A-P position along the length of the embryo and the y-axis shows the intensity levels which correspond to the amount of Bcd expressed at that location.

To evaluate the performance of the newly proposed method, this study examines 385 raw Bcd expression profiles from cleavage cycle 10 to 14(8) which were unprocessed for any noise reduction methods. Similar to the simulation results, the entire RRMSE values achieved for all Bcd profiles are less than one. To save space, Figures 7, 8 and 9 below portray a selection of RRMSE values versus different values of L ($3 \leq L \leq N/2$) achieved on the actual data for a group of embryos in cleavage cycle 10-12 [1].

It is evident that the new method provides more accurate signal extractions in comparison to the SDD model's signal extractions, with noteworthy superior performance at small Window Lengths L. In addition, the ratio of **w**-correlations for these Bcd profiles are reported in Tables 2, 3 and 4. In accordance with results from the simulation study, the entire ratios are less than one following the empirical application. Therefore, it can also be concluded that the new method can separate the signal from noise better than the SDD-NLS method.

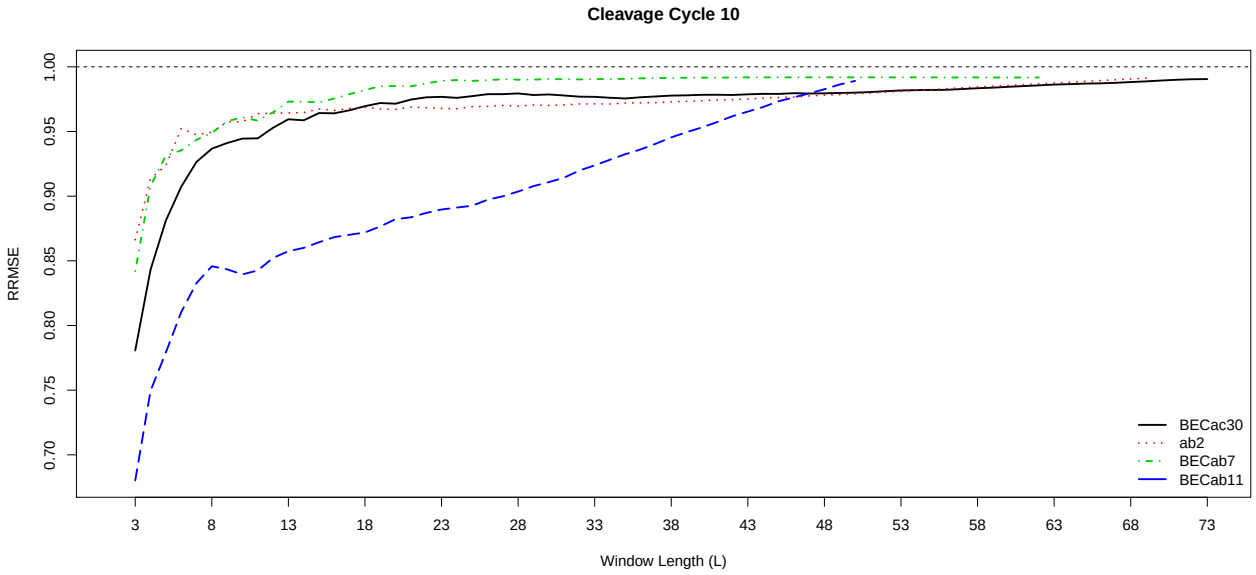

Figure 7: RRMSE for Bcd profiles from cleavage cycle 10.

	L									
Embryo	5	10	15	20	25	30	35	40	45	50
BECac30	0.635	0.267	0.399	0.865	0.202	0.135	0.123	0.114	0.109	0.104
ab2	0.314	0.146	0.059	0.049	0.053	0.065	0.074	0.082	0.089	0.090
BECab7	0.417	0.119	0.139	0.156	0.301	0.291	0.272	0.269	0.261	0.264
BECab11	0.401	0.086	0.067	0.073	0.088	0.093	0.092	0.088	0.077	0.063

Table 2: Ratio of **w**-correlation for Bcd profiles from cleavage cycle 10.

[1]Results for other cleavage cycles are available upon request.

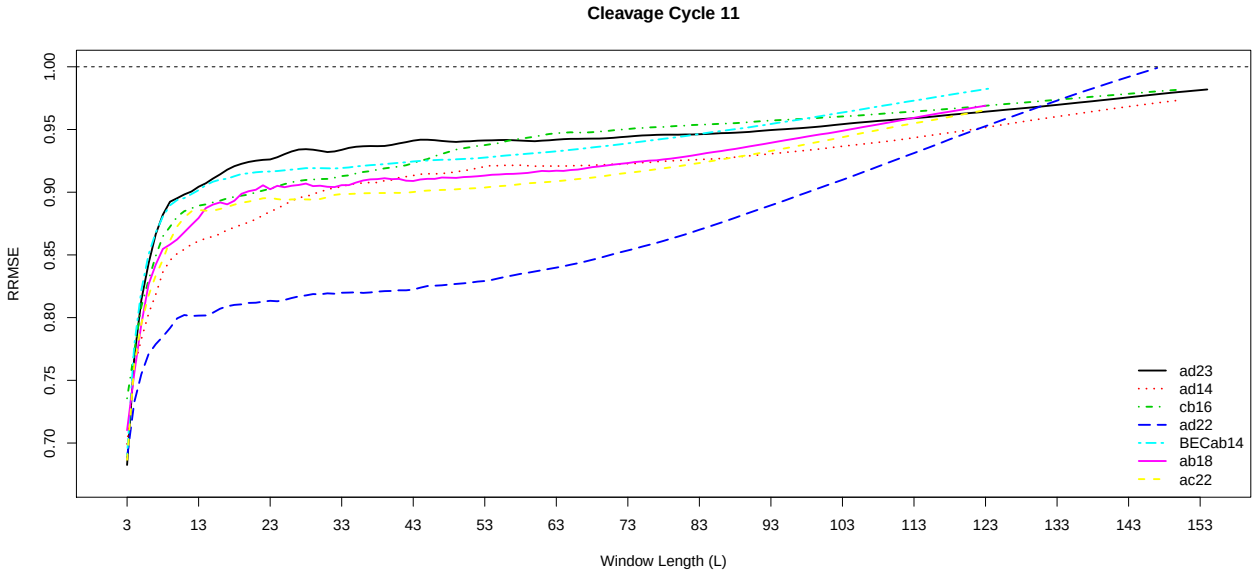

Figure 8: RRMSE for Bcd profiles from cleavage cycle 11.

	L											
Embryo	5	15	25	35	45	55	65	75	85	95	105	115
ad23	0.498	0.183	0.215	0.125	0.080	0.074	0.076	0.078	0.080	0.081	0.079	0.075
ad14	0.160	0.104	0.096	0.413	0.259	0.144	0.125	0.124	0.124	0.122	0.115	0.105
cb16	0.203	0.083	0.081	0.128	0.224	0.265	0.407	0.252	0.210	0.190	0.174	0.158
ad22	0.223	0.031	0.035	0.053	0.075	0.101	0.125	0.149	0.171	0.190	0.203	0.208
BECab14	0.286	0.224	0.082	0.087	0.083	0.087	0.091	0.093	0.093	0.090	0.081	0.068
ab18	0.562	0.117	0.079	0.114	0.159	0.210	0.260	0.282	0.291	0.283	0.291	0.224
ac22	0.347	0.065	0.041	0.040	0.049	0.056	0.063	0.069	0.074	0.075	0.071	0.065

Table 3: Ratio of **w**-correlation for Bcd profiles from cleavage cycle 11.

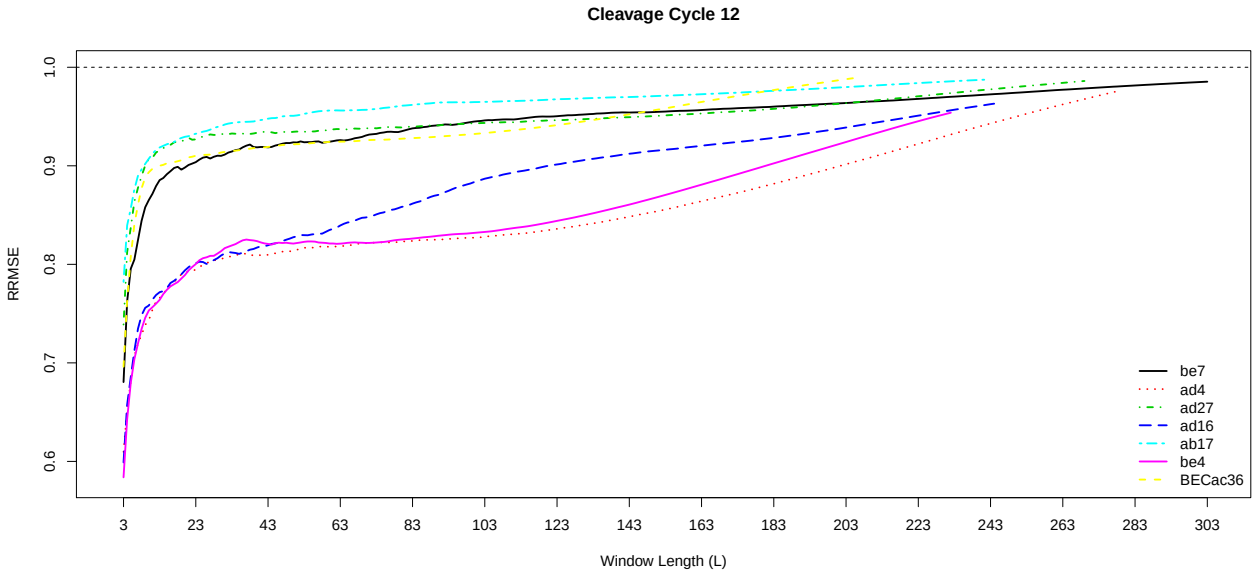

Figure 9: RRMSE for Bcd profiles from cleavage cycle 12.

	L													
Embryo	5	20	35	50	65	80	95	110	125	140	155	170	185	200
be7	0.017	0.023	0.018	0.028	0.046	0.084	0.288	0.438	0.178	0.131	0.116	0.108	0.102	0.096
ad4	0.326	0.313	0.046	0.038	0.036	0.041	0.051	0.063	0.074	0.083	0.092	0.099	0.105	0.108
ad27	0.650	0.277	0.431	0.211	0.145	0.112	0.097	0.087	0.085	0.084	0.083	0.081	0.078	0.072
ad16	0.265	0.035	0.059	0.115	0.245	0.396	0.216	0.370	0.276	0.242	0.223	0.210	0.194	0.173
ab17	0.602	0.059	0.043	0.075	0.248	0.349	0.193	0.149	0.135	0.127	0.122	0.115	0.106	0.094
be4	0.322	0.114	0.141	0.085	0.105	0.130	0.156	0.190	0.225	0.254	0.278	0.296	0.303	0.298
BECac36	0.258	0.139	0.083	0.063	0.062	0.071	0.081	0.090	0.097	0.098	0.093	0.085	0.075	0.063

Table 4: Ratio of **w**-correlation for Bcd profiles from cleavage cycle 12.

It is pertinent to note that the length of data, Bcd expression level and level of noise associated with Bcd profiles varies from one cleavage cycle to the other. The diffusion of Bcd molecules begin to appear at cleavage cycle 10. Hence, the concentration of these morphogens is lower at the initial cleavage cycles compared to final stages of cleavage cycle 14. Therefore, the constant superior performance of the new method throughout all cleavage cycles should be regarded as an important feature of this method.

Finally, SDD-NLS and SSA-NLS approaches perform differently at noise reduction and signal extraction steps which results in providing different estimations of A and λ. Since a more reliable noise reduction method gives a more accurate estimation of parameters of interest, for those researchers who wish to rely on the SDD model, it is suggested that they first seek to filter the noise using SSA-NLS before estimating the parameters using the SDD model.

4 Conclusion

The diffusion of Bcd along the embryo of *Drosophila melanogaster* provides a concentration gradient of signalling molecules which induces downstream genes at different concentration thresholds and provides embryonic tissues with essential positional information. Over the last few decades, the exponential shape of Bcd gradient had been characterised using several parametric models including SDD. However, whether this model can precisely estimate the associated parameters remains unclear.

The central aim of this paper was to introduce a new method based on SSA for filtering Bcd profiles and consequently enhance the estimation of parameters A and λ. To that end, we bring together empirical evidence via a simulation study and application to real data to evaluate the performance of the two methods of SDD-NLS and SSA-NLS.

The results we obtain demonstrate the feasibility and potential advantages of the SSA-NLS approach for filtering Bcd profiles and improving the estimation of parameters present in the exponential decay function. Moreover, as the length of the Bcd profile increases, the fluctuations and the level of noise in data also increases. Therefore, when extracting the signal from such data, it is reasonable to expect that the signal derived from a profile with large N, does not truly capture the information of the profile. However, as confirmed by the w-correlation results, signal components are precisely extracted for all cleavage cycles, indicating the robustness of SSA to the amount and the level of noise in data.

Further work, however, is encouraged to ensure the method can scale up to expression profiles of other genes in the segmentation network.

References

[1] Driever, W., and Nüsslein-Volhard, C. (1988). The bicoid protein determines position in the Drosophila embryo in a concentration-dependent manner. *Cell*, **54**(1), 95-104.

[2] Frohnhöfer, H., G., and Nüsslein-Volhard, C. (1986). Organization of anterior pattern in the Drosophila embryo by the maternal gene bicoid. *Nature*, **324**, 120-125.

[3] Berleth, T., Burri, M., Thoma, G., Bopp, D., Richstein, S., Frigerio, G., Noll, M. and Nüsslein-Volhard, C. (1988). The role of localization of bicoid RNA in organizing the anterior pattern of the Drosophila embryo. *The EMBO journal*, **7**(6), 1749-1756.

[4] Ghodsi, Z., Hassani, H., and McGhee, K. (2015). Mathematical approaches in studying bicoid gene. *Quantitative Biology*, **3**(4), 182-192.

[5] Gregor, T., Wieschaus, E. F., McGregor, A. P., Bialek, W., and Tank, D. W. (2007). Stability and nuclear dynamics of the bicoid morphogen gradient. *Cell*, **130**(1), 141-152.

[6] Drocco, J. A., Wieschaus, E. F., and Tank, D. W. (2012). The synthesis-diffusion-degradation model explains Bicoid gradient formation in unfertilized eggs. *Physical biology*, **9**(5), 055004.

[7] Pisarev, A., Poustelnikova, E., Samsonova, M., and Reinitz, J. (2009). FlyEx, the quantitative atlas on segmentation gene expression at cellular resolution. *Nucleic acids research*, **37**(suppl-1), D560-D566.

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

[9] Hassani, H., Silva, E.S. and Ghodsi, Z., 2017. Optimizing bicoid signal extraction. *Mathematical Biosciences*, 294, 46-56.

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

[11] Ghodsi, Z. and Hassani, H., 2017. Evaluating the Analytical Distribution of bicoid Gene Expression Profile. *Meta Gene*, **14**, 91–99.

[12] Grimm, O., Coppey, M., and Wieschaus, E. (2010). Modelling the Bicoid gradient. *Development*, **137**(14), 2253-2264.

[13] Sanei, S. and Hassani, H. (2016). *Singular Spectrum Analysis of Biomedical Signals*, Taylor & Francis, CRC Press.

[14] Kozlov, K., Myasnikova, E., Samsonova, M., Reinitz, J., and Kosman, D. (2000). Method for spatial registration of the expression patterns of Drosophila segmentation genes using wavelets. *Computational Technologies*, **5**, 112-119.

[15] Poustelnikova, E., Pisarev, A., Blagov, M., Samsonova, M., and Reinitz, J. (2004). A database for management of gene expression data in situ. *Bioinformatics*, **20**(14), 2212-2221.

[16] Hassani, H., Silva, E. S., Gupta, R., and Das, S. (2018). Predicting global temperature anomaly: A definitive investigation using an ensemble of twelve competing forecasting model. *Physica A*, **509**, 121-139.

[17] Silva, E. S., Hassani, H., and Heravi, S. (2018). Modeling European industrial production with multivariate singular spectrum analysis: A cross-industry analysis. *Journal of Forecasting*, **37**(3), 371-384.