Signal Processing 93 (2013) 3001–3013

Contents lists available at SciVerse ScienceDirect

Signal Processing

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

Improving time series modeling by decomposing and analyzing stochastic and deterministic influences

Ricardo Araújo Rios *, Rodrigo Fernandes de Mello

Institute of Mathematics and Computer Science, University of São Paulo, Avenida Trabalhador Sãocarlense, 400, P.O. Box 668, São Carlos, SP, Brazil

ARTICLE INFO

Article history:
Received 19 April 2012
Received in revised form
19 March 2013
Accepted 15 April 2013
Available online 6 May 2013

Keywords:
Time series decomposition
Stochastic component
Deterministic component
Empirical mode decomposition
Recurrence plot

ABSTRACT

This paper proposes a new approach to improve time series modeling by considering stochastic and deterministic influences. Assuming such influences are present in observations, a first decomposition step is required to split them into two components: one stochastic and another deterministic. As second step, models are adjusted on each component and combined to form a hybrid model improving time series analysis. The proposed approach considers the Empirical Mode Decomposition method and a Recurrence Plot-based measurement to decompose and assess stochastic and deterministic influences. Experiments confirmed improvements in time series modeling.

1. Introduction

Time series modeling permits understanding relations among observations over time [1,2], what is useful in several application domains such as: weather forecast, decision making, drug interaction with organisms and climate effects on agriculture [3]. There are two typical approaches to model time series, in which the first assumes deterministic and, the second, stochastic relations among observations. Dynamical systems are considered in the first approach [4] and statistical models were designed to support the latter [5]. However, real-world systems produce time series presenting a mixture of both types of relations [6–10]. In such situation, the model accuracy tends to be affected when one of the two approaches is considered, i.e., dynamical system methods may produce malformed attractors, whereas statistical models may underestimate deterministic influences [11].

This drawback motivated us to adjust individual models on the stochastic and deterministic influences, also referred to as components, and combine them to create a hybrid model in attempt to represent time series at higher accuracy. This requires a decomposition step, which is here performed using the Empirical Mode Decomposition (EMD) method [12].

EMD decomposes a time series into several components under different influences. By adding the components, the original time series is reconstructed. This characteristic of the EMD method makes feasible to analyze time series, whose observations are influenced by additive noise, i.e., when observations are composed by the sum of stochastic and deterministic influences. Aiming at separating the stochastic and the deterministic components, our approach uses Recurrence Quantification Analysis (RQA)[1] [14,15] to assess every component extracted by EMD. Then, dynamical system methods are used to induce a model for the deterministic component, while statistical models are adjusted on the stochastic. Both

* Corresponding author. Tel.: +55 1633736624.
E-mail addresses: rios@icmc.usp.br, rios.fsa@gmail.com (R.A. Rios), mello@icmc.usp.br (R.F. de Mello).

[1] The Recurrence Quantification Analysis is a set of measurements based upon the Recurrence Plot [13].

0165-1684/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.sigpro.2013.04.017

models compose what is referred to as hybrid model, which considers all influences present in time series.

The proposed approach was assessed using synthetic scenarios, comparing two situations: (i) by applying either dynamical system or statistical methods on the original time series and (ii) by decomposing time series and modeling their deterministic and stochastic components. Experimental results confirmed the decomposition improves time series modeling when there is a significant mixture of both influences.

The remaining of this paper is organized as follows: Section 2 presents a discussion on time series decomposition methods; Section 3 details our approach; the evaluation strategy is presented in Section 4; in Sections 5 and 6, we present experimental results and discussions, respectively; finally, Section 7 presents concluding remarks.

2. Time series decomposition

At beginning of this study, we evaluated different spectral-based decomposition methods, such as Fourier (FT) [16,12] and Wavelet (WT) [17,12] transforms. FT and WT are restricted to linear time series. In addition, FT imposes an extra bias for stationary series [12].

FT and WT drawbacks motivated Huang et al. [12] to propose the Empirical Mode Decomposition (EMD) method, which supports the decomposition of time series regardless their linearity, stationarity, and stochasticity. EMD decomposes time series into a set of components, also referred to as Intrinsic Mode Functions (IMFs), which correspond to the most important series features (such as Fourier coefficients, i.e., they have a dominant frequency and amplitude) [12].

The key point to perform this decomposition is the sifting process, which initially analyzes a time series $x(t)$, identifying local maxima and minima values. Afterwards, the cubic spline method is applied on the maxima and minima to compose the upper $u(t)$ and lower $l(t)$ envelopes, respectively [12]. Next, the approximation values obtained using both cubic splines (upper and lower) are used to compute the mean envelope $m(t)$.

Later, $m(t)$ is removed from the original time series $x(t)$, producing the first component candidate $h_{1,1}(t)=x(t)-m(t)$, in which the first index corresponds to the IMF identifier (as this is the first IMF to be extracted, this index is one) and the second refers to the candidate identifier (as this is the first candidate, the index is also one). This candidate is used in place of the original data and all sifting process is repeated until the candidate satisfies the IMF definition, which must agree with one of the following requirements: (i) the number of extrema and the number of zero-crossings must be either equal or differ at most by one; or (ii) at every point, the mean $m(t)$ is zero.

After obtaining the candidate satisfying the IMF definition, the first IMF is defined according to $h_1(t)=h_{1,k}(t)$, assuming k candidates were produced until reaching the IMF definition. This first IMF is then removed from data, i.e., $x(t)-h_1(t)$, and the resultant time series is again analyzed by the whole process, producing further IMFs until reaching a stop criterion. This criterion usually occurs when the last IMF becomes a monotonic function, avoiding the extraction of further components. Hence, this last component is called final residue, $r(t)$ [12]. In summary, according to EMD, a time series $x(t)$ is composed of a set of IMFs plus a residue as shown in Eq. (1).

$$x(t) = \sum_{j=1} h_j(t) + r(t) \tag{1}$$

After executing the EMD method on a time series, we assess the determinism level of every resultant IMF,[2] $h_j(t)$, in order to separate the stochastic from the deterministic influence present in every observation. Aiming at performing this task, we considered the Recurrence Quantification Analysis (RQA), which is a set of measurements based on the Recurrence Plot (RP) [18]. RP requires the reconstruction of the time series into a multidimensional space, also referred to as phase space, which maps the relationships among observations. After this reconstruction, relationships are organized in a two-dimensional binary matrix, called Recurrence Matrix,[3] as defined in Eq. (2), in which ε is a distance threshold, $\|\cdot\|$ is a norm used to calculate the distance between observations, and $\Theta(\cdot)$ is a heaviside function as defined in Eq. (3) [18].

$$R_{i,j} = \Theta(\varepsilon - \| \vec{x_i} - \vec{x_j} \|) \tag{2}$$

$$\Theta(\alpha) = \begin{cases} 0, & \alpha < 0 \\ 1, & \alpha \geq 0 \end{cases} \tag{3}$$

The structures generated by the Recurrence Matrix provide information about the time series under study [13,18] as, for instance: (i) isolated points mean system states are rarely repeated, i.e., the time series is highly stochastic; and (ii) diagonal lines occur when there is persistent behavior, i.e., the determinism rate of a time series is directly related to the number of diagonal lines. However, such information relies on the visual inspection of the Recurrence Plot, what is not a simple task even for humans, adding subjectivity to the estimation of stochastic and deterministic levels. In order to simplify this process and automatize it, a set of measurements called Recurrence Quantification Analysis (RQA) was developed [14,15]. Among all proposed measurements, we are particularly interested in one that quantifies the determinism rate of time series as defined in Eq. (4), in which $P(\varepsilon, l)$ corresponds to the frequency of diagonal lines of length l present in the RP structures. This frequency is calculated using Eq. (5).

$$DET = \frac{\sum_{l=l_{min}}^{N} l P(\varepsilon, l)}{\sum_{i=1}^{N} \sum_{j=1}^{N} R_{i,j}, \forall i \neq j} \tag{4}$$

$$P(\varepsilon, l) = \sum_{i,j=1}^{N} (1 - R_{i-1,j-1(\varepsilon)})(1 - R_{i+l,j+l(\varepsilon)}) \prod_{k=0}^{l-1} R_{i+k,j+k(\varepsilon)} \tag{5}$$

Then, we considered the EMD method to extract components from time series and the determinism rate measurement, provided by RQA, to design a new approach

[2] IMF is also referred to as component in this paper.

[3] The Recurrence Plot corresponds to the Recurrence Matrix plotted.