Expanding the transfer entropy to identify information circuits in complex systems

S. Stramaglia,[1,2] Guo-Rong Wu,[3,4] M. Pellicoro,[1,2] and D. Marinazzo[3]
[1]*Istituto Nazionale di Fisica Nucleare, Sezione di Bari, Bari, Italy*
[2]*Dipartimento di Fisica, University of Bari, Bari, Italy*
[3]*Faculty of Psychology and Educational Sciences, Department of Data Analysis, Ghent University, Henri Dunantlaan 1, B-9000 Ghent, Belgium*
[4]*Key Laboratory for NeuroInformation of Ministry of Education, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China*
(Received 15 June 2012; published 20 December 2012)

We propose a formal expansion of the transfer entropy to put in evidence irreducible sets of variables which provide information for the future state of each assigned target. Multiplets characterized by a large contribution to the expansion are associated to the informational circuits present in the system, with an informational character which can be associated to the sign of the contribution. For the sake of computational complexity, we adopt the assumption of Gaussianity and use the corresponding exact formula for the conditional mutual information. We report the application of the proposed methodology on two electroencephalography (EEG) data sets.

DOI: 10.1103/PhysRevE.86.066211 PACS number(s): 05.45.Tp, 87.19.L−

I. INTRODUCTION

The inference of couplings between dynamical subsystems, from data is a topic of general interest. Transfer entropy [1], which is related to the concept of Granger causality [2], has been proposed to distinguish effectively driving and responding elements and to detect asymmetry in the interaction of subsystems. By appropriate conditioning of transition probabilities this quantity has been shown to be superior to the standard time delayed mutual information, which fails to distinguish information that is actually exchanged from shared information due to common history and input signals [3,4]. On the other hand, Granger formalized the notion that, if the prediction of one time series could be improved by incorporating the knowledge of past values of a second one, then the latter is said to have a *causal* influence on the former. Initially developed for econometric applications, Granger causality has gained popularity also in neuroscience (see, e.g., [5–9]). A discussion about the practical estimation of information theoretic indexes for signals of limited length can be found in [10]. Transfer entropy and Granger causality are equivalent in the case of Gaussian stochastic variables [11]: They measure the information flow between variables [12]. Recently it has been shown that the presence of redundant variables influences the estimate of the information flow from data, and that the maximization of the total causality is connected to the detection of groups of redundant variables [13].

In recent years, the information theoretic treatment of groups of correlated degrees of freedom has been used to reveal their functional roles as memory structures or those capable of processing information [14]. Information theory suggests quantities that reveal if a group of variables is mutually redundant or synergetic [15,16]. Most approaches for the identification of functional relations among nodes of complex networks rely on the statistics of motifs; subgraphs of k nodes that appear more abundantly than expected in randomized networks with the same number of nodes and degree of connectivity [17,18].

An interesting approach to identifing functional subgraphs in complex networks, relying on an exact expansion of the mutual information with a group of variables, has been presented in Ref. [19]. In this work we generalize these results to show a formal expansion of the transfer entropy which puts in evidence irreducible sets of variables which provide information for the future state of the target. Multiplets of variables characterized by an high value, unjustifiable by chance, will be associated to informational circuits present in the system. Additionally, in applications where linear models are sufficient to explain the phenomenology, we propose to use the exact formula for the conditioned mutual information among Gaussian variables so as to get a computationally efficient approach. An approximate procedure is also developed to find informational circuits of variables starting from a few variables of the multiplet by means of a greedy search. We illustrate the application of the proposed expansion to a toy model and two real electroencephalogram (EEG) data sets.

The paper is organized as follows. In the next section we describe our approach. In Sec. III we report the applications of the approach and describe our greedy search algorithm. In Sec. IV we draw our conclusions.

II. EXPANSION

We started describing the work in [19]. Given a stochastic variable X and a family of stochastic variables $\{Y_k\}_{k=1}^n$, the following expansion for the mutual information, analogous to a Taylor series, has been derived there:

$$\begin{aligned} &S(X|\{Y\}) - S(X) \\ &\quad = -I(X;\{Y\}) = \sum_i \frac{\Delta S(X)}{\Delta Y_i} + \sum_{i>j} \frac{\Delta^2 S(X)}{\Delta Y_i \Delta Y_j} \\ &\qquad + \cdots + \frac{\Delta^n S(X)}{\Delta Y_i, \ldots, \Delta Y_n}, \end{aligned} \tag{1}$$

where the variational operators are defined as

$$\frac{\Delta S(X)}{\Delta Y_i} = S(X|Y_i) - S(X) = -I(X;Y_i), \tag{2}$$

$$\frac{\Delta^2 S(X)}{\Delta Y_i \Delta Y_j} = -\frac{\Delta I(X;Y_i)}{\Delta Y_j} = I(X;Y_i) - I(X;Y_i|Y_j), \tag{3}$$

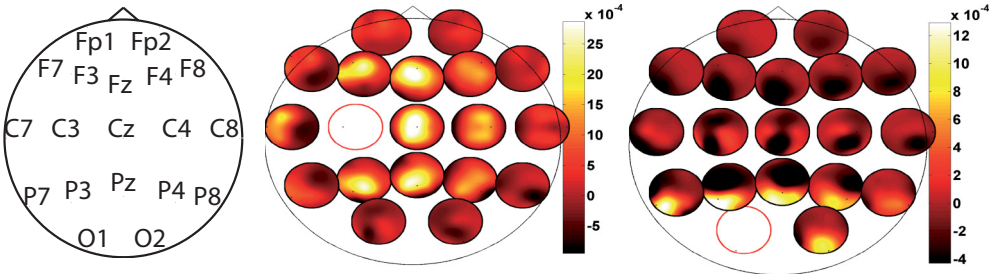

FIG. 1. (Color online) Left: Scalp map with electrode names. The instantaneous components Z^0_{ij} for two target electrodes, $C3$ at the center and $O1$ on the right. The target electrode is in white, and for each of the other electrodes i on the map, the value of Z^0_{ij} is displayed.

$$\frac{\Delta^3 S(X)}{\Delta Y_i \Delta Y_j \Delta Y_k} = I\left(X; Y_i | Y_k\right) - I\left(X; Y_i | Y_j, Y_k\right) - I\left(X; Y_i\right) + I\left(X; Y_i | Y_j\right), \tag{4}$$

and so on.

Now, let us consider an $n+1$ time series $\{x_\alpha(t)\}_{\alpha=0,\ldots,n}$. The lagged state vectors are denoted

$$Y_\alpha(t) = [x_\alpha(t-m), \ldots, x_\alpha(t-1)],$$

with m being the window length.

First we may use the expansion (1) to model the statistical dependencies among the x variables at equal times. We take x_0 as the target time series, and the first terms of the expansion are

$$W^0_i = -I\left(x_0; x_i\right) \tag{5}$$

for the first order

$$Z^0_{ij} = I(x_0; x_i) - I(x_0; x_i | x_j) \tag{6}$$

for the second order and so on. We note that

$$Z^0_{ij} = -\mathcal{I}(x_0; x_i; x_j),$$

where $\mathcal{I}(x_0; x_i; x_j)$ is the *interaction information*, a well-known information measure for sets of three variables [20]; it expresses the amount of information (redundancy or synergy) bound up in a set of variables, beyond that which is present in any subset of those variables. Unlike the mutual information, the interaction information can be either positive or negative. Common-cause structures lead to negative interaction information. As a typical example of positive interaction information one may consider the three variables of the following system: the output of an XOR gate with two independent random inputs (however, some difficulties may arise in the interpretation of the interaction information, see [21]). It follows that positive (negative) Z^0_{ij} corresponds to redundancy (synergy) among the three variables x_0, x_i, and x_j.

To go beyond equal time correlations, here we propose to consider the flow of information from multiplets of variables to a given target. Accordingly, we consider

$$S\left(x_0 | \{Y_k\}_{k=1}^n\right) - S(x_0) = -I\left(x_0; \{Y_k\}_{k=1}^n\right), \tag{7}$$

which measures to what extent all the remaining variables contribute to specifying the future state of x_0. This quantity can be expanded according to (1)

$$S\left(x_0 | \{Y_k\}_{k=1}^n\right) - S(x_0) = \sum_i \frac{\Delta S(x_0)}{\Delta Y_i} + \sum_{i>j} \frac{\Delta^2 S(x_0)}{\Delta Y_i \Delta Y_j} + \cdots + \frac{\Delta^n S(x_0)}{\Delta Y_i, \ldots, \Delta Y_n}. \tag{8}$$

A drawback of the expansion (7) is that it does not remove shared information due to common history and input signals; therefore we choose to condition it on the past of x_0 (i.e., Y_0). To this aim we introduce the conditioning operator $\mathcal{C}_{Y_0}$

$$\mathcal{C}_{Y_0} S(X) = S(X | Y_0),$$

and observe that $\mathcal{C}_{Y_0}$ and the variational operators (2) commute. It follows that we can condition the expansion (8) term by term, thus obtaining

$$S\left(x_0 | \{Y_k\}_{k=1}^n, Y_0\right) - S(x_0 | Y_0) = -I\left(x_0; \{Y\}_{k=1}^n | Y_0\right) = \sum_i \frac{\Delta S(x_0|Y_0)}{\Delta Y_i} + \sum_{i>j} \frac{\Delta^2 S(x_0|Y_0)}{\Delta Y_i \Delta Y_j} + \cdots + \frac{\Delta^n S(x_0|Y_0)}{\Delta Y_i, \ldots, \Delta Y_n}. \tag{9}$$

The first order terms in the expansion are given by

$$A^0_i = \frac{\Delta S(x_0|Y_0)}{\Delta Y_i} = -I\left(x_0; Y_i | Y_0\right), \tag{10}$$

and coincide with the bivariate transfer entropies $i \to 0$ (times -1). The second order terms are

$$B^0_{ij} = I(x_0; Y_i | Y_0) - I(x_0; Y_i | Y_j, Y_0), \tag{11}$$

and may be seen as a generalization of the interaction information $\mathcal{I}$; hence a positive (negative) B^0_{ij} corresponds to a redundant (synergetic) flow of information $\{i, j\} \to 0$. The typical examples of synergy and redundancy, in the present framework of network analysis, are the same as in the static case, plus a delay for the flow of information towards the target. The third order terms are

$$C^0_{ijk} = I(x_0; Y_i | Y_j, Y_0) + I\left(x_0; Y_i | Y_k, Y_0\right) - I\left(x_0; Y_i | Y_0\right) - I(x_0; Y_i | Y_j, Y_k, Y_0), \tag{12}$$

and so on.

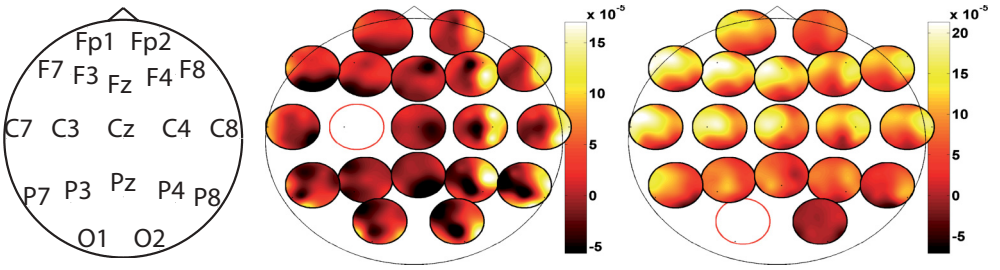

FIG. 2. (Color online) Left: Scalp map with electrode names. The lagged components B^0_{ij} for two target electrodes, $C3$ at the center and $O1$ on the right. The target electrode is in white, and for each of the other electrodes i on the map, the value of B^0_{ij} is displayed.

The generic term in the expansion (9)

$$\Omega_k = \frac{\Delta^k S(x_0|Y_0)}{\Delta Y_i, \ldots, \Delta Y_k} \tag{13}$$

is symmetrical under permutations of the Y_i and, remarkably, statistical independence among any of the Y_i results in a vanishing contribution to that order. Therefore each nonvanishing accounts for an irreducible set of variables providing information for the specification of the target: the search for informational multiplets is thus equivalent to the search for terms (13) which are significantly different from zero. Another property of Eq. (9) is that the sign of each term is connected to the informational character of the corresponding set of variables, see [19].

For practical applications, a reliable estimate of conditional mutual information from data should be used. Nonparametric methods are recommendable when nonlinear effects are relevant. However, a conspicuous amount of phenomenology in the brain can be explained by linear models. Therefore, for the sake of computational load, in this work we adopt the assumption of Gaussianity and use the exact expression that holds in this case [11], which reads as follows. Given multivariate Gaussian random variables X, W, and Z, the conditioned mutual information is

$$I(X;W|Z) = \frac{1}{2} \ln \frac{|\Sigma(X|Z)|}{|\Sigma(X|W \oplus Z)|}, \tag{14}$$

where $|\cdot|$ denotes the determinant, and the partial covariance matrix is defined as

$$\Sigma(X|Z) = \Sigma(X) - \Sigma(X,Z)\Sigma(Z)^{-1}\Sigma(X,Z)^{\top}, \tag{15}$$

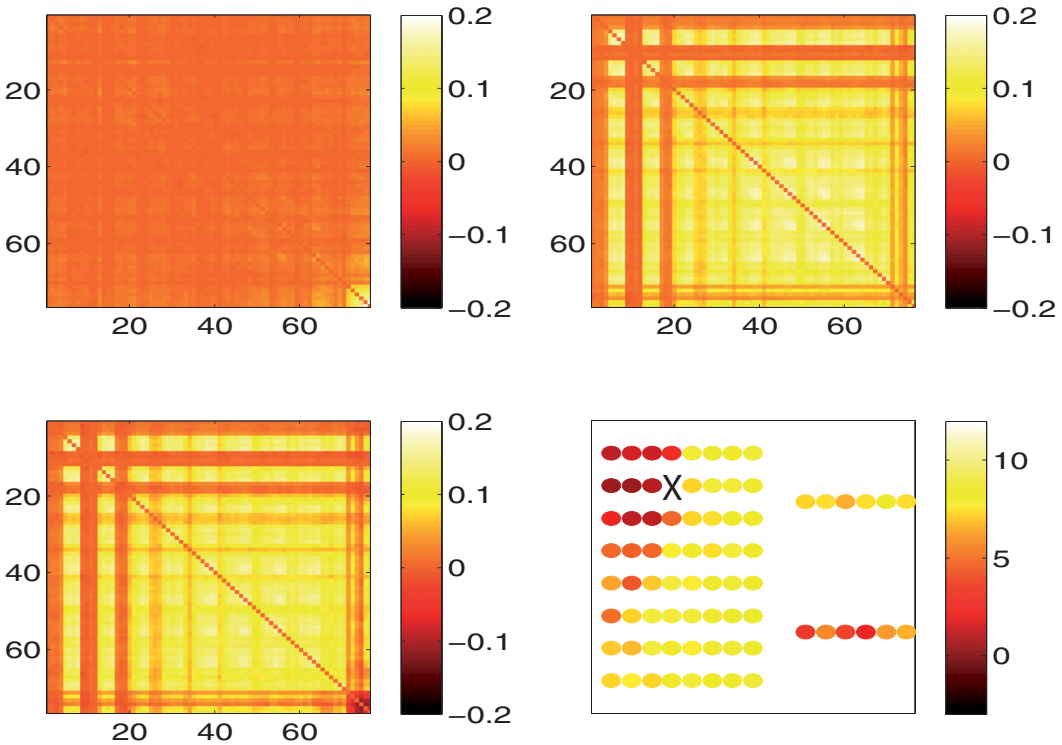

FIG. 3. (Color online) The lagged second order terms B^0_{ij} for a cortical electrode right before (top left) and during (top right) the clinical onset of a seizure, and their difference (bottom left). Bottom right: A map of the contacts (cortical grid on the left and two intracranial electrodes on the right) in which, for a cortical target (indicated with an X), the sum of B^0_{ij} over the second dimension is reported.

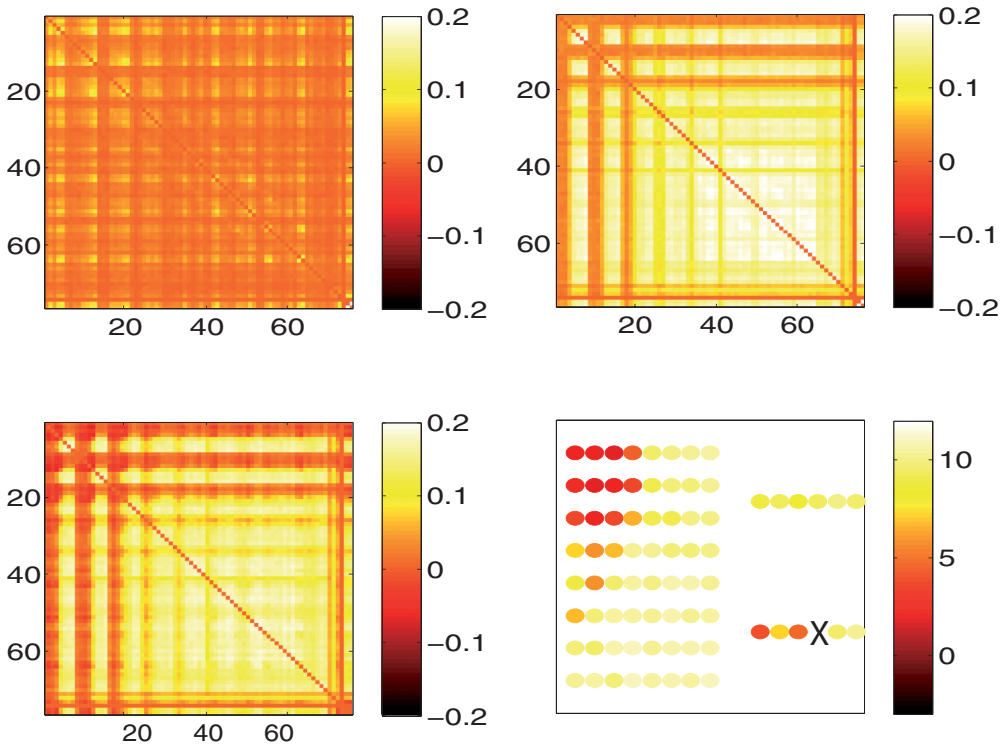

FIG. 4. (Color online) The lagged second order terms B^0_{ij} for a cortical electrode right before (top left) and during (top right) the clinical onset of a seizure, and their difference (bottom left). Bottom right: A map of the contacts (cortical grid on the left and two intracranial electrodes on the right) in which, for a intracranial target (indicated with an X), the sum of B^0_{ij} over the second dimension is reported.

in terms of the covariance matrix $\Sigma(X)$ and the cross covariance matrix $\Sigma(X,Z)$; the definition of $\Sigma(X|W \oplus Z)$ is analogous.

The statistical significance of Eq. (13) can be assessed by observing that it is the sum of terms like (14) which, under the null hypothesis $I(X;W|Z) = 0$, have a χ^2 distribution. Alternatively, statistical testing may be done using surrogate data obtained by the random temporal shuffling of the target vector x_0; this strategy is the one we use in this work.

III. APPLICATIONS

A. Second order terms

In this section we show the application of the proposed expansion, truncated at the second order. To this aim we turn to real EEG data, the window length m being fixed by cross validation. First we consider recordings obtained at rest from ten healthy subjects. During the experiment, which lasted for 15 min, the subjects were instructed to relax and keep their eyes closed. To avoid drowsiness, every minute the subjects were asked to open their eyes for 5 s. EEG was measured with a standard 10–20 system consisting of 19 channels whose names and locations are reported in Figs. 1 and 2. Data were sampled at 256 Hz and analyzed using the linked mastoids reference [22], and are available from Ref. [23].

For each subject we consider several epochs of 4 s in which the subjects kept their eyes closed. For each epoch we compute the second order terms at equal times Z^0_{ij} and the lagged ones B^0_{ij}; then we average the results over epochs. To visualize these results, for each target electrode we plot on a topographic scalp map the pairs of electrodes which are redundant or synergetic with respect to it. Both quantities are distributed with a clear pattern across the scalp. Interactions at equal times are one order of magnitude higher than the lagged interactions, and are dominated by the effect of spatial proximity, see Fig. 1. On the other hand, B^0_{ij} show a richer dynamics, such as interhemispheric communications and predominance redundancy to and from the occipital channels, see Fig. 2, reflecting the prominence of the occipital rhythms when the subjects rest with their eyes closed.

As another example we consider intracranial EEG recordings from a patient with drug-resistant epilepsy and which has thus been implanted with an array of 8×8 cortical electrodes and two depth electrodes with six contacts. The data are available at Ref. [24] and described in Ref. [25]. Data were sampled at 400 Hz. For each seizure data are recorded from the preictal period, the 10 s preceding the clinical onset of the seizure, and the ictal period, 10 s from the clinical onset of the seizure. We analyze data corresponding to eight seizures and average the corresponding results.

For each electrode we compute the lagged influences B^0_{ij}, obtaining for each electrode the pair of other electrodes with a redundant or synergetic contribution to its future. The patient has a putative epileptic focus in a deep hippocampal region, with the seizure that then spreads to the cortical areas. In Fig. 3 we report the values of coefficients B taking as the target a cortical electrode located on the putative cortical focus: We report the values of B^0_{ij} corresponding to all the couples of the electrodes, as well as their sum over electrode j. It is

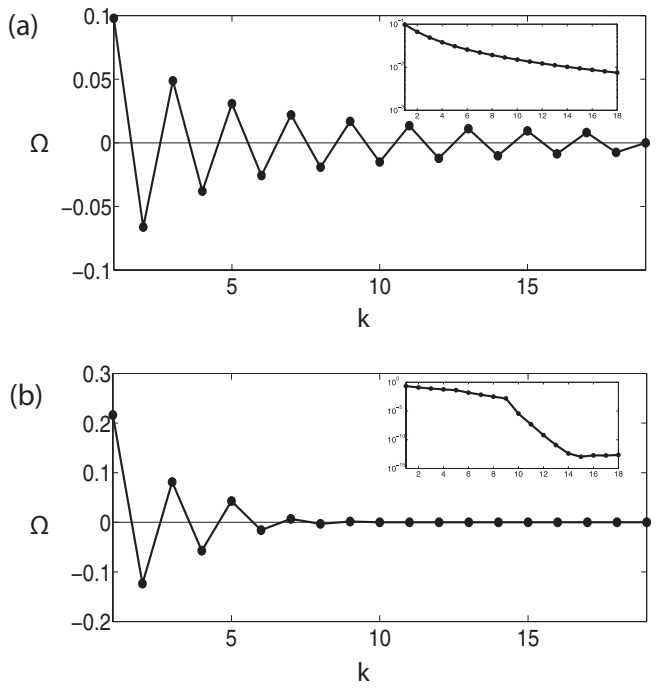

FIG. 5. Ω_k as a function of the multiplet size k for a model in which one variable is influenced by all the other variables or by part of them. In (a) $m = 20$ and $M = 0$: all the 20 variables influence the target with unitary weight. In (b) $m = 10$ and $M = 10$; the weights b_α are [1.75 1.75 1 1 1 1.5.5.5 0.5]. The insets show the logarithm of the absolute value of Ω_k. The first point $k = 1$, in both plots, represents the initial pair of variables chosen as the seed (i.e., {1,2}). The other parameters are, in both cases, $a,\sigma,\sigma_1,\sigma_2 = 0.5$.

clear how the redundancy increases during the seizure. On the other hand, for sensors from 70 to 76, corresponding to a depth electrode, the redundancy is higher in the preictal period, reflecting the fact that the seizure is already active in its primary focus even if it is not yet clinically observable. The values of B corresponding to this electrode are reported in Fig. 4.

B. Greedy search of multiplets

Given a target variable, the time required for the exhaustive search of all the subsets of variables, with a statistically significant information flow (13), is exponential in the size of the system. It follows that the exact search for large multiplets is computationally unfeasible, hence we adopt the following approximate strategy. We start from a pair of variables with the nonvanishing second order term B with regard to the given target. We consider these two variables as a *seed*, and aggregate other variables to them so as to construct a multiplet. The third variable of the subset is selected among the remaining ones as those that, jointly with the previously chosen variable, maximize the modulus $|C|$ of the corresponding third order term. Then, one keeps adding the rest of the variables by iterating this procedure. Calling Z_{k-1} the selected set of $k-1$ variables, the set Z_k is obtained adding, to Z_{k-1}, the variable, among the remaining ones, with the greatest modulus of Ω_k. These iterations stop when Ω_k, corresponding to Z_k, is not significantly different from zero [26]; Z_{k-1} is then recognized as the multiplet originated by the initial pair of variables chosen as the seed.

We apply this strategy to the following toy model

$$\begin{aligned} x_0(t) &= a\ \eta(t-1) + \sigma\xi_0(t), \\ x_\alpha(t) &= b_\alpha\ \eta(t) + \sigma_1\xi_\alpha(t), \quad \alpha = 1,\ldots,m, \\ x_\beta(t) &= \sigma_2\xi_\beta(t), \quad \beta = m+1,\ldots,m+M, \end{aligned} \tag{16}$$

where ξ and η are independent and identically distributed (i.i.d.) unit variance Gaussian variables. In this model the target x_0 is influenced by the process η; variables x_α, $\alpha = 1,\ldots,m$, are a mixture of η and noise ξ, while the remaining M variables are pure noise. Estimates of Ω_k are based on time series, generated from Eq. (16) and 1000 samples long. The results are displayed in Fig. 5. First we consider the case $m = 20$ and $M = 0$, with all the 20 variables driving the target with equal couplings b_α; in Fig. 5(a) we depict the term Ω_k corresponding to the kth iteration of the greedy search. We note that Ω_k has alternating sign and its modulus decreases with k. In Fig. 5(b) we consider another situation, with $m = 10$ and $M = 10$, the ten nonzero couplings b_α being nonuniform. Ω_k still shows alternating sign, and Ω_k vanishes for $k > 9$; hence the multiplet of ten variables is correctly identified. The order of selection is

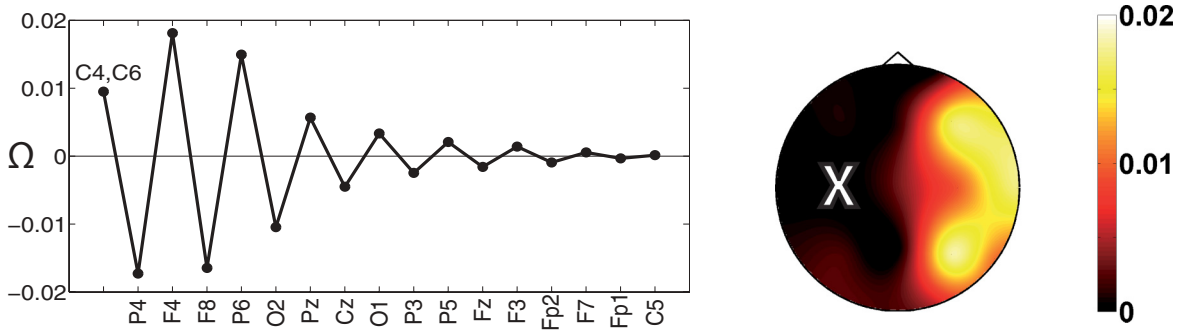

FIG. 6. (Color online) Informative contributions to the target electrode $C3$. Left: Information contributing from the resulting multiplet when time series from a given electrode are added to the existing multiplet, starting from the pair ($C4$,$C6$) which is the one which shares the most of information on the future of the target time series. Channels $P4$, $F4$, $F8$, $P6$, $O2$, Pz, and Cz are recognized as belonging to the same multiplet as $C4$ and $C6$, while including $O2$ leads to a Ω_k which is not significantly different from zero. Right: The absolute value of these contributions plotted on a scalp map, with the target indicated by an X.

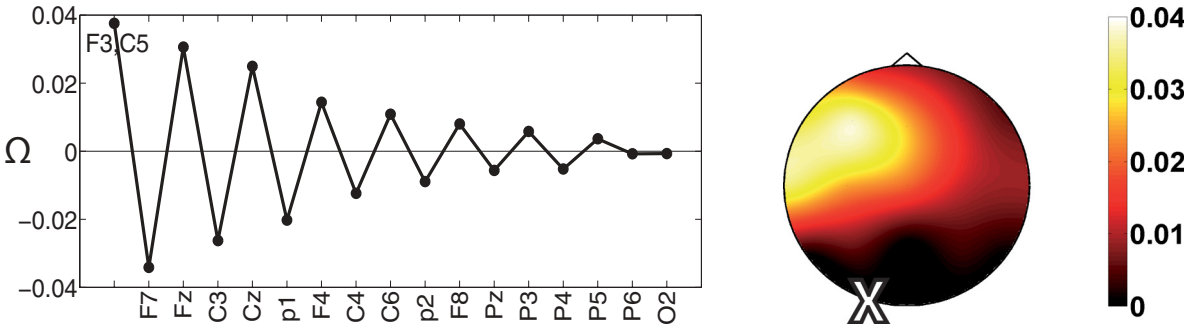

FIG. 7. (Color online) Informative contributions to the target electrode $O1$. Left: Information contributing from the resulting multiplet when time series from a given electrode are added to the existing multiplet, starting from the pair $(F3,C5)$ which is the one which shares the most of information on the future of the target time series. Nine channels ($F7$, Fz, $C3$, Cz, $Fp1$, $F4$, $C4$, $C6$, and $Fp2$) are recognized as belonging to the same multiplet, for the remaining variables Ω_k are not significantly different from zero. Right: The absolute value of these contributions plotted on a scalp map, with the target indicated by an X.

related to the strength of the coupling: variables with stronger coupling are selected first.

In Fig. 6 we consider again the EEG data from healthy subjects with closed eyes [23], and apply the greedy search taking $C3$ as the target and $\{C4,C6\}$ as the seed. We find a subset of nine variables influencing the target. The fact that the sign of Ω_k is alternating, as in the previous model, suggests that the channels in this set correspond to a single source which is responsible for the interhemispheric communication towards the target electrode $C3$. In Fig. 7 we take $O1$ as the target and $\{F3,C5\}$ as the seed. A subset of 11 variables is found which describes the information flow from the frontal to the occipital cortex.

IV. CONCLUSION

Summarizing, we have proposed to describe the flow of information in a system by means of multiplets of variables which send information to each assigned target node. We used a recently proposed expansion of the mutual information between a stochastic variable and a set of other variables to measure the character and the strength of multiplets of variables. Indeed, terms of the proposed expansion put in evidence irreducible sets of variables which provide information for the future state of the target channel. The sign of the contributions is related to their informational character; for the second order terms, synergy and redundancy correspond to negative and positive sign, respectively. For higher orders, we have shown that groups of variables, related to the same source of information, lead to contributions with alternating signs as the number of variables is increased. A decomposition with similarities to the present work has been reported in [27], where for multiple sources the distinction between unique, redundant, and synergistic transfer has been proposed; in Ref. [28] the inference of an effective network structure, given a multivariate time series, using incrementally conditioned transfer entropy measurements has been discussed. The main purpose of this paper is to introduce an information based decomposition, and we did that in a framework unifying Granger causality and transfer entropy, thus using a formula which is exact for linear models. In cases in which a nonlinear model is required, the entropy has to be computed, requiring a high enough number of time points for statistical validation; nonetheless the expansion that we proposed remains valid and exact in both cases.

We have reported the results of the applications to two EEG examples. The first data set is from *resting brains* and we found signatures of interhemispherical communications and the frontal to occipital flow of information. Concerning a data set from an epileptic subject, our analysis puts in evidence that the seizure is already active, close to the primary lesion, before it is clinically observable.

[1] T. Schreiber, Phys. Rev. Lett. **85**, 461 (2000).
[2] C. W. J. Granger, Econometrica **37**, 424 (1969).
[3] M. Staniek and K. Lehnertz, Phys. Rev. Lett. **100**, 158101 (2008).
[4] M. Lindner, R. Vicente, V. Priesemann, and M. Wibral, BMC Neuroscience **12**, 119 (2011).
[5] K. J. Blinowska, R. Kus, and M. Kaminski, Phys. Rev. E **70**, 050902(R) (2004).
[6] D. A. Smirnov and B. P. Bezruchko, Phys. Rev. E **79**, 046204 (2009).
[7] M. Dhamala, G. Rangarajan, and M. Ding, Phys. Rev. Lett. **100**, 018701 (2008).
[8] D. Marinazzo, M. Pellicoro, and S. Stramaglia, Phys. Rev. Lett. **100**, 144103 (2008).
[9] L. Faes, A. Porta, and G. Nollo, Phys. Rev. E **78**, 026201 (2008).
[10] A. Porta *et al.*, Methods of Information in Medicine **49**, 506 (2010).
[11] L. Barnett, A. B. Barrett, and A. K. Seth, Phys. Rev. Lett. **103**, 238701 (2009).
[12] K. Hlavackova-Schindler, M. Palus, M. Vejmelka, and J. Bhattacharya, Phys. Rep. **441**, 1 (2007).
[13] L. Angelini, M. de Tommaso, D. Marinazzo, L. Nitti, M. Pellicoro, and S. Stramaglia, Phys. Rev. E **81**, 037201 (2010).
[14] A. Borst and F. E. Theunissen, Nat. Neurosci. **2**, 947 (1999).

[15] E. Schneidman, W. Bialek, and M. J. Berry II, J. Neuroscience **23**, 11539 (2003).
[16] L. M. A. Bettencourt, G. J. Stephens, M. I. Ham, and G. W. Gross, Phys. Rev. E **75**, 021915 (2007).
[17] R. Milo *et al.*, Science **298**, 824 (2002).
[18] E. Yeger-Lotem *et al.*, Proc. Natl. Acad. Sci. USA **101**, 5934 (2004).
[19] L. M. A. Bettencourt, V. Gintautas, and M. I. Ham, Phys. Rev. Lett. **100**, 238701 (2008).
[20] W. J. McGill, Psychometrika **19**, 97 (1954).
[21] A. J. Bell, in *Proceedings of ICA 2003*, Nara, Japan, April 2003.
[22] G. Nolte, A. Ziehe, V. V. Nikulin, A. Schlögl, N. Krämer, T. Brismar, and K. R. Müller, Phys. Rev. Lett. **100**, 234101 (2008).
[23] http://clopinet.com/causality/data/nolte/.
[24] http://math.bu.edu/people/kolaczyk/datasets.html.
[25] M. A. Kramer, E. D. Kolaczyk, and H. E. Kirsch, Epilepsy Research **79**, 173 (2008).
[26] The Bonferroni correction for multiple comparisons has been applied at each iteration; it consists in recognizing Ω_k as significantly different from zero when the corresponding probability, under the null hypothesis, is less than $0.05/\ell$, where ℓ is the total number of comparisons for that iteration.
[27] P. L. Williams and R. D. Beer, arXiv:1102.1507.
[28] J. T. Lizier and M. Rubinov, "Multivariate construction of effective computational networks from observational data," preprint MPI MIS Preprint 25/2012.