entropy

MDPI

Article

On Geometry of Information Flow for Causal Inference

Sudam Surasinghe [1,*] **and Erik M. Bollt** [2,*]

[1] Department of Mathematics, Clarkson University, Potsdam, NY 13699, USA
[2] Department of Electrical and Computer Engineering, Clarkson Center for Complex Systems Science (C^3S^2), Clarkson University, Potsdam, NY 13699, USA
[*] Correspondence: surasinc@clarkson.edu (S.S.); ebollt@clarkson.edu (E.M.B.)

Received: 5 February 2020; Accepted: 27 March 2020 ; Published: 30 March 2020

Abstract: Causal inference is perhaps one of the most fundamental concepts in science, beginning originally from the works of some of the ancient philosophers, through today, but also weaved strongly in current work from statisticians, machine learning experts, and scientists from many other fields. This paper takes the perspective of information flow, which includes the Nobel prize winning work on Granger-causality, and the recently highly popular transfer entropy, these being probabilistic in nature. Our main contribution will be to develop analysis tools that will allow a geometric interpretation of information flow as a causal inference indicated by positive transfer entropy. We will describe the effective dimensionality of an underlying manifold as projected into the outcome space that summarizes information flow. Therefore, contrasting the probabilistic and geometric perspectives, we will introduce a new measure of causal inference based on the fractal correlation dimension conditionally applied to competing explanations of future forecasts, which we will write $GeoC_{y\rightarrow x}$. This avoids some of the boundedness issues that we show exist for the transfer entropy, $T_{y\rightarrow x}$. We will highlight our discussions with data developed from synthetic models of successively more complex nature: these include the Hénon map example, and finally a real physiological example relating breathing and heart rate function.

Keywords: causal inference; transfer entropy; differential entropy; correlation dimension; Pinsker's inequality; Frobenius–Perron operator

1. Introduction

Causation Inference is perhaps one of the most fundamental concepts in science, underlying questions such as "what are the causes of changes in observed variables". Identifying, indeed even defining causal variables of a given observed variable is not an easy task, and these questions date back to the Greeks [1,2]. This includes important contributions from more recent luminaries such as Russel [3], and from philosophy, mathematics, probability, information theory, and computer science. We have written that [4], "a basic question when defining the concept of information flow is to contrast versions of reality for a dynamical system. Either a subcomponent is closed or alternatively there is an outside influence due to another component". Claude Granger's Nobel prize [5] winning work leading to Granger Causality (see also Wiener [6]) formulates causal inference as a concept of quality of forecasts. That is, we ask, does system X provide sufficient information regarding forecasts of future states of system X or are there improved forecasts with observations from system Y? We declare that X is not closed, as it is receiving influence (or information) from system Y, when data from Y improve forecasts of X. Such a reduction of uncertainty perspective of causal inference is not identical to the interventionists' concept of allowing perturbations and experiments to decide what changes indicate influences. This data oriented philosophy of causal inference is especially appropriate when (1) the system is a dynamical system of some form producing data streams in time, and (2) a score of influence may be needed. In particular, contrasting forecasts is the defining concept underlying Granger Causality (G-causality), and it is closely related to the concept of information flow as defined by transfer entropy [7,8], which can be proved as a nonlinear version of Granger's otherwise linear (ARMA) test [9]. In this spirit, we find methods such as Convergent Cross-Mapping method (CCM) [10], and causation entropy (CSE) [11] to disambiguate direct versus indirect influences [11–18]. On the other hand, closely related to information flow are concepts of counter factuals: "what would happen if ..." [19] that are foundational questions for another school leading to the highly successful Pearl "Do-Calculus" built on a specialized variation of Bayesian analysis [20]. These are especially relevant for nondynamical questions (inputs and outputs occur once across populations), such as a typical question of the sort, "why did I get fat" may be premised on inferring probabilities of removing influences of saturated fats and chocolates. However, with concepts of counter-factual analysis in mind, one may argue that Granger is less descriptive of causation inference, but rather more descriptive of information flow. In fact, there is a link between the two notions for so-called "settable" systems under a conditional form of exogeneity [21,22].

This paper focuses on the information flow perspective, which is causation as it relates to G-causality. The role of this paper is to highlight connections between the probabilistic aspects of information flow, such as Granger causality and transfer entropy, to a less often discussed geometric picture that may underlie the information flow. To this purpose, here we develop both analysis and data driven concepts to serve in bridging what have otherwise been separate philosophies. Figure 1 illustrates the two nodes that we tackle here: causal inference and geometry. In the diagram, the equations that are most central in serving to bridge the main concepts are highlighted, and the main role of this paper then could be described as building these bridges.

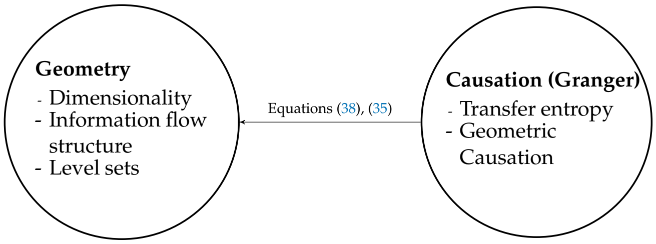


Figure 1. Summary of the paper and relationship of causation and geometry.

When data are derived from a stochastic or deterministic dynamical system, one should also be able to understand the connections between variables in geometric terms. The traditional narrative of information flow is in terms of comparing stochastic processes in probabilistic terms. However, the role of this paper is to offer a unifying description for interpreting geometric formulations of causation together with traditional statistical or information theoretic interpretations. Thus, we will try to provide a bridge between concepts of causality as information flow to the underlying geometry since geometry is perhaps a natural place to describe a dynamical system.

Our work herein comes in two parts. First, we analyze connections between information flow by transfer entropy to geometric quantities that describe the orientation of underlying functions of a corresponding dynamical system. In the course of this analysis, we have needed to develop a new "asymmetric transfer operator" (asymmetric Frobenius–Perron operator) evolving ensemble densities of initial conditions between spaces whose dimensionalities do not match. With this, we proceed to give a new exact formula for transfer entropy, and from there we are able to relate this Kullback–Leibler divergence based measure directly to other more geometrically relevant divergences, specifically total variation divergence and Hellinger divergence, by Pinsker's inequality. This leads to a succinct upper bound of the transfer entropy by quantities related to a more geometric description of the underlying dynamical system. In the second part of this work, we present numerical interpretations of transfer entropy $TE_{y\to x}$ in the setting of a succession of simple dynamical systems, with specifically designed underlying densities, and eventually we include a heart rate versus breathing rate data set. Then, we present a new measure in the spirit of G-causality that is more directly motivated by geometry. This measure, $GeoC_{y\to x}$, is developed in terms of the classical fractal dimension concept of correlation dimension.

In summary, the main theme of this work is to provide connections between probabilistic interpretations and geometric interpretations of causal inference. The main connections and corresponding sections of this paper are summarized as a dichotomy: Geometry and Causation (information flow structure) as described in Figure 1. Our contribution in this paper is as follows:

- In traditional methods, causality is estimated by probabilistic terms. In this study, we present analytical and data driven approach to identify causality by geometric methods, and thus also a unifying perspective.
- We show that a derivative (if it exists) of the underlining function of the time series has a close relationship to the transfer entropy (Section 2.3).
- We provide a new tool called *geoC* to identify the causality by geometric terms (Section 3).
- Correlation dimension can be used as a measurement for dynamics of a dynamical system. We will show that this measurement can be used to identify the causality (Section 3).

Part I: Analysis of Connections between Probabilistic Methods and Geometric Interpretations

2. The Problem Setup

For now, we assume that x, y are real valued scalars, but the multi-variate scenario will be discussed subsequently. We use a shorthand notation, $x := x_n$, $x' := x_{n+1}$ for any particular time n, where the prime ($'$) notation denotes "next iterate". Likewise, let $z = (x, y)$ denote the composite variable, and its future composite state, z'. Consider the simplest of cases, where there are two coupled dynamical systems written as discrete time maps,

$$x' = f_1(x, y), \tag{1}$$

$$y' = f_2(x, y). \tag{2}$$

The definition of transfer entropy [7,8,23], measuring the influence of coupling from variables y onto the future of the variables x, denoted by x' is given by:

$$T_{y\to x} = D_{KL}(p(x'|x)||p(x'|x,y)). \tag{3}$$

This hinges on the contrast between two alternative versions of the possible origins of x' and is premised on deciding one of the following two cases: Either

$$x' = f_1(x), \quad \text{or} \quad x' = f_1(x, y) \tag{4}$$

is descriptive of the actual function f_1. The definition of $T_{y\to x}$ is defined to decide this question by comparing the deviation from a proposed Markov property,

$$p(x'|x) \overset{?}{=} p(x'|x, y). \tag{5}$$

The Kullback–Leibler divergence used here contrasts these two possible explanations of the process generating x'. Since D_{KL} may be written in terms of mutual information, the units are as any entropy, bits per time step. Notice that we have overloaded the notation writing $p(x'|x)$ and $p(x'|x, y)$. Our practice will be to rely on the arguments to distinguish functions as otherwise different (likewise distinguishing cases of $f_1(x)$ versus $f_1(x, y)$.

Consider that the coupling structure between variables may be characterized by the directed graph illustrated in Figure 2.

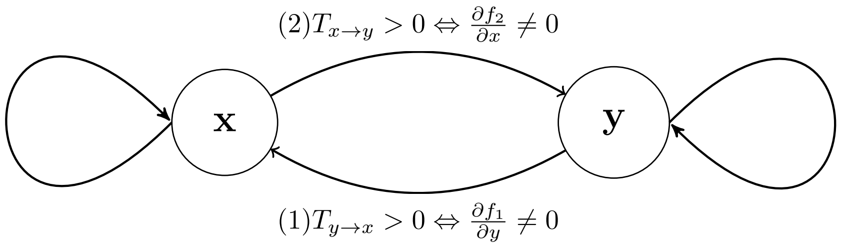

$(2)T_{x\to y} > 0 \Leftrightarrow \frac{\partial f_2}{\partial x} \neq 0$
$\mathbf{x}$
$\mathbf{y}$
$(1)T_{y\to x} > 0 \Leftrightarrow \frac{\partial f_1}{\partial y} \neq 0$

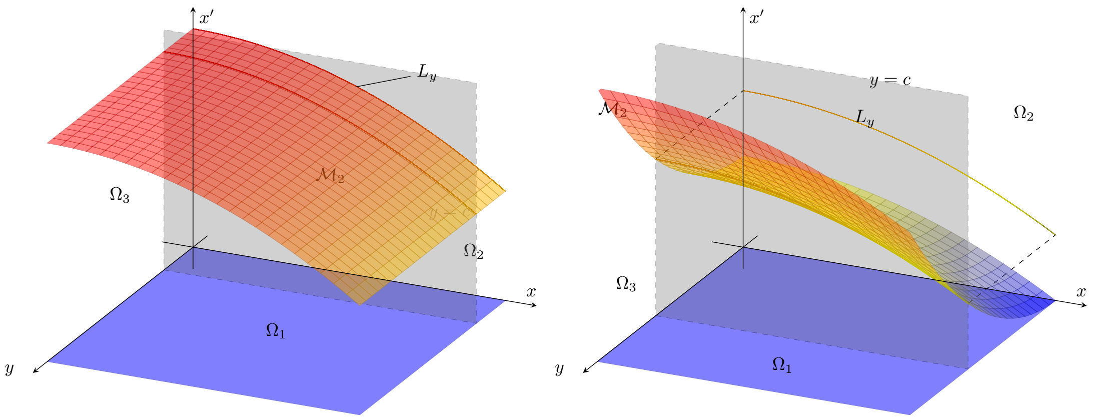
Figure 2. A directed graph presentation of the coupling stucture questions corresponding to Equations (1) and (2).

In one time step, without loss of generality, we may decide on Equation (4), the role of y on x', based on $T_{y\to x} > 0$, exclusively in terms of the details of the argument structure of f_1. This is separate from the reverse question of f_2 as to whether $T_{x\to y} > 0$. In geometric terms, assuming $f_1 \in C^1(\Omega_1)$, it is clear that, unless the partial derivative $\frac{\partial f_1}{\partial y}$ is zero everywhere, then the y argument in $f_1(x,y)$ is relevant. This is not a necessary condition for $T_{y\to x} > 0$, which is a probabilistic statement, and almost everywhere is sufficient.

2.1. In Geometric Terms

Consider a manifold of points $(x,y,x') \in X \times Y \times X'$ as the graph over Ω_1, which we label $\mathcal{M}_2$. In the following, we assume $f_1 \in C^1(\Omega_1), \Omega_1 \subset X \times Y$. Our primary assertion here is that the geometric aspects of the set (x,y,x') projected into (x,x') distinguishes the information flow structure. Refer to Figure 3 for notation. Let the level set for a given fixed y be defined,

$$L_y := \{(x,x') : x' = f(x,y), y = constant\} \in \Omega_2 = X \times X' \tag{6}$$



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

Figure 3. $\Omega_2 = X \times X'$ manifold and L_y level set for (**a**) $x' = f_1(x) = -0.005x^2 + 100$, (**b**) $x' = f_1(x,y) = -0.005x^2 + 0.01y^2 + 50$. The dimension of the projected set of (x,x') depends on the causality as just described. Compare to Figure 4 and Equation (27).

When these level sets are distinct, then the question of the relevance of y to the outcome of x' is clear:

- If $\frac{\partial f_1}{\partial y} = 0$ for all $(x,y) \in \Omega_1$, then $L_y = L_{\tilde{y}}$ for all $y, \tilde{y}$.

Notice that, if the y argument is not relevant as described above, then $x' = f_1(x)$ better describes the associations, but if we nonetheless insist to write $x' = f_1(x, y)$, then $\frac{\partial f_1}{\partial y} = 0$ for all $(x, y) \in \Omega_1$. The converse is interesting to state explicitly,

- If $L_y \neq L_{\tilde{y}}$ for some $y, \tilde{y}$, then $\frac{\partial f_1}{\partial y} \neq 0$ for some $(x, y) \in \Omega_1$, and then $x' = f_1(x)$ is not a sufficient description of what should really be written $x' = f_1(x, y)$. We have assumed $f_1 \in C^1(\Omega_1)$ throughout.

2.2. In Probabilistic Terms

Considering the evolution of x as a stochastic process [8,24], we may write a probability density function in terms of all those variables that may be relevant, $p(x, y, x')$. Contrasting the role of the various input variables requires us to develop a new singular transfer operator between domains that do not necessarily have the same number of variables. Notice that the definition of transfer entropy (Equation (3)) seems to rely on the absolute continuity of the joint probability density $p(x, y, x')$. However, that joint distribution of $p(x, y, f(x, y))$ is generally not absolutely continuous, noticing its support is $\{(x, y, f(x, y)) : (x, y) \in \Omega_x \times \Omega_y \subseteq \mathbb{R}^2\}$, a measure 0 subset of $\mathbb{R}^3$. Therefore, the expression $h(f(X, Y)|X, Y)$ is not well defined as a differential entropy and hence there is a problem with transfer entropy. We expand upon this important detail in the upcoming subsection. To guarantee existence, we interpret these quantities by convolution to smooth the problem. Adding an "artificial noise" with standard deviation parameter ϵ allows definition of the conditional entropy at the singular limit ϵ approaches to zero, and likewise the transfer entropy follows.

The probability density function of the sum of two continuous random variables (U, Z) can be obtained by convolution, $P_{U+Z} = P_U * P_Z$. Random noise (Z with mean $\mathbb{E}(Z) = 0$ and variance $\mathbb{V}(Z) = C\epsilon^2$) added to the original observable variables regularizes, and we are interested in the singular limit, $\epsilon \to 0$. We assume that Z is independent of X, Y. In experimental data from practical problems, we argue that some noise, perhaps even if small, is always present. Additionally, noise is assumed to be uniform or normally distributed in practical applications. Therefore, for simplicity of the discussion, we mostly focused on those two distributions. With this concept, Transfer Entropy can now be calculated by using $h(X'|X, Y)$ and $h(X'|X)$ when

$$X' = f(X, Y) + Z, \tag{7}$$

where now we assume that $X, Y, Z \in \mathbb{R}$ are independent random variables and we assume that $f : \Omega_x \times \Omega_y \to \mathbb{R}$ is a component-wise monotonic (we will consider the monotonically increasing case for consistent explanations, but one can use monotonically decreasing functions in similar manner) continuous function of X, Y and $\Omega_x, \Omega_y \subseteq \mathbb{R}$.

Relative Entropy for a Function of Random Variables

Calculation of transfer entropy depends on the conditional probability. Hence, we will first focus on conditional probability. Since for any particular values x, y the function value $f(x, y)$ is fixed, we conclude that $X'|x, y$ is just a linear function of Z. We see that

$$p_{X'|X,Y}(x'|x, y) = Pr(Z = x' - f(x, y)) = p_Z(x' - f(x, y)), \tag{8}$$

where p_Z is the probability density function of Z.

Note that the random variable $X'|x$ is a function of (Y, Z). To write $U + Z$, let $U = f(x, Y)$. Therefore, convolution of densities of U and Z gives the density function for $p(x'|x)$ (See Section 4.1 for examples). Notice that a given value of the random variable, say $X = \alpha$, is a parameter in U.

Therefore, we will denote $U = f(Y;\alpha)$. We will first focus on the probability density function of U, $p_U(u)$, using the Frobenius–Perron operator,

$$p_U(u) = \sum_{y:u=f(y;\alpha)} \frac{p_Y(f(y;\alpha))}{|f'(f(y;\alpha))|}. \tag{9}$$

In the multivariate setting, the formula is extended similarly interpreting the derivative as the Jacobian matrix, and the absolute value is interpreted as the absolute value of the determinant. Denote $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$, $\mathbf{g}(\mathbf{Y};\alpha) = (g_1, g_2, \ldots, g_n)$ and $U = f(\alpha, \mathbf{Y}) := g_1(\mathbf{Y};\alpha)$; and the vector $\mathbf{V} = (V_1, V_2, \ldots, V_{n-1}) \in \mathbb{R}^{n-1}$ such that $V_i = g_{i+1}(\mathbf{Y}) := Y_{i+1}$ for $i = 1, 2, \ldots, n-1$. Then, the absolute value of the determinate of the Jacobian matrix is given by: $|J_g(\mathbf{y})| = |\frac{\partial g_1(\mathbf{y};\alpha)}{\partial y_1}|$. As an aside, note that J is lower triangular with diagonal entries $d_{ii} = 1$ for $i > 1$. The probability density function of U is given by

$$p_U(u) = \int_S p_{\mathbf{Y}}(g^{-1}(u, \mathbf{v}; \alpha)) \Big| \frac{\partial g_1}{\partial y_1}(g^{-1}(u, \mathbf{v}; \alpha)) \Big|^{-1} d\mathbf{v}, \tag{10}$$

where S is the support set of the random variable $\mathbf{V}$.

Since the random variable $X'|x$ can be written as a sum of U and Z, we find the probability density function by convolution as follows:

$$p_{X'|x}(x'|x) = \int p_U(u) p_Z(x' - u) du. \tag{11}$$

Now, the conditional differential entropy $h(Z|X,Y)$ is in terms of these probability densities. It is useful that translation does not change the differential entropy, $h_\epsilon(f(X,Y) + Z|X,Y) = h(Z|X,Y)$. In addition, Z is independent from X, Y, $h(Z|X,Y) = h(Z)$. Now, we define

$$h(f(X,Y)|X,Y) := \lim_{\epsilon \to 0^+} h_\epsilon(f(X,Y) + Z|X,Y) \tag{12}$$

if this limit exists.

We consider two scenarios: (1) Z is a uniform random variable or (2) Z is a Gaussian random variable. If it is uniform in the interval $[-\epsilon/2, \epsilon/2]$, then the differential entropy is $h(Z) = \ln(\epsilon)$. If specifically, Z is Gaussian with zero mean and ϵ standard deviation, then $h(Z) = \frac{1}{2}\ln(2\pi e \epsilon^2)$. Therefore, $h_\epsilon(f(X,Y) + Z|X,Y) \to -\infty$ as $\epsilon \to 0^+$ in both cases. Therefore, $h(f(X,Y)|X,Y))$ is not finite in this definition (Equation (12)) as well. Thus, instead of calculating $X' = f(X,Y)$, we need to use a noisy version of data $X' = f(X,Y) + Z$. For that case,

$$h(X'|X,Y) = h(Z) = \begin{cases} \ln(\epsilon); & Z \sim U(-\epsilon/2, \epsilon/2) \\ \frac{1}{2}\ln(2\pi e \epsilon^2); & Z \sim \mathcal{N}(0, \epsilon^2) \end{cases}, \tag{13}$$

where $U(-\epsilon/2, \epsilon/2)$ is the uniform distribution in the interval $[-\epsilon/2, \epsilon/2]$, and $\mathcal{N}(0, \epsilon^2)$ is a Gaussian distribution with zero mean and ϵ standard deviation.

Now, we focus on $h(X'|X)$. If X' is just a function of X, then we can similarly show that: if $X' = f(X)$, then

$$h(f(X) + Z|X) = h(Z) = \begin{cases} \ln(\epsilon); & Z \sim U(-\epsilon/2, \epsilon/2) \\ \frac{1}{2}\ln(2\pi e \epsilon^2); & Z \sim \mathcal{N}(0, \epsilon^2). \end{cases} \tag{14}$$

In addition, notice that, if $X' = f(X,Y)$, then $h(X'|X)$ will exist, and most of the cases will be finite. However, when we calculate $T_{y \to x}$, we need to use the noisy version to avoid the issues in calculating $h(X'|X,Y)$. We will now consider the interesting case $X' = f(X,Y) + Z$ and calculate

$h(X'|X)$. We require $p_{X'|X}$ and Equation (11) can be used to calculate this probability. Let us denote $I := \int p_U(u)p_Z(x'-u)du$; then,

$$\begin{aligned} h_\epsilon(X'|X) &= \int\int I\, p_X(x)\ln(I)dx'dx \\ &= \int p_X(x)\int I\ln(I)dx'dx \\ &= \mathbb{E}_X(Q), \end{aligned} \tag{15}$$

where $Q = \int I\ln(I)dx'$. Notice that, if Q does not depend on x, then $h(X'|X) = Q\int p_X dx = Q$ because $\int p_X dx = 1$(since p_x is a probability density function). Therefore, we can calculate $h_\epsilon(X'|X)$ by four steps. First, we calculate the density function for $U = f(x,Y)$ (by using Equation (9) or (10)). Then, we calculate $I = p_{X'|X}$ by using Equation (11). Next, we calculate the value of Q, and finally we calculate the value of $h_\epsilon(X'|X)$.

Thus, the transfer entropy from y to x follows in terms of comparing conditional entropies,

$$T_{y\to x} = h(X'|X) - h(X'|X,Y). \tag{16}$$

This quantity is not well defined when $X' = f(X,Y)$, and therefore we considered the $X' = f(X,Y) + Z$ case. This interpretation of transfer entropy depends on the parameter ϵ, as we define

$$T_{y\to x} := \lim_{\epsilon\to 0^+} T_{y\to x}(\epsilon) = \lim_{\epsilon\to 0^+} h_\epsilon(X'|X) - h_\epsilon(X'|X,Y) \tag{17}$$

if this limit exists.

Note that

$$T_{y\to x} = \begin{cases} \lim_{\epsilon\to 0^+} h(Z) - h(Z) = 0; & X' = f(X) \\ \infty; & X' = f(X,Y) \neq f(X). \end{cases} \tag{18}$$

Thus, we see that a finite quantity is ensured by the noise term. We can easily find an upper bound for the transfer entropy when $X' = f(X,Y) + Z$ is a random variable with finite support (with all the other assumptions mentioned earlier) and suppose $Z \sim U(-\epsilon/2, \epsilon/2)$. First, notice that the uniform distribution maximizes entropy amongst all distributions of continuous random variables with finite support. If f is component-wise monotonically increasing continuous function, then the support of $X'|x$ is $[f(x,y_{min}) - \epsilon/2, f(x,y_{min}) + \epsilon/2]$ for all $x \in \Omega_x$. Here, y_{min} and y_{max} are minimum and maximum values of Y. Then, it follows that

$$h_\epsilon(X'|X) \leq ln(|f(x_{max},y_{max}) - f(x_{max},y_{min}) + \epsilon|), \tag{19}$$

where x_{max} is the maximum x value. We see that an interesting upper bound for transfer entropy follows:

$$T_{y\to x}(\epsilon) \leq \ln\left(\left|\frac{f(x_{max},y_{max}) - f(x_{max},y_{min})}{\epsilon} + 1\right|\right). \tag{20}$$

2.3. Relating Transfer Entropy to a Geometric Bound

Noting that transfer entropy and other variations of the G-causality concept are expressed in terms of conditional probabilities, we recall that

$$\rho(x'|x,y)\rho(x,y) = \rho(x,y,x'). \tag{21}$$

Again, we continue to overload the notation on the functions ρ, the details of the arguments distinguishing to which of these functions we refer.

Now, consider the change of random variable formulas that map between probability density functions by smooth transformations. In the case that $x' = f_1(x)$ (in the special case that f_1 is one-one), then

$$\rho(x') = \frac{\rho(x)}{|\frac{df_1}{dx}(x)|} = \frac{\rho(f_1^{-1}(x'))}{|\frac{df_1}{dx}(f_1^{-1}(x'))|}. \tag{22}$$

In the more general case, not assuming one-one-ness, we get the usual Frobenius–Perron operator,

$$\rho(x') = \sum_{x:x'=f_1(x)} \rho(x,x') = \sum_{x:x'=f_1(x)} \frac{\rho(x)}{|\frac{df_1}{dx}(x)|}, \tag{23}$$

in terms of a summation over all pre-images of x'. Notice also that the middle form is written as a marginalization across x of all those x that lead to x'. This Frobenius–Perron operator, as usual, maps densities of ensembles of initial conditions under the action of the map f_1.

Comparing to the expression

$$\rho(x,x') = \rho(x'|x)\rho(x), \tag{24}$$

we assert the interpretation that

$$\rho(x'|x) := \frac{1}{|\frac{df_1}{dx}(x)|}\delta(x' - f_1(x)), \tag{25}$$

where δ is the Dirac delta function. In the language of Bayesian uncertainty propagation, $p(x'|x)$ describes the likelihood function, if interpreting the future state x' as data, and the past state x as parameters, in a standard Bayes description, $p(\text{data}|\text{parameter}) \times p(\text{parameter})$. As usual for any likelihood function, while it is a probability distribution over the data argument, it may not necessarily be so with respect to the parameter argument.

Now, consider the case where x' is indeed nontrivially a function with respect to not just x, but also with respect to y. Then, we require the following asymmetric space transfer operator, which we name here an asymmetric Frobenius–Perron operator for smooth transformations between spaces of dissimilar dimensionality:

Theorem 1 (Asymmetric Space Transfer Operator)**.** *If $x' = f_1(x,y)$, for $f_1 : \Omega_1 \to \Upsilon$, given bounded open domain $(x,y) \in \Omega_1 \subset \mathbb{R}^{2d}$, and range $x' \in \Upsilon \subset \mathbb{R}^d$, and $f_1 \in C^1(\Omega_1)$, and the Jacobian matrices, $\frac{\partial f_1}{\partial x}(x,y)$, and $\frac{\partial f_1}{\partial y}(x,y)$ are not both rank deficient at the same time, then taking the initial density $\rho(x,y) \in L^1(\Omega_1)$, the following serves as a transfer operator mapping asymmetrically defined densities $P : L^1(\Omega_1) \to L^1(\Upsilon)$*

$$\rho(x') = \sum_{(x,y):x'=f_1(x,y)} \rho(x,y,x') = \sum_{(x,y):x'=f_1(x,y)} \frac{\rho(x,y)}{|\frac{\partial f_1}{\partial x}(x,y)| + |\frac{\partial f_1}{\partial y}(x,y)|}. \tag{26}$$

The proof of this is in Appendix A. Note also that, by similar argumentation, one can formulate the asymmetric Frobenius–Perron type operator between sets of dissimilar dimensionality in an integral form.

Corollary 1 (Asymmetric Transfer Operator, Kernel Integral Form)**.** *Under the same hypothesis as Theorem 1, we may alternatively write the integral kernel form of the expression,*

$$\begin{aligned} P : L^2(\mathbb{R}^2) &\to L^2(\mathbb{R}) && (27)\\ \rho(x,y) &\mapsto \quad \rho'(x') = P[\rho](x,y)] &&\\ &= &&\\ &= \quad \int_{L_{x'}} \rho(x,y,x')dxdy = \int_{L_{x'}} \rho(x'|x,y)\rho(x,y)dxdy &&\\ &= \quad \int_{L_{x'}} \frac{1}{|\frac{\partial f_1}{\partial x}(x,y)| + |\frac{\partial f_1}{\partial y}(x,y)|}\rho(x,y)dxdy. && (28) \end{aligned}$$

This is in terms of a line integration along the level set, $L_{x'}$. See Figure 4:

$$L_{x'} = \{(x,y) \in \Omega_1 : f(x,y) = x' \text{ a chosen constant.}\} \tag{29}$$

In Figure 4, we have shown a typical scenario where a level set is a curve (or it may well be a union of disjoint curves), whereas, in a typical FP-operator between sets of the same dimensionality, generally the integration is between pre-images that are usually either singletons, or unions of such points, $\rho'(x') = \int \delta(s - f(x))\rho(s)ds = \sum_{x:f(x)=x'} \frac{\rho(x)}{|Df(x)|}$.

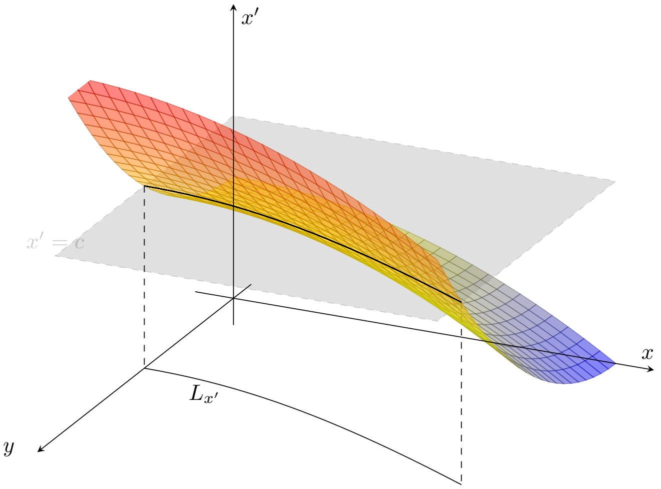


Figure 4. The asymmetric transfer operator, Equation (27), is written in terms of intefration over the level set, $L_{x'}$ of $x' = f_1(x,y)$ associated with a fixed value x', Equation (29).

Contrasting standard and the asymmetric forms of transfer operators as described above, in the next section, we will compute and bound estimates for the transfer entropy. However, it should already be apparent that, if $\frac{\partial f_1}{\partial y} = 0$ in probability with respect to $\rho(x,y)$, then $T_{y \to x} = 0$.

Comparison to other statistical divergences reveals geometric relevance: Information flow is quite naturally defined by the KL-divergence, in that it comes in the units of entropy, e.g., bits per second. However, the well-known Pinsker's inequality [25] allows us to more easily relate the transfer entropy to a quantity that has a geometric relevance using the total variation, even if this is only by an inequality estimate.

Recall that Pinsker's inequality [25] relates random variables with probability distributions p and q over the same support to the total variation and the KL-divergence as follows:

$$0 \le \frac{1}{2}TV(P,Q) \le \sqrt{D_{KL}(P||Q)}, \tag{30}$$

written as probability measures P, Q. The total variation distance between probability measures is a maximal absolute difference of possible events,

$$TV(P,Q) = \sup_{A} |P(A) - Q(A)|, \tag{31}$$

but it is well known to be related to $1/2$ of the L1-distance in the case of a common dominating measure, $p(x)d\mu = dP$, $q(x)d\mu = dQ$. In this work, we only need absolute continuity with respect to Lebesgue measure, $p(x) = dP(x)$, $q(x) = dQ(x)$; then,

$$TV(P,Q) = \frac{1}{2}\int |p(x) - q(x)|dx = \frac{1}{2}\|p - q\|_{L^1}, \tag{32}$$

here with respect to Lebesgue measure. In addition, we write $D_{KL}(P||Q) = \int p(x)\log\frac{p(x)}{q(x)}dx$; therefore,

$$\frac{1}{2}\|p - q\|^2_{L^1} \leq \int p(x)\log\frac{p(x)}{q(x)}dx. \tag{33}$$

Thus, with the Pinsker inequality, we can bound the transfer entropy from below by inserting the definition Equation (3) into the above:

$$0 \leq \frac{1}{2}\|p(x'|x,y) - p(x'|x)\|^2_{L^1} \leq T_{y\rightarrow x}. \tag{34}$$

The assumption that the two distributions correspond to a common dominating measure requires that we interpret $p(x'|x)$ as a distribution averaged across the same $\rho(x,y)$ as $p(x'|x,y)$. (Recall by definition [26] that λ is a common dominating measure of P and Q if $p(x) = dP/d\lambda$ and $q(x) = dQ/d\lambda$ describe corresponding densities). For the sake of simplification, we interpret transfer entropy relative to a uniform initial density, $\rho(x,y)$, for both entropies of Equation (16). With this assumption, we interpret

$$0 \leq \frac{1}{2}\|\frac{1}{|\frac{\partial f_1}{\partial x}(x,y)| + |\frac{\partial f_1}{\partial y}(x,y)|} - \frac{1}{|\frac{df_1}{dx}(x)|}\|^2_{L^1(\Omega_1,\rho(x,y))} \leq T_{y\rightarrow x}. \tag{35}$$

In the special case that there is very little information flow, we would expect that $|\frac{\partial f_1}{\partial y}| < b << 1$, and $b << |\frac{\partial f_1}{\partial x}|$, almost every x, y; then, a power series expansion in small b gives

$$\frac{1}{2}\|\frac{1}{|\frac{\partial f_1}{\partial x}(x,y)| + |\frac{\partial f_1}{\partial y}(x,y)|} - \frac{1}{|\frac{df_1}{dx}(x)|}\|^2_{L^1(\Omega_1,\rho(x,y))} \approx \frac{Vol(\Omega_1)}{2}\frac{<|\frac{\partial f_1}{\partial y}|>^2}{<|\frac{\partial f_1}{\partial x}|>^4}, \tag{36}$$

which serves approximately as the TV-lower bound for transfer entropy where have used the notation $<\cdot>$ to denote an average across the domain. Notice that, therefore, $\delta(p(x'|x,y), p(x'|x)) \downarrow$ as $|\frac{\partial f_1}{\partial y}| \downarrow$. While Pinsker's inequality cannot guarantee that $T_{y\rightarrow x}\downarrow$, since TV is only an upper bound, it is clearly suggestive. In summary, comparing inequality Equation (35) to the approximation (36) suggests that, for $|\frac{\partial f_1}{\partial y}| << b << |\frac{\partial f_1}{\partial x}|$, for $b > 0$, for a.e. x, y, then $T_{y\rightarrow x}\downarrow$ as $b\downarrow$.

Now, we change to a more computational direction of this story of interpreting information flow in geometric terms. With the strong connection described in the following section, we bring to the problem of information flow between geometric concepts to information flow concepts, such as entropy, it is natural to turn to studying the dimensionality of the outcome spaces, as we will now develop.

Part II: Numerics and Examples of Geometric Interpretations

Now, we will explore numerical estimation aspects of transfer entropy for causation inference in relationship to geometry as described theoretically in the previous section, and we will compare this numerical approach to geometric aspects.

3. Geometry of Information Flow

As theory suggests, see the sections above, there is a strong relationship between the information flow (causality as measured by transfer entropy) and the geometry, encoded for example in the estimates leading to Equation (36). The effective dimensionality of the underlying manifold as projected into the outcome space is a key factor to identify the causal inference between chosen variables. Indeed, any question of causality is in fact observer dependent. To this point, suppose x' only depends on x, y and $x' = f(x,y)$, where $f \in C^1(\Omega_1)$. We noticed that (Section 2) $T_{y\to x} = 0 \iff \frac{\partial f}{\partial y} = 0, \ \forall (x,y) \in \Omega_1$. Now, notice that $\frac{\partial f}{\partial y} = 0, \ \forall (x,y) \in \Omega_1 \iff x' = f(x,y) = f(x)$. Therefore, in the case that Ω_1 is two-dimensional, then (x,x') would be a one-dimensional, manifold if and only if $\frac{\partial f}{\partial y} = 0, \ \forall (x,y) \in \Omega_1$. See Figure 3. With these assumptions,

$$T_{y\to x} = 0 \iff (x,x') \text{ data lie on a} 1D \text{ manifold.}$$

Likewise, for more general dimensionality of the initial Ω_1, the story of the information flow between variables is in part a story of how the image manifold is projected. Therefore, our discussion will focus on estimating the dimensionality in order to identify the nature of the underlying manifold. Then, we will focus on identifying causality by estimating the dimension of the manifold, or even more generally of the resulting set if it is not a manifold but perhaps even a fractal. Finally, this naturally leads us to introduce a new geometric measure for characterizing the causation, which we will identify as $Geo_{y\to x}$.

3.1. Relating the Information Flow as Geometric Orientation of Data.

For a given time series $x := x_n \in \mathbb{R}^{d_1}, y := y_n \in \mathbb{R}^{d_2}$, consider the $x' := x_{n+1}$ and *contrast* the dimensionalities of (x,y,x') versus (x,x'), in order to identify that $x' = f(x)$ or $x' = f(x,y)$. Thus, in mimicking the premise of Granger causality, or likewise of Transfer entropy, contrasting these two versions of the explanations of x', in terms of either (x,y) or x, we decide the causal inference, but this time, by using only the geometric interpretation. First, we recall how fractal dimensionality evolves under transformations, [27].

Theorem 2 ([27])**.** *Let A be a bounded Borel subset of $\mathbb{R}^{d_1}$. Consider the function $F : A \to \mathbb{R}^{d_1} \times \mathbb{R}^{d_1}$ such that $F(x) = (x,x')$ for some $x' \in \mathbb{R}^{d_1}$. The correlation dimension $D_2(F(A)) \le d_1$, if and only if there exists a function $f : A \to \mathbb{R}^{d_1}$ such that $x' = f(x)$ with $f \in C^1(A)$.*

The idea of the arguments in the complete proof found in Sauer et. al., [27], are as follows. Let A be bounded Borel subset of $\mathbb{R}^{d_1}$ and $f : A \to \mathbb{R}^{d_1}$ with $f \in C^1(A)$. Then, $D_2(f(A)) = D_2(A)$, where D_2 is the correlation dimension [28]. Note that $D_2(A) \le d_1$. Therefore, $D_2(F(A)) = D_2(A) \le d_1$, with $F : A \to \mathbb{R}^{d_1} \times \mathbb{R}^{d_1}$ if and only if $F(x) = (x, \ f(x))$.

Now, we can describe this dimensional statement in terms of our information flow causality discussion, to develop an alternative measure of inference between variables. Let $(x,x') \in \Omega_2 \subset \mathbb{R}^{2d_1}$ and $(x,y,x') \in \Omega_3 \subset \mathbb{R}^{2d_1+d_2}$. We assert that there is a causal inference from y to x, if $dim(\Omega_2) > d_1$ and $d_1 < dim(\Omega_3) \le d_1 + d_2$, (Theorem 1). In this paper, we focus on time series $x_n \in \mathbb{R}$ which might also depend on time series $y_n \in \mathbb{R}$, and we will consider the geometric causation from y to x, for $(x,y) \in A \times B = \Omega_1 \subset \mathbb{R}^2$. We will denote geometric causation by $GeoC_{y\to x}$ and assume that A, B are Borel subsets of $\mathbb{R}$. Correlation dimension is used to estimate the dimensionality. First, we identify the causality using the dimensionality of on (x,x') and (x,y,x'). Say, for example, that $(x,x') \in \Omega_2 \subset \mathbb{R}^2$ and $(x,y,x') \in \Omega_3 \subset \mathbb{R}^3$; then, clearly we would enumerate a correlation dimension causal inference from y to x, if $dim(\Omega_2) > 1$ and $1 < dim(\Omega_3) \le 2$ (Theorem 1).

3.2. Measure Causality by Correlation Dimension

As we have been discussing, the information flow of a dynamical system can be described geometrically by studying the sets (perhaps they are manifolds) $X \times X'$ and $X \times Y \times X'$. As we noticed in the last section, comparing the dimension of these sets can be interpreted as descriptive of information flow. Whether dimensionality be estimated from data or by a convenient fractal measure such as the correlation dimension ($D_2(.)$), there is an interpretation of information flow when contrasting $X \times X'$ versus $X \times Y \times X'$, in a spirit reminiscent of what is done with transfer entropy. However, these details are geometrically more to the point.

Here, we define $GeoC_{y \to x}$ (geometric information flow) by $GeoC(.|.)$ as conditional correlation dimension.

Definition 1 (Conditional Correlation Dimensional Geometric Information Flow)**.** *Let $\mathcal{M}$ be the manifold of data set $(X_1,\ X_2,\ \ldots,\ X_n,\ X')$ and let Ω_1 be the data set $(X_1,\ X_2,\ \ldots,\ X_n)$. Suppose that the $\mathcal{M},\ \Omega_1$ are bounded Borel sets. The quantity*

$$GeoC(X'|X_1,\ \ldots,\ X_n) := D_2(\mathcal{M}) - D_2(\Omega_1) \tag{37}$$

is defined as "Conditional Correlation Dimensional Geometric Information Flow". Here, $D_2(.)$ is the usual correlation dimension of the given set, [29–31].

Definition 2 (Correlation Dimensional Geometric Information Flow)**.** *Let $x := x_n, y = y_n \in \mathbb{R}$ be two time series. The correlation dimensional geometric information flow from y to x as measured by the correlation dimension and denoted by $GeoC_{y \to x}$ is given by*

$$GeoC_{y \to x} := GeoC(X'|X) - GeoC(X'|X, Y). \tag{38}$$

A key observation is to notice that, if X' is a function of $(X_1,\ X_2,\ \ldots,\ X_n)$, then $D_2(\mathcal{M}) = D_2(\Omega_1)$; otherwise, $D_2(\mathcal{M}) > D_2(\Omega_1)$ (Theorem 1). If X is not influenced by y, then $GeoC(X'|X) = 0$, $GeoC(X'|X, Y) = 0$ and therefore $GeoC_{y \to x} = 0$. In addition, notice that $GeoC_{y \to x} \leq D_2(X)$, where $X = \{x_n | n = 1, 2, \ldots\}$. For example, if $x_n \in \mathbb{R}$, then $GeoC_{y \to x} \leq 1$. Since we assume that influence of any time series $z_n \neq x_n, y_n$ to x_n is relatively small, we can conclude that $GeoC_{y \to x} \geq 0$, and, if $x' = f(x, y)$, then $GeoC(X'|X, Y) = 0$. Additionally, the dimension ($GeoC(X'|X)$) in the (X, X') data scores how much additional (other than X) information is needed to describe the X' variable. Similarly, the dimension $GeoC(X'|X, Y)$ in the (X, Y, X') data describes how much additional (other than X, Y) information is needed to define X'. However, when the number of data points $N \to \infty$, the value $GeoC_{y \to x}$ is not negative (equal to the dimension of X data). Thus, theoretically, $GeoC$ identifies a causality in the geometric sense we have been describing.

4. Results and Discussion

Now, we present specific examples to contrast the transfer entropy with our proposed geometric measure to further highlight the role of geometry in such questions. Table 1 provides a summary of our numerical results. We use synthetic examples with known underlining dynamics to understand the accuracy of our model. Calculating transfer entropy has theoretical and numerical issues for those chosen examples while our geometric approach accurately identifies the causation. We use the correlation dimension of the data because data might be fractals. Using a Hénon map example, we demonstrate that fractal data will not affect our calculations. Furthermore, we use a real-world application that has a positive transfer entropy to explain our data-driven geometric method. Details of these examples can be found in the following subsections.

Table 1. Summary of the results. Here, we experiment our new approach by synthetics and real world application data.

Data	**Transfer Entropy (Section 4.1)**	**Geometric Approach**
Synthetic: f(x,y)=$aX + bY + C$, $a, b, c \in \mathbb{R}$	Theoretical issues can be noticed. Numerical estimation have boundedness issues when $b << 1$.	Successfully identify the causation for all the cases (100%).
Synthetic: f(x,y)=$ag_1(X) + bg_2(Y) + C, a, b, c \in \mathbb{R}$	Theoretical issues can be noticed. Numerical estimation have boundedness issues when $b << 1$.	Successfully identify the causation for all the cases (100%).
Hénon map: use data set invariant under the map.	special case of $aX^2 + bY + C$ with $a = -1.4$, $b = c = 1$. Estimated transfer entropy is positive.	Successfully identify the causation.
Application: heart rate vs. breathing rate	Positive transfer entropy.	Identify positive causation. It also provides more details about the data.

4.1. Transfer Entropy

In this section, we will focus on analytical results and numerical estimators for conditional entropy and transfer entropy for specific examples (see Figures 5 and 6). As we discussed in previous sections starting with Section 2.2, computing the transfer entropy for $X' = f(X, Y)$ has technical difficulties due to the singularity of the quantity $h(X'|X, Y)$. First, we will consider the calculation of $h(X'|X)$ for $X' = f(X, Y)$, and then we will discuss the calculation for noisy data. In the following examples, we assumed that X, Y are random variables such that $X, Y \overset{iid}{\sim} U([1, 2])$. A summary of the calculations for a few examples are listed in Table 2.

Table 2. Conditional entropy $h(X'|X)$ for $X' = f(X, Y)$, for specific parametric examples listed, under the assumption that $X, Y \overset{iid}{\sim} U([1, 2])$.

$\boldsymbol{f(X,Y)}$	$\boldsymbol{h(X'\|X)}$
$g(X) + bY$	$\ln(b)$
$g(X) + bY^2$	$\ln(8b) - 5/2$
$g(X) + b\ \ln(Y)$	$\ln\left(\frac{b\ e}{4}\right)$

We will discuss the transfer entropy with noisy data because making $h(X'|X, Y)$ well defined requires absolute continuity of the probability density function $p(x, y, x')$. Consider, for example, the problem form $X' = g(X) + bY + C$, where X, Y are uniformly distributed independent random variables over the interval $[1, 2]$ (the same analysis can be extend to any finite interval) with b being a constant, and g a function of random variable X. We will also consider C to be a random variable, which is distributed uniformly on $[-\epsilon/2, \epsilon/2]$. Note that it follows that $h(X'|X, Y) = \ln \epsilon$. To calculate the $h(X'|X)$, we need to find the conditional probability $p(X'|x)$ and observe that $X'|x = U + C$, where $U = g(x) + bY$. Therefore,

$$p_U(u) = \begin{cases} \frac{1}{b} & ; g_1(x) + b \leq X' \leq g_1(x) + 2b \\ 0 & ; otherwise. \end{cases} \tag{39}$$

and

$$p_{X'|X}(X'|x) = \begin{cases} \frac{x'+\epsilon/2-g(x)}{b\epsilon} & ; g(x)-\epsilon/2 \le X' \le g(x)+\epsilon/2 \\ \frac{1}{b} & ; g(x)+\epsilon/2 \le X' \le b+g(x)-\epsilon/2 \\ \frac{-x'+\epsilon/2+g(x)+b}{b\epsilon} & ; b+g(x)-\epsilon/2 \le X' \le b+g(x)+\epsilon/2 \\ 0 & ; otherwise \end{cases} . \tag{40}$$

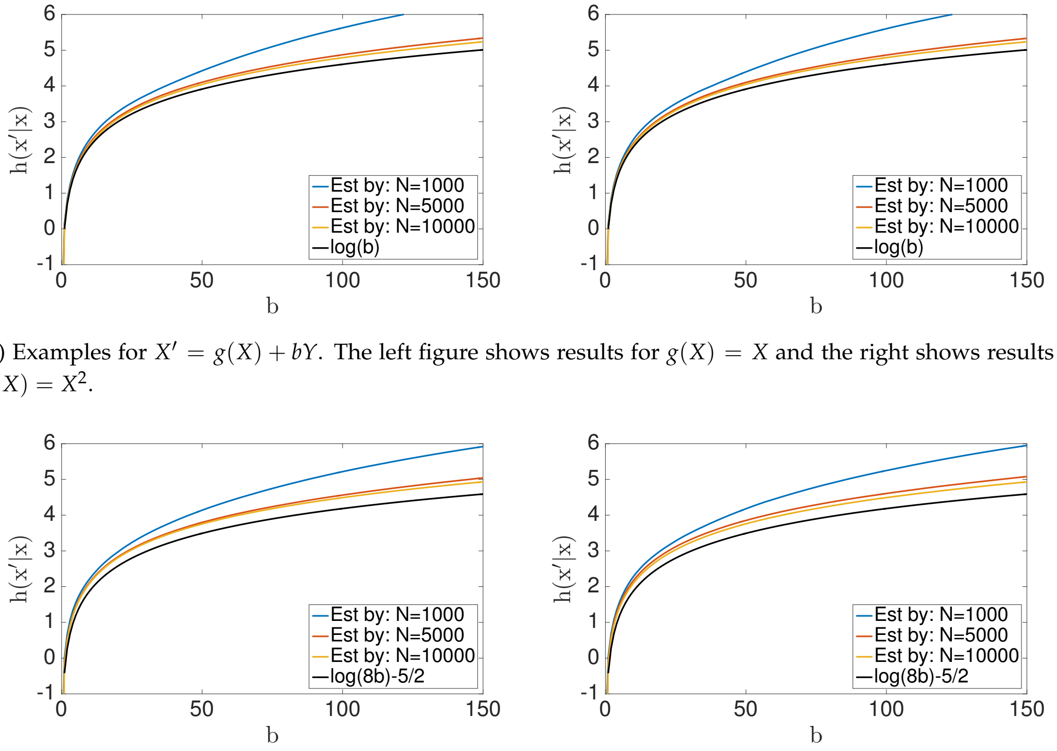
By the definition of transfer entropy, we can show that

$$h(X'|X) = \ln b + \frac{\epsilon}{2b} \tag{41}$$

and hence transfer entropy of this data are given by

$$T_{y\to x}(\epsilon; b) = \begin{cases} \ln\frac{b}{\epsilon} + \frac{\epsilon}{2b}; & b \ne 0 \\ 0; & b = 0. \end{cases} \tag{42}$$

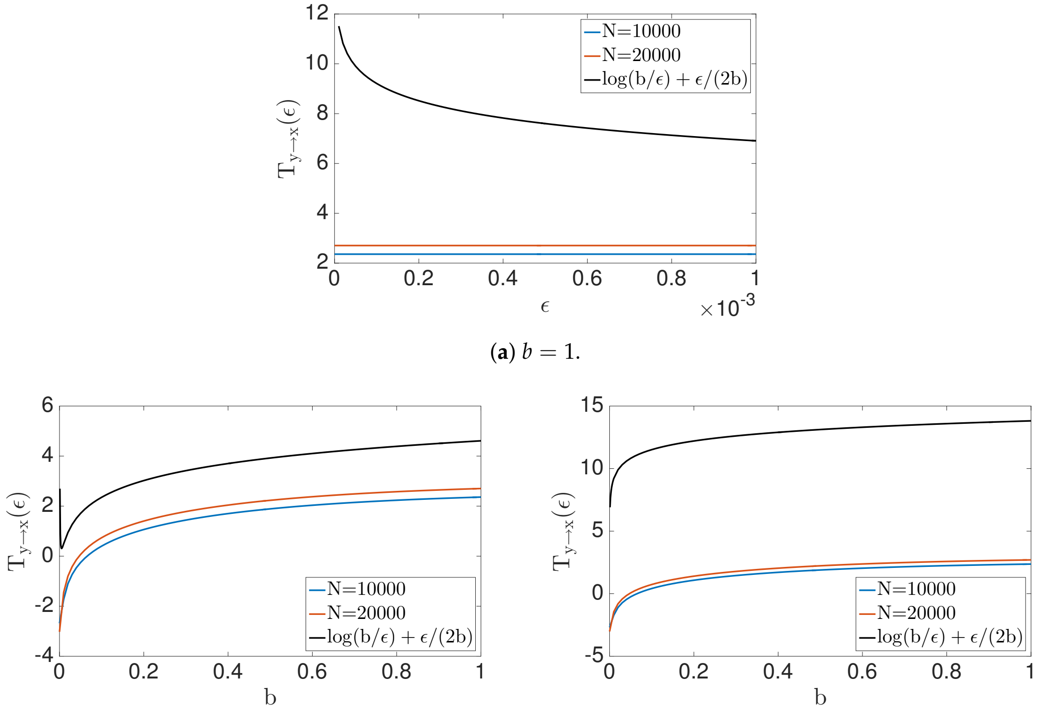
Therefore, when $b = 0$, the transfer entropy $T_{y\to x} = \ln\epsilon - \ln\epsilon = 0$. In addition, notice that $T_{y\to x}(\epsilon; b) \to \infty$ as $\epsilon \to 0$. Therefore, convergence of the numerical estimates is slow when $\epsilon > 0$ is small (see Figure 6).



(**a**) Examples for $X' = g(X) + bY$. The left figure shows results for $g(X) = X$ and the right shows results for $g(X) = X^2$.

(**b**) Examples for $X' = g(X) + bY^2$. The left figure shows results for $g(X) = X$ and the right shows results for $g(X) = e^x$.

Figure 5. Conditional entropy $h(X'|X)$. Note that these numerical estimates for the conditional entropy by the KSG method [32], converge (as $N \to \infty$) to the analytic solutions (see Table 2).



(**a**) $b = 1$.

(**b**) $\epsilon = 0.01$

(**c**) $\epsilon = 10^{-6}$

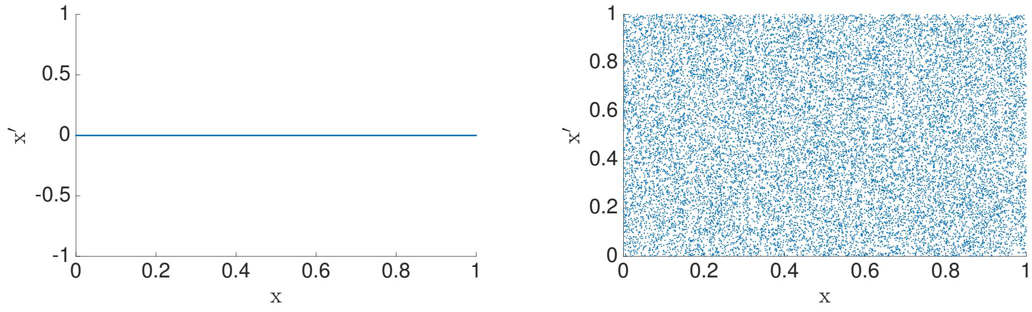
Figure 6. Numerical results and analytical results for transfer entropy $T_{y\to x}(\epsilon; b)$ to the problem $X' = X + bY + \epsilon$. Transfer entropy vs. ϵ shows in (**a**) for fixed b value. (**b**) and (**c**) show the behavior of the transfer entropy for b values with fixed ϵ values. Notice that convergence of numerical solution is slow when epsilon is small.

4.2. Geometric Information Flow

Now, we focus on quantifying the geometric information flow by comparing dimensionalities of the outcomes' spaces. We will contrast this to the transfer entropy computations for a few examples of the form $X' = g(X) + bY + C$.

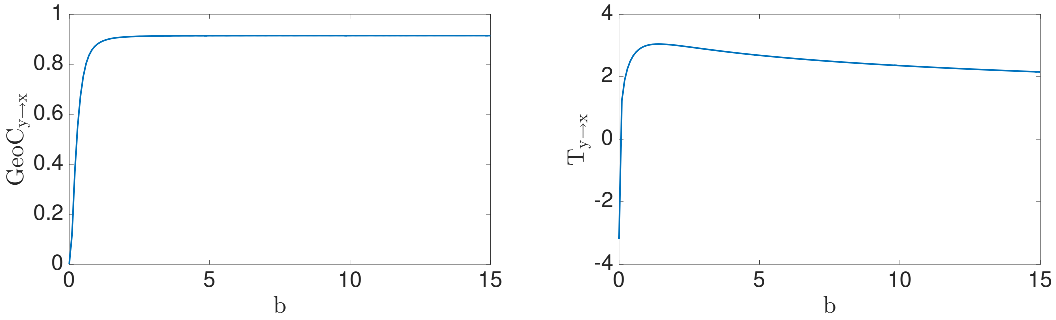
To illustrate the idea of geometric information flow, let us first consider a simple example, $x' = ax + by + c$. If $b = 0$, we have $x' = f(x)$ and, when $b \neq 0$, we have the $x' = f(x, y)$ case. Therefore, dimensionality of the data set (x', x) will change with parameter b (see Figure 7). When the number of data points $N \to \infty$ and $b \neq 0$, then $GeoC_{y\to x} \to 1$. Generally, this measure of causality depends on the value of b, but also the initial density of initial conditions.

In this example, we contrast theoretical solutions with the numerically estimated solutions (Figure 8). Theoretically, we expect $T_{y\to x} = \begin{cases} 0 & ; b = 0 \\ \infty & ; b \neq 0 \end{cases}$ as $N \to \infty$. In addition, the transfer entropy for noisy data can be calculated by Equation (42).



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

Figure 7. Manifold of the data (x', x) with $x' = by$ and y is uniformly distributed in the interval $[0, 1]$. Notice that, when (a) $b = 0$, we have a 1D manifold, (b) $b \neq 0$ we have 2D manifold, in the (x', x) plane.



(**a**) $GeoC_{y\to x}$ (**b**) $T_{y\to x}$(Numerical results)

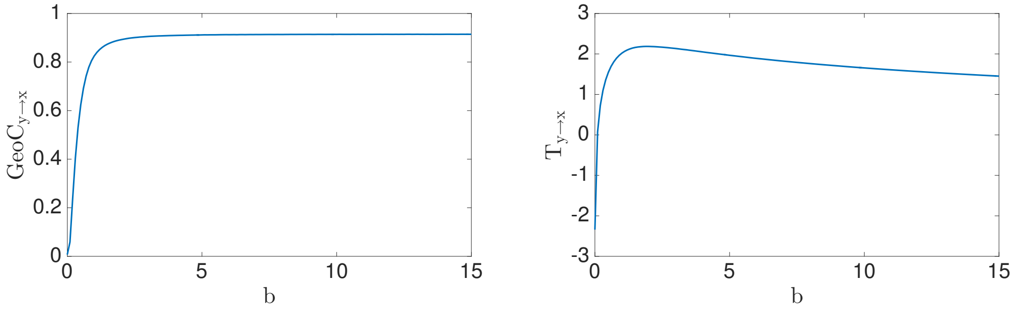
Figure 8. Geometric information flow vs. Transfer entropy for $X' = bY$ data.

4.3. Synthetic Data: $X' = aX + bY$ with $a \neq 0$

The role of the initial density of points in the domain plays an important role in how the specific information flow values are computed depending on the measure used. To illustrate this point, consider the example of a unit square, $[0, 1]^2$, that is uniformly sampled, and mapped by

$$X' = aX + bY, \text{ with } a \neq 0. \tag{43}$$

This fits our basic premise that (x, y, x') data embeds in a 2D manifold, by ansatz of Equations (1) and (43), assuming for this example that each of x, y and x' are scalar. As the number of data point grows, $N \to \infty$, we can see that $GeoC_{y\to x} = \begin{cases} 0 & ; b = 0 \\ 1 & ; b \neq 0 \end{cases}$ because (X, X') data are on 2D manifold iff $b \neq 0$ (numerical estimation can be seen in Figure 9b). On the other hand, the conditional entropy $h(X'|X, Y)$ is not defined, becoming unbounded when defined by noisy data. Thus, it follows that transfer entropy shares this same property. In other words, boundedness of transfer entropy depends highly on the $X'|X, Y$ conditional data structure, while, instead, our geometric information flow measure highly depends on $X'|X$ conditional data structure. Figure 9c demonstrates this observation with estimated transfer entropy and analytically computed values for noisy data. The slow convergence can be observed, Equation (42), Figure 6.



(**a**) $GeoC_{y\to x}$ (**b**) $T_{y\to x}$

Figure 9. (**a**) shows the geometric information flow and (**b**) represents the Transfer entropy for $x' = x + by$ data. The figures show the changes with parameter b. We can notice that the transfer entropy has similar behavior to the geometric information flow of the data.

4.4. Synthetic Data: Nonlinear Cases

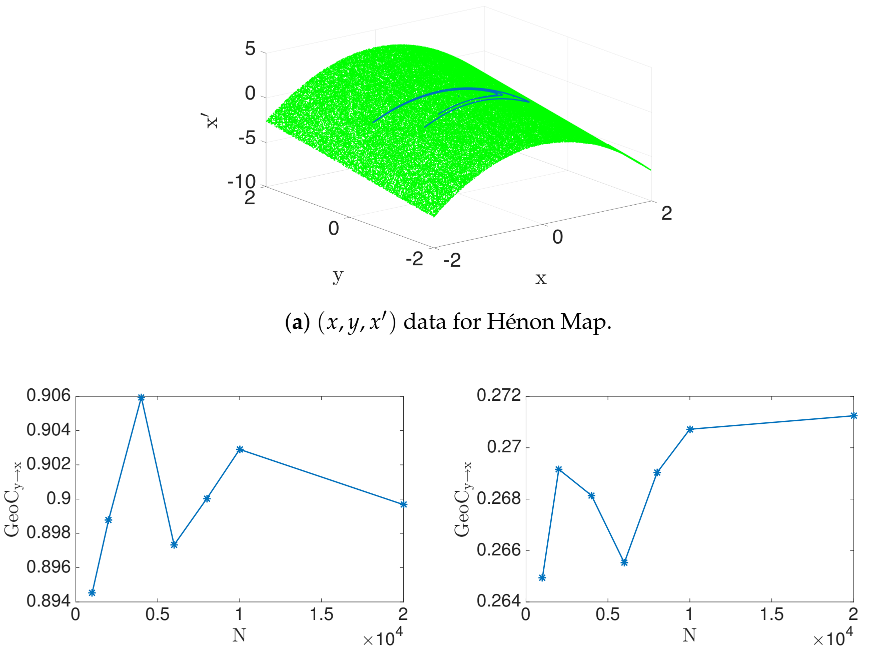
Now, consider the Hénon map,

$$\begin{aligned} x' &= 1 - 1.4x^2 + y \\ y' &= x \end{aligned} \tag{44}$$

as a special case of a general quadratic relationship, $x' = ax + by^2 + c$, for discussing how x' may depend on $(x, y) \in \Omega_1$. Again, we do not worry here if y' may or may not depend on x and or y when deciding dependencies for x'. We will discuss two cases, depending on how the $(x, y) \in \Omega_1$ data are distributed. For the first case, assume (x, y) is uniformly distributed in the square, $[-1.5, 1.5]^2$. The second and dynamically more realistic case will assume that (x, y) lies on the invariant set (the strange attractor) of the Hénon map. The geometric information flow is shown for both cases in Figure 10. We numerically estimate the transfer entropy for both cases, which gives $T_{y\to x} = 2.4116$ and 0.7942, respectively. (However, recall that the first case for transfer entropy might not be finite analytically, and there is slow numerical estimation—see Table 3).

Table 3. Hénon Map Results. Contrasting geometric information flow versus transfer entropy in two different cases, 1st relative to uniform distribution of initial conditions (reset each time) and 2nd relative to the natural invariant measure (more realistic).

Domain	***GeoC***	$T_{y\to x}$
$[-1.5, 1.5]^2$	0.90	2.4116
Invariant Set	0.2712	0.7942



(**a**) (x, y, x') data for Hénon Map.

(**b**) $(x, y) \sim U([-1.5, 1.5]^2)$ (**c**) (x, y) is in invariant set of Hénon map

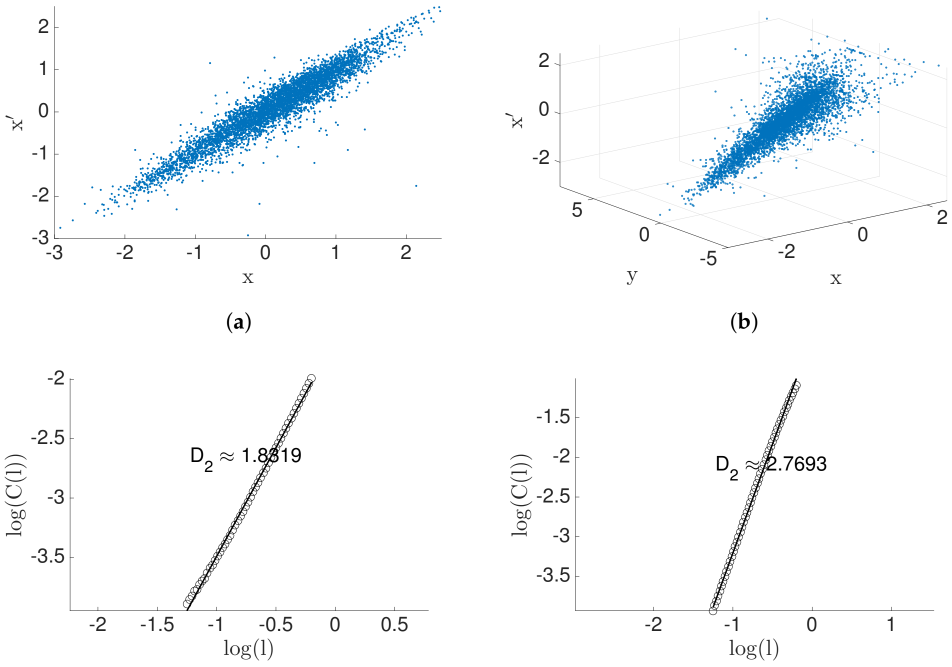
Figure 10. Consider the Hénon map, Equation (44), within the domain $[-1.5, 1.5]^2$ and the invariant set of Hénon map. (**a**) the uniform distribution case (green) as well as the natural invariant measure of the attractor (blue) are shown regarding the (x, y, x') data for both cases; (**b**) when $(x, y) \in [-1.5, 1.5]^2$, notice that $GeoC_{y\to x} = 0.9$, and (**c**) if (x, y) is in an invariant set of Hénon map, then $GeoC_{y\to x} = 0.2712$.

4.5. Application Data

Now, moving beyond bench-marking with synthetic data, we will contrast the two measures of information flow in a real world experimental data set. Consider heart rate (x_n) vs. breathing rate (y_n) data (Figure 11) as published in [33,34], consisting of 5000 samples. Correlation dimension of the data X is $D_2(X) = 1.00$, and $D_2(X, X') = 1.8319 > D_2(X)$. Therefore, $X' = X_{n+1}$ depends not only on x, but also on an extra variable (Theorem 2). In addition, correlation dimension of the data (X, Y) and (X, Y, X') is computed $D_2(X, Y) = 1.9801$ and $D_2(X, Y, X') = 2.7693 > D_2(X, Y)$, respectively. We conclude that X' depends on extra variable(s) other that (x, y) (Theorem 2) and the correlation dimension geometric information flow, $GeoC_{y\to x} = 0.0427$, is computed by Equations (38) and (37). Therefore, this suggests the conclusion that there is a causal inference from breathing rate to heart rate. Since breathing rate and heart rate share the same units, the quantity measured by geometric information flow can be described without normalizing. Transfer entropy as estimated by the KSG method [32] with parameter $k = 30$ is $T_{y\to x} = 0.0485$, interestingly relatively close to the $GeoC$ value. In summary, both measures for causality $(GeoC,\ T)$ are either zero or positive together. It follows that there exists a causal inference (see Table 4).

Table 4. Heart rate vs. breathing rate data—contrasting geometric information flow versus transfer entropy in breath rate to heart rate.

$GeoC_{y\to x}$	$T_{y\to x}$
0.0427	0.0485



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

(**c**) (**d**)

Figure 11. Result for heart rate(x_n) (**a**,**c**) vs. breathing rate(y_n) data (**b**,**d**). The top row is the scatter plot of the data, and the second row represents the dimension of the data.

5. Conclusions

We have developed here a geometric interpretation of information flow as a causal inference as usually measured by a positive transfer entropy, $T_{y \to x}$. Our interpretation relates the dimensionality of an underlying manifold as projected into the outcome space and summarizes the information flow. Furthermore, the analysis behind our interpretation involves standard Pinsker's inequality that estimates entropy in terms of total variation, and, through this method, we can interpret the production of information flow in terms of details of the derivatives describing relative orientation of the manifolds describing inputs and outputs (under certain simple assumptions).

A geometric description of causality allows for new and efficient computational methods for causality inference. Furthermore, this geometric perspective provides a different view of the problem and facilitates the richer understanding that complements the probabilistic descriptions. Causal inference is weaved strongly throughout many fields and the use of transfer entropy has been a popular black box tool for this endeavor. Our method can be used to reveal more details of the underling geometry of the data-set and provide a clear view of the causal inference. In addition, one can use the hybrid method of this geometric aspect and existing other methods in their applications.

We provided a theoretical explanation (part I: Mathematical proof of the geometric view of the problem) and numerical evidence (part 2: A data-driven approach for mathematical framework) of a geometric view for the causal inference. Our experiments are based on synthetic (toy problems) and practical data. In the case of synthetic data, the underlining dynamics of the data and the actual solution to the problem are known. For each of these toy problems, we consider a lot of cases by setting a few parameters. Our newly designed geometric approach can successfully capture these cases. One major problem may be if data describes a chaotic attractor. We prove theoretically (Theorem 2) and experimentally (by Hénon map example: in this toy problem, we also know actual causality) that correlation dimension serves to overcome this issue. Furthermore, we present a practical example based on heart rate vs. breathing rate variability, which was already shown to have positive transfer entropy, and here we relate this to show positive geometric causality.

Furthermore, we have pointed out that transfer entropy has analytic convergence issues when future data (X') are exactly a function of current input data (X, Y) versus more generally (X, Y, X'). Therefore, referring to how the geometry of the data can be used to identify the causation of the time series data, we develop a new causality measurement based on a fractal measurement comparing inputs and outputs. Specifically, the correlation dimension is a useful and efficient way to define what we call correlation dimensional geometric information flow, $GeoC_{y\to x}$. The $GeoC_{y\to x}$ offers a strongly geometric interpretable result as a global picture of the information flow. We demonstrate the natural benefits of $GeoC_{y\to x}$ versus $T_{y\to x}$, in several synthetic examples where we can specifically control the geometric details, and then with a physiological example using heart and breathing data.

Author Contributions: Conceptualization, Sudam Surasinghe and E.M.B.; Data curation, Sudam Surasinghe and E.M.B.; Formal analysis, S.S. and E.M.B.; Funding acquisition, E.M.B.; Methodology, S.S. and E.M.B.; Project administration, E.M.B.; Resources, E.M.B.; Software, S.S. and E.M.B.; Supervision, E.M.B.; Validation, S.S. and E.M.B.; Visualization, S.S. and E.M.B.; Writing—original draft, S.S. and E.M.B.; Writing—review and editing, S.S. and E.M.B.

Funding: Erik M. Bollt gratefully acknowledges funding from the Army Research Office W911NF16-1-0081 (Samuel Stanton) as well as from DARPA.

Acknowledgments: We would like to thank Ioannis Kevrekidis for his generous time, feedback, and interest regarding this project.

Appendix A. On the Asymmetric Spaces Transfer Operators

In this section we prove Theorem 1 concerning a transfer operator for smooth transformations between sets of perhaps dissimilar dimensionality. In general, the marginal probability density can be found by integrating (or summation in the case of a discrete random variable) to marginalize the joint probability densities. When $x' = f(x, y)$, the joint density (x, y, x') is non-zero only at points on $x' = f(x, y)$. Therefore, $\rho(x') = \sum_{(x,y):x'=f(x,y)} \rho(x, y, x')$ and notice that $\rho(x, y, x') = \rho(x'|x, y)\rho(x, y)$ (By Bayes theorem). Hence, $\rho(x') = \sum_{(x,y):x'=f(x,y)} \rho(x'|x, y)\rho(x, y)$ and we only need to show the following claims. We will discuss this by two cases. First, we consider $x' = f(x)$ and then we consider more general case $x = f(x, y)$. In higher dimensions we can consider similar scenarios of input and output variables, and correspondingly the trapezoidal bounding regions would need to be specified in which we can analytically control the variables.

Proposition A1 (Claim). *Let $X \in \mathbb{R}$ be a random variable with probability density function $\rho(x)$. Suppose $\rho(x), \rho(.|x)$ are Radon–Nikodym derivatives (of induced measure with respect to some base measure μ) which is bounded above and bounded away from zero. In addition, let $x' = f(x)$ for some function $f \in C^1(\mathbb{R})$. Then,*

$$\rho(x'|X = x_0) = \lim_{\epsilon\to 0} d_\epsilon(x' - f(x_0))$$

where $d_\epsilon(x' - f(x_0)) = \begin{cases} \frac{1}{2\epsilon|f'(x_0)|} & ; |x' - f(x_0)| < \epsilon|f'(x_0)| \\ 0 & ;\ \textit{otherwise} \end{cases}$.

Proof. Let $1 >> \epsilon > 0$ and $x \in I_\epsilon = (x_0 - \epsilon, x_0 + \epsilon)$. Since ρ is a Radon–Nikodym derivative with bounded above and bounded away from zero, $\rho(I_\epsilon) = \int_{I_\epsilon} \frac{d\rho}{d\mu} d\mu \geq \frac{m}{2\epsilon}$ where m is the infimum of the Radon–Nikodym derivative. Similarly $\rho(I_\epsilon) \leq \frac{M}{2\epsilon}$ where M is the supremum of the Radon–Nikodym derivative. In addition, $|x' - f(x_0)| \approx |f'(x_0)||x - x_0|$ for $x \in I_\epsilon$. Therefore, $x' \in (f(x_0) - \epsilon|f'(x_0)|, f(x_0) + \epsilon|f'(x_0)|) = I'_\epsilon$ when $x \in I_\epsilon$. Hence, $\rho(x'|x \in I_\epsilon) = \rho(x' \in I'_\epsilon)$ and $\frac{m}{2\epsilon|f'(x_0)|} \leq \rho(x'|x \in I_\epsilon) \leq \frac{M}{2\epsilon|f'(x_0)|}$. Therefore, $\rho(x'|X = x_0) = \lim_{\epsilon\to 0} d_\epsilon(x' - f(x_0))$ □

Proposition A2 (Claim). *2 Let $X, Y \in \mathbb{R}$ be random variables with joint probability density function $\rho(x, y)$. Suppose $\rho(x, y)$ and $\rho(.|x, y)$ are Radon–Nikodym derivatives (of induced measure with respect to some base* *measure μ) which is bounded above and bounded away from zero. In addition, let $x' = f(x,y) \in \mathbb{R}$ for some function $f \in C^1(\mathbb{R})$. Then,*

$$\rho(x'|X = x_0, Y = y_0) = \lim_{\epsilon \to 0} d_\epsilon(x' - f(x_0, y_0))$$

where $d_\epsilon(x' - f(x_0,y_0)) = \begin{cases} \frac{1}{2\epsilon(|f_x(x_0,y_0)|+|f_y(x_0,y_0)|)} & ; |x' - f(x_0,y_0)| < \epsilon(|f_x(x_0,y_0)| + |f_y(x_0,y_0)|) \\ 0 & ;\ \textit{otherwise} \end{cases}.$

Proof. Let $1 >> \epsilon > 0$ and $A_\epsilon = \{(x,y)|x \in (x_0 - \epsilon, x_0 + \epsilon), y \in (y_0 - \epsilon, y_0 + \epsilon)$. Since ρ is a Radon–Nikodym derivative with bounded above and bounded away from zero, $\rho(A_\epsilon) = \int_{A_\epsilon} \frac{d\rho}{d\mu} d\mu \geq \frac{m}{4\epsilon^2}$ where m is the infimum of the Radon–Nikodym derivative. Similarly, $\rho(A_\epsilon) \leq \frac{M}{4\epsilon^2}$ where M is the supremum of the Radon–Nikodym derivative. In addition, $|x' - f(x_0,y_0)| \approx |f_x(x_0,y_0)||x - x_0| + |f_y(x_0,y_0)||y - y_0|$ for $(x,y) \in A_\epsilon$. Therefore, $x' \in (f(x_0,y_0) - \epsilon(|f_x(x_0,y_0)| + |f_y(x_0,y_0)|), f(x_0,y_0) + \epsilon(|f_x(x_0,y_0)| + |f_y(x_0,y_0)|)) = I'_\epsilon$ when $(x,y) \in A_\epsilon$. Hence, $\rho(x'|(x,y) \in A_\epsilon) = \rho(x' \in I'_\epsilon)$ and $\frac{m}{2\epsilon(|f_x(x_0,y_0)|+|f_y(x_0,y_0)|)} \leq \rho(x'|x \in I_\epsilon) \leq \frac{M}{2\epsilon(|f_x(x_0,y_0)|+|f_y(x_0,y_0)|)}$. Therefore, $\rho(x'|X = x_0, Y = y_0) = \lim_{\epsilon \to 0} d_\epsilon(x' - f(x_0,y_0))$.

□

If f only depends on x, then the partial derivative of f with respect to y is equal to zero and which leads to the same result as clam 1.

References

1. Williams, C.J.F. "Aristotle's Physics, Books I and II", Translated with Introduction and Notes by W. Charlton. *Mind* **1973**, *82*, 617. [CrossRef]
2. Falcon, A. Aristotle on Causality. In *The Stanford Encyclopedia of Philosophy*, Spring 2019 ed.; Zalta, E.N., Ed.; Metaphysics Research Lab, Stanford University: Stanford, CA, USA, 2019.
3. Russell, B. I.—On the Notion of Cause. *Proc. Aristot. Soc.* **1913**, *13*, 1–26. [CrossRef]
4. Bollt, E.M. Open or closed? Information flow decided by transfer operators and forecastability quality metric. *Chaos Interdiscip. J. of Nonlinear Sci.* **2018**, *28*, 075309. doi:10.1063/1.5031109. [CrossRef] [PubMed]
5. Hendry, D.F. The Nobel Memorial Prize for Clive W. J. Granger. *Scand. J. Econ.* **2004**, *106*, 187–213. [CrossRef]
6. Wiener, N. The theory of prediction. In *Mathematics for the Engineer*; McGraw-Hill: New York, NY, USA, 1956.
7. Schreiber, T. Measuring Information Transfer. *Phys. Rev. Lett.* **2000**, *85*, 461–464. doi:10.1103/PhysRevLett.85.461. [CrossRef] [PubMed]
8. Bollt, E.; Santitissadeekorn, N. *Applied and Computational Measurable Dynamics*; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2013. doi:10.1137/1.9781611972641. [CrossRef]
9. Barnett, L.; Barrett, A.B.; Seth, A.K. Granger Causality and Transfer Entropy Are Equivalent for Gaussian Variables. *Phys. Rev. Lett.* **2009**, *103*, 238701. doi:10.1103/PhysRevLett.103.238701. [CrossRef]
10. Sugihara, G.; May, R.; Ye, H.; Hsieh, C.h.; Deyle, E.; Fogarty, M.; Munch, S. Detecting Causality in Complex Ecosystems. *Science* **2012**, *338*, 496–500. doi:10.1126/science.1227079. [CrossRef] [PubMed]
11. Sun, J.; Bollt, E.M. Causation entropy identifies indirect influences, dominance of neighbors and anticipatory couplings. *Phys. D Nonlinear Phenom.* **2014**, *267*, 49–57. [CrossRef]
12. Sun, J.; Taylor, D.; Bollt, E. Causal Network Inference by Optimal Causation Entropy. *SIAM J. Appl. Dyn. Syst.* **2015**, *14*, 73–106. doi:10.1137/140956166. [CrossRef]
13. Bollt, E.M.; Sun, J.; Runge, J. Introduction to Focus Issue: Causation inference and information flow in dynamical systems: Theory and applications. *Chaos Interdiscip. J. Nonlinear Sci.* **2018**, *28*, 075201. [CrossRef]
14. Runge, J.; Bathiany, S.; Bollt, E.; Camps-Valls, G.; Coumou, D.; Deyle, E.; Glymour, C.; Kretschmer, M.; Mahecha, M.D.; Muñoz-Marí, J.; et al. Inferring causation from time series in Earth system sciences. *Nat. Commun.* **2019**, *10*, 1–13. [CrossRef] [PubMed]

15. Lord, W.M.; Sun, J.; Ouellette, N.T.; Bollt, E.M. Inference of causal information flow in collective animal behavior. *IEEE Trans. Mol. Biol. Multi-Scale Commun.* **2016**, *2*, 107–116. [CrossRef]
16. Kim, P.; Rogers, J.; Sun, J.; Bollt, E. Causation entropy identifies sparsity structure for parameter estimation of dynamic systems. *J. Comput. Nonlinear Dyn.* **2017**, *12*, 011008. [CrossRef]
17. AlMomani, A.A.R.; Sun, J.; Bollt, E. How Entropic Regression Beats the Outliers Problem in Nonlinear System Identification. *arXiv* **2019**, arXiv:1905.08061.
18. Sudu Ambegedara, A.; Sun, J.; Janoyan, K.; Bollt, E. Information-theoretical noninvasive damage detection in bridge structures. *Chaos Interdiscip. J. Nonlinear Sci.* **2016**, *26*, 116312. [CrossRef]
19. Hall, N. Two Concepts of Causation. In *Causation and Counterfactuals*; Collins, J., Hall, N., Paul, L., Eds.; MIT Press: Cambridge, MA, USA, 2004; pp. 225–276.
20. Pearl, J. Bayesianism and Causality, or, Why I Am Only a Half-Bayesian. In *Foundations of Bayesianism*; Corfield, D., Williamson, J., Eds.; Kluwer Academic Publishers: Dordrecht, the Netherlands, 2001; pp. 19–36.
21. White, H.; Chalak, K.; Lu, X. Linking Granger Causality and the Pearl Causal Model with Settable Systems. *JMRL Workshop Conf. Proc.* **2011**, *12*, 1–29.
22. White, H.; Chalak, K. Settable Systems: An Extension of Pearl's Causal Model with Optimization, Equilibrium, and Learning. *J. Mach. Learn. Res.* **2009**, *10*, 1759–1799.
23. Bollt, E. Synchronization as a process of sharing and transferring information. *Int. J. Bifurc. Chaos* **2012**, *22*, doi:10.1142/S0218127412502616. [CrossRef]
24. Lasota, A.; Mackey, M. *Chaos, Fractals, and Noise: Stochastic Aspects of Dynamics*; Springer: New York, NY, USA, 2013.
25. Pinsker, M.S. Information and information stability of random variables and processes. *Dokl. Akad. Nauk SSSR* **1960** *133*, 28–30.
26. Boucheron, S.; Lugosi, G.; Massart, P. *Concentration Inequalities: A Nonasymptotic Theory of Independence*; Oxford University Press: Oxford, UK, 2013.
27. Sauer, T.; Yorke, J.A.; Casdagli, M. Embedology. *Stat. Phys.* **1991**, *65*, 579–616. [CrossRef]
28. Sauer, T.; Yorke, J.A. Are the dimensions of a set and its image equal under typical smooth functions? *Ergod. Theory Dyn. Syst.* **1997**, *17*, 941–956. [CrossRef]
29. Grassberger, P.; Procaccia, I. Measuring the strangeness of strange attractors. *Phys. D Nonlinear Phenom.* **1983**, *9*, 189–208. doi:10.1016/0167-2789(83)90298-1. [CrossRef]
30. Grassberger, P.; Procaccia, I. Characterization of Strange Attractors. *Phys. Rev. Lett.* **1983**, *50*, 346–349. doi:10.1103/PhysRevLett.50.346. [CrossRef]
31. Grassberger, P. Generalized dimensions of strange attractors. *Phys. Lett. A* **1983**, *97*, 227–230. doi:/10.1016/0375-9601(83)90753-3. [CrossRef]
32. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. *Phys. Rev. E* **2004**, *69*, 066138. doi:10.1103/PhysRevE.69.066138. [CrossRef]
33. Rigney, D.; Goldberger, A.; Ocasio, W.; Ichimaru, Y.; Moody, G.; Mark, R. Multi-channel physiological data: Description and analysis. In *Time Series Prediction: Forecasting the Future and Understanding the Past*; Addison-Wesley: Boston, MA, USA, 1993; pp. 105–129.
34. Ichimaru, Y.; Moody, G. Development of the polysomnographic database on CD-ROM. *Psychiatry Clin. Neurosci.* **1999**, *53*, 175–177. [CrossRef] [PubMed]