Biomolecular Detection and Quantification 4 (2015) 17–21

Contents lists available at ScienceDirect

Biomolecular Detection and Quantification

journal homepage: www.elsevier.com/locate/bdq

Review Article

A glance at the applications of Singular Spectrum Analysis in gene expression data

Hossein Hassani[a,b,*], Zara Ghodsi[b]

[a] Institute for International Energy Studies (I.I.E.S), 1967743711 Tehran, Iran
[b] The Statistical Research Centre, Bournemouth University, UK

ARTICLE INFO

Article history:
Received 7 January 2015
Received in revised form 12 March 2015
Accepted 8 April 2015
Available online 29 May 2015

Keywords:
Singular Spectrum Analysis
Filtering
Signal extraction

ABSTRACT

In recent years Singular Spectrum Analysis (SSA) has been used to solve many biomedical issues and is currently accepted as a potential technique in quantitative genetics studies. Presented in this article is a review of recent published genetics studies which have taken advantage of SSA. Since Singular Value Decomposition (SVD) is an important stage of this technique which can also be used as an independent analytical method in gene expression data, we also briefly touch upon some areas of the application of SVD. The review finds that at present, the most prominent area of applying SSA in genetics is filtering and signal extraction, which proves that SSA can be considered as a valuable aid and promising method for genetics analysis.

Contents

1. Introduction

Nowadays there exists a large amount of datasets in the field of genetics and expression measurement and there are many different methods and techniques for analysing these datasets [1–3]. However, it has been widely accepted that the major difficulties in working with such data no longer is the validity of expression measurements, but the reliability of inferences from the data as the achieved data is difficult to understand without the use of proper analytical tools [4] and if research result is obtained from an inappropriate model it can never be translated into a valid scientific context [5].

Historically, such data have been analysed using parametric methods [6,7]. However, constraining pre-assumptions needed for parametric approaches led towards the growing popularity of nonparametric methods. Recently it has been concluded that nonparametric techniques can be used as an alternative approach for analysing genetics data because of their inherent nature [8], and accordingly the applications in biomedical and genetics fields have expanded.

Among many non-parametric methods, Singular Spectrum Analysis (SSA) is a relatively new approach which has proven to be very successful. SSA has already transformed itself into a standard tool in the analysis of biomedical, mathematical, geometrical and several other time series [9–11] and recently it has also been applied in genetics which has illustrated its strong potential for such studies [12,13].

The emergence of SSA is usually associated with the work by Broomhead in 1986 [14]. However, the ideas of SSA were independently developed in Russia and in several groups in the UK and USA. Since after, several papers on the methodology and applications of SSA have been published (see, for example, [15,16]). An introduction to this technique can be identified in the paper by Elsner and

* Corresponding author at: The Statistical Research Centre, Bournemouth University, UK. Tel.: +44 7703367456; fax: +44 7703367456.
E-mail address: hhassani@bournemouth.ac.uk (H. Hassani).

http://dx.doi.org/10.1016/j.bdq.2015.04.001

Tsonis [17] and a comprehensive description with several examples of theoretical and practical aspects of SSA can also be found [8,18].

The main advantages of the SSA technique in the field of genetics can be attributed to its signal extraction and filtering capabilities [19], batch processing of a set of similar series [20] and derivation of an analytical formula of the signal [21]. The application of SSA for noise reduction in microarrays, and signal extraction in gene expression data has received more interest. The reason underlying the significant interest in the SSA technique's filtering capabilities are due to the fact that genetics data is often characterised by the existence of considerable noise, filtering this noisy data is considered as one of the most arduous tasks when analysing genetics data [22,23].

For example, microarray is a very useful method for acquiring quantitative data in genetics and researchers today are conducting most of their studies using this method. The main advantage of microarray is the capability of studying thousands of genes simultaneously. However, microarray data usually contains a high level of noise, which can reduce the performance of the results [24].

This article categorises and summarises almost all recently published articles associated with the application of SSA in genetics.

Presented below, is a short description of SSA technique in doing so we mainly follow [9] where a more detailed description is made available. Moreover, the R package for this technique including decomposition, forecasting and gap filling for univariate and multivariate time series can be downloaded via [25]

Consider a set of genetics observations in a series of $Y_N = (y_1, \ldots, y_N)$ with length of N. After choosing a window length L where $(2 \le L \le N-1)$, we construct the L-lagged vectors $X_j = (y_j, \ldots, y_{L+j-1})^T$, $j = 1, \ldots, K$ where $K = N - L + 1$. Define the matrix $\mathbf{X} = (x_{ij})_{i,j=1}^{L,K} = (X1, \ldots, X_K)$. Now $\mathbf{X}$ is our multivariate data with L characteristics and K observations. The columns X_j of $\mathbf{X}$, are the vectors, positioned in an L-dimensional space $\mathbb{R}^L$. Define the matrix $\mathbf{X}\mathbf{X}^T$: SVD of $\mathbf{X}\mathbf{X}^T$ gives us the collections of L eigenvalues $(\lambda_1 \ge \cdots \ge \lambda_L \ge 0)$ and the corresponding eigenvectors $U_1 \ldots U_L$ where U_i is the normalised eigenvector corresponding to the eigenvalue $\lambda_i (i = 1, \ldots, L)$. A group of r (with $1 \le r < L$) eigenvectors determines an r-dimensional hyperplane in the L-dimensional space RL of vectors X_j. By choosing the first r eigenvectors $U_1, \ldots, U_r$, then the squared L_2-distance between this projection and $\mathbf{X}$ is equal to $\Sigma_{j=r+1}^{L} \lambda_j$. Based on the SSA process, the L-dimensional data are projected onto this r-dimensional subspace and the final diagonal averaging gives us an appropriate approximation of the first one dimensional series.

The remainder of this paper is organised as follows. In the following section we present a review of papers involving the application of SSA and SVD[1] on signal extraction and noise reduction. The SSA based on minimum variance and a hybrid modelling approach combining SSA and autoregressive (AR) model are also discussed in depth in that section. Section 3 provides theoretical developments leading to what is termed as "two-dimensional SSA" and the paper ends with some conclusions in Section 4.

2. Signal extraction and filtering

In this section we identify existing applications of SSA for signal extraction and noise filtering in genetics.

The first such application is reported in 2006 where SSA was used for signal extraction of *Drosophila melanogaster*'s gene expression profile [19]. The idea of using SSA for signal extraction was then followed in an approximately similar study in [26] which led to an improved result. By 2008 a more technical study conducted on the methodology of signal extraction from the noisy *Bicoid (Bcd)* protein profile in *Drosophila melanogaster* was presented in [21]. The problem under investigation in that study was complicated by two facts: (i) the data contained outliers and (ii) that the data was exceedingly noisy and the noise consisted of an unknown structure. The author examined two approaches for reconstructing signal more precisely: the use of small window length and improvements to separability by adding a constant to the series.

In addition, the activation of the *hunchback (hb)* gene in response to different concentrations of *Bcd* gradient was studied in [27] and SSA was applied for filtering two kinds of noise; experimental noise and the noise caused by variability in nuclear order [27].

2.1. SSA based on minimum variance

Recently, a modified version of SSA was examined for filtering and extracting the *bcd* gene expression signal [28] and the results illustrated that SSA based on minimum variance can significantly outperform the previous methods used for filtering noisy *Bcd* [28].

SSA based on minimum variance mainly relies on the concept that by dividing the given noisy time series into the mutually orthogonal noise and signal+noise components, an enhanced estimation of the signal can be achieved. Thus, after performing SVD, by adapting the weights for different obtained singular components, an estimation of the Hankel matrix $\mathbf{X}$, will be achieved which in turn is corresponds to a filtered series.

A short description of the SSA based on minimum variance is given below. For more details, see [28,29].

Let us begin with the Singular Value Decomposition (SVD) of the trajectory matrix $\mathbf{X}$:

$$\mathbf{X} = \mathbf{U}(\mathrm{W}\Sigma)\mathbf{V}^T, \tag{1}$$

where $\mathbf{W}$ is the diagonal matrix of the weights to be determined. The SVD of the matrix $\mathbf{X}$ can be written as:

$$\mathbf{X} = [\mathbf{U}_1 \;\; \mathbf{U}_2] \begin{bmatrix} \mathbf{\Sigma}_1 & 0 \\ 0 & \mathbf{\Sigma}_2 \end{bmatrix} \begin{bmatrix} \mathbf{V}_1^T \\ \mathbf{V}_2^T \end{bmatrix}, \tag{2}$$

where $\mathbf{U}_1 \in \mathbb{R}^{L \times r}$, $\mathbf{\Sigma}_1 \in \mathbb{R}^{r \times r}$ and $\mathbf{V}_1 \in \mathbb{R}^{K \times r}$.

Now, the issue is in selecting the weight matrix $\mathbf{W}$. If we represent the SVD of the Hankel matrix related to the signal as $\mathbf{S}$, by considering different criteria in choosing this matrix, different estimation of $\mathbf{S}$ can be achieved. The LS Estimate of $\mathbf{S}$ is the current widely used approach in selecting the weight matrix $\mathbf{W}$. This approach is based on the idea of removing the noise subspace but keeping the noisy signal uncorrelated in the signal+noise subspace. However, the accuracy of this estimator is mainly dependent on the estimation of the signal rank r since selecting singular values in LS follows a binary approach. Although in using this estimator, considering any assumptions is not needed.

In MV Estimate of $\mathbf{S}$ proposed by Hassani in [29], removing the noise components in the signal + noise subspace has been improved considerably. However, to obtain the MV estimate considering some assumptions is essential (for more details see [29]).

Let us now consider the weight matrix $\mathbf{W}$ of the LS and MV estimates:

$$\begin{aligned} \hat{\mathbf{S}}_{LS} &= \mathbf{U}_1(\mathbf{W}_{LS}\mathbf{\Sigma}_1)\mathbf{V}_1^T \\ \hat{\mathbf{S}}_{MV} &= \mathbf{U}_1(\mathbf{W}_{MV}\mathbf{\Sigma}_1)\mathbf{V}_1^T, \end{aligned} \tag{3}$$

[1] The SVD algorithms used in this paper are either based on Hankel or Teoplitz matrix.