Physica A 436 (2015) 387–398

Contents lists available at ScienceDirect

Physica A

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

Estimation of connectivity measures in gappy time series

G. Papadopoulos [a], D. Kugiumtzis [b,*]

[a] *Economics Department, Democritus University of Thrace, Komotini, Greece*

[b] *Department of Electrical and Computer Engineering, Faculty of Engineering, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece*

HIGHLIGHTS

- The problem of gaps in time series is solved without any assumption of the underlying dynamics.
- The accuracy of connectivity estimates is the same as for non-gappy time series of reduced length.
- The method compares favorably to a number of standard gap-filling techniques and the gap closure.
- The method can be applied to any time domain method of analysis of multivariate time series.
- The accuracy of the method is confirmed in a financial application.

ARTICLE INFO

Article history:
Received 23 January 2015
Received in revised form 29 April 2015
Available online 15 May 2015

Keywords:
Multivariate time series analysis
Connectivity measures
Gaps in time series
Transfer entropy

ABSTRACT

A new method is proposed to compute connectivity measures on multivariate time series with gaps. Rather than removing or filling the gaps, the rows of the joint data matrix containing empty entries are removed and the calculations are done on the remainder matrix. The method, called measure adapted gap removal (MAGR), can be applied to any connectivity measure that uses a joint data matrix, such as cross correlation, cross mutual information and transfer entropy. MAGR is favorably compared using these three measures to a number of known gap-filling techniques, as well as the gap closure. The superiority of MAGR is illustrated on time series from synthetic systems and financial time series.

1. Introduction

In the analysis of multivariate time series, the primary interest is in investigating interactions among the observed variables. For this a number of measures have been proposed under different terms, such as inter-dependence, coupling, Granger causality and connectivity. There are certain distinctions of these measures, as correlation and causality measures, linear and nonlinear measures, and measures on the time and frequency domain [1–4]. Examples of such measures that we use in our study are the correlation measures of cross correlation and cross mutual information, and the causality measure of transfer entropy [5].

All these methods are developed under the assumption that the time series being analyzed are evenly spaced, meaning the measurements are taken at a fixed sampling rate. However, this is not always the case and in many applications the time series have gaps, as in environmental sciences (occurrence of gaps is a common problem with geophysical [6–8], ecological [9], and oceanographic [10] time series), astronomy [11] and socio-economics [12,13]. Sampling at irregular or uneven time intervals regards a different class of problems and it is not studied here, e.g. for spectral estimation see Refs. [14,15] and for Granger causality see Ref. [16].

* Corresponding author.
E-mail address: dkugiu@auth.gr (D. Kugiumtzis).

http://dx.doi.org/10.1016/j.physa.2015.05.032

The common approach of all proposed techniques for gappy time series is first to fill the gaps in some way and then apply the method of choice to the new evenly spaced time series. The techniques range from relatively simple ones, such as the "gap closure" joining the edges of the gaps, cubic spline and k-nearest neighbors interpolation, to more complex ones such as Single Spectrum Analysis (SSA) [6], neural networks [7], and state space reconstruction under the hypothesis of chaos [17], among others. Comparisons of these methods on different real-world applications can be found in Refs. [8,18,19].

For bivariate and multivariate time series, methods such as SSA and neural networks can be extended to incorporate information from all time series to recover the gaps [6], along with other recent approaches making use of the concepts of nonlinear dynamics and surrogate data [9]. Especially for the application of transfer entropy, Kulp and Tracy [20] examined a stochastic gap-filling technique called "random replacement" in gappy data from harmonic oscillators.

In our paper we take a different route and address the problem in a method specific manner. Instead of filling the gappy time series, we modify the measure to be used, accounting for the gaps in the time series, thus leaving the underlying dynamics to the time series intact, free of artificial intervention. Our approach, called measure adapted gap removal (MAGR), is general for any measure of multivariate time series, and we exemplify it here on two correlation measures, cross correlation and cross mutual information, and one Granger causality measure, the transfer entropy. We demonstrate the effectiveness of our approach in comparison to gap-filling methods on a linear stochastic multivariate autoregressive (MVAR) system and a nonlinear system, the coupled Henon map. We randomly remove samples of the generated time series and estimate each measure on the gappy time series using our approach as well as different gap filling methods. Further, we consider also the case of missing blocks of consecutive samples, of fixed or varying size, often met in applications.

The remainder of the paper is structured as follows. Section 2 gives briefly the theoretical framework of the correlation and causality measures used in this study. Section 3 describes our approach for computing the measures on gappy time series. Section 4 presents the simulation results on the linear and nonlinear systems for the estimation of the measures with our approach and the gap-filling methods. Section 5 presents an application of MAGR to real financial data. Finally, the results are discussed in Section 6.

2. Correlation and causality measures

In the following, we present briefly the three measures used to demonstrate our approach when the multivariate time series contain single missing values, called single gaps, or blocks of missing values, called block gaps. We denote the variables with capital letters and the sample values with small letters. The measures considered in this study are bivariate and for multivariate time series they are applied to each pair of time series. The implementation of our approach to multivariate measures, e.g. the partial transfer entropy [21,22], is straightforward.

2.1. Cross correlation

For two simultaneously measured variables X and Y giving the time series $\{x_t, y_t\}_{t=1}^{N}$, the cross correlation measures the linear correlation of X and Y at the same time t, or at a delay τ, defined as

$$r_{XY}(\tau) = \mathrm{Corr}(X_t, Y_{t+\tau}) = \frac{\sum_{t=1}^{N-\tau}(x_t - \bar{x})(y_{t+\tau} - \bar{y})}{\sqrt{\sum_{t=1}^{N-\tau}(x_t - \bar{x})^2 \sum_{t=1}^{N-\tau}(y_t - \bar{y})^2}}, \tag{1}$$

where $\bar{x}$ and $\bar{y}$ are the mean values of the two time series. For $\tau = 0$, $r_{XY}(0)$ is the standard Pearson correlation coefficient of X and Y.

2.2. Cross mutual information

Cross mutual information is an appropriate analogue to cross correlation if also nonlinear correlation is to be estimated. First, mutual information of two variables X and Y is defined in terms of entropies as

$$I(X;Y) = H(X) + H(Y) - H(X,Y), \tag{2}$$

where $H(X)$ and $H(X, Y)$ are the Shannon entropy of X and the joint entropy of (X, Y), respectively [23]. For the estimation of the entropies, we first discretize X and Y, and then compute the standard frequency estimates of the probability mass function of X, Y and (X, Y) denoted p_X, p_Y and $p_{X,Y}$, respectively, giving

$$I(X;Y) = \sum_x \sum_y p_{X,Y}(x,y) \log \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}. \tag{3}$$

We consider here the discretization of X and Y using equiprobable partition of their domains in b intervals. So, there is an equal occupancy at each interval, and the probability $p_X(x)$ of the x-th element of the partition of X is $p_X(x) = 1/b$ and