Noise correction in gene expression data: A new approach based on subspace method

Nader Alharbi*,Zara Ghodsi*,Hossein Hassani *

* *The Statistical Research Centre, Bournemouth University, 89 Holdenhurst Road, Bournemouth BH8 8EB, UK*

Abstract

We present a new approach for removing the nonspecific noise from *Drosophila* segmentation genes. The algorithm used for filtering here is an enhanced version of Singular Spectrum Analysis (SSA) method which decomposes a gene profile into the sum of a signal and noise. Since the main issue in extracting signal using SSA procedure lies in identifying the number of eigenvalues needed for signal reconstruction, this paper seeks to explore the applicability of the new proposed method for eigenvalues identification in four different gene expression profiles. Our findings indicate that when extracting signal from different genes, for optimised signal and noise separation, different number of eigenvalues need to be chosen for each gene.

1 Introduction

Segmentation in *Drosophila melanogaster* is arguably the best studied example of gene regulatory network in developmental studies [1]. In this network, it is widely accepted that the patterns of the segmentation factors which have been activated by the primary morphogens direct the early embryo development. However, due to the presence of noise, finding the pattern of segmentation factors is not a simple task [2–4] and even a small level of noise in gene expression patterns will considerably affect our understanding of the embryo developmental fate. Hence, it is important to probe the gene expression signal using a method which effectively enables us to filter the fluctuations of the related gene protein profile.

These profiles can mostly be achieved by using fluorescence imaging technique [5]. Such quantification relies on the assumption that the actual protein concentrations detected by the Fluorescence in situ hybridization (FISH) technique are linearly related to the embryos natural protein concentration. However, the obtained profile contains different levels of noise which need to be removed first. Among several noise removal models Singular Spectrum Analysis (SSA) is a relatively new method which has recently transformed itself into a valuable tool for gene expression signal extraction. The first such application of SSA was seen in 2006 when Holloway et al. studied the relation

between maternal protein gradients and segmentation process in *Drosophila*, by analysing gene expression patterns extracted by SSA [6]. It is worthy to mention two powerful characteristics of SSA; no requirement of any assumptions about the data and related residuals and its effective performance in noise filtering [7] made SSA a valuable method in analysing segmentation genes profile. Since then SSA and related theoretical developments of this method such as two-dimensional SSA and SSA based on minimum variance has been used in several other studies, see for example [3, 5, 8–13].

Even though the signal extraction by SSA appears to be simple, in practice it is a complicated task since for some cases the trend cannot be separated from noise or cyclic components just by choosing the first eigenvalue. This issue for the first time raised by Alexandrov et al. in [3] in which the author suggested using either a small window length or adding a constant to the series for improving the separability between signal and noise [3]. Although the practical possibility of these suggestions there is still an open question related with the identification of the number of eigenvalues required for series reconstruction.

To address this question we mainly follow the idea proposed in [14–16]. The proposed method has been mainly used for noise reduction, filtering, signal extraction and distinguishing chaos from noise in time series analysis [17]. In identifying the number of the eigenvalues this method mainly relies on the distribution of the scaled Hankel matrix eigenvalues. Here, we apply the method for signal extraction in four different genes; bicoid (bcd), caudal (cad), giant (gt) and even-skipped (eve) which are among the most important zygotic segmentation genes. The approach enables us to decide the appropriate number of eigenvalues related to the gene signal.

The remainder of this paper is organised such that Section 2 describe the methodology which is followed by the empirical results from the simulated and real data applications in Section 3. The paper concludes with a concise summary in Section 4.

2 Methodology

2.1 A brief description

Presented below is a short description of the method used in this study. The main aim of the SSA technique is to analyse the original series into a sum of series, so that each component can be identified as either a main component or noise. Our interest here is to consider the signal as a whole so that we can determine the eigenvalues related to the signal component. The SSA approach consists of two main stages: decomposition and reconstruction; of which each stage consists of two compatible steps. Embedding and singular value decomposition (SVD) in the first stage, grouping and diagonal averaging in the second stage [7]. The proposed approach is a novel step that can be used between the first and second stages of SSA to select the proper value of eigenvalues r.

In doing so, let us consider a one-dimensional series $Y_N = (y_1, \ldots, y_N)$ of length N. Mapping 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 ($2 \leqslant$

$L \leqslant N/2$) and $K = N - L + 1$. The matrix $\mathbf{X}$ is a Hankel matrix, which means all the elements along the diagonal $i + j = const$ are equal. Set $\mathbf{A} = \mathbf{X}\mathbf{X}^T$ and denote by λ_i $(i = 1, \ldots, L)$ the eigenvalues of $\mathbf{A}$ taken in the 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 the matrix $\mathbf{A}$ corresponding to these eigenvalues. Set

$$d = rank\, \mathbf{X} = \max(i, \text{ such that } \lambda_i > 0).$$

The SVD of the trajectory matrix can be written as:

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

where $\mathbf{X}_i = \sqrt{\lambda_i} U_i V_i^T$. The elementary matrices $\mathbf{X}_i$ have rank 1, U_i and V_i are the left and right eigenvectors of the trajectory matrix. 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 $||\mathbf{X}_i|| = \lambda_i$, where $||\ ||_F$ denotes the Frobenius norm.

Let us now consider the step that comes between the two stages in SSA, that is to divide the matrix $\mathbf{A}$ by its trace, $\mathbf{A}/tr(\mathbf{A})$. Let $\zeta_1, \ldots, \zeta_L$ denote the eigenvalues of the matrix $\mathbf{A}/tr(\mathbf{A})$ in decreasing order of magnitude ($1 \geq \zeta_1 \geq \ldots \geq \zeta_L \geq 0$). In this step, we perform the simulation technique to gain the distribution of ζ_i, so we can understand the behaviour of each eigenvalue, which can be useful for obtaining the proper value of r. In this paper, our aim is to ascertain the distribution of ζ_i and its related forms that can be used directly for choosing the optimal value of r for the genes signal extraction.

Once r is obtained, the grouping step splits the matrices $\mathbf{X}_i$ into two groups. Therefore, (1) can be written as

$$\mathbf{X} = \mathbf{S} + \mathbf{E}, \tag{2}$$

where $\mathbf{S} = \sum_{i=1}^{r} \mathbf{X}_i$ is the signal matrix and $\mathbf{E} = \sum_{i=r+1}^{d} \mathbf{X}_i$ is the noise matrix. At the final step, we use the diagonal averaging to transform the matrix $\mathbf{S}$ into a new series of length N (for more information see [18,19]).

2.2 Algorithm

The Algorithm is divided in two stages. At the first stage, we use skewness coefficient and coefficient of variation of ζ_i as the main indicators to find the optimal value of r for the separability between signal and noise, and then at the second stage, we reconstruct the time series.

2.2.1 Stage 1:

1. Transfer a one-dimensional time series $Y_N = (y_1, \ldots, y_N)$ into the 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 $K = N - L + 1$, and the window length L is an integer such that $2 \leq L \leq N/2$. This steps provides the trajectory matrix $\mathbf{X} = [X_1, \ldots, X_K] = (x_{ij})_{i,j=1}^{L,K}$.

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

3. Compute the eigenvalues and eigenvectors of the matrix $\mathbf{A}$ and represent it in the form $\mathbf{A} = \mathbf{P}\mathbf{\Gamma}\mathbf{P}^T$. Here, $\mathbf{\Gamma} = diag(\zeta_1, \ldots, \zeta_L)$ 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 $\mathbf{P} = (P_1, P_2, \ldots, P_L)$ is the corresponding orthogonal matrix of the eigenvectors of $\mathbf{A}$.

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 $b =\mid y_i - y_{i+1} \mid$. In order to obtain a noisy series similar to the real one, random error ϵ with a normal distribution with zero mean and variance σ_ϵ^2 with different amplitudes were added to different part of the series.

5. Calculate the coefficient of skewness for each eigenvalue, $skew(\zeta_i)$. If $skew(\zeta_c)$ is the maximum, then select $r = c - 1$.

6. Calculate the coefficient of variation, $CV(\zeta_i)$. This can split the eigenvalues in two groups, from ζ_1 to ζ_{c-1} which are corresponding to the signal and the rest which has almost a U shape which are corresponding to the noise component.

7. Calculate the absolute values of the correlation between ζ_i and ζ_{i+1}, and plot them in one figure. If $\rho(\zeta_{c-1}, \zeta_c)$ is the minimum, and the pattern for $\rho(\zeta_c, \zeta_{c+1})$ to $\rho(\zeta_{L-1}, \zeta_L)$ has the same pattern for the white noise, then choose $r = c - 1$.

2.2.2 Stage 2

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

2. Transition to the one dimensional series can now be achieved by averaging over the diagonals of the matrix $\widetilde{\mathbf{S}}$.

3 Real data

3.1 Data description

The gene expression data in wild-type *Drosophila melanogaster* embryos achieved by fluorescently tagged antibodies technique and is available via [20] where a more detailed description on the biological characteristics, method and data is made available. This data was extracted from the nuclear intensities of %10 longitudinal strips and the data was not processed for any other noise removal.

Of the many segmentation genes, we are only concerned with four different genes in this study; *bcd*, *cad*, *gt* and *eve* which among them *bcd* is maternal

and, *cad* has both maternal and zygotic origin and *gt* and *eve* are respectively related to gap and pair rule categories of zygotic genes [5, 6, 8].

bcd mRNA is completely maternal and the Bcd protein gradient is formed at cleavage cycle 9 [5, 6, 8]. Figure 1(a) depicts a typical example of *Bcd* gradient related to cleavage cycle 14(3). Although this figure suggests *Bcd* follows an exponential trend due to the high volatility seen in the series, the extraction of this trend is not a simple task.

cad mRNA has both maternal and zygotic origin and the maternal transcripts begin to translate immediately after fertilization. However, proteins encoded by *gt* and *eve* were reported to appear at cycle 12 and 10 respectively and it is accepted that the posterior stripe of *gt* expression is regulated by *bcd* and *cad* [5, 6, 8, 21].

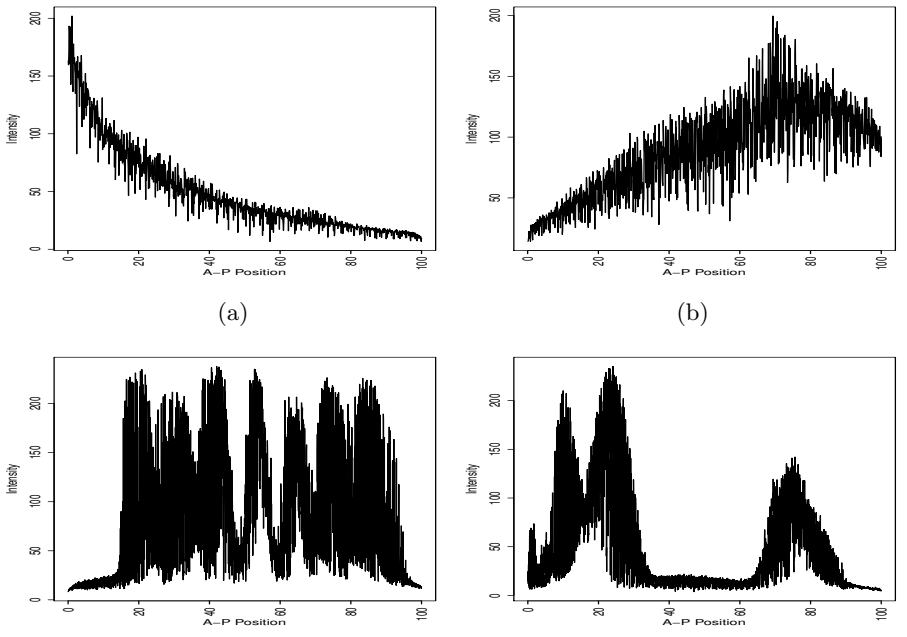


Figure 1: Experimental data from *Drosophila melanogaster* embryo; (a): *bcd*, (b): *cad*, (c): *eve*, (d): *gt* .

3.2 Main results

In this paper, a series of simulated data are used to evaluate the performance of the newly introduced approach. To generate the simulated noisy profiles with similar structure, shape and distribution to the real gene expression profiles we mainly follow the algorithm presented in [17, 18]. This algorithm has also been explained briefly in Section 2.2. Although the gene expression profiles are slightly different from embryo to embryo, as the obtained results in terms of number of eigenvalues are similar, we only consider ten different embryos

for studying each gene. In this regard, each copy of gene expression data was simulated 10^4 times. Studying the distribution of each eigenvalue provides the capacity to obtain an accurate and deep intuitive understanding of selecting the proper value of r. The first data for each gene is analysed and discussed in more detail whilst the results of the other data are summarised based on the outcomes of the skewness, variation and correlation coefficients. The window length used for analysing the *bcd*, *cad*, *gt* and *eve* genes series is 200 (for more information for the selection of the window length refer to [22]).

We mainly focus on the third moment, that is the skewness of the distribution for each eigenvalue:

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

and the coefficient of variation, $CV(\zeta_i)$, which is defined as the ratio of the standard deviation $\sigma(\zeta_i)$ and $\overline{\zeta}_i$:

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

In addition, the Spearman correlation ρ between ζ_i and ζ_{i+1} is also evaluated to enhance the results obtained by *skew* and CV measures. The absolute value of the correlation between ζ_i and ζ_{i+1} is considered, 1 indicates that ζ_i and ζ_{i+1} have perfect positive correlation whilst 0 shows there is no correlation between them.

Figure 2 illustrates the results of $skew(\zeta_i)$ (left) and $CV(\zeta_i)$ (right) for the first data series for each gene type. It can be seen from the left column that the maximum value of *skew* is obtained for ζ_2 in both *bcd* and *cad* data. Whereas, $skew(\zeta_4)$ is the maximum for both *eve* and *gt* series. In the right column, the results of CV splits the eigenvalues into two groups for each data; the second group looks like a U shape which is related to the noise component. The results indicate that $r = 2, 2, 3, 3$ for extracting the *bcd*, *cad*,*eve* and *gt* signal, respectively.

Furthermore, the result of ρ can be used as a decision or test tool if the *skew* and CV measures give different results. However, in these typical examples, the results of those two measures are the same which also supported by the results of the correlation coefficient. It is obvious that the minimum value of ρ are emerged between (ζ_2, ζ_3), (ζ_2, ζ_3), (ζ_3, ζ_4) and (ζ_3, ζ_4) for *bcd*, *cad*, *eve* and *gt*, respectively. Therefore, the results enhance that $r = 2, 2, 3, 3$ for the first data for each gene (see Fig. 3).

Tables. 1, 2, 3, and 4 depict the results of r based on those three measures for all 40 series. For the *bcd* signal extraction, all the outputs show $r = 2$ for all *bcd* data (see Table. 1). Similar results was emerged for extracting the *cad* signal, most of the outcomes indicate $r = 2$.

For the *eve* data, $r = 3$ for five series as all the three measures give the same result. However, for example; for series 2, the results of *skew* and CV are different, $r = 3$ and $r = 4$, respectively. To overcome this, we look at the

result of ρ, which confirms $r = 4$. In this regards, the decision that $r = 3$ is for six series of ten *eve* data. Table. 4 demonstrates that $r = 3$ for all *gt* series except the last series, because all measures have the same results. As a result, for $L = 200$, the required eigenvalues to extract the *bcd*, *cad*, *eve*, and *gt* signals are 2, 2, 3, 3, respectively. Table. 5 shows the final results for all four genes along with the most frequent reported *skew*, *CV* and ρ.

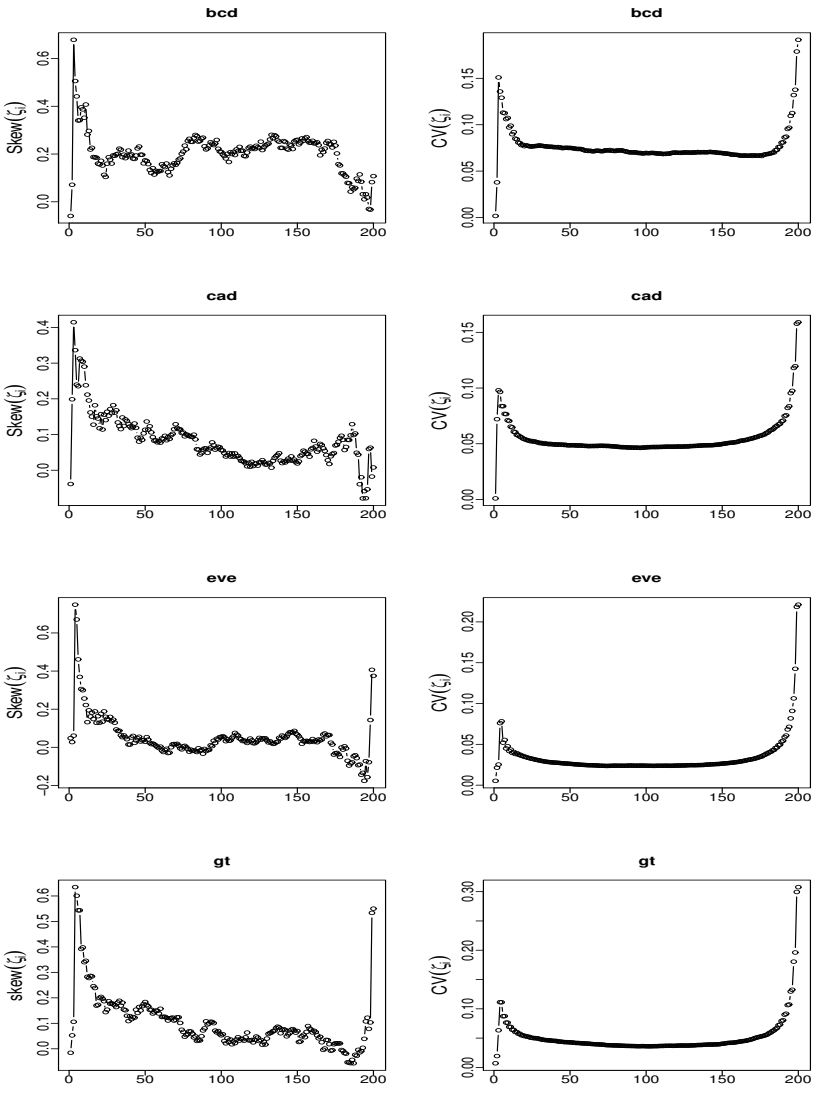


Figure 2: The skewness coefficient (left) and the variation coefficient of ζ_i (right) for the first series of *bcd*, *cad*, *gt* and *eve* data.

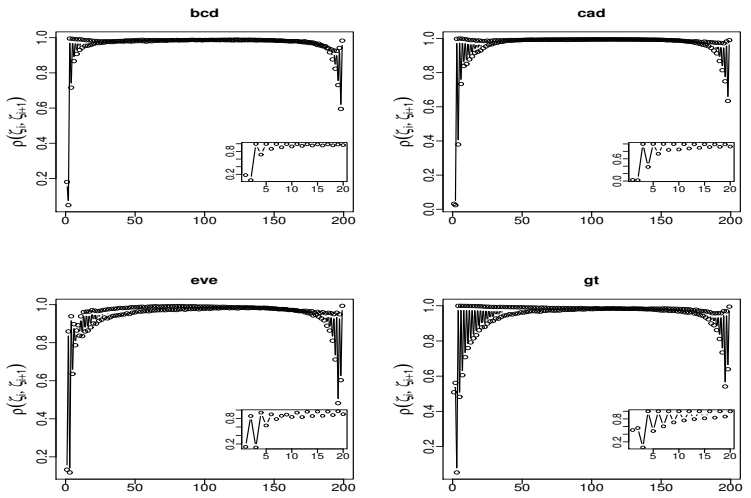


Figure 3: The correlation between ζ_i and ζ_{i+1} for the first series from each data.

Table 1: The values of r based on $Skew$ and CV for the ten *bcd* series.

Series	r $(skew)$	r (CV)	r (ρ)	Series	r $(skew)$	r (CV)	r (ρ)
1	2	2	2	6	2	2	2
2	2	2	2	7	2	2	2
3	2	2	2	8	2	2	2
4	2	2	2	9	2	2	2
5	2	2	2	10	4	2	2

Table 2: The values of r based on $skew$ and CV for the ten *cad* series.

Series	r $(skew)$	r (CV)	r (ρ)	Series	r $(skew)$	r (CV)	r (ρ)
1	2	2	2	6	1	2	1
2	2	2	2	7	2	2	2
3	2	2	2	8	2	2	2
4	1	2	1	9	2	1	2
5	2	2	2	10	3	3	3

After the step of identifying the value of r, we can use those leader eigenvalues in the second stage of the SSA approach (Grouping and Diagonal averaging) to reconstruct the first typical data for each gene. Fig. 4 shows the result of the gene signal extraction or reconstruction series without noise. The red and the black lines correspond to the reconstructed series and the original series respectively. As a results, the considered r for the reconstruction of the original series is obtained properly.

Table 3: The values of r based on *skew* and CV for the ten *eve* series.

Series	r (*skew*)	r (CV)	r (ρ)	Series	r (*skew*)	r (CV)	r (ρ)
1	3	3	3	6	3	4	4
2	4	3	4	7	3	3	3
3	6	6	6	8	4	4	4
4	6	4	4	9	3	3	3
5	3	3	3	10	3	3	3

Table 4: The values of r based on *skew* and CV for the ten *gt* series.

Series	r (*skew*)	r (CV)	r (ρ)	Series	r (*skew*)	r (CV)	r (ρ)
1	3	3	3	6	3	3	3
2	3	3	3	7	3	3	3
3	3	3	3	8	3	3	3
4	3	3	3	9	3	3	3
5	3	3	3	10	5	3	5

Table 5: The final result obtained in noise-signal separation study.

Gene type	r (*skew*)	r (CV)	r (ρ)
bcd	2	2	2
cad	2	2	2
eve	3	3	3
gt	3	3	3

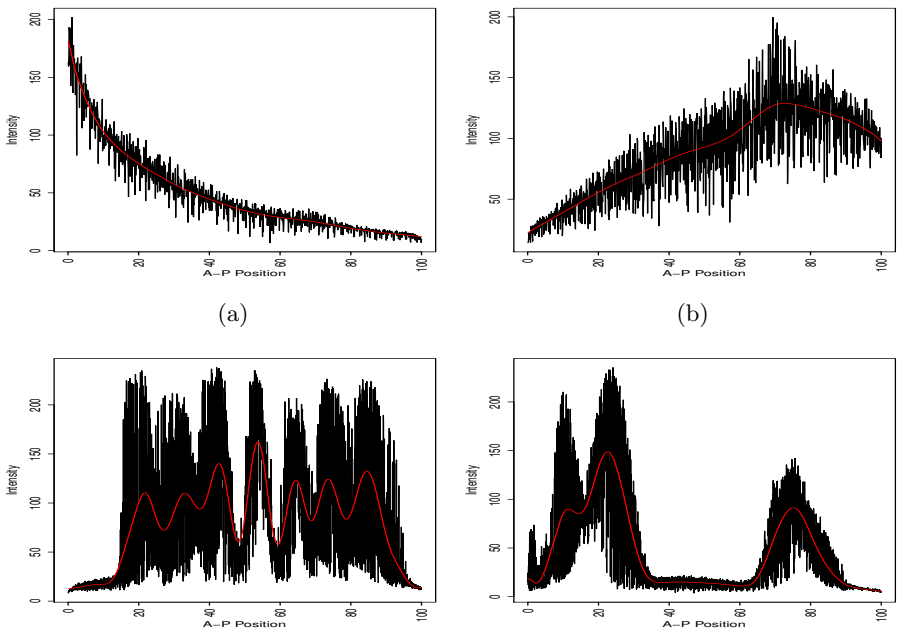
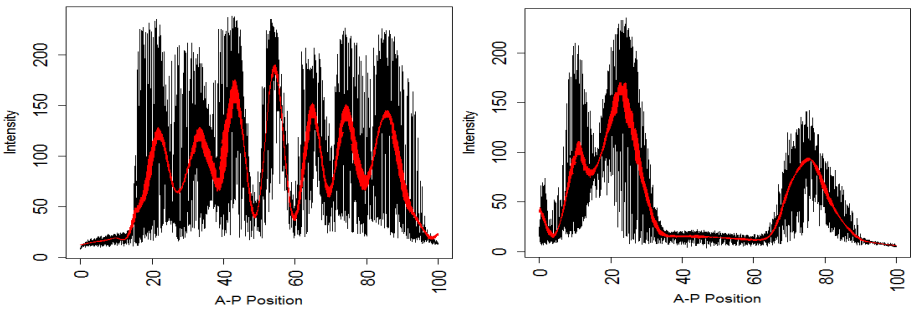


Figure 4: Original (black) and extracted signal (red); (a): *bcd*, (b): *cad*, (c): *gt*, (d): *eve* .

Taking a closer look at Fig 4, it is imperative to note that the extracted signal profiles of *eve* and *gt* do not follow the expression data satisfactorily when the data series changes sharply. Therefore, in order to solve this issue and capture the peaks of the profiles, we used sequential SSA. The main idea underlying this approach is to apply SSA recursively on the residuals with different window length L [23]. By doing so we extract some components of the initial series using basic SSA and then extract the remaining components related to the signal by applying SSA on residuals. Such a recursive SSA application produces a gradual extraction of the signal present in the noise. Fig 5 shows the result after applying sequential SSA. As can be seen signal extraction and peak capturing has been improved accordingly.



(a) (b)

Figure 5: Improving signal extraction using sequential SSA. Original (black) and extracted signal (red);(a): *eve*, (b): *gt* .

4 Conclusion

In this study, a new approach for removing noise and signal extraction in four different Drosophila segmentation genes was applied. The approach is based on the distribution of the eigenvalues of a scaled Hankel matrix. The skewness and variation coefficients of the eigenvalue distribution was used as a new criteria and indicator for the cut-off point in the eigenvalue spectra between signal and noise components. The results confirm that the proposed approach gives a promising output for the gene expression signal extraction and also indicates that the method used for removing noise from the protein profile of gene expression should be flexible enough for different type of genes, as in this study we have obtained different number of eigenvalues needed for signal extraction in each gene.

References

[1] Porcher A, Dostatni N. The bicoid morphogen system. *Current Biology*.2010; **20**(5): 249-254.

[2] Houchmandzadeh B, Wieschaus E, Leibler S. Establishment of developmental precision and proportions in the early Drosophila embryo. *Nature*. 2002; **415**: 798-802.

[3] Alexandrov T, Golyandina N, and Spirov A. Singular Spectrum Analysis of gene expression profiles of early Drosophila embryo: exponential-in-distance patterns. *Research Letters in Signal Processing*. 2008; **2008**(12): 1-5.

[4] Myasnikova E, Samsonova M, Kosman D, & Reinitz J. Removal of background signal from in situ data on the expression of segmentation genes in Drosophila. *Development genes and evolution*, 2005; **215**(6): 320-326.

[5] Surkova S, Kosman D. et al. Characterization of the Drosophila segment determination morphome. *Developmental Biology*. 2007; **313**: 844-862.

[6] Holloway DM, Harrison LG, Kosman D, Vanario Alonso CE, Spirov AV. Analysis of pattern precision shows that Drosophila segmentation develops substantial independence from gradients of maternal gene products. *Developmental Dynamics*. 2006; **235**(11): 2949-2960.

[7] Hassani H. Singular Spectrum Analysis: Methodology and Comparison .*Journal of Data Science*. 2007; **5**:239-257.

[8] Holloway DM, Harrison LG. et al. Analysis of Pattern Precision Shows That *Drosophila* Segmentation Develops Substantial Independence From Gradients of Maternal Gene Products. *Developmental Dynamics*. 2006; **235**: 2949-2960.

[9] Alexandrov T. A method for trend extraction using Singular Spectrum Analysis. *Revstat*. 2009; **7**(1): 1-22.

[10] Golyandina NE, Holloway DM, Lopesc, FJP et al. Measuring gene expression noise in early *Drosophila* embryos: nucleus-to-nucleus variability. *International Conference on Computational Science*. 2012; **9**: 373-382.

[11] Spirov AV, Golyandina NE, Holloway, DM et al. Measuring Gene Expression Noise in Early *Drosophila* Embryos: The Highly Dynamic Compartmentalized Micro-environment of the Blastoderm Is One of the Main Sources of Noise. *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*. 2012; **7246**:177-188

[12] Holloway DM, Lopes FJP, da Fontoura Costa L et al. Gene Expression Noise in Spatial Patterning: hunchback Promoter Structure Affects Noise Amplitude and Distribution in Drosophila Segmentation. *PLoS Computational Biology*. 2011; **7**(2): e1001069. doi:10.1371/journal.pcbi.1001069

[13] Hassani H, Ghodsi Z. Pattern Recognition of Gene Expression with Singular Spectrum Analysis. *Medical Sciences*.2014; **2**(3): 127-139.

[14] Hassani H, Alharbi N, and Ghodsi M. A short Note on the Pattern of Singular Values of a Scaled Random Hankel Matrix. *International Journal of Applied Mathematics*. 2014; **27**(3): 237-243.

[15] Hassani H, Alharbi N, and Ghodsi M. A Study on the Empirical Distribution of the Scaled Hankel Matrix Eigenvalues. *Journal of Advanced Research*. 2014; doi:10.1016/j.jare.2014.08.008

[16] Ghodsi M, Alharbi N, and Hassani H. The Emperical Distribution of the Singular Values of a Random Hankel Matrix. *Fluctuation and Noise Letters*. 2015; **14**(3).

[17] Hassani H, Alharbi N, and Ghodsi M. Distinguishing Chaos from Noise: A New Approach. *International Journal of Energy and Statistics*. 2014; **2**(2): 137-150.

[18] Ghodsi Z, Silva E S, and Hassani, H. Bicoid Signal Extraction with a Selection of Parametric and Nonparametric Signal Processing Techniques. *Genomics, proteomics & bioinformatics*. 2015; doi:10.1016/j.gpb.2015.02.006.

[19] Golyandina N, Nekrutkin V, and Zhigljavsky A. *Analysis of Time Series Structure: SSA and Related Techniques*. Chapman and Hall/CRC. 2001.

[20] Available online: http://urchin.spbcas.ru/flyex/ (accessed on Feburary 2014).

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

[22] Golyandina N, and Zhigljavsky A. *Singular Spectrum Analysis for Time Series*. Springer Briefs in Statistics. Springer. 2013.

[23] Lahiri K, Vaughan D R, and Wixon B. Modeling SSA's sequential disability determination process using matched SIPP data. *Social Security Bulletin*.1995; **58**: 3.