Hindawi Publishing Corporation
Journal Name
Volume XXX, Article ID XXXXX, XX pages
doi:10.1155/2007/XXXXX

Research Article

Singular Spectrum Analysis of gene expression profiles of early Drosophila embryo: exponential-in-distance patterns

T. Alexandrov[1], **N. Golyandina**[2], **and A. Spirov**[3]

[1] *Center for Industrial Mathematics, University of Bremen, D-28334 Bremen, Germany*
[2] *Department of Mathematics and Mechanics, St. Petersburg State University, St.Petersburg, 198504 Russia*
[3] *Department of Applied Mathematics and Statistics, and Center for Developmental Genetics, State University of New York, Stony Brook, NY 11794, USA*

Correspondence should be addressed to Theodore Alexandrov, theodore@math.uni-bremen.de

Received ???; Accepted ???

Recommended by XXX

We present investigation of gene expression profiles by means of Singular Spectrum Analysis (SSA). The biological problem under investigation is the decomposition of *Bicoid* protein profiles of *Drosophila melanogaster* into the sum of a signal and noise, where the former consists of an exponential-in-distance pattern and close to constant nonspecific component, or "background". The signal processing problems addressed are: (i) trend extraction from a noisy signal, (ii) batch processing of similar data, and (iii) analytical approximation of the signal components by the sum of exponential and constant-like functions. The proposed methods are evaluated on the given 17 series.

1 INTRODUCTION

Singular Spectrum Analysis is a method intended to perform decomposition of a sequence of measurements (usually a time series) into a sum of interpretable components, such as a trend, cycles and noise [2]. SSA is recognized in geosciences, and is giving promising results in other areas, see the collection of references on the website SSAwiki: *http://www.math.uni-bremen.de/~theodore/ssawiki/*. This work presents the use of SSA for signal extraction from spatial one-dimensional gene expression data. SSA was chosen as a compromise between parametric methods like regression, which can lead to wrong results if the model is not valid, and frequency methods like filtering.

The study of the activity of diverse genes has become one of key approaches in modern functional genomics and is crucial for our understanding of embryo development. The expression of genes is traced either in time, or in space (as in our case), along different tissues and organs, or even a whole embryo. The research of gene expression is aimed at biomedical problems, but first it is systematically tested and developed on so-called model organisms. *Drosophila melanogaster* (fruit fly) is one such organism and the gene ensemble governing early events of fly embryo segmentation is one of the best studied genetic networks.This network of cross-regulating genes makes complicated patterns of their products, the segmentation factors. These patterns are directing the embryo developmental processes. Exponential-in-distance patterns are common at the very beginning of segmentation and the primary morphogenetic gradient of the protein *Bicoid* is the most known and best studied [3, 4].

The biological problem under investigation is extraction of a signal from the noisy *Bicoid* protein profile. Following [3], we assume Bicoid to generate exponential pattern. The measured protein profile contains also the smooth residual referred to as "background" [5] and the measurement noise. In general, the form of the background is still an open question [3, 4, 5]; moreover, it includes an unknown additive function depending on the confocal microscopy settings used. In this paper, we extend the model of [3] allowing the background to differ from constant. The problem is complicated by the fact that (i) the data contain outliers and (ii) the data are very noisy and the noise has unknown structure; although the noise appears to be multiplicative, the carried out study showed that it is true in very rough approximation only.

The SSA features demonstrated are: (i) signal extraction with no parametric models of signal/noise specified; (ii) robustness to outliers; (iii) taking into account a parametric model of signal, if specified; (iv) interactive analysis with control over quality of separation of signal and noise; (v) batch processing of a set of similar series; (vi) derivation of an analytical formula of the signal.

Section 2 introduces the data, SSA and the related methods used. The results of data processing are presented in Section 3. Finally, the conclusions are provided.

2 METHODS AND APPROACHES

Biological Data. Expression level of protein in wild-type fruit fly embryos (*Drosophila melanogaster*, Oregon-R) was measured using fluorescently-tagged antibodies. For each embryo a 1024×1024 pixel image with 8 bits of fluorescence data was obtained. Image processing transforms each image into an ASCII table containing a series of records (fluorescence intensity), one for each nucleus. About 1100-1300 nuclei are obtained from each image. Each nucleus is characterized by a unique identification number, the anteroposterior (AP) and dorsoventral (DV) coordinates of its centroid, and the average fluorescence level of the gene product. Because the expression of segmentation genes is largely a function of position along the AP axis, it is natural to use the AP profiles of gene expression. We use straightened data from a rectangle 50% of the DV height of the embryo, centred on the AP axis, for details see [3]. This captures approximately 700-800 nuclei. As in [3], we investigate only the interval of the AP coordinate between 20 and 80 percent egg length (%EL). For examples of plotting nuclear intensity vs. AP position, see Figure 1a,c. The considered test set consists of 17 embryo images belonging to cleavage cycle 13 and thoroughly studied in [3]. The data and software used in this study are available at *http://www.math.uni-bremen.de/~theodore/GENESSA*. For a description of the embryos, see [6], *http://flyex.ams.sunysb.edu/FlyEx/*, or *http://urchin.sbpcas.ru/FlyEx/*.

SSA. Let us describe the basic algorithm of SSA for extraction of signal from a one-dimensional series $F = (f_0, \ldots, f_{N-1})$. For the given data, f_n represents the intensity measured at the n'th nucleus inside the considered interval of 20%EL–80%EL. The first step of SSA has only the parameter L, the *window length*, $1 < L < N$, and consists of the construction of the Hankel matrix of size $L \times K$, $K = N - L + 1$, with column vectors $X_j = (f_{j-1}, \ldots, f_{j+L-2})^{\mathrm{T}}$, $j = 1, \ldots, K$, which is called the *trajectory matrix*. Note that there is a one-to-one relation between series of length N and Hankel matrices of size $L \times K$: each secondary diagonal of a Hankel matrix has equal values and produces a term of the series. The trajectory matrix is then decomposed into the sum of the ordered elementary matrices, $\mathbf{X} = \sum_{i=1}^{d} \mathbf{X}_i$, where $\mathbf{X}_i = \sqrt{\lambda_i} U_i V_i^{\mathrm{T}}$, λ_i are nonzero eigenvalues of $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ in decreasing order, U_i are the corresponding eigenvectors, and V_i are the factor vectors. This is the so-called *Singular Value Decomposition* (SVD) and each SVD component generates an *elementary reconstructed component (elementary RC)* of the series F as follows. The matrix $\mathbf{X}_i$ is hankelized by averaging the entries with indices $i + j = const$, and the corresponding series of length N is reconstructed by the above-mentioned one-to-one relation. Thus we decompose F into the sum of elementary RCs, $F = \widetilde{F}_1 + \ldots + \widetilde{F}_d$, where d is the number of nonzero eigenvalues λ_i (the so-called *SSA rank* of F). Then we choose a group $\mathcal{J}$ of r indices of the desirable components of F (signal in our case) and gather the *reconstructed signal* as $\widetilde{F} = \sum_{i \in \mathcal{J}} \widetilde{F}_i$.

The signal extraction problem is thus reduced to (i) choice of window length L and (ii) selection of the subgroup $\mathcal{J}$ of SVD components for reconstruction. These questions are briefly discussed in the next paragraph.

Trend extraction in SSA. SSA needs no a priori specification of models of series and signals, neither deterministic nor stochastic ones. In this paper, we are interested in extraction of a slowly-varying signal, usually called the *trend*. Hereinafter, we refer to trend instead of signal.

Generally, SSA is able to extract different kinds of trends. Note that any trend can be approximated by a finite-rank series as the class of finite-rank series includes all kinds of sums of products of polynomials, exponentials and sinusoids. Let us assume that the trend is (or is approximated by) a series of rank r. With large enough N and L ($L \leq N/2$), the trend is separable from noise and is reconstructed by the r leading SVD components. The subspace spanned by the r corresponding eigenvectors contains information about the finite-rank structure and, in particular, allows to derive the approximate analytical formula of the trend. If N is not large enough (for strong noise or high rank r) and separability is bad, then the trend still can be extracted using small L. In this case trend is determined by a few leading components and SSA works like a smoothing adaptive linear filter. However, the subspace spanned by the corresponding eigenvectors does not reflect the finite-rank structure of the trend and is liable to be affected by noise and outliers.

Grouping of SVD components is based on the fact that the slowly-varying component of the series generates eigenvectors and factor vectors of slowly-varying form [1, 2] and therefore is composed of similar elementary RCs. Thus, the identification of the components of a trend consists in identification (visual or automatic) of slowly-varying eigenvectors, factor vectors or elementary RCs.

AutoSSA for trend extraction. In this paragraph, we present a method of identification of slowly-varying eigenvectors. This method is easy to use as it has only two parameters (if the window length is fixed).

Firstly, we introduce the periodogram $I_Y(\omega)$ of a vector $Y \in \mathbb{R}^M$, $Y = (y_0, \ldots, y_{M-1})^{\mathrm{T}}$: $I_Y(k/M) = \frac{1}{M} \left| \sum_{n=0}^{M-1} e^{-i2\pi nk/M} y_n \right|^2$, $k = 0, \ldots, \lfloor M/2 \rfloor$, which can be interpreted as the contribution of the frequency k/M. The cumulative contribution is evaluated as $\pi_Y(\omega) = \sum_{k:0 \leq k/M \leq \omega} I_Y(k/M)$, $\omega \in [0, 0.5]$. For $\omega_0 \in (0, 0.5)$ the contribution of low frequencies to $Y \in \mathbb{R}^M$ is defined as $\mathcal{C}(Y, \omega_0) = \pi_Y(\omega_0)/\pi_Y(0.5)$.

Let us consider eigenvectors U_i. Then, given $\omega_0 \in$

$(0, 0.5)$ and $\mathcal{C}_0 \in [0, 1]$, we select SVD components with eigenvectors satisfying: $\mathcal{C}(U_i, \omega_0) \geq \mathcal{C}_0$. One may interpret this method as selection of SVD components characterized mostly by low-frequency fluctuations.

The low-frequency boundary ω_0 manages the scale of the extracted trend: the lower ω_0, the slower the trend varies. The parameter $\mathcal{C}_0$ regulates an acceptable share of higher frequencies in the extracted component and is usually chosen close to 1. For more details on selecting ω_0 and for description of a fitting procedure to choose $\mathcal{C}_0$ see [1].

Derivation of the analytical form of a component. A series consisting of a sum of exponentials and sinusoids has SSA-rank r and is represented in the complex-valued form as $g_n = \sum_{j=1}^{r} C_j \mu_j^n$. The subspace spanned by the r leading eigenvectors of the trajectory matrix determines the values μ_j. The methods calculating μ_j through this subspace are called the *subspace methods.* Real-valued μ_j correspond to exponential components; complex conjugate μ_j generate sinusoids. For a noisy signal, the subspace of the signal can be estimated using SSA as the space spanned by the eigenvectors $\{U_i\}_{i \in \mathcal{J}}$. Subspace methods for calculation of exponentials μ_j (mostly complex-valued) have been known for a long time. Among their advantages are: (i) high resolution (estimation of close frequencies of summand sinusoids), (ii) robustness to outliers, (iii) little prior information (no signal model specified, only its rank r).

In this paper we use the method ESPRIT [7], which was chosen to illustrate application of subspace methods. Note that there are modifications of ESPRIT for more precise results, e.g. weighted/total least square ESPRIT. Having μ_j estimated, the coefficients C_i can be computed by means of one of the least squares methods. In terms of SSA, ESPRIT exploits the rotational invariance property of the subspace of the signal found by SSA.

3 RESULTS AND DISCUSSION

3.1 Methodology

An ideal case (trend is separable from noise with $L = N/2$). Let us consider the data *ms19* (data from a single embryo) as an example, see Figure 1a. The only parameter of SSA is the window length L. SSA theory [2] implies that for better separability between components of a given series one should choose L close to the half-length $N/2$. Having performed SVD with $L = 246 \approx N/2$, we visually examine the elementary RCs produced, see Figure 1b. As only the two leading elementary RCs vary slowly, we reconstruct the trend with the two leading SVD components. The resulting trend is depicted in Figure 1a. One can easily see that the result reasonably reconstructs the trend, but at the same time is robust to the apparent outliers.

A bad case (trend is not separable from noise with $L = N/2$). However, for some cases the trend can not be separated from noise or cyclic components. For the data *ac2* (see Figure 1c) with $L = 285 = N/2$, the third and fourth elementary RCs contain both trend and noise, see Figure 1d. The contribution of these SVD components is small and they can be omitted for a tentative trend reconstruction. But for background estimation, loss of even 1–2 intensity units is sizeable. Let us consider two additional tricks which help to reconstruct trend more precisely.

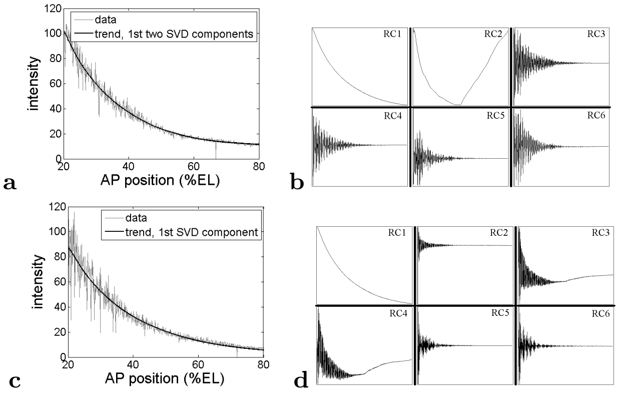

Figure 1: **a**, **b** *ms19*, **c**, **d** *ac2*; **a**, **c** original data with SSA trend and **b**, **d** elementary RCs produced with $L = N/2$

Use of small window length. The trend and noise components can be mixed between each other due to the complicated trend shape or strong noise, and the latter is observed in our study. The first strategy to cope with the mixing is to choose small window length L. With small L SSA works like a smoothing adaptive linear filter. With $L = 35$ for *ac2*, we get only the one leading SVD component corresponding to the trend. Due to small window length, this SVD component includes all terms of the trend. This method is suitable for different kinds of signals and overcomes the mixing of trend and noise. However, the resultant decomposition does not allow us to split the trend into the pattern and background.

Improvement of separability by the addition of a constant. Recall that we suppose the trend to be the sum of an exponential pattern and an almost constant background, where the latter is approximated by an exponential function with small rate. In this particular case another trend extraction strategy can be exploited. Such a trend generates two SVD components [2]. In order to enlarge the contribution of the second SVD component and therefore to reduce mixture with the rest, we add a constant to the given data, thus artificially increasing the background. After that we use the theoretically best window length $L = N/2$ with no effect of mixing. This strategy greatly helps; having added $A = 50$ to the given data, the results for *ac2* (with $L = 285$) become visually the same as in the good case *ms19* depicted in Figure 1b. The value A equal to 50 was chosen to provide separability for all series from the considered test set and therefore to allow us to extract trends being splitted into patterns and backgrounds. As for *ms19*, smaller values of A that enough for separability can be used.

3.2 Batch identification of trend components

Above we investigated the properties of the SSA representation of *Bicoid* gene expression data, which contain the exponential pattern and a close to constant background. Let us apply AutoSSA for batch-processing the whole dataset taking into account the experience of the previous section. First, we set the parameter ω_0. As mentioned above, ω_0 defines the low-frequency interval $[0, \omega_0]$. Examining the eigenvector periodogram, we can guess ω_0 as a value bounding the interval of large periodogram values next to the zero frequency. Taking the data *ms19* as an example, we consider periodograms of its eigenvectors ($L = N/2$). Exactly the two leading SVD components of *ms19* are to be identified. The first six frequencies $0, 1/L, \ldots, 5/L$ of the periodograms of both eigenvectors contain the prevailing contribution. Thus, we select $\omega_0 = 5/L$. For five randomly selected test series (*ad36*,*as27*, *cb23*,*hx8*,*iz13*), the procedure of choice of $\mathcal{C}_0$ presented in [1] ($\mathcal{C}_{\min} = 0.5$, $\mathcal{C}_{\max} = 1$, $\Delta\mathcal{C} = 0.01$, $\Delta\mathcal{R} = 0.01$) produces the following $\mathcal{C}_0$: 0.85,0.87,0.88,0.85,0.88 of which the smallest $\mathcal{C}_0 = 0.85$ is selected. AutoSSA with $\omega_0 = 5/L$, $\mathcal{C}_0 = 0.85$ identifies the same SVD components as those visually identified above: two leading SVD components for *ms19* ($L = N/2$) and for *ac2* increased by $A = 50$ ($L = N/2$), as well as only the leading component for *ac2* with $L = 35$.

Having added $A = 50$ to all series from the given set, AutoSSA identifies exactly the two leading components. The visual check of the identified eigenvectors proves their slowly-varying shape; this substantiates the use of AutoSSA, especially for those data where addition of a constant has enhanced the separability. Moreover, this uniformity of results over the whole dataset allows us to derive the analytical approximation using these SVD components.

3.3 Analytical trend approximation: exponential pattern plus background

Let us consider two-rank approximation of the trend. A two-rank series is either an exponentially-modulated sinusoid or a sum of two real exponentials [2], and a constant function is a special case of an exponential. Note that we specify no approximation model (from the two given above) but only the rank. For the data *ac2* the resulting formula for the trend is $g_n = 79.29 \cdot 0.994^n + (59.88 \cdot 0.9998^n - 50)$, where n runs through the nucleus numbers in the considered AP interval. After transformation from nucleus number n to AP coordinate x, we obtain the approximation for the exponential pattern $p(x)$ and the background $b(x)$: $p(x) = 302.35e^{-0.062x}$, $b(x) = 11.7 - 0.0865x$, see Figure 2.

3.4 Summary results of the analytical approximation

The considered dataset contains 17 series and we extracted exponential patterns using ESPRIT for all of them increased by $A = 50$ in advance. That the patterns tend to zero close to the posterior end is confirmed by the biological interpretation of the *Bicoid* gradient. To generate reference results, we fitted the curve $g_n = Ce^{Ln} + B$ to each original series using the least squares (LS) method, like [3]. The patterns produced with ESPRIT and with LS-fitting are very similar that confirms potential of ESPRIT in extracting exponential patterns without fixing the model of constant background. Both the Matlab function *fminsearch* (v.7.4) and the *Nonlinear Estimation* module of STATISTICA (v.6.0) with randomly chosen initial values produce substantially incorrect results. Thus, the initial values used are crucial. It turns out that using ESPRIT estimates of C, L and $B = 1$ as the initial values, the procedure of the LS-estimation becomes stable and precise. This shows the usefulness of ESPRIT in combination with LS-methods.

The SSA-based procedure presented here is more flexible than the usual *Bicoid* profile modeling with constant background. On the other hand, as simultaneous estimation of parameters of two exponents is less stable in general, the variation of the resulting pattern parameters can be potentially larger than that in the model with a constant background. However, even for modeling with fixed shape of the background, SSA can be useful for setting initial values of the corresponding fitting procedure.

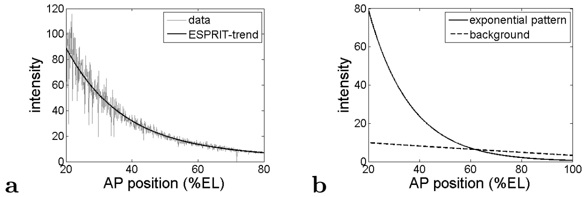

Figure 2: *ac2*: **a** initial series and its ESPRIT-approximation of trend; **b** trend components: exponential pattern and background

4 CONCLUSIONS

First, we developed the SSA-based technique for signal extraction from one-dimensional spatial gene expression profiles containing exponential-in-distance patterns and constant-like backgrounds. The obtained results are consistent with the state-of-the-art results for the given data, though the data contain strong noise, outliers and we do not assume models of profile, pattern, and background to be known a priori. Moreover, the feasibility of batch processing of the given data using AutoSSA is demonstrated.

Second, using the SSA-related method ESPRIT, we obtained an analytical representation of the signal as a sum of two exponential functions. The first is the well-known exponential pattern of the *Bicoid* protein, and the second is the background approximated by an exponential or linear function in our case. The employed method produces stable parameter estimates, even for noisy series. Moreover, these estimates can be used as initial values for nonlinear least square fitting procedures in which the model is assumed a priori.

Acknowledgements. We thank David Holloway for comments and Alexey Timofeyev for ESPRIT software. This work was supported by the NSF/NIGMS BioMath 1-R01-GM072022 and CRDF RUB1-1643-ST-06 grants.

REFERENCES

[1] T. Alexandrov. A method of trend extraction using Singular Spectrum Analysis. Preprint, 2008. http://arxiv.org/abs/0804.3367.

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

[3] D. M. Holloway, L. G. Harrison, D. Kosman, C. E. Vanario-Alonso, and A. Spirov. Analysis of pattern precision shows that *Drosophila* segmentation develops substantial independence from gradients of maternal gene products. *Dev. Dynam.*, 235:2949–2960, 2006.

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

[5] E. Myasnikova, M. Samsonova, D. Kosman, and J. Reinitz. Removal of background signal from *in situ* data on the expression of segmentation genes in *Drosophila*. *Dev. Genes Evol.*, 215:320–326, 2005.

[6] E. Poustelnikova, A. Pisarev, M. Blagov, M. Samsonova, and J. Reinitz. A database for management of gene expression data in situ. *Bioinformatics*, 20:2212–2221, 2004.

[7] R. Roy and T. Kailath. ESPRIT: estimation of signal parameters via rotational invariance techniques. *IEEE Trans. Acoust.*, 37:984–995, 1989.