# Classification-Based Causality Detection in Time Series

Danilo Benozzo[1,2(✉)], Emanuele Olivetti[1,2], and Paolo Avesani[1,2]

[1] NeuroInformatics Laboratory (NILab), Bruno Kessler Foundation, Trento, Italy
[2] Center for Mind and Brain Sciences (CIMeC), University of Trento, Trento, Italy
{benozzo,olivetti,avesani}@fbk.eu

**Abstract.** Brain effective connectivity aims to detect causal interactions between distinct brain units and it can be studied through the analysis of magneto/electroencephalography (M/EEG) signals. Methods to evaluate effective connectivity belong to the large body of literature related to detecting causal interactions between multivariate autoregressive (MAR) data, a field of signal processing. Here, we reformulate the problem of causality detection as a supervised learning task and we propose a classification-based approach for it. Our solution takes advantage of the MAR model by generating a labeled data set that contains trials of multivariate signals for each possible configuration of causal interactions. Through the definition of a proper feature space, a classifier is trained to identify the causality structure within each trial. As evidence of the efficacy of the proposed method, we report both the cross-validated results and the details of our submission to the causality detection competition of Biomag2014, where the method reached the 2nd place.

## 1 Introduction

A main part of neuroscience research concerns brain connectivity and aims to investigate the pattern of interactions between distinct units within the brain [10]. The concept of brain units is strongly related to the level of the adopted scale. Thus, brain connectivity can be studied from the microscopic level of single synaptic connections to the macroscopic level of brain regions. Depending on the type of interactions that we focus on, the topic of brain connectivity is divided into structural, functional and effective connectivity. In the first case the connectivity patterns are referred to anatomical links i.e. neural pathways, in the second case to the statistical dependences between brain activity in different units and in the last one to the causal interactions between them [15].

In particular, effective connectivity provides information about the direct influence that one or more units exert over another and aims to establish causal interactions among them [7]. To achieve this goal the usefulness of brain signals measured by magneto/electroencephalography (M/EEG) has been largely shown [3]. In fact, M/EEG record high temporal resolution signals that directly measure the brain activity. A large body of work was developed about methods to quantify the effective connectivity, mainly in the field of signal processing


I. Rish et al. (Eds.): MLINI 2014, LNAI 9444, pp. 85–93, 2016.
DOI: 10.1007/978-3-319-45174-9_9

where it is known as the problem of *inferring causality among time series*. An overview of the literature is provided below.

A first distinction that can be made in the available methods for causality detection, is between linear and nonlinear methods. Linear approaches are largely used both in time and frequency domain. An example of time domain technique is the Granger causality index. Granger causality is one of the most widespread measure to estimate the direction of causal influence in time series and its basic assumption, that a cause has to precede its effect, has been adopted in many other methods [8]. More precisely, if one or more time series $x_0(t), \ldots, x_k(t)$ are causing the time series $y(t)$, then a future value of $y(t)$ is better predicted by considering also the past values of $x_0(t), \ldots, x_k(t)$ than only those of $y(t)$. Most of the other time domain methods have the property that their multivariate extension is based on the partial auto- and cross-spectra estimation done by frequency-domain methods [16]. Thus, these latter have great adoption in causality assessment [5]. Examples are: the direct transfer function (DTF) [11,12], the direct coherence (DC) [2] and the partial direct coherence [1].

In situations in which the nonlinear component of the causal interaction is expected to be important, nonlinear multivariate methods are used [14]. A first attempt to deal with nonlinearity was done by the local application of linear multivariate methods in order to perform nonlinear prediction [6]. Further approaches are based on information theory [9], phase synchronization [4] and state space synchronization [13].

The intricate structure of interconnections, the enormous amount of dependence that brain units can exert over each other and, last but not least, the lack of a ground truth, make the assessment of the causal interactions a very complex problem. In general new methods to estimate causal interactions are assessed and validated on a limited set of signals and often by using data simulated by multivariate autoregressive (MAR) model. This is a common premise that allows researchers to analyse the performance of their techniques in the fully controlled environment of the MAR model. An example of the interest that has been addressed to causality in multivariate time series is the Biomag2014 Causality Challenge (Causal2014)[1]. The purpose of the contest was to estimate the direct causal interactions in a data set of simulated trials. One trial is meant as three multivariate time series, generated by a known MAR model, that is expected to simulate the behaviour of three neuronal populations.

In this paper we propose a new approach for the causality detection in time series by attacking the problem from a different prospective. Instead of developing a solution in the context of signal processing, as in the previous literature, we faced the problem from the machine learning point of view. Since modelling causal interactions with a MAR model is a common practice in the literature, we used the competition MAR model to create a set of trials for each possible causal configuration among the time series. Then a classifier was trained on those data in order to

[1] http://www.biomag2014.org/competition.shtml, see "**Challenge 2**: Causality Challenge".

discriminate between causal configurations. Finally, it was applied to the competition data set providing a solution that reached the second place of Causal2014.

## 2 Materials

The competition organizers provided the code of the MAR model together with the data set of which to estimate the direct causal interactions. Here, we will describe them both.

The final output of the MAR model is the multivariate time series $\mathbf{X} = \{X(t), t = 0, 1, \ldots, N-1\}, X(t) \in \mathbb{R}^{M\times 1}$ that is defined as the linear combination of two $M$-dimensional multivariate time series $\mathbf{X_s}$ and $\mathbf{X_n}$

$$\mathbf{X} = (1-\gamma)\mathbf{X_s} + \gamma\mathbf{X_n} \tag{1}$$

$\mathbf{X_s}$ carries the causal information, $\mathbf{X_n}$ represents the noise corruption and $\gamma \in [0, 1]$ tunes the signal-to-noise ratio. Each time point of $\mathbf{X_s}$ and $\mathbf{X_n}$ is computed by following the MAR model

$$\begin{aligned} X_s(t) &= \sum_{\tau=1}^{\min(P,t)} A_s^{(\tau)\top} X_s(t-\tau) + \varepsilon_s(t) \\ X_n(t) &= \sum_{\tau=1}^{\min(P,t)} A_n^{(\tau)\top} X_n(t-\tau) + \varepsilon_n(t) \end{aligned} \tag{2}$$

where $P$ is the order of the MAR model and represents the maximal time lag. $\varepsilon_s(t)$ and $\varepsilon_n(t)$ are realizations from a $M$-dimensional standard normal distribution. And $A_s^{(\tau)}, A_n^{(\tau)} \in \mathbb{R}^{M\times M}, \tau = 1, 2, \ldots, P$ are the coefficient matrices modelling the influence of the signal values at time $t-\tau$ on the current signal values, i.e. at time $t$. The coefficient matrices $\{A_s^{(\tau)}\}_\tau$ are involved in the process of causal-informative data generation. They are computed by randomly corrupting the non-zero elements of the $M \times M$ binary matrix $A$, called configuration matrix. In essence, the configuration matrix $A$ contains the causal structure that leads the MAR model. Specifically $A_{i,j} = 1$ means signal $i$ causes the signal $j$. On the other hand, coefficient matrices $A_n^{(\tau)}$ lead the noisy part of the signals and they are obtained by randomly generating $P$ diagonal matrices. The diagonality of these latter matrices is needed to avoid noise regressive dependencies across signals. After that, if both sets of matrices $A_s^{(\tau)}$ and $A_n^{(\tau)}$ fulfil the stationarity condition, each time point of $\mathbf{X_s}$ and $\mathbf{X_n}$ can be generated by Eq. 2.

In essence, given $P$, $\gamma$ and $A$, it is possible to generate $\mathbf{X}$ following Eqs. 1 and 2. The goal of the competition is to reconstruct $A$ given $\mathbf{X}$.

The competition data set was built by generating 1000 trials with the following parameter assignments: the number of time series in each trial is $M = 3$, the MAR model order is $P = 10$ and the time series length is $N = 6000$. The trial-specific parameters $\gamma$ and $A$ were randomly sampled from a standard uniform distribution for each trial and kept secret by the organizer of the competition. From now on, we will refer to the competition data set as $\mathbf{C}$.

## 3 Methods

The solution that we propose to the causality detection problem is based on a supervised approach. Indeed, this task to reconstruct $A$ from the data can be formulated as a classification problem. In a general setting, each trial is composed by $M$ time series and the final goal is to estimate its $M \times M$ binary configuration matrix $A$. Thus, there are $M(M-1)$ free binary parameters and $2^{M(M-1)}$ possible causal configurations[2].

Our supervised approach aims to train a classifier in order to discriminate between trials that were generated by different configuration matrices. And since we aim to predict $A$ given a trial, the classifier is going to treat $A$ as the trial's class label.

The training of the classifier is done on a new simulated data set generated by the MAR model described in Sect. 2. This new data set, that we will call $\mathbf{L}$, is meant to better represent the entire population of causal configurations that can be obtained by the adopted model. Therefore, $\mathbf{L}$ contains multiple trials for each of the possible $2^{M(M-1)}$ causal configurations.

Before the training, a proper feature space has to be defined in order to extract the causal structure that led the generation of the trial. And once $\mathbf{L}$ has been mapped on that feature space, a classifier $f$ is trained on it.

The classifier $f$ and the benefit that the feature space provides, are evaluated by estimating the discriminative power of $f$ through cross-validation. The discriminative power can be maximized by trying different types of classifiers, by tuning the related parameters and also by adjusting the feature space. Such way of proceeding does not introduce circularity because we are not using $\mathbf{C}$.

In the end, $f$ is applied to the competition data set $\mathbf{C}$ to predict the configuration matrix of each trial.

The feature space that we built, is strongly based on the concept of Granger causality. Indeed, it is a collection of measures that quantifies the ability to predict the value at a given time point of a certain time series (effect) from the past values of each possible subset of the $M$ time series in the trial (causes). The pair, made by causes and effect, is called causality scenario and, for $M$ time series, there are $\sum_{i=1}^{M} \binom{M}{i} M$ scenarios. In the case of the competition, where $M = 3$, the possible causality scenarios are 21, and they are summarized in Table 1, where $x_i(t), i = 0, 1, 2$, denotes each of the time series that defines a trial.

For each causality scenario, a plain linear regression problem was built by selecting, as dependent variable, a set of time points from the signal in the *effect* column. Each of these dependent variables has a regressor vector composed by the $P$ previous time points selected from the signals in the *causes* column. Table 2 shows how the regression problems were defined when $M = 3$, by specifying from which time series and time points, regressors and dependent variables are extracted. In the following, in order to simplify the notation, we will use $x_i^t$ instead of $x_i(t)$, $i = 0, 1, 2$ and $t \in \mathbf{T}, \mathbf{T} \subseteq \{P, P+1, \ldots, N-1\}$. Figure 1 explains

[2] The diagonal is not relevant since by definition the time series are autoregressive.

**Table 1.** The possible causality scenarios for three time series $x_i(t), i = 0, 1, 2$.

| Causes | Effect |
|---|---|
| $x_0(t)$ | $x_i(t)$ |
| $x_1(t)$ | $x_i(t)$ |
| $x_2(t)$ | $x_i(t)$ |
| $x_0(t), x_1(t)$ | $x_i(t)$ |
| $x_0(t), x_2(t)$ | $x_i(t)$ |
| $x_1(t), x_2(t)$ | $x_i(t)$ |
| $x_0(t), x_1(t), x_2(t)$ | $x_i(t)$ |

how, for the specific time point $t = 30$, the input of the regression problem is built in the case of the last causality scenario of the Table 2 with $i = 2$. More precisely, this example shows how the input of the regression problem is defined in order to quantify the plausibility of the causality scenario: "$x_0, x_1$ and $x_2$ are causing $x_2$".

**Table 2.** Description of how the 21 linear regression problems are defined for each trial. $x_i^t$, $i = 0, 1, 2$ and $t \in \mathbf{T}, \mathbf{T} \subseteq \{10, 11, \ldots, N-1\}$, are the three time series of a trial.

| Regressors (causes) | Dependent variable (effect) |
|---|---|
| $[x_0^{t-10}, \ldots, x_0^{t-1}]$ | $x_i^t$ |
| $[x_1^{t-10}, \ldots, x_1^{t-1}]$ | $x_i^t$ |
| $[x_2^{t-10}, \ldots, x_2^{t-1}]$ | $x_i^t$ |
| $[x_0^{t-10}, \ldots, x_0^{t-1}, x_1^{t-10}, \ldots, x_1^{t-1}]$ | $x_i^t$ |
| $[x_0^{t-10}, \ldots, x_0^{t-1}, x_2^{t-10}, \ldots, x_2^{t-1}]$ | $x_i^t$ |
| $[x_1^{t-10}, \ldots, x_1^{t-1}, x_2^{t-10}, \ldots, x_2^{t-1}]$ | $x_i^t$ |
| $[x_0^{t-10}, \ldots, x_0^{t-1}, x_1^{t-10}, \ldots, x_1^{t-1}, x_2^{t-10}, \ldots, x_2^{t-1}]$ | $x_i^t$ |

The regression problem of each causality scenario was cross-validated and its performance was quantified through multiple regression metrics, e.g. mean square error. The ensemble of the regression metrics of each causality scenario defined the initial feature vector of the trial. We then applied standard feature engineering techniques on the initial feature vector to enrich the feature space. The choice of using multiple regression metrics and in particular which ones including in the initial feature vectors, as well the choice of the feature engineering techniques, are driven by the goal to maximize the discriminative power of $f$. See Sect. 4 for further details.

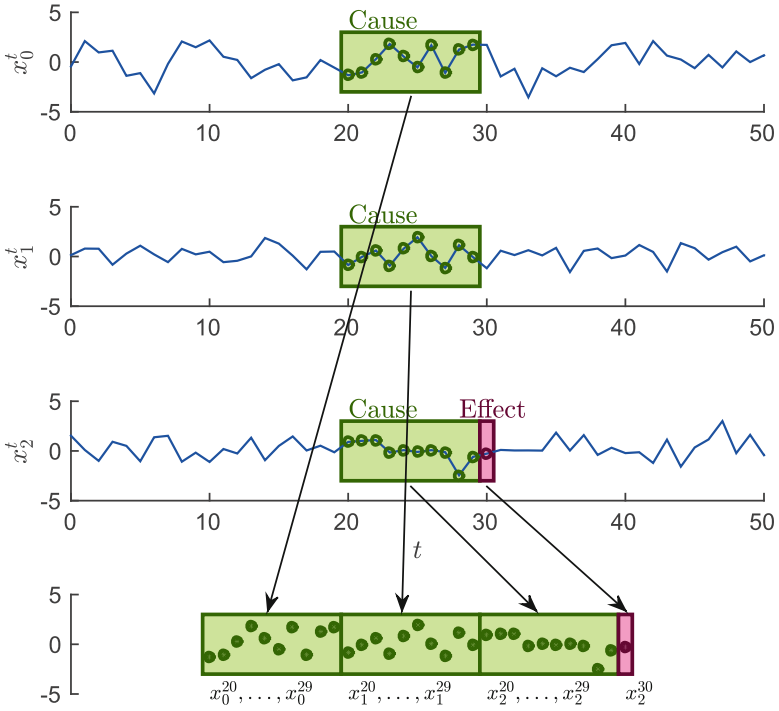


**Fig. 1.** Example of how the sample associated at the time point $t = 30$ is built in order to form the input of the last regression problem of the Table 2, for the case $i = 2$.

## 4 Experiments

In this section we present the technical details and results of the experiments that were conducted to evaluate the method described in Sect. 3. In particular, we show two different types of results. The first one is an estimate of the discriminative power of a classifier trained on the **L** data set and it provides a quantification of how well the defined feature space is able to express the causal structure behind a trial. The second result is the competition score obtained by our submission, which gives us insights into how our approach works compared to the ones adopted by the other participants.

Results are presented both in terms of confusion matrices and competition score. The competition score was defined in the following way. For each entry $\hat{A}_{ij}$, $i \neq j$, of each predicted $\hat{A}$, if $\hat{A}_{ij}$ was 1 and correct, then $+1$ point was given. If $\hat{A}_{ij}$ was 1 but incorrect, then $-3$ points were given. If $\hat{A}_{ij}$ was 0, then 0 points were given. In practice, false discoveries were punished three times more than what true discoveries were rewarded.

In order to take into account the strong false positive penalisation, we added a cost model to our predictions, by combining the probability of each of the 64

classes with the cost of predicting one class instead of another. Given $S_{ij}$ the cost of predicting $i$ when the true class was $j$, the optimal way to assign the class $l$ to a trial is

$$l = \underset{i=1,2,\ldots,64}{\operatorname{argmax}} \sum_{j=1}^{64} S_{i,j} p_j \tag{3}$$

where $p_j$ is the probability of class $j$ for the trial, as estimated by the classifier.

The new simulated labeled data set $\mathbf{L}$ was generated by keeping the same parameter initialization[3] of $\mathbf{C}$, except for the number of trials that was increased to 64000 in order to have 1000 trials for each class. Indeed, since $M = 3$ the amount of causal configurations is $2^6 = 64$. The regression metrics used to build the feature space are the mean square error and the coefficient of determination $r^2$. Both were included because we noticed a significant improvement in the cross-validated score, although, intuitively, they could seem redundant. We also added an estimate of the Granger causality coefficients[4] to the feature space.

As a final step we increased the number of features through standard feature engineering techniques by applying simple basis functions. This consisted in extracting the 2nd power, 3rd power and square root of the previously defined features, together with the pairwise product of all features. Adding extracted features was motivated both by the need to overcome the limitation of the adopted linear classifier and because they proved to be effective in increasing the cross-validated score.

Both the data sets, $\mathbf{L}$ and $\mathbf{C}$, were mapped to the proposed feature space. Then the performance of the logistic regression classifier[5], with $\ell_2$ regularisation, was evaluated on $\mathbf{L}$ through 5-folds cross-validation. In this way we quantified the discriminative capability of the proposed method.

Tables 3 and 4 show the cross-validated classification results in $\mathbf{L}$ by means of confusion matrices. In particular, Table 3 is related to the percentage of causal interactions predicted by assigning to each test trial the most probable class, i.e. $l = \operatorname{argmax} p_i$, and its accuracy is 81 %. In Table 4 the assignments are done by Eq. 3 according to the cost matrix, i.e. by penalizing the false positives,

**Table 3.** Confusion matrix computed by assigning to each test trial the most probable class.

| | | Predicted | |
|---|---|---|---|
| | | 1 | 0 |
| True | 1 | 79 % | 21 % |
| | 0 | 17 % | 83 % |

**Table 4.** Confusion matrix in which the test trial class labels are computed by Eq. 3.

| | | Predicted | |
|---|---|---|---|
| | | 1 | 0 |
| True | 1 | 56 % | 44 % |
| | 0 | 1 % | 99 % |

[3] Excluding the trial-specific parameter $\gamma$ which was randomly uniformly generated for each trail.

[4] http://nipy.org/nitime.

[5] http://scikit-learn.org.

and the related accuracy is 77.5 %. Through their comparison, the effect of $S$ is evident since in Table 4, false positives are strongly decreased, due to the score penalization, but to the detriment of some true positives.

Finally logistic regression was trained on **L** and tested on **C** to predict the configuration matrices of the competition. According to the number of trials in **C** and the assumptions of the generative process, the expected range of the score is $[-9000, 3000]$. The score of our submission was 1571, which reached the 2nd place in the final ranking of the Causal2014 competition.

## 5 Discussion, Conclusion and Future Works

In this paper, we proposed a new approach to detect causal interactions in multivariate time series. Specifically, we developed a classification-based causality detection method by defining a feature space based on the concept of Granger causality and by exploiting the MAR model as data generator. Aside from the novelty of the method itself, the interesting aspect of our solution is that it is a supervised method. Thus, it belongs to the machine learning field and not to the signal processing as traditionally was for that type of problem.

The proposed method was assessed by cross-validating the generated labeled data set and it provided promising results, as shown in Tables 3 and 4 by means of confusion matrices. Then, the submitted solution to the Causal2014 competition was computed by a classifier trained on the generated labeled data set. The achieved results, both in terms of cross-validation and competition ranking, are evidence that classification-based techniques are a feasible alternative to the signal processing methods for inferring causality between time series. And furthermore, that the defined feature space is able to well capture the causal structures among signals.

As an improvement of our approach, we are working on a tractable extension to the case of detecting causality in more than three time series.

## References

1. Baccalà, L.A., Sameshima, K.: Partial directed coherence: a new concept in neural structure determination. Biol. Cybern. **84**(6), 463–474 (2001). http://view.ncbi.nlm.nih.gov/pubmed/11417058
2. Baccalà, L.A., Sameshima, K., Ballester, G., Do Valle, A.C., Timo-Iaria, C.: Studying the interaction between brain structures via directed coherence and granger causality. Appl. Signal Process. **5**, 40–48 (1998). http://www.lcs.poli.usp.br/~baccala/pdc/papers/asp.pdf
3. Brookes, M.J., Woolrich, M.W., Barnes, G.R.: Measuring functional connectivity in MEG: a multivariate approach insensitive to linear source leakage. NeuroImage **63**(2), 910–920 (2012). http://view.ncbi.nlm.nih.gov/pubmed/22484306
4. Butler, S.R., Glass, A.: Asymmetries in the electroencephalogram associated with cerebral dominance. Electroencephalogr. Clin. Neurophysiol. **36**(5), 481–491 (1974). http://view.ncbi.nlm.nih.gov/pubmed/4135345

5. Faes, L., Erla, S., Nollo, G.: Measuring connectivity in linear multivariate processes: definitions, interpretation, and practical analysis. Comput. Math. Methods Med. **2012**, 1–18 (2012). doi:10.1155/2012/140513
6. Freiwald, W.A., Valdes, P., Bosch, J., Biscay, R., Jimenez, J.C., Rodriguez, L.M., Rodriguez, V., Kreiter, A.K., Singer, W.: Testing non-linearity and directedness of interactions between neural groups in the macaque inferotemporal cortex. J. Neurosci. Methods **94**(1), 105–119 (1999). http://view.ncbi.nlm.nih.gov/pubmed/10638819
7. Friston, K.J.: Functional and effective connectivity: a review. Brain Connectivity **1**(1), 13–36 (2011). doi:10.1089/brain.2011.0008
8. Granger, C.W.J.: Investigating causal relations by econometric models and cross-spectral methods. Econometrica **37**(3), 424–438 (1969). doi:10.2307/1912791
9. Hlavackovaschindler, K., Palus, M., Vejmelka, M., Bhattacharya, J.: Causality detection based on information-theoretic approaches in time series analysis. Phys. Rep. **441**(1), 1–46 (2007). doi:10.1016/j.physrep.2006.12.004
10. Horwitz, B.: The elusive concept of brain connectivity. NeuroImage **19**(2 Pt 1), 466–470 (2003). http://view.ncbi.nlm.nih.gov/pubmed/12814595
11. Kamiński, M., Ding, M., Truccolo, W.A., Bressler, S.L.: Evaluating causal relations in neural systems: granger causality, directed transfer function and statistical assessment of significance. Biol. Cybern. **85**(2), 145–157 (2001). http://view.ncbi.nlm.nih.gov/pubmed/11508777
12. Kaminski, M.J., Blinowska, K.J.: A new method of the description of the information flow in the brain structures. Biol. Cybern. **65**(3), 203–210 (1991). doi:10.1007/bf00198091
13. Papana, A., Kugiumtzis, D., Larsson, P.G.: Reducing the bias of causality measures. Phys. Rev. E **83**(3) (2011). http://dx.doi.org/10.1103/physreve.83.036207
14. Pereda, E., Quiroga, R.Q.Q., Bhattacharya, J.: Nonlinear multivariate analysis of neurophysiological signals. Prog. Neurobiol. **77**(1–2), 1–37 (2005). doi:10.1016/j.pneurobio.2005.10.003
15. Sakkalis, V.: Review of advanced techniques for the estimation of brain connectivity measured with EEG/MEG. Comput. Biol. Med. **41**(12), 1110–1117 (2011). doi:10.1016/j.compbiomed.2011.06.020
16. Winterhalder, M., Schelter, B., Hesse, W., Schwab, K., Leistritz, L., Klan, D., Bauer, R., Timmer, J., Witte, H.: Comparison of linear signal processing techniques to infer directed interactions in multivariate neural systems. Sig. Process. **85**(11), 2137–2160 (2005). doi:10.1016/j.sigpro.2005.07.011