Synergy, redundancy and unnormalized Granger causality

S. Stramaglia[1], L. Angelini[2], J.M. Cortes[3] and D. Marinazzo[4]

***Abstract*— We analyze by means of Granger causality the effect of synergy and redundancy in the inference (from time series data) of the information flow between subsystems of a complex network. Whilst fully conditioned Granger causality is not affected by synergy, the pairwise analysis fails to put in evidence synergetic effects. We show that maximization of the total Granger causality to a given target, over all the possible partitions of the set of driving variables, puts in evidence redundant multiplets of variables influencing the target, provided that an *unnormalized* definition of Granger causality is adopted. Along the same lines we also introduce a pairwise index of synergy (w.r.t. to information flow to a third variable) which is zero when two independent sources additively influence a common target; thus, this definion differs from previous definitions of synergy.**

I. INTRODUCTION

The inference of dynamical networks from time series data is related to the estimation of the information flow between variables [1]. Granger causality (GC) [2] has emerged as a major tool to address this issue. GC is based on prediction: if the prediction error of the first time series is reduced by including measurements from the second one in the linear regression model, then the second time series is said to have a Granger causal influence on the first one.

The pairwise Granger analysis consists in assessing GC between each pair of variables, independently of the rest of the system. It is well known that the pairwise analysis cannot disambiguate direct and indirect interactions among variables. The most straightforward extension, the conditioning approach, removes indirect influences by evaluating to which extent the predictive power of the driver on the target decreases when the conditioning variable is removed. As a convenient alternative to this suboptimal solution, a partially conditioned approach has been proposed [3], consisting in conditioning on a small number of variables chosen as the most informative ones for the driver node. Sometimes though, a fully conditioned approach can also encounter conceptual limitations (in addition to be practically unfeasible and computationally expensive): in the presence of redundant variables the application of the standard analysis leads to underestimating the influences [4]. Redundancy and synergy are intuitive yet elusive concepts, which have been investigated in different fields, including pure information theory [5], [6], [7], [8], [9], [10]. A complementary concept to redundancy is synergy. The synergetic effects that we address here, related to the analysis of dynamical influences in multivariate time series, are similar to those encountered in sociological and psychological modeling, where *suppressor* is the name given to variables that increase the predictive validity of another variable after its inclusion into a linear regression equation [11]. For further details, see also [12], [7], where information-based approaches were applied to address collective influences.

[1]Sebastiano Stramaglia is with the Physics Department of the University in Bari and Istituto Nazionale di Fisica Nucleare - Sezione di Bari, and with the Basque Center for Applied Mathematics, Bilbao, Spain sebastiano.stramaglia@ba.infn.it

[2]Leonardo Angelini is with the Physics Department of the University in Bari and Istituto Nazionale di Fisica Nucleare - Sezione di Bari angelini@ba.infn.it

[3]Jesus M Cortes is with the Computational Neuroimaging Lab, Biocruces Health Research Institute, Cruces University Hospital, Barakaldo, Spain and with Ikerbasque, The Basque Foundation for Science, Bilbao, Spain jesus.m.cortes@gmail.com

[4]Daniele Marinazzo is with the Faculty of Psychology and Educational Sciences, Department of Data Analysis, Ghent University, Henri Dunantlaan 1, B-9000, Ghent, Belgium daniele.marinazzo@ugent.be

II. GRANGER CAUSALITY

Granger causality is a powerful and widespread data-driven approach to determine whether and how two time series exert direct dynamical influences on each other [13]. A convenient nonlinear generalization of GC has been implemented in [14] by exploiting the kernel trick, which makes computation of dot products in high-dimensional feature spaces possible using simple functions (kernels) defined on pairs of input patterns. This trick allows the formulation of nonlinear variants of any algorithm that can be cast in terms of dot products, for example Support Vector Machines [15]. Thus, although the aim in [14] is still to perform linear GC, it does it within a space defined by the nonlinear features of the data. This projection is conveniently and implicitly performed through kernel functions [16] in addition to use a statistical procedure to avoid over-fitting.

Quantitatively, let us consider n time series $\{x_\alpha(t)\}_{\alpha=1,\ldots,n}$; the lagged state vectors are denoted

$$X_\alpha(t) = (x_\alpha(t-m),\ldots,x_\alpha(t-1)),$$

where m is the order of the model (window length). Let $\varepsilon(x_\alpha|\mathbf{X})$ be the mean squared error prediction of x_α on the basis of all the vectors $\mathbf{X} = \{X_\beta\}_{\beta=1}^n$ (corresponding to the kernel approach described in [17]). The fully conditioned GC index $\delta_{FCGC}(\beta \to \alpha)$ is defined as follows: consider the prediction of x_α on the basis of all the variables except X_β and similarly the prediction of x_α using all the variables, then FCGC is the logarithm variation of the error for the two conditions, i.e.

$$\delta_{FCGC}(\beta \to \alpha) = \log \frac{\varepsilon\left(x_\alpha|\mathbf{X}\setminus X_\beta\right)}{\varepsilon\left(x_\alpha|\mathbf{X}\right)}. \qquad (1)$$

It was shown in [18] that not all the kernels are suitable to estimate GC. Indeed, two important classes of kernels used

to construct nonlinear GC measures are the *inhomogeneous polynomial kernel* (whose features are all the monomials in the input variables up to the p-th degree; $p=1$ corresponds to linear Granger causality) and the *Gaussian kernel*.

The pairwise Granger causality is given by:

$$\delta_{PWGC}(\beta \to \alpha) = \log \frac{\varepsilon\left(x_\alpha | X_\alpha\right)}{\varepsilon\left(x_\alpha | X_\alpha, X_\beta\right)}. \tag{2}$$

III. UNNORMALIZED GRANGER CAUSALITY

Interaction information is a classical measure of the amount of information (redundancy or synergy) bound up in a set of three variables [19], [20]. A generalization of the interaction information to the case of lagged interactions was addressed in [21]. It is important to emphasize that the sign of the interaction information corresponds to synergy or redundancy, and that this interpretation implies that synergy and redundancy are taken to be mutually exclusive qualities of the interactions between variables [22]. Other approaches, instead, consider that synergy and redundancy are separated entities; for example, the *partial information decomposition* (PID) approach in [23] showed that the information that two source variables Y and Z hold about a third target variable X can be decomposed into four parts: (i) the unique information that only Y (out of Y and Z) holds about X; (ii) the unique information that only Z holds about X; (iii) the redundant information that both Y and Z hold about X; and (iv) the synergetic information about X that only arises from knowing both Y and Z. For Gaussian variables, where these four contributions were calculated analytically [24], it was shown some undesirable results, e.g., redundancy reduces to the minimum information provided by either source variable, and hence it is independent on the correlation between sources. As suggested in [24], this occurrence may be related to the fact that to evaluate the Shannon information for continuous random variables is more convenient to do it by using the differential entropy, as the limit to the continuum is not straightforward [25]. This is why here, for the case of continuous variables, we propose to describe the informational character of a subset of variables in terms of the reduction of residuals variance of the target due to inclusion of driver variables, similar to the strategy followed in [4]. The informational character of each multiplet will be associated to a single number, which may be seen as the difference of redundancy and synergy in every formalism where these two notions are separately defined (see the discussion in [22]).

First of all, we note that the straightforward generalization of Granger causality for driving sets of variables is

$$\delta_{\mathbf{X}}(B \to \alpha) = \log \frac{\varepsilon\left(x_\alpha | X_\alpha, \mathbf{X} \setminus B\right)}{\varepsilon\left(x_\alpha | X_\alpha, \mathbf{X}\right)}, \tag{3}$$

where B are is a subset of variables, x_α is the target variable and $\mathbf{X} \setminus B$ means the set of all variables except for those X_β with $\beta \in B$. Note that we have isolated the variable X_α, i.e. the present state of the target. The subscript $\mathbf{X}$ has been included to put in evidence the conditioning variables used to evaluate GC.

On the other hand, an unnormalized version of it is given by

$$\delta^u_{\mathbf{X}}(B \to \alpha) = \varepsilon\left(x_\alpha | X_\alpha, \mathbf{X} \setminus B\right) - \varepsilon\left(x_\alpha | X_\alpha, \mathbf{X}\right). \tag{4}$$

It can be easily shown that it satisfies the following interesting property: if $\{X_\beta\}_{\beta \in B}$ are statistically independent and their contributions in the model for x_α are additive, then

$$\delta^u_{\mathbf{X}}(B \to \alpha) = \sum_{\beta \in B} \delta^u_{\mathbf{X}}(\beta \to \alpha). \tag{5}$$

We remark that this property does not hold for the standard definition of Granger causality; neither for entropy-rooted quantities [22], due to the presence of the logarithm.

In order to identify the informational character of a set of variables B, concerning the causal relationship $B \to \alpha$, we remind that, in general, synergy occurs if B contributes to α with more information than the sum of all its variables, whilst redundancy corresponds to situations with the same information being shared by the variables in B [20]. We can render quantitatively these notions and define the variables in B to be *redundant* if $\delta^u_{\mathbf{X}}(B \to \alpha) > \sum_{\beta \in B} \delta^u_{\mathbf{X}}(\beta \to \alpha)$, and *synergetic* if $\delta^u_{\mathbf{X}}(B \to \alpha) < \sum_{\beta \in B} \delta^u_{\mathbf{X}}(\beta \to \alpha)$. In order to justify these definitions, firstly we observe that the case of independent variables (and additive contributions) does not fall in the redundancy case nor in the synergetic case, due to (5), as it should be. Moreover, we describe the following example for two variables X_1 and X_2. If X_1 and X_2 are redundant, then removing X_1 from the input variables of the regression model does not have a great effect, as X_2 provides the same information as X_1; this implies that $\delta^u_{\mathbf{X}}(X_1 \to \alpha)$ is nearly zero. The same reasoning holds for X_2, hence we expect that $\delta^u_{\mathbf{X}}(\{X_1, X_2\} \to \alpha) > \delta^u_{\mathbf{X}}(X_1 \to \alpha) + \delta^u_{\mathbf{X}}(X_2 \to \alpha)$. Conversely, let us suppose that X_1 and X_2 are synergetic, i.e. they provide some information about α only when both the variables are used in the regression model; in this case $\delta^u_{\mathbf{X}}(\{X_1, X_2\} \to \alpha)$, $\delta^u_{\mathbf{X}}(X_1 \to \alpha)$ and $\delta^u_{\mathbf{X}}(X_2 \to \alpha)$ are almost equal and therefore $\delta^u_{\mathbf{X}}(\{X_1, X_2\} \to \alpha) < \delta^u_{\mathbf{X}}(X_1 \to \alpha) + \delta^u_{\mathbf{X}}(X_2 \to \alpha)$.

Two analytically tractable cases are now reported as examples. Consider two stationary and Gaussian time series $x(t)$ and $y(t)$ with $\langle x^2(t) \rangle = \langle y^2(t) \rangle = 1$ and $\langle x(t) y(t) \rangle = \mathscr{C}$; they correspond, e.g., to the asymptotic regime of the autoregressive system

$$\begin{aligned} x_{t+1} &= a x_t + b y_t + \sigma \xi^{(1)}_{t+1} \\ y_{t+1} &= b x_t + a y_t + \sigma \xi^{(2)}_{t+1}, \end{aligned} \tag{6}$$

where ξ are i.i.d. unit variance Gaussian variables, $\mathscr{C} = 2ab/(1-a^2-b^2)$ and $\sigma^2 = 1 - a^2 - b^2 - 2ab\mathscr{C}$. Considering the time series $z_{t+1} = A(x_t + y_t) + \sigma' \xi^{(3)}_{t+1}$ with $\sigma' = \sqrt{1 - 2A^2(1+\mathscr{C})}$, we obtain for $m = 1$:

$$\delta^u_{\mathbf{X}}(\{x,y\} \to z) - \delta^u_{\mathbf{X}}(x \to z) - \delta^u_{\mathbf{X}}(y \to z) = A^2(\mathscr{C} + \mathscr{C}^2). \tag{7}$$

Hence x and y are redundant (synergetic) for z if $\mathscr{C}$ is positive (negative). Turning to consider $w_{t+1} = B\, x_t \cdot y_t + \sigma'' \xi^{(4)}_{t+1}$ with $\sigma'' = \sqrt{1 - B^2(1+2\mathscr{C})^2}$, and using the polynomial kernel with $p = 2$, we have

$$\delta^u_{\mathbf{X}}(\{x,y\} \to z) - \delta^u_{\mathbf{X}}(x \to z) - \delta^u_{\mathbf{X}}(y \to z) = B^2(4\mathscr{C}^2 - 1); \tag{8}$$

x and y are synergetic (redundant) for w if $|\mathscr{C}| < 0.5$ ($|\mathscr{C}| > 0.5$).

The presence of redundant variables leads to underestimation of their Granger causality when the standard multivariate approach is applied (as it is clear from the discussion above, this is not the case for synergetic variables). Redundant variables should then be grouped to get a reliable measure of Granger causality, and to characterize interactions in a more compact way. As it is clear from the discussion above, grouping redundant variables is connected to maximization of the un-normalized Granger causality index (4) and, in the general setting, can be made as follows. For a given target α, we call B the set of the remaining $n-1$ variables. The partition $\{A_\ell\}$ of B, maximizing the total Granger causality

$$\Delta = \sum_\ell \delta^u_{\mathbf{X}}(A_\ell \to x_\alpha),$$

consists of groups of redundant variables.

A. PAIRWISE SYNERGY INDEX

The discussion in the previous section suggests to describe quantitatively the informational character of two variables i and j, providing information for the future state of the variable x_α, by the following pairwise synergy index (PSI):

$$\begin{aligned}\Psi_\alpha(i,j) &= \delta^u_{\mathbf{X}\setminus\mathbf{j}}(i\to\alpha) - \delta^u_{\mathbf{X}}(i\to\alpha) \\ &= \delta^u_{\mathbf{X}}(\{i,j\}\to\alpha) - \delta^u_{\mathbf{X}}(i\to\alpha) - \delta^u_{\mathbf{X}}(j\to\alpha),\end{aligned}$$

where $\mathbf{X}$ is the set of conditioning variables. Ψ is negative for increased unnormalized causality $i\to\alpha$ due to the inclusion of j in the conditioning variables (positive PSI corresponds to redundancy). Note that if i and j are statistically independent and they cause α additively then PSI is zero, differently from interaction information, where a common effect of two causes induces a dependency among the causes that did not formerly exist [26].

Another interpretation of Ψ is given by the cumulant expansion of the prediction error of x_α:

$$\varepsilon(x_\alpha|X_\alpha) - \varepsilon(x_\alpha|\{X_\alpha,\mathbf{X}\}) = \sum_{B\subset\mathbf{X}} S(B). \tag{9}$$

Equation (9) can be solved by a Moebius inversion, which yields

$$S(B) = \sum_{\Gamma\subset B} (-1)^{|n_B|+|n_\Gamma|}\ \delta^u_B(\Gamma\to\alpha), \tag{10}$$

where $|n_B|$ and $|n_\Gamma|$ are the number of variables in the subsets B and Γ. The first order cumulant is then

$$S(i) = \delta^u_i(i\to\alpha), \tag{11}$$

the second cumulant is

$$S(i,j) = \delta^u_{ij}(\{ij\}\to\alpha) - \delta^u_{ij}(i\to\alpha) - \delta^u_{ij}(j\to\alpha), \tag{12}$$

the third cumulant is

$$\begin{aligned}S(i,j,k) = {}& \delta^u_{ijk}(\{ijk\}\to\alpha) - \delta^u_{ijk}(\{ij\}\to\alpha) \\ & -\delta^u_{ijk}(\{jk\}\to\alpha) - \delta^u_{ijk}(\{ik\}\to\alpha) \\ & +\delta^u_{ijk}(i\to\alpha) + \delta^u_{ijk}(j\to\alpha) + \delta^u_{ijk}(k\to\alpha),\end{aligned} \tag{13}$$

and so on. The index PSI may then be seen as the order two cumulant of the expansion of the prediction error of the target variable; equation (10) allows also the generalization to higher order terms. Obviously PSI also depends also on the choice of the kernel, i.e. on the choice of the regression model.

IV. CONCLUSIONS

In this paper we have considered the inference, from time series data, of the information flow between subsystems of a complex network, an important problem in medicine and biology. In particular we have analyzed the effects that synergy and redundancy induce on the Granger causal analysis of time series; it is well known that the presence of redundancy and synergy degrades the performance of GC methods. Here we have introduced a frame for data analysis based on unnormalized Granger causality, i.e. the reduction of variance of the residuals of each target variables when candidate driver variables are included in the regression model. Maximizing the total unnormalized Granger causality leads to group redundant variables. Finally, we have introduced a pairwise index of synergy, which for each pair of variables measures how much they interact to provide better predictions of the target. Such index can be seen as the second cumulant in the expansion of the prediction error of the target variable, to be compared with the expansion of the transfer entropy in [21] which provides the interaction information as the second cumulant. The advantages provided by the present cumulant expansion are (i) conceptual problems found in the Gaussian case [24] are avoided, and (ii) the nonlinearity of PSI can be easily controlled by varying the kernel in the regression model. We have thus introduced a novel frame to study interdependencies among subcomponents of complex systems from data. A pitfall of unnormalized GC is the occurrence that the connection with information theory is lost, but the aim of the present approach is to identify redundant and synergetic circuits rather than quantifying the information flow in the system.

REFERENCES

[1] E. Pereda, R. Quiroga, J. Bhattacharya, Progress in Neurobiology **77**, 1 (2005)

[2] C.W.J. Granger, Econometrica **37**, 424 (1969).

[3] D. Marinazzo, M. Pellicoro, and S. Stramaglia, Computational and Mathematical Methods in Medicine, Volume 2012 (2012), Article ID 303601.

[4] L. Angelini et al., Phys. Rev. **E 81**, 037201 (2010).

[5] V. Griffith and C. Koch, (2014). Quantifying synergistic mutual information, in Guided Self-Organization: Inception, Vol. 9, ed. M. Prokopenko (Berlin: Springer), 159?190.

[6] M. Harder, C. Salge and D. Polani, Phys. Rev. E **87**, 012130 (2013)

[7] J.T. Lizier, B. Flecker, P.L. Williams Artificial Life (ALIFE), 2013 IEEE Symposium on , pp.43,51, doi: 10.1109/ALIFE.2013.6602430 (2013)

[8] P. Wollstadt, U. Meyer, M. Wibral, A Graph Algorithmic Approach to Separate Direct from Indirect Neural Interactions, arXiv:1504.00156 [cs.IT].

[9] S. Stramaglia et al 2014 New J. Phys. 16 105003

[10] L Faes, D Kugiumtzis, A Montalto, G Nollo, D Marinazzo, Phys. Rev. E 2015; 91:032904.

[11] AJ Conger, Educational and Psychological Measurement April 1974 vol. 34 no. 1 35-46

[12] D. Chicharro, A. Ledberg Physical Review E, 86, 041901 (2012)

[13] K. Hlavackova-Schindler, M. Palus, M. Vejmelka, J. Bhattacharya, Physics Reports **441**, 1 (2007).
[14] D. Marinazzo, M. Pellicoro and S. Stramaglia, Phys. Rev. E **77**, 056215 (2008).
[15] V. Vapnik. The Nature of Statistical Learning Theory. Springer, N.Y., 1995.
[16] J. Shawe-Taylor and N. Cristianini, *Kernel Methods For Pattern Analysis*. (Cambridge University Press, London, 2004)
[17] D. Marinazzo, M. Pellicoro, S. Stramaglia, Phys. Rev. Lett. **100**, 144103 (2008).
[18] N. Ancona and S. Stramaglia, Neural Comput. **18**, 749 (2006).
[19] W.J. McGill, Multivariate information transmission Psychometrika 19, 97-116 (1954).
[20] E. Schneidman, W. Bialek, M.J. Berry, Journal of Neuroscience **23** 11539 (2003).
[21] S. Stramaglia, G. Wu, M. Pellicoro, D. Marinazzo Physical Review E, 86, 066211 (2012)
[22] N. Timme, W. Alford, B. Flecker, JM Beggs, J Comput Neurosci. 36(2):119-40. (2014)
[23] P. L. Williams and R. D. Beer, Nonnegative Decomposition of Multivariate Information, arXiv:1004.2515 [cs.IT].
[24] A. Barrett, An exploration of synergistic and redundant information sharing in static and dynamical Gaussian systems, arXiv:1411.2832 [cs.IT].
[25] N. Cufaro Petroni, Entropy 2014, 16(7), 4044-4059.
[26] Pearl, J (1988), Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann, San Mateo, CA.