Impact of lag information on network inference

Nicolás Rubido[1] and Cristina Masoller[2, *]

[1] Universidad de la República, Instituto de Física de la Facultad de Ciencias, Iguá 4225, 11400 Montevideo, Uruguay
[2] Departament de Fisica, Universitat Politecnica de Catalunya,
Rambla Sant Nebridi 22, ES-08222 Terrassa, Barcelona, Spain

Extracting useful information from data is a fundamental challenge across disciplines as diverse as climate, neuroscience, genetics, and ecology. In the era of "big data", data is ubiquitous, but appropriated methods are needed for gaining reliable information from the data. In this work we consider a complex system, composed by interacting units, and aim at inferring which elements influence each other, directly from the observed data. The only assumption about the structure of the system is that it can be modeled by a network composed by a set of N units connected with L un-weighted and un-directed links, however, the structure of the connections is not known. In this situation the inference of the underlying network is usually done by using interdependency measures, computed from the output signals of the units. We show, using experimental data recorded from randomly coupled electronic Rössler chaotic oscillators, that the information of the lag times obtained from bivariate cross-correlation analysis can be useful to gain information about the real connectivity of the system.

INTRODUCTION

Network inference involves discovering, from observations, the underlying connectivity between the elements of a complex system. Reliable inference is important because it allows to understand, predict, and control complex behaviours. Main challenges involve the fact that usually one can only observe a single scalar variable (but the evolution of the system depends on other –unobserved variables), during a limited time-interval, with limited resolution, and with considerable measurement noise. In recent years many network inference methods (or network reconstruction) have been proposed [1–15], whose success depends, not only on the previous knowledge of the system (e.g., weighted or unweighted interactions, directed or undirected, instantaneous or lagged), but also, on data availability (e.g., hidden nodes or unobserved variables and limited temporal or spatial resolution).

Relevant examples of network inference include brain functional networks and climate networks. Brain functional networks, which have shed light into many neurological conditions, such as Alzheimer, Parkinson or Epilepsy, are inferred from recorded brain signals (magneto-encephalography, MEG, electro-encephalography, EEG, and functional magnetic resonance imaging, fMRI) by correlating different brain regions and linking the ones that exhibit the highest correlations [16–18]. Similarly, climate networks have shed light into climate phenomena (such as long range teleconnections or atmosphere-ocean interactions) by correlating time-series of climate variables and linking the geographical regions that exhibit significant correlation. [19–23].

When trying to infer a system's connectivity, the statistical similarity of the time series recorded from different units is commonly measured by using bivariate time series analysis, such as cross-correlation or mutual information. Typically, time series are mutually lagged in order to find the maximum of the similarity measure, S_{ij}, but the information contained in the set of lag times, τ_{ij} has not yet been used to infer the links of the network. In climate network studies, the lag times have been used to infer the directionality of the links; in addition, lag analysis has received attention in the context of financial and ecological data analysis [24–26]. In this work we investigate if the discovery of the real interactions in a complex network can be improved if these lag times are taken into account. The working assumption is that, when the strength of the coupling is increased, the transition to synchronized behavior occurs [27]. During this transition, the lags between the time-series of nodes which have direct interactions can be smaller than the lags between nodes that are not directly coupled. In other words, if two nodes have a direct link between them, they can synchronize with a lag that is smaller than the lag between nodes that are not directly connected, and this difference can be used for improving network inference.

Here, we analyze under which conditions the lag information can be used to complement the information of the similarity measure for improving the discovery of the existing links (true positives), for decreasing the number of wrongly inferred links (false positives), for improving the detection of non-existing links (true negatives) and for decreasing the number of mistakes due to undetected links (false negatives).

We investigate an experimental dataset from a network of electronic Rössler chaotic oscillators [6] composed by $N = 12$ units, which are randomly coupled with $L = 19$ links. The real underlying adjacency matrix, A_{ij}, is known, while an inferred matrix, A_{ij}^*, is extracted by using bivariate time series analysis. We propose three criteria for classifying links as existing or non-existing and, by comparing the inferred and the known coupling

matrices, we discuss the effectiveness of the different criteria for uncovering the real connectivity of the system. We conclude that when using an OR criterion, namely, one that detects a link when either a small lag or a high similarity value is found, non-existent links are almost always correctly discarded. However, this criterion also discards direct links more than the other two criteria.

DATA

The data was described in [6]. It is generated from $N = 12$ Rössler electronic oscillators randomly coupled with $L = 19$ links. The coupling between units i and j is kA_{ij}, where k is the coupling strength and A_{ij} is the adjacency matrix [$A_{ij} = 1$ if the oscillators i and j are coupled and $A_{ij} = 0$ if they are not]. The dataset contains the time series recorded for 31 values of the coupling strength (the minimum is $k = 0$ and the maximum is $k = 0.15$). Each time series has 30000 data points. In order to reduce the effects of noise, each time series is divided into non-overlapping segments of length T. To avoid transient effects the first segment is disregarded, and the following segments are used for the analysis. The results presented are obtained with $T = 5000$, so we have five segments for computing the mean values and the error bar of the measures described in the following section.

METHODS

The lagged cross-correlation is used to quantify the similarity between the time series recorded from nodes i and j, $x_i(t)$ and $x_j(t)$, with $t \in [1, T]$ and $i, j \in [1, N]$. Specifically, each time series is first normalized to zero-mean and unit variance. Then, we calculate the Pearson coefficient,

$$C_{ij}(\tau) = \frac{1}{T - \tau_{max}} \left| \sum_{t=1}^{T - \tau_{max}} x_i(t) \, x_j(t + \tau) \right|, \quad (1)$$

varying τ in the interval $[0, \tau_{max}]$ with $\tau_{max} = T/5$ [6]. We define the lag, τ_{ij}, between nodes i and j as the value of τ that maximizes $C_{ij}(\tau)$, and we define the correlation strength, S_{ij}, as $S_{ij} \equiv C_{ij}(\tau_{ij}) = \max_\tau C_{ij}(\tau)$.

Next, we test whether the information contained in the matrices τ_{ij} and S_{ij} is useful for inferring the existing links. We define two thresholds, one for the lags, τ_{th}, and one for the correlation strengths, S_{th}, and use the following criteria for classifying the links between the existing and the non-existing ones.

1. **SIM**: Only the similarity measure (CC) is used to infer the links. The link between i and j exists ($A_{ij}^* = 1$) if $S_{ij} > S_{th}$, else, the link does not exist ($A_{ij}^* = 0$).

2. **AND**: The link between i and j exists ($A_{ij}^* = 1$) if $\tau_{ij} < \tau_{th}$ and $S_{ij} > S_{th}$, else, the link does not exist ($A_{ij}^* = 0$).

3. **OR**: The link between i and j exists ($A_{ij}^* = 1$) if $\tau_{ij} < \tau_{th}$ or $S_{ij} > S_{th}$, else, the link does not exist ($A_{ij}^* = 0$).

For these criteria, the thresholds S_{th} and τ_{th} are chosen such that they return a number of links as close as possible to the (known) number of existing links.

To quantify the efficiency of these criteria for uncovering the real connectivity of the network we use the following measures

- True negatives (TN): number of non-existing links which are correctly classified as not existing, relative to the number of non-existing links;

- False negatives (FN): number of existing links which are incorrectly classified as not existing, relative to the number of existing links;

- True positives (TP): number of existing links which are correctly classified as existing, relative to the number of existing links;

- False positives (FP): number of non-existing links which are incorrectly classified as existing, relative to the number of non-existing links.

We also quantify the global success of the inference method by calculating the total wrongly predicted existent, FP, and non-existent, FN, links relative to the total number of links:

$$\Delta = \frac{2}{N(N-1)} \sum_i \sum_{j>i} |A_{ij} - A_{ij}^*| = \frac{FN + FP}{N(N-1)/2}. \quad (2)$$

Finally, since both τ_{ij} and S_{ij} depend on the length of the segment, T, of the time series, and of the maximum lag, τ_{max}, we analyze if the results are robust with respect to the choice of these parameters.

RESULTS

In Fig. 1 we show the values of the similarity measure, S_{ij}, [see Eq. (1)] and the corresponding lag, τ_{ij} separating the links that exist ($A_{ij} = 1$, left column panels) and the links that do not exist ($A_{ij} = 0$, right column panels). We clearly observe a different variation as the coupling strength is increased: for the existing links, S_{ij} tends to increase faster in comparison with the non-existing links, and the opposite happens with τ_{ij}.

Next, we show how including the lag information into the bivariate analysis for the network inference can be
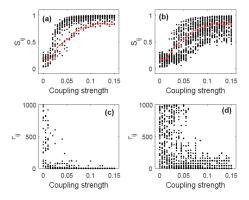
FIG. 1. Analysis of experimental data ($N = 12$ Rössler electronic oscillators coupled in a random network): comparison of the values of the similarity measure, S_{ij}, and lags , τ_{ij}, of the existing links [panels **(a)** and **(c)**, respectively] and of the non-existing links [panels **(b)** and **(d)**, respectively]. It is observed that the variation of S_{ij} and τ_{ij} with the coupling tends to be different: for the existing links, S_{ij} increases faster (and τ_{ij} decreases faster) than for the non-existing links. In panels (a), (b) the solid lines indicate the value of the similarity measure averaged over all (existing and non-existing) links.
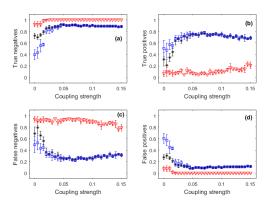

FIG. 2. Quantitative comparison of the inference of the network obtained with the three criteria described in Sec. : SIM (*), AND (squares) and OR (triangles). The analysis is performed in six non-superposed segments of $T = 5000$ data points each. In each segment a network is inferred and then the true positives (TPs), false positives (FPs), true negatives (TNs) and false negatives (FNs) are calculated. The symbols and the error bars indicate the corresponding mean values and standard deviations, computed from the TP, FP, TN and FN values obtained in the six segments.
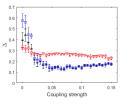

FIG. 3. Comparison of the total number of mistakes, as defined in Eq. (2), for the three inference criteria: SIM (*), AND (squares) and OR (triangles).

useful. Figure 2 displays the different types of errors that are made when applying the criteria described in Sec. . We see that the SIM and AND only differ for weak coupling, but as the coupling increases, the number of correctly inferred links and mistakes made are the same for the two criteria. At weak coupling, adding the lag information (AND) improves the detection of the existing links (true positives), at the cost of also improving the detection of not-existing links (false positives). When considering the total mistakes, as defined in Eq. (2), we can see in Fig. 3 that the AND criteria produces, at low coupling, more mistakes than the SIM criteria. Consequently, in this system the lag information is helpful, at low coupling, to decrease particular types of mistakes of the inference process. However, using similarity values alone is better if the main goal is to minimize the total number of mistakes, i.e., the sum of wrongly predicted existing, FP, and non-existing, FN, links.

Interestingly, the OR criteria gives very different results: avoids the false positives at the cost of giving a large number of false negatives. This is due to the fact that, regardless of the coupling strength, many τ_{ij} values are small for both, existing and non-existing links (see Fig. 1).

The analysis of how relevant the particular choice of T and τ_{max} is, indicates that these results are robust. In Fig. 4 we consider a given coupling strength ($k = 0.1$) and change the length of the segment, T, of the time series, while in Fig. 5 we change the interval of lag values where we search for the maximum of the cross correlation. In both cases we see that for T and τ_{max} large enough, the number of FPs, TPs, FNs, and TNs are independent of

the choice of T and τ_{max}.

CONCLUSIONS AND DISCUSSION

We have explored the possibility of using the lag information from pairwise cross-correlation for improving the inference of the connectivity of a system composed by interacting units. We have limited the study to a simple situation: the number of units and the number of links between pairs of units are known, and the links either exist or do not exist (i.e., they are undirected and unweighted).

We have used data recorded from by 12 electronic Rössler chaotic oscillators, coupled in a random network that has 19 links. We have found that lag information can be useful for reducing certain types of mistakes, but it does not improve network inference when all the mistakes are added up. In particular, we find that we can decrease the number of false positive detections and detect all true negatives in a robust way – independent of the coupling
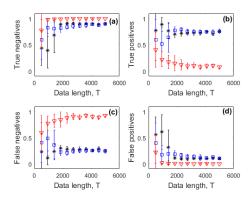
FIG. 4. Robustness of the results obtained with the three inference criteria, with respect to the length, T, of the data segment for coupling strength $k = 0.1$. For $k = 0$ there is no effect of T (not shown).
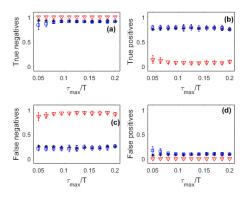

FIG. 5. Robustness of the results obtained with the three inference criteria, with respect to the length, τ_{max}, of the interval where we search for the maximum of the cross-correlation, normalized to the length of the data segment. The coupling strength is $k = 0.1$; for $k = 0$ there is no effect of τ_{max} (not shown).

strength value – when using an OR criterion. Regarding the total number of errors, guided by Fig. 3 we conclude that, in our system, for weak coupling the OR criterion is the best option, while for intermediate coupling SIM performs the same as AND and both outperform OR. If the coupling is large enough to synchronize the system, it is not possible, using these criteria, to infer the network.

In general, a drawback of including the lag information is that one needs to select two inference thresholds, S_{th}, for the similarity values, and τ_{th}, for the lags. Here we used the simplest approach: we varied them simultaneously [S_{th} was increased linearly from $\min(S_{ij})$ while τ_{th} was decreased linearly from $\max(\tau_{ij})$], and the pair of values S_{th} and τ_{th} that returned a number of links closer to the known number of existing links were used for the inference. This approach requires a minimal knowledge from the network, namely, its link density – number of

links and network size. However, a set of S_{th} and τ_{th} values that returned the target number of links was not always found. A possible way to improve the inference method is by considering two independent thresholds; however, an important consideration is the shape of the distributions the S_{ij} and τ_{ij} values: if they are bimodal or long-tailed, there might not be any combination of thresholds that returns the target number of links, because a small variation of one of the thresholds might result in either too many or too few links being classified as existent.

In a realistic situation the number of existing links is unknown. Therefore, choosing a set of thresholds S_{th} and τ_{th} that return a pre-defined number of links is not an appropriated inference strategy. In this situation a possible alternative for using lag information for network inference is by taking into account how the lags and the similarity measures vary with the coupling strength. Here we have not used the fact that when the link between units i and j indeed exists, S_{ij} (τ_{ij}) tends to increase (decrease) with the coupling faster than when the link does not exist. However, classifying links according to the variation of S_{ij} and τ_{ij} with the coupling, increases the data requirements, as the values of S_{ij} and τ_{ij} will need to be compared for different coupling conditions. In addition, in systems with non-instantaneous interactions (i.e., coupling delays) or in systems where the units display periodic behavior, the lag information will not be useful for network inference because in such systems the units can synchronize with lags between them which do not have a clear relation with the underlying interactions.

ACKNOWLEDGMENTS

———————

* cristina.masoller@upc.edu

[1] A. A. Margolin, I. Nemenman, K. Basso, et al. "ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context", BMC Bioinformatics **7**, (2006) S7.

[2] M. Timme, "Revealing network connectivity from response dynamics", Phys. Rev. Lett. **98**, (2007) 224101.

[3] W. X. Wang, Y. C. Lai, C. Grebogi, et al. "Network reconstruction based on evolutionary-game data via compressive sensing", Phys. Rev. X **1**, (2011) 021021.

[4] N. Rubido, A. C. Marti, E. Bianco-Martinez, et al. "Exact detection of direct links in networks of interacting dynamical units", New J. Phys. **16**, (2014) 093010.

[5] E. Bianco-Martínez, N. Rubido, Ch. G. Antonopoulos, and M. S. Baptista, "Successful network inference from time-series data using mutual information rate", Chaos **26**, 043102 (2015).

[6] G. Tirabassi, R. Sevilla-Escoboza, J. M. Buldú and C. Masoller, "Inferring the connectivity of coupled oscillators from time series statistical similarity analysis", Sci. Rep. **5**, (2015) 10829.

[7] J. F. Donges, J. Heitzig, B. Beronov, et al. "Unified functional network and nonlinear time series analysis for complex systems science: The pyunicorn package", Chaos **25**, (2015) 113101.

[8] W. Wiedermann, J. F. Donges, J. Kurths, et al. "Spatial network surrogates for disentangling complex system structure from spatial embedding of nodes", Phys. Rev. E **93**, (2016) 042308.

[9] A. Pikovsky, "Reconstruction of a neural network from a time series of firing rates", Phys. Rev. E **93**, (2016) 062313.

[10] R. Cestnik and M. Rosenblum, "Reconstructing networks of pulse-coupled oscillators from spike trains", Phys. Rev. E **96**, (2017) 012209.

[11] E. A. Martin, J. Hlinka, A. Meinke, et al. "Network Inference and Maximum Entropy Estimation on Information Diagrams", Sci. Rep. **7**, (2017) 7062.

[12] B. J. Stolz, H. A. Harrington, and M. A. Porter, "Persistent homology of time-dependent functional networks constructed from coupled time series", Chaos **27**, (2017) 047410.

[13] J. Casadiego, N. Nitzan, S. Hallerberg et al. "Model-free inference of direct network interactions from nonlinear collective dynamics", Nat. Comm. **8**, (2017) 2192.

[14] E. S. C. Ching and H. C. Tam, "Reconstructing links in directed networks from noisy dynamics", Phys. Rev. E **95**, (2017) 010301.

[15] L. Li, D. Xu, H. Peng, et al. "Reconstruction of complex network based on the noise via QR decomposition and compressed sensing", Sci. Rep. **7**, (2017) 15036.

[16] V. M. Eguiluz, D. R. Chialvo, G. A. Cecchi, et al. "Scale-free brain functional networks", Phys. Rev. Lett. **94**, (2005) 018102.

[17] B. T. Bullmore and O. Sporns, "Complex brain networks: graph theoretical analysis of structural and functional systems", Nat. Rev. Neuroscience **10**, (2009) 186–198.

[18] K. Lehnertz, G. Ansmann, S. Bialonski, et al. "Evolving networks in the human epileptic brain", Physica D **267**, (2014) 7–15.

[19] K. Yamasaki, A. Gozolchiani and S. Havlin, "Climate networks around the globe are significantly affected by El Nino", Phys. Rev. Lett. **100**, (2008) 228501.

[20] A. A. Tsonis and K. L. Swanson, "Topology and predictability of El Nino and La Nina networks", Phys. Rev. Lett. **100**, (2008) 228502.

[21] J. F. Donges, Y. Zou, N. Marwan and J. Kurths, Eur. Phys. J. Spec. Top. **174**, (2009) 157–179.

[22] S. Bialonski, M. T. Horstmann, and K. Lehnertz, "From brain to earth and climate systems: Small-world interaction networks or not?", Chaos **20**, (2010) 013134.

[23] J. I. Deza, M. Barreiro and C. Masoller, "Inferring interdependencies in climate networks constructed at inter-annual, intra-season and longer time scales", Eur. Phys. J. Spec. Top. **222**, (2013) 511–523.

[24] J. D. Olden and B. D. Neff, "Cross-correlation bias in lag analysis of aquatic time series", Marine Biology **138**, 1063–1070 (2001).

[25] C. Curme, *Lagged correlation networks* (Doctoral dissertation, Boston University, 2015).

[26] P. Damos, "A stepwise algorithm to detect significant time lags in ecological time series in terms of auto-correlation functions and ARMA model optimisation of pest population seasonal outbreaks", Stochastic Environmental Research and Risk Assessment **30**(7), 1961–1980 (2016).

[27] A. Pikovsky, M. Rosenblum, and J. Kurths, *Synchronization: A universal concept in nonlinear sciences* (Cambridge University Press, 2001).

[28] R. Sevilla-Escoboza and J. M. Buldú, Data in Brief **7**, (2016) 1185–1189.