Estimating conditional transfer entropy in time series using mutual information and nonlinear prediction

Baboukani, Payam Shahsavari; Graversen, Carina; Alickovic, Emina; Østergaard, Jan

Published in:
Entropy

DOI (link to publication from Publisher):
10.3390/e22101124

Publication date:
2020

Document Version
Publisher's PDF, also known as Version of record

entropy

MDPI

Article

Estimating Conditional Transfer Entropy in Time Series Using Mutual Information and Nonlinear Prediction

Payam Shahsavari Baboukani [1,*], Carina Graversen [2], Emina Alickovic [2,3] and Jan Østergaard [1]

[1] Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark; jo@es.aau.dk
[2] Eriksholm Research Centre, Oticon A/S, 3070 Snekkersten, Denmark; cagr@eriksholm.com (C.G.); eali@eriksholm.com (E.A.)
[3] Department of Electrical Engineering, Linköping University, 581 83 Linköping, Sweden
* Correspondence: pasba@es.aau.dk

Received: 14 August 2020; Accepted: 28 September 2020; Published: 3 October 2020

Abstract: We propose a new estimator to measure directed dependencies in time series. The dimensionality of data is first reduced using a new non-uniform embedding technique, where the variables are ranked according to a weighted sum of the amount of new information and improvement of the prediction accuracy provided by the variables. Then, using a greedy approach, the most informative subsets are selected in an iterative way. The algorithm terminates, when the highest ranked variable is not able to significantly improve the accuracy of the prediction as compared to that obtained using the existing selected subsets. In a simulation study, we compare our estimator to existing state-of-the-art methods at different data lengths and directed dependencies strengths. It is demonstrated that the proposed estimator has a significantly higher accuracy than that of existing methods, especially for the difficult case, where the data are highly correlated and coupled. Moreover, we show its false detection of directed dependencies due to instantaneous couplings effect is lower than that of existing measures. We also show applicability of the proposed estimator on real intracranial electroencephalography data.

Keywords: directed dependency; conditional transfer entropy; non-uniform embedding; nonlinear prediction; mutual information

1. Introduction

Real-world interconnected technological systems such as car traffic and distributed power grids as well as biological systems such as the human brain can be represented in terms of complex dynamical systems that contain subsystems. Characterizing the subsystems and their interdependencies can help understanding the overall system behavior on a local and global scale. For example, different regions of the brain such as the cortices can be considered as subsystems. An assessment of the interaction between the cortices may provide insights into how the brain functions [1]. In order to identify the interactions, several time series analyses methods ranging from information theoretical to signal processing approaches have been proposed in the literature [2–4]. In particular, the directional methods have gained increasing attention because, unlike symmetric measures such as mutual information [2] and phase synchronization [3,5], directional measures are generally able to assess the direction in addition to the strength of the interactions between subsystems [4,6–9].

A popular approach used in the literature to assess directed dependencies uses Wiener's definition, which is based on the concept of prediction [10]. According to the Wiener's definition, if the prediction of the future value of a time series X_t from its own past values can be improved by incorporating past values of another time series Y_t, then there are causal dependencies from Y_t to X_t [10]. Although the

term "causal" was used in Wiener's definition, it has been shown that measures quantifying the Wiener's definition over- or under-estimate the causal effect in certain cases [11,12]. In this paper, we use the term "directed dependencies" to refer to the property of time series or processes satisfying Wiener's definition.

Schreiber [4] formalized directed dependencies by using the concept of conditional mutual information (CMI) and proposed a new measure called transfer entropy (TE). TE does not depend on any model in its formulation, which makes this method able to assess both linear and nonlinear interactions [13]. Additionally, estimating TE by using the combination of data-efficient and model-free estimators like Kraskov–Stögbauer–Grassberger (KSG) [14], and uniform embedding state space reconstruction schemes [15,16] has increased the popularity of TE. TE has been used for quantifying directed dependencies between joint processes in neuro-physiological [15,16] and economical [17] applications.

As an example, assume that we are interested in measuring TE between processes which, for example, represent sensor measurement data from different regions of the brain, e.g., multi-channel electroencephalography (EEG) data. The recorded EEG data are spatially auto-correlated due to the phenomenon known as the volume conduction effect in neuro-physiological time series [18]. The spatial auto-correlation in such data can lead to overestimate in the estimated TE and eventually lead to false positives detection of TE. A possible approach to reduce such effect is to use a conditional version of TE [19,20], which is referred to as conditional transfer entropy (CTE).

It is preferred to condition out all other variables in the network to ensure that the obtained CTE values reflect the true directed dependencies from an individual source to the target. On the other hand, the more variables we include in the conditioning, the higher the dimension of the problem becomes and the less accurate CTE estimators are, since we only have access to a limited number of realizations. Considering the fact that we are interested in estimating directed dependencies and we need to condition out past variables related to the remaining variables, the dimension of the conditioning process increases even more and reliable estimation of CTE in multi-channel data (such as EEG data) by using the classical uniform embedding technique is limited by the so-called "curse of dimensionality" problem [13,21–23].

Non-uniform embedding (NUE) approaches reconstruct the past of the system with respect to a target variable by selecting the most relevant past and thereby decreases the dimensionality [13,19,22,24–26]. The information theoretical-based NUE algorithm proposed in [13] is a greedy strategy, which uses CMI for selecting the most informative candidates. The authors in [13] showed a significant improvement of NUE over uniform embedding. The author in [21] stated that, as the iteration of the NUE algorithm increases and more variables are selected, estimation of the higher dimensional CMI may become less accurate. The author in [21] then suggested to use a low-dimensional approximation (LA) of the CMI, and proposed a new NUE algorithm.

Adding more variables in the conditioning process decreases accuracy of the CTE estimator. The key problem is therefore how to decide whether we should include more variables, or terminate the algorithm. The existing NUE algorithms terminate if they fulfill a termination criterion defined by a bootstrap statistical-based test [13,21,23,26]. The bootstrap test is used to approximate a confidence bound (or a critical value) by which the NUE algorithm is terminated. A higher bootstrap size, up to a threshold, generally leads to better approximation of the confidence bound [27], which can further influence the accuracy of the NUE algorithms. A bootstrap size of at most 100 is generally used in the literature [13,19,21,22] due to computational complexity reasons. It has been shown that using an alternative to the bootstrap-based termination criterion can improve the accuracy and computational efficiency of the greedy algorithms [27,28]. For example, the Akaike information criterion (AIC) and kernel density estimation (KDE)-based regression were proposed in [27] as an alternative to bootstrap methods for input variable selection techniques

In the present study, inspired by [27] and originated from our initial work in [29], we propose an alternative approach to the bootstrap-based termination criterion used in the existing NUE algorithms.

Specifically, to aid in making the decision of whether to include a variable or terminate the algorithm, we propose to measure the relevance of the new candidate variable by assessing the effect of it on the accuracy of the nonlinear prediction of the target variable. The nonlinear prediction is based on nearest neighbor (NN)-based regression [30]. We show that it is also advantageous to use the nonlinear prediction strategy for selecting the pool of candidates in the first place. We then introduce a new NUE algorithm which uses a weighted combination of CMI and the accuracy of the nonlinear prediction for selection of candidates and present the new termination criterion for stopping the algorithm. Finally, we demonstrate that our proposed NUE procedure is more accurate than the existing NUE algorithms on both synthetic and real-world data.

The effect of instantaneous coupling (IC) on the NUE algorithms will also be investigated. IC can occur due to simultaneous (zero lag) information sharing like source mixing as a result of volume conduction in EEG signals [19,31] and may lead to spurious detection of TE or CTE.

The remainder of this paper is structured as follows. In Section 2, the necessary background on CTE and the existing NUE algorithms will be briefly reviewed. Then, the proposed termination criterion and NUE procedure will be introduced in Sections 3 and 4, respectively. This is followed by the description of our simulation study in Section 5, which is based on Henon maps and nonlinear autoregressive (AR) models. The results of applying the proposed NUE algorithm on real EEG data will be reported in Section 6. Section 7 will discuss the results. The same section will also conclude the paper.

2. Background

2.1. Conditional Transfer Entropy

Let us consider a complex system which consists of L interacting subsystems. We assume that we are interested in assessing the directed dependencies between subsystems $\mathcal{X}$ and $\mathcal{Y}$. Let stationary stochastic processes $X = (X_1, X_2, \ldots, X_N)$ and $Y = (Y_1, Y_2, \ldots, Y_N)$ describe the state visited by the subsystem $\mathcal{X}$ and $\mathcal{Y}$ over time, respectively. We denote $X_n \in \mathbb{R}$ and $Y_n \in \mathbb{R}$ as stochastic variables obtained by sampling the processes X and Y at the present time n, respectively. Furthermore, we denote the past of X up until X_{n-1} by a random vector $X_n^- = [X_{n-1}, X_{n-2}, \ldots]$. TE from X to Y is then defined as [4].

$$\mathrm{TE}(X \to Y) \triangleq I(Y_n; X_n^- \mid Y_n^-), \tag{1}$$

where $I(\,.\,;\,.\mid.\,)$ is CMI. However, in a complex network, it is not guaranteed that (1) only describes the directed dependencies from X to Y. For example, there could be a third process, say Z, through which shared information is mediated to X and Y. In this case, the shared information will lead to an increase in TE. To reduce the effect of common information being shared through other process, it has been suggested to use CTE [13,19]. Let us consider the $L = 6$ nodes network in Figure 1, where we are interested in assessing the directed dependencies from node $\mathcal{X}$ to $\mathcal{Y}$ and which is not due to indirect paths through the remaining nodes $\boldsymbol{\mathcal{Z}} = \{\mathcal{Z}^1, \mathcal{Z}^2, \mathcal{Z}^3, \mathcal{Z}^4\}$. We denote $Z^i = (Z_1^i, Z_2^i, \ldots, Z_N^i)$ as a stochastic process describes the state visited by $\mathcal{Z}^i$ and $\boldsymbol{Z} = [Z^1, Z^2, \ldots, Z^4]$ as a 4-variate stochastic process which describes state visited by $\boldsymbol{\mathcal{Z}}$ over time. CTE from an individual source X to the target Y excluding information from $\boldsymbol{Z}$ is then defined as

$$\mathrm{CTE}(X \to Y|\boldsymbol{Z}) \triangleq I(Y_n; X_n^- \mid Y_n^-, \boldsymbol{Z}_n^-), \tag{2}$$

where $\boldsymbol{Z}_n^- = [\boldsymbol{Z}_{n-1}, \boldsymbol{Z}_{n-2}, \ldots]$ denotes the past of up $\boldsymbol{Z}$ until but not including $\boldsymbol{Z}_n$.

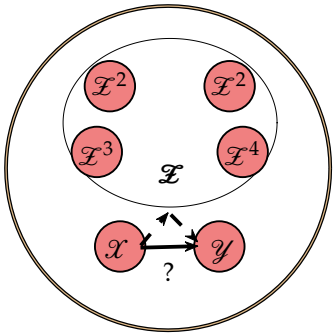


Figure 1. An example of $L = 6$ nodes network where indirect paths through the remaining channels $\boldsymbol{\mathcal{Z}}$ may cause a falsely (dashed line) detected directed dependency (solid line) from X to Y.

2.2. Existing Non-Uniform Embedding Algorithm

Prior to estimating CTE in (2), it is mandatory to approximate the possibly infinite-dimensional random vectors which represent the past of the processes. Let us denote the approximated past vector variable X_n^- by V_n^X. The same notation applies to V_n^Y and $V_n^{\mathbf{Z}}$. The basic idea behind reconstructing the past of the processes X, Y, and $\boldsymbol{Z}$ by assuming Y as the target process is to form a low dimensional embedding vector $\mathcal{S}$ comprising the most informative past variables about the present state of the target Y. Traditionally, the past of the system is reconstructed by using the uniform embedding scheme in which each component of $\mathcal{S}$ is approximated separately. For example, V_n^Y is approximated as $V_n^Y = [Y_{n-m}, Y_{n-2m}, \ldots, Y_{n-dm}]$, where m and d are the embedding delay and embedding dimension, respectively [13,15]. Then, the V_n^X and $V_n^{\mathbf{Z}}$ are estimated using the same approach and the final embedding vector $\mathcal{S} = [V_n^X, V_n^Y, V_n^{\mathbf{Z}}]$ is formed and utilized to estimated CTE in (2).

The uniform embedding scheme may lead to selection of redundant past variables and ignore relevant variables, as a result decrease the accuracy of the CTE estimation. This can limit applications in high dimensional data [13,21,23]. Alternatively, the NUE schemes try to select the most relevant and least redundant past variables and form a new embedding vector [13,21,23].

2.2.1. Bootstrap-Based Non-Uniform Embedding Algorithm

The NUE algorithm, as suggested in [13], can be described as follows:

1. Choose embedding delay d and embedding dimension m and construct the candidate set $\mathscr{C} = [X_{n-m}, \ldots, X_{n-md}, Y_{n-m}, ..Y_{n-md}, \mathbf{Z}_{n-m}, \ldots, \mathbf{Z}_{n-md}]$.
2. Initialize the algorithms by an empty set of the selected candidates $\mathcal{S}_n^0 = \emptyset$.
3. Run a forward search to find the most informative candidate among the candidate set $\mathscr{C}$. This can be achieved by quantifying the amount of information that each candidate W_n has about Y_n which is not provided by the selected candidates from the last iteration $\mathcal{S}_n^{k-1}$. To formalize this, at each iteration $k \geq 1$, select the candidate W_n^k, such that CMI between W_n and Y_n conditioned on $\mathcal{S}_n^{k-1}$ is maximized
$$W_n^k = \underset{W_n \in \mathscr{C} \setminus \mathcal{S}_n^{k-1}}{\operatorname{argmax}} \mathrm{I}\left(Y_n; W_n \mid \mathcal{S}_n^{k-1}\right), \tag{3}$$
where $\mathcal{S}_n^{k-1} = \bigcup_{i=0}^{k-1} W_n^i$ denotes the set of the selected candidates up till iteration $k-1$ and $\mathscr{C} \setminus \mathcal{S}_n^{k-1}$ denotes the remaining candidates in $\mathscr{C}$. We estimate the CMI given in (3) by using the KSG approach [13,14,32] in this study (cf. Appendix A.1).
4. Stop the iteration if the termination criterion is fulfilled and return $\mathcal{S}_n^{k-1}$ as the desired embedding vector.

The flow chart of the NUE algorithm is shown in Figure 2. After obtaining the embedding vector $\mathcal{S}_n^{k-1}$, CTE is estimated by using (2) in which case $[X_n^-, Y_n^-, \mathbf{Z}_n^-]$ is replaced by $\mathcal{S}_n^{k-1}$ and $[Y_n^-, \mathbf{Z}_n^-]$ is

replaced by $\mathcal{S}_n^{k-1}$ excluding the past of X_n. CTE is written as the sum/difference of four differential entropies and is estimated by using KSG approach (In this paper, we use the KSG approach to estimate CTE and CMI. The KSG estimator is designed to estimates differential entropies. Therefore, we assumed that variables used in this paper are continuous.) [13,14,32] (cf. Appendix A.2).

The existing NUE algorithm proposed in [13] utilizes a bootstrap-based termination criterion. The goal of the bootstrap test in the NUE algorithm is to estimate an upper bound on the CMI between independently selected candidate $\widehat{W_n^k}$ and the target variable $\widehat{Y_n}$ given $\mathcal{S}_n^{k-1}$, $I(\widehat{W_n^k}; \widehat{Y_n}|\mathcal{S}_n^{k-1})$. The estimation is accomplished by drawing 100 independent randomly shuffled realizations of Y_n and W_n^k, estimating the CMI between the randomized W_n^k and the randomized Y_n given the original $\mathcal{S}_n^{k-1}$, and then finding the 95th percentile I^{95} of the generated distribution. The obtained value I^{95} can be used as a critical value (at 5% confidence level) of $I(W_n^k; Y_n|\mathcal{S}_n^{k-1})$ so that if $I(W_n^k; Y_n|\mathcal{S}_n^{k-1}) > I^{95}$ then the candidate is included in the embedding vector and the algorithm continues to search for more candidates in iteration $k+1$. Otherwise, the termination criterion is fulfilled and the algorithm is ended and $\mathcal{S}_n^{k-1}$ is returned as the embedding vector.

2.2.2. Low-Dimensional Approximation-Based Non-Uniform Embedding Algorithm

The LA-based strategy follows the same flow chart as the existing NUE algorithm, shown in Figure 2, except that the CMI in (3) is substituted by its LA [21]. It is suggested in [21,23] that using LA of the CMI in (3) can increase the accuracy of estimation of the CMI and may outperform the accuracy of the NUE algorithm. The author in [21] proposed two LA alternatives to the CMI and concluded based on a simulation study that the LA of the CMI used in this study for the sake of comparison with our proposed NUE algorithm, outperforms another LA of the CMI. The criterion for finding the most informative candidates (i.e., Equation (3)) in the LA-based NUE algorithm is then given by

$$W_n^k = \underset{W_n \in \mathscr{C} \setminus \mathcal{S}_n^{k-1}}{\operatorname{argmax}} \left\{ I\left(W_n; Y_n\right) - \frac{2}{|\mathcal{S}_n^{k-1}|} \sum_{W_j \in \mathcal{S}_n^{k-1}} I\left(W_n; W_j\right) + \frac{2}{|\mathcal{S}_n^{k-1}|} \sum_{W_j \in \mathcal{S}_n^{k-1}} I\left(W_n; W_j | Y_n\right) \right\}, \tag{4}$$

where $|.|$ denotes the cardinality of a set. The mutual information and CMI are estimated using the KSG approach [13,14,32] (cf. Appendix A.1). The LA-based NUE algorithm also uses the bootstrap-based termination criterion. It should be noted that the LA of the CMI (i.e., Equation (4)) is used to estimate I^{95}.

2.2.3. Akaike Information Criterion-Based Non-Uniform Embedding Algorithm

AIC is used to assess the trade-off between accuracy and complexity of a model. It was adapted to quantify the trade-off between accuracy and complexity of a KDE-based prediction as an alternative to the bootstrap termination criterion in an input variable selection approach in [27,28]. AIC can also be adapted to act as a termination criterion for stopping the NUE algorithm. Therefore, an AIC-based NUE algorithm could follow the same flow chart as the existing NUE algorithm, shown in Figure 2, except that the the termination criterion will be replaced with the AIC-based termination criterion as is described below.

After selecting the most informative candidate W_n^k by using (3), the target variable Y_n is predicted given $\mathcal{U}_n^k = [W_n^k, \mathcal{S}_n^{k-1}] \in \mathbb{R}^k$, by using KDE-based prediction (cf. Appendix B). Let $y_n = (y_n(1), y_n(1), \ldots, y_n(N))$ be N realizations of Y_n. The AIC at iteration k is then given as:

$$AIC_k = N \log\left(\frac{1}{N}\sum_{i=1}^{N}(y_n(i) - \widehat{y}_n(i|\mathcal{U}_n^k))^2\right) + 2p, \tag{5}$$

where the ith realization of Y_n is denoted by $y_n(i)$ and $\widehat{y}_n(i|\mathcal{U}_n)$ is an estimator for the prediction of $y_n(i)$ given $\mathcal{U}_n$. The total number of realization of Y_n is N and p is the measure of complexity and for KDE-based regression, it is given as [27,33]:

$$p = \sum_{n=1}^{N} \frac{K_h(u_n^k(i), u_n^k(i))}{\sum_{j=1}^{N} K_h(u_n^k(i), u_n^k(j))}, \tag{6}$$

where $u_n^k(i)$ is ith realization of $\mathcal{U}_n^k$ (see Equation (7) for more details) and K_h is a Gaussian kernel with Mahalonobis distance and Gaussian reference kernel bandwidth (cf. Appendix B). During the NUE algorithm, if $AIC_k > AIC_{k-1}$ then, W_n^k is included in the embedding vector $\mathcal{S}_n^k$. Otherwise, the algorithm stops and $\mathcal{S}_n^{k-1}$ will be considered as the desired reconstructed past state of the system.

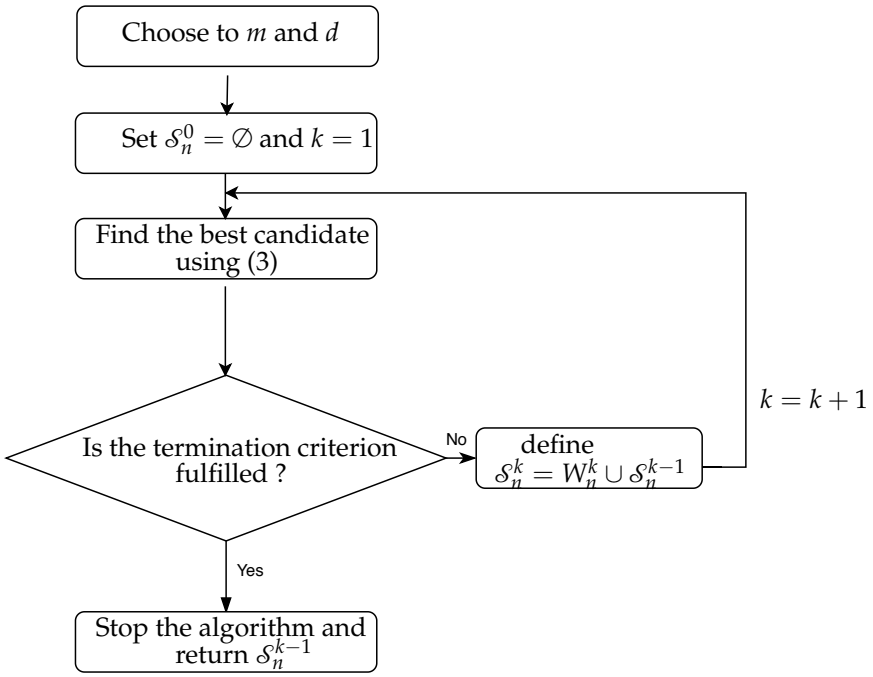


Figure 2. The flow chart of the NUE algorithm.

3. Proposed Termination Criterion

In this section, inspired by [27], we present a new termination criterion. Our proposed criterion is based on nonlinear prediction of the target variable, similar to the AIC approach. We modify NN-based regression [30] in order to be able to assess the effect of the selected candidate W_n^k on the accuracy of the prediction of Y_n.

We are interested in nonlinear prediction of the random variable Y_n given the random vector $\mathcal{U}_n^k = [W_n^k, \mathcal{S}_n^{k-1}] \in \mathbb{R}^k$. We denote the set of N realizations of W_n^k by $w_n^k = (w_n^k(1), w_n^k(2) \ldots, w_n^k(N))$ and set of N realizations of $\mathcal{U}_n^k$ be the $N \times k$ matrix

$$u_n^k = \begin{bmatrix} w_n^k(1) & w_n^{k-1}(1) & \cdots & w_n^1(1) \\ w_n^k(2) & w_n^{k-1}(2) & \cdots & w_n^1(2) \\ \vdots & \vdots & \ddots & \vdots \\ w_n^k(N) & w_n^{k-1}(N) & \cdots & w_n^1(N) \end{bmatrix}. \tag{7}$$

The ith row of the matrix u_n^k is a realization of the random vector $\mathcal{U}_n^k$. Let $t(i)$ be the set of indices of the T nearest neighbors of the ith realization of $\mathcal{U}_n^k$. For example, $t(i) = \{3, 7, 9\}$ shows that 3rd, 7th, and 9th rows of u_n^k are the $T = 3$ nearest neighbors of its ith row. The Euclidean distance is used as the distance metric for finding the nearest neighbors in the NN-based prediction. The prediction of the ith realization of Y_n (i.e., $y_n(i)$) given $\mathcal{U}_n^k$ is then calculated as an average of the realizations of Y_n whose indices are specified by the neighbor search in u_n^k. The average of the y-values having the same conditioned past is not an optimal estimator. However, it is simple, works well in the cases that we

have considered, and has also been used in previous work on non-conditional NN-based prediction. The $\widehat{y_n}(i|\mathcal{U}_n^k)$ is given as:

$$\widehat{y_n}(i|\mathcal{U}_n^k) \triangleq \frac{1}{T}\sum_{v\in t(i)} y_n(v). \tag{8}$$

For example, if $t(i) = \{3,7,9\}$ then $\widehat{y}(i|\mathcal{U}_n^k)$ is equal to the mean of $\{y_n(3), y_n(7), y_n(9)\}$. The residual $r(i|\mathcal{U}_n^k)$ can be computed as:

$$r(i|\mathcal{U}_n^k) = y_n(i) - \widehat{y_n}(i|\mathcal{U}_n^k). \tag{9}$$

In the NUE algorithm, the most informative candidate at iteration k, W_n^k, will be included in the embedding vector, if it significantly improves the accuracy of the prediction of the target variable Y_n given $\mathcal{U}_n^k$ compared to the prediction accuracy from the iteration $k-1$. The accuracy of the prediction can be calculated as the mean of the squared prediction residual (MSR):

$$\mathrm{MSR}(Y_n \mid \mathcal{U}_n^k) = \frac{1}{N}\sum_{i=1}^{N} r(i|\mathcal{U}_n^k)^2, \tag{10}$$

where the smaller MSR, the better prediction.

We first assume that the NUE algorithm contains at least $k = 2$ iterations and the termination test is performed from the second iteration. Accordingly, at each iteration $k \geq 2$, if $\mathrm{MSR}(Y_n|\mathcal{U}_n^{k-1}) - \mathrm{MSR}(Y_n|\mathcal{U}_n^k) > \gamma$, then W_n^k is included in $\mathcal{S}_n^k$ and the algorithm proceeds to search for more candidates at iteration $k+1$. Otherwise, the algorithm ends and $\mathcal{S}_n^{k-1}$ is considered as the desired embedding vector. The non-negative parameter γ defines how much the accuracy of the prediction needs to be improved before a variable is selected. Basically, by increasing the non-negative parameter γ which we have introduced, our proposed algorithm terminates sooner, and hence less variables are selected. In other words, the parameter γ controls the balance between true positives and true negatives, which can be useful, for example, in taking care of the confounder effects like IC. We will show in Section 5.2.2 that, by choosing a proper γ value, the number of true negatives significantly increases while the number of true positives does not decrease significantly in data in which the IC may cause spurious detection of directed dependencies.

4. Proposed Non-Uniform Embedding Algorithm

Our proposed NUE algorithm (referred to as MSR-based) uses a weighted combination of the CMI and MSR for selecting the most informative candidate and our proposed termination criterion for ending the algorithm. The details of the proposed NUE algorithm are as follows:

1. Choose γ, λ, embedding delay d and embedding dimension m and construct the candidate set $\mathscr{C} = [X_{n-m}, \ldots, X_{n-md}, Y_{n-m}, ..Y_{n-md}, \mathbf{Z}_{n-m}, \ldots, \mathbf{Z}_{n-md}]$.
2. Initialize by setting $\mathcal{S}_n^0 = \emptyset$,
3. At first iteration $k = 1$, find the first most relevant candidate W_n^1 by using a weighted combination of MSR and mutual information as:

$$W_n^1 = \underset{W_n\in\mathscr{C}}{\operatorname{argmax}} \left[(1-\lambda)\,\mathrm{I}\,(Y_n; W_n) - \lambda\,\mathrm{MSR}(Y_n \mid W_n)\right], \tag{11}$$

   where $0 \leq \lambda \leq 1$ is the weight. Then, set $\mathcal{S}_n^1 = [W_n^1]$.
4. At each iteration $k \geq 2$, run a search procedure to select the candidate which leads to the highest amount of new information about target variable Y_n and the best prediction of Y_n given the random vector $\mathcal{U}_n^k = [W_n, \mathcal{S}_n^{k-1}]$. It can be formalized by:

$$W_n^k = \underset{W_n\in\mathscr{C}\setminus\mathcal{S}_n^{k-1}}{\operatorname{argmax}} \left[(1-\lambda)\,\mathrm{I}\left(Y_n; W_n \mid \mathcal{S}_n^{k-1}\right) - \lambda\,\mathrm{MSR}\left(Y_n \mid \mathcal{U}_n^k\right)\right], \tag{12}$$

where $\mathcal{S}_n^{k-1} = \bigcup_{i=0}^{k-1} W_n^i$ denotes the set of selected candidates up till iteration $k-1$ and $\mathscr{C} \setminus \mathcal{S}_n^{k-1}$ refers to all elements of $\mathscr{C}$ except the elements of $\mathcal{S}_n^{k-1}$. Similar to the existing NUE algorithms, mutual information and CMI are estimated using the KSG approach [13,14,32] (cf. Appendix A.1).

5. Include the candidate W_n^k in the embedding vector $\mathcal{S}_n^k$ if $\text{MSR}(Y_n|\mathcal{U}_n^{k-1}) - \text{MSR}(Y_n|\mathcal{U}_n^k) > \gamma$ and continue the algorithm to find more candidate. Otherwise, terminate the algorithm and return $\mathcal{S}_n^{k-1}$ as the desired embedding vector.

The flow chart of the proposed algorithm is shown in Figure 3. CTE is then estimated by replacing $[X_n^-, Y_n^-, \mathbf{Z}_n^-]$ and $[Y_n^-, \mathbf{Z}_n^-]$ with $\mathcal{S}_n^{k-1}$ and $\mathcal{S}_n^{k-1}$ excluding the past of X_n, respectively. The CTE is finally estimated using the KSG approach [13,14,32] (cf. Appendix A.2).

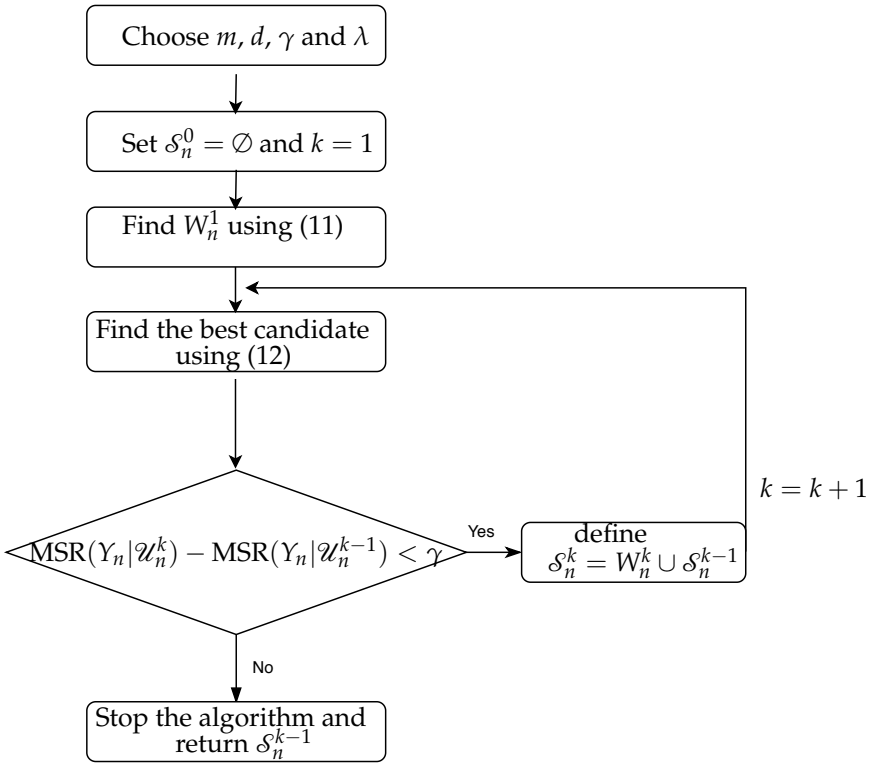


Figure 3. The flow chart of our proposed NUE algorithm.

5. Simulation Study

In this section, we use simulated data in order to compare the performance of our proposed NUE algorithm with the existing algorithms described in Section 2.2. We investigate the effect of the data length, strength of directed dependency and instantaneous coupling effect on the NUE algorithms. The execution time of the NUE algorithms are also investigated. The main reason for using simulated data are to be able to obtain well-defined ground truth. Therefore, it is possible to compare the NUE algorithms by computing their accuracies. The termination criterion of the NUE algorithms is also utilized for testing the significance of the estimated CTE in the simulation study: if the embedding vector $\mathcal{S}_n$ of the target variable Y_n does not include any lagged component of the node $\mathcal{X}$, then CTE from $\mathcal{X}$ to $\mathcal{Y}$ is zero and, otherwise its CTE is positive. The results are used to calculate true positive (TP), i.e., number of truly detected directed coupled nodes, true negative (TN), false positive (FP), and false negative (FN). The accuracy (ACC), true positive rate (TPR), and true negative rate (TPR) of the NUE algorithms are then defined as:

$$\begin{aligned} \text{ACC} &= 100 \times \tfrac{TP+TN}{TP+TN+FP+FN} \\ \text{TNR} &= 100 \times \tfrac{TN}{TN+FP} \\ \text{TPR} &= 100 \times \tfrac{TP}{TP+FN}. \end{aligned} \tag{13}$$

The TPR shows the ability of NUE algorithms to include the candidates in the embedding vector related to correctly coupled nodes, and TNR represents the ability to exclude the candidates related to uncoupled nodes. The ACC, TPR and TNR are computed as an average over 100 generated realizations because the simulated data depends on the random initial condition. The embedding delay m and dimension d are chosen as 1 and 5 samples, respectively. For estimation of the CMI and MSR, $T = 10$ nearest neighbors are considered.

5.1. Henon Map Model

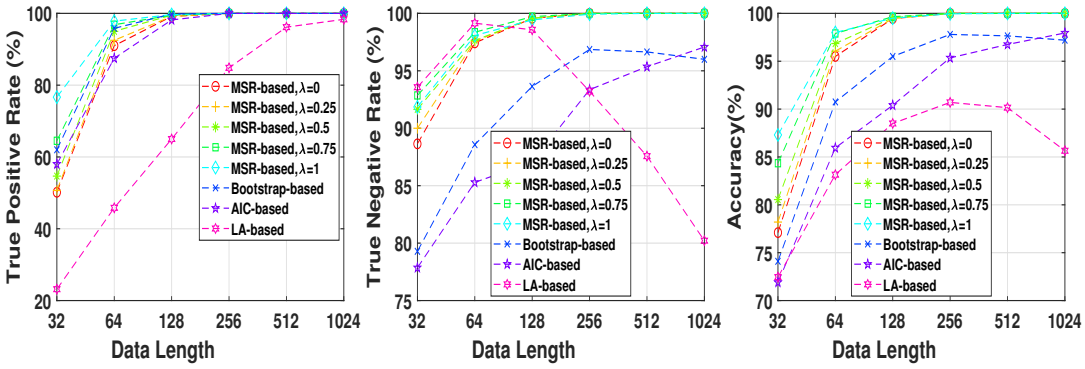
The Henon map model has been frequently utilized in the literature to generate multivariate data with a controlled amount of directed interaction [13,21,22]. A 5 nodes Henon map can be defined as [13,21,22]:

$$\begin{aligned} Y_{l,n} &= 1.4 - Y_{l,n-1}^2 + 0.3Y_{l,n-2}, \quad \text{for } l = 1,5 \\ Y_{l,n} &= 1.4 - \left[0.5Q(Y_{l-1,n-1} + Y_{l+1,n-1}) + (1-Q)Y_{l,n-1}\right]^2 + 0.3Y_{l,n-2}\,, \quad \text{for } l = 2,3,4, \end{aligned} \tag{14}$$

where Q is the coupling strength and it varies between 0.2 to 0.8 in this study; it is guaranteed that the complete synchronization between any pair nodes is avoided [34]. The first and last nodes (Y_1 and Y_5) depend only on their own past (first row of (14)) and therefore they do not depend on other nodes. On the other hand, nodes $l = 2,3,4$ depend on the past of nodes Y_{l-1} and Y_{l+1}. Consequently, there are nonlinear directed dependencies with strength Q from nodes Y_{l-1} and Y_{l+1} to node Y_l for $l = 2,3,4$ (second row of (14)). The aforementioned connectivity is considered as the ground truth when comparing the performance of the NUE algorithms.

5.1.1. Data Length Effect

Henon map data sequences were generated at a fixed normal strength $Q = 0.6$ and different lengths, $N = 2^h, h = 5,6,\ldots,10$, in order to evaluate the effect of the data length on the performance of the NUE algorithms. The proposed NUE algorithm were used with five different weights, $\lambda = 0, 0.25, 0.5, 0.75, 1$, to demonstrate the effect of the weight. According to the fact that in this simulation there is no unobserved confounder effect like IC, we set the parameter $\gamma = 0$. Figure 4 shows TPRs, TNRs and accuracies of the MSR-based NUE algorithm with five different λ's. In addition, shown in the figure, are the performances of the existing NUE algorithms. As Figure 4c demonstrates, the accuracy of our proposed NUE algorithm (for any λ) increases as the data length increases up to 256 samples where the accuracy is nearly 100%. The proposed algorithm with higher λ attains better performance at data length under 128 samples. Figure 4a,b show that the improvement of the accuracy by changing λ is mostly due to the better TPRs. As we can see in Figure 4b, TNRs of bootstrap-based and LA-based algorithms decreases for data lengths greater than 256 and 64, respectively. The accuracy, TPR and TNR of the AIC-based algorithm increases by increasing the data length. Overall, the proposed algorithm with $\lambda = 1$ attains the greatest accuracy and the LA-based algorithm has the worst accuracy for all data lengths.

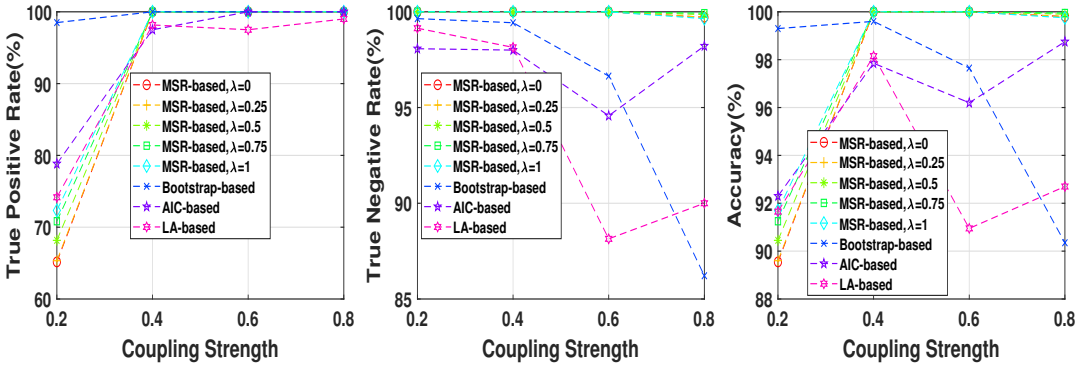


(**a**) (**b**) (**c**)

Figure 4. (**a**) true positive rates, (**b**) true negative rates and (**c**) accuracies of MSR-based, bootstrap-based, AIC-based, and LA-based NUE algorithms for the Henon map model at moderate fixed coupling strength $Q = 0.6$ and data length ranging from 32 to 1024. The results are shown as an average over 100 realizations.

5.1.2. Coupling Strength Effect

The Henon map model at 512 data length was generated with different coupling strengths ranging from 0.2 to 0.8 in step of 0.2 in order to evaluate the NUE algorithms as a function of the strength of the directed dependencies. As Figure 5 shows TNRs of the MSR-based algorithm (for any λ) is almost 100 % while the TNRs of the existing NUE algorithms tend to decrease as the strength of the directed dependency increase, which also causes a decrease in the accuracy. TPRs of the NUE algorithms are nearly equal except that at very low coupling strength the bootstrap-based algorithm has higher TPR. Changing λ at $Q = 0.2$ leads to slightly better TPR and accuracy. Overall, our proposed MSR-based algorithm has better accuracy compared to that of the existing NUE algorithms, except for $Q = 0.2$ where bootstrap-based algorithms yields better performance.



(**a**) (**b**) (**c**)

Figure 5. (**a**) true positive rates, (**b**) true negative rates and (**c**) accuracies of MSR-based, bootstrap-based, AIC-based, and LA-based NUE algorithms for the Henon map model at fixed data length $N = 512$ and coupling strength ranging from 0.2 to 0.8. The results are shown as an average over 100 realizations.

5.1.3. Execution Time

In this section the execution time of the proposed MSR-based algorithm with $\lambda = 1$ and $\lambda = 0$ (at fixed $\gamma = 0$) is compared with that of the existing NUE algorithms. The Henon Map data at length 512 samples and coupling strength $Q = 0.6$ was generated and execution time of the NUE algorithms are reported as an average over 100 realizations. The execution time was calculated in a single block-wise code where each NUE algorithms has a block. The function *tic* of MATLAB was set before each block and the function *toc* was used to calculate the execution time of the blocks related to

the NUE algorithms. The code was run by a Intel(R) core(TM) i7-7600 CPU@2.10 GHz. We use the ITS toolbox (available at http://www.lucafaes.net/its.html) for implementation of bootstrap-based NUE algorithm. The ITS toolbox was also modified for implementation of the LA-based algorithm by using a MATLAB code provided in [21]. We also modified ITS toolbox in order to implement the AIC-based and MSR-based NUE algorithms. The results are reported in Table 1. In addition to execution time, the total number of iterations k that the algorithms were performed before they terminated, are reported.

Table 1. The execution time and and total iterations before termination of the proposed MSR-based with $\lambda = 0, 1$ (at fixed $\gamma = 0$) as well as existing NUE algorithms for the Henon map data at data length 512. The results are reported as an average over 100 realizations.

NUE Algorithm	Bootstrap-Based	LA-Based	AIC-Based	MSR-Based, $\lambda = 1$	MSR-Based, $\lambda = 0$
Execution Time (second)	40.59	117.23	11.29	2.26	5.34
Total Number of Iterations	19.18	16.94	24.08	16.19	16.64

As Table 1 indicates, the execution time of MSR-based with the known parameters $\lambda = 1$ and $\lambda = 0$, and AIC-based NUE algorithms are significantly less than that of the bootstrap-based and LA-based ones. However, the total number of iterations of the AIC-based algorithm before termination is on average higher in comparison with that of the MSR-based algorithm. The higher total number of iterations of the AIC-based algorithm increases its execution time. It is important to note that the execution time of the MSR-based with $\lambda = 1$ is less than that of with $\lambda = 0$. Overall, our proposed MSR-based NUE algorithm with $\lambda = 1$ and $\gamma = 0$ attains the best and the LA-based has the worst execution time.

5.2. Autoregressive Model

AR models have been widely used to generate multivariate data with controlled directed dependencies among them [13,21,22]. The considered nonlinear AR model is given as:

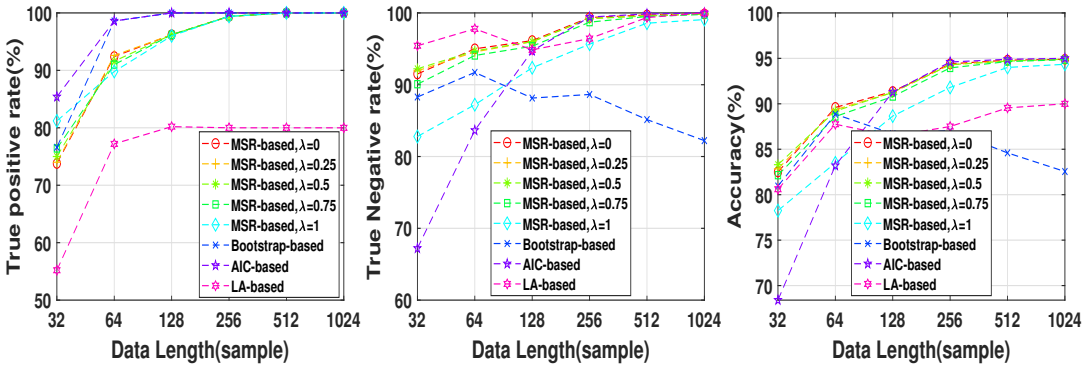
$$\begin{aligned} Y_{1,n} &= 0.95\sqrt{2}Y_{1,n-1} - 0.9125Y_{1,n-2} + \varepsilon_1 \\ Y_{2,n} &= 0.5Y_{1,n-2}^2 + \varepsilon_2 \\ Y_{3,n} &= -0.4Y_{1,n-3} + 0.4Y_{2,n-1} + \varepsilon_3 \\ Y_{4,n} &= -0.5Y_{1,n-1}^2 + 0.25\sqrt{2}Y_{4,n-1} + \varepsilon_4 \\ Y_{5,n} &= -0.25\sqrt{2}Y_{4,n-1} + 0.25\sqrt{2}Y_{5,n-2} + \varepsilon_5, \end{aligned} \tag{15}$$

where $\varepsilon_1, \ldots, \varepsilon_5$ are mutually independent zero mean and unit variance white Gaussian noise processes. In accordance with (15), node 1 only depends on its own past and therefore there is no directed dependency from other nodes to node 1 (first row of (15)). On the other hand, nodes 2, 3 and 4 depend on the past of node 1 and therefore there are nonlinear directed dependencies from node 1 and to nodes 2 and 4 (second and fourth rows of (15)) and linear directed dependencies from node 1 to node 3 (third row of (15)). There are also linear directed dependencies from nodes 2 and 4 to nodes 3 and 5, respectively (third and fifth rows of (15)). These dependencies describe the ground truth couplings when comparing TPR, TNR, and ACC of the NUE algorithms.

5.2.1. Data Length Effect

Nonlinear AR data series were first generated for 100 realizations at different lengths, $N = 2^h$, $h = 5, 6, \ldots, 10$, in order to evaluate the effect of data length on the performance of the NUE algorithms using AR data. We set the parameter $\gamma = 0$ since in this simulation there is no IC effect. Figure 6 shows TPRs, TNRs and accuracies of the NUE algorithms for the AR model as a function of data lengths. As Figure 6a illustrates, the LA-based NUE algorithm has significantly lower TPR compared to that

of the other algorithms. It is also noteworthy that the TNR of the bootstrap-based algorithm tends to decrease as the data length increases. The MSR-based algorithm, for all λ except $\lambda = 1$, presents higher accuracy than that of the bootstrap-based and LA-based algorithms at all data lengths and higher accuracy than that of the AIC-based algorithms at data length smaller than 128.



(**a**) (**b**) (**c**)

Figure 6. (**a**) true positive rates, (**b**) true negative rates and (**c**) accuracies of MSR-based, bootstrap-based, AIC-based, and LA-based NUE algorithms for the AR at data length ranging from 32 to 1024. The results are shown as an average over 100 realizations.

5.2.2. Instantaneous Coupling Effect

IC can happen due to sharing information at same lag. In other words, it can occur due to fast sharing information [31]. For example, in neuro-physiological time series like EEG, the recorded electrical activity at each electrode located at the scalp, is considered to be a mixture of the source generators because the sources pass through the volume conductor [18]. The volume conduction can be considered as the zero lag coupling which may lead to detection of false directed dependency by the NUE algorithms.

Let us consider the AR model defined in (15) at length N as the sources, which are instantly mixed to simulate the effect of IC. The considered mixing matrix is given as

$$A = \begin{bmatrix} (1-\alpha) & \alpha & \alpha & \alpha & \alpha \\ \alpha & (1-\alpha) & \alpha & \alpha & \alpha \\ \alpha & \alpha & (1-\alpha) & \alpha & \alpha \\ \alpha & \alpha & \alpha & (1-\alpha) & \alpha \\ \alpha & \alpha & \alpha & \alpha & (1-\alpha) \end{bmatrix}. \tag{16}$$

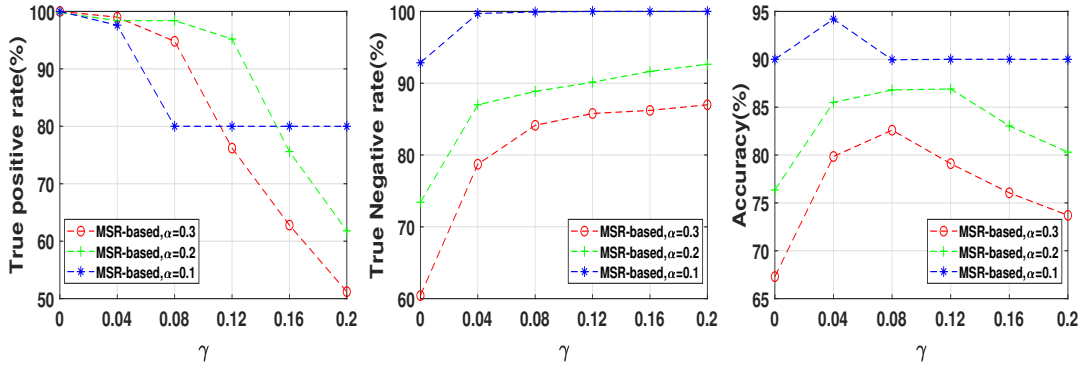
where α varies between 0.1 and 0.3 in step of 0.1 in this paper. The greater α, the greater IC between the sources. Let $Y = [Y_1, Y_2, \ldots, Y_5]^T$ be $N \times 5$ matrix which includes all sequences (they are considered to simulate sources in the brain) generated by the AR model (15). The mixed matrix (it is considered to simulate the EEG signals recorded at the scalp level which is the mixture of all sources) is then defined as the matrix product between Y and A that is

$$Y^{mixed} = YA. \tag{17}$$

Each column of A defines how the sources $Y_1, \ldots Y_5$ are mixed. As expected, for the nth mixed data sequence Y_n^{mixed}, the most important term is Y_n. This is more clear by looking at the main diagonal of the A.

The nonlinear AR data series were first generated for 100 realization at data lengths 512 using (15) and then mixed using (17) in order to evaluate the effect of IC on the performance of the NUE algorithms. As it was mentioned in Section 3, selecting a decent γ can control the balance between true positives and true negatives which can be useful, for example, to increase the accuracy of our

proposed MSR-based NUE algorithm when there is an unobserved confounder effect like IC effect. Therefore, the proposed algorithm was implemented using six γs. We set a fixed $\lambda = 0.5$ since in this section the goal is to investigate effect of γ on the performance of the MSR-based algorithm. Figure 7 demonstrates the TPRs, TNRs and accuracies of the MSR-based with six γ when they are applied on the data with three instantaneous couplings, i.e., $\alpha = 0.1, 0.2, 0.3$. As we can see in Figure 7, the TNR of the MSR-based algorithm increases by increasing γ while the TPR gradually decreases up to a certain γ (e.g., $\gamma = 0.04$ for $\alpha = 0.1$) and then it significantly declines. Accordingly, the accuracy increases up to a certain γ due to the increasing of the TNR compensating for the slight decrease of the TPR. Table 2 illustrates accuracies of the existing NUE algorithms as well as the best accuracy of the MSR-based algorithm which is obtained by a reported γ in the table. As Table 2 demonstrates, accuracies of the NUE algorithms decrease by increasing instantaneous effect strength. Our proposed MSR-based NUE algorithm attains the greatest accuracy compared to the existing algorithms for all αs.



(**a**) (**b**) (**c**)

Figure 7. (**a**) true positive rates, (**b**) true negative rates and (**c**) accuracies of the MSR-based algorithm with different γ ranging from 0 to 0.2 in step of 0.04 when applied to the mixed AR data sequences at length 512 with different instantaneous coupling strength $\alpha = 0.1, 0.2, 0.3$. The results are shown as an average over 100 realizations.

Table 2. Accuracies of the bootstrap-based, LA-based and AIC-based algorithms as well as the proposed MSR-based algorithm. The γ leads to the best accuracies of the MSR-based algorithm are also reported in parenthesis after the accuraies. The results are reported as an average over 100 realizations.

NUE Algorithm	Bootstrap-Based	LA-Based	AIC-Based	MSR-Based (Best γ)
$\alpha = 0.1$	71.10	86.30	88.65	94.20 ($\gamma = 0.04$)
$\alpha = 0.2$	71.80	77.75	73.20	86.90 ($\gamma = 0.12$)
$\alpha = 0.3$	63.55	75.10	60.55	82.60 ($\gamma = 0.08$)

5.2.3. Execution Time

In this section the AR model data at length 512 was generated and execution time of the NUE algorithms is reported as an average over 100 realizations in Table 3. Similar to the results reported in Section 5.1.3, the MSR-based algorithm with $\lambda = 1$ (at fixed $\gamma = 0$) is the fastest algorithm and LA-based one is the slowest one. Although the total number of iterations of the AIC-based and MSR-based algorithms with $\lambda = 0$ before termination are almost the same (around 10 iterations), the execution time of the AIC-based is slightly higher. It can be due to the fact that we did not have access to optimal code for calculating the KDE-based regression while for the NN-based prediction we have used a mex file for the neighbor search which is provided by the ITS toolbox [19].

Table 3. The execution time of our proposed MSR-based with $\lambda = 0, 1$ (at fixed $\gamma = 0$) as well as existing NUE algorithms for the AR data at length 512. The results are reported as an average over 100 realizations.

NUE Algorithm	Bootstrap-Based	LA-Based	AIC-Based	MSR-Based, $\lambda = 1$	MSR-Based, $\lambda = 0$
Execution Time	28.96	38.02	5.09	1.62	3.64
Total Number of Iterations	14.13	7.65	10.15	10.61	10.12

6. Application

In this section, we demonstrate the applicability of our proposed MSR-based algorithm on a real-world data. We consider a publicly available high dimensional intracranial EEG data from an epileptic patient. While our proposed estimator is defined for stationary stochastic processes, at least for this particular case of real world EEG data, our estimator is also able to provide good results when applied on non-stationary signals. The overall goal here is to apply NUE algorithms to estimate CTE and find patterns related to the onset and spread of the seizure. A total of 76 implanted electrodes was recorded, resulting in 76 time series. Electrodes 1–64 are cortical electrode grid and electrodes 65–76 are in-depth electrodes (six electrodes on each side). The data comprises 8 epileptic seizures (Ictal) and 8 periods just before the seizure onset (Pre-ictal) segments. Each segment is 10 seconds intracranial EEG data recorded at 400 Hz sampling frequency (more details about this data can be found in [35]). In this work, an anti-aliasing low-pass filter with a cutoff frequency of 50 Hz was applied prior to downsampling the signals to 100 Hz (Slow temporal auto-correlation of signals can induce a bias in the estimated conditional TE, nonlinear prediction and CMI in the NUE algorithms [36]. An approach used to correct this bias is called Theiler correction based on which too close observations in time should be discarded from the NN searches included in the estimation of TE, CMI and MSR [36]. In this paper, we down-sample the EEG data to avoid slow auto-correlation bias. In other words, the Theiler window is 4 samples.). The embedding delay and dimension were chosen as 1 and 8, respectively.

Epileptologists recognized the regions corresponding to one of the depth strips (electrodes 70 to 76) and the lower left corner of the grid (electrodes 1–4, 9–11 and 17) were resected during anterior temporal lobectomy as the seizure onset zone, which means synchronous activity of neurons in the specific regions of the brain becomes so strong, so that it can propagate its own activity to other distant regions [7,13,21,23]. From an information theory point of view, these nodes send information to other nodes, resulting in seizure onset. The amount of information each node sends to other nodes can be computed by the summation over each row of the directed dependencies matrix.

We applied our proposed in addition to bootstrap-based and LA-based NUE algorithms to estimate CTE in real high dimensional and redundant intracranial EEG data. The overall goal here is to compare advantages of our proposed NUE algorithms over the other algorithms reported in the literature. The MSR-based NUE algorithm was implemented with $\lambda = 1$ and $\gamma = 0.005$. The directed dependencies matrices obtained by our proposed algorithm as well as the existing algorithms are shown in Figure 8. The directed dependencies matrices obtained by the bootstrap-based NUE algorithm (Figure 8b,e) contain many connections in both pre-ictal and ictal conditions. Specifically, the diagonal pattern observed in the matrices obtained by the bootstrap-based NUE algorithm can be due to the volume conduction and conduction effect of the grid. On the other hand, our proposed (Figure 8a,d) and LA-based NUE algorithms (Figure 8c,f) are less sensitive to the volume conduction effect in comparison to that of the bootstrap-based algorithm.

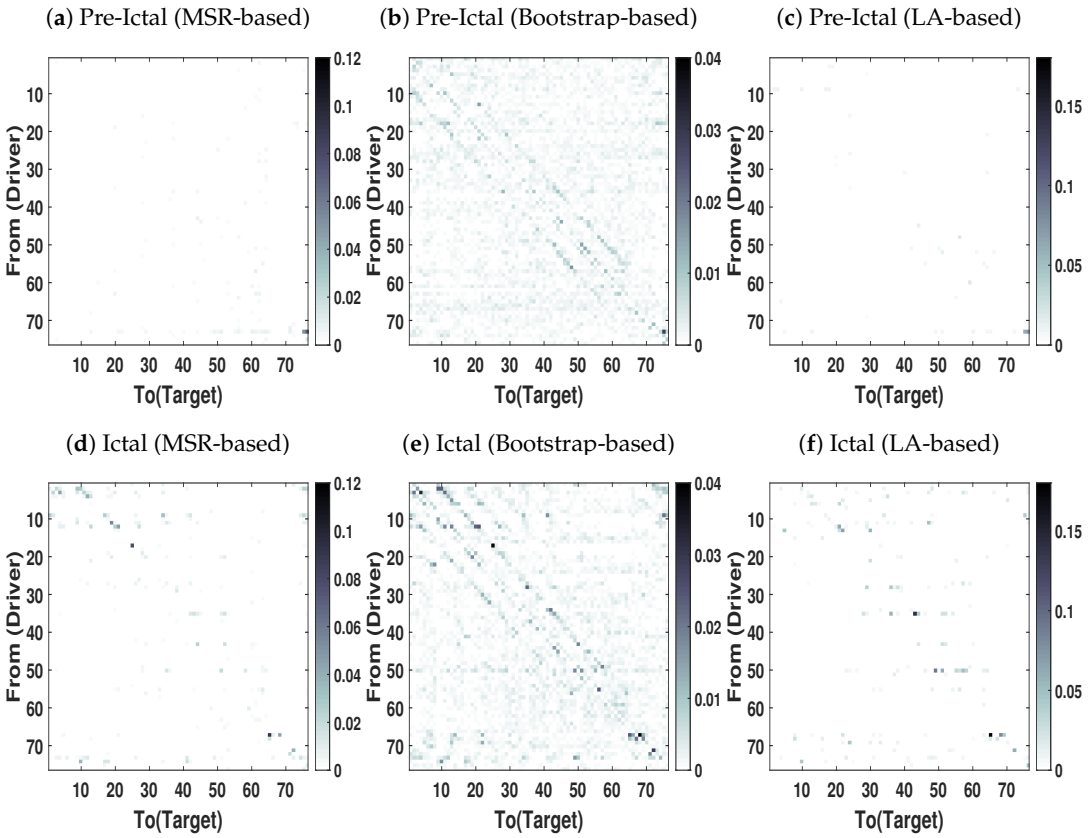


Figure 8. Directed dependency Matrices obtained by applying NUE algorithms on intracranial EEG data at epileptic seizures (Ictal) and just before the seizure onset (Pre-ictal) conditions. The directed dependency is shown from rows (driver) to the colomns (Targets). The darker color of an element, the higher the directed dependency is. The results are shown as an average over 8 segments.

Figure 9 represents the total amount of information each electrode sends to other electrodes. As Figure 9b demonstrates, due to the volume conduction effect there are some peaks even in the pre-ictal condition. On the other hand, the amount of information each electrode sends in the pre-ictal condition obtained by the MSR-based (Figure 9a) and LA-based (Figure 9c) NUE algorithms is approximately zero except for electrode 73. This electrode can be associated with the seizure onset although it is not yet clinically observable.

As mentioned earlier, electrodes 2–4, 9–11 and 17 are the seizure onset zones. Figure 9d,f show that the magnitude of the peaks at electrodes 2–4 and 9–11 for the MSR-based algorithm is higher than the one of the LA-based procedure. It is also important to mention that the existing LA-based and bootstrap-based NUE algorithms are not able to detect the peak at electrode 17 as opposed to that of our proposed MSR-based NUE algorithm.

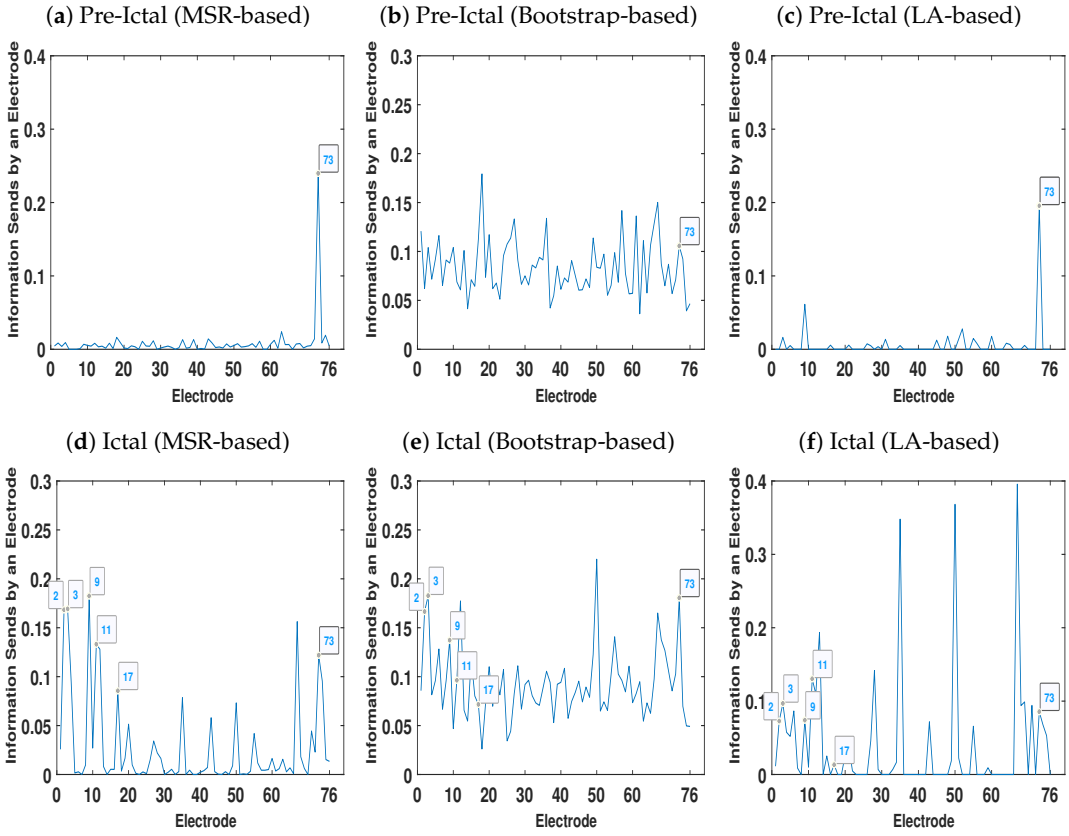


Figure 9. Total Information each electrode sends to other contacts at ictal and pre-ictal conditions.

7. Discussion and Conclusions

Reliable estimation of the directed dependencies in conditional high dimensional data are limited by the so-called ´´curse of dimensionality" problem. A greedy approach called non-uniform embedding (NUE) algorithm was proposed in [13] to select the most relevant variables and reduce the dimension of the reconstructed state-space of the data. Then, the model-free directed dependencies measure, conditional transfer entropy (CTE) is estimated using the reconstructed state-space. The NUE strategy based on sequentially selecting the best candidates in a greedy way will generally not lead to the same performance as would be obtained by using a brute-force combinatorial approach, where the performance is maximized over all possible sets of candidates. It has, however, been shown that NUE approaches often lead to an improved accuracy of the CTE compared to that of uniform embedding approaches [13,19]. The NUE algorithm has been widely utilized to estimate the directed dependencies in neuro-physiological [15,16] and economical [17] applications. It still has some obstacles like using a bootstrap-based termination criterion which highly depends on the bootstrap size [27]. It has been shown in [9,28] that using an alternative to the bootstrap statistical test can be more accurate and computationally efficient.

In this paper, we proposed a new modification for the NUE algorithm which uses a weighted sum of conditional mutual information (CMI) and nearest neighbor(NN)-based prediction for ranking the candidates and the algorithm is terminated if the highest ranked candidate is not relevant enough to significantly improve the accuracy of the prediction of the target variable. It should be noted that while our simulations on synthetic and real world data indicate that using prediction accuracy can lead to better assessment of directed dependency, we have not been able to prove this from an estimation theoretic point of view. It should also be noted that for the linear Gaussian processes, accuracy of the prediction of the target variable given selected candidates $\mathrm{MSR}(Y_n|\mathcal{U}_n)$, is monotonically equivalent to the conditional entropy $H(Y_n|\mathcal{U}_n)$ [37].

The proposed NUE procedure was compared with the original bootstrap-based NUE algorithm in [13], low-dimensional approximation(LA)-based [21] and Akaike information criterion (AIC)-based [27]. Performance analysis using simulation data generated by Henon map and autoregressive (AR) models at different lengths and coupling strengths revealed that the proposed mean of the squared (MSR)-based NUE algorithm tends to outperform the existing ones for detecting the directed dependencies. Specifically, the higher true negative rate (TNR) of the proposed MSR-based NUE compared to that of the existing ones may represent better ability of the proposed algorithm to terminate at the correct iteration and, as a result, better functionality of the proposed termination criterion. The poor selectivity (or TNR) of the bootstrap-based is in line with the results observed in [21], where they also found higher false positives for the bootstrap-based procedure compared to that of the LA-based one. The proposed algorithm also attains less false positive in comparison to that of the LA-based approach. The greater true positive rate (TPR) of the MSR-based algorithm with higher λ for small simulated data length and low coupling strength can justify using the weighted sum for ranking candidates. However, the limitation of the proposed NUE algorithm is that, for very low coupling strength, the accuracy of the proposed estimator was not as good as for the the bootstrap-based one.

The applicability of the NUE algorithms in real-word data can be affected by unobserved confounder effects like instantaneous information sharing which can be falsely detected as directed dependencies [31]. The data sequences generated by the AR model were instantly mixed at different mixing strengths in order to simulate an instantaneous coupling (IC) effect. The results showed that, by choosing a proper parameter γ, the proposed MSR-based measure attains significantly better performance than the existing ones. The simulated data results were consistent with the real-data used in this paper where the best results also were obtained for positive γ. The better performance can be of particular importance for such real-world applications like electroencephalography (EEG) and magnetoencephalography in which the volume conduction effect can cause IC [18]. There are also other frameworks like compensated transfer entropy [31], which tries to improve the estimation of the TE in the presence of IC. This measure modified the definition of the transfer entropy to compensate the effect of IC. The NUE algorithms are defined to find the embedding vector for estimating transfer entropy. Therefore, comparison or even modification of the proposed NUE algorithm for restructuring the state-space to estimate compensated transfer entropy deserves an independent and comprehensive study and will be considered in future works.

The proposed MSR-based algorithm with known parameter γ achieved a significant improvement in the computational efficiency. This can be due to the elimination of the computation effort of the bootstrap test which is not included in the proposed MSR-based algorithm. If we consider that the estimation of the CMI dominates the computation of the NUE algorithms (except for the MSR-based with $\lambda = 1$), then the overall computational requirement of the NUE algorithms which uses bootstrap-based test in the worst case will be $k|\mathscr{C}| + 100k$, where k is the number iterations reported in Tables 1 and 2 and $|\mathscr{C}|$ is the cardinality of $\mathscr{C}$. On the other hand, the computational requirement of the proposed NUE algorithm with known parameter γ can be expressed as $k|\mathscr{C}|$. The computational effort of the MSR-based NUE algorithm with $\lambda = 1$ can be considered to be dominated by the estimation of MSR. It is computationally less complex than that of the CMI since it only includes a neighbor search while CMI estimation contains a neighbor search and range searches. Therefore, in very high-dimensional data (like the intracranial EEG data used in the application part where $|\mathscr{C}| = 608$), where execution of the NUE algorithm can be very time-consuming, it is suggested to use $\lambda = 1$, since it will be significantly faster. The proposed NUE algorithm with $\lambda = 1$ also achieved better execution times than that of with $\lambda = 0$ in the simulation data used this paper. As already mentioned in [21,23], the LA-based approximation of the CMI used in the LA-based NUE algorithm is computationally more expensive, and this is consistent with the execution time reported in this paper where LA-based procedure attains the worst execution time. Better execution time can be especially important for such applications like a scalp EEG-based brain-computer interface where faster time

series analyses methods are required. We also consider testing the performance our proposed estimator on high dimensional scalp EEG data in future works.

Another parameter of our proposed NUE algorithm over which one needs to scan is the positive parameter γ. The parameter γ and $MSR(Y_n|\mathcal{U}_n)$ have the same units, and it defines the required amount of improvement in the accuracy of prediction prior to selecting a variable. Intuitively, the prediction accuracy $MSR(Y_n|\mathcal{U}_n^k)$ can vary between 0 and var(y_n), where 0 shows that one can perfectly predict Y_n by incorporating $\mathcal{U}_n^k$. The intuition of the worst case of the accuracy of the prediction $MSR(Y_n|\mathcal{U}_n^k)$ can be the case that incorporating $\mathcal{U}_n^k$ does not help the prediction at all and the indices specified by neighbor search in $\mathcal{u}_n$ will be uniformly distributed. The obtained $\widehat{y_n}(i|\mathcal{U})$ will be an approximation of mean of y_n and as a result MSR will be approximately var(y_n). In this paper, we normalize time series related to the realizations of the target processes to have zero mean and unit variance. We therefore scan the parameter γ in the interval between 0 and 1 to tune the algorithm. Therefore, another limitation of our proposed NUE algorithm is that it needs to be tuned by scanning over the parameter γ. The optimal choice of γ will be data dependent. The more accurate investigation of the criterion with which the parameter γ can be selected will be considered in the future works. Moreover, scanning over γ can increase execution time of our proposed algorithm. We suggest tuning the algorithm by using small subset of segments and use the tuned algorithm for the rest of segments. The reason is that the parameter γ can take care of confounder effects found in the data and will not vary during the segments such as volume conduction effect in neuro-physiological time series [18].

In this paper, TE has been used to assess the directed dependencies. Estimated TE in networks consisting of more than two nodes can be affected by other nodes through, for example, an indirect path or common shared information. One possible approach to reduce such effects is to condition out information coming from other nodes. However, this approach can present bias in the estimated directed dependencies in data in which there is the collider condition [38]. There are other approaches to assess directed dependencies in the network like decomposing TE into unique, synergistic, and redundant information [39]. However, comparison of the estimated conditional TE and decomposing TE deserves an independent and comprehensive study and it is out of the scope of this paper.

Author Contributions: Conceptualization, P.S.B. and J.Ø.; Methodology, P.S.B. and J.Ø.; Software, P.S.B.; Validation, P.S.B., C.G., E.A. and J.Ø.; Formal Analysis, P.S.B., C.G., E.A. and J.Ø.; Investigation, P.S.B., C.G., E.A. and J.Ø.; Writing—Original Draft Preparation, P.S.B.; Writing—Review & Editing, P.S.B., C.G., E.A. and J.Ø.; Visualization, P.S.B., C.G., E.A. and J.Ø.; Supervision, C.G. and J.Ø. All authors have read and agreed to the published version of the manuscript

Funding: This research was partially funded by Centre for Acoustic Signal Processing Research (CASPR).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Kraskov–Grassberger–Stögbauer Estimator

Appendix A.1. Conditional Mutual Information Estimation

The Kraskov–Grassberger–Stögbauer approach [14] is an NN-based estimator which was originally developed in order to estimate mutual information. It was adapted to estimate CMI in the NUE algorithm in [13,19,32]. The CMI in (3) can be rewritten as the sum/difference of four joint entropies [13,32]

$$\mathrm{I}\left(Y_n; W_n \mid \mathcal{S}_n^{k-1}\right) = h(Y_n, \mathcal{S}_n^{k-1}) - h(\mathcal{S}_n^{k-1}) - h(Y_n, W_n, \mathcal{S}_n^{k-1}) + h(W_n, \mathcal{S}_n^{k-1}). \tag{A1}$$

Then, the CMI is estimated by using a NN approach in which the entropy of the higher dimension $h(Y_n, W_n, \mathcal{S}_n^{k-1})$ is estimated through a neighbor search as [13,32]

$$h(Y_n, W_n, \mathcal{S}_n^{k-1}) \approx -\psi(T) + \psi(N) + (d+1)\langle ln(\epsilon_n(i))\rangle, \tag{A2}$$

where ψ is the digamma function, and N is the total number of observations of the vector variable $[Y_n, W_n, \mathcal{S}_n^{k-1}]$. Twice the amount of distance (maximum norm) of ith observation of $[Y_n, W_n, \mathcal{S}_n^{k-1}]$ from its Tth neighbor is denoted by $\epsilon_n(i)$ and $\langle . \rangle$ is the average over all observations. The rest of entropies in (A1) are estimated by using a range search as

$$\begin{aligned} h(W_n, \mathcal{S}_n^{k-1}) &\approx -\psi(T) + d \left\langle \psi \left(N_{[W_n, \mathcal{S}_n^{k-1}]} + 1 \right) \right\rangle \\ h(Y_n, \mathcal{S}_n^{k-1}) &\approx -\psi(T) + d \left\langle \psi \left(N_{[Y_n, \mathcal{S}_n^{k-1}]} + 1 \right) \right\rangle \\ h(\mathcal{S}_n^{k-1}) &\approx -\psi(T) + (d-1) \left\langle \psi \left(N_{\mathcal{S}_n^{k-1}} + 1 \right) \right\rangle . \end{aligned} \tag{A3}$$

The number of realizations of $[W_n, \mathcal{S}_n^{k-1}]$ whose maximum norm from the ith realization of $[W_n, \mathcal{S}_n^{k-1}]$ is strictly less than $\epsilon_n/2$, which is denoted by $N_{[W_n, \mathcal{S}_n^{k-1}]}$. A similar notation applies to $N_{[Y_n, \mathcal{S}_n^{k-1}]}$ and $N_{\mathcal{S}_n^{k-1}}$. The CMI is finally estimated by replacing (A2) and (A3) in (A1)

$$I \left(Y_n; W_n \mid \mathcal{S}_n^{k-1} \right) = \psi(T) + \left\langle \psi \left(N_{\mathcal{S}_n^{k-1}} + 1 \right) - \psi \left(N_{[W_n, \mathcal{S}_n^{k-1}]} + 1 \right) - \psi \left(N_{[Y_n, \mathcal{S}_n^{k-1}]} + 1 \right) \right\rangle . \tag{A4}$$

Appendix A.2. Conditional Transfer Entropy Estimation

After selecting the most informative candidates and forming the embedding vector $\mathcal{S}_n^k$ using the NUE algorithms, the CTE in (2) can be estimated using the same approach explained in Appendix A.1. The CTE can also be expressed as the sum of four joint entropies as

$$\mathrm{CTE}(\mathcal{X} \to \mathcal{Y} | \mathcal{Z}) = h(Y_n, Y_n^-, \mathbf{Z}_n^-) - h(Y_n^-, \mathbf{Z}_n^-) - h(Y_n, Y_n^-, X_n^-, \mathbf{Z}_n^-) + h(Y_n^-, X_n^-, \mathbf{Z}_n^-), \tag{A5}$$

where $[X_n^-, Y_n^-, \mathbf{Z}_n^-]$ is replaced by $\mathcal{S}_n^{k-1}$ and $[Y_n^-, \mathbf{Z}_n^-]$ is substituted by $\mathcal{S}_n^{k-1}$ without any past variables of X_n. Then, by using range search in the higher dimension $[Y_n, Y_n^-, \mathbf{Z}_n^-]$ and range search in the rest of dimensions, the CTE can be estimated as

$$\mathrm{CTE}(\mathcal{X} \to \mathcal{Y} | \mathcal{Z}) = \psi(T) + \left\langle \psi \left(N_{[Y_n^-, \mathbf{Z}_n^-]} + 1 \right) - \psi \left(N_{[Y_n, Y_n^-, \mathbf{Z}_n^-]} + 1 \right) - \psi \left(N_{[X_n^-, Y_n^-, \mathbf{Z}_n^-]} + 1 \right) \right\rangle , \tag{A6}$$

where $N_{[Y_n^-, \mathbf{Z}_n^-]}$ denotes the number of realizations of whose maximum norm from its ith realization of is strictly less than $\epsilon_n/2$. The same notation applies to $N_{[Y_n, Y_n^-, \mathbf{Z}_n^-]}$ and $N_{[X_n^-, Y_n^-, \mathbf{Z}_n^-]}$.

Appendix B. Kernel Density Estimation-Based Prediction

In the Akaike information criterion-based termination criterion which is adapted in this paper to stop the NUE algorithm, one needs to predict the target variable Y_n given $\mathcal{U}_n^k = [W_n^k, \mathcal{S}_n^{k-1}]$ by using the kernel density estimation (KDE) approach. The KDE-based prediction is performed as:

$$\widehat{y}_n(i | \mathcal{U}_n^k) = \sum_{i=1}^{N} \frac{y_n(i) K_h(u_n^k, u_n^k(i))}{\sum_{i=1}^{M} K_h(u_n^k, u_n^k(i))}, \tag{A7}$$

where $\widehat{y}_n(i | \mathcal{U}_n^k)$ denotes for estimated ith observation of Y_n and K_h is the Gaussian kernel with Mahalonobis distance (Equation (A9)) [27]:

$$K_h(u_n^k, u_n^k(i)) = \frac{1}{(\sqrt{2\pi}h)^d} \exp \left(\frac{-\| u_n^k - u_n^k(i) \|}{2h^2} \right), \tag{A8}$$

$$\| u_n^k - u_n^k(i) \| = (u_n^k - u_n^k(i))^T \Sigma^{-1} (u_n^k - u_n^k(i)), \tag{A9}$$

where d and Σ are dimension (number of columns) and covariance of w_n^k, respectively. The bandwidth of the kernel function h is chosen for unit variance data as [27,28]:

$$h = 1.5 \left(\frac{1}{d+2} \right)^{1/(d+4)} N^{-1/(d+4)}. \tag{A10}$$

References

1. Omidvarnia, A.; Azemi, G.; Boashash, B.; O'Toole, J.M.; Colditz, P.B.; Vanhatalo, S. Measuring time-varying information flow in scalp EEG signals: orthogonalized partial directed coherence. *IEEE Trans. Biomed. Eng.* **2013**, *61*, 680–693. [CrossRef] [PubMed]
2. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
3. Baboukani, P.S.; Azemi, G.; Boashash, B.; Colditz, P.; Omidvarnia, A. A novel multivariate phase synchrony measure: Application to multichannel newborn EEG analysis. *Digit. Signal Process.* **2019**, *84*, 59–68. [CrossRef]
4. Schreiber, T. Measuring information transfer. *Phys. Rev. Lett.* **2000**, *85*, 461. [CrossRef] [PubMed]
5. Baboukani, P.S.; Mohammadi, S.; Azemi, G. Classifying Single-Trial EEG During Motor Imagery Using a Multivariate Mutual Information Based Phase Synchrony Measure. In Proceedings of the 2017 24th National and 2nd International Iranian Conference on Biomedical Engineering (ICBME), Tehran, Iran, 30 November–1 December 2017; pp. 1–4.
6. Gençağa, D. Transfer Entropy. *Entropy* **2018**, *20*, 288. [CrossRef]
7. Faes, L.; Marinazzo, D.; Stramaglia, S. Multiscale information decomposition: Exact computation for multivariate Gaussian processes. *Entropy* **2017**, *19*, 408. [CrossRef]
8. Derpich, M.S.; Silva, E.I.; Østergaard, J. Fundamental inequalities and identities involving mutual and directed informations in closed-loop systems. *arXiv* **2013**, arXiv:1301.6427.
9. Massey, J. Causality, feedback and directed information. In Proceedings of the 1990 International Symposium on Information Theory and its Applications (ISITA-90), Waikiki, HI, USA, 27–30 November 1990; pp. 303–305.
10. Wiener, N. *The Theory of Prediction. Modern Mathematics for Engineers*; McGraw-Hill: New York, NY, USA, 1956; pp. 165–190.
11. James, R.G.; Barnett, N.; Crutchfield, J.P. Information flows? A critique of transfer entropies. *Phys. Rev. Lett.* **2016**, *116*, 238701. [CrossRef]
12. Lizier, J.T.; Prokopenko, M. Differentiating information transfer and causal effect. *Eur. Phys. J. B* **2010**, *73*, 605–615. [CrossRef]
13. Montalto, A.; Faes, L.; Marinazzo, D. MuTE: A MATLAB toolbox to compare established and novel estimators of the multivariate transfer entropy. *PLoS ONE* **2014**, *9*, e109462. [CrossRef]
14. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. *Phys. Rev. E* **2004**, *69*, 066138. [CrossRef]
15. Lindner, M.; Vicente, R.; Priesemann, V.; Wibral, M. TRENTOOL: A Matlab open source toolbox to analyse information flow in time series data with transfer entropy. *BMC Neurosci.* **2011**, *12*, 119. [CrossRef] [PubMed]
16. Wibral, M.; Pampu, N.; Priesemann, V.; Siebenhühner, F.; Seiwert, H.; Lindner, M.; Lizier, J.T.; Vicente, R. Measuring information-transfer delays. *PLoS ONE* **2013**, *8*, e55809. [CrossRef] [PubMed]
17. Bossomaier, T.; Barnett, L.; Harré, M.; Lizier, J.T. *An Introduction to Transfer Entropy*; Springer International Publishing: Cham, Switzerland, 2016; pp. 65–95.
18. Ruiz-Gómez, S.J.; Hornero, R.; Poza, J.; Maturana-Candelas, A.; Pinto, N.; Gómez, C. Computational modeling of the effects of EEG volume conduction on functional connectivity metrics. Application to Alzheimer's disease continuum. *J. Neural Eng.* **2019**, *16*, 066019. [CrossRef]
19. Faes, L.; Marinazzo, D.; Nollo, G.; Porta, A. An information-theoretic framework to map the spatiotemporal dynamics of the scalp electroencephalogram. *IEEE Trans. Biomed. Eng.* **2016**, *63*, 2488–2496. [CrossRef] [PubMed]
20. Mehta, K.; Kliewer, J. Directional and Causal Information Flow in EEG for Assessing Perceived Audio Quality. *IEEE Trans. Mol. Biol. Multi-Scale Commun.* **2017**, *3*, 150–165. [CrossRef]
21. Zhang, J. Low-dimensional approximation searching strategy for transfer entropy from non-uniform embedding. *PLoS ONE* **2018**, *13*, e0194382. [CrossRef] [PubMed]

22. Xiong, W.; Faes, L.; Ivanov, P.C. Entropy measures, entropy estimators, and their performance in quantifying complex dynamics: Effects of artifacts, nonstationarity, and long-range correlations. *Phys. Rev. E* **2017**, *95*, 062114. [CrossRef]
23. Jia, Z.; Lin, Y.; Jiao, Z.; Ma, Y.; Wang, J. Detecting causality in multivariate time series via non-uniform embedding. *Entropy* **2019**, *21*, 1233. [CrossRef]
24. Kugiumtzis, D. Direct-coupling information measure from nonuniform embedding. *Phys. Rev. E* **2013**, *87*, 062918. [CrossRef]
25. Olejarczyk, E.; Marzetti, L.; Pizzella, V.; Zappasodi, F. Comparison of connectivity analyses for resting state EEG data. *J. Neural Eng.* **2017**, *14*, 036017. [CrossRef]
26. Novelli, L.; Wollstadt, P.; Mediano, P.; Wibral, M.; Lizier, J.T. Large-scale directed network inference with multivariate transfer entropy and hierarchical statistical testing. *Netw. Neurosci.* **2019**, *3*, 827–847. [CrossRef] [PubMed]
27. May, R.J.; Maier, H.R.; Dandy, G.C.; Fernando, T.G. Non-linear variable selection for artificial neural networks using partial mutual information. *Environ. Model. Softw.* **2008**, *23*, 1312–1326. [CrossRef]
28. Li, X.; Maier, H.R.; Zecchin, A.C. Improved PMI-based input variable selection approach for artificial neural network and other data driven environmental and water resource models. *Environ. Model. Softw.* **2015**, *65*, 15–29. [CrossRef]
29. Baboukani, P.S.; Graversen, C.; Østergaard, J. Estimation of Directed Dependencies in Time Series Using Conditional Mutual Information and Non-linear Prediction. In Proceedings of the European Signal Processing Conference (EUSIPCO), European Association for Signal Processing (EURASIP), 2020, in press.
30. Altman, N.S. An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Stat.* **1992**, *46*, 175–185.
31. Faes, L.; Nollo, G.; Porta, A. Compensated transfer entropy as a tool for reliably estimating information transfer in physiological time series. *Entropy* **2013**, *15*, 198–219. [CrossRef]
32. Faes, L.; Kugiumtzis, D.; Nollo, G.; Jurysta, F.; Marinazzo, D. Estimating the decomposition of predictive information in multivariate systems. *Phys. Rev. E* **2015**, *91*, 032904. [CrossRef] [PubMed]
33. Danafar, S.; Fukumizu, K.; Gomez, F. Kernel-based Information Criterion. *arXiv* **2014**, arXiv:1408.5810.
34. Faes, L.; Marinazzo, D.; Montalto, A.; Nollo, G. Lag-specific transfer entropy as a tool to assess cardiovascular and cardiorespiratory information transfer. *IEEE Trans. Biomed. Eng.* **2014**, *61*, 2556–2568. [CrossRef]
35. Kramer, M.A.; Kolaczyk, E.D.; Kirsch, H.E. Emergent network topology at seizure onset in humans. *Epilepsy Res.* **2008**, *79*, 173–186. [CrossRef]
36. Wibral, M.; Vicente, R.; Lizier, J.T. *Directed Information Measures in Neuroscience*; Springer: Berlin, Germany, 2014.
37. Barnett, L.; Barrett, A.B.; Seth, A.K. Granger causality and transfer entropy are equivalent for Gaussian variables. *Phys. Rev. Lett.* **2009**, *103*, 238701. [CrossRef]
38. Cole, S.R.; Platt, R.W.; Schisterman, E.F.; Chu, H.; Westreich, D.; Richardson, D.; Poole, C. Illustrating bias due to conditioning on a collider. *Int. J. Epidemiol.* **2010**, *39*, 417–420. [CrossRef] [PubMed]
39. Williams, P.L.; Beer, R.D. Generalized measures of information transfer. *arXiv* **2011**, arXiv:1102.1507.