Assessment of Resampling Methods for Causality Testing

Angeliki Papana[*]
Department of Economics, University of Macedonia, Greece
apapana@uom.edu.gr, angeliki.papana@gmail.com

Catherine Kyrtsou
Department of Economics, University of Macedonia, Greece;
University of Strasbourg, BETA; University of Paris 10, France
ckyrtsou@uom.gr

Dimitris Kugiumtzis
Faculty of Engineering, Aristotle University of Thessaloniki, Greece
dkugiu@gen.auth.gr

Cees Diks
CeNDEF, University of Amsterdam, The Netherlands
C.G.H.Diks@uva.nl

May 21, 2014

Abstract

Different resampling methods for the null hypothesis of non-causality are assessed. As test statistic the partial transfer entropy (PTE), an information and model-free measure, is used. Two resampling techniques, 1) the time shifted surrogates and 2) the stationary bootstrap, are combined with the following three independence settings (giving in total six resampling schemes), all consistent to the null hypothesis

[*]Corresponding author

of non-causality: A) only the driving variable is resampled, B) both the driving and response variable are independently resampled, and C) both the driving and response variable are resampled while also the dependence of the future of the response variable and the vector of its past values is destroyed. The empirical null distribution of the PTE as the surrogate and bootstrapped time series become more independent is examined along with the size and power of the respective tests. Further, we consider the resampling method of contemporaneously resampling the the driving and the response time series using the stationary bootstrap. Although this resampling method does not comply with the non-causality hypothesis, one can obtain an accurate sampling distribution of the mean of the test statistic since the mean value of the test statistic is zero under H_0. This resampling scheme performs well in terms of size and power, provided that the null distribution of the bootstrap values of the test statistic is shifted to have mean zero.

1 Introduction

The investigation of the direction of the causal relationships between the variables of a multivariate system allows us to better understand its structure and improve the predictions of the variables that comprise it. When evaluating the causal influence between two variables from a multivariate system, it is necessary to take the effects of the remaining variables into account in order to distinguish between direct and indirect causal effects.

Theoretically, the causality measures should be zero if there are no causal interactions, otherwise give positive values. However, there is usually bias, which may be due to the estimation method of the test statistic, the selection of the parameters, the finite sample size, the level of the noise, etc. To determine the extent to which a positive value of a measure indicates a weak coupling or not, it is necessary to determine the statistical significance of a test statistic.

When the asymptotic distribution of a test statistic cannot be established, resampling techniques are employed for the construction of its empirical null distribution. The resampled time series should satisfy the corresponding null hypothesis, i.e. H_0: no causal effects, however they also have to capture the statistical properties of the original time series. The statistical significance of the causality tests can be assessed by resampling methods, which include bootstraps or surrogates.

Bootstrapping is a statistical technique that has been introduced by [1], that aims to estimate the properties of a test statistic when sampling from an approximating distribution. Bootstrapping is utilized when the theoretical distribution of a test statistic is not known. It provides a method to assess the properties of an estimator, such as the variance or the sampling distribution, by resampling the data from an empirical distribution. For time series, bootstraps must be carried out in a way that they suitably capture the dependence structure of the data generation process consistent to the H_0, and be otherwise random e.g. [5, 10, 9].

Surrogate statistical tests utilize surrogate data, which are modified samples of the original data, to empirically estimate the expected probability distribution of the estimator. These randomization methods preserve the dependence structure consistent to H_0 when randomly shuffling the time series [13, 11, 3].

In this paper, we perform a comparative study of different resampling methods for assessing the statistical significance of the information causality measure, partial transfer entropy. The transfer entropy (TE) is a non-parametric measure that quantifies the amount of directed transfer of information between two random processes [12]. The partial transfer entropy (PTE) is the multivariate extension of the TE [14, 7]. Since the asymptotic distribution of the PTE is not known, resampling methods are required.

The appropriateness of seven resampling methods for the null hypothesis H_0 of non-causality is examined. Specifically, we combine two resampling methods: 1) the time shifted surrogates [11] and 2) the stationary bootstrap [9], with three independence settings of the time series adapted for the non-causality test (giving six resampling methods): A) resampling only the time series of the driving variable, B) resampling independently the driving and the response time series, and C) resampling separately the driving and the response time series, while destroying the dependence of the future and past of the response variable. Further, we also consider the following resampling method: resampling contemporaneously the driving and the response time series. For the last resampling method, we only employ the method of the stationary bootstrap. In this case, the bootstrap PTE values are centered to zero since the H_0 of no causal effects is not satisfied. The properties of the PTE for the seven in total resampling methods, i.e. the empirical distribution of PTE, the size and power of the test, are assessed in a simulation study.

The structure of the paper is as follows. In Sec. 2, the PTE is briefly discussed. In Sec. 3 the seven resampling methods are presented. In Sec. 4, the resampling methods are evaluated on a simulation study using different coupled and uncoupled multivariate systems. Finally, the conclusions are drawn in Sec. 5.

2 Partial Transfer Entropy

The TE quantifies the amount of information explained in a response variable Y at h time steps ahead from the state of a driving variable X accounting for the concurrent state of Y. Let $\{x_t, y_t\}$, $t = 1, \ldots, n$ be the observed time series of two variables. We define the reconstructed vectors of the state space of the variables as $\mathbf{x}_t = (x_t, x_{t-\tau}, \ldots, x_{t-(m-1)\tau})'$ and $\mathbf{y}_t = (y_t, y_{t-\tau}, \ldots, y_{t-(m-1)\tau})'$, where m is the embedding dimension and τ is the time delay. The TE from X to Y is the conditional mutual information $I(y_{t+h}; \mathbf{x}_t|\mathbf{y}_t)$ given as

$$\begin{aligned} \mathrm{TE}_{X\to Y} &= I(y_{t+h}; \mathbf{x}_t|\mathbf{y}_t) = \sum p(y_{t+h}, \mathbf{x}_t, \mathbf{y}_t) \log \frac{p(y_{t+h}|\mathbf{x}_t, \mathbf{y}_t)}{p(y_{t+h}|\mathbf{y}_t)} \qquad (1) \\ &= H(\mathbf{x}_t, \mathbf{y}_t) - H(y_{t+h}, \mathbf{x}_t, \mathbf{y}_t) + H(y_{t+h}, \mathbf{y}_t) - H(\mathbf{y}_t), \end{aligned}$$

where TE is expressed either based on the probability distributions, $p(\cdot)$, or based on entropy terms, $H(\cdot)$, where $H(\mathbf{x}) = -\int f(\mathbf{x}) \log f(\mathbf{x}) \mathrm{d}\mathbf{x}$ is the differential entropy of the vector variable $\mathbf{x}$ with probability density function $f(\mathbf{x})$).

The partial transfer entropy (PTE) has been introduced as the multivariate extension of of transfer entropy (TE) [12] in [14, 7]. The PTE accounts for the direct coupling of X to Y conditioning on the remaining variables of a multivariate system, collectively denoted Z. It is defined as

$$\begin{aligned} \mathrm{PTE}_{X\to Y|Z} &= I(y_{t+h}; \mathbf{x}_t|\mathbf{y}_t, \mathbf{z}_t) \qquad (2) \\ &= H(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t) - H(y_{t+h}, \mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t) + H(y_{t+h}, \mathbf{y}_t, \mathbf{z}_t) - H(\mathbf{y}_t, \mathbf{z}_t). \end{aligned}$$

Different types of estimators for the TE exist, and thus for the PTE, such as histogram-based (e.g. by discretizing the state space to equidistant intervals at each axis), kernel-based and using correlation sums. Here, we use the nearest neighbor estimator [4], which is specifically effective for high-dimensional data [15].

Theoretically, the causality measures and namely the PTE should be zero in the case of no causal effects. However, a bias can be present due to various reasons, e.g. the estimation method for the entropies and subsequently densities, the selection of the embedding parameters, the finite sample size and the noise level as well [6]. In order to determine whether a PTE value indicates a weak coupling or whether it is not statistically significant, resampling methods are employed to assess its statistical significance.

3 Resampling Methods

We first present the two resampling methods that have been used here, i.e. 1) the time shifted surrogates and 2) the stationary bootstrap. Then, we discuss the three independence settings that we examine in combination with the two resampling methods. Finally, we discuss the seven resampling method, based on the stationary bootstrap, which is however the only method here that does not comply with the null hypothesis of no causal effects.

Let us consider two variables X and Y and their corresponding time series $\{x_1, \dots, x_n\}$ and $\{y_1, \dots, y_n\}$. The time shifted surrogates are generated so that they preserve the dynamics of the original time series, i.e. $\{x_1, \dots, x_n\}$, while the couplings between X and Y are destroyed [11]. They are formed by cyclically time shifting the components of the time series. To formulate them from the time series $\{x_1, \dots, x_n\}$, an integer d is randomly chosen (with $d < n$) and the d first values of the time series are moved to the end: $\{x_t^*\} = \{x_{d+1}, \dots, x_n, x_1, \dots, x_d\}$. For testing the causality relation $X \to Y$ in a bivariate time series, the pair $\{x_t^*, y_t\}$ is consistent with the H_0 of non-causality.

The stationary bootstrap has been introduced in [9] for correlated data. The block bootstrap has been used mainly with data correlated in time (i.e. time series) but can also be used with data correlated in space, or among groups (so-called cluster data). By construction, the stationary bootstrap does not destroy the time dependence of the data. This method tries to replicate the correlations by resampling blocks of data.

The bootstrap series are generated by resampling blocks of random size, where the length of each block has a geometric distribution. For a fixed probability p, block lengths L_i are generated with probability $p(L_i = k) = (1-p)^{(k-1)}$ for $k = 1, 2, \dots$. The starting time points of the blocks I_i are drawn from the discrete uniform distribution on $\{1, \dots, n-k\}$. A bootstrap time series $\{x_t^*\}$ is formed by first starting with a random block as defined above $B_{I_1,L_1} = \{x_{I_1}, x_{I_1+1}, \dots, x_{I_1+L_1-1}\}$, and blocks are added until length n is reached.

Three independence settings are considered in combination with the aforementioned resampling methods, all consistent with the H_0 of non-causality from X to Y conditioned on Z. The first setting, denoted A, is to resample only the time series of the driving variable X. This is the standard approach for surrogate test for the significance of causality measures [11, 2, 15, 8]. The intrinsic dynamics of the variable X is preserved in the resampled time series $\{x_t^*\}$ but the coupling between X^* and Y is destroyed, so that H_0 is fulfilled and $\mathrm{PTE}_{X^* \to Y|Z} = 0$. The variables X and Y as well as X and Z are independent, however the pair of variables (Y, Z) preserves its interdependence.

The second setting, denoted B, suggests to randomize both the driving variable X and the response Y, i.e. the resampled time series $\{x_t^*\}$ and $\{y_t^*\}$ are generated. Again, the intrinsic dynamics of both X and Y are preserved but the coupling between them is destroyed, H_0 is fulfilled and $\mathrm{PTE}_{X^* \to Y^*|Z} = 0$. In this case, independence holds for all variable pairs (X, Y), (Y, Z) and (X, Z). However, there is still no complete independence between all arguments in the definition of PTE, as y_{t+h} preserves by construction of $\{y_t^*\}$ its dependence on $\mathbf{y}_t$.

The third setting we consider, regards complete independence of all variables involved in the definition of PTE, denoted C, i.e. in addition to the resampling of X and Y, also y_{t+h} is resampled separately. Thus all terms in PTE, i.e. y_{t+h}, $\mathbf{x}_t$, $\mathbf{y}_t$ and $\mathbf{z}_t$ are independent, and H_0 is again fulfilled.

Combining the two resampling methods, i.e. 1) the time shifted surrogates and 2) the stationary bootstraps) and the three independence settings (A, B and C), six resampling methods are formulated that are utilized to test the null hypothesis of no causal effects among the variables of multivariate systems.

Last but not least, the stationary bootstrap is again utilized in a different scheme in order to test the non-causality hypothesis (case 3). We formulate the null distribution of the test statistic by contemporaneously resampling the driving and the response time series. Although this resampling method does not comply with H_0, one can obtain an accurate sampling distribution of the mean of the test statistic since the mean value of the test statistic is zero under H_0. The idea is that $\sqrt{n}(\mathrm{PTE} - E(\mathrm{PTE}))$ can be distributed similar for series that comply to H_0 ($E(\mathrm{PTE}) = 0$) as for series that do not comply to H_0 ($E(\mathrm{PTE}) > 0$). By centering distribution of the bootstrap PTE values around zero, we get an approximation of the null distribution of PTE. Thus, this resampling method can be employed to test H_0, provided that the null distribution of the bootstrap values of the test statistic is shifted to have mean zero.

Considering also the last resampling method where the driving and the response time series are contemporaneously resampled, we end up with seven resampling methods in total that are examined here.

4 Simulation study

We apply the significance test for the PTE with the seven different resampling schemes to multiple realizations of different simulation systems. Specifically, we estimate the PTE from 1000 realizations per simulation system. For each realization and each resampling scheme, $M = 100$ resampled time series are generated. Let us denote q_0 the PTE value from one realization of a system and $q_1, q_2, \dots, q_M$ the PTE values from the resampled time series for this particular realization and for a specific resampling scheme. The rejection of H_0 of no causal effects is decided by the rank ordering of the PTE values computed on the original time series, q_0, and the resampled time series, $q_1, q_2, \dots, q_M$. For the one-sided test, if r_0 is the rank of q_0 when ranking the list $q_0, q_1, \dots, q_M$ in ascending order, the p-value of the test is $1 - (r_0 - 0.326)/(M + 1 + 0.348)$, by applying the correction in [17].

The simulation systems we considered in this study are the following:

1. Three coupled Hénon maps, with nonlinear couplings ($X_1 \to X_2$, $X_2 \to X_3$)

$$\begin{aligned}
x_{1,t} &= 1.4 - x_{1,t-1}^2 + 0.3x_{1,t-2} \\
x_{2,t} &= 1.4 - cx_{1,t-1}x_{2,t-1} - (1-c)x_{2,t-1}^2 + 0.3x_{2,t-2} \\
x_{3,t} &= 1.4 - cx_{2,t-1}x_{3,t-1} - (1-c)x_{3,t-1}^2 + 0.3x_{3,t-2},
\end{aligned}$$

   with equal coupling strengths c for $X_1 \to X_2$ and $X_2 \to X_3$. We set $c = 0$ (uncoupled case), $c = 0.3$ and $c = 0.5$ (strong coupling). We note that the time series of this system become completely synchronized for coupling strengths $c \geq 0.7$.

2. A vector autoregressive process of 4 variables and order 5, VAR(5), with linear couplings ($X_1 \to X_3$, $X_2 \to X_1$, $X_2 \to X_3$, $X_4 \to X_2$)

$$x_{1,t} = 0.8x_{1,t-1} + 0.65x_{2,t-4} + \epsilon_{1,t}$$

$$
\begin{aligned}
x_{2,t} &= 0.6x_{2,t-1} + 0.6x_{4,t-5} + \epsilon_{2,t} \\
x_{3,t} &= 0.5x_{3,t-3} - 0.6x_{1,t-1} + 0.4x_{2,t-4} + \epsilon_{3,t} \\
x_{4,t} &= 1.2x_{4,t-1} - 0.7x_{4,t-2} + \epsilon_{4,t},
\end{aligned}
$$

where $\epsilon_{i,t}$, $i = 1, \dots, 4$, are independent to each other Gaussian white noise processes with unit standard deviation (Eq.(12) in [16]).

3. Five coupled Hénon maps, with nonlinear couplings ($X_1 \to X_2$, $X_2 \to X_3$, $X_3 \to X_4$, $X_4 \to X_5$) defined similarly to system 1. We consider again equal coupling strengths c, and set $c = 0$ (uncoupled case), $c = 0.2$ and $c = 0.4$ (strong coupling).

We consider the time series lengths $n = 512$ and 2048. To estimate the PTE, we set the embedding dimension m to appropriate values for each system, i.e. $m = 2$ for system 1 and 3, $m = 5$ for system 2, the delay time $\tau = 1$ and the time step ahead $h = 1$ (as defined in [12]). The number of nearest neighbors for the estimation of the probability distributions is 10.

In terms of presentation, we focus on the sensitivity of the PTE, i.e the percentage of rejection of H_0 when there is true direct causality, as well as the specificity of the PTE, i.e. the percentage of no rejection of H_0 when there is no direct causality, at the significance level $\alpha = 0.05$. The notation $X_2 \to X_1|Z$ denotes the Granger causality from X_2 to X_1, accounting for the presence of confounding variables $Z = X_3, \dots$. For brevity, we use the notation $X_2 \to X_1$ instead of $X_2 \to X_1|Z$, implying the conditioning on the confounding variables. The notation of Granger causality for other pairs of variables is analogous.

System 1. The mean PTE values are negatively biased in the uncoupled case ($c = 0$) (see Table 1). For $c = 0.3$ and $c = 0.5$, they are much larger when direct couplings exist ($X_1 \to X_2$, $X_2 \to X_3$) than the rest of the directions, and increase with n. Regarding the indirect coupling $X_1 \to X_3$, the PTE slightly increases with n as the coupling strength increases, reaching the highest mean value for $c = 0.5$ (mean $PTE_{X_1 \to X_3} = 0.0004$ for $n = 512$ and $PTE_{X_1 \to X_3} = 0.0071$ for $n = 2048$).

We evaluate how the null distribution of the PTE from the seven resampling schemes differs with respect to the original PTE values. For $c = 0$, all the resampling schemes correctly indicate the absence of couplings (see Table 2). Considering $c = 0.3$, the couplings are denoted by all cases, however also spurious and indirect ones are also indicated for the settings A and B.

Table 1: Mean PTE values from 1000 realizations of system 1 for $n = 512$ / 2048. A single number is displayed when the same percentage corresponds to both n. The mean PTE values at the directions of the true couplings are highlighted.

$n = 512$	$X_1 \to X_2$	$X_2 \to X_1$	$X_2 \to X_3$	$X_3 \to X_2$	$X_1 \to X_3$	$X_3 \to X_1$
$c = 0$	-0.0059	-0.0062	-0.0062	-0.0061	-0.0058	-0.0056
$c = 0.3$	**0.0802**	-0.0042	**0.0885**	-0.0064	-0.0045	-0.0074
$c = 0.5$	**0.2324**	-0.0071	**0.1557**	-0.0044	0.0004	-0.0079
$n = 2048$	$X_1 \to X_2$	$X_2 \to X_1$	$X_2 \to X_3$	$X_3 \to X_2$	$X_1 \to X_3$	$X_3 \to X_1$
$c = 0$	-0.0086	-0.0088	-0.0087	-0.0085	-0.0087	-0.0088
$c = 0.3$	**0.1736**	-0.0024	**0.1725**	-0.0059	-0.0039	-0.0094
$c = 0.5$	**0.3649**	-0.0026	**0.2601**	-0.0049	0.0071	-0.0078

Similar performance is observed also in the case of strong coupling strength ($c = 0.5$). In this case, large percentages are obtained for the indirect coupling $X_1 \to X_3$.

Concerning the first six resampling methods, the percentage of erroneously rejected H_0 for non-existing or indirect couplings tends to increase with c and the time series length for all resampling schemes, the most robust being 1C and 2C. It turns out that when the resampled time series become more independent (case A to case C), the percentage of spurious couplings decreases. The most independent resampling schemes 1C and 2C give smallest rejection rate since the null distribution for the test is more spread and displaced to the right as the resampling changes from the least independent scheme (scheme A) to the most independent one (C) (e.g. see Fig. 1). Case 3 seems to be the most effective one. It does not require large time series length to give high percentage of rejection of H_0 at the directions of the true couplings, while the lowest percentages of significant PTE values are estimated at the uncoupled directions.

System 2. The mean PTE values from 1000 realizations of the second system for the directions of the true couplings are larger than for the other directions and increase with n, with the exception of $X_2 \to X_3$ that is at a lower level and does not significantly increase with n (see Table 4). Concerning the uncoupled directions, the mean PTE values vary from 0.0014 to 0.0097 for both n, while they decrease with n.

The true couplings $X_2 \to X_1$, $X_1 \to X_3$, $X_4 \to X_2$ are well established by the significance test (see Table 4). No spurious causalities are identified by

Table 2: Percentage of significant PTE values for system 1 for $n = 512$ / 2048, for all resampling schemes.

$c = 0$	$X_1 \to X_2$	$X_2 \to X_1$	$X_2 \to X_3$	$X_3 \to X_2$	$X_1 \to X_3$	$X_3 \to X_1$
1A	5.7 / 4.2	5.6 / 5.2	4.7 / 4.9	5.3 / 5.6	5.8 / 5.5	5.5 / 5.2
1B	5.2 / 4.8	4.6 / 5.6	4 / 5.2	4.3 / 6.6	4.6 / 5	5.8 / 5.5
1C	0.7 / 0	0.8 / 0	0.4 / 0	0.7 / 0	0.3 / 0	0.5 / 0
2A	4.4 / 3.8	3.1 / 3.9	3.4 / 4.1	4.5 / 4.5	4.5 / 4.3	4.1 / 5.1
2B	1.9 / 0.4	1.9 / 0.7	1.8 / 0.6	2.1 / 0.3	1.9 / 0.5	2.4 / 1
2C	0.6 / 0	0.6 / 0	0.3 / 0	0.5 / 0	0.4 / 0	0.1 / 0
3	0.6 / 0	0.7 / 0	0.3 / 0	0.7 / 0	0.2 / 0	0.2 / 0
$c = 0.3$	$\mathbf{X_1 \to X_2}$	$X_2 \to X_1$	$\mathbf{X_2 \to X_3}$	$X_3 \to X_2$	$X_1 \to X_3$	$X_3 \to X_1$
1A	**100**	11.8/ 40.1	**100**	9.5 / 17.2	12.8/ 34	6.1 / 5.5
1B	**100**	9 / 37.2	**100**	2.7 / 1.8	5.4 / 6.7	5 / 4.3
1C	**100**	0.9 / 0.5	**86.3** / **100**	0	0.2 / 0	0.4 / 0.1
2A	**100**	8.7 / 32.8	**100**	6.9 / 13.5	8.9 / 28	4.7 / 4.1
2B	**100**	2.9 / 13.7	**100**	0.9 / 0.3	1.2 / 1.7	1.2 / 0.5
2C	**100**	0.8 / 0.6	**99.9** / **100**	0	0 / 0.1	0.3 / 0.1
3	**100**	1.2 / 0.9	**100**	0.1 / 0	0.2 / 0.3	0.4 / 0.1
$c = 0.5$	$\mathbf{X_1 \to X_2}$	$X_2 \to X_1$	$\mathbf{X_2 \to X_3}$	$X_3 \to X_2$	$X_1 \to X_3$	$X_3 \to X_1$
1A	**100**	8.1 / 33.8	**100**	10.2 / 21.5	31 / 96.3	6.2 / 8.3
1B	**100**	4.3 / 30.4	**100**	1.7 / 1.4	9.1 / 67.3	4.5 / 4.8
1C	**100**	0.7 / 0.4	**100**	0	1.9 / 25.4	0.1
2A	**100**	5.1 / 29.2	**100**	7.7 / 17	24.1 / 94.7	4 / 7.1
2B	**100**	2 / 11	**100**	0.8 / 0.2	5.2 / 53.3	1.3 / 0.8
2C	**100**	0 / 0.2	**100**	0	1.2 / 24.3	0 / 0.1
3	**100**	0.2 / 0.6	**100**	0 / 0.1	1.4 / 11.6	0.1

Table 3: Mean PTE values from 1000 realizations of system 2.

	$X_1 \to X_2$	$\mathbf{X_2 \to X_1}$	$\mathbf{X_1 \to X_3}$	$X_3 \to X_1$	$X_1 \to X_4$	$X_4 \to X_1$
$n = 512$	0.0044	**0.0914**	**0.0757**	0.0032	0.0057	0.0038
$n = 2048$	0.0026	**0.1232**	**0.0960**	0.0014	0.0038	0.0021
	$\mathbf{X_2 \to X_3}$	$X_3 \to X_2$	$X_2 \to X_4$	$\mathbf{X_4 \to X_2}$	$X_3 \to X_4$	$X_4 \to X_3$
$n = 512$	**0.0056**	0.0052	0.0097	**0.1002**	0.0069	0.0033
$n = 2048$	**0.0058**	0.0029	0.0064	**0.1348**	0.0045	0.0014

the first six resampling methods (percentage of significant PTE vary from 0% to 6% at the uncoupled directions), however case 3 indicates the couplings $X_2 \to X_4$ and $X_3 \to X_4$, giving much higher percentage than the nominal size 5%. The weak coupling $X_2 \to X_3$ is detected only by the scheme A (1A

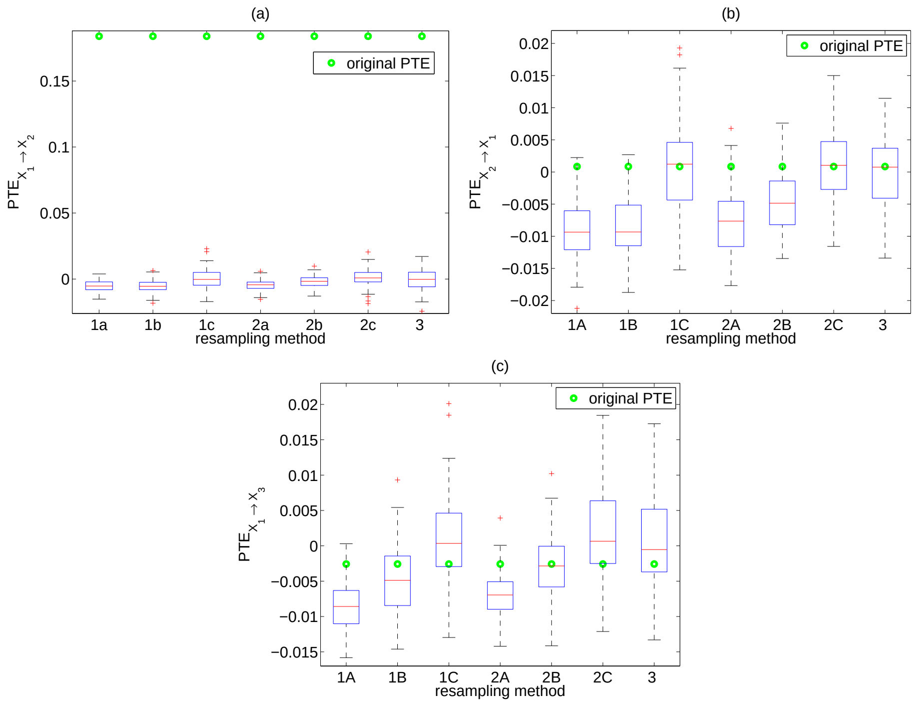

Figure 1: Boxplots of surrogate / bootstrap PTE values and original PTE value from one realization of system 1 for $c = 0.3$ and $n = 2048$, for the directions (a) $X_1 \to X_2$ (direct coupling), (b) $X_2 \to X_1$ (no coupling) and (c) $X_1 \to X_3$ (indirect coupling).

and 2A), with a power of the test increasing with n.

We note that the surrogate / bootstrap PTE values increase as the resampled time series become more independent (see Figure 2 and Figure 3). The bootstrap PTE values are centered around zero by construction, while on the other hand for the six first resampling methods, the surrogate / bootstrap PTE values are positively biased. The distribution of the the surrogate / bootstrap PTE values becomes wider as the resampling method gets more independent (case A to case C), with case 3 having the the wider one. The failure of the resampling methods B and C to detect the coupling $X_2 \to X_3$ is due to the fact that both the original PTE value and the surrogate / bootstrap PTE values are positively biased. Case 3 performs poorly because the distribution of the bootstrap PTE values is much wider compared to the other cases and the original PTE value falls within the tail of this distribution

Table 4: Percentage of significant PTE values from 1000 realizations of system 2 for $n = 512$ / 2048.

	$X_1 \to X_2$	$\mathbf{X_2 \to X_1}$	$\mathbf{X_1 \to X_3}$	$X_3 \to X_1$	$X_1 \to X_4$	$X_4 \to X_1$
1A	0.4 / 0	**100**	**100**	0.6 / 0.3	0.1 / 0	4.6 / 3.2
1B	0	**100**	**99.4** / **100**	0	0	0
1C	0	**100**	**100**	0	0	0
2A	0.4 / 0	**100**	**100**	0.5	0.1 / 0	2.8 / 3.7
2B	0	**100**	**100**	0	0.2 / 0	0 / 0
2C	0	**100**	**99.7** / **100**	0	0	0
3	2.3 / 3.8	**100**	**100**	0.8 / 0.5	8.2 / 15.7	1.9 / 1.8
	$\mathbf{X_2 \to X_3}$	$X_3 \to X_2$	$X_2 \to X_4$	$\mathbf{X_4 \to X_2}$	$X_3 \to X_4$	$X_4 \to X_3$
1A	**18.8** / **62.4**	1.1 / 0.2	3.5 / 2	**100**	0.8 / 0	6 / 4.2
1B	**0** / **0.1**	0	2.1 / 1.3	**99.9** / **100**	0.4 / 0	0
1C	**3.7** / **10.1**	0	0	**100**	0	0.9 / 0
2A	**11.7** / **60.1**	0.6 / 0.1	2.6 / 3.2	**100**	0.4 / 0	3.1
2B	**0** / **0.2**	0	3.1	**100**	0.8 / 0.1	0
2C	**3.4** / **18.5**	0	0	**100**	0	0.2 / 0
3	**2.7** / **24**	4.7 / 6.7	21.6 / 37.8	**100**	15.7 / 24.2	0.8 / 0.6

(see Figure 3g).

System 3. No couplings are noted in the uncoupled case ($c = 0$) for system 3 (see Table 5); the percentage of significant PTE values range from 0% to 5.6% for all the resampling schemes and both time series lengths. The PTE is also effective when couplings are present. For $c = 0.2$, the sensitivity of PTE increases with n. When strong couplings exist, the percentage of significant PTE values is not that sensitive to the time series length as for $c = 0.2$.

As resampled time series become more independent, a loss in the power of the test is observed for $n = 512$, especially when couplings are not very strong. On the other hand, when $c = 0.4$, indirect and spurious couplings are observed, especially for $n = 2048$ and mainly for the resampling scheme. For example, we obtain for scheme 1A and $n = 2048$: 50.5% for $X_1 \to X_3$ (indirect coupling), 22.2% for $X_2 \to X_1$ (spurious coupling), 56.8% for $X_2 \to X_4$ (indirect coupling), 19.7% for $X_3 \to X_2$ (spurious coupling), 62.2% for $X_3 \to X_5$ (indirect coupling), 22.9% for $X_4 \to X_3$ (spurious coupling) and 14.1% for $X_5 \to X_4$ (spurious coupling). Similar results are indicated by scheme 2A. Considering more independent resampled time series, the corresponding percentages of indirect and spurious couplings decrease, e.g.

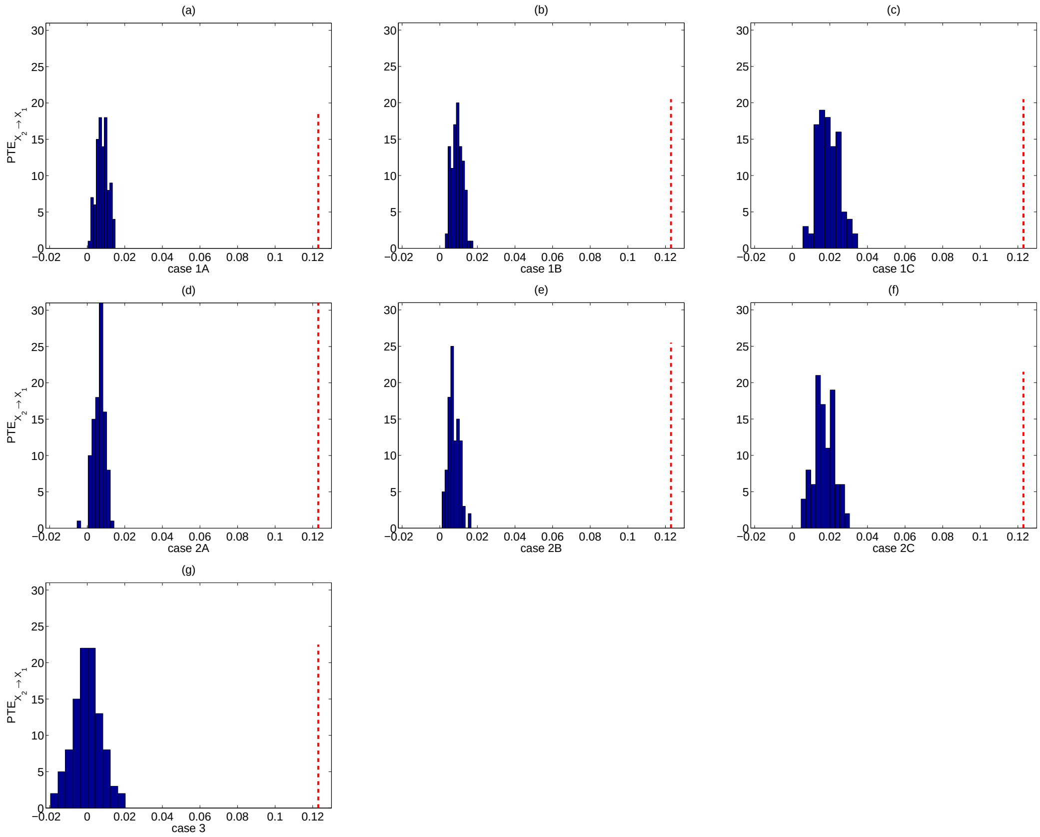

Figure 2: Distribution of surrogate / bootstrap PTE values and original PTE value (vertical dotted line) from one realization of system 2 with $n = 2048$, for the direction $X_2 \to X_1$.

for scheme 1B and $n = 2048$: 27.5% for $X_1 \to X_3$, 20% for $X_2 \to X_1$, 21.4% for $X_2 \to X_4$, 3.7% for $X_3 \to X_2$, 28% for $X_3 \to X_5$, 4.1% for $X_4 \to X_3$ and 4.7% for $X_5 \to X_4$. Similar results are observed for scheme 2A. Only the true couplings are indicated using the resampling methods C; the percentages of the significant PTE values for the uncoupled cases vary from 0% to 4.7% for both schemes 1C and 2C and both n.

The PTE performs equivalently to case C when considering the resampling scheme 3. All the true couplings are denoted, while the percentages of rejecting H_0 in the uncoupled directions range from 0% to 15.4% for both n; the three highest percentages of significant PTE values in case of the uncoupled directions are obtained for $n = 2048$ and $X_1 \to X_3$ (5.8%), $X_2 \to X_4$

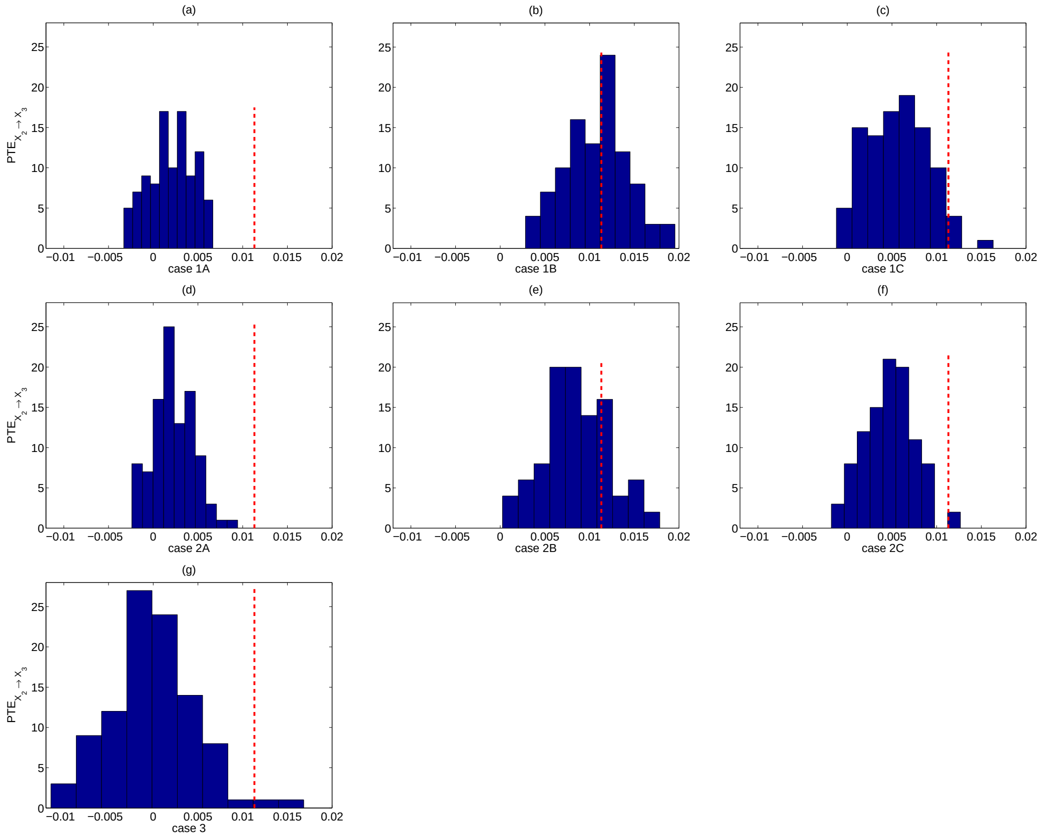

Figure 3: As Fig. 2 but for the direction $X_2 \rightarrow X_3$.

(9.4%) and $X_3 \rightarrow X_5$ (15.4%).

5 Discussion

The importance of assessing the correct statistical significance for the partial transfer entropy (PTE) has been explored in a simulation study. Concerning the resampled time series, by definition, the mutual information of X and Y conditioned on Z should be in theory zero, i.e. $I(Y;X|Z) = 0$. The formulation of more independent resampled data (schemes B and C) compared to the standard technique (scheme A) seems to improve the bias of the test statistic and helps prevent false indications of couplings in the case of the nonlinear coupled systems. The size and the power of the test are improved,

Table 5: Percentage of significant PTE values from 1000 realizations of system 3 for $n = 512$ / 2048, for the true couplings, an indirect coupling ($X_2 \rightarrow X_4$) and an uncoupled case ($X_5 \rightarrow X_4$).

$c = 0$	$X_1 \rightarrow X_2$	$X_2 \rightarrow X_3$	$X_3 \rightarrow X_4$	$X_4 \rightarrow X_5$	$X_2 \rightarrow X_4$	$X_5 \rightarrow X_4$
1A	4.5 / 5.1	5.8 / 5.6	5.6 / 5.4	5.4 / 4.1	4.9 / 4.6	3.8 / 4.8
1B	4.5 / 4.3	5.8 / 5.6	5.9 / 5.5	5.2 / 4.8	4.9 / 4.5	3.9 / 4.4
1C	1.9 / 0.6	2 / 0.5	2.2 / 0.5	2.1 / 0.5	2.2 / 0.4	1.5 / 0.6
2A	4.4 / 4.3	4.8 / 5.5	5.1 / 4.8	4.9 / 4.2	4.7 / 4.4	4.3
2B	3.3 / 2.6	3.6 / 3.3	3.7 / 2.9	3 / 2.9	2.9 / 2.3	3.2 / 3
2C	1 / 0.7	1.4 / 0.3	1.5 / 0.5	1.3 / 0.2	2.1 / 0.6	0.9 / 0.4
3	1.9 / 0.8	2.4 / 0.9	2.3 / 0.9	1.6 / 0.8	1.7 / 0.6	1.4 / 0.7
$c = 0.2$	$\mathbf{X_1} \rightarrow \mathbf{X_2}$	$\mathbf{X_2} \rightarrow \mathbf{X_3}$	$\mathbf{X_3} \rightarrow \mathbf{X_4}$	$\mathbf{X_4} \rightarrow \mathbf{X_5}$	$X_2 \rightarrow X_4$	$X_5 \rightarrow X_4$
1A	**58** / **100**	**51.8** / **100**	**57** / **100**	**52.7** / **100**	6.5 / 6.6	8.1 / 10.8
1B	**57.5** / **100**	**50.6** / **100**	**54.5** / **100**	**49.2** / **100**	4.9	5.6 / 7
1C	**34.3** / **100**	**17.5** / **100**	**18.9** / **100**	**16.6** / **100**	0.5 / 0	0.5 / 0
2A	**57.1** / **100**	**56.9** / **100**	**62.1** / **100**	**57** / **100**	6.2 / 5.3	8.7 / 11.3
2B	**49.8** / **100**	**52.1** / **100**	**58.1** / **100**	**52.2** / **100**	4.9 / 2.4	6.2 / 4.2
2C	**30.6** / **100**	**24.2** / **99.8**	**26** / **99.9**	**24.3** / **99.8**	0.5 / 0	1 / 0.1
3	**31.3** / **100**	**34.4** / **100**	**38.9** / **100**	**33.5** / **100**	3.2 / 0.8	3.4 / 0.8
$c = 0.4$	$\mathbf{X_1} \rightarrow \mathbf{X_2}$	$\mathbf{X_2} \rightarrow \mathbf{X_3}$	$\mathbf{X_3} \rightarrow \mathbf{X_4}$	$\mathbf{X_4} \rightarrow \mathbf{X_5}$	$X_2 \rightarrow X_4$	$X_5 \rightarrow X_4$
1A	**100**	**99.7** / **100**	**99.8** / **100**	**99.4** / **100**	14 / 56.8	14.1 / 23
1B	**100**	**99.8** / **100**	**99.6** / **100**	**99.1** / **100**	5.9 / 21.4	5 / 4.7
1C	**100**	**85.2** / **100**	**87.7** / **100**	**84** / **100**	0.4 / 0.6	0.8 / 0.2
2A	**100**	**99.9** / **100**	**100**	**99.8** / **100**	18.2 / 59.5	16.9 / 25
2B	**100**	**99.9** / **100**	**99.9** / **100**	**99.8** / **100**	11 / 26.3	9.6 / 6.4
2C	**99.8** / **100**	**97.1** / **100**	**97.6** / **100**	**95.1** / **100**	1.5 / 2.7	2.4 / 0.3
3	**99.8** / **100**	**99.1** / **100**	**99.1** / **100**	**98.5** / **100**	5.1 / 9.4	5.1 / 2.2

especially if strong couplings exist. However, when the couplings are linear, scheme A seems to be more efficient in identifying weak couplings. Scheme 3 seems to also effective in case of the nonlinear simulation systems, however does not seem to be effective in case of linear coupled systems as spurious couplings are denoted.

It turns out that when the PTE is estimated for an increasing level of randomness in the resampled time series, the estimated PTE values also increase, while the distribution of PTE from the resampled time series gets wider and less spurious couplings are thus detected. This higher specificity comes at the cost of lower sensitivity, and vice versa. Thus, none of the first six resampling schemes turns out to be optimal, but it becomes clear that the significance test for the PTE gets more conservative as resampling is more random. Regarding case 3, the bootstrap PTE values are centered by construction around zero. Since the original PTE value is biased, spurious indications may appear.

The aforementioned resampling schemes can be utilized for any test statistic in order to examine the null hypothesis of no causal effects. Since the efficiency of a causality measure is determined in terms of the corresponding resampling technique that is used, the usefulness of each of the examined resampling schemes will be further investigated for different causality measures.

Acknowledgements

The research project is implemented within the framework of the Action "Supporting Postdoctoral Researchers" of the Operational Program "Education and Lifelong Learning" (Action's Beneficiary: General Secretariat for Research and Technology), and is co-financed by the European Social Fund (ESF) and the Greek State.

References

[1] B. Efron. Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7(1):1–26, 1979.

[2] L. Faes, A. Porta, and G. Nollo. Mutual nonlinear prediction as a tool to evaluate coupling strength and directionality in bivariate time series: comparison among different strategies based on k nearest neighbors. *Physical Review E*, 78:026201, 2008.

[3] L. Faes, A. Porta, and G. Nollo. Testing frequency-domain causality in multivariate time series. *IEEE Transanctions on Biomedical Engineering*, 57(8):1897–1906, 2010.

[4] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical Review E*, 69(6):066138, 2004.

[5] H.R. Künsch. The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, 17:1217–1241, 1989.

[6] A. Papana, D. Kugiumtzis, and P.G. Larsson. Reducing the bias of causality measures. *Physical Review E*, 83:036207, 2011.

[7] A. Papana, D. Kugiumtzis, and P.G. Larsson. Detection of direct causal effects and application to electroencephalogram analysis. *International Journal of Bifurcation and Chaos*, 22(9):1250222, 2012.

[8] A. Papana, C. Kyrtsou, D. Kugiumtzis, and C. Diks. Simulation study of direct causality measures in multivariate time series. *Entropy*, 15(7):2635–2661, 2013.

[9] D. Politis and J.P. Romano. The stationary bootstrap. *Journal of Americal Statistical Association*, 89(428):1303–1313, 1994.

[10] D. Politis, J.P. Romano, and T. Lai. Bootstrap confidence bands for spectra and cross-spectra. *IEEE Transanctions on Signal Processing*, 40:1206–1215, 1992.

[11] R. Quian Quiroga, A. Kraskov, T. Kreuz, and P. Grassberger. Performance of different synchronization measures in real data: A case study on electroencephalographic signals. *Physical Review E*, 65:041903, 2002.

[12] T. Schreiber. Measuring information transfer. *Physical Review Letter*, 85(2):461–464, 2000.

[13] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J.D. Farmer. Testing for nonlinearity in time series: the method of surrogate data. *Physica D*, 58:77–94, 1992.

[14] V.A. Vakorin, O.A. Krakovska, and A.R. McIntosh. Confounding effects of indirect connections on causality estimation. *Journal of Neuroscience Methods*, 184:152–160.

[15] I. Vlachos and D. Kugiumtzis. Non-uniform state space reconstruction and coupling detection. *Physical Review E*, 82:016207, 2010.

[16] M. Winterhalder, B. Schelter, W. Hesse, K. Schwab, L. Leistritz, D. Klan, R. Bauer, J. Timmer, and H. Witte. Comparison of linear signal processing techniques to infer directed interactions in multivariate neural systems. *Signal Proccessing*, 85(11):2137–2160, 2005.

[17] G.H. Yu and C.C. Huang. A distribution free plotting position. *Stochastic Environmental Research and Risk Assessment*, 15(6):462–476, 2001.