Entropy **2015**, *17*, 4173-4201; doi:10.3390/e17064173

OPEN ACCESS
entropy
ISSN 1099-4300
www.mdpi.com/journal/entropy

Article

Contribution to Transfer Entropy Estimation via the k-Nearest-Neighbors Approach

Jie Zhu [1,2,3], Jean-Jacques Bellanger [1,2], Huazhong Shu [3,4] and Régine Le Bouquin Jeannès [1,2,3,*]

[1] Institut National de la Santé Et de la Recherche Médicale (INSERM), U 1099, Rennes F-35000, France; E-Mails: jie.zhu.1@etudiant.univ-rennes1.fr (J.Z.); jean-jacques.bellanger@univ-rennes1.fr (J.-J.B.)
[2] Université de Rennes 1, LTSI, Rennes F-35000, France
[3] Centre de Recherche en Information Biomédicale sino-français (CRIBs), Rennes F-35000, France
[4] Laboratory of Image Science and Technology (LIST), School of Computer Science and Engineering, Southeast University, Nanjing 210018, China; E-Mail: shu.list@seu.edu.cn

* Author to whom correspondence should be addressed; E-Mail: regine.le-bouquin-jeannes@univ-rennes1.fr; Tel.: +33-2-23236919; Fax.: +33-2-23236917.

Academic Editor: Deniz Gencaga

Received: 31 December 2014 / Accepted: 10 June 2015 / Published: 16 June 2015

Abstract: This paper deals with the estimation of transfer entropy based on the k-nearest neighbors (k-NN) method. To this end, we first investigate the estimation of Shannon entropy involving a rectangular neighboring region, as suggested in already existing literature, and develop two kinds of entropy estimators. Then, applying the widely-used error cancellation approach to these entropy estimators, we propose two novel transfer entropy estimators, implying no extra computational cost compared to existing similar k-NN algorithms. Experimental simulations allow the comparison of the new estimators with the transfer entropy estimator available in free toolboxes, corresponding to two different extensions to the transfer entropy estimation of the Kraskov–Stögbauer–Grassberger (KSG) mutual information estimator and prove the effectiveness of these new estimators.

Keywords: entropy estimation; k nearest neighbors; transfer entropy; bias reduction

1. Introduction

Transfer entropy (TE) is an information-theoretic statistic measurement, which aims to measure an amount of time-directed information between two dynamical systems. Given the past time evolution of a dynamical system $\mathcal{A}$, TE from another dynamical system $\mathcal{B}$ to the first system $\mathcal{A}$ is the amount of Shannon uncertainty reduction in the future time evolution of $\mathcal{A}$ when including the knowledge of the past evolution of $\mathcal{B}$. After its introduction by Schreiber [1], TE obtained special attention in various fields, such as neuroscience [2–8], physiology [9–11], climatology [12] and others, such as physical systems [13–17].

More precisely, let us suppose that we observe the output $X_i \in \mathbb{R}$, $i \in \mathbb{Z}$, of some sensor connected to $\mathcal{A}$. If the sequence X is supposed to be an m-th order Markov process, *i.e.*, if considering subsequences $X_i^{(k)} = (X_{i-k+1}, X_{i-k+2}, \cdots, X_i)$, $k > 0$, the probability measure $\mathcal{P}_X$ (defined on measurable subsets of real sequences) attached to X fulfills the m-th order Markov hypothesis:

$$\forall i : \forall m' > m : \mathrm{d}\mathcal{P}_{X_{i+1}|X_i^{(m)}}\left(x_{i+1}|x_i^{(m)}\right) = \mathrm{d}\mathcal{P}_{X_{i+1}|X_i^{(m')}}\left(x_{i+1}|x_i^{(m')}\right), x_{i+1} \in \mathbb{R}, x_i^{(k)} \in \mathbb{R}^k, \quad (1)$$

then the past information $X_i^{(m)}$ (before time instant $i+1$) is sufficient for a prediction of X_{i+k}, $k \geq 1$, and can be considered as an m-dimensional state vector at time i (note that, to know from X the hidden dynamical evolution of $\mathcal{A}$, we need a one-to-one relation between $X_i^{(m)}$ and the physical state of $\mathcal{A}$ at time i). For the sake of clarity, we introduce the following notation: $\left(X_i^p, X_i^-, Y_i^-\right)$, $i = 1, 2, \ldots, N$, is an independent and identically distributed (IID) random sequence, each term following the same distribution as a random vector $(X^p, X^-, Y^-) \in \mathbb{R}^{1+m+n}$ whatever i (in X^p, X^-, Y^-, the upper indices "p" and "-" correspond to "predicted" and "past", respectively). This notation will substitute for the notation $\left(X_{i+1}, X_i^{(m)}, Y_i^{(n)}\right)$, $i = 1, 2, \ldots, N$, and we will denote by $\mathcal{S}_{X^p,X^-,Y^-}$, $\mathcal{S}_{X^p,X^-}$, $\mathcal{S}_{X^-,Y^-}$ and $\mathcal{S}_{X^-}$ the spaces in which (X^p, X^-, Y^-), (X^p, X^-), (X^-, Y^-) and X^- are respectively observed.

Now, let us suppose that a causal influence exists from $\mathcal{B}$ on $\mathcal{A}$ and that an auxiliary random process $Y_i \in \mathbb{R}$, $i \in \mathbb{Z}$, recorded from a sensor connected to $\mathcal{B}$, is such that, at each time i and for some $n > 0$, $Y_i^- \triangleq Y_i^{(n)}$ is an image (not necessarily one-to-one) of the physical state of $\mathcal{B}$. The negation of this causal influence implies:

$$\forall\,(m > 0, n > 0) : \forall i : \mathrm{d}\mathcal{P}_{X_i^p|X_i^{(m)}}\left(x_i^p|x_i^{(m)}\right) = \mathrm{d}\mathcal{P}_{X_i^p|X_i^{(m)},Y_i^{(n)}}\left(x_i^p|x_i^{(m)}, y_i^{(n)}\right). \quad (2)$$

If Equation (2) holds, it is said that there is an absence of information transfer from $\mathcal{B}$ to $\mathcal{A}$. Otherwise, the process X can be no longer considered strictly a Markov process. Let us suppose the joint process (X, Y) is Markovian, *i.e.*, there exist a given pair (m', n'), a transition function f and an independent random sequence e_i, $i \in \mathbb{Z}$, such that $[X_{i+1}, Y_{i+1}]^T = f\left(X_i^{(m\prime)}, Y_i^{(n')}, e_{i+1}\right)$, where the random variable e_{i+1} is independent of the past random sequence (X_j, Y_j, e_j), $j \leq i$, whatever i. As $X_i = g\left(X_i^{(m)}, Y_i^{(n)}\right)$ where g is clearly a non-injective function, the pair $\left\{\left(X_i^{(m)}, Y_i^{(n)}\right), X_i\right\}$, $i \in \mathbb{Z}$, corresponds to a hidden Markov process, and it is well known that this observation process is not generally Markovian.

The deviation from this assumption can be quantified using the Kullback pseudo-metric, leading to the general definition of TE at time i:

$$\mathrm{TE}_{Y\to X,i} = \int_{\mathbb{R}^{m+n+1}} \log\left[\frac{\mathrm{d}\mathcal{P}_{X_i^p|X_i^-,Y_i^-}\left(x_i^p|x_i^-, y_i^-\right)}{\mathrm{d}\mathcal{P}_{X_i^p|X_i^-}\left(x_i^p|x_i^-\right)}\right] \mathrm{d}\mathcal{P}_{X_i^p,X_i^-,Y_i^-}\left(x_i^p, x_i^-, y_i^-\right), \quad (3)$$

where the ratio in Equation (3) corresponds to the Radon–Nikodym derivative [18,19] (*i.e.*, the density) of the conditional measure $\mathrm{d}\mathcal{P}_{X_i^p|X_i^-,Y_i^-}\left(\cdot|x_i^-,y_i^-\right)$ with respect to the conditional measure $\mathrm{d}\mathcal{P}_{X_i^p|X_i^-}\left(\cdot|x_i^-\right)$. Considering "log" as the natural logarithm, information is measured in natural units (nats). Now, given two observable scalar random time series X and Y with no *a priori* given model (as is generally the case), if we are interested in defining some causal influence from Y to X through TE analysis, we must specify the dimensions of the past information vectors X^- and Y^-, *i.e.*, m and n. Additionally, even if we impose them, it is not evident that all of the coordinates in $X_i^{(m)}$ and $Y_i^{(n)}$ will be useful. To deal with this issue, variable selection procedures have been proposed in the literature, such as uniform and non-uniform embedding algorithms [20,21].

If the joint probability measure $\mathcal{P}_{X_i^p,X_i^-,Y_i^-}\left(x_i^p,x_i^-,y_i^-\right)$ is derivable with respect to the Lebesgue measure μ^{n+m+1} in $\mathbb{R}^{1+m+n}$ (*i.e.*, if $\mathcal{P}_{X_i^p,X_i^-,Y_i^-}$ is absolutely continuous with respect to μ^{n+m+1}), then the pdf (joint probability density function) $p_{X_i^p,X_i^-,Y_i^-}\left(x_i^p,x_i^-,y_i^-\right)$ and also the pdf for each subset of $\left\{X_i^p,X_i^-,Y_i^-\right\}$ exist, and $\mathrm{TE}_{Y\to X,i}$ can then be written (see Appendix A):

$$\begin{aligned}\mathrm{TE}_{Y\to X,i} = \; & -E\left[\log\left(p_{X_i^-,Y_i^-}\left(X_i^-,Y_i^-\right)\right)\right] - E\left[\log\left(p_{X_i^p,X_i^-}\left(X_i^p,X_i^-\right)\right)\right] \\ & +E\left[\log\left(p_{X_i^p,X_i^-,Y_i^-}\left(X_i^p,X_i^-,Y_i^-\right)\right)\right] + E\left[\log\left(p_{X_i^-}\left(X_i^-\right)\right)\right]\end{aligned} \tag{4}$$

or:

$$\mathrm{TE}_{Y\to X,i} = \mathcal{H}\left(X_i^-,Y_i^-\right) + \mathcal{H}\left(X_i^p,X_i^-\right) - \mathcal{H}\left(X_i^p,X_i^-,Y_i^-\right) - \mathcal{H}\left(X_i^-\right), \tag{5}$$

where $\mathcal{H}(U)$ denotes the Shannon differential entropy of a random vector U. Note that, if the processes Y and X are assumed to be jointly stationary, for any real function $g:\mathbb{R}^{m+n+1}\to\mathbb{R}$, the expectation $E\left[g\left(X_{i+1},X_i^{(m)},Y_i^{(n)}\right)\right]$ does not depend on i. Consequently, $\mathrm{TE}_{Y\to X,i}$ does not depend on i (and so can be simply denoted by $\mathrm{TE}_{Y\to X}$), nor all of the quantities defined in Equations (3) to (5). In theory, TE is never negative and is equal to zero if and only if Equation (2) holds.

According to Definition (3), TE is not symmetric, and it can be regarded as a conditional mutual information (CMI) [3,22] (sometimes also named partial mutual information (PMI) in the literature [23]). Recall that mutual information between two random vectors X and Y is defined by:

$$\mathcal{I}(X;Y) = \mathcal{H}(X) + \mathcal{H}(Y) - \mathcal{H}(X,Y), \tag{6}$$

and TE can be also written as:

$$\mathrm{TE}_{Y\to X} = \mathcal{I}\left(X^p,Y^-|X^-\right). \tag{7}$$

Considering the estimation $\widehat{\mathrm{TE}_{Y\to X}}$ of TE, $\mathrm{TE}_{Y\to X}$, as a function defined on the set of observable occurrences (x_i,y_i), $i=1,\ldots,N$, of a stationary sequence (X_i,Y_i), $i=1,\ldots,N$, and Equation (5), a standard structure for the estimator is given by (see Appendix B):

$$\begin{aligned}\widehat{\mathrm{TE}_{Y\to X}} &= \widehat{\mathcal{H}\left(X^-,Y^-\right)} + \widehat{\mathcal{H}\left(X^p,X^-\right)} - \widehat{\mathcal{H}\left(X^p,X^-,Y^-\right)} - \widehat{\mathcal{H}\left(X^-\right)} \\ &= -\frac{1}{N}\sum_{n=1}^{N}\widehat{\log\left(p_{U_1}\left(u_{1n}\right)\right)} - \frac{1}{N}\sum_{n=1}^{N}\widehat{\log\left(p_{U_2}\left(u_{2n}\right)\right)} + \frac{1}{N}\sum_{n=1}^{N}\widehat{\log\left(p_{U_3}\left(u_{3n}\right)\right)} \\ &\quad + \frac{1}{N}\sum_{n=1}^{N}\widehat{\log\left(p_{U_4}\left(u_{4n}\right)\right)},\end{aligned} \tag{8}$$

where U_1, U_2, U_3 and U_4 stand respectively for (X^-, Y^-), (X^p, X^-), (X^p, X^-, Y^-) and X^-. Here, for each n, $\widehat{\log\left(p_U\left(u_n\right)\right)}$ is an estimated value of $\log\left(p_U\left(u_n\right)\right)$ computed as a function $f_n\left(u_1, \ldots, u_N\right)$ of the observed sequence u_n, $n = 1, \ldots, N$. With the k-NN approach addressed in this study, $f_n\left(u_1, \ldots, u_N\right)$ depends explicitly only on u_n and on its k nearest neighbors. Therefore, the calculation of $\widehat{\mathcal{H}(U)}$ definitely depends on the chosen estimation functions f_n. Note that if, for N fixed, these functions correspond respectively to unbiased estimators of $\log\left(p\left(u_n\right)\right)$, then $\widehat{\mathrm{TE}_{Y\to X}}$ is also unbiased; otherwise, we can only expect that $\widehat{\mathrm{TE}_{Y\to X}}$ is asymptotically unbiased (for N large). This is so if the estimators of $\log\left(p_U\left(u_n\right)\right)$ are asymptotically unbiased.

Now, the theoretical derivation and analysis of the most currently used estimators $\widehat{\mathcal{H}(U)}\left(u_1, \ldots, u_N\right) = -\frac{1}{N}\sum_{n=1}^{N}\widehat{\log\left(p(u_n)\right)}$ for the estimation of $\mathcal{H}\left(U\right)$ generally suppose that $u_1, \ldots, u_N$ are N independent occurrences of the random vector U, *i.e.*, $u_1, \ldots, u_N$ is an occurrence of an independent and identically distributed (IID) sequence $U_1, \ldots, U_N$ of random vectors $(\forall i = 1, \ldots, N : \mathcal{P}_{U_i} = \mathcal{P}_U)$. Although the IID hypothesis does not apply to our initial problem concerning the measure of TE on stationary random sequences (that are generally not IID), the new methods presented in this contribution are extended from existing ones assuming this hypothesis, without relaxing it. However, the experimental section will present results not only on IID observations, but also on non-IID stationary autoregressive (AR) processes, as our goal was to verify if some improvement can be nonetheless obtained for non-IID data, such as AR data.

If we come back to mutual information (MI) defined by Equation (6) and compare it with Equations (5), it is obvious that estimating MI and TE shares similarities. Hence, similarly to Equation (8) for TE, a basic estimation $\widehat{\mathcal{I}\left(X;Y\right)}$ of $\mathcal{I}\left(X;Y\right)$ from a sequence (x_i, y_i), $i = 1, \ldots, N$, of N independent trials is:

$$\widehat{\mathcal{I}\left(X;Y\right)} = -\frac{1}{N}\sum_{n=1}^{N}\widehat{\log\left(p_X\left(x_n\right)\right)} - \frac{1}{N}\sum_{n=1}^{N}\widehat{\log\left(p_Y\left(y_n\right)\right)} + \frac{1}{N}\sum_{n=1}^{N}\log\left(\widehat{p_{X,Y}\left(x_n, y_n\right)}\right). \tag{9}$$

In what follows, when explaining the links among the existing methods and the proposed ones, we refer to Figure 1. In this diagram, a box identified by a number k in a circle is designed by box ⓚ.

Improving performance (in terms of bias and variance) of TE and MI estimators (obtained by choosing specific estimation functions $\widehat{\log\left(p\left(\cdot\right)\right)}$) in Equations (8) and (9), respectively) remains an issue when applied on short-length IID (or non-IID) sequences [3]. In this work, we particularly focused on bias reduction. For MI, the most widely-used estimator is the Kraskov–Stögbauer–Grassberger (KSG) estimator [24,31], which was later extended to estimate transfer entropy, resulting in the k-NN TE estimator [25–27,32–35] (adopted in the widely-used TRENTOOL open source toolbox, Version 3.0). Our contribution originated in the Kozachenko–Leonenko entropy estimator summarized in [24] and proposed beforehand in the literature to get an estimation $\widehat{\mathcal{H}\left(X\right)}$ of the entropy $\mathcal{H}(X)$ of a continuously-distributed random vector X, from a finite sequence of independent outcomes x_i, $i = 1, \ldots, N$. This estimator, as well as another entropy estimator proposed by Singh *et al.* in [36] are briefly described in Section 2.1, before we introduce, in Section 4, our two new TE estimators based on both of them. In Section 2.2, Kraskov MI and standard TE estimators derived in literature from the Kozachenko–Leonenko entropy estimator are summarized, and the passage from a square to rectangular neighboring region to derive new entropy estimation is detailed in Section 3. Our methodology is depicted in Figure 1.

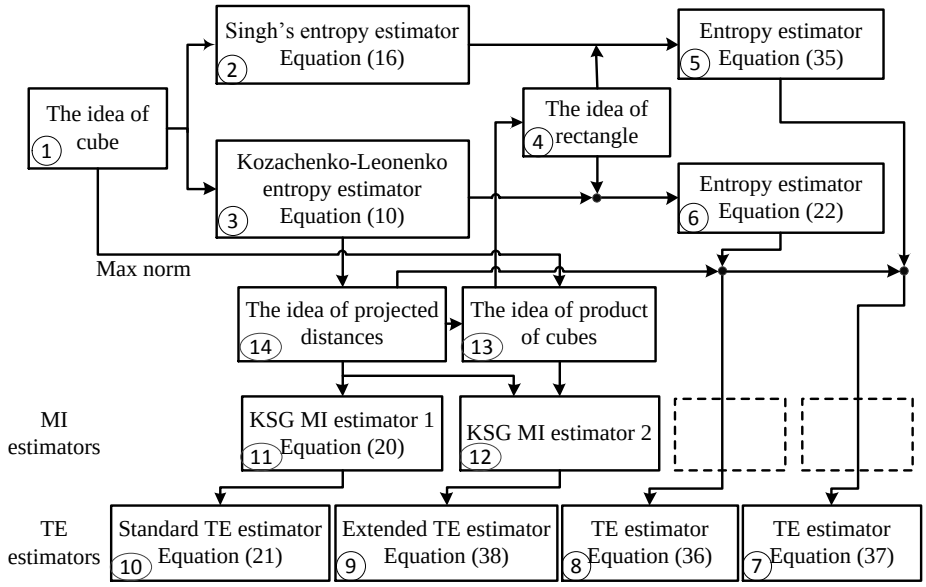


Figure 1. Concepts and methodology involved in k-nearest-neighbors transfer entropy (TE) estimation. Standard k-nearest-neighbors methods using maximum norm for probability density and entropy non-parametric estimation introduce, around each data point, a minimal (hyper-)cube (Box ①), which includes the first k-nearest neighbors, as is the case for two entropy estimators, namely the well-known Kozachenko–Leonenko estimator (Box ③) and the less commonly used Singh's estimator (Box ②). The former was used in [24] to measure mutual information (MI) between two signals X and Y by Kraskov *et al.*, who propose an MI estimator (Kraskov–Stögbauer–Grassberger (KSG) MI Estimator 1, Box ⑪) obtained by summing three entropy estimators (two estimators for the marginal entropies and one for the joint entropy). The strategy was to constrain the three corresponding (hyper-)cubes, including nearest neighbors, respectively in spaces $\mathcal{S}_X$, $\mathcal{S}_Y$ and $\mathcal{S}_{X,Y}$, to have an identical edge length (the idea of projected distances, Box ⑭) for a better cancellation of the three corresponding biases. The same approach was used to derive the standard TE estimator [25–29] (Box ⑩), which has been implemented in the TRENTOOL toolbox, Version 3.0. In [24], Kraskov *et al.* also suggested, for MI estimation, to replace minimal (hyper-)cubes with smaller minimal (hyper-)rectangles equal to the product of two minimal (hyper-)cubes built separately in subspaces $\mathcal{S}_X$ and $\mathcal{S}_Y$ (KSG MI Estimator 2, Box ⑫) to exploit more efficiently the Kozachenko–Leonenko approach. An extended algorithm for TE estimation based on minimal (hyper-)rectangles equal to products of (hyper-)cubes was then proposed in [27] (extended TE estimator, Box ⑨) and implemented in the JIDT toolbox [30]. Boxes ⑩ and ⑨ are marked as "standard algorithm" and "extended algorithm". The new idea extends the idea of the product of cubes (Box ⑬). It consists of proposing a different construction of the neighborhoods, which are no longer minimal (hyper-)cubes, nor products of (hyper-)cubes, but minimal (hyper-)rectangles (Box ④), with possibly a different length for each dimension, to get two novel entropy estimators (Boxes ⑤ and ⑥), respectively derived from Singh's entropy estimator and the Kozachenko–Leonenko entropy estimator. These two new entropy estimators lead respectively to two new TE estimators (Box ⑦ and Box ⑧) to be compared with the standard and extended TE estimators.

2. Original k-Nearest-Neighbors Strategies

2.1. Kozachenko–Leonenko and Singh's Entropy Estimators for a Continuously-Distributed Random Vector

2.1.1. Notations

Let us consider a sequence x_i, $i = 1, \ldots, N$ in $\mathbb{R}^{d_X}$ (in our context, this sequence corresponds to an outcome of an IID sequence $X_1, \ldots, X_N$, such that the common probability distribution will be equal to that of a given random vector X). The set of the k nearest neighbors of x_i in this sequence (except for x_i) and the distance between x_i and its k-th nearest neighbor are respectively denoted by χ_i^k and $d_{x_i,k}$. We denote $\mathcal{D}_{x_i}\left(\chi_i^k\right) \subset \mathbb{R}^{d_X}$ a neighborhood of x_i in $\mathbb{R}^{d_X}$, which is the image of $\left(x_i, \chi_i^k\right)$ by a set valued map. For a given norm $\|\cdot\|$ on $\mathbb{R}^{d_X}$ (Euclidean norm, maximum norm, *etc.*), a standard construction $\left(x_i, \chi_i^k\right) \in \left(\mathbb{R}^{d_X}\right)^{k+1} \to \mathcal{D}_{x_i}\left(\chi_i^k\right) \subset \mathbb{R}^{d_X}$ is the (hyper-)ball of radius equal to $d_{x_i,k}$, *i.e.*, $\mathcal{D}_{x_i}\left(\chi_i^k\right) = \{x : \|x - x_i\| \leq d_{x_i,k}\}$. The (hyper-)volume (*i.e.*, the Lebesgue measure) of $\mathcal{D}_{x_i}\left(\chi_i^k\right)$ is then $v_i = \int_{\mathcal{D}_{x_i}\left(\chi_i^k\right)} \mathrm{d}x$ (where $\mathrm{d}x \triangleq \mathrm{d}\mu^{d_X}(x)$).

2.1.2. Kozachenko–Leonenko Entropy Estimator

The Kozachenko–Leonenko entropy estimator is given by (Box ③ in Figure 1):

$$\widehat{\mathcal{H}(X)}_{KL} = \psi(N) + \frac{1}{N}\sum_{i=1}^{N} \log\left(v_i\right) - \psi(k), \tag{10}$$

where v_i is the volume of $\mathcal{D}_{x_i}\left(\chi_i^k\right) = \{x : \|x - x_i\| \leq d_{x_i,k}\}$ computed with the maximum norm and $\psi(k) = \frac{\Gamma'(k)}{\Gamma(k)}$ denotes the digamma function. Note that using Equation (10), entropy is measured in natural units (nats).

To come up with a concise presentation of this estimator, we give hereafter a summary of the different steps to get it starting from [24]. First, let us consider the distance $d_{x_i,k}$ between x_i and its k-th nearest neighbor (introduced above) as a realization of the random variable $D_{x_i,k}$, and let us denote by $q_{x_i,k}(x)$, $x \in \mathbb{R}$, the corresponding probability density function (conditioned by $X_i = x_i$). Secondly, let us consider the quantity $h^{x_i}(\varepsilon) = \int_{\|u - x_i\| \leq \varepsilon/2} \mathrm{d}P_X(u)$. This is the probability mass of the (hyper-)ball with radius equal to $\varepsilon/2$ and centered on x_i. This probability mass is approximately equal to:

$$h^{x_i}(\varepsilon) \simeq p_X(x_i) \int_{\|\xi\| \leq \varepsilon/2} \mathrm{d}\mu^d(\xi) = p_X(x_i)\, c_d \varepsilon^d, \tag{11}$$

if the density function is approximately constant on the (hyper-)ball. The variable c_d is the volume of the unity radius d-dimensional (hyper-)ball in $\mathbb{R}^d$ ($c_d = 1$ with maximum norm). Furthermore, it can be established (see [24] for details) that the expectation $E\left[\log\left(h^{X_i}\left(D_{X_i,k}\right)\right)\right]$, where h^{X_i} is the random variable associated with h^{x_i}, $D_{X_i,k}$ (which must not be confused with the notation $\mathcal{D}_{x_i}\left(\chi_i^k\right)$ introduced previously) denotes the random distance between the k-th neighbor selected in the set of random vectors $\{X_k, 1 \leq k \leq N, k \neq i\}$, and the random point X_i is equal to $\psi(k) - \psi(N)$ and does not depend on $p_X(\cdot)$. Equating it with $E\left[\log\left(p_X\left(X_i\right) c_d D_{X_i,k}\right)\right]$ leads to:

$$\begin{aligned}\psi(k) - \psi(N) &\simeq E\left[\log\left(p_X\left(X_i\right)\right)\right] + E\left[\log\left(c_d D_{X_i,k}^d\right)\right] \\ &= -\mathcal{H}(X_i) + E\left[\log\left(V_i\right)\right]\end{aligned} \tag{12}$$

and:

$$\mathcal{H}\left(X_i\right) \simeq \psi(N) - \psi(k) + E\left[\log\left(c_d D_{X_i,k}^d\right)\right]. \tag{13}$$

Finally, by using the law of large numbers, when N is large, we get:

$$\begin{aligned}\mathcal{H}\left(X_i\right) &\simeq \psi(N) - \psi(k) + \frac{1}{N}\sum_{i=1}^{N}\log\left(v_i\right) \\ &= \widehat{\mathcal{H}\left(X\right)}_{KL},\end{aligned} \tag{14}$$

where v_i is the realization of the random (hyper-)volume $V_i = c_d D_{x_i,k}^d$.

Moreover, as observed in [24], it is possible to make the number of neighbors k depend on i by substituting the mean $\frac{1}{N}\sum_{i=1}^{N}\psi(k_i)$ for the constant $\psi(k)$ in Equation (14), so that $\widehat{\mathcal{H}\left(X\right)}_{KL}$ becomes:

$$\widehat{\mathcal{H}\left(X\right)}_{KL} = \psi(N) + \frac{1}{N}\sum_{i=1}^{N}\left(\log\left(v_i\right) - \psi(k_i)\right). \tag{15}$$

2.1.3. Singh's Entropy Estimator

The question of k-NN entropy estimation is also discussed by Singh *et al.* in [36], where another estimator, denoted by $\widehat{\mathcal{H}(X)}_S$ hereafter, is proposed (Box ② in Figure 1):

$$\widehat{\mathcal{H}(X)}_S = \log(N) + \frac{1}{N}\sum_{i=1}^{N}\log\left(v_i\right) - \psi(k). \tag{16}$$

Using the approximation $\psi(N) \approx \log(N)$ for large values of N, the estimator given by Equation (16) is close to that defined by Equation (10). This estimator was derived by Singh *et al.* in [36] through the four following steps:

(1) Introduce the classical entropy estimator structure:

$$\widehat{\mathcal{H}(X)} \triangleq -\frac{1}{N}\sum_{i=1}^{N}\log\widehat{p_X\left(X_i\right)} = \frac{1}{N}\sum_{i=1}^{N}T_i, \tag{17}$$

where:

$$\widehat{p_X(x_i)} \triangleq \frac{k}{N v_i}. \tag{18}$$

(2) Assuming that the random variables T_i, $i = 1, \ldots, N$ are identically distributed, so that $E\left[\widehat{\mathcal{H}(X)}\right] = E\left(T_1\right)$ (note that $E\left(T_1\right)$ depends on N, even if the notation does not make that explicit), compute the asymptotic value of $E\left(T_1\right)$ (when N is large) by firstly computing its asymptotic cumulative probability distribution function and the corresponding probability density p_{T_1}, and finally, compute the expectation $E\left(T_1\right) = \int_{\mathbb{R}} t p_{T_1}(t)\mathrm{d}t$.

(3) It appears that $E(T_1) = E\left[\widehat{\mathcal{H}(X)}\right] = \mathcal{H}(X) + B$ where B is a constant, which is identified with the bias.

(4) Subtract this bias from $\widehat{\mathcal{H}(X)}$ to get $\widehat{\mathcal{H}(X)}_S = \widehat{\mathcal{H}(X)} - B$ and the formula given in Equation (16).

Note that the cancellation of the asymptotic bias does not imply that the bias obtained with a finite value of N is also exactly canceled. In Appendix C, we explain the origin of the bias for the entropy estimator given in Equation (17).

Observe also that, as for the Kozachenko–Leonenko estimator, it is possible to adapt Equation (16) if we want to consider a number of neighbors k_i depending on i. Equation (16) must then be replaced by:

$$\widehat{\mathcal{H}(X)}_S = \log(N) + \frac{1}{N}\sum_{i=1}^{N}\left(\log\left(v_i\right) - \psi(k_i)\right). \tag{19}$$

2.2. Standard Transfer Entropy Estimator

Estimating entropies separately in Equations (8) and (9) leads to individual bias values. Now, it is possible to cancel out (at least partially) the bias considering the algebraic sums (Equations (8) and (9)). To help in this cancellation, on the basis of Kozachenko–Leonenko entropy estimator, Kraskov *et al.* proposed to retain the same (hyper-)ball radius for each of the different spaces instead of using the same number k for both joint space $\mathcal{S}_{X,Y}$ and marginal spaces ($\mathcal{S}_X$ and $\mathcal{S}_Y$ spaces) [24,37], leading to the following MI estimator (Box ⑪ in Figure 1):

$$\widehat{\mathcal{I}}_K = \psi(k) + \psi(N) - \frac{1}{N}\sum_{i=1}^{N}\left[\psi(n_{X,i}+1) + \psi(n_{Y,i}+1)\right], \tag{20}$$

where $n_{X,i}$ and $n_{Y,i}$ denote the number of points that strictly fall into the resulting distance in the lower-dimensional spaces $\mathcal{S}_X$ and $\mathcal{S}_Y$, respectively.

Applying the same strategy to estimate TE, the number of neighbors in the joint space $\mathcal{S}_{X^p,X^-,Y^-}$ is first fixed, then for each i, the resulting distance $\varepsilon_i \triangleq d_{(x_i^p,x_i^-,y_i^-),k}$ is projected into the other three lower dimensional spaces, leading to the standard TE estimator [25,27,28] (implementation available in the TRENTOOL toolbox, Version 3.0, Box ⑩ in Figure 1):

$$\widehat{\mathrm{TE}_{Y\to X}}_{SA} = \psi(k) + \frac{1}{N}\sum_{i=1}^{N}\left[\psi(n_{X^-,i}+1) - \psi(n_{(X^-,Y^-),i}+1) - \psi(n_{(X^p,X^-),i}+1)\right], \tag{21}$$

where $n_{X^-,i}$, $n_{(X^-,Y^-),i}$ and $n_{(X^p,X^-),i}$ denote the number of points that fall into the distance ε_i from x_i^-, $\left(x_i^-,y_i^-\right)$ and $\left(x_i^p,x_i^-\right)$ in the lower dimensional spaces $\mathcal{S}_{X^-}$, $\mathcal{S}_{X^-,Y^-}$ and $\mathcal{S}_{X^p,X^-}$, respectively. This estimator is marked as the "standard algorithm" in the experimental part.

Note that a generalization of Equation (21) was proposed in [28] to extend this formula to the estimation of entropy combinations other than MI and TE.

3. From a Square to a Rectangular Neighboring Region for Entropy Estimation

In [24], to estimate MI, as illustrated in Figure 2, Kraskov *et al.* discussed two different techniques to build the neighboring region to compute $\widehat{\mathcal{I}(X;Y)}$: in the standard technique (square $ABCD$ in Figure 2a,b), the region determined by the first k nearest neighbors is a (hyper-)cube and leads to Equation (20), and in the second technique (rectangle $A'B'C'D'$ in Figure 2a,b), the region determined by the first k nearest neighbors is a (hyper-)rectangle. Note that the TE estimator mentioned in the previous section (Equation (21)) is based on the first situation (square $ABCD$ in Figure 2a or 2b). The introduction of the second technique by Kraskov *et al.* was to circumvent the fact that Equation (15) was not applied rigorously to obtain the terms $\psi(n_{X,i}+1)$ or $\psi(n_{Y,i}+1)$ in Equation (20). As a matter of fact, for one of these terms, no point x_i (or y_i) falls exactly on the border of the (hyper-)cube $\mathcal{D}_{x_i}$ (or $\mathcal{D}_{y_i}$) obtained by the distance projection from the $\mathcal{S}_{X,Y}$ space. As clearly illustrated in Figure 2 (rectangle $A'B'C'D'$ in Figure 2a,b), the second strategy prevents that issue, since the border of the (hyper-)cube (in this case, an interval of $\mathbb{R}$) after projection from $\mathcal{S}_{X,Y}$ space to $\mathcal{S}_X$ space (or $\mathcal{S}_Y$ space) contains one point. When the dimensions of $\mathcal{S}_X$ and $\mathcal{S}_Y$ are larger than one, this strategy leads to building an (hyper-)rectangle equal to the product of two (hyper-)cubes, one of them in $\mathcal{S}_X$ and the other one in $\mathcal{S}_Y$. If the maximum distance of the k-th NN in $\mathcal{S}_{X,Y}$ is obtained in one of the directions in $\mathcal{S}_X$, this maximum distance, after multiplying by two, fixes the size of the (hyper-)cube in $\mathcal{S}_X$. To obtain the size of the second (hyper-)cube (in $\mathcal{S}_Y$), the k neighbors in $\mathcal{S}_{X,Y}$ are first projected on $\mathcal{S}_Y$, and then, the largest of the distances calculated from these projections fixes the size of this second (hyper-)cube.

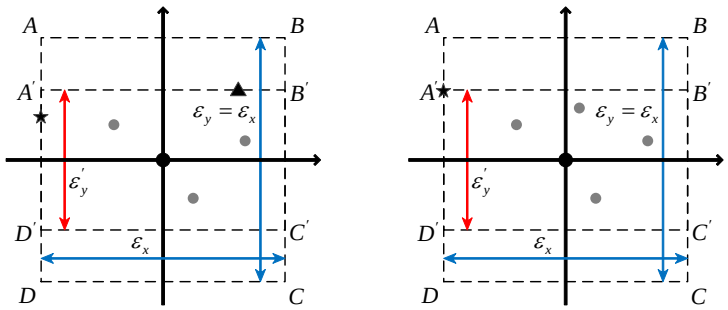


Figure 2. In this two-dimensional example, $k = 5$. The origin of the Cartesian axis corresponds to the current point x_i. Only the five nearest neighbors of this point, *i.e.*, the points in the set χ_i^k, are represented. The fifth nearest neighbor is symbolized by a star. The neighboring regions $ABCD$, obtained from the maximum norm around the center point, are squares, with equal edge lengths $\varepsilon_x = \varepsilon_y$. Reducing one of the edge lengths, ε_x or ε_y, until one point falls onto the border (in the present case, in the vertical direction), leads to the minimum size rectangle $A'B'C'D'$, where $\varepsilon_x \neq \varepsilon_y$. Two cases must be considered: (**a**) the fifth neighbor is not localized on a node, but between two nodes, contrary to (**b**). This leads to obtaining either two points (respectively the star and the triangle in (a)) or only one point (the star in(b)) on the border of $A'B'C'D'$. Clearly, it is theoretically possible to have more than two points on the border of $A'B'C'D'$, but the probability of such an occurrence is equal to zero when the probability distribution of the random points X_j is continuous.

In the remainder of this section, for an arbitrary dimension d, we propose to apply this strategy to estimate the entropy of a single multidimensional variable X observed in $\mathbb{R}^d$. This leads to introducing a d-dimensional (hyper-)rectangle centered on x_i having a minimal volume and including the set χ_i^k of neighbors. Hence, the rectangular neighboring is built by adjusting its size separately in each direction in the space $\mathcal{S}_X$. Using this strategy, we are sure that, in any of the d directions, there is at least one point on one of the two borders (and only one with probability one). Therefore, in this approach, the (hyper-)rectangle, denoted by $\mathcal{D}_{x_i}^{\varepsilon_1,\ldots,\varepsilon_d}$, where the sizes $\varepsilon_1,\ldots,\varepsilon_d$ in the respective d directions are completely specified from the neighbors set χ_i^k, is substituted for the basic (hyper-)square $\mathcal{D}_{x_i}\left(\chi_i^k\right) = \{x : \|x - x_i\| \leq d_{x_i,k}\}$. It should be mentioned that the central symmetry of the (hyper-)rectangle around the center point allows for reducing the bias in the density estimation [38] (*cf.* Equation (11) or (18)). Note that, when $k < d$, there must exist neighbors positioned on some vertex or edges of the (hyper-)rectangle. With $k < d$, it is impossible that, for any direction, one point falls exactly inside a face (*i.e.*, not on its border). For example, with $k = 1$ and $d > 1$, the first neighbor will be on a vertex, and the sizes of the edges of the reduced (hyper-)rectangle will be equal to twice the absolute value of its coordinates, whatever the direction.

Hereafter, we propose to extend the entropy estimators by Kozachenko–Leonenko and Singh using the above strategy before deriving the corresponding TE estimators and comparing their performance.

3.1. Extension of the Kozachenko–Leonenko Method

As indicated before, in [24], Kraskov *et al.* extended the Kozachenko–Leonenko estimator (Equations (10) and (15)) using the rectangular neighboring strategy to derive the MI estimator. Now, focusing on entropy estimation, after some mathematical developments (see Appendix D), we obtain another estimator of $\mathcal{H}(X)$, denoted by $\widehat{\mathcal{H}(X)}_K$ (Box ⑥ in Figure 1),

$$\widehat{\mathcal{H}(X)}_K = \psi(N) + \frac{1}{N}\sum_{i=1}^{N}\log\left(v_i\right) - \psi(k) + \frac{d-1}{k}. \tag{22}$$

Here, v_i is the volume of the minimum volume (hyper-)rectangle around the point x_i. Exploiting this entropy estimator, after substitution in Equation (8), we can derive a new estimation of TE.

3.2. Extension of Singh's Method

We propose in this section to extend Singh's entropy estimator by using a (hyper-)rectangular domain, as we did for the Kozachenko–Leonenko estimator extension introduced in the preceding section. Considering a d-dimensional random vector $X \in \mathbb{R}^d$ continuously distributed according to a probability density function p_X, we aim at estimating the entropy $\mathcal{H}(X)$ from the observation of a p_X distributed IID random sequence X_i, $i = 1,\ldots,N$. For any specific data point x_i and a fixed number k ($1 \leq k \leq N$), the minimum (hyper-)rectangle (rectangle $A'B'C'D'$ in Figure 2) is fixed, and we denote this region by $\mathcal{D}_{x_i}^{\varepsilon_1,\ldots,\varepsilon_d}$ and its volume by v_i. Let us denote ξ_i ($1 \leq \xi_i \leq \min(k,d)$) the number of points on the border of the (hyper-)rectangle that we consider as a realization of a random variable Ξ_i. In the situation described in Figure 2a,b, $\xi_i = 2$ and $\xi_i = 1$, respectively. According to [39] (Chapter 6, page 269),

if $\mathcal{D}_{x_i}\left(\chi_i^k\right)$ corresponds to a ball (for a given norm) of volume v_i, an unbiased estimator of $p_X(x_i)$ is given by:

$$\widehat{p_X(x_i)} = \frac{k-1}{Nv_i}, i = 1, 2, \ldots, N. \tag{23}$$

This implies that the classical estimator $\widehat{p_X(x_i)} = \frac{k}{Nv_i}$ is biased and that presumably $\log\left(\frac{k}{Nv_i}\right)$ is also a biased estimation of $\log\left(p_X(x_i)\right)$ for N large, as shown in [39].

Now, in the case $\mathcal{D}_{x_i}\left(\chi_i^k\right)$ is the minimal (*i.e.*, with minimal (hyper-)volume) (hyper-)rectangle $\mathcal{D}_{x_i}^{\varepsilon_1,\ldots,\varepsilon_d}$, including χ_i^k, more than one point can belong to the border, and a more general estimator $\widetilde{p_X(x_i)}$ of $p_X(x_i)$ can be *a priori* considered:

$$\widetilde{p_X(x_i)} = \frac{\tilde{k}_i}{Nv_i}, \tag{24}$$

where $\tilde{k}_i$ is some given function of k and ξ_i. The corresponding estimation of $\mathcal{H}(X)$ is then:

$$\widehat{\mathcal{H}(X)} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\widetilde{p_X(x_i)}\right) = \frac{1}{N}\sum_{i=1}^{N} t_i, \tag{25}$$

with:

$$t_i = \log\left(\frac{Nv_i}{\tilde{k}_i}\right), i = 1, 2, \ldots, N, \tag{26}$$

t_i being realizations of random variables T_i and $\tilde{k}_i$ being realizations of random variables $\widetilde{K}_i$. We have:

$$\forall i = 1, \ldots, N : E\left[\widehat{\mathcal{H}(X)}\right] = E\left(T_i\right) = E\left(T_1\right). \tag{27}$$

Our goal is to derive $E\left[\widehat{\mathcal{H}(X)}\right] - \mathcal{H}(X) = E\left(T_1\right) - \mathcal{H}(X)$ for N large to correct the asymptotic bias of $\widehat{\mathcal{H}(X)}$, according to Steps (1) to (3), explained in Section 2.1.3. To this end, we must consider an asymptotic approximation of the conditional probability distribution $\mathcal{P}\left(T_1 \leq r | X_1 = x_1, \Xi_1 = \xi_1\right)$ before computing the asymptotic difference between the expectation $E\left[T_1\right] = E\left[E\left[T_1 | X_1 = x_1, \Xi_1 = \xi_1\right]\right]$ and the true entropy $\mathcal{H}(X)$.

Let us consider the random Lebesgue measure V_1 of the random minimal (hyper-)rectangle $\mathcal{D}_{x_1}^{\epsilon_1,\ldots,\epsilon_d}$ ($(\epsilon_1, \ldots, \epsilon_d)$ denotes the random vector for which $(\varepsilon_1, \ldots, \varepsilon_d) \in \mathbb{R}^d$ is a realization) and the relation $T_1 = \log\left(\frac{NV_1}{\widetilde{K}_1}\right)$. For any $r > 0$, we have:

$$\begin{aligned}\mathcal{P}\left(T_1 > r | X_1 = x_1, \Xi_1 = \xi_1\right) &= \mathcal{P}\left(\log\left(\frac{NV_1}{\widetilde{K}_1}\right) > r | X_1 = x_1, \Xi_1 = \xi_1\right) \\ &= \mathcal{P}\left(V_1 > v_r | X_1 = x_1, \Xi_1 = \xi_1\right),\end{aligned} \tag{28}$$

where $v_r = e^r \frac{\tilde{k}_1}{N}$, since, conditionally to $\Xi_1 = \xi_1$, we have $\widetilde{K}_1 = \tilde{k}_1$.

In Appendix E, we prove the following property.

Property 1. *For N large,*

$$\mathcal{P}\left(T_1 > r | X_1 = x_1, \Xi_1 = \xi_1\right) \simeq \sum_{i=0}^{k-\xi_1}\binom{N-\xi_1-1}{i}\left(p_X(x_1)v_r\right)^i\left(1-p_X(x_1)v_r\right)^{N-\xi_1-1-i}. \tag{29}$$

The Poisson approximation (when $N \to \infty$ and $v_r \to 0$) of the binomial distribution summed in Equation (29) leads to a parameter $\lambda = (N - \xi_1 - 1)\, p_X(x_1) v_r$. As N is large compared to $\xi_1 + 1$, we obtain from Equation (26):

$$\lambda \simeq \tilde{k}_1 e^r p_X(x_1), \tag{30}$$

and we get the approximation:

$$\lim_{N\to\infty} \mathcal{P}\left(T_1 > r | X_1 = x_1, \Xi_1 = \xi_1\right) \simeq \sum_{i=0}^{k-\xi_1} \frac{\left[\tilde{k}_1 e^r p_X(x_1)\right]^i}{i!} e^{-\tilde{k}_1 e^r p_X(x_1)}. \tag{31}$$

Since $\mathcal{P}(T_1 \leq r | X_1 = x_1, \Xi_1 = \xi_1) = 1 - \mathcal{P}(T_1 > r | X_1 = x_1, \Xi_1 = \xi_1)$, we can get the density function of T_1, noted $g_{T_1}(r)$, by deriving $\mathcal{P}\left(T_1 \leq r | X_1 = x_1, \Xi_1 = \xi_1\right)$. After some mathematical developments (see Appendix F), we obtain:

$$\begin{aligned} g_{T_1}(r) &= \mathcal{P}'(T_1 \leq r | X_1 = x_1, \Xi_1 = \xi_1) \\ &= -\mathcal{P}'(T_1 > r | X_1 = x_1, \Xi_1 = \xi_1) \\ &= \frac{\left[\tilde{k}_1 e^r p_X(x_1)\right]^{(k-\xi_1+1)}}{(k-\xi_1)!} e^{-\tilde{k}_1 e^r p_X(x_1)} \end{aligned} \quad , r \in \mathbb{R}, \tag{32}$$

and consequently (see Appendix G for details),

$$\begin{aligned} \lim_{N\to\infty} E\left[T_1 | X_1 = x_1, \Xi_1 = \xi_1\right] &= \int_{-\infty}^{\infty} r \frac{\left[\tilde{k}_1 e^r p_X(x_1)\right]^{(k-\xi_1+1)}}{(k-\xi_1)!} e^{-\tilde{k}_1 e^r p_X(x_1)} \mathrm{d}r \\ &= \psi(k - \xi_1 + 1) - \log\left(\tilde{k}_1\right) - \log\left(p_X(x_1)\right). \end{aligned} \tag{33}$$

Therefore, with the definition of differential entropy $\mathcal{H}(X_1) = E[-\log\left(p_X(X_1)\right)]$, we have:

$$\lim_{N\to\infty} E\left[T_1\right] = \lim_{N\to\infty} E\left[E\left[T_1 | X_1, \Xi_1\right]\right] = E\left[\psi(k - \Xi_1 + 1) - \log\left(\widetilde{K}_1\right)\right] + \mathcal{H}(X_1). \tag{34}$$

Thus, the estimator expressed by Equation (25) is asymptotically biased. Therefore, we consider a modified version, denoted by $\widehat{\mathcal{H}(X)}_{NS}$, obtained by subtracting an estimation of the bias $E\left[\psi(k - \Xi_1 + 1) - \log\left(\widetilde{K}_1\right)\right]$ given by the empirical mean $\frac{1}{N}\sum_{i=1}^{N} \psi(k - \xi_i + 1) + \frac{1}{N}\sum_{i=1}^{N} \log\left(\tilde{k}_i\right)$ (according to the large numbers law), and we obtain, finally (Box ⑤ in Figure 1):

$$\begin{aligned} \widehat{\mathcal{H}(X)}_{NS} &= \frac{1}{N}\sum_{i=1}^{N} t_i - \frac{1}{N}\sum_{i=1}^{N} \psi(k - \xi_i + 1) + \frac{1}{N}\sum_{i=1}^{N} \log\left(\tilde{k}_i\right) \\ &= \frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{N v_i}{\tilde{k}_i}\right) - \frac{1}{N}\sum_{i=1}^{N} \psi(k - \xi_i + 1) + \frac{1}{N}\sum_{i=1}^{N} \log\left(\tilde{k}_i\right) \\ &= \log(N) + \frac{1}{N}\sum_{i=1}^{N} \log\left(v_i\right) - \frac{1}{N}\sum_{i=1}^{N} \psi(k - \xi_i + 1). \end{aligned} \tag{35}$$

In comparison with the development of Equation (22), we followed here the same methodology, except we take into account (through a conditioning technique) the influence of the number of points on the border.

We observe that, after cancellation of the asymptotic bias, the choice of the function of k and ξ_i to define $\tilde{k}_i$ in Equation (24) does not have any influence on the final result. In this way, we obtain an expression for $\widehat{\mathcal{H}(X)}_{NS}$, which simply takes into account the values ξ_i that could *a priori* influence the entropy estimation.

Note that, as for the original Kozachenko–Leonenko (Equation (10)) and Singh (Equation (16)) entropy estimators, both new estimation functions (Equations (22) and (35)) hold for any value of k, such that $k \ll N$, and we do not have to choose a fixed k while estimating entropy in lower dimensional spaces. Therefore, under the framework proposed in [24], we built two different TE estimators using Equations (22) and (35), respectively.

3.3. Computation of the Border Points Number and of the (Hyper-)Rectangle Sizes

We explain more precisely hereafter how to determine the numbers of points ξ_i on the border. Let us denote $x_i^j \in \mathbb{R}^d$, $j = 1, \ldots, k$, the k nearest neighbors of $x_i \in \mathbb{R}^d$, and let us consider the $d \times k$ array D_i, such that for any $(p, j) \in \{1, \ldots, d\} \times \{1, \ldots, k\}$, $D_i(p, j) = \left| x_i^j(p) - x_i(p) \right|$ is the distance (in $\mathbb{R}$) between the p-th component $x_i^j(p)$ of x_i^j and the p-th component $x_i(p)$ of x_i. For each p, let us introduce $J_i(p) \in \{1, \ldots, k\}$ defined by $D_i(p, J_i(p)) = \max\left(D_i(p, 1), \ldots, D_i(p, k)\right)$ and which is the value of the column index of D_i for which the distance $D_i(p, j)$ is maximum in the row number p. Now, if there exists more than one index $J_i(p)$ that fulfills this equality, we select arbitrarily the lowest one, hence avoiding the $\max(\cdot)$ function to be multi-valued. The MATLAB implementation of the $\max$ function selects such a unique index value. Then, let us introduce the $d \times k$ Boolean array B_i defined by $B_i(p, j) = 1$ if $j = J_i(p)$ and $B_i(p, j) = 0$, otherwise. Then:

(1) The d sizes ε_p, $p = 1, \ldots, d$ of the (hyper-)rectangle $\mathcal{D}_{x_i}^{\varepsilon_1, \ldots, \varepsilon_d}$ are equal respectively to $\varepsilon_p = 2D_i(p, J_i(p))$, $p = 1, \ldots, d$.

(2) We can define ξ_i as the number of non-null column vectors in B_i. For example, if the k-th nearest neighbor x_i^k is such that $\forall j \neq k$, $\forall p = 1, \ldots, d : \left| x_i^j(p) - x_i(p) \right| < \left| x_i^k(p) - x_i(p) \right|$, *i.e.*, when the k-th nearest neighbor is systematically the farthest from the central point x_i for each of the d directions, then all of the entries in the last column of B_i are equal to one, while all other entries are equal to zero: we have only one column including values different from zero and, so, only one point on the border ($\xi_i = 1$), which generalizes the case depicted in Figure 2b for $d = 2$.

N.B.: this determination of ξ_i may be incorrect when there exists a direction p, such that the number of indices j for which $D_i(p, j)$ reaches the maximal value is larger than one: the value of ξ_i obtained with our procedure can then be underestimated. However, we can argue that, theoretically, this case occurs with a probability equal to zero (because the observations are continuously distributed in the probability) and, so, it can be *a priori* discarded. Now, in practice, the measured quantification errors and the round off errors are unavoidable, and this probability will differ from zero (although remaining small when the aforesaid errors are small): theoretically distinct values $D_i(p, j)$ on the row p of D_i may be erroneously confounded after quantification and rounding. However, the $\max(\cdot)$ function then selects on row p only one value for $J_i(p)$ and, so, acts as an error correcting procedure. The fact that the maximum distance

in the concerned p directions can then be allocated to the wrong neighbor index has no consequence for the correct determination of ξ_i.

4. New Estimators of Transfer Entropy

From an observed realization $\left(x_i^p, x_i^-, y_i^-\right) \in \mathcal{S}_{X^p,X^-,Y^-}$, $i = 1, 2, \ldots, N$ of the IID random sequence $\left(X_i^p, X_i^-, Y_i^-\right)$, $i = 1, 2, \ldots, N$ and a number k of neighbors, the procedure could be summarized as follows (distances are from the maximum norm):

(1) similarly to the MILCA [31] and TRENTOOL toolboxes [34], normalize, for each i, the vectors x_i^p, x_i^- and y_i^-;

(2) in joint space $\mathcal{S}_{X^p,X^-,Y^-}$, for each point $\left(x_i^p, x_i^-, y_i^-\right)$, calculate the distance $d_{\left(x_i^p,x_i^-,y_i^-\right),k}$ between $\left(x_i^p, x_i^-, y_i^-\right)$ and its k-th neighbor, then construct the (hyper-)rectangle with sizes $\varepsilon_1, \ldots, \varepsilon_d$ (d is the dimension of the vectors $\left(x_i^p, x_i^-, y_i^-\right)$), for which the (hyper-)volume is $v_{(X^p,X^-,Y^-),i} = \varepsilon_1 \times \ldots \times \varepsilon_d$ and the border contains $\xi_{(X^p,X^-,Y^-),i}$ points;

(3) for each point (x_i^p, x_i^-) in subspace $\mathcal{S}_{X^p,X^-}$, count the number $k_{(X^p,X^-),i}$ of points falling within the distance $d_{\left(x_i^p,x_i^-,y_i^-\right),k}$, then find the smallest (hyper-)rectangle that contains all of these points and for which $v_{(X^p,X^-),i}$ and $\xi_{(X^p,X^-),i}$ are respectively the volume and the number of points on the border; repeat the same procedure in subspaces $\mathcal{S}_{X^-,Y^-}$ and $\mathcal{S}_{X^-}$.

From Equation (22) (modified to k not constant for $\mathcal{S}_{X^-}$, $\mathcal{S}_{X^p,X^-}$ and $\mathcal{S}_{X^-,Y^-}$), the final TE estimator can be written as (Box ⑧ in Figure 1):

$$\begin{aligned}\widehat{\mathrm{TE}}_{Y\to X p1} =& \frac{1}{N}\sum_{i=1}^{N}\log\frac{v_{(X^p,X^-),i}\cdot v_{(X^-,Y^-),i}}{v_{(X^p,X^-,Y^-),i}\cdot v_{X^-,i}} \\ &+ \frac{1}{N}\sum_{i=1}^{N}\left(\psi(k)+\psi(k_{X^-,i})-\psi(k_{(X^p,X^-),i})-\psi(k_{(X^-,Y^-),i})\right. \\ &\left.+\frac{d_{X^p}+d_{X^-}-1}{k_{(X^p,X^-),i}}+\frac{d_{X^-}+d_{Y^-}-1}{k_{(X^-,Y^-),i}}-\frac{d_{X^p}+d_{X^-}+d_{Y^-}-1}{k}-\frac{d_{X^-}-1}{k_{X^-,i}}\right),\end{aligned} \tag{36}$$

where $d_{X^p} = \dim\left(\mathcal{S}_{X^p}\right), d_{X^-} = \dim\left(\mathcal{S}_{X^-}\right), d_{Y^-} = \dim\left(\mathcal{S}_{Y^-}\right)$, and with Equation (35), it yields to (Box ⑦ in Figure 1):

$$\begin{aligned}\widehat{\mathrm{TE}}_{Y\to X p2} =& \frac{1}{N}\sum_{i=1}^{N}\log\frac{v_{(X^p,X^-),i}\cdot v_{(X^-,Y^-),i}}{v_{(X^p,X^-,Y^-),i}\cdot v_{X^-,i}} \\ &+ \frac{1}{N}\sum_{i=1}^{N}\left(\psi(k-\xi_{(X^p,X^-,Y^-),i}+1)+\psi(k_{X^-,i}-\xi_{X^-,i}+1)-\psi(k_{(X^p,X^-),i}\right. \\ &\left.-\xi_{(X^p,X^-),i}+1)-\psi(k_{(X^-,Y^-),i}-\xi_{(X^-,Y^-),i}+1)\right).\end{aligned} \tag{37}$$

In Equations (36) and (37), the volumes $v_{(X^p,X^-),i}$, $v_{(X^-,Y^-),i}$, $v_{(X^p,X^-,Y^-),i}$, $v_{X^-,i}$ are obtained by computing, for each of them, the product of the edges lengths of the (hyper-)rectangle, *i.e.*, the product

of d edges lengths, d being respectively equal to $d_{X^p}+d_{X^-}$, $d_{X^-}+d_{Y^-}$, $d_{X^p}+d_{X^-}+d_{Y^-}$ and d_{X^-}. In a given subspace and for a given direction, the edge length is equal to twice the largest distance between the corresponding coordinate of the reference point (at the center) and each of the corresponding coordinates of the k nearest neighbors. Hence a generic formula is $v_U = \prod_{j=1}^{\dim(\mathcal{U})} \varepsilon_{Uj}$, where U is one of the symbols (X^p, X^-), (X^-, Y^-), (X^p, X^-, Y^-) and X^- and the ε_{Uj} are the edge lengths of the (hyper-)rectangle.

The new TE estimator $\widehat{\mathrm{TE}}_{Y\to X p1}$ (Box ⑧ in Figure 1) can be compared with the extension of $\widehat{\mathrm{TE}}_{Y\to X SA}$, the TE estimator proposed in [27] (implemented in the JIDT toolbox [30]). This extension [27], included in Figure 1 (Box ⑨), is denoted here by $\widehat{\mathrm{TE}}_{Y\to X EA}$. The main difference with our $\widehat{\mathrm{TE}}_{Y\to X p1}$ estimator is that our algorithm uses a different length for each sub-dimension within a variable, rather than one length for all sub-dimensions within the variable (which is the approach of the extended algorithm). We introduced this approach to make the tightest possible (hyper-)rectangle around the k nearest neighbors. $\widehat{\mathrm{TE}}_{Y\to X EA}$ is expressed as follows:

$$\begin{aligned}\widehat{\mathrm{TE}}_{Y\to X EA} =& \frac{1}{N}\sum_{i=1}^{N}(\psi(k) - \frac{2}{k} + \psi(l_{X^-,i}) - \psi(l_{(X^p,X^-),i}) \\ &+ \frac{1}{l_{(X^p,X^-),i}} - \psi(l_{(X^-,Y^-),i}) + \frac{1}{l_{(X^-,Y^-),i}}).\end{aligned} \tag{38}$$

In the experimental part, this estimator is marked as the "extended algorithm". It differs from Equation (36) in two ways. Firstly, the first summation on the right hand-side of Equation (36) does not exist. Secondly, compared with Equation (36), the numbers of neighbors $k_{X^-,i}$, $k_{(X^p,X^-),i}$ and $k_{(X^-,Y^-),i}$ included in the rectangular boxes, as explained in Section 3.1, are replaced respectively with $l_{X^-,i}$, $l_{(X^p,X^-),i}$ and $l_{(X^-,Y^-),i}$, which are obtained differently. More precisely, Step (2) in the above algorithm becomes:

(2') For each point (x_i^p, x_i^-) in subspace $\mathcal{S}_{X^p,X^-}$, $l_{(X^p,X^-),i}$ is the number of points falling within a (hyper-)rectangle equal to the Cartesian product of two (hyper-)cubes, the first one in $\mathcal{S}_{X^p}$ and the second one in $\mathcal{S}_{X^-}$, whose edge lengths are equal, respectively, to $d_{x_i^p}^{\max} = 2 \times \max\left\{\|x_k^p - x_i^p\| : (x^p, x^-, y^-)_k \in \chi^k_{(x^p,x^-,y^-)_i}\right\}$ and $d_{x_i^-}^{\max} = 2 \times \max\left\{\left\|x_k^- - x_i^-\right\| : (x^p, x^-, y^-)_k \in \chi^k_{(x^p,x^-,y^-)_i}\right\}$, *i.e.*, $l_{(X^p,X^-),i} = card\left\{\left(x_j^p, x_i^-\right) : j \in \{\{1,\ldots,N\} - \{i\}\} \;\&\; \left\|x_j^p - x_i^p\right\| \le d_{x_i^p}^{\max} \;\&\; \left\|x_j^- - x_i^-\right\| \le d_{x_i^-}^{\max}\right\}$. Denote by $v_{(X^p,X^-),i}$ the volume of this (hyper-)rectangle. Repeat the same procedure in subspaces $\mathcal{S}_{X^-,Y^-}$ and $\mathcal{S}_{X^-}$.

Note that the important difference between the construction of the neighborhoods used in $\widehat{\mathrm{TE}}_{Y\to X EA}$ and in $\widehat{\mathrm{TE}}_{Y\to X p1}$ is that, for the first case, the minimum neighborhood, including the k neighbors, is constrained to be a Cartesian product of (hyper-)cubes and, in the second case, this neighborhood is a (hyper-)rectangle whose edge lengths can be completely different.

5. Experimental Results

In the experiments, we tested both Gaussian IID and Gaussian AR models to compare and validate the performance of the TE estimators proposed in the previous section. For a complete comparison,

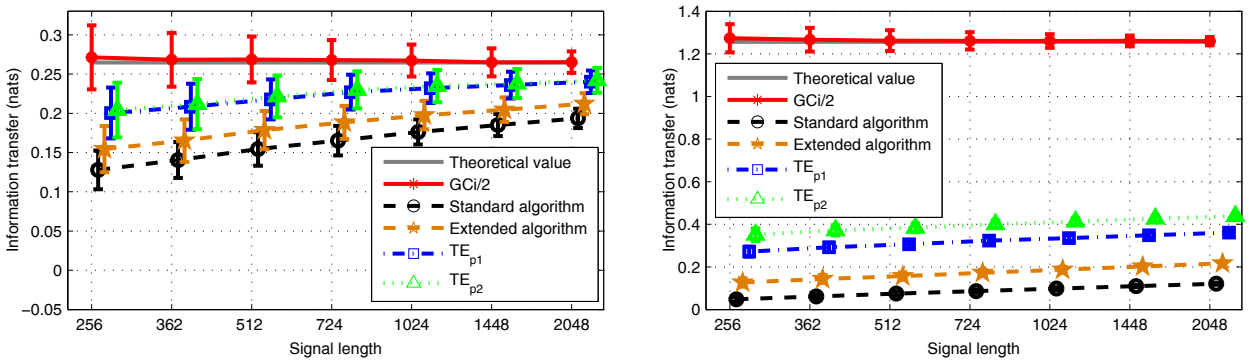
beyond the theoretical value of TE, we also computed the Granger causality index as a reference (as indicated previously, in the case of Gaussian signals TE and Granger causality index are equivalent up to a factor of two; see Appendix H). In each following figure, GCi/2 corresponds to the Granger causality index divided by two; TE estimated by the free TRENTOOL toolbox (corresponding to Equation (21)) is marked as the standard algorithm; that estimated by JIDT (corresponding to Equation (38)) is marked as the extended algorithm; TE_{p1} is the TE estimator given by Equation (36); and TE_{p2} is the TE estimator given by Equation (37). For all of the following results, the statistical means and the standard deviations of the different estimators have been estimated using an averaging on 200 trials.

5.1. Gaussian IID Random Processes

The first model we tested, named Model 1, is formulated as follows:

$$X_t = aY_t + bZ_t + W_t, \quad W_t \in \mathbb{R}, Y \in \mathbb{R}^{d_Y}, Z \in \mathbb{R}^{d_Z}, \tag{39}$$

where $Y_t \sim \mathcal{N}(0, C_Y)$, $Z_t \sim \mathcal{N}(0, C_Z)$, $W_t \sim \mathcal{N}(0, \sigma_W^2)$, the three processes Y, Z, and W being mutually independent. The triplet (X_t, Y_t, Z_t) corresponds to the triplet $\left(X_i^p, X_i^-, Y_i^-\right)$ introduced previously. C_U is a Toeplitz matrix with the first line equal to $[1, \alpha, \ldots, \alpha^{d_U - 1}]$. For the matrix C_Y, we chose $\alpha = 0.5$, and for C_Z, $\alpha = 0.2$. The standard deviation σ_W was set to 0.5. The vectors a and b were such that $a = 0.1 * [1, 2, \ldots, d_Y]$ and $b = 0.1 * [d_Z, d_Z - 1, \ldots, 1]$. With this model, we aimed at estimating $\mathcal{H}(X|Y) - \mathcal{H}(X|Y, Z)$ to test if the knowledge of signals Y and Z could improve the prediction of X compared to only the knowledge of Y.



(a) (b)

Figure 3. Information transfer from Z to X (Model 1) estimated for two different dimensions with $k = 8$. The figure displays the mean values and the standard deviations: **(a)** $d_Y = d_Z = 3$; **(b)** $d_Y = d_Z = 8$.

Results are reported in Figure 3 where the dimensions d_Y and d_Z are identical. We observe that, for a low dimension and a sufficient number of neighbors (Figure 3a), all TE estimators tend all the more to the theoretical value (around 0.26) that the length of the signals is large, the best estimation being obtained by the two new estimators. Compared to Granger causality, these estimators display a greater bias, but a lower variance. Due to the "curse of dimensionality", with an increasing dimension (see Figure 3b), it

becomes much more difficult to obtain an accurate estimation of TE. For a high dimension, all estimators reveal a non-negligible bias, even if the two new estimators still behave better than the two reference ones (standard and extended algorithms).

5.2. Vectorial AR Models

In the second experiment, two AR models integrating either two or three signals have been tested. The first vectorial AR model (named Model 2) we tested was as follows:

$$\begin{cases} x_t = 0.45\sqrt{2}x_{t-1} - 0.9x_{t-2} - 0.6y_{t-2} + e_{x,t} \\ y_t = 0.6x_{t-2} - 0.175\sqrt{2}y_{t-1} + 0.55\sqrt{2}y_{t-2} + e_{y,t}. \end{cases} \tag{40}$$

The second vectorial AR model (named Model 3) was given by:

$$\begin{cases} x_t = -0.25x_{t-2} - 0.35y_{t-2} + 0.35z_{t-2} + e_{x,t} \\ y_t = -0.5x_{t-1} + 0.25y_{t-1} - 0.5z_{t-3} + e_{y,t} \\ z_t = -0.6x_{t-2} - 0.7y_{t-2} - 0.2z_{t-2} + e_{z,t}. \end{cases} \tag{41}$$

For both models, e_x, e_y and e_z denote realizations of independent white Gaussian noises with zero mean and a variance of 0.1. As previously, we display in the following figures not only the theoretical value of TE, but also the Granger causality index for comparison. In this experiment, the prediction orders m and n were equal to the corresponding regression orders of the AR models. For example, when estimating $\mathrm{TE}_{Y\to X}$, we set $m = 2$, $n = 2$, and $\left(X_i^p, X_i^-, Y_i^-\right)$ corresponds to $\left(X_{i+1}, X_i^{(2)}, Y_i^{(2)}\right)$.

For Figures 4 and 5, the number k of neighbors was fixed to eight, whereas, in Figure 6, this number was set to four and three (respectively Figures 6a,b) to show the influence of this parameter. Figures 4 and 6 are related to Model 2, and Figure 5 is related to Model 3.

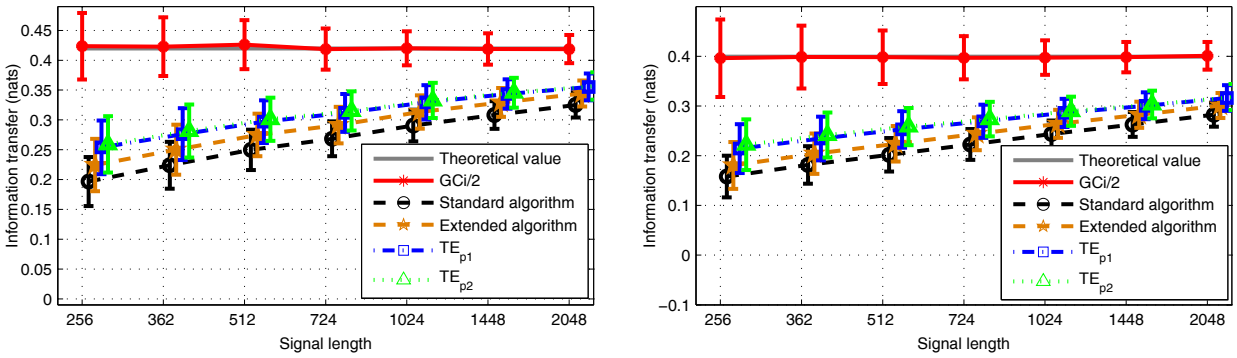


Figure 4. Information transfer (Model 2), mean values and standard deviations, $k = 8$. (**a**) From X to Y; (**b**) from Y to X.

As previously, for large values of k (*cf.* Figures 4 and 5), we observe that the four TE estimators converge towards the theoretical value. This result is all the more true when the signal length increases. As expected in such linear models, Granger causality outperforms the TE estimators at the expense

of a slightly larger variance. Contrary to Granger causality, TE estimators are clearly more impacted by the signal length, even if their standard deviations remain lower. Here, again, when comparing the different TE estimators, it appears that the two new estimators achieve improved behavior compared to the standard and extended algorithms for large k.

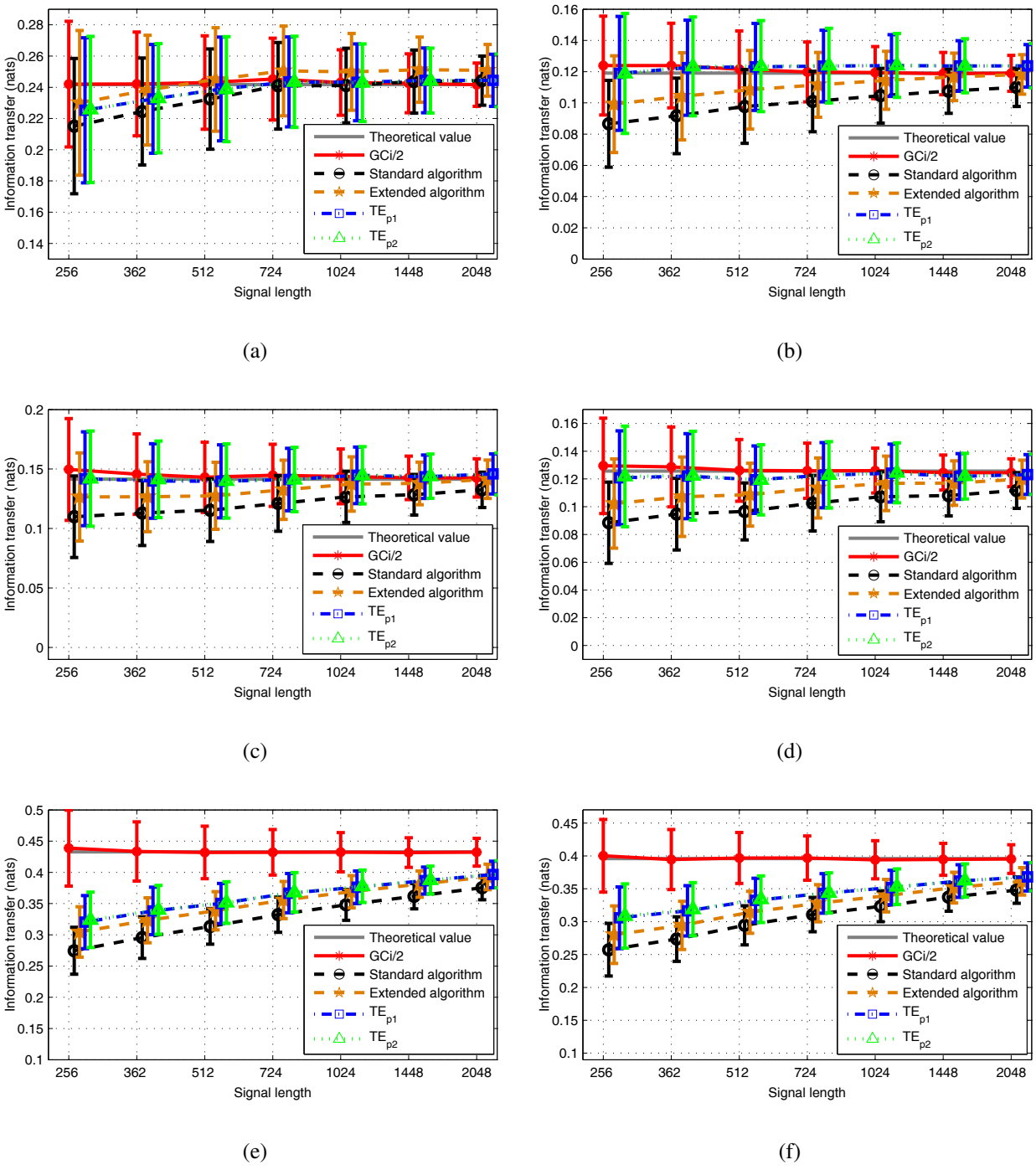
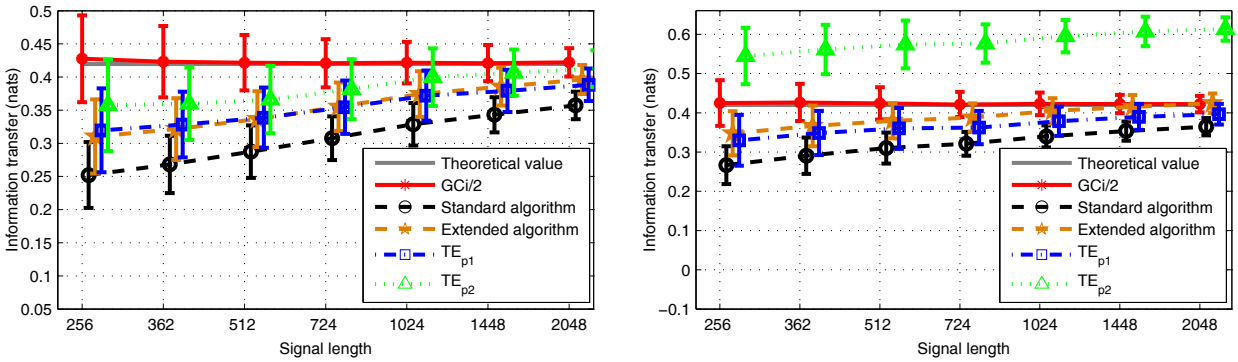


Figure 5. Information transfer (Model 3), mean values and standard deviations, $k = 8$. (**a**) From X to Y; (**b**) from Y to X; (**c**) from X to Z; (**d**) from Z to X; (**e**) from Y to Z; (**f**) from Z to Y.



(a) (b)

Figure 6. Information transfer from X to Y (Model 2), mean values and standard deviations: (**a**) $k = 4$; (**b**) $k = 3$.

In the scope of k-NN algorithms, the choice of k must be a tradeoff between the estimation of bias and variance. Globally, when the value of k decreases, the bias decreases for the standard and extended algorithms and for the new estimator TE_{p1}. Now, for the second proposed estimator TE_{p2}, it is much more sensitive to the number of neighbors (as can be seen when comparing Figures 4 and 6). As shown in Figures 3 to 5, the results obtained using TE_{p2} and TE_{p1} are quite comparable when the value of k is large ($k = 8$). Now, when the number of neighbors decreases, the second estimator we proposed, TE_{p2}, is much less reliable than all of the other ones (Figure 6). Concerning the variance, it remains relatively stable when the number of neighbors falls from eight to three, and in this case, the extended algorithm, which displays a slightly lower bias, may be preferred.

When using $k = 8$, a possible interpretation of getting a lower bias with our algorithms could be that, once we are looking at a large enough number of k nearest neighbors, there is enough opportunity for the use of different lengths on the sub-dimensions of the (hyper-)rectangle to make a difference to the results, whereas with $k = 3$, there is less opportunity.

To investigate the impact on the dispersion (estimation error standard deviation) of (i) the estimation method and (ii) the number of neighbors, we display in Figures 7a,b the boxplots of the absolute values of the centered estimation errors (AVCE) corresponding to experiments reported in Figures 4a and 6b for a 1024-point signal length. These results show that neither the value of k, nor the tested TE estimator dramatically influence the dispersions. More precisely, we used a hypothesis testing procedure (two-sample Kolmogorov–Smirnov goodness-of-fit hypothesis, KSTEST2 in MATLAB) to test if two samples (each with 200 trials) of AVCE are drawn from the same underlying continuous population or not. The tested hypothesis corresponds to non-identical distributions and is denoted $H = 1$, and $H = 0$ corresponds to the rejection of this hypothesis. The confidence level was set to 0.05.

(1) Influence of the method:

   (a) Test between the standard algorithm and TE_{p1} in Figure 7a: $H = 0$, p-value = 0.69 $\rightarrow$ no influence

   (b) Test between the extended algorithm and TE_{p1} in Figure 7a: $H = 0$, p-value = 0.91 $\rightarrow$ no influence

(c) Test between the standard algorithm and TE_{p1} in Figure 7b: $H = 0$, p-value = 0.081 $\rightarrow$ no influence

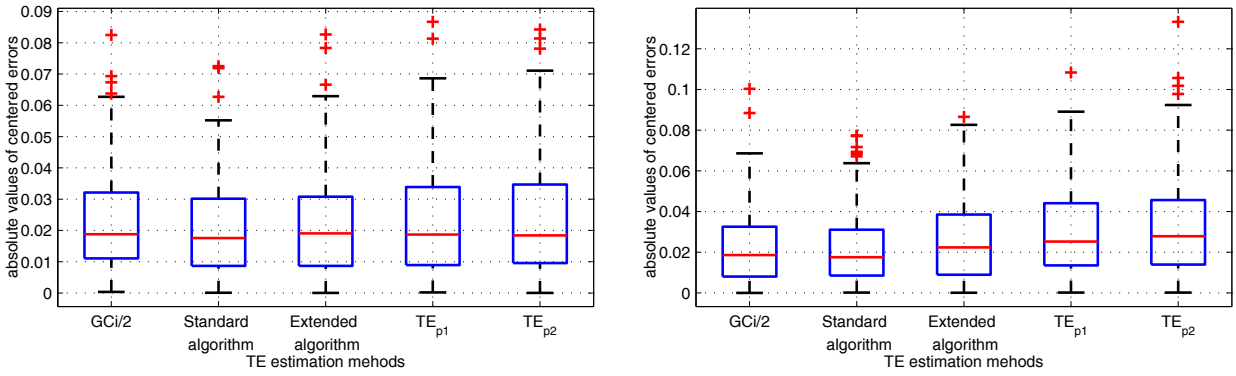
(d) Test between the extended algorithm and TE_{p1} in Figure 7b: $H = 1$, p-value = 0.018 $\rightarrow$ influence exists.

(2) Influence of the neighbors' number k:

(a) Test between $k = 8$ (Figure 7a) and $k = 3$ (Figure 7b) for the standard algorithm: $H = 0$, p-value = 0.97 $\rightarrow$ no influence

(b) Test between $k = 8$ (Figure 7a) and $k = 3$ (Figure 7b) for TE_{p1}: $H = 0$, p-value = 0.97 $\rightarrow$ no influence.

For these six tested cases, the only case where a difference between distributions (and so, between the dispersions) corresponds to a different distribution is when comparing the extended algorithm and TE_{p1} in Figure 7b.



(a) (b)

Figure 7. Box plots of the centered errors obtained with the five methods for Model 2, $X \rightarrow Y$: (**a**) $k = 8$ (corresponding to Figure 4a); (**b**) $k = 3$ (corresponding to Figure 6b).

6. Discussion and Summary

In the computation of k-NN based estimators, the most time-consuming part is the procedure of nearest neighbor searching. Compared to Equations (10) and (16), Equations (22) and (35) involve supplementary information, such as the maximum distance of the first k-th nearest neighbor in each dimension and the number of points on the border. However, most currently used neighbor searching algorithms, such as k-d tree (k-dimensional tree) and ATRIA (A TRiangle Inequality based Algorithm) [40], provide not only information on the k-th neighbor, but also on the first $(k-1)$ nearest neighbors. Therefore, in terms of computation cost, there is no significant difference among the three TE estimators (Boxes ⑦, ⑧, ⑨, ⑩ in Figure 1).

In this contribution, we discussed TE estimation based on k-NN techniques. The estimation of TE is always an important issue, especially in neuroscience, where getting large amounts of stationary data is problematic. The widely-used k-NN technique has been proven to be a good choice for the

estimation of information theoretical measurement. In this work, we first investigated the estimation of Shannon entropy based on the k-NN technique involving a rectangular neighboring region and introduced two different k-NN entropy estimators. We derived mathematically these new entropy estimators by extending the results and methodology developed in [24] and [36]. Given the new entropy estimators, two novel TE estimators have been proposed, implying no extra computation cost compared to existing similar k-NN algorithm. To validate the performance of these estimators, we considered different simulated models and compared the new estimators with the two TE estimators available in the free TRENTOOL and JIDT toolboxes, respectively, and which are extensions of two Kraskov–Stögbauer–Grassberger (KSG) MI estimators, based respectively on (hyper-)cubic and (hyper-)rectangular neighborhoods.

Under the Gaussian assumption, experimental results showed the effectiveness of the new estimators under the IID assumption, as well as for time-correlated AR signals in comparison with the standard KSG algorithm estimator. This conclusion still holds when comparing the new algorithms with the extended KSG estimator. Globally, all TE estimators satisfactorily converge to the theoretical TE value, *i.e.*, to half the value of the Granger causality, while the newly proposed TE estimators showed lower bias for k sufficiently large (in comparison with the reference TE estimators) with comparable variances estimation errors.

As the variance remains relatively stable when the number of neighbors falls from eight to three, in this case, the extended algorithm, which displays a slightly lower bias, may be preferred.

Now, one of the new TE estimators suffered from noticeable error when the number of neighbors was small. Some experiments allowed us to verify that this issue already exists when estimating the entropy of a random vector: when the number of neighbors k falls below the dimension d, then the bias drastically increases. More details on this phenomenon are given in Appendix I.

As expected, experiments with Model 1 showed that all three TE estimators under examination suffered from the "curse of dimensionality", which made it difficult to obtain accurate estimation of TE with high dimension data. In this contribution, we do not present the preliminary results that we obtained when simulating a nonlinear version of Model 1, for which the three variables X_t, Y_t and Z_t were scalar and their joint law was non-Gaussian, because a random nonlinear transformation was used to compute X_t from Y_t, Z_t. For this model, we computed the theoretical TE (numerically, with good precision) and tuned the parameters to obtain a strong coupling between X_t and Z_t. The theoretical Granger causality index was equal to zero. We observed the same issue as that pointed out in [41], *i.e.*, a very slow convergence of the estimator when the number of observations increases, and noticed that the four estimators $\widehat{\mathrm{TE}_{Y\to XSA}}$, $\widehat{\mathrm{TE}_{Y\to XEA}}$, $\widehat{\mathrm{TE}_{Y\to Xp1}}$ and $\widehat{\mathrm{TE}_{Y\to Xp2}}$, revealed very close performance. In this difficult case, our two methods do not outperform the existing ones. Probably, for this type of strong coupling, further improvement must be considered at the expense of an increasing computational complexity, as that proposed in [41].

This work is a first step in a more general context of connectivity investigation for neurophysiological activities obtained either from nonlinear physiological models or from clinical recordings. In this context, partial TE has also to be considered, and future work would address a comparison of the techniques presented in this contribution in terms of bias and variance. Moreover, considering the

practical importance to know statistical distributions of the different TE estimators for independent channels, this point should be also addressed.

Author Contributions

All authors have read and approved the final manuscript.

Appendix

A. Mathematical Expression of Transfer Entropy for Continuous Probability Distributions

Here, we consider that the joint probability measure $\mathcal{P}_{X_i^p,X_i^-,Y_i^-}$ is absolutely continuous (with respect to the Lebesgue measure in $\mathbb{R}^{m+n+1}$ denoted by μ^{m+n+1}) with the corresponding density:

$$p_{X_i^p,X_i^-,Y_i^-}\left(x_i^p,x_i^-,y_i^-\right)=\frac{\mathrm{d}\mathcal{P}_{X_i^p,X_i^-,Y_i^-}\left(x_i^p,x_i^-,y_i^-\right)}{\mathrm{d}\mu^{n+m+1}\left(x_i^p,x_i^-,y_i^-\right)}. \tag{42}$$

Then, we are sure that the two following conditional densities probability functions exist:

$$\begin{aligned} p_{X_i^p|X_i^-}\left(x_i^p|x_i^-\right)&=\frac{\mathrm{d}\mathcal{P}_{X_i^p|X_i^-}\left(x_i^p|x_i^-\right)}{\mathrm{d}\mu^1\left(x_i^p\right)}\\ p_{X_i^p|X_i^-,Y_i^-}\left(x_i^p|x_i^-,y_i^-\right)&=\frac{\mathrm{d}\mathcal{P}_{X_i^p|X_i^-,Y_i^-}\left(x_i^p|x_i^-,y_i^-\right)}{\mathrm{d}\mu^1\left(x_i^p\right)}. \end{aligned} \tag{43}$$

and Equation (3) yields to:

$$\begin{aligned} \mathrm{TE}_{Y\to X,i}&=\int_{\mathbb{R}^{m+n+1}}p_{X^p,X_i^-,Y_i^-}\left(x_i^p,x_i^-,y_i^-\right)\log\left[\frac{p_{X^p|X_i^-,Y_i^-}\left(x_i^p|x_i^-,y_i^-\right)}{p_{X^p|X_i^-}\left(x_i^p|x_i^-\right)}\right]\mathrm{d}x_i^p\mathrm{d}x_i^-y_i^-\\ &=\int_{\mathbb{R}^{m+n+1}}p_{X^p,X_i^-,Y_i^-}\left(x_i^p,x_i^-,y_i^-\right)\log\left[\frac{p_{X^p,X_i^-,Y_i^-}\left(x_i^p,x_i^-,y_i^-\right)p_{X_i^-}\left(x_i^-\right)}{p_{X_i^-,Y_i^-}\left(x_i^-,y_i^-\right)p_{X^p,X_i^-}\left(x_i^p,x_i^-\right)}\right]\mathrm{d}x_i^p\mathrm{d}x_i^-y_i^-. \end{aligned} \tag{44}$$

Equation (44) can be rewritten:

$$\begin{aligned} \mathrm{TE}_{Y\to X,i}=&-E\left[\log\left(p_{X_i^-,Y_i^-}\left(X_i^-,Y_i^-\right)\right)\right]-E\left[\log\left(p_{X_i^p,X_i^-}\left(X_i^p,X_i^-\right)\right)\right]\\ &+E\left[\log\left(p_{X_i^p,X_i^-,Y_i^-}\left(X_i^p,X_i^-,Y_i^-\right)\right)\right]+E\left[\log\left(p_{X_i^-}\left(X_i^-\right)\right)\right]. \end{aligned} \tag{45}$$

B. Basic Structure of TE Estimators

From Equation (8), assuming that X and Y are jointly strongly ergodic leads to:

$$\begin{aligned} \mathrm{TE}_{Y\to X}=\lim_{N\to\infty}\frac{1}{N}\sum_{i=1,\ldots,N}&\left[-\log\left(p_{X_i^-,Y_i^-}\left(X_i^-,Y_i^-\right)\right)-\log\left(p_{X_i^p,X_i^-}\left(X_i^p,X_i^-\right)\right)\right.\\ &\left.+\log\left(p_{X_i^p,X_i^-,Y_i^-}\left(X_i^p,X_i^-,Y_i^-\right)\right)+\log\left(p_{X_i^-}\left(X_i^-\right)\right)\right], \end{aligned} \tag{46}$$

where the convergence holds with probability one. Hence, as a function of an observed occurrence (x_i, y_i), $i = 1, \ldots, N$, of (X_i, Y_i), $i = 1, \ldots, N$, a standard estimation $\widehat{\mathrm{TE}_{Y\to X}}$ of $\mathrm{TE}_{Y\to X}$ is given by:

$$\begin{aligned}\widehat{\mathrm{TE}_{Y\to X}} &= \widehat{\mathcal{H}\left(X^-, Y^-\right)} + \widehat{\mathcal{H}\left(X^p, X^-\right)} - \widehat{\mathcal{H}\left(X^p, X^-, Y^-\right)} - \widehat{\mathcal{H}\left(X^-\right)} \\ &= -\frac{1}{N}\sum_{n=1}^{N}\widehat{\log\left(p_{U_1}\left(u_{1n}\right)\right)} - \frac{1}{N}\sum_{n=1}^{N}\widehat{\log\left(p_{U_2}\left(u_{2n}\right)\right)} + \frac{1}{N}\sum_{n=1}^{N}\widehat{\log\left(p_{U_3}\left(u_{3n}\right)\right)} \\ &\quad + \frac{1}{N}\sum_{n=1}^{N}\widehat{\log\left(p_{U_4}\left(u_{4n}\right)\right)},\end{aligned} \tag{47}$$

where U_1, U_2, U_3 and U_4 stand respectively for (X^-, Y^-), (X^p, X^-), (X^p, X^-, Y^-) and X^-.

C. The Bias of Singh's Estimator

Let us consider the equalities $E\left(T_1\right) = -E\left[\log\left(\widehat{p_X\left(X_1\right)}\right)\right] = -E\left[\log\left(\frac{k}{NV_1}\right)\right]$ where V_1 is the random volume for which v_1 is an outcome. Conditionally to $X_1 = x_1$, if we have $\frac{k}{NV_1} \xrightarrow[N\to\infty]{pr} p_X\left(x_1\right)$ (convergence in probability), then $E\left(T_1/X_1 = x_1\right) \xrightarrow[N\to\infty]{} -\log\left(p_X\left(x_1\right)\right)$, and by deconditioning, we obtain $E\left(T_1\right) \xrightarrow[N\to\infty]{} -E\left(\log\left(p_X\left(X_1\right)\right)\right) = \mathcal{H}(X)$. Therefore, if $\frac{k}{NV_1} \xrightarrow[N\to\infty]{pr} p_X\left(x_1\right)$, the estimation of $\mathcal{H}(X)$ is asymptotically unbiased. Here, this convergence in probability does not hold, even if we assume that $E\left(\frac{k}{NV_1}\right) \xrightarrow[N\to\infty]{} p_X\left(x_1\right)$ (one order mean convergence), because we do not have $\mathrm{var}\left(\frac{k}{NV_1}\right) \xrightarrow[N\to\infty]{} 0$. The ratio $\frac{k}{NV_1}$ remains fluctuating when $N \to \infty$, because the ratio $\frac{\sqrt{\mathrm{var}(V_1)}}{E(V_1)}$ does not tend to zero, even if V_1 tends to be smaller: when N increases, the neighborhoods become smaller and smaller, but continue to 'fluctuate'. This explains informally (see [37] for a more detailed analysis) why the naive estimator given by Equation (17) is not asymptotically unbiased. It is interesting to note that the Kozachenko–Leonenko entropy estimator avoids this problem, and so it does not need any bias subtraction.

D. Derivation of Equation (22)

As illustrated in Figure 2, for $d = 2$, there are two cases to be distinguished: (1) ε_x and ε_y are determined by the same point; (2) ε_x and ε_y are determined by distinct points.

Considering the probability density $q_{i,k}\left(\epsilon_x, \epsilon_y\right), \left(\epsilon_x, \epsilon_y\right) \in \mathbb{R}^2$ of the pair of random sizes $(\varepsilon_x, \varepsilon_y)$ (along x and y, respectively), we can extend it to the case $d > 2$. Hence, let us denote by $q^d_{x_i,k}\left(\varepsilon_1, \ldots, \varepsilon_d\right)$, $\left(\varepsilon_1, \ldots, \varepsilon_d\right) \in \mathbb{R}^d$ the probability density (conditional to $X_i = x_i$) of the d-dimensional random vector whose d components are respectively the d random sizes of the (hyper-)rectangle built from the random k nearest neighbors, and denote by $h^{x_i}\left(\varepsilon_1, \ldots, \varepsilon_d\right) = \int_{u\in\mathcal{D}^{\varepsilon_1,\ldots,\varepsilon_d}_{x_i}} \mathrm{d}P_X\left(u\right)$ the probability mass (conditional to $X_i = x_i$) of the random (hyper-)rectangle $\mathcal{D}^{\epsilon_1,\ldots,\epsilon_d}_{x_i}$. In [24], the equality $E\left[\log\left(h^{x_i}\left(D_{x_i,k}\right)\right)\right] = \psi(k) - \psi(N)$ obtained for an (hyper-)cube is extended for the case $d > 2$ to:

$$E\left[\log\left(h^{x_i}\left(\epsilon_1, \ldots, \epsilon_d\right)\right)\right] = \psi(k) - \frac{d-1}{k} - \psi(N). \tag{48}$$

Therefore, if p_X is approximately constant on $\mathcal{D}^{\varepsilon_1,\ldots,\varepsilon_d}_{x_i}$, we have:

$$h^{x_i}\left(\varepsilon_1, \ldots, \varepsilon_d\right) \simeq v_i p_X\left(x_i\right), \tag{49}$$

where $v_i = \int_{\mathcal{D}_{x_i}^{\varepsilon_1,\ldots,\varepsilon_d}} \mathrm{d}\mu^d(\xi)$ is the volume of the (hyper-)rectangle, and we obtain:

$$\log p_X(x_i) \approx \psi(k) - \psi(N) - \frac{d-1}{k} - \log(v_i). \tag{50}$$

Finally, by taking the experimental mean of the right term in Equation (50), we obtain an estimation of the expectation $E[\log p_X(X)]$, *i.e.*,:

$$\widehat{\mathcal{H}(X)} = -\psi(k) + \psi(N) + \frac{d-1}{k} + \frac{1}{N}\sum_{i=1}^{N} \log(v_i). \tag{51}$$

E. Proof of Property 1

Let us introduce the (hyper-)rectangle $\mathcal{D}_{x_1}^{\epsilon'_1,\ldots,\epsilon'_d}$ centered on x_1 for which the random sizes along the d directions are defined by $(\epsilon'_1,\ldots,\epsilon'_d) = (\epsilon_1,\ldots,\epsilon_d) \times \left(\frac{v_r}{\epsilon_1\times\ldots\times\epsilon_d}\right)^{1/d}$, so that $\mathcal{D}_{x_1}^{\epsilon'_1,\ldots,\epsilon'_d}$ and $\mathcal{D}_{x_1}^{\epsilon_1,\ldots,\epsilon_d}$ are homothetic and $\mathcal{D}_{x_1}^{\epsilon'_1,\ldots,\epsilon'_d}$ has a (hyper-)volume constrained to the value v_r. We have:

$$\int_{x\in\mathcal{D}_{x_1}^{\varepsilon_1,\ldots,\varepsilon_d}} \mathrm{d}\mu^d(x) > v_r \Leftrightarrow \mathcal{D}_{x_1}^{\varepsilon'_1,\ldots,\varepsilon'_d} \subset \mathcal{D}_{x_1}^{\varepsilon_1,\ldots,\varepsilon_d} \Leftrightarrow card\left\{x_j : x_j \in \mathcal{D}_{x_1}^{\varepsilon'_1,\ldots,\varepsilon'_d}\right\} \leq k - \xi_1, \tag{52}$$

where the first equivalence (the inclusion is a strict inclusion) is clearly implied by the construction of $\mathcal{D}_{x_1}^{\epsilon'_1,\ldots,\epsilon'_d}$ and the second equivalence expresses the fact that the (hyper-)volume of $\mathcal{D}_{x_1}^{\epsilon_1,\ldots,\epsilon_d}$ is larger than v_r if and only if the normalized domain $\mathcal{D}_{x_1}^{\epsilon'_1,\ldots,\epsilon'_d}$ does not contain more than $(k-\xi_1)$ points x_j (as ξ_1 of them are on the border of $\mathcal{D}_{x_1}^{\epsilon_1,\ldots,\epsilon_d}$, which is necessarily not included in $\mathcal{D}_{x_1}^{\epsilon'_1,\ldots,\epsilon'_d}$). These equivalences imply the equalities between conditional probability values:

$$\begin{aligned}\mathcal{P}(T_1 > r | X_1 = x_1, \Xi_1 = \xi_1) &= \mathcal{P}\left(\log\left(\frac{NV_1}{\widetilde{K}_1}\right) > r | X_1 = x_1, \Xi_1 = \xi_1\right)\\ &= \mathcal{P}(V_1 > v_r | X_1 = x_1, \Xi_1 = \xi_1)\\ &= \mathcal{P}\left(card\left\{X_j : X_j \in \mathcal{D}_{x_1}^{\epsilon'_1,\ldots,\epsilon'_d}\right\} \leq k - \xi_1\right).\end{aligned} \tag{53}$$

Only $(N-1-\xi_1)$ events $\left\{X_j : X_j \in \mathcal{D}_{x_1}^{\epsilon'_1,\ldots,\epsilon'_d}\right\}$ are to be considered, because the variable X_1 and the ξ_1 variable(s) on the border of $\mathcal{D}_{x_1}^{\epsilon_1,\ldots,\epsilon_d}$ must be discarded. Moreover, these events are independent. Hence, the probability value in (53) can be developed as follows:

$$\begin{aligned}\mathcal{P}(T_1 > r | X_1 = x_1, \Xi_1 = \xi_1) \simeq \sum_{i=0}^{k-\xi_1} \binom{N-\xi_1-1}{i} &\left(\mathcal{P}\left(X \in \mathcal{D}_{x_1}^{\epsilon'_1,\ldots,\epsilon'_d}\right)\right)^i\\ &\left(1 - \mathcal{P}\left(X \in \mathcal{D}_{x_1}^{\epsilon'_1,\ldots,\epsilon'_d}\right)\right)^{N-\xi_1-1-i}.\end{aligned} \tag{54}$$

If $p_X(x_1)$ is approximately constant on $\mathcal{D}_{x_1}^{\epsilon'_1,\ldots,\epsilon'_d}$, we have $\mathcal{P}\left(X \in \mathcal{D}_{x_1}^{\epsilon'_1,\ldots,\epsilon'_d}\right) \simeq p_X(x_1)v_r$ (note that the randomness of $(\epsilon'_1,\ldots,\epsilon'_d)$ does not influence this approximation as the (hyper-)volume of $\mathcal{D}_{x_1}^{\epsilon'_1,\ldots,\epsilon'_d}$ is imposed to be equal to v_r). Finally, we can write:

$$\mathcal{P}(T_1 > r | X_1 = x_1, \Xi_1 = \xi_1) \simeq \sum_{i=0}^{k-\xi_1} \binom{N-\xi_1-1}{i} (p_X(x_1)v_r)^i (1 - p_X(x_1)v_r)^{N-\xi_1-1-i}. \tag{55}$$

F. Derivation of Equation (32)

With $\mathcal{P}(T_1 \leq r|X_1 = x_1, \Xi_1 = \xi_1) = 1 - \mathcal{P}(T_1 > r|X_1 = x_1, \Xi_1 = \xi_1)$, we take the derivative of $\mathcal{P}(T_1 \leq r|X_1 = x_1, \Xi_1 = \xi_1)$ to get the conditional density function of T_1:

$$\begin{aligned}
&\mathcal{P}'(T_1 \leq r|X_1 = x_1, \Xi_1 = \xi_1) \\
&= -\mathcal{P}'(T_1 > r|X_1 = x_1, \Xi_1 = \xi_1) \\
&= -\left[\sum_{i=0}^{k-\xi_1} \frac{[\tilde{k}_1 p_X(x_1)e^r]^i}{i!} e^{-\tilde{k}_1 p_X(x_1)e^r}\right]' \\
&= -\sum_{i=0}^{k-\xi_1} \left(\left[\frac{[\tilde{k}_1 p_X(x_1)e^r]^i}{i!}\right]' e^{-\tilde{k}_1 p_X(x_1)e^r} + \frac{[\tilde{k}_1 p_X(x_1)e^r]^i}{i!}\left[e^{-\tilde{k}_1 p_X(x_1)e^r}\right]'\right) \\
&= -\sum_{i=0}^{k-\xi_1} \left(\frac{i[\tilde{k}_1 p_X(x_1)e^r]^{i-1}(\tilde{k}_1 p_X(x_1)e^r)}{i!} e^{-\tilde{k}_1 p_X(x_1)e^r} + \frac{[\tilde{k}_1 p_X(x_1)e^r]^i}{i!} e^{-\tilde{k}_1 p_X(x_1)e^r}(-\tilde{k}_1 p_X(x_1)e^r)\right) \\
&= -\sum_{i=0}^{k-\xi_1} e^{-\tilde{k}_1 p_X(x_1)e^r} \left(\frac{[\tilde{k}_1 p_X(x_1)e^r]^i}{(i-1)!} - \frac{[\tilde{k}_1 p_X(x_1)e^r]^{i+1}}{i!}\right).
\end{aligned} \tag{56}$$

Defining:

$$a(i) = \frac{[\tilde{k}_1 p_X(x_1)e^r]^i}{(i-1)!} \quad \text{and} \quad a(0) = 0, \tag{57}$$

we have:

$$\begin{aligned}
\mathcal{P}'(T_1 \leq r) &= -\sum_{i=0}^{k-\xi_1} e^{-\tilde{k}_1 p_X(x_1)e^r}\left(a(i) - a(i+1)\right) \\
&= -e^{-\tilde{k}_1 p_X(x_1)e^r}\left(a(0) - a(k-\xi_1+1)\right) \\
&= e^{-\tilde{k}_1 p_X(x_1)e^r} a(k-\xi_1+1) \\
&= \frac{\left[\tilde{k}_1 p_X(x_1)e^r\right]^{(k-\xi_1+1)}}{(k-\xi_1)!} e^{-\tilde{k}_1 p_X(x_1)e^r}.
\end{aligned} \tag{58}$$

G. Derivation of Equation (33)

$$\begin{aligned}
\lim_{n\to\infty} E\left(T_1|X_1 = x_1\right) &= \int_{-\infty}^{\infty} r \frac{[\tilde{k}_1 p_X(x_1)e^r]^{(k-\xi_1+1)}}{(k-\xi_1)!} e^{-\tilde{k}_1 p_X(x_1)e^r} \mathrm{d}r \\
&= \int_0^{\infty} \left[\log(z) - \log\left(\tilde{k}_1\right) - \log p_X(x_1)\right] \frac{z^{k-\xi_1}}{(k-\xi_1)!} e^{-z} \mathrm{d}z \\
&= \frac{1}{\Gamma(k-\xi_1+1)} \int_0^{\infty} \left[\log(z) z^{k-\xi_1} e^{-z}\right] \mathrm{d}z - \log\left(\tilde{k}_1\right) - \log p_X(x_1) \\
&= \frac{1}{\Gamma(k-\xi_1+1)} \int_0^{\infty} \left[\log(z) z^{(k-\xi_1+1)-1} e^{-z}\right] \mathrm{d}z - \log\left(\tilde{k}_1\right) - \log p_X(x_1) \\
&= \frac{\Gamma'(k-\xi_1+1)}{\Gamma(k-\xi_1+1)} - \log\left(\tilde{k}_1\right) - \log p_X(x_1) \\
&= \psi(k-\xi_1+1) - \log\left(\tilde{k}_1\right) - \log p_X(x_1).
\end{aligned} \tag{59}$$

H. Transfer Entropy and Granger Causality

TE can be considered as a measurement of the degree to which the history Y^- of the process Y disambiguates the future X^p of X beyond the degree to how its history X^- disambiguates this future [22]. It is an information theoretic implementation of Wiener's principle of observational causality. Hence, TE reveals a natural relation to Granger causality. As is well known, Granger causality emphasizes the concept of reduction of the mean square error of the linear prediction of X_i^p when adding Y_i^- to X_i^- by introducing the Granger causality index:

$$\mathrm{GC}_{Y\to X} = \log\left[\frac{\mathrm{var}\left(lpe_{X_i^p|X_i^-}\right)}{\mathrm{var}\left(lpe_{X_i^p|X_i^-,Y_i^-}\right)}\right], \tag{60}$$

where $lpe_{X_i^p|U}$ is the error when predicting linearly X_i^p from U. TE is framed in terms of the reduction of the Shannon uncertainty (entropy) of the predictive probability distribution. When the probability distribution of $\left(X_i^p, X_i^-, Y_i^-\right)$ is assumed to be Gaussian, TE and Granger causality are entirely equivalent, up to a factor of two [42]:

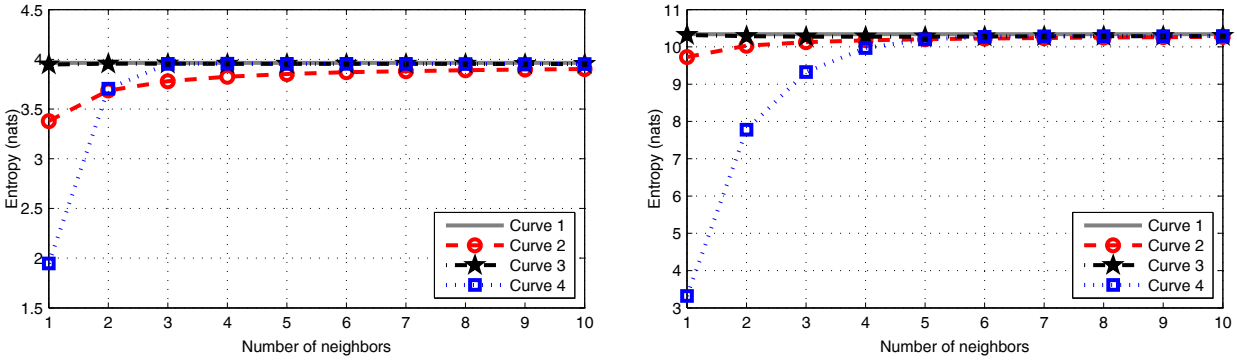
$$\mathrm{TE}_{Y\to X} = \frac{1}{2}\mathrm{GC}_{Y\to X}. \tag{61}$$

Consequently, in the Gaussian case, TE can be easily computed from a statistical second order characterization of $\left(X_i^p, X_i^-, Y_i^-\right)$. This Gaussian assumption obviously holds when the processes Y and X are jointly normally distributed and, more particularly, when they correspond to a Gaussian autoregressive (AR) bivariate process. In [42], Barnett *et al.* discussed the relation between these two causality measures, and this work bridged information-theoretic methods and autoregressive ones.

I. Comparison between Entropy Estimators

Figure 8 displays the values of entropy for a Gaussian d-dimensional vector as a function of the number of neighbors k, for $d = 3$ in Figure 8a and $d = 8$ in Figure 8b, obtained with different estimators. The theoretical entropy value is compared with its estimation from the Kozachenko–Leonenko reference estimator (Equation (10), red circles), its extension (Equation (22), black stars) and the extension of Singh's estimator (Equation (35), blue squares). It appears clearly that, for the extended Singh's estimator, the bias (true value minus estimated value) increases drastically when the number of neighbors decreases under a threshold slightly lower than the dimension d of the vector. This allows us to interpret some apparently surprising results obtained with this estimator in the estimation of TE, as reported in Figure 6b. TE estimation is a sum of four separate vector entropy estimations, $\widehat{\mathrm{TE}_{Y\to X}} = \widehat{\mathcal{H}\left(X^-, Y^-\right)} + \widehat{\mathcal{H}\left(X^p, X^-\right)} - \widehat{\mathcal{H}\left(X^p, X^-, Y^-\right)} - \widehat{\mathcal{H}\left(X^-\right)}$. Here, the dimensions of the four vectors are $d\left(X^-, Y^-\right) = m + n = 4$, $d\left(X^p, X^-\right) = 1 + m = 3$, $d\left(X^p, X^-, Y^-\right) = 1 + m + n = 5$, $d\left(X^-\right) = m = 2$, respectively. Note that, if we denote by X_{M2} and Y_{M2} the two components in Model 2, the general notation $\left(X^p, X^-, Y^-\right)$ corresponds to $\left(Y_{M2}^p, Y_{M2}^-, X_{M2}^-\right)$, because in Figure 6b, the analyzed direction is $X \to Y$ and not the reverse. We see that, when considering the estimation of $\mathcal{H}\left(X^p, X^-, Y^-\right)$, we have $d = 5$ and $k = 3$, which is the imposed neighbors number in the global space. Consequently, from the results shown in Figure 8, we can expect that in Model 2, the

quantity $\mathcal{H}\left(X^p, X^-, Y^-\right)$ will be drastically underestimated. For the other components $\widehat{\mathcal{H}\left(X^-, Y^-\right)}$, $\widehat{\mathcal{H}\left(X^p, X^-\right)}$, $\widehat{\mathcal{H}\left(X^-\right)}$, the numbers of neighbors to consider are generally larger than three (as a consequence of Kraskov's technique, which introduces projected distances) and $d \leq 5$, so that we do not expect any underestimation of these terms. Therefore, globally, when summing the four entropy estimations, the resulting positive bias observed in Figure 6b is understandable.



(a) (b)

Figure 8. Comparison between four entropy estimators: (**a**) $d = 3$; (**b**) $d = 8$. The covariance matrix of the signals is a Toeplitz matrix with first line $\beta^{[0:d-1]}$, where $\beta = 0.5$. "Curve 1" stands for the true value; "Curve 2", "Curve 3" and "Curve 4" correspond to the values of entropy obtained using respectively Equations (10), (22) and (35).

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Schreiber, T. Measuring information transfer. *Phys. Rev. Lett.* **2000**, *85*, doi:10.1103/PhysRevLett.85.461.
2. Gourévitch, B.; Eggermont, J.J. Evaluating information transfer between auditory cortical neurons. *J. Neurophysiol.* **2007**, *97*, 2533–2543.
3. Hlaváčková-Schindler, K.; Paluš, M.; Vejmelka, M.; Bhattacharya, J. Causality detection based on information-theoretic approaches in time series analysis. *Phys. Rep.* **2007**, *441*, 1–46.
4. Sabesan, S.; Narayanan, K.; Prasad, A.; Iasemidis, L.; Spanias, A.; Tsakalis, K. Information flow in coupled nonlinear systems: Application to the epileptic human brain. In *Data Mining in Biomedicine*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 483–503.
5. Ma, C.; Pan, X.; Wang, R.; Sakagami, M. Estimating causal interaction between prefrontal cortex and striatum by transfer entropy. *Cogn. Neurodyn.* **2013**, *7*, 253–261.
6. Vakorin, V.A.; Krakovska, O.A.; McIntosh, A.R. Confounding effects of indirect connections on causality estimation. *J. Neurosci. Methods* **2009**, *184*, 152–160.

7. Yang, C.; Le Bouquin Jeannes, R.; Bellanger, J.J.; Shu, H. A new strategy for model order identification and its application to transfer entropy for EEG signals analysis. *IEEE Trans. Biomed. Eng.* **2013**, *60*, 1318–1327.
8. Zuo, K.; Zhu, J.; Bellanger, J.J.; Jeannès, R.L.B. Adaptive kernels and transfer entropy for neural connectivity analysis in EEG signals. *IRBM* **2013**, *34*, 330–336.
9. Faes, L.; Nollo, G. Bivariate nonlinear prediction to quantify the strength of complex dynamical interactions in short-term cardiovascular variability. *Med. Biol. Eng. Comput.* **2006**, *44*, 383–392.
10. Faes, L.; Nollo, G.; Porta, A. Non-uniform multivariate embedding to assess the information transfer in cardiovascular and cardiorespiratory variability series. *Comput. Biol. Med.* **2012**, *42*, 290–297.
11. Faes, L.; Nollo, G.; Porta, A. Information-based detection of nonlinear Granger causality in multivariate processes via a nonuniform embedding technique. *Phys. Rev. E* **2011**, *83*, 051112.
12. Runge, J.; Heitzig, J.; Petoukhov, V.; Kurths, J. Escaping the curse of dimensionality in estimating multivariate transfer entropy. *Phys. Rev. Lett.* **2012**, *108*, 258701.
13. Duan, P.; Yang, F.; Chen, T.; Shah, S.L. Direct causality detection via the transfer entropy approach. *IEEE Trans. Control Syst. Technol.* **2013**, *21*, 2052–2066.
14. Bauer, M.; Thornhill, N.F.; Meaburn, A. Specifying the directionality of fault propagation paths using transfer entropy. In Proceedings of the 7th International Symposium on Dynamics and Control of Process Systems (DYCOPS 7), Cambridge, MA, USA, 7–9 July 2004; pp. 203–208.
15. Bauer, M.; Cox, J.W.; Caveness, M.H.; Downs, J.J.; Thornhill, N.F. Finding the direction of disturbance propagation in a chemical process using transfer entropy. *IEEE Trans. Control Syst. Technol.* **2007**, *15*, 12–21.
16. Kulp, C.; Tracy, E. The application of the transfer entropy to gappy time series. *Phys. Lett. A* **2009**, *373*, 1261–1267.
17. Overbey, L.; Todd, M. Dynamic system change detection using a modification of the transfer entropy. *J. Sound Vib.* **2009**, *322*, 438–453.
18. Gray, R.M. *Entropy and Information Theory*; Springer: Berlin/Heidelberg, Germany, 2011.
19. Roman, P. *Some Modern Mathematics for Physicists and Other Outsiders: An Introduction to Algebra, Topology, and Functional Analysis*; Elsevier: Amsterdam, The Netherlands, 2014.
20. Kugiumtzis, D. Direct-coupling information measure from nonuniform embedding. *Phys. Rev. E* **2013**, *87*, 062918.
21. 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.
22. Paluš, M.; Komárek, V.; Hrnčíř, Z.; Štěrbová, K. Synchronization as adjustment of information rates: Detection from bivariate time series. *Phys. Rev. E* **2001**, *63*, 046211.
23. Frenzel, S.; Pompe, B. Partial mutual information for coupling analysis of multivariate time series. *Phys. Rev. Lett.* **2007**, *99*, 204101.
24. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. *Phys. Rev. E* **2004**, *69*, 066138.
25. Vicente, R.; Wibral, M.; Lindner, M.; Pipa, G. Transfer entropy—A model-free measure of effective connectivity for the neurosciences. *J. Comput. Neurosci.* **2011**, *30*, 45–67.

26. 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.
27. Wibral, M.; Vicente, R.; Lindner, M. Transfer Entropy in Neuroscience. In *Directed Information Measures in Neuroscience*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 3–36.
28. Gómez-Herrero, G.; Wu, W.; Rutanen, K.; Soriano, M.C.; Pipa, G.; Vicente, R. Assessing coupling dynamics from an ensemble of time series. **2010**, arXiv:1008.0539.
29. Vlachos, I.; Kugiumtzis, D. Nonuniform state-space reconstruction and coupling detection. *Phys. Rev. E* **2010**, *82*, 016207.
30. Lizier, J.T. JIDT: An information-theoretic toolkit for studying the dynamics of complex systems. **2014**, arXiv:1408.3270.
31. MILCA Toolbox. Available online: http://www.ucl.ac.uk/ion/departments/sobell/Research/RLemon/MILCA/MILCA (accessed on 11 June 2015).
32. Wibral, M.; Rahm, B.; Rieder, M.; Lindner, M.; Vicente, R.; Kaiser, J. Transfer entropy in magnetoencephalographic data: Quantifying information flow in cortical and cerebellar networks. *Prog. Biophys. Mol. Biol.* **2011**, *105*, 80–97.
33. Wollstadt, P.; Martínez-Zarzuela, M.; Vicente, R.; Díaz-Pernas, F.J.; Wibral, M. Efficient transfer entropy analysis of non-stationary neural time series. *PLoS One* **2014**, *9*, e102833.
34. Wollstadt, P.; Lindner, M.; Vicente, R.; Wibral, M.; Pampu, N.; Martinez-Zarzuela, M. Trentool Toolbox. Available online: www.trentool.de (accessed on 11 June 2015).
35. 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.
36. Singh, H.; Misra, N.; Hnizdo, V.; Fedorowicz, A.; Demchuk, E. Nearest neighbor estimates of entropy. *Am. J. Math. Manag. Sci.* **2003**, *23*, 301–321.
37. Zhu, J.; Bellanger, J.J.; Shu, H.; Yang, C.; Jeannès, R.L.B. Bias reduction in the estimation of mutual information. *Phys. Rev. E* **2014**, *90*, 052714.
38. Fukunaga, K.; Hostetler, L. Optimization of k nearest neighbor density estimates. *IEEE Trans. Inf. Theory* **1973**, *19*, 320–326.
39. Fukunaga, K. *Introduction to Statistical Pattern Recognition*; Academic Press: Waltham, MA, USA, 1990.
40. Merkwirth, C.; Parlitz, U.; Lauterborn, W. Fast nearest-neighbor searching for nonlinear signal processing. *Phys. Rev. E* **2000**, *62*, 2089–2097.
41. Gao, S.; Steeg, G.V.; Galstyan, A. Efficient Estimation of Mutual Information for Strongly Dependent Variables. In Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS), San Diego, CA, USA, 9–12 May 2015; pp. 277–286.
42. Barnett, L.; Barrett, A.B.; Seth, A.K. Granger causality and transfer entropy are equivalent for Gaussian variables. *Phys. Rev. Lett.* **2009**, *103*, 238701.