2015 International Symposium on Nonlinear Theory and its Applications
NOLTA2015, Kowloon, Hong Kong, China, December 1-4, 2015

On constructing networks from multivariate time series

Tomomichi Nakamura† and Toshihiro Tanizawa‡

†Graduate School of Simulation Studies, University of Hyogo
7-1-28 Minatojima-minamimachi, Chuo-ku, Kobe, Hyogo 650-0047, Japan
‡Kochi National College of Technology
200-1 Monobe-Otsu, Nankoku, Kochi 783-8508 Japan
Email: tomo@sim.u-hyogo.ac.jp, tanizawa@ee.kochi-ct.ac.jp

Abstract—We introduce a method of constructing networks from multivariate time series. The method enables us to construct networks even if a given multivariate time series do not have strong similarity. We show a simple example where a common method based on similarity does not work. The method we introduce is demonstrated for numerical data generated by a known system and applied to actual time series with unknown dynamics.

1. Introduction

To tackle phenomena in the real world, one of the first valid approaches is to consider that the phenomena move by systems and to assume the underlying systems [1]. Elements in the system interact with each other. To understand the details of the interaction among the elements, the concept of complex networks has been widely recognized to be useful [2, 3].

There are work to construct a network from multivariate time series. As time series usually show irregular fluctuations, it is difficult to know the precise relationship among them on the first impressions. For constructing a network from multivariate time series the cross correlation function with a constant threshold is used most extensively [4, 5], which we refer to as "the common method." The cross correlation function is one of the useful statistics that can directly investigate some kind of relations between two signals. When the statistic has strong peaks or has large absolute values between -1 and $+1$ at some time lags the result is a good indication that the data have similarities. Then, we expect that there are correlation structures between the two signals (or that similar factors may influence both systems). Then, the pair is considered to be connected with an undirected link. However, the patchwork of many two-body information might not be the same as the many-body information as a whole. Also, although periodicities in multivariate time series contain important information, the cross correlation function cannot treat it directly. Furthermore, we cannot the direction between the nodes on the network constructed by the common method. Hence, it might be preferable if we could capture many-body periodicities and directions among dynamical elements in a system as a whole from a dynamical system-wide perspective, even if the time series do not have large values of cross correlation. In this paper, we introduce such a method for constructing directed networks from multivariate time series based on the linear modeling technique.

2. Common technology

The most extensively used method to construct networks from multivariate time series can be reduced to the following three steps [4, 5].

1. Each time series is considered as a basic node of a network.
2. To investigate the relationship among multivariate time series, the cross correlation between each pair of time series (i.e. two time series) taken from the whole multivariate time series is estimated.
3. The pair of nodes corresponding to the chosen two time series is connected with an undirected edge when the value of the cross correlation is larger than an appropriately chosen threshold.

We referred this method to as "the common method", as mentioned above. Although the common method based on the concept of the cross correlation has been proved to be effective in various cases [4, 5], the range of applicability might be restrictive because "no similarity" is not equivalent to "no correlation" [6].

3. Our proposed algorithm

To construct directed networks from multivariate time series from a dynamical system-wide perspective, even if the time series do not have large values of cross correlation, we use the reduced auto-regressive (RAR) model [7].

Periodic or nearly periodic behavior is an important nature for many time-dependent phenomena in the real world. Without including such (nearly) periodic effects, we cannot reproduce the time-dependent phenomena properly [8]. There are several widely accepted techniques to estimate the period of behavior, such as spectral estimation, autocorrelation, wavelet transforms and so on [7]. However, these methods cannot provide accurate and decisive periodicities.

Small and Judd have proposed a method to identify precise periodicities directly from the model [7]. The technique is based on an information theoretic reduction of AR models, which is referred to as the reduced autoregressive (RAR) model [7]. RAR models include minimal number of terms indispensable for describing time series as assessed by an information criterion. There are also strong information theoretic arguments to support that the RAR model can detect any periodicities built into a given time series [7]. Moreover, the RAR model has proven to be effective in modeling both linear and nonlinear dynamics [9, 10]. Hence, we consider to construct directed networks based on information included in RAR models from multivariate time series.

The building of an RAR model from given time series proceeds as follows. Given a scalar time series $\{x(t)\}_{t=1}^{n}$ of n observations, an RAR model with the largest time delay l_w can be expressed by

$$x(t) = a_0 + \sum_{i=1}^{w} a_i x(t - l_i) + \varepsilon(t), \tag{1}$$

where $1 \leq l_1 < l_2 \cdots < l_w$, a_i $(i = 0, 1, 2 \ldots, w)$ are parameters to be determined, and $\varepsilon(t)$ is assumed to be independent and identically distributed Gaussian random variables, which are interpreted as fitting errors. The parameters a_i are chosen to minimize the sum of the squares of fitting errors. To build an RAR model we prepare candidate basis functions used in the modeling, in the form of a dictionary, and select the most appropriate basis functions that can extract the peculiarities of the time series as much as possible [11]. As RAR models are linear, the basis functions are a constant function and linear terms.

This methodology can be applied equally to multivariate time series straightforwardly [7, 9, 10]. A set of multivariate RAR models is expressed by

$$x_i(t) = a_{i,0} + \sum_{j=1}^{N} \left(\sum_{k=1}^{w_j} a_{i,j,k}\ x_j(t - l_k) \right) + \varepsilon_i(t), \tag{2}$$

where $i = 1, 2, \ldots, N$, N is the number of components and $l_{w_i} (\geq 0)$ is the largest time delay of the i-th component.

To know the best model we apply the the concept of description length in the information theory. An approximation to description length takes the form

$$DL(k) = \left(\frac{n}{2} - 1\right) \ln \frac{\mathbf{e}^T \mathbf{e}}{n} + (k + 1)\left(\frac{1}{2} + \ln \gamma\right) - \sum_{i=1}^{k} \ln \delta_i \tag{3}$$

where n is the length of the time series to be fitted, $\mathbf{e}$ stands for the vector composed from fitting errors, k is the number of parameters (or model size), γ is related to the scale of the data, and the variables δ can be interpreted as the relative precision to which the parameters are specified. The factor γ is a constant and typically fixed to be $\gamma = 32$ [9]. When a model has the smallest value of the description length among many models, we treat the model as the "best model (optimal model)." See more details in [9, 10].

After building the multivariate RAR model corresponding to the systems under consideration, we use the information contained these models to construct a directed network representing the system. The model for the i-th variable $x_i(t)$ takes the form as

$$\begin{aligned} x_i(t) = a_{i,0} &+ a_{i,i,1} x_i(t - l_1) + a_{i,i,2} x_i(t - l_2) \\ &+ a_{i,j,3} x_j(t - l_3) + a_{i,k,4} x_k(t - l_4) + \varepsilon_i(t), \end{aligned} \tag{4}$$

indicating that, to determine the value of x_i at time t, we need the information of the values of x_i, x_j, and x_k at some previous times. We pack the information of interdependency of the components contained in Eq. (4) into the form,

$$x_i = f_i(x_i, x_j, x_k), \tag{5}$$

representing that component x_i is a function of components, x_i, x_j and x_k, where f_i stands for the function representing the time dependency of the i-th component, x_i. When we construct a network from this expression, each component of the multivariate time series such as x_i is translated to a node. Next, we draw directed arrows from x_j to x_i and from x_k to x_i, if the right hand side of the model of x_i contains x_j and x_k. This basic idea enables us to construct a directed network embodying the entire relationship among the components represented in a multivariate RAR model. In Eq. (5), the component x_i itself is included in the right hand side and the node x_i has a directed self-loop from x_i to x_i in the network. Such a case indicating that a component drives its own dynamics often happens.

4. Numerical Examples

We now demonstrate the application of our algorithm to simulated time series data, and confirm our theoretical arguments.

The system consists of four dynamical variables, $x_1(t)$, $x_2(t)$, $x_3(t)$ and $x_4(t)$, and their time dependencies are described by the following expressions:

$$\begin{aligned} x_1(t) = 0.7\ x_1(t-1) &- 0.4\ x_1(t-3) + 0.3\ x_2(t-4) \\ &+ 0.2\ x_4(t-7) + \varepsilon_1(t), \end{aligned} \tag{6}$$

$$x_2(t) = 3.0 + 0.6\ x_2(t-1) - 0.2\ x_2(t-6) + \varepsilon_2(t), \tag{7}$$

$$x_3(t) = 0.5\ x_1(t-2) + 0.3\ x_4(t-9) + \varepsilon_3(t), \tag{8}$$

$$\begin{aligned} x_4(t) = 0.2\ x_1(t-2) &+ 0.5\ x_4(t-1) \\ &- 0.3\ x_4(t-3) + \varepsilon_4(t), \end{aligned} \tag{9}$$

where $\varepsilon_i(t)$ $(i = 1, 2, 3, 4)$ are dynamic noise, independent and identically distributed Gaussian random variables with mean zero and standard deviation 1.0. In this paper, we distinguish "component" and "variable" as different technical terms. The term "component" is used for representing x_i and the term "variable" for representing $x_i(t - l)$ including its time delay. For instance, Eq. (6) has 3 components (x_1,

	x_1	x_2	x_3	x_4
x_1	1.0000	—	—	—
x_2	0.3606	1.0000	—	—
x_3	0.6691	0.1930	1.0000	—
x_4	0.4448	0.1790	0.4900	1.0000

Table 1: The largest absolute values of the cross correlation of all possible pairs between the time lag −30 and 30. The data are generated by Eqs. (6)–(9), and the values are estimated using 1000 data points with Gaussian observational noise with the mean zero and the standard deviation 0.01.

x_2 and x_4) and 4 variables, $x_1(t-1)$, $x_1(t-3)$, $x_2(t-4)$ and $x_4(t-7)$.

To construct the network by the common method, we estimate the cross correlation (CC) of all pairs using 1000 data points generated by Eqs. (6)–(9). Since we have four time series corresponding to the components, x_1, x_2, x_3, and x_4, the network contains four nodes. All the values are shown in Table 1. The network constructed from these CCs with threshold 0.5 is shown in Fig. 1(a). With this threshold, only the nodes, x_1 and x_3, are connected. The connection itself seems to be reasonable, because Eq. (8) representing the dynamics of $x_3(t)$ includes $x_1(t-2)$. However, Eq. (6) representing the dynamics of $x_1(t)$ does not include the component x_3. The undirectedness of the connection thus cannot capture this one-way relationship between x_1 and x_3. Furthermore, we would conclude that the pair, x_1 and x_4 are independent, since the value of CC between these components, 0.4452, is below the threshold 0.5. However, it is clearly untrue, because Eq. (9) representing the dynamics of $x_4(t)$ does include the variable $x_1(t-2)$. This simple example shows two insufficiencies of the common method: (i) the undirectedness of the edges that cannot capture the directions of the relationship between components and (ii) the arbitrariness of the threshold value that cannot always recover existing relationships. Hence, we consider that the network constructed using the values of the cross correlation function does not properly represent the exact relationship between components defined by Eqs. (6)–(9).

We apply the multivariate RAR models to the same data used for estimating the CCs. We obtain the multivariate RAR model exactly the same as Eqs. (6)–(9). The packed expressions such as Eq. (5) corresponding to this model become

$$x_1 = f_1(x_1, x_2, x_4), \tag{10}$$
$$x_2 = f_2(x_2), \tag{11}$$
$$x_3 = f_3(x_1, x_4), \tag{12}$$
$$x_4 = f_4(x_1, x_4). \tag{13}$$

Using this summarized information we construct a directed network, and the network is shown in Fig. 1(b). We consider that this structure appears to be a more faithful and straightforward network representation of the time structure of the system defined by Eqs. (6)–(9) than that constructed using the common method with an arbitrary value of threshold. Also, using the information we can know the number of incoming edges (in-degree) and outgoing edges (out-degree) of each node.

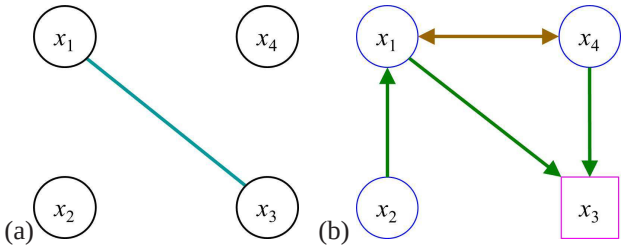

Figure 1: The constructed network: (a) The undirected netowk is based on the values of the cross correlation shown in Table 1 with threshold 0.5, (b) The directed network is based on the result of multvariate RAR models, Eqs. (10)–(13). In Fig. 1(b) the notation ○ means that the model for a component includes the component itself, and notation □ means that the component is not included in the model.

5. Applications

We apply the proposed method to multivariate time series from meteorological time series in the south pole [1]. The data we use are five different time series: the atmospheric pressure, the atmospheric temperature, the dew-point temperature, the vapor pressure and the humidity, taken hourly from 4 January to 15 February in 2015. As shown in Fig. 2, all of them show irregular fluctuations.

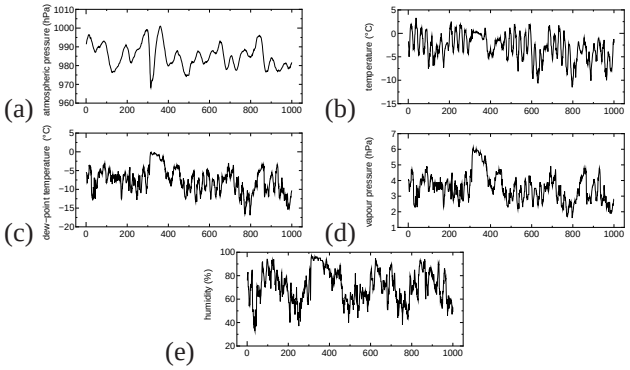

Figure 2: Meteorological hourly time series in the south pole from 4 January to 15 February in 2015: (a) atmospheric pressure, (b) temperature, (c) dew-point temperature, (d) vapor pressure and (e) humidity. These data are used for building multivariate RAR models.

We use 1000 data points (around 42 days) to build multivariate RAR models. As there are 5 time series, choosing

[1] The data can be obtained from Japan Meteorological Agency, http://www.jma.go.jp/jma/indexe.html

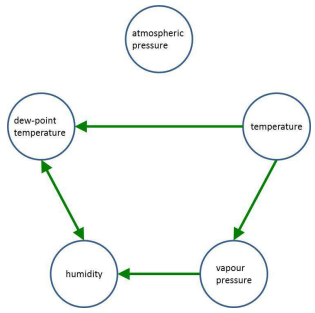

Figure 3: (Color online) The directed network constructed by multivariate RAR models of meteorological data in the south pole. All nodes are represented by ◯, as all models contain their own components. For the explanation of the notation, see Fig. 1(b).

Table 2: The number of in-degree and out-degree of the directed network of meteorological data shown in Fig. (3), where x_1 corresponds to atmospheric pressure, x_2 temperature, x_3 dew-point temperature, x_4 vapor pressure and x_5 humidity.

	x_1	x_2	x_3	x_4	x_5
in-degree	0	0	2	1	2
out-degree	0	2	1	1	1

a time delay up to 20 for time series of each data and the constant function give 101 candidate basis functions in the dictionary. Using the dictionary we build the multivariate RAR model for each data. The reduced expressions of the obtained 5 multivariate RAR models are

$$x_1 = f_1(x_1), \tag{14}$$
$$x_2 = f_2(x_2), \tag{15}$$
$$x_3 = f_3(x_2, x_3, x_5), \tag{16}$$
$$x_4 = f_4(x_2, x_4), \tag{17}$$
$$x_5 = f_5(x_3, x_4, x_5), \tag{18}$$

where x_1 corresponds to the atmospheric pressure, x_2 the atmospheric temperature, x_3 the dew-point temperature, x_4 the vapor pressure and x_5 the humidity.

In Fig. 3, we show the directed network constructed from these models, Eqs. (14)–(18), representing the relationship of interdependency among these five data. Note that all models contain their own components, which means that all nodes in Fig. 3 have self-loops. The numbers of in-degree and out-degree for each node in Fig. 3 are shown in Table 2. From these results we find that the atmospheric pressure is independent, the dew-point temperature and humidity are influenced eath other the vapor pressure are influenced by temperature and humidity.

6. Conclusion

We describe an algorithm for constructing directed networks from multivariate time series based on the RAR modeling technique. The strong point of this method is that it enables us to extract the hidden relationship among dynamical components from a dynamical system-wide perspective even if the time series do not have large values of cross correlation.

Acknowledgments

Tomo Nakamura would like to acknowledge the partial support of a Grant-in-Aid for Scientific Research (C) (No. 25282094) from the Japan Society for the Promotion of Science (JSPS), and Toshi Tanizawa also would like to acknowledge the support of a Grant-in-Aid for Scientific Research (C) (No. 24540419) from JSPS.

References

[1] D. H. Meadows. *Thinking in Systems: A Primer*. Chelsea Green Publishing, London, 2008.

[2] A-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[3] M. Small, D. M. Walker, and C. K. Tse. Scale-free distribution of avian influenza outbreaks. *Phys. Rev. Lett*, 99(18):188702, 2007.

[4] R. N. Mantegna. Hierarchical structure in financial markets. *The European Physical Journal B*, 11(1):193–197, 1999.

[5] K. Yamasaki, A. Gozolchiani, and S. Havlin. Climate networks around the globe are significantly affected by el niño. *Phys. Rev. Lett*, 100(22):228501, 2008.

[6] T. Nakamura, Y. Hirata, and M. Small. Testing for correlation structures in short-term variabilities with long-term trends of multivariate time series. *Phys. Rev. E*, 74:041114, 2006.

[7] M. Small and K. Judd. Detecting periodicity in experimental data using linear modeling techniques. *Phys. Rev. E*, 59:1379–1385, 1999.

[8] T. Ohira and T. Yamane. Delayed stochastic systems. *Phys. Rev. E*, 61:1247–1257, 2000.

[9] K. Judd and A. Mees. On selecting models for nonlinear time series. *Physica D*, 82:426–444, 1995.

[10] K. Judd and A. Mees. Embedding as a modeling problem. *Physica D*, 120:273–286, 1998.

[11] T. Nakamura, D. Kilminster, K. Judd, and A. Mees. A comparative study of model selection methods. *Int. J. Bifurcation Chaos*, 14(3):1129–1146, 2004.