Digital Signal Processing 51 (2016) 101–109

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

From nature to maths: Improving forecasting performance in subspace-based methods using genetics Colonial Theory

Hossein Hassani [a,*], Zara Ghodsi [b], Emmanuel Sirimal Silva [c], Saeed Heravi [d]

[a] *Institute for International Energy Studies, Tehran 1967743 711, Iran*
[b] *Translational Genetics Group, Bournemouth University, Fern Barrow, Poole, BH125BB, UK*
[c] *Fashion Business School, London College of Fashion, University of the Arts London, WC1V7EY, UK*
[d] *Cardiff Business School, Cardiff University, Cardiff, CF10 3EU, UK*

ARTICLE INFO

Article history:
Available online 22 January 2016

Keywords:
Colonial theory
Forecasting
Nature inspired algorithm
Subspace methods

ABSTRACT

Many scientific fields consider accurate and reliable forecasting methods as important decision-making tools in the modern age amidst increasing volatility and uncertainty. As such there exists an opportune demand for theoretical developments which can result in more accurate forecasts. Inspired by Colonial Theory, this paper seeks to bring about considerable improvements to the field of time series analysis and forecasting by identifying certain core characteristics of Colonial Theory which are subsequently exploited in introducing a novel approach for the grouping step of subspace based methods. The proposed algorithm shows promising results in terms of improved performances in noise filtering and forecasting of time series. The reliability and validity of the proposed algorithm is evaluated and compared with popular forecasting models with the results being thoroughly evaluated for statistical significance and thereby adding more confidence and value to the findings of this research.

1. Introduction

As a great source of inspiration, nature holds the key to many questions we face on a daily basis. Therefore, it is not entirely surprising that much of the novel problem solving techniques were initially inspired by nature (see for example [1–5]), even though credit is seldom given. In developing such intelligent solutions, nature provides us with effective background knowledge which follows from the profound observation and questioning of a natural phenomenon.

Quantum computing [1], genetic algorithms [2], neural networks [3], swarm algorithms [4] and ant colony optimization algorithms [5] are among the most established nature inspired models which seek to imitate specific phenomenon from nature in order to provide simple solutions to complex problems. Although attempting to model natural phenomena has a long history, the recent application of nature inspired algorithms like firefly algorithm [6], neuro fuzzy technique [7] and genetic programming [8] in the area of soft computing and also recent improvements in forecasting approaches [9,10], has lead to more accurate analysis and predictions, and thereby causing a noticeable growth of interest in this field [11–18]. However, attempts at improving signal extraction and forecasting using bio-inspired algorithms is a relatively new area of research. Moreover, it is noteworthy that in most of the nature inspired algorithms, the natural phenomenon of interest is the strategy taken by biological organisms after facing an environmental change which enables the organism to exploit an ingenious solution to meet the new specific conditions.

Even though such biological solutions have been many times by organisms over the evolution process, the most prominent one can be referred to as the multicellularity phenomenon which describes how multicellular organisms arose from a single cell and generated multi-celled organisms. Despite there being various theories that may be able to explain this mechanism, Colonial Theory (CT) has received most credit by developmental scientists [19]. Inspired by CT and by identifying certain characteristics of this theory, in this paper we seek to draw a line between nature and mathematics. The mathematical procedures which we specifically seek to link with nature consist of Singular Value Decomposition (SVD) based methods and signal subspace (SS) methods which form the basis of a general class of subspace-based noise reduction algorithms. The superior performance of this class of algorithms in noise reduction and forecasting has been proved by several studies [20–22].

In this context, the Singular Spectrum Analysis (SSA) technique, which is a SVD and SS based method has been considered as a powerful nonparametric tool [23]. In brief, the SSA technique begins by decomposing the original series into the sum of a small

---

[*] Corresponding author.
E-mail address: hassani.stat@gmail.com (H. Hassani).

http://dx.doi.org/10.1016/j.dsp.2016.01.002

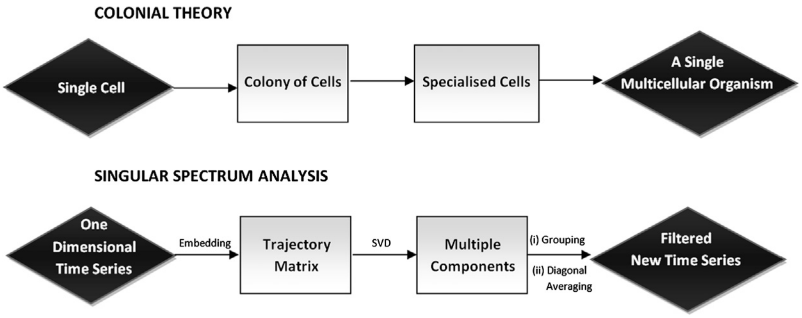

Fig. 1. The linkage between CT and SSA.

number of independent components. Thereafter, the selected components are used to reconstruct the less noisy series which can be used for forecasting future data points. However, due to the nature of least squares (LS) estimation method used in the current SSA procedure, the signal and noise separation is not optimum and the reconstructed series continues to hold some part of initial noise whilst the residual is not completely signal free. This paper seeks to consider an alternative approach which is based on CT in order to provide a more efficient outcome for the signal and noise separation issue in SSA.

The exploitation of CT towards improving the SSA process is made possible via our identification of a general similarity between CT and SSA. However, this similarity was not visible in the Grouping step of the basic SSA process, and therein lies our focus as we intend on defining a new approach to grouping in SSA by imitating one of the steps followed in CT. It is expected that this novel CT based approach to grouping will enable a more efficient separation of signal from noise which in turn will enhance the signal extraction and forecasting results.

The remainder of this paper is organized such that Section 2 presents an outline of CT along with the steps underlying the SSA process in order to clearly illustrate similarities and discuss how the different steps of CT are fully consistent with the procedure underlying SSA. Section 3 describes the newly introduced approach for grouping in SSA. Section 4 provides the theoretical presentation of the algorithm which is followed by several applications in Section 5 and the paper concludes with a concise summary in Section 6.

2. Similarities between SSA and CT

This section focuses on providing a clear view on the similarities between SSA and CT as portrayed in Fig. 1. In what follows, the information contained in Fig. 1 is expanded upon as we present a detailed explanation of the linkages between SSA and CT.

The SSA method is made up of two complementary stages: Decomposition and Reconstruction; each stage consists of two compatible steps. At the first stage a group of small number of independent and interpretable components is achieved by decomposing the main series [23], which is followed by the reconstruction of a less noisy series at the second stage [24]. Thereafter, this noise free series is used for forecasting future data points.

2.1. Stage 1: decomposition

We begin with a one dimensional time series, $Y_N = (y_1, \ldots, y_N)$ where N is the length of the series. The SSA technique consists of two choices, the window length L and the number of eigenvalues r [25]. In SSA the number of components are related to the selection of the proper window length L which should be defined such that it minimises the signal distortion and maximises the residual noise level. However, we cannot impose a general rule in selecting L for different time series with different structure. For example, in instances where there is a periodic component with an integer period like a seasonal component, to obtain a higher separability the tradition is to select L proportional to that period [23].

Likewise, the starting point of CT is a single cell which evolves over time and generates a multi-celled organism. Similar to SSA, there is also a limit on the number of cell types and different kinds of organisms necessitate different numbers of cell types. It is assumed that this number is determined by the balance between selective pressure and functional requirements, whilst variety is favoured by selection, functional needs limit the number of cell types [26].

2.1.1. 1st step: embedding

Here we take the one dimensional time series Y_N, and map it in order to create a multi-dimensional variable of $X_1, \ldots, X_K$ where $X_i = (y_i, \ldots, y_{i+L-1})' \in \mathbf{R}^L$. It is clear that this can be viewed as the creation of the colony from the initial single cell as we take Y_N and create multiple dimensions from the same series. This step provides us with a trajectory matrix, $\mathbf{X}$ which is a Hankel matrix that captures all information contained in Y_N.

$$\mathbf{X} = (x_{ij})_{i,j=1}^{L,K} = \begin{pmatrix} y_1 & y_2 & y_3 & \cdots & y_K \\ y_2 & y_3 & y_4 & \cdots & y_{K+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_L & y_{L+1} & y_{L+2} & \cdots & y_N \end{pmatrix}. \tag{1}$$

It should be noted that unlike the Symbiotic theory [27], which assumes that the symbiosis of various species caused a multicellular organism, in CT it is the symbiosis of many cells of the same species that forms a multicellular organism. This point is interesting as it can be referred to as the first and main difference between SSA and principal component analysis (PCA). In the latter, the obtained matrix is achieved by considering different time series (multiple cells) whilst in SSA we consider one time series (single cell).

Moreover, transferring a one dimensional time series into a trajectory matrix will enable us to significantly reduce the computation time required for running the algorithm, as it eliminates the need for running the algorithm over a wide range of values for the hidden state dimension. Furthermore, by analyzing the eigenvalues with the aim of filtering the signal and noise, the signal to noise ratio (SNR) will be optimized in the newly reconstructed time series. Likewise, increasing in size is initially favoured by individual cells since multicellular organisms do not have the size limit which is mainly imposed by diffusion. As the surface-to-volume ratio decreases in a given single cell, with increased size they will experience difficulty in obtaining the required nutrients and transporting the cellular waste products out the cell [26,28].

2.1.2. 2nd step: singular value decomposition

SVD is a procedure which is performed on $\mathbf{X}$ and provides us with several eigenvalues or components. The components obtained