Estimation of Directed Dependencies in Time Series Using Conditional Mutual Information and Non-linear Prediction

Payam Shahsavari Baboukani[1], Carina Graversen[2], Jan Østergaard[1]
[1]*Department of Electronic Systems, Aalborg University, Denmark*
[2]*Eriksholm Research Centre, Denmark*
{pasba,jo}@es.aau.dk, cagr@eriksholm.com

***Abstract*—It is well-known that estimation of the directed dependency between high-dimensional data sequences suffers from the "curse of dimensionality" problem. To reduce the dimensionality of the data, and thereby improve the accuracy of the estimation, we propose a new progressive input variable selection technique. Specifically, in each iteration, the remaining input variables are ranked according to a weighted sum of the amount of new information provided by the variable and the variable's prediction accuracy. Then, the highest ranked variable is included, if it is significant enough to improve the accuracy of the prediction. A simulation study on synthetic non-linear autoregressive and Henon maps data, shows a significant improvement over existing estimator, especially in the case of small amounts of high-dimensional and highly correlated data.**

***Index Terms*—directed dependency, input variable selection, non-linear prediction, conditional mutual information**

I. Introduction

Complex dynamical systems often consist of interacting subsystems with unknown inter-dependencies. A wide variety of time series analysis methods ranging from statistical and information theoretical to empirical based have been proposed in order to characterize the interaction between time-varying subsystems. These methods are primarily based on mutual information [1], phase synchrony [2], [3], and directed information flow [4], [5]. One limitation of mutual information and phase synchrony is that they are unidirectional and, therefore, they are not able to determine the direction of the information flow [6]. They are non-causal and cannot distinguish whether dependency is due to subsystem $\mathscr{A}$ is communicating with $\mathscr{B}$, or $\mathscr{B}$ with $\mathscr{A}$ [5]. Alternatively, directed information flow measures [6], [7] have been proposed for assessing directed dependencies.

Information theoretical-based transfer entropy (TE) [7] estimates the directed information flow from a variable X to another variable Y by using conditional mutual information (CMI). TE does not depend on any model in its formulation compared to other directed information flow measures like Granger causality [8], which makes this method capable of assessing both linear and non-linear interactions. This advantage has led to using TE in neuro-physiological [9], [10] and economical [11] applications.

Bivariate TE cannot determine whether information flow is due to direct path from X to Y [4], or it is due to information from another variable Z. Partial transfer entropy (PTE) [9], also called causally conditioned TE in [4], is an extension of bivariate TE which quantifies the amount of directed information flow from X to Yexcluding information coming from Z. Reliable estimation of PTE in high dimensional multivariate data by using classical probability distribution estimators like nearest neighbor (NN) [12], context tree weighting [13], and kernel density estimation (KDE) [14] is limited by the problem of "curse of dimensionality" [12], [14], [15]. To address this limitation, one can for example decrease the dimensionality through input variable selection techniques also known as non-uniform embedding (NUE) [9], [12], [14].

In [12], an NUE algorithm was proposed which uses CMI for selecting the most informative candidates about the current value of the target variable Y from a candidate set consisting of past values of X, Y, and Z. The termination criterion of the algorithm is defined by a confidence bound estimated by a bootstrap statistical test [12]. The size of bootstrap has an effect on the accuracy of the estimated confidence bound and, as a result, the estimated PTE. A higher bootstrap size, up to a threshold, generally leads to higher accuracy [16]. The size of the bootstrap depends on the quality of the signal [16]. The number of bootstraps were limited to around 100 in the literature [9], [12], [14], [15] because of computational complexity reasons.

An input variable selection technique using KDE-based non-linear prediction was proposed [16] as an alternative to the bootstrap test for terminating the algorithm.

In this paper, we propose a new NUE procedure, hereinafter will be referred to as mean of the squared residual (MSR)-based NUE, which uses MSR of the NN-based non-linear prediction as a termination criterion and utilizes a weighted sum of NN-based CMI and MSR for selecting candidates. Then, we compare its performance with the existing bootstrap-based approach which is part of the newly released MATLAB ITS toolbox [9], [14], [17].

The rest of the paper is organized as follows. In Section II, required background on PTE will be briefly reviewed. This is followed by the introduction of our proposed MSR-based estimator in Section III. Then, simulated data generated by Henon maps and non-linear autoregressive (AR) models will be described in Section IV. Results of the comparison between

the proposed and existing procedures will be presented in Section V. Section VI will discuss the results.

II. Background

A. Partial Transfer Entropy

Let us consider the M nodes network shown in Fig. 1 where we are interested in the information that flows from node $\mathscr{X}$ to node $\mathscr{Y}$ and which is not due to indirect paths through the remaining $M-2$ nodes $\boldsymbol{\mathscr{Z}} = \mathscr{Z}^1, ..., \mathscr{Z}^{M-2}$ (see Fig. 1a) or due to common shared information from $\boldsymbol{\mathscr{Z}}$ (see Fig. 1b). Let the output of node $\mathscr{X}$ be the stationary stochastic process $X_1, X_2, ..., X_N$, where $X_n \in \mathbb{R}, \forall n$, the same notation applies to the output of the nodes $\mathscr{Y}$ and $\boldsymbol{\mathscr{Z}}$. Moreover, let $X_n^- = [X_{n-1}, X_{n-2}, ..., X_1]$ describe the past of X_n. Similar notation applies to Y_n^- and $\mathbf{Z}_n^-$. The PTE from an individual node $\mathscr{X}$ to $\mathscr{Y}$ in the presence of $\boldsymbol{\mathscr{Z}}$ is then defined as:

$$\begin{aligned} \mathrm{PTE}(\mathscr{X} \to \mathscr{Y} \mid \boldsymbol{\mathscr{Z}}) &\triangleq \mathrm{CMI}(Y_n; X_n^- \mid Y_n^-, \mathbf{Z}_n^-) \\ &= h(Y_n \mid Y_n^-, \mathbf{Z}_n^-) - h(Y_n \mid X_n^-, Y_n^-, \mathbf{Z}_n^-), \end{aligned} \tag{1}$$

where $h(.)$ and CMI denote Shannon differential entropy and conditional mutual information, respectively. NUE approximates the past of, for example Y_n, by selecting the optimal candidate set from the embedding vector $V_n^Y = [Y_{n-m}, Y_{n-2m}, ..., Y_{n-dm}]$ which is a subsampled version of Y_n^- [12], where m and d are the embedding delay and embedding dimension, respectively [12], [17].

B. Bootstrap-based Non-uniform Embedding

In the bootstrap-based NUE technique, the most informative candidates about the target Y_n are selected from a candidate set $\mathscr{C} = [V_n^Y, V_n^X, V_n^{\mathbf{Z}}]$. The algorithm starts by the empty set of selected candidates, $\mathcal{S}_n^0 = \emptyset$. Then, at each step $k \geq 1$, the candidate W_n^k that maximizes the CMI is found by [12], [14]:

$$W_n^k = \underset{W_n \in \mathscr{C} \setminus \mathcal{S}_n^{k-1}}{\operatorname{argmax}} \mathrm{CMI}\left(Y_n; W_n \mid \mathcal{S}_n^{k-1}\right), \tag{2}$$

where $\mathcal{S}_n^{k-1} = \bigcup_{i=0}^{k-1} W_n^i$ is the set of selected candidates at last step $k-1$.

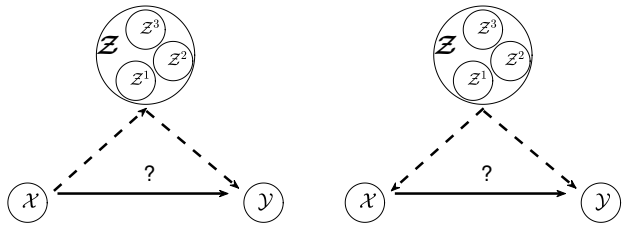

Fig. 1: An $M = 5$ nodes network where (a) indirect paths through remaining nodes $\mathbf{Z}$, (b) common shared information from $\mathbf{Z}$, may cause falsely (dashed line) detected directed dependency (solid line) from X to Y.

Then, a significance test is accomplished by 100 randomization of Y_n and W_n^k and estimating the 95^{th} percentile CMI^{95} of the empirical distribution generated by $\mathrm{CMI}(\widehat{Y}_n; \widehat{W}_n^k \mid \mathcal{S}_n^{k-1})$, where $\widehat{W}_n^k$ and $\widehat{Y}_n^k$ denote the randomized version of W_n^k and Y_n, respectively. If $\mathrm{CMI}(Y_n; W_n^k \mathcal{S}_n^{k-1})$ is larger than CMI^{95}, W_n^k will be included in the selected candidates $\mathcal{S}_n^k$ and the algorithm proceeds to search for more candidates in step $k+1$. Otherwise, the procedure is ended and $\mathcal{S}_n^{k-1}$ is returned as the optimal subsample of the vector variable $[X_n^-, Y_n^-, \mathbf{Z}_n^-]$. The $\mathcal{S}_n^{k-1}$ is finally used to estimate the PTE in (1).

C. Conditional Mutual Information Nearest Neighbor-based Estimation

NN-based estimation of the CMI in (2) has attracted attention in [9], [12], [15], [17] for performing bootstrap-based NUE. The CMI in (2) can be expressed as the sum of four unconditional entropies:

$$\begin{aligned} \mathrm{CMI}\left(Y_n; W_n \mid \mathcal{S}_n^{k-1}\right) &= h(Y_n, \mathcal{S}_n^{k-1}) - h(\mathcal{S}_n^{k-1}) \\ &\quad - h(Y_n, W_n, \mathcal{S}_n^{k-1}) + h(W_n, \mathcal{S}_n^{k-1}). \end{aligned} \tag{3}$$

Then, the CMI is estimated by using a NN approach in which the entropy $h(Y_n, W_n, \mathcal{S}_n^{k-1})$ is estimated through a neighbor search and the rest of the entropies in (3) are estimated by using a range search [12], [18].

III. Proposed Estimator

Our estimator uses the MSR of its prediction as a termination criterion and utilizes the weighted sum of the CMI and the MSR for selecting candidates. For the prediction, one can for example use KDE-based or NN-based non-linear prediction. In this work, we will focus on NN-based non-linear prediction, since it performed significantly better than KDE-based prediction on synthetic data.

The MSR-based estimator progressively selects the most informative candidates from the candidate set $\mathscr{C} = [V_n^Y, V_n^X, V_n^{\mathbf{Z}}]$. It initiates by the empty set of selected candidates, $\mathcal{S}_n^0 = \emptyset$, and then proceeds at each step $k \geq 1$ by finding the candidate W_n^k which maximizes a weighted sum of the CMI and MSR:

$$\begin{aligned} W_n^k = \underset{W_n \in \mathscr{C} \setminus \mathcal{S}_n^{k-1}}{\operatorname{argmax}} \big[&(1-\lambda)\,\mathrm{CMI}\left(Y_n; W_n \mid \mathcal{S}_n^{k-1}\right) \\ &- \lambda\,\mathrm{MSR}(Y_n \mid \mathscr{U}_n)\big], \end{aligned} \tag{4}$$

where $0 \leq \lambda \leq 1$ is the weight and $\mathscr{U}_n = [W_n, \mathcal{S}_n^{k-1}] \in \mathbb{R}^k$. The CMI in (4) is estimated by using the NN-based approach described in Section II-C. The MSR of the prediction of Y_n given the candidate W_n and the set of selected candidates at the last step, $\mathcal{S}_n^{k-1}$, is denoted by $\mathrm{MSR}(Y_n \mid \mathscr{U}_n)$.

Let $\mathscr{y}_n = [y_n(1), y_n(2)..., y_n(M)]^T$ be a set of M realizations of Y_n. Furthermore, let $M \times k$ matrix $\mathscr{u}_n$ be a set of M realizations of the vector variable $\mathscr{U}_n$. An estimator for the prediction of Y_n given $\mathscr{U}_n$ can be formed by the average of the realizations of Y_n which are found by using the neighbor search in the predictors (independent variables) $\mathscr{U}_n$. Let us denote by l_i the Euclidean distance from the i^{th} row of $\mathscr{u}_n$ (will be denoted by $\mathscr{u}_n(i)$ for simplicity) to its P^{th} neighbor.

Then the set of indices of rows of $\mathcal{u}_n$ whose distance from $\mathcal{u}_n(i)$ is less than d is given by $\mathcal{p}(i)$:

$$\mathcal{p}(i) = \{j \in \{1, 2, ..., M\} : \|u_n(i) - u_n(j)\|_2 < l_i \mid j \neq i\}, \tag{5}$$

where cardinality of the set $\mathcal{p}(i)$ is $P-1$. For instance, $\mathcal{p}(i) = \{1, 4, 15\}$ means that the Euclidean distance from the 1^{st}, 4^{th}, and 15^{th} rows of $\mathcal{u}_n$ to the i^{th} row of $\mathcal{u}_n$ is less than l_i. The prediction of the i^{th} realization of Y_n (i.e. $y_n(i)$) given $\mathcal{U}_n$ is then estimated as an average of realizations of Y_n whose indices are in $\mathcal{p}(i)$:

$$\widehat{y_n}(i|\mathcal{U}_n) \triangleq \frac{1}{P-1} \sum_{b \in \mathcal{p}(i)} y_n(b), \tag{6}$$

where $y_n(b)$ denotes b^{th} realization of Y_n. For example, if $\mathcal{p}(i) = \{1, 4, 15\}$ then $\widehat{y_n}(i|\mathcal{U})$ is equal to the average of $\{y_n(1), y_n(4), y_n(15)\}$. The residual $B_n(i|\mathcal{U}_n)$ can be calculated afterward as:

$$B_n(i|\mathcal{U}_n) = y_n(i) - \widehat{y_n}(i|\mathcal{U}_n). \tag{7}$$

The MSR of Y_n given $\mathcal{U}_n$ is then defined as the mean of the squared residuals $B_n(i|\mathcal{U}_n)$ over all M realizations:

$$\mathrm{MSR}(Y_n \mid \mathcal{U}_n) = \sum_{i=1}^{M} B_n(i|\mathcal{U}_n). \tag{8}$$

After calculating $\mathrm{MSR}(Y_n \mid \mathcal{U}_n)$ and replacing that in (4), W_n^k will be computed. Then, it will be included in the selected candidates, if it fulfills a new termination criterion which is also based on NN-based non-linear prediction. We first assume that $\mathcal{S}_n^1 = W_n^1$, which means that there is no significance test at the first step in the proposed algorithm. The main idea behind the new termination criterion is that if the candidate W_n^k is significant enough, it will decrease the MSR of the prediction. Accordingly, at each step $k \geq 2$, if $\mathrm{MSR}(Y_n|W_n^k, \mathcal{S}_n^{k-1}) < \mathrm{MSR}(Y_n|\mathcal{S}_n^{k-1})$ then W_n^K is included in $\mathcal{S}_n^k$ and the algorithm continues to search for more candidates at step $k+1$. Otherwise, the algorithm stops and $\mathcal{S}_n^{k-1}$ will be considered as the optimal subsample of the vector variable $[X_n^-, Y_n^-, \mathbf{Z}_n^-]$. Finally, the PTE is estimated by using (1) 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 $\mathcal{S}_n^{k-1}$ except any lagged component of X_n.

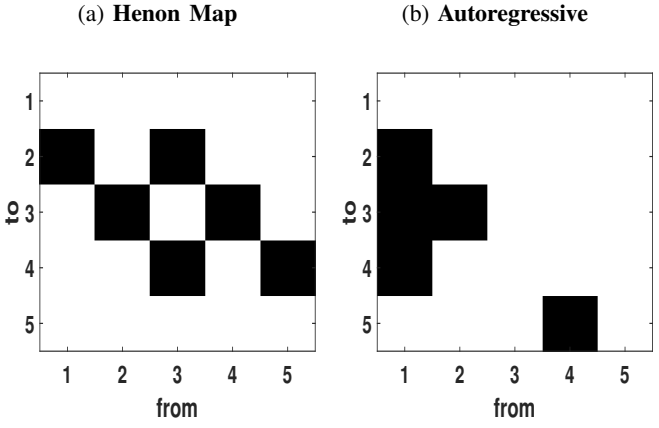

Fig. 2: Ground truth of the (a) Henon map and (b) AR model, used in this study. Information flow is shown from columns (driver) to the rows (Targets).Black squares show the information flow is significantly different from zero.

IV. Data For Performance Evaluation

A. Simulated data

Simulated data generated by a Henon map and a non-linear AR model will be used in order to evaluate the proposed MSR-based estimator and compare it to the existing bootstrap-based estimator. Henon maps and AR models have been widely used to generate multivariate data with controlled connectivity within them [12], [14], [15]. The Henon map is given as [12], [14], [15]:

$$\begin{aligned} Y_{m,n} &= 1.4 - Y_{m,n-1}^2 + 0.3Y_{m,n-2} \quad \text{for } m = 1, M \\ Y_{m,n} &= 1.4 - [0.5Q(Y_{m-1,n-1} + Y_{m+1,n-1}) \\ &\quad +(1-Q)Y_{m,n-1}]^2 + 0.3Y_{m,n-2} \\ &\quad \text{for } m = 2, ..., M-1, \end{aligned} \tag{9}$$

where Q is the coupling strength. The ground truth of the Henon map is shown in Fig. 2a. As the figure shows, a node m, except the first and last nodes, is coupled and therefore has non-linear information flow to the $(m-1)^{th}$ and $(m+1)^{th}$ nodes.

The considered non-linear AR model is defined as [12], [14], [15]:

$$\begin{aligned} Y_{1,n} &= 0.95\sqrt{2}Y_{1,n-1} - 0.9125Y_{1,n-2} + \varepsilon_1 \\ Y_{2,n} &= 0.5Y_{1,n-2}^2 + \varepsilon_2 \\ Y_{3,n} &= -0.4Y_{1,n-3} + 0.4Y_{2,n-1} + \varepsilon_3 \\ Y_{4,n} &= -0.5Y_{1,n-1}^2 + 0.25\sqrt{2}Y_{4,n-1} + \varepsilon_4 \\ Y_{5,n} &= -0.25\sqrt{2}Y_{4,n-1} + 0.25\sqrt{2}Y_{5,n-2} + \varepsilon_5, \end{aligned} \tag{10}$$

where $\varepsilon_{1,...,5}$ are mutually independent zero mean and unit variance white Gaussian noise. Node 1 has non-linear information flow to nodes 2 and 4. There is also linear information flow from nodes 4 and 2 to nodes 5 and 3, respectively. Aforementioned PTEs are considered as the ground truth for the considered AR data, as shown in Fig. 2b.

B. Statistical Tests

In order to compare the proposed NUE procedure with the existing one, we use the termination criterion of both bootstrap-based and MSR-based algorithms also for testing the significance of the estimated PTE: if the optimal candidate set $\mathcal{S}_n$ of the target variable Y_n does not include any lagged component of the node $\mathcal{X}$, then its PTE is zero and, otherwise the PTE from $\mathcal{X}$ to $\mathcal{Y}$ is positive. The termination criterion results were used to calculate true positive (TP),i.e for example, number of truly (defined by (9) and (10) and shown in Fig. 2) detected directed coupled nodes, true negative

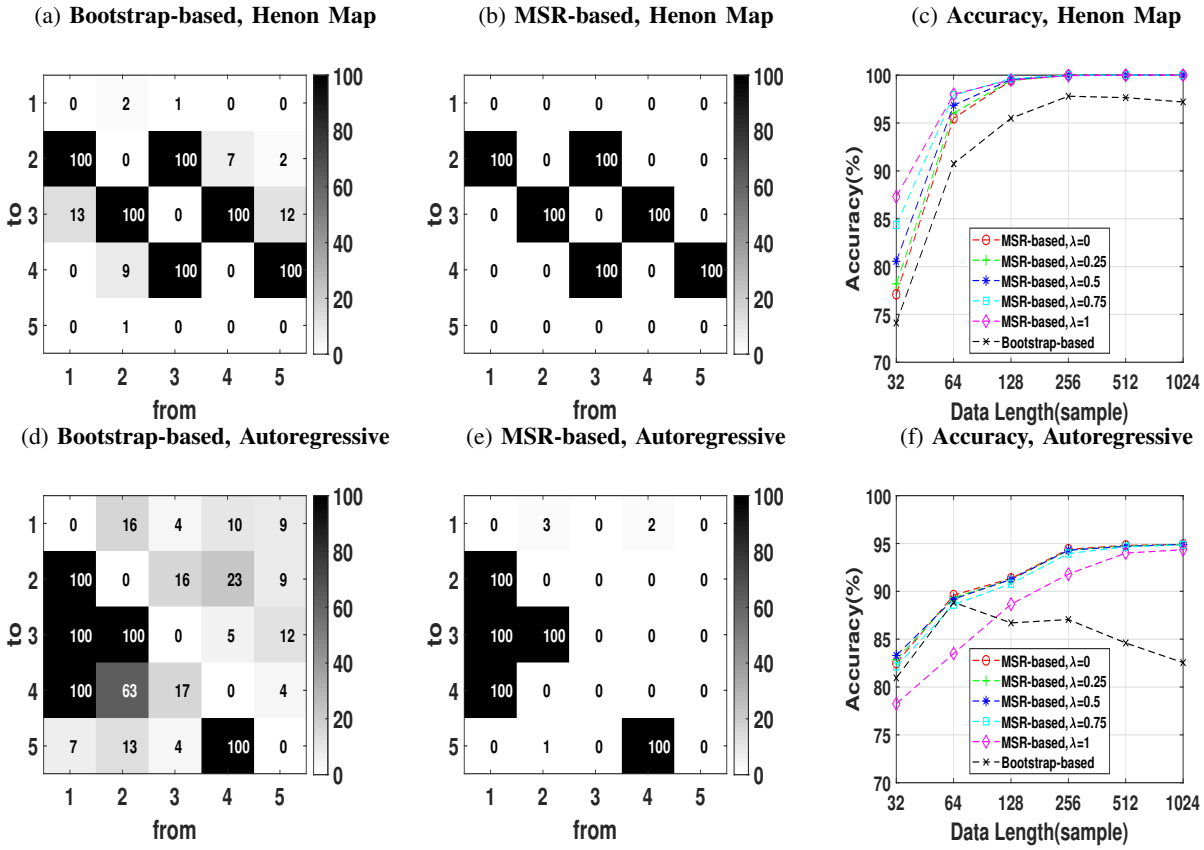

Fig. 3: Information flow matrices and accuracies of both MSR and bootstrap based estimators on both Henon map (first row) and AR model (second row). Information flow is shown from columns to the rows. The integers inside first two figures in each row, shows the number of time the associated information flow are selected over 100 realizations. Accuracies are shown as an average over all realizations.

(TN), false positive (FP), and false negative (FN). The ratio of correctly detected observations is then defined on the accuracy as ACC:

$$\text{ACC} = 100 \times \frac{TP + TN}{TP + TN + FP + FN}. \tag{11}$$

V. Simulation Study

Simulated data sequences containing both linear and non-linear interactions were generated at different lengths, $N = 2^l, l = 5, 6, ..., 10$, in order to evaluate the NUE algorithms. The embedding delay m and dimension d were chosen as 1 and 5 samples, respectively. The accuracy is computed as an average over 100 generated realizations. The MSR-based NUE is calculated with 5 different weights, $\lambda = 0, 0.25, 0.5, 0.75, 1$, to test the effect of the weight. For estimation of the CMI and prediction in (4) and (2), $P = 11$ nearest neighbor is considered.

The information flow matrices over 100 realizations of the Henon map, with medium-length $N = 512$, medium-coupling strength $Q = 0.6$, and for the MSR-based method, medium weight $\lambda = 0.5$, are shown in Figs. 3a and 3b. The information flows are depicted from columns to rows, for example, element $(2, 1)$ represents the PTE from node 1 to node 2. The darker color of an element, the more of the realizations this information flow is significant in. The integers inside the figures show the number of realizations in which the associated information flows are detected as significant. As can be seen, both the MSR-based and bootstrap-based estimators detect true directly connected pairs in all realizations, i.e. in 600 out of 600 information flows. The number of FP, meaning the pairs that are not directly connected but they are falsely detected, is for the MSR-based NUE zero and 47 for the existing NUE algorithm.

Fig. 3c represents the accuracy of the MSR-based with 5 different λ's and the existing bootstrap-based NUE algorithm. In this case, the data are the Henon map with medium coupling strength $Q = 0.6$. As the figure demonstrates, the accuracy of the proposed MSR-based NUE algorithm for any λ is higher than the bootstrap-based one (black line) at all data lengths. It is also noteworthy that the proposed estimator with higher λ attain better performance at data length under 128 samples.

The causal matrices of the 512 samples AR model data are summarized in Figs. 3d and 3e. It is observed that both estimators again detect all 500 directly connected pairs. Similar to the Henon map data results, the FP value of the bootstrap-based estimator is significantly higher than that of

the proposed MSR-based one; 212 in comparison to 6 pairs. Fig. 3f plots accuracies of the NUE algorithms for the AR model with data lengths ranging from 32 to 1024 samples. The MSR-based estimator, with all λ except $\lambda = 1$ at length 32 and 64 samples, presents higher accuracy than the existing one. It is worth noting that the accuracy of the bootstrap-based NUE algorithms significantly reduces at data length higher than 64 samples.

Finally, 512 samples Henon map data were generated with different coupling strengths ranging from 0.2 to 0.8 by 0.2 steps in order to evaluate both estimators as a function of the strength of the directed dependence. The results are shown in Fig. 4. Except for the very low coupling strength $Q = 0.2$, the proposed estimator outperforms the existing one. Higher λ for very low correlated data attains slightly better accuracy than for the lower weights. The performance of the existing procedure declines at coupling strength higher than $Q = 0.4$.

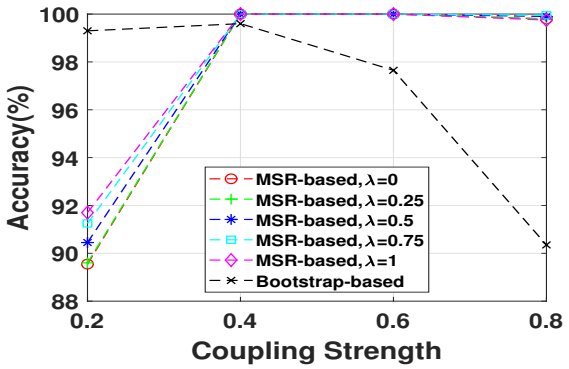

Fig. 4: Accuracy as a function of coupling strength.

VI. Discussion

The comparison between the existing and proposed estimators using simulated data including Henon maps and AR models, revealed a significantly higher FP of the bootstrap-based estimator, which is consistent with the results observed in previous work [15], where they also found high FP for the bootstrap-based estimator on some synthetic data. The higher FP appears to be due to the inefficient termination criterion of the bootstrap-based estimator, while the proposed procedure with the new termination criterion can stop the algorithm at the correct step that hinders selecting false directed dependent pairs. By training the MSR-based procedure and choosing a correct λ based on the application, it is possible to gain better performance specifically for small data length. Estimating the MSR is computationally less complex than for the CMI since it only includes a neighbor search, while CMI estimation contains a neighbor search and range searches. Therefore, for very high dimensional data where estimating the PTE would be time consuming, it is suggested to use $\lambda = 1$ which only uses MSR for the selection step (sophisticated comparison of the computational load is planned to be performed in the future works). For very low coupling strength, the accuracy of the proposed estimator was not as good as for the bootstrap-based method.

References

[1] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[2] P. S. Baboukani, G. Azemi, B. Boashash, P. Colditz, and A. Omidvarnia, "A novel multivariate phase synchrony measure: Application to multichannel newborn eeg analysis," *Digital Signal Processing*, vol. 84, pp. 59–68, 2019.

[3] P. S. Baboukani, S. Mohammadi, and G. Azemi, "Classifying single-trial eeg during motor imagery using a multivariate mutual information based phase synchrony measure," in *2017 24th National and 2nd International Iranian Conference on Biomedical Engineering (ICBME)*, pp. 1–4, IEEE.

[4] K. Mehta and J. Kliewer, "Directional and causal information flow in eeg for assessing perceived audio quality," *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, vol. 3, no. 3, pp. 150–165, 2017.

[5] M. S. Derpich, E. I. Silva, and J. Østergaard, "Fundamental inequalities and identities involving mutual and directed informations in closed-loop systems," *arXiv preprint arXiv:1301.6427*, 2013.

[6] J. Massey, "Causality, feedback and directed information," in *Proc. Int. Symp. Inf. Theory Applic.(ISITA-90)*, pp. 303–305, Citeseer, 1990.

[7] T. Schreiber, "Measuring information transfer," *Physical review letters*, vol. 85, no. 2, p. 461, 2000.

[8] P.-O. Amblard and O. J. Michel, "On directed information theory and granger causality graphs," *Journal of computational neuroscience*, vol. 30, no. 1, pp. 7–16, 2011.

[9] L. Faes, D. Marinazzo, G. Nollo, and A. Porta, "An information-theoretic framework to map the spatiotemporal dynamics of the scalp electroencephalogram," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 12, pp. 2488–2496, 2016.

[10] Z. Cai, C. L. Neveu, D. A. Baxter, J. H. Byrne, and B. Aazhang, "Inferring neuronal network functional connectivity with directed information," *Journal of neurophysiology*, vol. 118, no. 2, pp. 1055–1069, 2017.

[11] T. Bossomaier, L. Barnett, M. Harré, and J. T. Lizier, "An introduction to transfer entropy," *Cham: Springer International Publishing*, pp. 65–95, 2016.

[12] A. Montalto, L. Faes, and D. Marinazzo, "Mute: a matlab toolbox to compare established and novel estimators of the multivariate transfer entropy," *PloS one*, vol. 9, no. 10, p. e109462, 2014.

[13] J. Jiao, H. H. Permuter, L. Zhao, Y.-H. Kim, and T. Weissman, "Universal estimation of directed information," *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 6220–6242, 2013.

[14] W. Xiong, L. Faes, and P. C. Ivanov, "Entropy measures, entropy estimators, and their performance in quantifying complex dynamics: Effects of artifacts, nonstationarity, and long-range correlations," *Physical Review E*, vol. 95, no. 6, p. 062114, 2017.

[15] J. Zhang, "Low-dimensional approximation searching strategy for transfer entropy from non-uniform embedding," *PloS one*, vol. 13, no. 3, p. e0194382, 2018.

[16] R. J. May, H. R. Maier, G. C. Dandy, and T. G. Fernando, "Non-linear variable selection for artificial neural networks using partial mutual information," *Environmental Modelling & Software*, vol. 23, no. 10-11, pp. 1312–1326, 2008.

[17] L. Faes, D. Kugiumtzis, G. Nollo, F. Jurysta, and D. Marinazzo, "Estimating the decomposition of predictive information in multivariate systems," *Physical Review E*, vol. 91, no. 3, p. 032904, 2015.

[18] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical review E*, vol. 69, no. 6, p. 066138, 2004.