Optimizing Bicoid Signal Extraction

Abstract

Signal extraction and analysis is of great importance not only in fields such as economics and meteorology but also in genetics and even biomedicine. There exists a range of parametric and nonparametric techniques which can perform signal extractions. However, the aim of this paper is to define a new criterion for optimising signal extraction from bicoid gene expression profile. Having studied both parametric and nonparametric signal extraction techniques, we identified the lack of specific criterion to enable users to select the optimal signal extraction parameters. Exploiting the expression profile of *bicoid* gene, which is a maternal segmentation coordinate gene found in *Drosophila melanogaster*, we introduce a new criterion for optimising the signal extraction with a nonparametric technique. This criterion is based on the distribution of the residual, more specifically its skewness.

Keywords: signal extraction; Optimisation; Residual distribution; Bicoid.

1 Introduction

Signal extraction is an important and challenging task in the field of time series analysis and forecasting. Signals can take various forms with the most common being trends and seasonal fluctuations. Trend extraction in particular enables analysts to smooth out a time series and remove the seasonal and cyclical variations so as to determine the long-run behaviour of the underlying data. A trend can be formally defined as a smooth additive component which contains information relating to the global change in a time series [4], and the term 'smooth' is a vital characteristic of any given signal. In the field of genetics and gene expression studies, signal extraction and noise reduction are crucial as genetic data is often characterised by the existence of considerable noise [5].

Our interest in this topic is motivated by the findings in [5] where the authors evaluated a variety of parametric and nonparametric signal processing techniques for extracting the signal in bicoid (*bcd*)[1], which is a morphogen localised at the anterior end of the egg. After fertilisation, the distribution of Bcd along the embryo –the signal under study in this paper– determines the cell's destiny in a concentration-dependent mode. Here, the authors found that a nonparametric approach produced the most efficient extraction of the Bcd signal [5].As Ghodsi et al. [5] point out, the Bcd signal extraction process is

[1]In what follows, the italic lower-case *bcd* presents either the gene or the mRNA and Bcd refers to the protein.

complex as the data associates with both observational and biological noise, and the extracted residual is not normally distributed as required by parametric techniques. Figure 1 below shows an example of a typical noisy Bcd. As noted in [5], the distribution of Bcd follows an exponential trend, and the high volatility seen in the profile ensures that the extraction of this signal remains an arduous task.

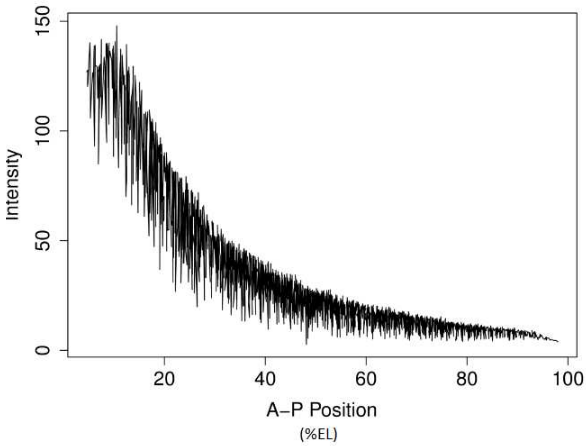

Figure 1: A typical example of noisy Bcd [9].

The aim of this paper is to introduce and define a new criterion for optimising Bcd signal extraction. At present, there exist no definitive criterion to aid researchers and scientists interested in extracting the Bcd signal for analysis. Since the Bcd signal defines what positional information is available for morphogen readout, studying the characteristics of this signal expects to improve our knowledge on several critical developmental processes such as embryogenesis, regional specification and canalisation. It should be noted that the set criterion is tailored for the sole purpose of extracting an accurate Bcd signal based on the knowledge disseminated through the work in [5] with regard to the distribution of the residual following Bcd signal extraction. Therefore, the criterion presented herewith may not be directly suitable for other applications.

In addition to covering the main aim of this paper, we also present readers with two other interesting concepts related to Bcd expression profile. These are sequential and hybrid signal extraction processes which are explained in Section 2. Accordingly, this paper is able to present readers with three different approaches for Bcd signal extraction based on their requirements and interests. The first approach is suitable for those who wish to rely on a single model for Bcd signal extraction. We have tailored the criterion presented in this paper to enable a swift and accurate Bcd signal extraction using the nonparametric approach identified as best in [5]. Should the extracted signal appear to have captured some unnecessary fluctuations, then the sequential process described can be applied on the original signal to generate a refined and smoother signal line. Even though the findings in [5] suggests that the Bcd residual is skewed, we appreciate that statisticians who subscribe to classical methods would find it difficult to agree with such outcomes. Therefore, as a second approach, we propose a hybrid parametric signal extraction process which can ensure that the residual is in fact white noise. Finally, for those who wish to exploit hybrid modelling from a purely nonparametric perspective with the possibility of capturing the maximum variation via a smooth signal line, we present the hybrid nonparametric approach and show that it can produce far better results when combined with the optimized signal extraction criteria presented herewith. The above three approaches also represent the core contributions of this research.

The remainder of this paper is organised such that Section 2 focuses on optimising Bcd signal extraction with Section 3 presenting the empirical results. This is followed by an interesting discussion in Section 4, and the paper concludes in Section 5.

2 Optimising Bcd Signal Extraction

2.1 Singular Spectrum Analysis

The Singular Spectrum Analysis (SSA) technique is a nonparametric filtering technique that is dependent upon its choice of Window Length L and the number of eigenvalues r. SSA was successfully introduced for Bcd signal extraction in [16] and exploited in more detail in [5]. This particular study found that the residual following signal extraction in Bcd is not normally distributed or stationary, and also showed that the residual itself has a complex pattern which adds further to the difficulty in smoothing and signal extraction. However, SSA is unique as it can extract several signals for any given time series depending on the chosen value of L. In fact, the choice could be any L such that $2 \leq L \leq N/2$ where N is the length of the series. As such, the findings in [5] which shows SSA as the best option for Bcd signal extraction (in relation to Synthesis Diffusion Degradation, Exponential Smoothing, Autoregressive Integrated Moving Average (ARIMA), Fractionalized ARIMA, and Neural Networks) falls short of defining the optimal SSA model choices for Bcd signal extraction.

Through our work we intend to fill this gap by introducing a new criterion which enables optimisation of the Bcd signal extraction process with SSA. The importance of defining such a criterion is further evidenced by the fact that SSA has been applied for extracting the Bcd and other segmentation gene's signal since 2006, see for example [5, 12–17]. Therefore, it is clear that researchers and scientists alike can benefit from some formal criterion for the selection of SSA choices when using same for Bcd signal extraction. Whilst the remainder of this paper focuses entirely on SSA, we find it pertinent to acknowledge and comment on the comparative preferability of SSA over other filtering techniques such as Hilbert-Huang (HH) [18] and Hodrick-Prescott (HP) [19]. Firstly, the SSA technique (as detailed below) is a Singular Value Decomposition based method and as such is very effective for noise reduction [20]. Secondly, the HH approach is closely associated with Empirical Mode Decomposition which is related to the setting of intrinsic mode functions. Thirdly, the signal process in the HP filtering approach has two instead of one unit root and is therefore most suitable for time series with two unit roots [20]. A direct comparison of both SSA and HP under equal conditions showed that SSA performs on par with the HP filter [20].

The basic SSA technique consists of two complementary stages referred to as decomposition and reconstruction, and each of these stages includes two separate steps [21]. In brief, at the first stage the Bcd is decomposed into the sum of a small number of independent and interpretable components such as a slowly varying trend and a structureless noise [5,21], and at the second stage the noise free Bcd is reconstructed [5,22]. It should be noted that the use of SSA in this paper is solely intended towards obtaining the optimal decomposition of Bcd using SSA and then extracting the signal component alone. Figure 2 summarises the basic SSA process as a flowchart.

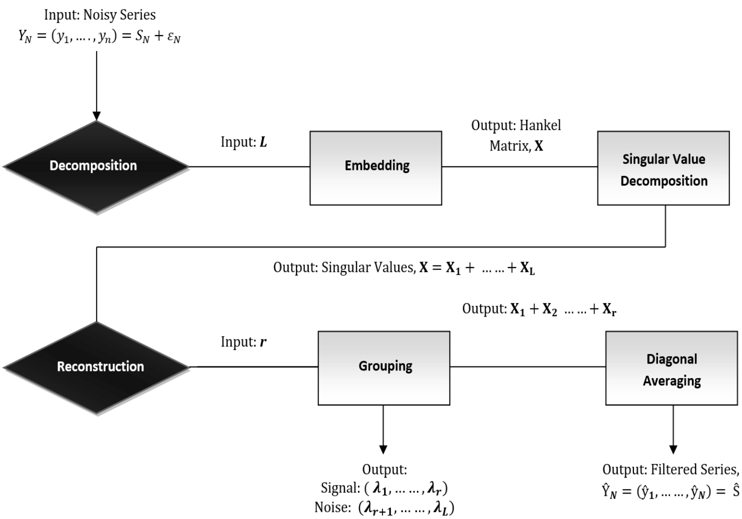

Figure 2: A flowchart of the basic SSA process. Figure adapted from [21].

A more detailed explanation of the steps underlying SSA for bicoid signal extraction is provided below, and in doing so we mainly follow [5,21].

The first step maps a one dimensional time series $Y_N = (y_1, \dots, y_N)$ into a multi-dimensional series $X_1, \dots, X_K$ with vectors $X_i = (y_i, \dots, y_{i+L-1})^T \in \mathbf{R}^L$, where $K = N - L + 1$. Whilst the process itself is referred to as embedding, the vectors X_i are called *L-lagged vectors*. The single choice of the embedding stage is the Window Length L, which is an integer such that $2 \leq L \leq N/2$. This step results in the trajectory matrix $\mathbf{X}$, which is also a Hankel matrix and takes the form: $\mathbf{X} = [X_1, \dots, X_K] = (x_{ij})_{i,j=1}^{L,K}$.

Thereafter, we obtain the singular value decomposition (SVD) of the trajectory matrix and represent it as a sum of rank-one bi-orthogonal elementary matrices. The eigenvalues of $\mathbf{XX}^T$ are denoted by $\lambda_1, \dots, \lambda_L$ in decreasing order of magnitude ($\lambda_1 \geq \dots \lambda_L \geq 0$) and by $U_1, \dots, U_L$ the orthonormal system. Then, we set

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

If we denote $V_i = \mathbf{X}^T U_i/\sqrt{\lambda_i}$, then the SVD of the trajectory matrix can be written as:

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

where $\mathbf{X}_i = \sqrt{\lambda_i} U_i V_i^T$ $(i = 1, \ldots, d)$. The matrices $\mathbf{X}_i$ are elementary matrices as they have rank 1, U_i and V_i denotes the left and right eigenvectors of the trajectory matrix. The collection $(\sqrt{\lambda_i}, U_i, V_i)$ is called the i-th eigentriple of the matrix $\mathbf{X}$, $\sqrt{\lambda_i}$ $(i = 1, \ldots, d)$ are the singular values of the matrix $\mathbf{X}$ and the set $\{\sqrt{\lambda_i}\}$ is called the spectrum of the matrix $\mathbf{X}$. The expansion (1) is said to be uniquely defined if all the eigenvalues have a multiplicity of one. The process of splitting the elementary matrices $\mathbf{X}_i$ into several groups and summing the matrices within each group is called grouping and transfusing each resultant matrix from grouping step to a less noisy series is called diagonal averaging.

As specifically noted in [5], in general the first eigenvalue corresponds to the trend of a given time series when using SSA. In order to illustrate this more clearly to the reader, we show a couple of examples in Figures 3 and 4. Moreover, in [10, 11] the authors extract and illustrate the trend for tourist arrivals using SSA based decomposition and the first eigenvalue. Thus, we extract the first eigenvalue alone and consider the remainder as noise, and then perform diagonal averaging to transform the matrix containing the first eigenvalue into a series which will now provide the extracted signal from Bcd.

2.1.1 New Approach for Optimising Bcd Signal with SSA

In this section we present the new approach for optimising Bcd signal extraction with SSA and provide justification for the process. The proposed criteria are developed as follows.

1) The extracted Bcd trend must be smooth. This is in accordance with the widely accepted definition of a trend which states that it must be a 'smooth' additive component [4].

2) Setting L sufficiently large enables the first eigenvalue, i.e. $r = 1$ (in some cases, $r = 1, 2$) to extract a smooth signal for a given series, however the value of L must not be too small or too large. By theory, L must lie between $2 \leq L \leq N/2$ [21]. Yet, when it comes to Bcd signal extraction, setting L at $N/2$ can have negative implications, as with setting L too small.

For example, let us first consider the scenario in Figure 3 whereby in a series with length 301 we consider SSA choices of $L = 2$ and $r = 1$ for Bcd signal extraction. Notice how the extracted signal fails to meet the 'smooth' criteria as per the definition of a signal in [4]. Accordingly, it is evident that setting L too small fails to achieve an optimal signal extraction with SSA for Bcd.

Secondly, let us consider what happens when we set L too large for the same data set. Here, the maximum possible value of L is 150. As such, we set $L = 150$ and seek to extract the signal in our data. Figure 4 shows the resulting outcome. In this case, notice how the signal line is smooth (confirming that

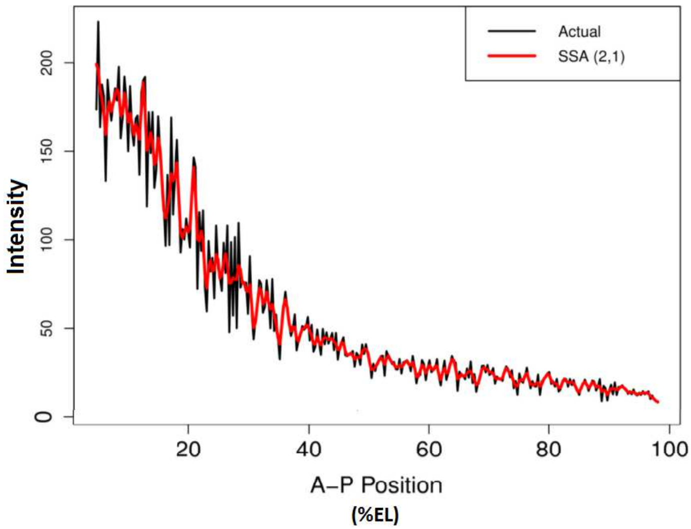

Figure 3: signal extraction from noisy Bcd with SSA choices of $L = 2$ and $r = 1$.

setting L large can provide a smoother line) but the extracted signal fails to fit well to the actual data especially towards the tail of the series.

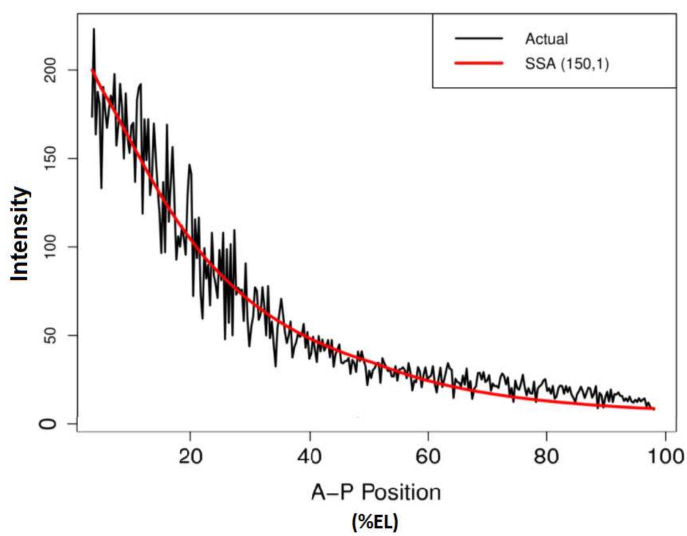

Figure 4: signal extraction from noisy Bcd with SSA choices of $L = 150$ and $r = 1$.

3) Based on points 1) and 2), we suggest the following threshold for the selection of L for Bcd signal extraction purposes. The window length L should be some value between $10 \leq L \leq N/4$. Whilst this assumption helps restrict the selection of L, on its own it fails to provide the researcher with an exact value for L. Therefore, we call upon the nonparametric nature of SSA to provide the final closing argument for the criteria.

4) As a nonparametric technique, the SSA residual can be skewed. Based on the findings in [5] which was an extensive study into signal extraction in Bcd, the residual from the process was in fact found to be skewed. As such, we propose using the skewness statistic as an indicator, and finding L which corresponds to the minimum skewness for a given Bcd series within the threshold $10 \leq L \leq N/4$ and coupling this with $r = 1$ or $r = 1, 2$ as appropriate optimal Bcd signal extraction with SSA.

2.2 Sequential and Hybrid Signal Extraction

Section 4 in this paper is dedicated to a discussion which focuses on the exploitation of Sequential SSA and a hybrid signal extraction process for Bcd signal extraction. In what follows we present the ideas that are evaluated later on with empirical data.

2.2.1 Nonparametric Approach

Signal extraction in Bcd data can be an arduous task owing to the complex structure portrayed by the data [5]. Sequential SSA is a relatively new concept which is of great benefit when faced with weak separability between signal and noise as a result of such complexities. For example, when faced with problems in separating a signal of complex form and seasonality, Sequential SSA can be exploited to obtain a more accurate decomposition from the residual after signal extraction [23]. Whilst historically, Sequential SSA was performed on a residual, in this paper we suggest the use of Sequential SSA for refining the Bcd signal further.

The basic idea underlying Sequential SSA is to perform a second round of SSA based decomposition and reconstruction on data that has already undergone an initial round of SSA, with the aim to refine the signal of interest further. Suppose that we exploit the optimised Bcd signal extraction algorithm explained above and extract some signal line. However, if the Bcd data in question has a highly complex structure, it is possible to end up with a signal line that is not as smooth as one would like. In such instances, we suggest exploiting Sequential SSA, not on the residual, but on the extracted signal to smooth it further and obtain a new and refined signal curve. This approach is greatly beneficial to those who wish to rely on a single model for Bcd signal extraction and enjoy the benefits of a nonparametric technique.

2.2.2 Hybrid Signal Extraction with SSA

It is possible that some statisticians may not be convinced or used to subspace-based methods such as SSA. Therefore, we find it pertinent to present the possibility of obtaining a hybrid signal extraction process which will combine the optimised SSA signal extraction algorithm for Bcd with other automated signal processing techniques from both parametric and nonparametric backgrounds.

The basic idea underlying the hybrid signal extraction process is as follows:

1. Extract the Bcd signal via the optimised SSA signal extraction algorithm.

2. Fit a different time series model to the residuals following SSA signal extraction and obtain the fitted values.

3. Add the fitted values to the original SSA signal to create the Hybrid SSA signal.

2.2.2.1 Hybrid SSA Signal: Parametric Approach The idea underlying the hybrid SSA signal with a parametric approach is to combine the nonparametric SSA signal with the fitted values on residuals from a parametric signal processing model. As most classical statisticians welcome and subscribe to the ARIMA model, here we choose an automated ARIMA model as provided via the forecast package in R [24]. It is important to note that in this paper we do not rely on ARIMA for its forecasting capabilities. Instead, we consider ARIMA as a tool for extracting any hidden signals within the residual following the initial filtering by SSA. This in turn enables one to ensure that the residual is indeed white noise, as required by parametric models. This approach is useful as it ensures that the residual following hybrid signal extraction will indeed be white noise.

The modelling equations for ARIMA relevant to this study can be described by following [25]. A non-seasonal ARIMA model may be written as:

$$(1-\phi_1 B - \ldots - \phi_p B^p)(1-B)^d y_t = c + (1+\theta_1 B + \ldots + \theta_q B^q)e_t, \qquad (2)$$

where B is the backshift operator, c is a constant, p is the order of the autoregressive part, q is the degree of first differencing, d is the order of the moving average part of the model, and e_t is white noise [25]. In the R software, the inclusion of a constant in a non-stationary ARIMA model is equivalent to inducing a polynomial signal of order d in the forecast function.

2.2.2.2 Hybrid SSA Signal: Nonparametric Approach Whilst the underlying idea remains the same, in this instance, as opposed to relying on a parametric time series analysis model, we can combine the nonparametric and optimised SSA signal with fitted values on residuals from a nonparametric time series analysis model in order to obtain the hybrid SSA signal. The benefits of this approach would be that it enables to overcome the parametric restrictions of normality and stationarity of residuals of which the former condition was found to be irrelevant in the case of Bcd data where the residual following signal extraction is skewed according to [5]. In this case we rely on the automated Exponential Smoothing (ETS) model found in the forecast package in R.

Those interested in the several ETS formula's that are evaluated through the forecast package when selecting the best model to fit the residuals are referred to Chapter 7, Table 7.8 in [25].

3 Empirical Results

3.1 Data

The evaluation in this study is performed on 17 *Drosophila melanogaster* embryos introduced by Alexandrov et al. [1] which was originally obtained from FlyEx database [6, 7]. This dataset has been widely used as a valuable source of information for studying the dynamics of segment determination of early *Drosophila development* [8].

In FlyEx, the quantitative Bcd data was obtained using the confocal scanning microscopy of fixed embryos immunostained for segmentation proteins [2]. To that aim, A 1024 1024 pixel confocal image with 8 bits of fluorescence data was achieved for each embryo which 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 anterior to a posterior position along the length of the egg expressed as the percentage, and y-axis shows the intensity levels which correspond to the amount of expressed *bcd* gene.

It is of note that in the study conducted by Alexandrov et al. the out of focus regions were removed by excluding the utmost anterior and posterior areas. After removing the upper and lower values, to get a complete profile along the A-P axis of the embryo, a curve was fitted to the interval of the A-P coordinate between 20 and 80% of egg length (a complete explanation of the method and biological characteristics of this data can be found in [1, 3]). However, to introduce a signal processing method capable of both noise filtering and signal extraction, this paper considers the whole data which is unprocessed for any noise reduction methods.

3.2 Signal Extraction

Here, we consider real Bcd data and seek to extract the signal with SSA using the newly proposed criteria as outlined in Section 2.1. Figure 5 below portrays a selection of the actual data and extracted signal with the optimized SSA algorithm, and also outlines the SSA choices which have been used in each case. For the examples in Figure 5, note how the extracted signal is not only smooth, but also well centred around the data, thereby providing the reader with a very accurate outlook for the long term prospects of the Bcd gradient. However, it is evident that on its own, SSA appears to have difficulties in accurately capturing the signal curve initially when it is faced with very high levels of fluctuations as clearly visible within the first few observations of the Bcd profile. We consider this aspect further in the discussion which follows in Section 4.

Even though signal extraction is the primary focus of this study, it is no secret that the residual can often enlighten us to crucial information pertaining to any given data set. As such, we follow up the signal extractions with a sound residual analysis.

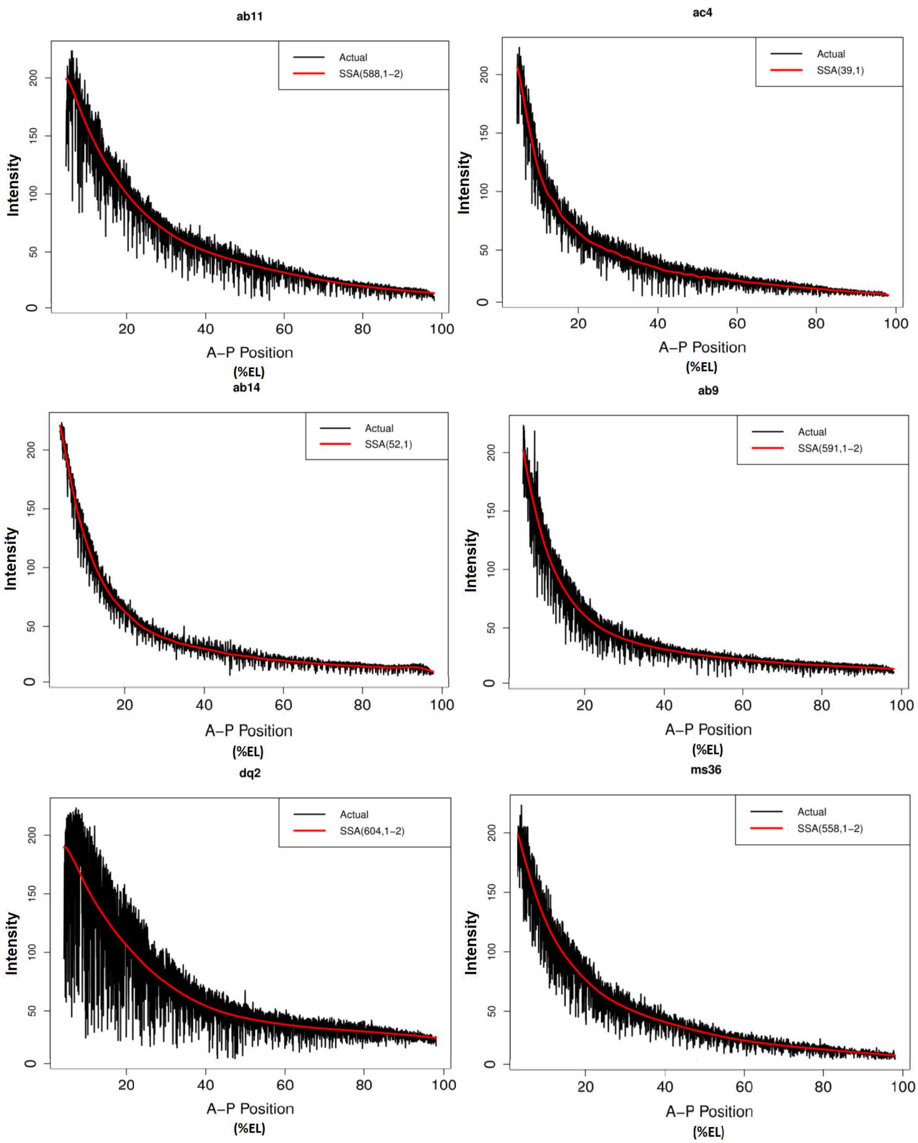

Figure 5: Optimised signal extraction with SSA for a selection of Bcd data.

3.3 Residual Analysis

In order to save space, via Figure 6 we only show the residuals corresponding to the signal extractions shown in Figure 5. A first look at the structure and distribution of the residual over time helps us understand the difficulty in extracting the signal from Bcd profiles. This is largely to do with the the highly volatile nature of the data which results in fluctuating amplitudes over time in a particular pattern. In fact, the general patterns appears such that all residuals portray amplitudes which are initially high and then gradually decrease. This in turn means that the techniques adopted for Bcd signal extraction should be able to cope well with such variation and fluctuations in data if it is to accurately perform its task. Moreover, it appears to the naked eye that there is indeed some signal contained within these residuals. Whilst it is expected that a residual following signal extraction would result in capturing the other signals, in some instances there also appears to be a small signal pattern hidden within this data.

However, as visual inspections fall short of providing sound evidence, we also consider some statistics for analysing the residuals further. These are reported via Table 1 for all the Bcd data considered in this study. The residuals are initially tested for normality via the Kolmogorov-Smirnov (KS) test for normality. The choice of KS test was as opposed to using the popular Shapiro-Wilk (SW) test for normality was because when faced with large samples the KS test is likely to be comparatively more accurate than the SW test [26]. As expected, all residuals failed to pass the normality test reporting probability values of less than 0.001, and thereby leading to a rejection of the null hypothesis of normality. This lets us conclude with 99% confidence that the Bcd residuals following signal extraction are in fact skewed and these results are consistent with the findings in [5].

Finally, we go a step further and fit optimal ARIMA models [25] to the residuals. This was done in order to ascertain the randomness of the residuals following Bcd signal extraction with optimised SSA. Statisticians who rely on classical signal extraction techniques would be overly concerned with the parametric assumptions of normality and stationarity of the residuals. Whilst we have assessed the normality of residuals via the KS test and justified based on [5] that the residuals from this signal extraction exercise should be skewed, fitting of optimal ARIMA models enables us to easily show whether the residuals meet the stationary criteria. We fit automated and optimised ARIMA models (as provided via the forecast package in R) on the residuals and report the outcomes in Table 1. A non-seasonal ARIMA model is represented in the form ARIMA(p, d, q) where p indicates the order of the autoregressive parts, d the degree of first differencing and q the order of the moving average part of the model [25]. If the data is non-stationary, then within the ARIMA(p, d, q) process the value of $d \geq 1$. If the data is stationary, then no differencing is required, and so $d = 0$. In this case, we notice that $d = 0$ in all instances, and thereby proves that the residuals are indeed stationary.

However, the fitting of ARIMA models on the residuals also highlight another interesting point. Notice how for 27 Bcd residuals there have been a variety of 14 different ARIMA models which have been fitted. This in turn indicates the complexity and difficulty associated with the selection of a single technique for extracting Bcd signal, and most certainly highlights the difficulties which any technique would when seeking to extract a signal from data with such complex fluctuations. In addition, except for where the model reads $ARIMA(0, 0, 0)$, in all other instances we notice that the residuals are not white noise. We discuss this, and provide a possible solution within the discussion.

Table 1: Residual analysis for Bcd signal extractions.

Embryo	n	SW	ARIMA
ab2	138	<0.001	ARIMA(0,0,1) with zero mean
hz15	85	<0.001	ARIMA(0,0,0) with zero mean
hz28	79	<0.001	ARIMA(2,0,2) with zero mean
ad14	301	<0.001	ARIMA(2,0,5) with zero mean
ad22	294	<0.001	ARIMA(4,0,3) with zero mean
ad23	308	<0.001	ARIMA(1,0,3) with non-zero mean
ab17	485	<0.001	ARIMA(1,0,3) with non-zero mean
ad4	556	<0.001	ARIMA(4,0,4) with zero mean
ad6	566	<0.001	ARIMA(2,0,2) with non-zero mean
ab12	2284	<0.001	ARIMA(4,0,2) with zero mean
ab10	2263	<0.001	ARIMA(1,0,2) with zero mean
ac5	2404	<0.001	ARIMA(4,0,4) with non-zero mean
ab1	2570	<0.001	ARIMA(4,0,4) with zero mean
ac7	2268	<0.001	ARIMA(1,0,2) with zero mean
ad13	2235	<0.001	ARIMA(4,0,2) with non-zero mean
ad29	2193	<0.001	ARIMA(1,0,2) with zero mean
ad32	2183	<0.001	ARIMA(2,0,1) with zero mean
ab7	2346	<0.001	ARIMA(1,0,2) with zero mean
ac3	2356	<0.001	ARIMA(0,0,1) with zero mean
ac9	2215	<0.001	ARIMA(4,0,1) with zero mean
ms14	2305	<0.001	ARIMA(4,0,2) with zero mean
ab11	2355	<0.001	ARIMA(4,0,2) with zero mean
ac4	2383	<0.001	ARIMA(3,0,1) with zero mean
ab14	2218	<0.001	ARIMA(1,0,2) with zero mean
ab9	2369	<0.001	ARIMA(2,0,1) with zero mean
dq2	2423	<0.001	ARIMA(2,0,4) with zero mean
ms36	2239	<0.001	ARIMA(5,0,1) with zero mean

4 Discussion

4.1 Sequential SSA on Bcd signal

Note how the signal extraction in ac3, Figure 7, appears to have captured some other fluctuations apart from the signal alone. As such, this extraction, in particular, fails to meet our criteria for a smooth signal. When faced with such situations, we are able to find a solution via sequential SSA. Sequential SSA enables users to take the extracted signal (the signal in our example) and filter same with SSA once more to obtain a more refined output. In what follows we have applied Sequential SSA on the initially extracted Bcd signal.

As visible via Figure 8, following sequential SSA we have been able to extract a smoother signal. In this instance, we used the signal extracted via the opti-

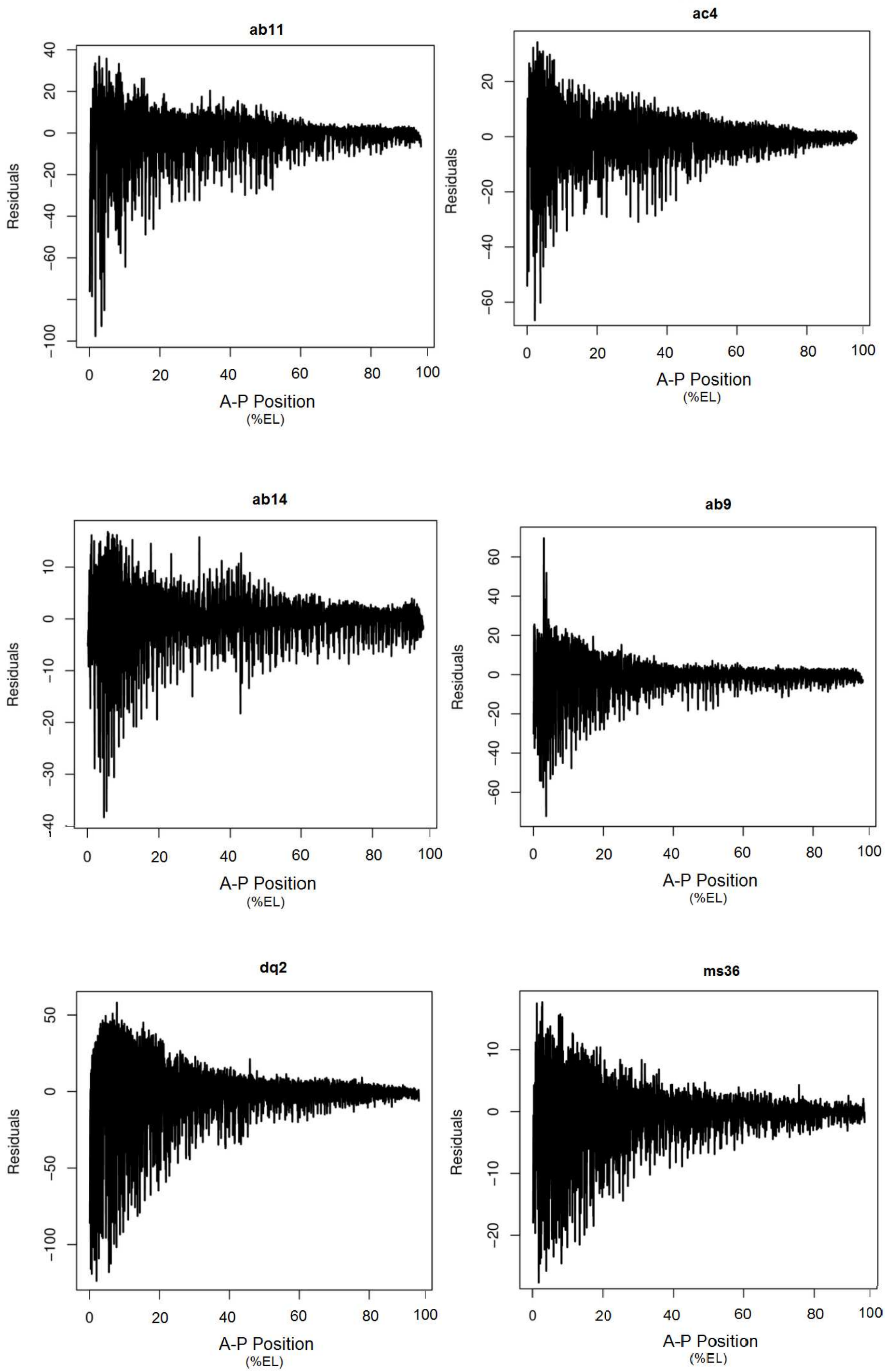

Figure 6: Residuals following optimised signal extraction with SSA for a selection of Bcd data.

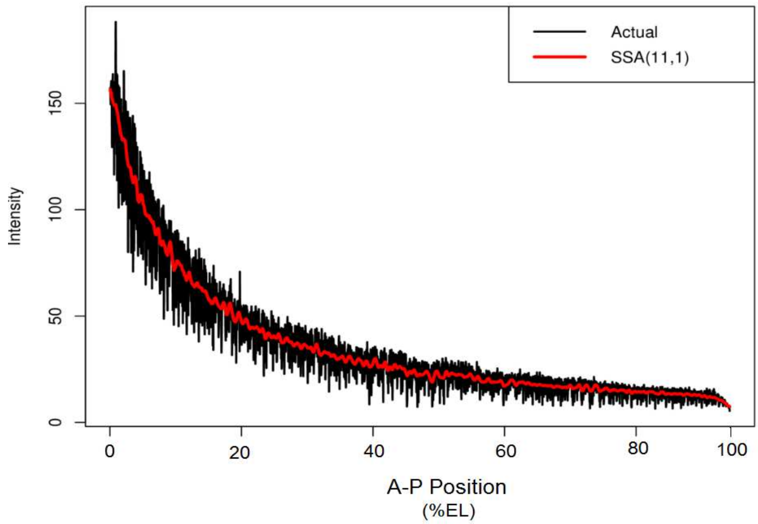

Figure 7: SSA based optimal trend extraction for ac3.

mised SSA signal extraction algorithm for Bcd and refined this signal further via Sequential SSA. Here we have used $L = N/2$ and $r = 1$ for signal extraction with Sequential SSA. In line with good practice, the residual was once again tested for normality via the KS test which indicated that the residual is skewed at a 1% significance level, and fitting of an ARIMA model showed that the residual is stationary as well.

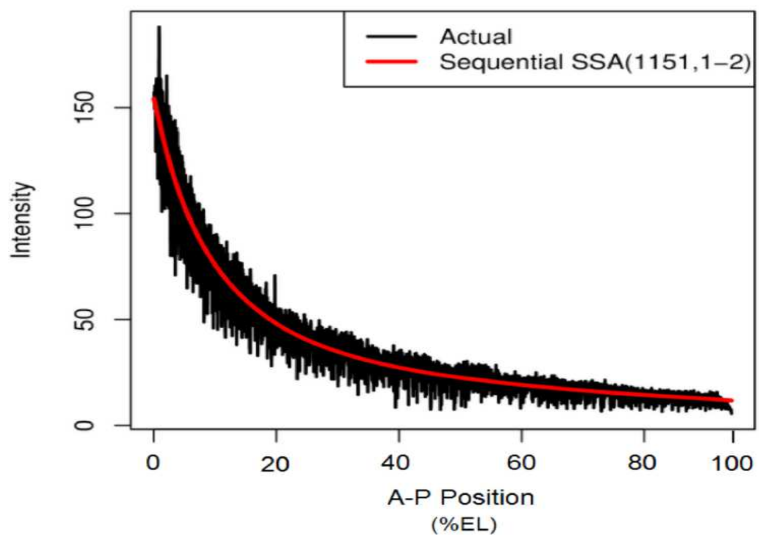

Figure 8: Refined signal extraction with sequential SSA on ac3 signal.

4.2 Hybrid SSA signal Extraction for Bicoid

4.2.1 Hybrid SSA signal: Parametric Approach

The residual analysis in Table 1 indicates that ARIMA models could be fitted to all but one of the residuals following signal extraction with the optimised SSA signal algorithm. This means that only one of the residuals are pure white noise as it stands. Whilst some might argue that this is acceptable given that the objective is to extract the signal component alone, there may be others who subscribe to an alternate view along the lines of obtaining a random residual following signal extraction. The first hybrid SSA signal approach we present is one which enables users who wish to obtain white noise achieve this following Bcd signal extraction with SSA. We begin by fitting the ARIMA models as identified via Table 1 to the data and extract the fitted values which are then combined with our original SSA Bcd signal to create a hybrid SSA-ARIMA signal for Bcd. We consider the examples discussed in text so far and generate the following results. Figure 9 shows the hybrid SSA-ARIMA signals for Bcd data. In comparison to the optimised SSA signals in Figure 5, the hybrid SSA signal with ARIMA fit fails to meet the smooth criteria. As such, it is evident that on its own, the hybrid SSA-ARIMA approach is only beneficial for those who wish to capture all the signal in the data whilst ensuring that the residual following Bcd signal extraction is white noise. It clearly comes at a high cost of lost smoothness in signal curves. However, it is of note that as previously mentioned, noise in gene expression data enters not only from the data acquisition and processing procedures [27] but also the fluctuations seen in an expression pattern can be a consequence of biological noise which may also introduce error into the data [28]. Therefore, the source of the natural biological variability is different from the experimental noise [28]. Biological noise arises from the active molecular transport, compartmentalization, and the mechanics of cell division [29]. Therefore, the hybrid SSA with the ARIMA model can be applied in studies such as segmentation network analysis where the combination of Bcd signal with its biological noise needs to be considered as an input to the system.

4.2.2 Hybrid SSA signal: Nonparametric Approach

Here, we apply the same process as above, but instead of ARIMA, we rely on the nonparametric time series analysis model of ETS. This enables the entire hybrid SSA signal approach to remain nonparametric in nature. The resulting hybrid SSA signals with ETS fit are shown via Figure 10.

There is an interesting point to note here. In comparison to the parametric hybrid signal extraction approach, it is clear that the nonparametric hybrid approach has resulted in much smoother signal curves as one would expect and like to see following a signal extraction exercise. As such, out of the two hybrid approaches, for the purposes of Bcd signal extraction, it is likely that users will prefer the nonparametric approach over the parametric approach.

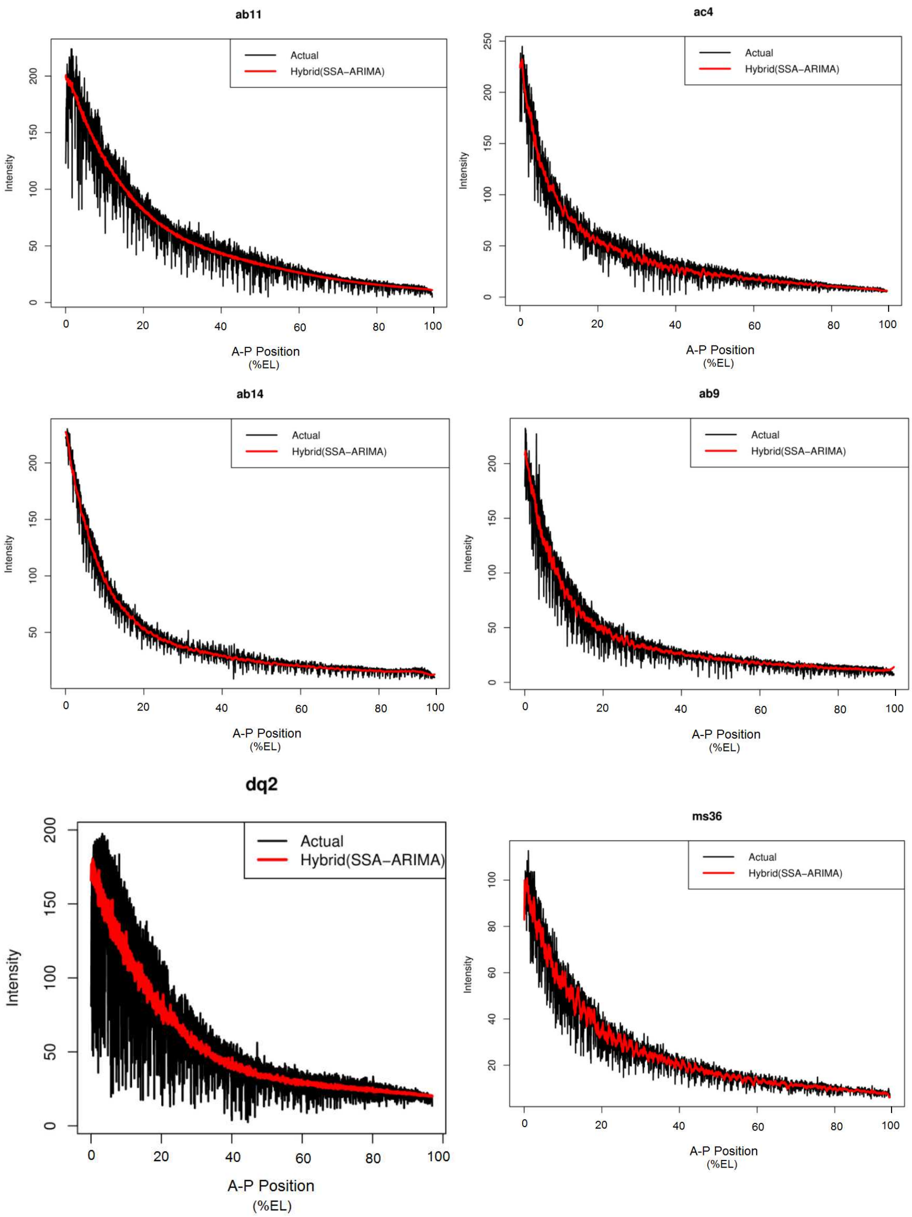

Figure 9: Hybrid SSA signal with ARIMA fit for Bcd data.

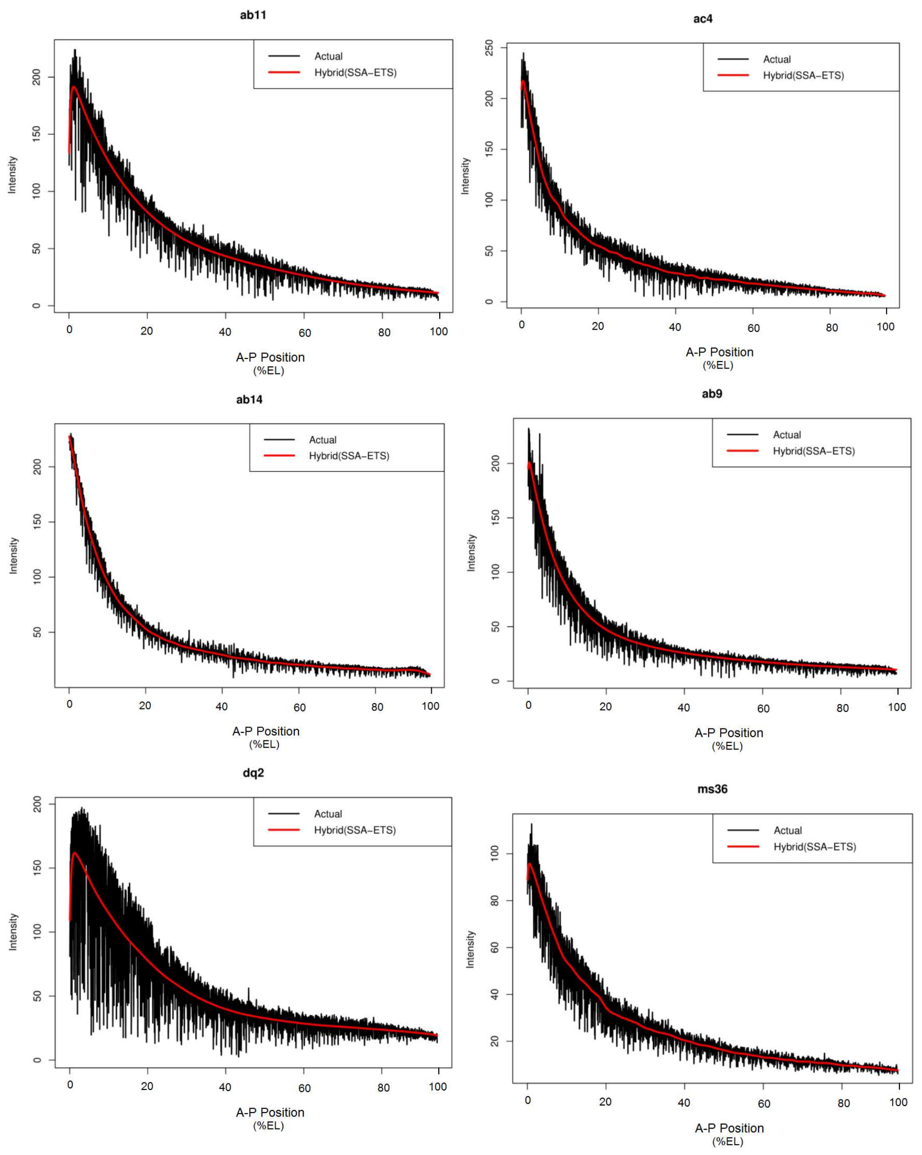

Figure 10: Hybrid SSA signal with ETS fit for bicoid data.

5 Conclusion

This paper begins with the core aim of introducing new criteria for optimising Bcd signal extraction. Motivated by the findings in [5], we opt to tailor the new Bcd signal extraction criteria for use with the Singular Spectrum Analysis technique which Ghodsi et al. [5] found to be the best option for Bcd signal extraction in relation to SDD, ARIMA, ETS, ARFIMA and NN models. In line with our aim, we initially produce an algorithm for optimising the Bcd signal extraction process with SSA. In brief, the algorithm is optimised based on minimising the skewness statistic for the SSA residual. We suggest that setting L equal to the minimum skewness within the threshold $10 \geq L \geq N/4$ and combine this SSA choice with $r = 1$ or $r = 1, 2$ as appropriate will enable users to obtain the optimal Bcd signal extraction with SSA.

Through this research, we have succeeded in presenting several contributions to the field of Bcd signal extraction. The first and most important of which deals with the application of the newly proposed algorithm to 27 real Bcd data to show that it can enable researchers to select the appropriate SSA choices to extract a smooth and accurate Bcd signal quickly and easily without the need to spend an increased amount of time for the selection of L for decomposing the data. However, we notice that given the highly complex nature of the Bcd data, on one occasion the SSA algorithm fails to extract an absolutely smoothed signal. As a solution to this problem, we introduce for the first time, the concept of Sequential SSA on signals which is also the second contribution of this research. Via this approach, we are able to refine and smoothen further the initial signal which had captured some of the observational and biological noise in Bcd data.

In line with good practice, in addition to evaluating the signal extractions alone, this study also pays attention to the residuals. The analysis of the residuals motivated us to introduce hybrid SSA based signal extraction processes for Bcd. In brief, when extracting the signal from any given data set, one would reasonably expect other signals to end up within the noise component. However, this would mean that the residual is no longer random and some statisticians could find it difficult to accept such techniques. Accordingly, the first hybrid SSA signal process (and the third contribution) is focussed on providing a Bcd signal extraction procedure which will ensure the residual is white noise. This was achieved by combining the optimised SSA signal with optimised ARIMA models being fitted to the residuals. Whilst the results did provide the necessary outcomes in terms of residuals with white noise, it comes at a cost - i.e., a loss in the smoothness of the extracted signal.

The SSA-ARIMA hybrid approach is a combination of parametric and nonparametric techniques. For those who wish to rely on nonparametric techniques alone so that one is not restricted by the parametric assumptions, we present the SSA-ETS hybrid Bcd signal extraction approach. This process also produces the fourth and second most important contribution of this research as we find a solution to the problem of modelling accurately the initial curve in Bcd data which was not only experienced in this paper when we employed the optimised SSA signal extraction process, but was also experienced in [5]. Accordingly, we are able to present the hybrid SSA-ETS process which is a combination of the optimised SSA signal extraction algorithm with an optimised ETS algorithm as the most efficient approach for Bcd signal extraction.

We believe that the findings of this research and the information contained within this paper opens up several avenues for future research. For example, future research should evaluate the possibility of optimizing the SSA signal extraction process based on different criteria in order to determine whether a more improved signal extraction can be produced. For example, as we are seeking to introduce a novel approach for optimizing Bicoid signal extraction, in this paper we have relied on a binary decomposition. However, future studies could consider the Colonial Theory based approach to decomposition as presented in [30]. In addition, more extensive research into hybrid signal extraction processes are likely to result in positive, vital and interesting outcomes as clearly shown via this paper. Researchers should evaluate a variety of different signal extraction techniques within the hybrid framework proposed in this paper to ascertain whether outcomes could be further improved.

References

[1] Alexandrov, T., Golyandina, N. and Spirov, A., 2008. Singular spectrum analysis of gene expression profiles of early Drosophila embryo: exponential-in-distance patterns. Research letters in signal processing, 2008, p.12.

[2] 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), pp.D560-D566.

[3] Surkova, S., Kosman, D., Kozlov, K., Myasnikova, E., Samsonova, A.A., Spirov, A., Vanario-Alonso, C.E., Samsonova, M. and Reinitz, J., 2008. Characterization of the Drosophila segment determination morphome. Developmental biology, 313(2), pp.844-862.

[4] Alexandrov, T. (2009). A method of trend extraction using Singular Spectrum Analysis. *REVSTAT*, **7**(1), 1–22.

[5] 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**, 183–191.

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

[7] Pisarev, A., Poustelnikova, E., Samsonova, M., & Reinitz, J. (2008). FlyEx, the quantitative atlas on segmentation gene expression at cellular resolution. Nucleic acids research, 37(suppl-1), D560-D566.

[8] 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), pp.2212-2221.

[9] Hassani, H., and Ghodsi, Z. (2014). Pattern Recognition of Gene Expression with Singular Spectrum Analysis. *Medical Sciences*, **2**(3), 127–139.

[10] Hassani, H., Silva, E. S., Antonakakis, N., Filis, G., and Gupta, R. (2017). Forecasting accuracy evaluation of tourist arrivals. *Annals of Tourism Research*, **63**, 112–127.

[11] Hassani, H., Webster, A., Silva, E. S., and Heravi, S. (2017). Forecasting U.S. tourist arrivals using optimal Singular Spectrum Analysis. *Tourism Management*, **46**, 322–335.

[12] Golyandina, N.E; Holloway, D. M.; Lopesc, F.J.P.; *et al.*. Measuring gene expression noise in early *Drosophila* embryos: nucleus-to-nucleus variability. *International Conference on Computational Science,* **2012**, *9*, 373-382.

[13] Spirov, A.V.; Golyandina, N.E.; Holloway, D.M.; *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.

[14] Holloway, D.M.; Lopes, F.J.P; da Fontoura Costa, L.; *et al.* Gene Expression Noise in Spatial Patterning: hunchback Promoter Structure Affects Noise Amplitude and Distribution in Drosophila Segmentation. PLoS Comput Biol, **2011** 7(2): e1001069. doi:10.1371/journal.pcbi.1001069

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

[16] 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: 2949-2960.

[17] Alexandrov, T.; Golyandina, N.; Sprov, A. Singular spectrum analysis of gene expression profiles of early drosophila embryo: Exponential-in-distance patterns. *Res. Lett. Signal Process* **2010**, *2008*, 5.

[18] Huang, N. E., Long, S. R., and Shen, Z. (1996). The Mechanism for Frequency Downshift in Nonlinear Wave Evolution. *Advances in Applied Mechanics*, **32**, 59–111.

[19] Hodrick, R., and Prescott E. C. (1997). Postwar U.S. Business Cycles: An Empirical Investigation. *Journal of Money, Credit and Banking*, **29** 1–16.

[20] Hassani, H., and Thomakos, D. (2010). A review on singular spectrum analysis for economic and financial time series. *Statistics and its Interface*, **3**, 377–397.

[21] Sanei, S., and Hassani, H. (2015). *Singular Spectrum Analysis of Biomedical Signals.* CRC Press.

[22] Hassani, H., Webster, A., Silva, E. S., and Heravi, S. (2015). Forecasting U.S. Tourist arrivals using optimal Singular Spectrum Analysis. *Tourism Management*, **46**, 322–335.

[23] Golyandina, N., and Shlemov, A. (2013). Variations of Singular Spectrum Analysis for Separability Improvement: Non-Orthogonal Decompositions of Time Series. *arxiv.org*. Available via: `https://arxiv.org/pdf/1308.4022.pdf`. [Accessed: 25.10.2016].

[24] Hyndman, R. J., and Khandakar, Y. (2008). Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software*, **27**:1–22.

[25] Hyndman, R. J., and Athanasopoulos, G. (2013). *Forecasting: principles and practice.* OTexts, Australia. Available via: `www.OTexts.com/fpp`.

[26] Silva, E. S., Ghodsi, M., Hassani, H., and Abbasirad, K. (2016). A quantitative exploration of the statistical and mathematical knowledge of university entrants into a UK Management School. *International Journal of Management Education*, In Press. `http://dx.doi.org/10.1016/j.ijme.2016.10.002`.

[27] Wu, Y.F., Myasnikova, E. and Reinitz, J., 2007. Master equation simulation analysis of immunostained Bicoid morphogen gradient. BMC systems biology, 1(1), p.52.

[28] Myasnikova, E., Surkova, S., Panok, L., Samsonova, M. and Reinitz, J., 2009. Estimation of errors introduced by confocal imaging into the data on segmentation gene expression in Drosophila. Bioinformatics, 25(3), pp.346-352.

[29] Spirov, A.V., Golyandina, N.E., Holloway, D.M., Alexandrov, T., Spirova, E.N. and Lopes, F.J., 2012, April. 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. In European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics (pp. 177-188). Springer Berlin Heidelberg.

[30] Hassani, H., Ghodsi, Z., Silva, E. S., and Heravi, S. (2016). From nature to maths: Improving forecasting performance in subspace-based methods using genetics Colonial Theory. *Digital Signal Processing*, **51**, 101–109.