entropy

MDPI

Article

Causal Inference in Time Series in Terms of Rényi Transfer Entropy

Petr Jizba [1,*,†], Hynek Lavička [1,†,‡] and Zlata Tabachová [2,†]

[1] Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, Břehová 7, 115 19 Prague, Czech Republic; hynek.lavicka@fjfi.cvut.cz
[2] Complexity Science Hub Vienna, Josefstädter Straße 39, 1080 Vienna, Austria; zlata.tabachova@fjfi.cvut.cz
* Correspondence: p.jizba@fjfi.cvut.cz; Tel.: +420-775-317-309
† These authors contributed equally to this work.
‡ Current address: Blocksize Capital GmbH, Taunusanlage 8, D-60329 Frankfurt am Main, Germany.

Abstract: Uncovering causal interdependencies from observational data is one of the great challenges of a nonlinear time series analysis. In this paper, we discuss this topic with the help of an information-theoretic concept known as Rényi's information measure. In particular, we tackle the directional information flow between bivariate time series in terms of Rényi's transfer entropy. We show that by choosing Rényi's parameter α, we can appropriately control information that is transferred only between selected parts of the underlying distributions. This, in turn, is a particularly potent tool for quantifying causal interdependencies in time series, where the knowledge of "black swan" events, such as spikes or sudden jumps, are of key importance. In this connection, we first prove that for Gaussian variables, Granger causality and Rényi transfer entropy are entirely equivalent. Moreover, we also partially extend these results to heavy-tailed α-Gaussian variables. These results allow establishing a connection between autoregressive and Rényi entropy-based information-theoretic approaches to data-driven causal inference. To aid our intuition, we employed the Leonenko et al. entropy estimator and analyzed Rényi's information flow between bivariate time series generated from two unidirectionally coupled Rössler systems. Notably, we find that Rényi's transfer entropy not only allows us to detect a threshold of synchronization but it also provides non-trivial insight into the structure of a transient regime that exists between the region of chaotic correlations and synchronization threshold. In addition, from Rényi's transfer entropy, we could reliably infer the direction of coupling and, hence, causality, only for coupling strengths smaller than the onset value of the transient regime, i.e., when two Rössler systems are coupled but have not yet entered synchronization.

Keywords: Rényi entropy; Rényi transfer entropy; Rössler system; multivariate time series

Citation: Jizba, P.; Lavička, H.; Tabachová, Z. Causal Inference in Time Series in Terms of Rényi Transfer Entropy. *Entropy* **2022**, *24*, 855. https://doi.org/10.3390/e24070855

Academic Editor: Joanna Olbryś

Received: 17 March 2022
Accepted: 11 June 2022
Published: 22 June 2022

1. Introduction

The time evolution of a complex system is usually recorded in the form of a time series. Time series analysis is a traditional field of mathematical statistics; however, the development of nonlinear dynamical systems and the theory of deterministic chaos have opened up new vistas in the analysis of nonlinear time series [1,2]. The discovery of the synchronization of chaotic systems [3] has changed the study of interactions and cooperative behavior of complex systems and brought new approaches to studying the relations between nonlinear time series [4]. During the process of synchronization, two systems can either mutually interact or only one can influence the other. In order to distinguish these two ways, and to find which system is the driver ("master") and which is the response ("slave") system, a number of approaches from the dynamical system theory have been proposed [5–8]. The aforementioned problem of synchronization can be seen as part of a broader framework known as *causality* or *causal relations between systems*, processes, or phenomena. The mathematical formulation of causality, in terms of predictability, was first proposed by Wiener [9] and formulated for the time series by

Granger [10]. In particular, Granger introduced what is now known as *Granger causality*, which is a statistical concept of causality that is based on the evaluation of predictability in bivariate autoregressive models.

Extracting causal interdependencies from observational data is one of the key tasks in a nonlinear time series analysis. Apart from the linear Granger causality and various nonlinear extensions thereof [11–13], existing methods for this purpose include state-space-based approaches, such as conditional probabilities of recurrence [14–16], or information-theoretic quantities, such as conditional mutual information [17,18] and *transfer entropies* [2,19–21]. In particular, the latter information-theoretic quantities represent powerful instruments in quantifying causality between time-evolving systems. This is because ensuing information-theoretic functionals (typically based on Shannon entropy) quantify—in a non-parametric and explicitly non-symmetric way—the flow of information between two (or more) time series. In particular, transfer entropies (TEs) have recently received considerable attention. The catalyst was the infusion of new (numerical and conceptional) ideas. For instance, the performances of the Shannon entropy-based conditional entropies and conditional mutual entropies have been, in recent years, extensively tested using numerically-generated time series [17,22]. Sophisticated algorithms have been developed to uncover direct causal relations in multivariate time series [23–25]. In parallel, increasing attention has been devoted to the development of reliable estimators of entropic functionals to detect causality from nonlinear time series [26]. At the same time, it has been recognized that information-theoretic approaches play important roles in dealing with complex dynamical systems that are multiscale and/or non-Gaussian [21,27–29]. The latter class includes complex systems with heavy-tailed probability distributions epitomized, e.g., in financial and climatological time series [30,31].

In this paper, we extend the popular Shannon entropy-based TE (STE), which represents a prominent tool for assessing directed information flow between joint processes, and quantifies information transfer in terms of *Rényi's TE* (RTE). RTE was introduced by one of us (PJ) in reference [21] in the context of a bivariate financial time series. The original idea was to use the RTE in order to exploit the theoretical formulation that could identify and quantify peculiar features in multiscale bivariate processes (e.g., multiscale patterns, generalized fractal dimensions, or multifractal cross-correlations) that are often seen in finance. In contrast to [21], where the focus was mostly on qualitative aspects of Rényian information flow between selected stock-market time series, in the present work, we wish to be more quantitative by analyzing coupled time series that are numerically generated from known dynamics. Specifically, we demonstrate how *the RTE method performs in the detection of the coupling direction and onset of synchronization between two Rössler oscillators* [32] *that are unidirectionally coupled in the first variable* x. The Rössler system (RS) is a paradigmatic and well-studied low-dimensional chaotic dynamical system. When coupled, RSs allow for *synchronization* as well as a subtle phenomenon known as "phase synchronization", i.e., when the amplitudes of both systems are not correlated while the phases are approximately equal. In this respect, the synthetic bivariate time series (generated from coupled RSs) serves as an excellent test-bed, allowing to numerically analyze, e.g., drive–response relationships or identify the ensuing onset (or threshold) of synchronization. In doing so, we identify factors and influences that can lead to either decreases in the RTE sensitivity or false detections and propose some ways to cope with them. The aforementioned issues have not been explicitly studied in the framework of the RTE; this work presents the first attempt in this direction.

To set the stage, we shall first, in Section 2, provide the information-theoretic background on Rényi entropy (RE), which will be needed in the main body of the text. For self-consistency of our exposition, we briefly review Shannon's transfer entropy of Schreiber and motivate and derive the core quantity of this work—the Rényi transfer entropy. The issue of causality (and its connection to RTE) is examined in Section 3. In particular, we prove that the Granger causality is entirely equivalent to the RTE for Gaussian processes and show how the Granger causality and the RTE are related in the case of heavy-tailed

(namely α-Gaussian) processes. Section 4 is dedicated to derived information-theoretic concepts, such as the balance of transfer entropy and effective transfer entropy that will be employed in our analysis. The proposed framework is then illustrated on two unidirectionally coupled Rössler systems as a paradigmatic example. To cultivate our intuition about the latter RSs, we discuss in Section 5 the inner workings of such RSs in terms of simple numerical experiments. The ensuing numerical analysis is presented in Section 6, where we discuss how the RTE can be used to detect causality and the onset of synchronization in the two coupled RSs. We also demonstrate how the RTE provides non-trivial insight into the structure of a transient regime that exists between the regions of chaotic correlations and the onset of synchronization. Finally, Section 7 summarizes our theoretical and numerical findings and discusses possible extensions of the present work. For the reader's convenience, we relegate some technical issues concerning the RE estimator employed and the statistical significance of results presented to Appendices A and B.

2. Rényi Entropy

Information theory approaches based on Shannon entropy currently belong in the portfolio of techniques and tools that are indispensable in addressing causality issues in complex dynamical systems. At the same time, Shannon's information theory is limited in its scope. In fact, since Shannon's seminal papers [33], it has been known that Shannon's information measure (or entropy) represents mere idealized information, appearing only in situations when the buffer memory (or storage capacity) of a transmitting channel is infinite. In particular, Shannon's source coding theorem (or noiseless coding theorem), which establishes the limits to possible data compression and, thus, provides operational meaning to the Shannon entropy, assumes that the *cost* of a codeword is a linear function of its length (so the optimal code has a minimal cost out of all codes). However, the linear costs of codewords are not always desirable. For instance, when the storage capacity is finite one would aim to penalize excessively lengthy codewords with a price that is, e.g., exponential rather than the linear function of the length.

For these reasons, information theorists have devised various remedies to deal with such cases. This usually consists of substituting Shannon's information measure with information measures of other types. Consequently, numerous generalizations of Shannon's entropy have started to proliferate in the information-theory literature, ranging from additive entropies [34,35] to a rich class of non-additive entropies [36–40], to more exotic types of entropies [41]. The one-parametric class of information measures, known as *Rényi entropies*, introduced by Hungarian mathematician and information theorist Alfred Rényi in the early 1960s [42,43], is particularly prominent among such generalizations. Applications of RE in information theory, namely its generalization to coding theorems, were carried over by Campbel [44], Csiszár [45,46], Aczél [47], and others. In a physical setting, RE was popularized in the context of chaotic dynamical systems by Kadanoff et al. [48] and in connection with multifractals by Mandelbrot [49]. RE is also indispensable in the quantum information theory where it quantifies multipartite entanglement [50].

In its essence, REs constitute a one-parametric family of information measures labeled by parameter α, fulfilling the additivity with respect to the composition of statistically independent systems. The special case with $\alpha = 1$ corresponds to ordinary Shannon's entropy. REs belong to a broader class of so-called Uffink entropic functionals [51,52], i.e., the most general class of solutions that satisfy Shorem–Johnson axioms for the maximum entropy principle in the statistical estimation theory. Moreover, it might be shown that Rényi entropies belong to the class of the so-called mixing homomorphic functions [53] and that they are analytic for $\alpha \in \mathbb{C}_{I \cup IV}$, cf. [34].

2.1. Definition

RE is defined as an exponentially weighted mean of the *Hartley information measure* $-\log p$ (i.e., elementary measure of information) [54]. In fact, it was shown by Rényi that, except for a linearly-weighted average (which leads to Shannon entropy), exponential

weighting is the only possible averaging that is both compatible with the Kolmogorov–Nagumo average prescription and leads to entropies that are additive, with respect to independent systems [42,43]. RE, associated with a system described with a probability distribution $\mathcal{P}$, reads

$$H_\alpha[\mathcal{P}] \;=\; \frac{1}{1-\alpha}\log_2\sum_{i=1}^{n} p_i^\alpha\,. \tag{1}$$

RE has the following properties [34,43]:

- RE is symmetric, i.e., $H_\alpha[\{p_1,\ldots,p_n\}] = H_\alpha[\{p_{\pi(1)},\ldots,p_{\pi(n)}\}]$;
- RE is non-negative, i.e., $H_\alpha \geq 0$;
- $\lim_{\alpha\to 1} H_\alpha = H_1$, where $H_1 = H$ is the Shannon entropy;
- $H_0 = \log_2 n$ is the *Hartley* entropy and $H_2 = -\log_2\sum_{i=1}^{n} p_i^2$ is the *Collision entropy*;
- $0 \leq H_\alpha[\mathcal{P}] \leq \log_2 n$;
- H_α is a positive, decreasing the function of $\alpha \geq 0$.

Let us mention that $H_\alpha[\mathcal{P}]$ with different αs complement each other. This is because for each specific α, the ensuing $H_\alpha[\mathcal{P}]$ carries extra information that is not present in any other $H_\beta[\mathcal{P}]$ with $\beta \neq \alpha$. In information theory, this fact is known as the *reconstruction theorem*, namely, the underlying distribution $\mathcal{P}$ can be uniquely reconstructed only if all $H_\alpha[\mathcal{P}]$ are known, [21,34,55]. In chaotic dynamical systems, the reconstruction theorem goes under the name *complementary generalized dimensions* [56] (cf. also next subsection).

2.2. Multifractals, Chaotic Systems, and Rényi Entropy

Another appealing property of the Rényi entropy is its close connection to *multifractals*, i.e., the mathematical paradigm that is often encountered in complex dynamical systems with examples ranging from turbulence and strange attractors to meteorology and finance, see, e.g., [57]. The aforementioned connection is established through the so-called *generalized dimensions*, which are defined as [2,48]

$$D_\alpha \;=\; -\lim_{\delta\to 0}\frac{H_\alpha(\delta)}{\log\delta} \tag{2}$$

where δ is a size of a $\delta-$mesh covering of a configuration space of a system. Generalized dimensions D_α are conjugate to the *multifractal spectrum* $f(\beta)$ through the Legendre transform [48]

$$(\alpha-1)D_\alpha \;=\; \alpha\beta \;-\; f(\beta). \tag{3}$$

The function $f(\beta)$ is called the multifractal spectrum because β plays the role of the scaling exponent in the local probability distribution, e.g., distribution with support on the i-th hypercube of a mesh size δ scale, as $p_i(\delta) \sim \delta^{\beta_i}$. The key assumption in the multifractal analysis is that in the small $\delta-$ limit, the local probability distribution depends smoothly on β. It can be argued that $f(\beta)$ corresponds to the (box-counting) fractal dimension of the portion of the configuration space where local probability distributions have the scaling exponent β, cf., e.g., reference [34]. In this way, the multifractal can be viewed as an ensemble of intertwined (uni)fractals, each with its own fractal dimension $f(\beta)$.

The multifractal paradigm is particularly pertinent in the *theory of chaotic systems*. For instance, chaotic dynamics and strange attractors, in particular, are uniquely characterized by the infinite sequences of generalized dimensions D_α, cf. reference [56]. In particular, the generalized dimensions can help to recognize (in a quantitative way) the main geometric features of chaotic systems. For instance, they may help to distinguish chaotic behavior from noisy behavior, determine the number of variables that are needed to model the dynamics of the system or classify systems into universality classes. On the other hand, dynamical features of chaotic systems are often analyzed through such quantifiers as *Lyapunov exponent*, which is a measure of the divergence of nearby trajectories, or ensuing *Kolmogorov-Sinai entropy rate* (KSE), which quantifies the change of entropy as the system evolves and is given by the sum of all positive Lyapunov exponents. The connection

between KSE and the time evolution of the information-theoretic or statistical entropy is quite delicate, see, e.g., the discussion in reference [58], though the upshot is clear, in order to describe the dynamics of a (complex) system, the temporal change or the difference in entropy is more relevant than the entropy itself. Consequently, while RE (alongside with D_α) is a suitable quantifier of geometric properties of chaotic systems, its temporal differences or temporal rates are useful for the description of the dynamics of such systems. Rényi's transfer entropy follows the latter route.

2.3. Shannon Transfer Entropy

In order to understand the concept of Rényi transfer entropy, we recall first its Shannon's counterpart.

Let $X = \{x_i\}_{i=1}^N$ be a discrete random variable with ensuing probability distribution $\mathcal{P}_X$, then the Shannon entropy of this process is

$$H(X) \equiv H(\mathcal{P}_X) = -\sum_{x\in\mathcal{X}} p(x)\log_2 p(x)\,. \tag{4}$$

Let $Y = \{y_i\}_{i=1}^N$ be another random variable, then *mutual information* between X and Y is

$$\begin{aligned} I(X:Y) &= \sum_{x\in X,\, y\in Y} p(x,y)\log_2\frac{p(x,y)}{p(x)p(y)} \\ &= H(X) - H(X|Y) = H(Y) - H(Y|X)\,, \end{aligned} \tag{5}$$

where quantity $H(X|Y)$ is the *conditional entropy*, defined as

$$H(X|Y) = -\sum_{x\in X,\, y\in Y} p(x,y)\log_2 p(x|y)\,. \tag{6}$$

Mutual information quantifies an average reduction in uncertainty (i.e., gain in information) about X resulting from the observation of Y, or vice versa. Since $I(X:Y) = I(Y:X)$, it cannot be used as a measure of directional information flow. Note also that the amount of information contained in X about itself is just the Shannon entropy, i.e., $I(X:X) = H(X)$.

The mutual information between two processes X and Y conditioned on the third process Z is called *conditional mutual information* and is defined as

$$I(X:Y|Z) = H(X|Z) - H(X|Y,Z) = I(X:(Y,Z)) - I(X:Y)\,. \tag{7}$$

Let us now consider two time sequences (e.g., two stock market time series) described by stochastic (possibly vector-type) random variables X_t and Y_t. Let us assume further that the time steps (e.g., data ticks) are discrete with the time step τ and with $t_n = t_0 + n\tau$ where t_0 is some reference time. For practical purposes, it is also useful to assume that X_t and Y_t represent discrete-time stochastic Markov processes of order k and l, respectively.

We wish to know what information will be gained on $X_{t_{n+1}}$ by observing Y_t up to time t_n. To this end, we introduce the joint process $X_{t_n}, X_{t_{n-1}}, \ldots, X_{t_{n-k+1}}$, which we denote as $X_n^{(k)}$, and similarly, we define the joint process $Y_n^{(l)} \equiv Y_{t_n}, Y_{t_{n-1}}, \ldots, Y_{t_{n-l+1}}$. By replacing X in (7) by $X_{t_{n+1}}$, Y by $Y_n^{(l)}$, and Z by $X_n^{(k)}$, we obtain the desired conditional mutual information

$$\begin{aligned} I(X_{t_{n+1}} : Y_n^{(l)}|X_n^{(k)}) &= H(X_{t_{n+1}}|X_n^{(k)}) - H(X_{t_{n+1}}|Y_n^{(l)}, X_n^{(k)}) \\ &= \sum_{x_n^{(k)}\in X_{n+1}^{(k)},\, y_n^{(l)}\in Y_n^{(l)}} p(x_{n+1}, x_n^{(k)}, y_n^{(l)})\log_2\left(\frac{p(x_{n+1}|x_n^{(k)}, y_n^{(l)})}{p(x_{n+1}|x_n^{(k)})}\right). \end{aligned} \tag{8}$$

The conditional mutual information (8) is also known as *Shannon transfer entropy* from Y_t to X_t (or simply from Y to X) and as a measure of the directed (time asymmetric) infor-

mation transfer between joint processes, it was introduced by Schreiber in reference [19]. The latter is typically denoted as

$$T_{Y\to X}(k,l) \;\equiv\; I(X_{t_{n+1}} : Y_n^{(l)} | X_n^{(k)})\,. \tag{9}$$

As already mentioned, for independent processes, TE is equal to zero. For a non-zero case transfer, entropy measures the deviation from the independence of the two processes. An important property of the transfer entropy is that it is directional, i.e., in general, $T_{Y\to X} \neq T_{X\to Y}$.

2.4. Rényi Transfer Entropy

In the same manner as in (7), we can introduce the *Rényi transfer entropy of order* α from Y to X (see also reference [21]) as

$$\begin{aligned} T^R_{\alpha,Y\to X}(k,l) &= H_\alpha(X_{t_{n+1}}|X_n^{(k)}) \;-\; H_\alpha(X_{t_{n+1}}|X_n^{(k)}, Y_n^{(l)}) \\ &= I_\alpha(X_{t_{n+1}} : Y_n^{(l)} | X_n^{(k)})\,, \end{aligned} \tag{10}$$

where $H_\alpha(X|Y)$ is the *conditional entropy of order* α and $I_\alpha(X:Y)$ is the *mutual information of order* α. These can be explicitly written as [21,43]

$$\begin{aligned} H_\alpha(X|Y) &= \frac{1}{1-\alpha}\log_2 \frac{\sum_{x\in X, y\in Y} p^\alpha(x,y)}{\sum_{y\in Y} p^\alpha(y)}\,, \\ I_\alpha(X:Y) &= \frac{1}{1-\alpha}\log_2 \frac{\sum_{x\in X, y\in Y} p^\alpha(x) p^\alpha(y)}{\sum_{x\in X, y\in Y} p^\alpha(x,y)}\,. \end{aligned} \tag{11}$$

It can be checked (via L'Hospital's rule) that Rényi's transfer α-entropy reduces to Shannon TE in the $\alpha \to 1$ limit, i.e.,

$$\lim_{\alpha\to 1} T^R_{\alpha,Y\to X} \;=\; T_{Y\to X}\,. \tag{12}$$

From (10), we see that $T^R_{\alpha,Y\to X}(k,l)$ may be intuitively interpreted as the degree of ignorance (or uncertainty) about $X_{t_{n+1}}$ resolved by the past states $Y_n^{(l)}$ and $X_n^{(k)}$, over and above the degree of ignorance about $X_{t_{n+1}}$ already resolved by its own past state alone. Here, the ignorance is quantified by the Rényi information measure (i.e., RE) of order α.

Rényi TE can also be negative (unlike the Shannon TE). This means that the uncertainty of the process X_t becomes bigger knowing the past of Y_t, i.e., $H_\alpha(X_{t_{n+1}}|X_n^{(k)}) \leq H_\alpha(X_{t_{n+1}}|X_n^{(k)}, Y_n^{(l)})$. If X_t and Y_t are independent, then $T^R_{\alpha,Y\to X} = T^R_{\alpha,X\to Y} = 0$. However, in contrast to Shannon's case, the fact that $T^R_{\alpha,Y\to X} = 0$ does necessarily imply the independence of the two underlying stochastic processes. Nonetheless, in Section 3, we prove that in case of Gaussian (Wiener) processes, 0-valued RTE is a clear signature of independence.

Due to the *reconstruction theorem* mentioned in Section 2.1, RTE $T^R_{\alpha,Y\to X}$ conveys for each α a different type of directional information from Y to X. The essence of this statement can be understood qualitatively by introducing the so-called *escort distribution*.

2.5. Escort Distribution

Because of the nonlinear way in which probability distributions enter in the definition of RE, cf. Equation (1), the RTE represents a useful measure of transmitted information that quantifies the dominant information flow between certain parts of underlying distributions. In fact, for $0 < \alpha < 1$, the corresponding information flow accentuates marginal events, while for $\alpha > 1$, more probable (close-to-average) events are emphasized [21]. In this respect, one can zoom or amplify different parts of probability density functions involved by merely choosing appropriate values of α. This is particularly useful in studies of time

sequences, where marginal events are of crucial importance, for instance, in financial time series.

In order to better understand the aforementioned "zooming" property of RTE, we rewrite (10) in the form

$$T^{R}_{\alpha,Y\to X}(k,l) \;=\; \frac{1}{1-\alpha}\log_2\left(\frac{\sum \frac{p^{\alpha}(x_n^{(k)})}{\sum p^{\alpha}(x_n^{(k)})}\, p^{\alpha}(x_{n+1}|x_n^{(k)})}{\sum \frac{p^{\alpha}(x_n^{(k)},y_n^{(l)})}{\sum p^{\alpha}(x_n^{(k)},y_n^{(l)})}\, p^{\alpha}(x_{n+1}|x_n^{(k)},y_n^{(l)})}\right). \tag{13}$$

This particular representation shows how the underlying distribution changes (or deforms) with the change of parameter α. The numerator and denominator inside the log-function contain the so-called *escort* (or *zooming*) *distributions* ρ_α

$$\rho_\alpha(x) \;\equiv\; \frac{p^{\alpha}(x)}{\sum_{x\in X} p^{\alpha}(x)}, \tag{14}$$

which emphasize less probable events for $0<\alpha<1$ and more probable events when $\alpha>1$, see Figure 1.

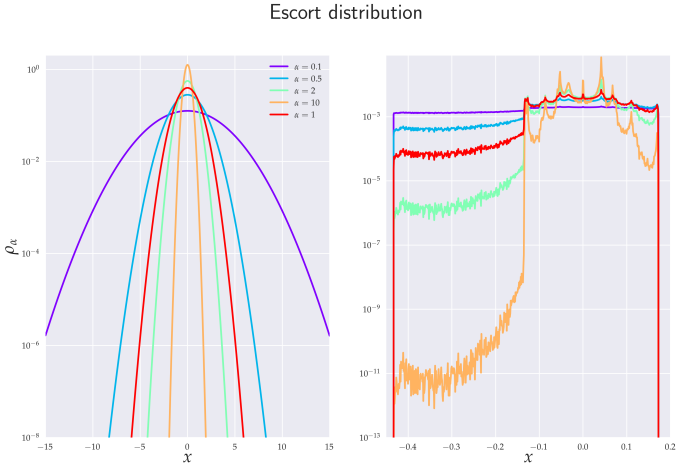

Figure 1. Illustration of the concept of escort distribution ρ_α on histograms. The left figure depicts log-scaled normal distribution $\mathcal{N}(0,1)$, while in the right figure, we show the log-scaled histogram for x_1-projection increments from the Rössler system (51). Both figures demonstrate that the escort distribution deforms the original distribution ($\alpha=1$) so that $0<\alpha<1$ less probable events are emphasized (the smaller, α the greater emphasis) while high probable events are accordingly suppressed. For $\alpha>1$, the situation is reversed.

Note also that $\rho_\alpha(x_n^{(k)},y_n^{(l)})$ is not the joint probability distribution of $X_n^{(k)}$ and $Y_n^{(l)}$ as it does not satisfy the Kolmogorov–de Finetti relation for conditional probabilities [59].

In connection with (13), we may note that for $0<\alpha<1$ the multiplicative factor is positive, and so the RTE is negative if, by learning $Y_n^{(l)}$, the rare events are (on average) more emphasized than in the case when only $X_n^{(k)}$ alone is known. Analogically, for $\alpha>1$ the RTE can be negative when—by learning $Y_n^{(l)}$—the more probable events are (on average) more accentuated in comparison with the situation when $Y_n^{(l)}$ is not known. It should be stressed that the analogous situation does not hold for Shannon's TE. This is because in the limit $\alpha\to 1$ we regain expression (8), which is nothing but relative entropy, and as such, it is always non-negative due to Gibbs inequality. At the same time, Shannon's TE is, by its very definition, also mutual information. While RTE is also defined to be a mutual information, it is not relative entropy (in the RE case, those two concepts do not coincide). It can be shown (basically via Jensen's inequality) [34] that the relative entropy based on

RE is also non-negative but this is not true for ensuing mutual information, which serves as a conceptual basis for the definition of RTE.

3. Rényi Transfer Entropy and Causality

As already seen, Rényi TE (analogously to Shannon TE) is a directional measure of information transfer. Let us now comment on the connection of the RTE with the causality concept.

3.1. Granger Causality—Gaussian Variables

The first general definition of causality, which could be quantified and measured computationally was given by Wiener in 1956, namely "*... For two simultaneously measured signals, if we can predict the first signal better by using the past information from the second one than by using the information without it, then we call the second signal causal to the first one...*" [9].

The introduction of the concept of causality into the experimental practice, namely into analyses of data observed in consecutive time instants (i.e., time series), is due to the Nobel prize winner (economy, 2003) C.W.J. Granger. The so-called *Granger causality* is defined so that the process Y_t *Granger causes* another process X_t if, in an appropriate statistical sense, Y_t assists in predicting the future of X_t beyond the degree to which X_t already predicts its own future.

The standard test of the Granger causality was developed by Granger himself [10] and it is based on a linear regression model, namely

$$X_t \;=\; a_{0t} \;+\; \sum_{\ell=1}^{k} a_{1\ell} X_{t-\ell} \;+\; \sum_{\ell=1}^{l} a_{2\ell} Y_{t-\ell} \;+\; e_t\,, \tag{15}$$

where $a_0, a_{1\ell}, a_{2\ell}$ are (constant) regression coefficients, l and k represent the maximum number of lagged observations included in the model (i.e., memory indices), t is a discrete time with the time step τ (ℓ is also quantified in units of τ) and e_t is the uncorrelated random variable (residual) with zero mean and variance σ^2. The *null hypothesis* that Y_t does not cause X_t (in the sense of Granger) is not rejected if and only if $a_{2\ell} = 0$ for $\ell = 1, \ldots, l$. In the latter case, we will call the ensuing regression model the *reduced regression model*.

It is not difficult to show that for Gaussian variables, the RTE and Granger causality are entirely equivalent. To see this, we use the *standard measure* of the Granger causality, which is defined as [60]

$$\mathcal{F}^{(k,l)}_{Y\to X} \;=\; \log_2 \frac{|\Sigma(e'_t)|}{|\Sigma(e_t)|}\,, \tag{16}$$

where $\Sigma(\ldots)$ is the covariance matrix, $|\ldots|$ denotes the matrix determinant, and e_t, e'_t are residuals in the full and reduced regression model, respectively. We chose the logarithm to the base 2, rather than e for technical convenience. We now prove the following theorem:

Theorem 1. *If the joint process X_t, Y_t is Gaussian, then there is an exact equivalence between the Granger causality and RTE, namely*

$$\mathcal{F}^{(k,l)}_{Y\to X} \;=\; 2T^{R}_{\alpha,Y\to X}(k,l)\,. \tag{17}$$

This can be proved in the following way (for an analogous proof for Shannon's TE, see [61]). We first define the *partial covariance* as

$$\Sigma(\mathbf{X}|\mathbf{Y}) \;=\; \mathbf{\Sigma}(\mathbf{X}) \;-\; \mathbf{\Sigma}(\mathbf{X},\mathbf{Y})\mathbf{\Sigma}(\mathbf{Y})^{-\mathbf{1}}\mathbf{\Sigma}(\mathbf{X},\mathbf{Y})^{\top}\,, \tag{18}$$

where $\Sigma(\mathbf{X})_{ij} = \mathrm{cov}(X_i, X_j)$ and $\Sigma(\mathbf{X},\mathbf{Y})_{ij} = \mathrm{cov}(X_i, Y_j)$ with $\mathbf{X}$ and $\mathbf{Y}$ being random vector (or multivariate) variables. Let $\mathbf{X}$ and $\mathbf{Y}$ be jointly distributed random vectors in the linear regression model

$$\mathbf{X} \;=\; \mathbf{a} \;+\; \mathbf{Y}\mathbb{A} \;+\; \mathbf{e}\,. \tag{19}$$

Here, $\mathbf{a}$ is a constant vector, $\mathbb{A}$ contains regression coefficients, and $\mathbf{e}$ is a residual random vector with zero mean. In the subsequent, we will identify both $\mathbf{X}$ and $\mathbf{Y}$ with stochastic vectors (see text after Equation (28)). In such a case, one can always choose a specified number of time lags, so that system (19) (or better (23) and, consequently, (22)) is uniquely solvable, as neither vector $\mathbf{a}$ nor matrix $\mathbb{A}$ are time-dependent.

We now apply the least square method to the mean square error

$$\mathcal{E}^2 \;\equiv\; \sum_i \mathbb{E}(e_i^2) \;=\; \sum_i \mathbb{E}\Big[(\mathbf{X}-\mathbf{Y}\mathbb{A}-\mathbf{a})_i^2\Big]\,, \tag{20}$$

Here, $\mathbb{E}(\ldots)$ denotes the average value. The ensuing least square equations

$$\frac{\partial \mathcal{E}^2}{\partial \mathbb{A}_{ij}} \;=\; 0 \quad \text{and} \quad \frac{\partial \mathcal{E}^2}{\partial a_k} \;=\; 0\,, \tag{21}$$

yield

$$a_l \;=\; \mathbb{E}(X_l) \;-\; \sum_k \mathbb{E}(Y_k)\mathbb{A}_{kl}\,, \tag{22}$$

$$\mathbb{A}_{li} \;=\; \sum_j [\Sigma(\mathbf{X})]^{-1}_{lj}\,\Sigma(\mathbf{Y},\mathbf{X})_{ji}\,. \tag{23}$$

From (19) follows that

$$\mathbb{E}(X_iX_j) \;=\; \mathbb{E}\big[(\mathbf{a} \;+\; \mathbf{Y}\mathbb{A} \;+\; \mathbf{e})_i(\mathbf{a} \;+\; \mathbf{Y}\mathbb{A} \;+\; \mathbf{e})_j\big]\,, \tag{24}$$

which after employing (22) can be equivalently rewritten as

$$\mathrm{cov}(X_i,X_j) \;=\; \sum_{l,k} \mathrm{cov}(Y_l,Y_k)\mathbb{A}_{li}\mathbb{A}_{kj} \;+\; \mathrm{cov}(e_i,e_j)\,, \tag{25}$$

or equivalently

$$\Sigma(\mathbf{X}) \;=\; \mathbb{A}^\top \Sigma(\mathbf{Y})\mathbb{A} \;+\; \Sigma(\mathbf{e})\,. \tag{26}$$

If we now insert (23)–(26), we obtain

$$\mathrm{cov}(e_i,e_j) \;=\; \mathrm{cov}(X_i,X_j) \;-\; \mathrm{cov}(X_i,Y_k)[\mathrm{cov}(Y_k,Y_i)]^{-1}[\mathrm{cov}(X_i,Y_j)]^\top\,, \tag{27}$$

which might be equivalently written as

$$\Sigma(\mathbf{e}) \;=\; \Sigma(\mathbf{X}|\mathbf{Y})\,. \tag{28}$$

If we now take $\mathbf{X} = (X_{t_{n+1}})$, $\mathbf{a} = (a_0)$, $\mathbf{Y} = (X^{(k)}, Y^{(l)})$, $\mathbb{A} = \mathrm{diag}(a_{1n}^{(k)}, a_{2n}^{(l)})$ for the full regression model and $\mathbf{Y} = (X_n^{(k)})$, $\mathbb{A} = \mathrm{diag}(a_1^{(k)})$ for the reduced regression model, we might write that

$$\mathcal{F}^{(k,l)}_{Y\to X} \;=\; \log_2 \frac{|\Sigma(e_t')|}{|\Sigma(e_t)|} \;=\; \log_2\left(\frac{|\Sigma(X_{t_{n+1}}|X_n^{(k)})|}{|\Sigma(X_{t_{n+1}}|X_n^{(k)},Y_n^{(l)})|}\right). \tag{29}$$

At this stage, we can use the fact that RE of the multivariate Gaussian variable $\mathbf{X}$ is [62]

$$H_\alpha(\mathbf{X}) \;=\; \frac{1}{2}\log_2|\Sigma(\mathbf{X})| \;+\; \frac{D_X}{2}\log_2\Big(2\pi\alpha^{\alpha'/\alpha}\Big)\,. \tag{30}$$

Here, $D_{\mathbf{X}}$ is the dimension of $\mathbf{X}$ and α' is a Hölder dual variable to α (i.e., $1/\alpha + 1/\alpha' = 1$). In particular, for jointly multivariate Gaussian variables $\mathbf{X}$ and $\mathbf{Y}$, we can use (11) to write

$$\begin{aligned} H_\alpha(\mathbf{X}|\mathbf{Y}) &= \left[\frac{1}{2}\log_2|\Sigma(\mathbf{X}\oplus\mathbf{Y})| + \frac{D_{\mathbf{X}}+D_{\mathbf{Y}}}{2}\log_2\left(2\pi\alpha^{\alpha'/\alpha}\right)\right] \\ &- \left[\frac{1}{2}\log_2|\Sigma(\mathbf{Y})| + \frac{D_{\mathbf{Y}}}{2}\log_2\left(2\pi\alpha^{\alpha'/\alpha}\right)\right] \\ &= \frac{1}{2}\log_2|\Sigma(\mathbf{X}|\mathbf{Y})| + \frac{D_{\mathbf{X}}}{2}\log_2\left(2\pi\alpha^{\alpha'/\alpha}\right). \end{aligned} \tag{31}$$

Here, $\oplus$ denotes the direct sum. Employing finally the defining relation (10), we obtain

$$\begin{aligned} T^R_{\alpha,Y\to X}(k,l) &= H_\alpha(X_{t_{n+1}}|X_n^{(k)}) - H_\alpha(X_{t_{n+1}}|X_n^{(k)},Y_n^{(l)}) \\ &= \frac{1}{2}\log_2\left(\frac{|\Sigma(X_{t_{n+1}}|X_n^{(k)})|}{|\Sigma(X_{t_{n+1}}|X_n^{(k)},Y_n^{(l)})|}\right). \end{aligned} \tag{32}$$

This confirms the statement of Theorem 1. In addition, since the standard measure of Granger causality (16) is typically defined only for the univariate target and source variables X_t and Y_t, we can omit $|\ldots|$ in (29) and (32).

Theorem 1 deserves two comments. First, the theorem is clearly true for any α. In fact, it is α independent, which means that for Gaussian processes we can employ any RTE to test the Granger causality. This naturally generalizes the classical result of Barnett et al. [61] (see also [1]) that is valid for Shannon's TE. When TE is phrased in terms of the Shannon entropy, it is typically easier to use various multivariate autoregressive model fitting techniques (e.g., the Lewinson–Wiggins–Robinson algorithm or the least-squares linear regression approach [63]) to derive $\mathcal{F}^{(k,l)}_{Y\to X}$ more efficiently than by employing direct entropy/mutual information-based estimators. On the other hand, since the efficiency and robustness of RTE estimators crucially hinge on the parameter α employed [64] (see also our discussion in Section 4), it might be, in many cases, easier to follow the information-theoretic route to the Granger causality (provided the Gaussian framework is justified). One can even test the Gaussian assumption in the actual time series by determining the RTE for various α parameters and checking if the results are α independent.

Second, the exact equivalence between the Granger causality and RTE can be (in the Gaussian case) retraced to the fact that in Equation (30) the second additive term on the RHS is proportional to $D_{\mathbf{X}}$. It is not difficult to see (by a direct inspection) that this proportionality will be preserved in many other exponential distributions that satisfy the Markov factorization property. In these cases, the equivalence between the Granger causality and RTE statistics will also be preserved. However, for generic distributions, the additive term in (30) will no longer be a linear function of $D_{\mathbf{X}}$ and, hence, it will not be canceled. This, in turn, spoils the desired equivalence. In the following section, we will discuss one possible generalization of Theorem 1 in the context of heavy-tailed distributions.

3.2. Granger Causality—Heavy-Tailed Variables

It is not difficult to find relations analogous to (32) in a more general setting. Here, we will illustrate this point with heavy-tailed (namely α-Gaussian) random variables, where computations can be conducted analytically.

It is well known that if variance and mean are the only statistical observables, then the conventional maximum entropy principle (MaxEnt) based on Shannon entropy yields Gaussian distribution. Similarly, if the very same MaxEnt is applied to Rényi entropy H_α, one obtains the so-called α-Gaussian distribution [34] (cf. also Figure 2)

$$p_i = \frac{1}{\mathcal{Z}_\alpha}\left[1-\beta(\alpha-1)x_i^2\right]_+^{1/(\alpha-1)}, \tag{33}$$

that decays asymptotically following power law. Here, $\beta \in \mathbb{R}^+$ and $[z]_+ = z$ if $z \geq 0$ and 0, otherwise, $\mathcal{Z}_\alpha$ is the normalization factor. It is more conventional to write (33) as

$$p_i \;=\; \mathcal{Z}_\alpha^{-1} \exp_{\{2-\alpha\}}\left(-\beta x_i^2\right), \tag{34}$$

where

$$e^x_{\{\alpha\}} \;=\; [1 \;+\; (1-\alpha)x]_+^{1/(1-\alpha)}, \tag{35}$$

is the Box–Cox α-exponential [30].

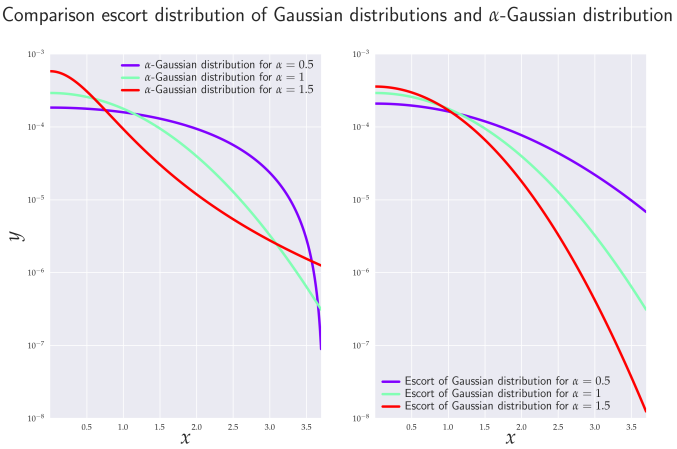

Figure 2. Comparison of the escort distributions ρ_α of the Gaussian (normal) distribution $\mathcal{N}(0,1)$ and α-Gaussian distributions (in log-linear plots) with a choice of β in (33), such that variances are the same for equal αs. For $\alpha = 1$, the two distributions correspond to the Gaussian distribution $\mathcal{N}(0,1)$. Even though ρ_α and α-Gaussian distributions deform the same underlying Gaussian distribution $\mathcal{N}(0,1)$, α-Gaussian is (save for $\alpha = 1$) heavy-tailed, while ρ_α remains Gaussian.

α-Gaussian distribution (33) has finite variance (and, more generally, the covariance matrix) for $\frac{D}{2+D} < \alpha \leq 1$. Let us now assume that Granger's linear (full/reduced) regression model is described by joint processes X_t and Y_t that are α-Gaussian. We now prove the following theorem:

Theorem 2. *If the joint process X_t, Y_t is α-Gaussian with $\alpha \in \left(\frac{1+k+l}{3+k+l}, 1\right]$ (i.e., a finite covariance matrix region) then $\mathcal{F}^{(k,l)}_{Y\to X} - 2T^R_{\alpha,Y\to X}(k,l)$ is a monotonically decreasing function of α (at fixed k and l) with zero reached at a stationary point $\alpha = 1$. The leading-order correction to the Granger causality is "k"-independent and has the form*

$$\mathcal{F}^{(k,l)}_{Y\to X} \;=\; 2T^R_{\alpha,Y\to X}(k,l) \;+\; \frac{l(\alpha-1)^2}{4} \;+\; \mathcal{O}((\alpha-1)^3). \tag{36}$$

This result explicitly illustrates how certain "soft" heavy-tailed processes can be related to the concept of the Granger causality via universal types of corrections that are principally discernible in data analysis.

Theorem 2 can be proved in close analogy with our proof of Theorem 1. In fact, all steps in the proof are identical up to Equation (29). For the D-dimensional α-Gaussian process, the scaling property (30) reads

$$H_\alpha(\mathbf{X}) \;=\; \frac{1}{2}\log_2|\Sigma(\mathbf{X})| \;+\; H_\alpha(\mathbf{Z}^{\mathbf{1},D}_\alpha). \tag{37}$$

Here, $\mathbf{Z}_\alpha^{\mathbf{1},D}$ represents an α-Gaussian random vector with zero mean and unit $(D \times D)$ covariance matrix. Relation (37) results from the following chain of identities

$$\begin{aligned} H_\alpha(\mathbf{X}) &= H_\alpha(\sqrt{\Sigma(\mathbf{X})}\,\mathbf{Z}_\alpha^{\mathbf{1},D}) \\ &= \frac{1}{1-\alpha}\log_2 \int_{\mathbb{R}^D} d^D\mathbf{y}\left(\int_{\mathbb{R}^D} d^D\mathbf{z}\,\delta\left(\mathbf{y}-\sqrt{\Sigma(\mathbf{X})}\,\mathbf{z}\right)\mathcal{F}(\mathbf{z})\right)^\alpha \\ &= \frac{1}{1-\alpha}\log_2\left[|\Sigma(\mathbf{X})|^{(1-\alpha)/2}\int_{\mathbb{R}^D} d^D\mathbf{y}\,\mathcal{F}^\alpha(\mathbf{y})\right] \\ &= \frac{1}{2}\log_2|\Sigma(\mathbf{X})| \;+\; H_\alpha(\mathbf{Z}_\alpha^{\mathbf{1},D})\,, \end{aligned} \tag{38}$$

which is clearly valid for any non-singular covariance matrix. The derivation $\mathcal{F}(\ldots)$ denoted the α-Gaussian probability density function with the unit covariance matrix and zero mean. We can now use the simple fact that

$$\begin{aligned} H_\alpha(\mathbf{Z}_\alpha^{\mathbf{1},D}) &= \log_2\left[\left(\frac{\pi}{\mathfrak{b}(1-\alpha)}\right)^{D/2}\frac{\Gamma\left(\frac{1}{1-\alpha}-\frac{D}{2}\right)}{\Gamma\left(\frac{1}{1-\alpha}\right)}\left(1-\frac{D}{2\alpha}(1-\alpha)\right)^{1/(\alpha-1)}\right] \\ &= \frac{D}{2}\log_2[2\pi\alpha] \;+\; \log_2\left[\frac{\Gamma\left(\frac{1}{1-\alpha}-\frac{D}{2}\right)}{(1-\alpha)^{D/2}\Gamma\left(\frac{1}{1-\alpha}\right)}\right] \;+\; \log_2\left[\left(1-\frac{D}{2\alpha}(1-\alpha)\right)^{\frac{D}{2}-\frac{1}{1-\alpha}}\right], \end{aligned} \tag{39}$$

(where $\mathfrak{b} = [2\alpha - D(1-\alpha)]^{-1}$), to write

$$H_\alpha(\mathbf{X}|\mathbf{Y}) \;=\; \frac{1}{2}\log_2|\Sigma(\mathbf{X}|\mathbf{Y})| \;+\; H_\alpha(\mathbf{Z}_\alpha^{\mathbf{1},D_\mathbf{X}+D_\mathbf{Y}}) \;-\; H_\alpha(\mathbf{Z}_\alpha^{\mathbf{1},D_\mathbf{Y}})\,. \tag{40}$$

At this stage, we note that

$$\begin{aligned} H_\alpha(\mathbf{Z}_\alpha^{\mathbf{1},D_\mathbf{X}+D_\mathbf{Y}}) &- H_\alpha(\mathbf{Z}_\alpha^{\mathbf{1},D_\mathbf{Y}}) - H_\alpha(\mathbf{Z}_\alpha^{\mathbf{1},D_\mathbf{X}}) \\ &= H_\alpha(\mathbf{Z}_\alpha^{\mathbf{1},D_\mathbf{X}}|\mathbf{Z}_\alpha^{\mathbf{1},D_\mathbf{Y}}) - H_\alpha(\mathbf{Z}_\alpha^{\mathbf{1},D_\mathbf{X}})\,, \end{aligned} \tag{41}$$

which is not zero as it was in the case of the Gaussian distribution. In fact, from the foregoing discussion, it is clear that for the α-Gaussian random variables, we can write the RTE in the form

$$\begin{aligned} T^R_{\alpha,Y\to X}(k,l) &= H_\alpha(X_{t_{n+1}}|X_n^{(k)}) - H_\alpha(X_{t_{n+1}}|X_n^{(k)},Y_n^{(l)}) \\ &= \frac{1}{2}\log_2\left(\frac{\Sigma(X_{t_{n+1}}|X_n^{(k)})}{\Sigma(X_{t_{n+1}}|X_n^{(k)},Y_n^{(l)})}\right) + H_\alpha(\mathbf{Z}_\alpha^{1,1}|\mathbf{Z}_\alpha^{\mathbf{1},k}) - H_\alpha(\mathbf{Z}_\alpha^{1,1}|\mathbf{Z}_\alpha^{\mathbf{1},k+l}) \\ &= \frac{1}{2}\mathcal{F}^{(k,l)}_{Y\to X} + I_\alpha(\mathbf{Z}_\alpha^{1,1}:\mathbf{Z}_\alpha^{\mathbf{1},l}|\mathbf{Z}_\alpha^{\mathbf{1},k})\,. \end{aligned} \tag{42}$$

Here, we have set $\mathbf{Z}_\alpha^{1,1}$ to correspond to the random variable $X_{t_{n+1}}$ with unit variance. Similarly, $\mathbf{Z}_\alpha^{\mathbf{1},k}$ and $\mathbf{Z}_\alpha^{\mathbf{1},l}$ correspond to unit covariance random variables $X_n^{(k)}$ and $Y_n^{(l)}$, respectively.

Clearly, when Y_t and X_t processes are independent (and, hence, *not causal* in the Granger sense), their joint distribution factorizes and, thus, $H_\alpha(\mathbf{Z}_\alpha^{\mathbf{1},D_\mathbf{X}+D_\mathbf{Y}}) \mapsto H_\alpha(\mathbf{Z}_\alpha^{\mathbf{1},D_\mathbf{X}} \times \mathbf{Z}_\alpha^{\mathbf{1},D_\mathbf{Y}})$. Additivity of the RE then ensures that $H_\alpha(\mathbf{Z}_\alpha^{1,1}|\mathbf{Z}_\alpha^{\mathbf{1},k}) = H_\alpha(\mathbf{Z}_\alpha^{1,1}|\mathbf{Z}_\alpha^{\mathbf{1},k+l})$ and, hence, $I_\alpha(\mathbf{Z}_\alpha^{1,1}:\mathbf{Z}_\alpha^{\mathbf{1},l}|\mathbf{Z}_\alpha^{\mathbf{1},k})$ is zero. In other words, when two processes are not Granger causal, their RTEs are zero. Actually, it is not difficult to see that this is true irrespective of a specific form of the distribution involved. However, the opposite is not true since $I_\alpha(\mathbf{Z}_\alpha^{1,1}:\mathbf{Z}_\alpha^{\mathbf{1},l}|\mathbf{Z}_\alpha^{\mathbf{1},k})$ might be (unlike in Shannon's case) negative; consequently, $T^R_{\alpha,Y\to X}(k,l)$ can be zero even if

$\mathcal{F}^{(k,l)}_{Y\to X}$ is not. To understand this point better, we explicitly evaluate $I_\alpha(\mathbf{Z}^{1,1}_\alpha : \mathbf{Z}^{\mathbf{1},l}_\alpha|\mathbf{Z}^{\mathbf{1},k}_\alpha)$ for our α-Gaussian random variables. Using (39), we can write

$$\begin{aligned} I_\alpha(\mathbf{Z}^{1,1}_\alpha : \mathbf{Z}^{\mathbf{1},l}_\alpha|\mathbf{Z}^{\mathbf{1},k}_\alpha) &= \log_2\left[\frac{\Gamma\left(\frac{1}{1-\alpha}-\frac{1+k}{2}\right)}{\Gamma\left(\frac{1}{1-\alpha}-\frac{k}{2}\right)}\frac{\Gamma\left(\frac{1}{1-\alpha}-\frac{k+l}{2}\right))}{\Gamma\left(\frac{1}{1-\alpha}-\frac{1+k+l}{2}\right)}\right] \\ &+ \log_2\left[\frac{\left(\frac{\alpha}{1-\alpha}-\frac{1+k}{2}\right)^{\frac{1+k}{2}-\frac{1}{1-\alpha}}}{\left(\frac{\alpha}{1-\alpha}-\frac{k}{2}\right)^{\frac{k}{2}-\frac{1}{1-\alpha}}}\frac{\left(\frac{\alpha}{1-\alpha}-\frac{k+l}{2}\right)^{\frac{k+l}{2}-\frac{1}{1-\alpha}}}{\left(\frac{\alpha}{1-\alpha}-\frac{1+k+l}{2}\right)^{\frac{1+k+l}{2}-\frac{1}{1-\alpha}}}\right]. \end{aligned} \tag{43}$$

By setting $\zeta = \frac{1}{1-\alpha}-\frac{k}{2}$ and $\xi = \frac{1}{1-\alpha}-\frac{k+l}{2}$, we can rewrite (43) as

$$\begin{aligned} I_\alpha(\mathbf{Z}^{1,1}_\alpha : \mathbf{Z}^{\mathbf{1},l}_\alpha|\mathbf{Z}^{\mathbf{1},k}_\alpha) &= \log_2\left[\frac{\Gamma\left(\zeta-\frac{1}{2}\right)}{\Gamma(\zeta)}\frac{(\zeta-1)^\zeta}{\left(\zeta-\frac{3}{2}\right)^{\zeta-\frac{1}{2}}}\frac{\Gamma(\xi)}{\Gamma\left(\xi-\frac{1}{2}\right)}\frac{\left(\xi-\frac{3}{2}\right)^{\xi-\frac{1}{2}}}{(\xi-1)^\xi}\right] \\ &= \log_2\left[\frac{\Gamma\left(\zeta-\frac{3}{2}\right)}{\Gamma(\zeta-1)}\frac{(\zeta-1)^{\zeta-1}}{\left(\zeta-\frac{3}{2}\right)^{\zeta-\frac{3}{2}}}\frac{\Gamma(\xi-1)}{\Gamma\left(\xi-\frac{3}{2}\right)}\frac{\left(\xi-\frac{3}{2}\right)^{\xi-\frac{3}{2}}}{(\xi-1)^{\xi-1}}\right] \\ &\leq -\frac{1}{2}\log_2\left[\frac{(\xi-1)}{\left(\xi-\frac{3}{2}\right)}\right] \leq 0, \end{aligned} \tag{44}$$

where on the last line we use the Kečkić–Vasić inequality [65]

$$\frac{(x+1)^{x+1}}{(x+s)^{x+s}}e^{s-1} \leq \frac{\Gamma(x+1)}{\Gamma(x+s)} \leq \frac{(x+1)^{x+\frac{1}{2}}}{(x+s)^{x+s-\frac{1}{2}}}e^{s-1}, \tag{45}$$

valid for $s \in (0,1)$. In addition, it can be numerically checked that $\frac{dI_\alpha(\mathbf{Z}^{1,1}_\alpha:\mathbf{Z}^{\mathbf{1},l}_\alpha|\mathbf{Z}^{\mathbf{1},k}_\alpha)}{d\alpha} > 0$, for all l,k from the definition, so the maximum of $I_\alpha(\mathbf{Z}^{1,1}_\alpha : \mathbf{Z}^{\mathbf{1},l}_\alpha|\mathbf{Z}^{\mathbf{1},k}_\alpha)$ is attained at $\alpha = 1$, see Figure 3. When α is close to 1, then one can employ the asymptotic relation $\Gamma[x+\gamma] \sim \Gamma[x]x^\gamma$ valid for $x \gg 1$, $\gamma \in \mathbb{C}$, and rewrite (39) in the form $(D/2)\log_2[2\pi\alpha e^\alpha]$. In this case, (43) tends to zero and we obtain equivalence between TE and the Granger causality. This result should not be so surprising because in the limit $\alpha \to 1$, RE tends to Shannon's entropy and the α-Gaussian distribution tends to the Gaussian distribution.

The leading order behavior near $\alpha = 1$ can be obtained directly from (43). The ensuing Taylor expansion gives

$$I_\alpha(\mathbf{Z}^{1,1}_\alpha : \mathbf{Z}^{\mathbf{1},l}_\alpha|\mathbf{Z}^{\mathbf{1},k}_\alpha) = -\frac{l(\alpha-1)^2}{8} + \mathcal{O}((\alpha-1)^3), \tag{46}$$

so, the point $\alpha = 1$ is a *stationary point* of $I_\alpha(\mathbf{Z}^{1,1}_\alpha : \mathbf{Z}^{\mathbf{1},l}_\alpha|\mathbf{Z}^{\mathbf{1},k}_\alpha)$. This closes the proof.

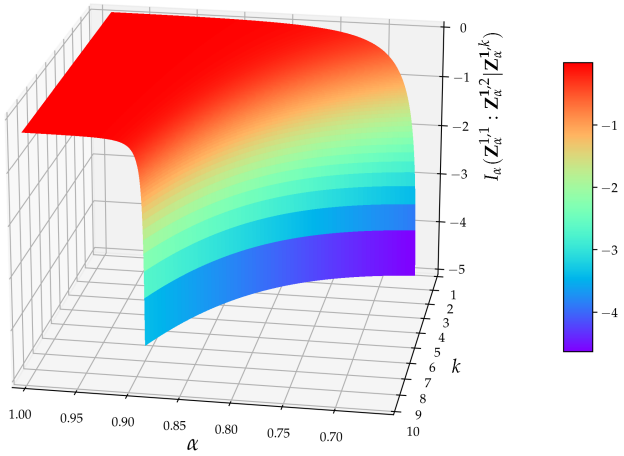

Figure 3. Example of $I_\alpha(\mathbf{Z}_\alpha^{1,1} : \mathbf{Z}_\alpha^{1,l}|\mathbf{Z}_\alpha^{1,k})$ for $l = 2$ and $k = 1, 2, \ldots, 10$. Range validity of α is thus between $\frac{3+k}{5+k}$ and 1.

4. Estimation of Rényi Entropy

4.1. RTE and Derived Concepts

From a data analysis point of view, it is not very practical to use the full joint processes $X_n^{(k)}$ and $Y_n^{(l)}$ (cf. the defining relation (10)) because (possibly) high values of k and l negatively influence the accuracy of estimation of RTE. In the following sections, we will thus switch to a more expedient definition of RTE given by

$$\begin{aligned} T^R_{\alpha,Y\to X}(\{k\},\{m\},\{l\}) &= H_\alpha(X_n^{\{m\},+}|X_n^{\{k\},-}) - H_\alpha(X_n^{\{m\},+}|X_n^{\{k\},-}, Y_n^{\{l\},-}) \\ &= I_\alpha(X_n^{\{m\},+} : Y_n^{\{l\},-}|X_n^{\{k\},-}), \end{aligned} \tag{47}$$

where $X_n^{\{k\},\Omega}$ is a subset of past ($\Omega = -$) or future ($\Omega = +$) values of X_{t_n} with the number of elements equal to k, such that $\{k\} = \{\kappa_1, ..., \kappa_k\}$ is a set of indices and $X_n^{\{k\},\Omega} \equiv X_{t_{n\Omega\kappa_1}}, X_{t_{n\Omega\kappa_2}}, \ldots, X_{t_{n\Omega\kappa_k}}$ is a selected subsequence of X_{t_n}, i.e., n_X-dimensional vectors. The same notational convention applies to $Y_n^{\{l\},\Omega}$ as a subsequence of Y_{t_n}, i.e., n_Y-dimensional vectors. In definition (47), we added a third parameter, m—the so-called *future step*. Though such a parametrization is often used in the literature on Shannon's TE, cf., e.g., reference [17], we will (in the following) only employ $m = \{1\}$ so as to conform with the definition (10). In such a case, we will often omit the middle index in $T^R_{\alpha,Y\to X}(\{k\},\{1\},\{l\})$.

4.1.1. Balance of Transfer Entropy

In order to compare RTE that flows in the direction from $Y \to X$ with the RTE that flows in the opposite direction $X \to Y$, we define the *balance of transfer entropy*

$$T^{R,\,\text{balance}}_{\alpha,Y\to X}(\{k\},\{l\}) = T^R_{\alpha,Y\to X}(\{k\},\{l\}) - T^R_{\alpha,X\to Y}(\{k\},\{l\}). \tag{48}$$

4.1.2. Effective Transfer Entropy

To mitigate the finite size effects, we employ the idea of a surrogate time series. To this end, we define the *effective transfer entropy*

$$T^{R,\,\text{effective}}_{\alpha,Y\to X}(\{k\},\{l\}) = T^R_{\alpha,Y\to X}(\{k\},\{l\}) - T^R_{\alpha,Y^{(\text{sur})}\to X}(\{k\},\{l\}), \tag{49}$$

where $Y^{(\text{sur})}$ stands for the randomized (reordered) time series—the surrogate data sequence. Such a series has the same mean, the same variance, the same autocorrelation function and, therefore, the same power spectrum as the original sequence, but (nonlinear) phase relations are destroyed. In effect, all the potential correlations between $X_n^{\{k\}}$ and $Y_n^{\{l\}}$

are removed, which means that $T^R_{\alpha,Y^{(\mathrm{sur})}\to X}(\{k\},\{l\})$ should be zero. In practice, this is not the case, despite the fact that there are no obvious structures in the data. The non-zero value of $T^R_{\alpha,Y^{(\mathrm{sur})}\to X}(\{k\},\{l\})$ must then be a byproduct of the finite data set. Definition (49) then ensures that spurious effects caused by finite k and l are removed. In our computations, we used the Fisher–Yates algorithm [66] together with Mersenne twister random generation algorithm [67] for the randomized surrogates. For a more technical exposition, see, e.g., refs. [68–70].

4.1.3. Balance of Effective Transfer Entropy

Finally, we combined both previous definitions to form the *balance effective transfer entropy*

$$\begin{aligned} T^{R,\,\mathrm{balance,\,effective}}_{\alpha,Y\to X}(\{k\},\{l\}) &= T^{R,\,\mathrm{effective}}_{\alpha,Y\to X}(\{k\},\{l\}) - T^{R,\,\mathrm{effective}}_{\alpha,X\to Y}(\{k\},\{l\}) \\ &= T^R_{\alpha,Y\to X}(\{k\},\{l\}) - T^R_{\alpha,Y^{(\mathrm{sur})}\to X}(\{k\},\{l\}) \\ &- T^R_{\alpha,X\to Y}(\{k\},\{l\}) + T^R_{\alpha,X^{(\mathrm{sur})}\to Y}(\{k\},\{l\}), \end{aligned} \tag{50}$$

to quantify the direction of flow of transfer entropy without finite size effects.

4.1.4. Choice of Parameters k and l

The choice of the parameters k and l is essential to reliably analyze the information transfer between variables in a system. So, a natural question arises as to how one should choose such parameters.

The order of k and l, both in the RTE and Shannon's TE, but also in approximating autoregression in the Granger case, is often (in practice) set rather arbitrarily at some moderately high number. In the literature, there are theoretical criteria for optimal choices of k and l—with no unique answer. In our numerical simulations, we employed two pragmatic criteria: (a) results should be stable under the increase of k and l and, additionally, (b) k, and l should be equal to—or higher than—those used in the literature for the analysis of Shannon's TE in Rössler systems, e.g., references [18,22], so that we could make a comparison with the existence results. The chosen values $(\{k\},\{l\}) \equiv (\{k\},\{1\},\{l\}) = (\{0,1\},\{1\},\{0\})$ often well-satisfied both aforementioned conditions. In Section 6.3, it was sufficient to set $\{k\} = \{0\}$ and $\{l\} = \{0\}$, in agreement with [18]. When a need has arisen to emphasize some finer details in the behavior of the RTE (cf. Figures 6 and 10), $\{k\}$ was chosen to be $\{0,1,2,3,4\}$ or even $\{0,1,2,3,4,5,6\}$.

5. Rössler System

5.1. Equations for Master System

In order to illustrate the use of RTE, we considered two unidirectionally coupled Rössler systems (oscillators). These often serve as testbeds for various measures of synchronization, including Shannon's TE [71–73]. Rössler's system is described by three non-linearly coupled partial differential equations

$$\begin{aligned} \dot{x}_1 &= -\omega_1 x_2 - x_3\,, \\ \dot{x}_2 &= \omega_1 x_1 + a x_2\,, \\ \dot{x}_3 &= b + x_3(x_1 - c)\,, \end{aligned} \tag{51}$$

with four coefficients ω_1, a, b, and c. Strictly speaking, only three coefficients are independent, as ω_1 can be set to one by appropriately rescaling x_2. RS was invented in 1976 by O.E. Rössler [32] and it likely represents the most elementary geometric construction of chaos in the continuous systems. In fact, since the Poincaré–Bendixson theorem precludes the existence of (other than) steady, periodic, or quasi-periodic attractors in autonomous systems, defined in one- or two-dimensional manifolds, the minimal dimension for chaos

is three [74]. The simplicity of the RS is bolstered by the fact that it only has one nonlinear (quadratic) coupling.

RS classifies as the *continuous (deterministic) chaotic system*, and more specifically as the *chaotic attractor*. The word "attractor" refers to the fact that whatever is the initial condition for the solution of the differential Equation (52), the trajectory $\mathbf{x}(t)$ ends up (after a short transient period) at the same geometrical structure (see Figure 5), which is neither a fixed point nor a limit cycle. This attractive geometrical structure is known as the Rössler attractor.

For future convenience, we will call the RS (51) as *driving* or *master system* and denote it as $\{X\}$.

5.2. Equations for the Slave System

In the following, we investigate RTE between two Rössler systems that are unidirectionally coupled in the variable x_1 via a small adjustable parameter ε. The corresponding second RS—*driven* or *slave system*, is defined as

$$\begin{aligned} \dot{y}_1 &= -\omega_2 y_2 - y_3 + \varepsilon(x_1 - y_2)\,, \\ \dot{y}_2 &= \omega_2 y_1 + a y_2\,, \\ \dot{y}_3 &= b + y_3(y_1 - c)\,. \end{aligned} \tag{52}$$

Here, we fix the coefficients so that $a = 0.15$, $b = 0.2$, $c = 10.0$, and frequencies $\omega_1 = 1.015$ and $\omega_2 = 0.985$, and initial conditions $(x_1(0), x_2(0), x_3(0)) = (0, 0, 0)$ and $(y_1(0), y_2(0), y_3(0)) = (0, 0, 1)$. This parametrization is adopted from reference [18] where Shannon's TE between systems (51) and (52) was studied. In the following, we will denote the slave system also as $\{Y\}$.

5.3. Numerical Experiments with Coupled RSs

Before we embark on the RTE analysis, let us first take a look at the phenomenology of the coupled RSs (51) and (52) by means of simple numerical experiments. In our numerical treatment, we simulate coupled RSs by using the integration method, which is implemented in a package SciPy named `solve_ivp` with the LSODA option that exploits the Addams/BDF method, see, e.g., reference [75]. Projections of the ε-dependent RSs dynamics to various planes are presented in Figure 5. For visualization purposes, we used the toolkit Matplotlib [76] that exploits toolkit NumPy [77]. The sources are part of the Pyclits project [78]. In the future, the work can be rebased. The resulting data set analyzed consisted of 100,000 data points. To gain insight into the transient region, we chose shorter time lags in the data set generated from RS with $0.1 \leq \varepsilon \leq 0.15$, namely, we reduced the time steps from 0.01 to 0.001. In parallel, we display in Figure 4 the behaviors of the corresponding Lyapunov exponents, as adapted from [22], which help to elucidate our discussion.

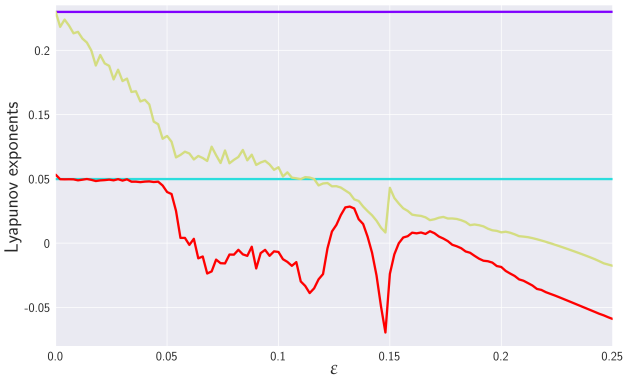

Figure 4. The two largest Lyapunov exponents of the master system (constant—violet and green) and the slave system (decreasing—red and yellow). So, for small ε, the signature of LE is $++\,00\,--$, while after synchronization, we end up with the signature $+0\,----$. After synchronization, there is a "collaps" of the dimension, in the sense that the slave system is completely dependent on the master system, so that there is only one dimension (direction) in which there is an expansion. Accordingly, there is only one LE with a positive sign. The LEs are measured in nats per time unit.

Projections

Instead of a conventional stereoscopic plotting, we found it more convenient (and illuminating) to focus on various plane projections of the coupled RSs. First, we noticed that, in Figure 5, the projections of RSs on the x_2- x_1, x_3-x_2, and x_1-x_3 planes do not depend on the coupling between systems (i.e., they are ε-independent), as expected, because the slave system (52) does not influence dynamics of the master system (51), which is autonomous (irrespective of ε). However, it is clear that signatures of the interaction between non-symmetrically coupled RSs (51) and (52) will show up in projections on the x_i-y_j and y_i-y_j planes.

Secondly, when the RSs are not coupled (i.e., when $\varepsilon = 0$), we have two autonomous RSs—in fact, two strange attractors that differ only by values of their frequency coefficients and initial values. The autonomies of the respective RSs are clearly seen in projections on the x_i-x_j and y_i-y_j planes (cf. Figure 5). A different density of trajectories (in a given time window t = 100,000) can be ascribed to the frequency mismatch. Projections on the x_1-y_1 and x_2-y_2 planes show how the ensuing chaotic and (component-wise) uncorrelated trajectories fill their support regions. In particular, we can observe that on the background of densely packed chaotic trajectories, clear vertical stripes of dominantly-visited regions appear in the slave system. Vertical stripes are clearly visible because limit cycles in the autonomous slave system are far more localized than in the master system. The projection on the x_3-y_3 plane indicates that (most of the time) the master system orbits venture to the x_3 direction, the slave system orbits are in the vicinity of the y_1-y_2 plane, and vice versa.

By continuously increasing the coupling strength ε from the zero value, we can observe that, already, a small interaction significantly changes the evolution of the slave system. For instance, in Figure 5, we see that when $\varepsilon = 0.01$, then the diffusive term $\varepsilon(x_1 - y_2)$ significantly disperses the limit cycles in the slave system. This is reflected not only in all projections on the y_i-y_j planes but also in projections on the x_1-y_1 and x_2-y_2 planes. In the latter two cases, the diffusion causes that horizontal stripes to completely disappear. Finally, the projection on the x_3-y_3 plane does not change significantly from the $\varepsilon = 0$ case.

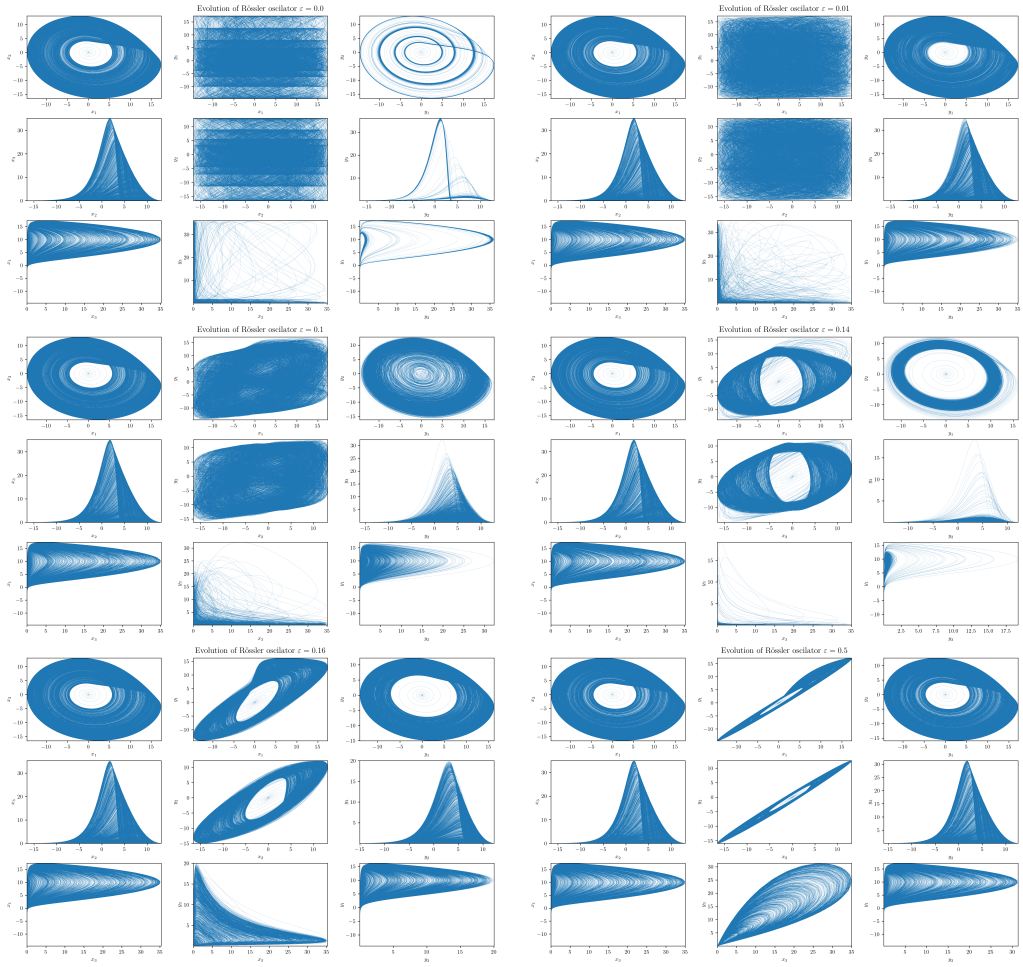

Figure 5. Projections of the RSs (51) and (52) on various planes. For each fixed ε, we depict nine figures that correspond (from top to bottom and left to right) to projections on the x_2-x_1, x_3-x_2, x_1-x_3, x_1-y_1, x_2-y_2, x_3-y_3, y_2-y_1, y_3-y_2, and y_1-y_3 planes. In the figure, we display, altogether, nine values of ε corresponding (from left to right and top to bottom) to $\varepsilon = 0, 0.01, 0.1, 0.14, 0.16$ and 0.5. The initial values are chosen as $x_1(0), x_2(0), x_3(0) = 0$, $y_1(0), y_2(0) = 0$, and $y_3(0) = 1$. Further projections for the transient region $0.12 \lesssim \varepsilon \lesssim 0.15$ are shown in Figure 8. All RSs are depicted in the time window $t = 10{,}000$.

When we further increase ε, we see that the behavior of the slave system starts to qualitatively depart from that of the master system. For ε, around 0.1, the slave system orbit diffuses to the region around the origin that is basically not visited (apart from an initial transient orbit) by the master system orbit (cf. projections on the y_i-y_j planes). In addition, projections on the x_1-y_1 and x_2-y_2 planes disclose that the ensuing support areas are not filled anymore. In fact, we can see a development of a slant stripe structure. On the other hand, the projection on the y_3-x_3 plane reveals that the slave system orbits stop visiting regions further from $y_3 = 0$. A yet higher ε (around 0.14) orbit of the system $\{Y\}$ first converges to a single limit cycle before it makes (again) a transition into a chaotic regime. Finally, we can observe that at $\varepsilon \sim 0.14$, the slave system rarely deviates far from $y_3 = 0$ and spends most of its time in the close vicinity of the y_1-y_2 plane—its evolution is "flattened".

Moreover, at $\varepsilon \sim 0.14$, we can also notice that projections on the y_1-x_1 and y_2-x_2 planes underwent a change in topology (in fact, this happened already at around $\varepsilon \sim 0.12$). The onset of this "topological phase transition" is closely correlated with the behavior of the largest Lyapunov exponent (LE) of the slave system. In fact, coupled RSs altogether have six Lyapunov exponents. The $\varepsilon = 0$ one has two autonomous RSs each with three LEs—one positive, one zero, and one negative (signature $+0-$ is a typical hallmark of a strange

attractor in three dimensions). While at $\varepsilon = 0$, the signature of LEs is $++\,00\,--$, increasing ε all three LEs associated with $\{Y\}$ decreasing (initially) monotonically, cf. Figure 4. After a transient negativity and a return to zero (red curve in Figure 4), the originally positive LE of the slave system monotonically decreases and the negative for $\varepsilon \gtrsim 0.15$. In particular, we see that the critical value $\varepsilon \sim 0.12$ at which the "topological phase transition" occurs coincides with the value at which the largest LE of the system $\{Y\}$ crosses zero.

What is particularly noteworthy is an abrupt (non-analytic) change in the behavior of LEs at the value $\varepsilon \sim 0.145$. At this value, the LE changes direction and starts to increase with increasing ε. The increase stops at $\varepsilon \sim 0.15$ when the yellow-colored LE in Figure 4 reaches (approximately) value zero, after which it monotonically decreases. Such a decrease also starts for the second red-colored LE, but at a slightly different value of ε.

For stronger interactions with $0.15 \lesssim \varepsilon \lesssim 0.2$, we see (cf. Figure 5 with $\varepsilon = 0.16$) that the slave system starts to approach the structure of the master system strange attractor (cf. x_i-x_j and y_i-y_j projections). From the tilt and thinning of projections on the x_1-y_1 and x_2-y_2 planes, one may deduce that the amplitude synchronizations in the x_1 and y_1 (as well as x_2 and y_2) directions increase. Projection on the x_3-y_3 plane shows that amplitudes in the x_3 and y_3 directions are also synchronized (being roughly a half-cycle behind each other).

Finally, for very strong interactions, e.g., for $\varepsilon \sim 0.5$, the synchronization is almost complete: the system $\{Y\}$ basically fully emulates the master system's behavior with both systems now being structurally identical (cf. x_i-x_j and y_i-y_j projections). Full synchronization is nicely seen in projections on the x_1-y_1 and x_2-y_2 planes. Note that the amplitudes in the x_3 and y_3 directions start to synchronize.

6. Numerical Analysis of RTE for Coupled RSs

In the previous section, we learned some essentials about the coupled RS (51) and (52). In order to demonstrate the inner workings of the RTE and to gain further insight into how the two RSs approach synchronization, we compute here the RTE for various salient situations, such as the RTE between the x_1- and y_1-component, between the x_1- and y_3-component, or RTE between the full master and slave system. In our numerical analysis, we employed the RE estimator introduced by Leonenko et al. [26]. Some fundamentals associated with this estimator are relegated to Appendix A.

6.1. Effective RTE between x_1 and y_1 Directions

In order to understand the dynamics of the two coupled nonlinear dynamical systems (51) and (52) on their routes to synchronization, we first analyzed the effective RTE between the x_1 and y_1 components. Corresponding plots for different coupling strengths ε and different orders α are depicted in Figure 6. We can observe first that the effective RTE from x_1 to y_1 gradually increases with the increasing coupling strength until $\varepsilon \sim 0.12$. The regime between $\varepsilon \sim 0.12$ and $\varepsilon \sim 0.15$, as seen from Figure 5, corresponds to a transient synchronization behavior, which stabilizes only after $\varepsilon \sim 0.15$. This can also be seen from the behavior of the LEs at Figure 4. It should also be noted that the behavior of effective RTEs in the transient regime is apparently almost identical for all α in both $T^{R,\,\text{effective}}_{\alpha,x_1\to y_1}(\{0,1\},\{1\},\{0\})$ and $T^{R,\,\text{effective}}_{\alpha,y_1\to x_1}(\{0,1\},\{1\},\{0\})$. This would, in turn, indicate that the information transfer is the same across all sectors of the underlying probability distributions. Upon closer inspection though, such a highly correlated behavior will disappear when more historic data on $\{X\}$ and $\{Y\}$ are included (cf. $T^{R,\,\text{effective}}_{\alpha,x_1\to y_1}(\{0,1,2,3,4,5,6\},\{1\},\{0\})$ and $T^{R,\,\text{effective}}_{\alpha,y_1\to x_1}(\{0,1,2,3,4,5,6\},\{1\},\{0\})$ in Figure 6).

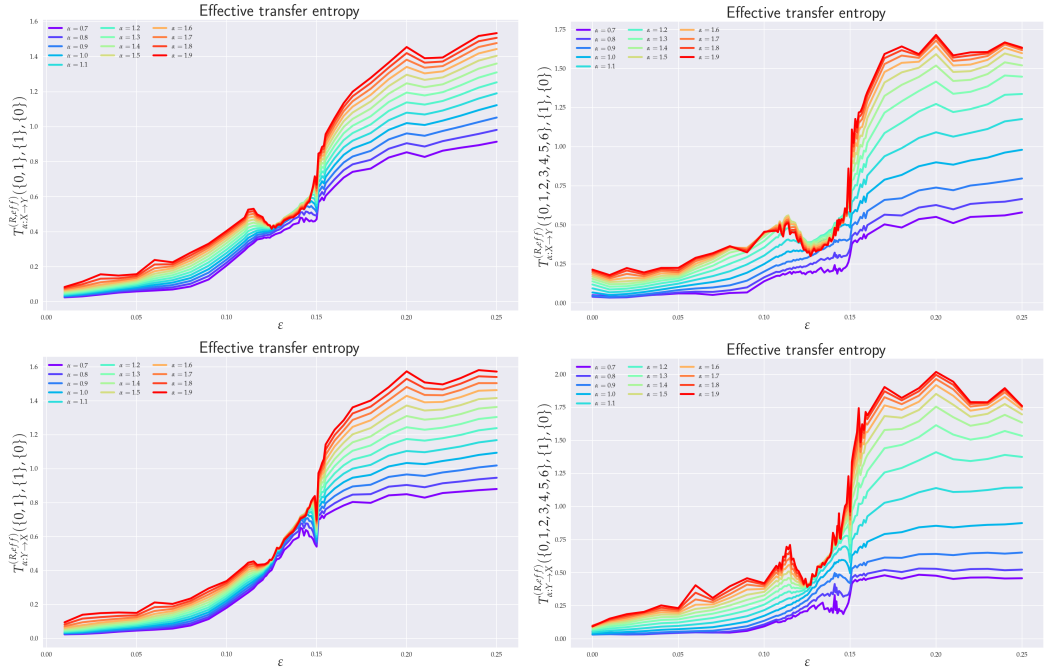

Figure 6. Effective RTE between x_1 and y_1 for two different histories of x_1, i.e., $T^{R,\,\text{effective}}_{\alpha,x_1\to y_1}(\{0,1\},\{1\},\{0\})$, $T^{R,\,\text{effective}}_{\alpha,x_1\to y_1}(\{0,1,2,3,4,5,6\},\{1\},\{0\})$, $T^{R,\,\text{effective}}_{\alpha,y_1\to x_1}(\{0,1\},\{1\},\{0\})$, $T^{R,\,\text{effective}}_{\alpha,y_1\to x_1}(\{0,1,2,3,4,5,6\},\{1\},\{0\})$, respectively, from left to right and top to bottom. RTE is measured in nats.

The same conclusion can be reached when the effective RTEs for the full six-dimensional systems are considered, cf. Figure 7.

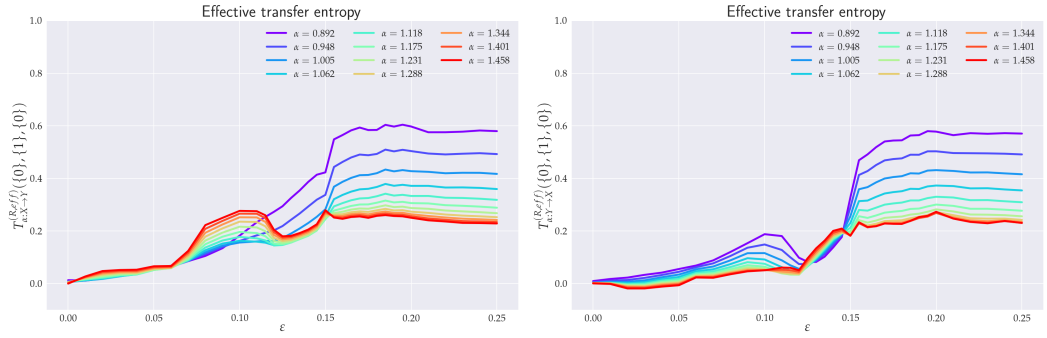

Figure 7. Effective transfer entropy for the full system ($n_X = 3$ and $n_Y = 3$) and for different values of α as functions of the coupling ε. We depict $T^{R,\,\text{effective}}_{\alpha,X\to Y}(\{0\},\{1\},\{0\})$ (**left**) and $T^{R,\,\text{effective}}_{\alpha,Y\to X}(\{0\},\{1\},\{0\})$ (**right**). RTE is measured in nats.

Nevertheless, from Figure 6, it can clearly be inferred that—in the transient region—strong correlations do exist, albeit not for all αs. In particular, one starts with the correlated flow for $\alpha \gtrsim 1.2$, which becomes stronger as ε increases. On the other hand, as ε approaches 0.15, the information flow decreases for $\alpha \lesssim 1$. This can be seen clearly in both Figures 6 and 7. At $\varepsilon = 0.15$, the information flow abruptly increases for all αs. This is similar to a first order phase transition in statistical physics. In this respect, our "topological phase transition" would be more similar to a second order phase transition due to a smooth change in the entropic flow across the critical point $\varepsilon = 0.12$. This scenario is also supported by Figure 8, where the actual behavior of the RS between the two critical points for four selected values of ε's is depicted. Note, in particular, how the increase in the RTE for $\alpha \gtrsim 1.2$ (as well as the decrease of RTE for $\alpha \lesssim 1$) are reflected in the contractions (measure concentrations) of the regions with denser orbit populations in the slave system. This, in

turn, reinforces the picture that RTEs with higher αs describe the transfer of information between more central parts of underlying distributions, which, in this case, relate to higher occupation densities of the $\{Y\}$ system orbit. From Figure 8, we can also note that, at the critical point $\varepsilon = 0.15$, the contracted orbit regions abruptly expand and the slave system starts its way toward full synchronization with the master system. This is again compatible with the fact that the RTE abruptly increases for all αs at this point—i.e., all parts of underlying distributions participate in this transition and, consequently, the occupation density of the $\{Y\}$ system orbit spreads. In this respect, point $\varepsilon = 0.15$ represents the *threshold to full synchronization* while point $\varepsilon = 0.12$ denotes the *threshold to transient behavior prior to full synchronization*. The latter can be identified with a phase synchronization threshold, which should be at (or very close to) this point [22].

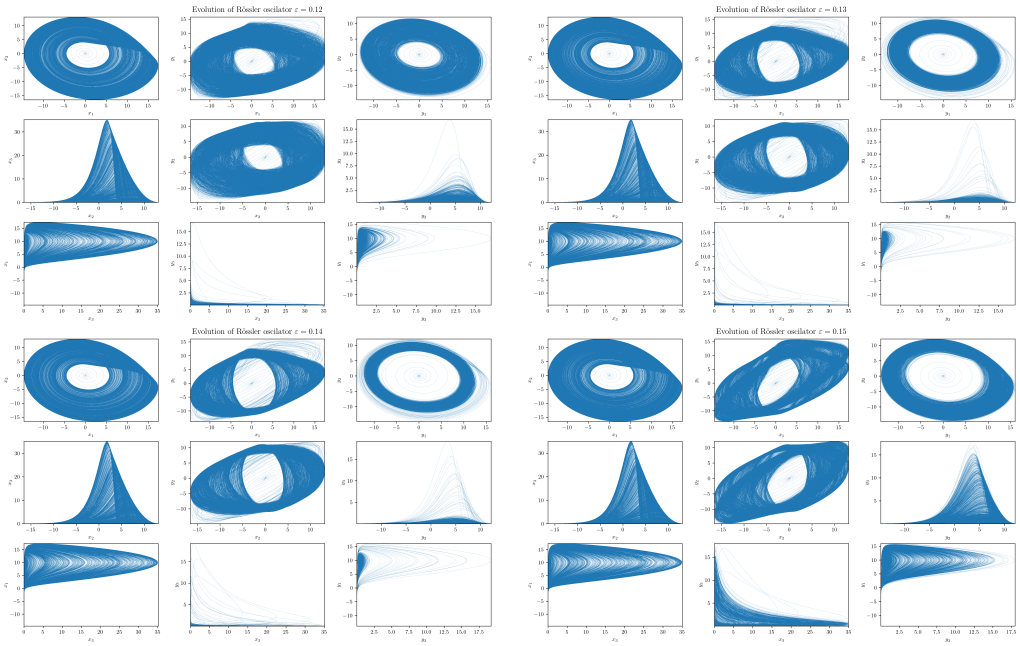

Figure 8. Four projections of the RSs (51) and (52) in the transient region $0.12 \lesssim \varepsilon \lesssim 0.15$. Depicted are projections (from left to right, from top to bottom) with $\varepsilon = 0.12$, 0.13, 0.14, and 0.15. With increasing ε, one can observe the contractions (measure concentrations) of the regions with denser orbit populations in the slave system. At the critical point $\varepsilon = 0.15$, the contracted orbit regions abruptly expand and the slave system starts its way toward full synchronization with the master system (cf. also Figure 5). All RSs are depicted in the time window $t = 10{,}000$.

After the critical point $\varepsilon \sim 0.15$, both RSs enter full synchronization. In fact, the full synchronization starts when the information flow from all sectors of underlying distributions (i.e., for all αs) starts to be (almost) ε-independent and when $T^{R,\,\text{balance, effective}}_{\alpha,X\to Y}$ approach zero—so there is a one-to-one relation between the states of the systems, and the time series of the $\{X\}$ system can be predicted from the time series $\{Y\}$ system, and vice versa. Indeed, from Figure 6 (cf. also Figures 7 and 9), we see that all $T^{R,\,\text{effective}}_{\alpha,Y\to X}$ proceed in a slow increase toward their asymptotic values in the fully-synchronized state.

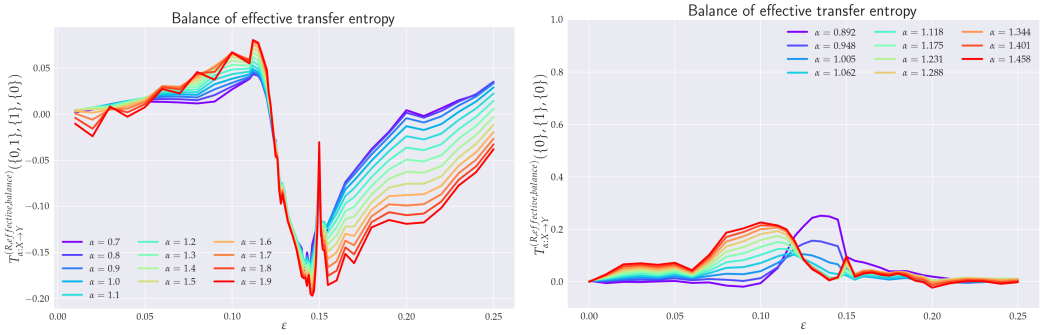

Figure 9. Balance of effective RTEs from x_1 to y_1 $T^{R,\,\text{balance, effective}}_{\alpha,x_1\to y_1}(\{0,1\},\{1\},\{0\})$ (**left**, $n_{x_1}=1$ and $n_{y_1}=1$) and the balance of effective RTEs for the full system $T^{R,\,\text{balance, effective}}_{\alpha,X\to Y}(\{0\},\{1\},\{0\})$ (**right**, $n_X=3$ and $n_Y=1$) with Y being y_1.

6.2. Effective RTE between x_3 and y_3 Directions

As already seen from Figures 5 and 8, projections in the x_3-y_3 plane are particularly distinct. In Figure 10, we see the ensuing effective RTE between x_3 and y_3 directions.

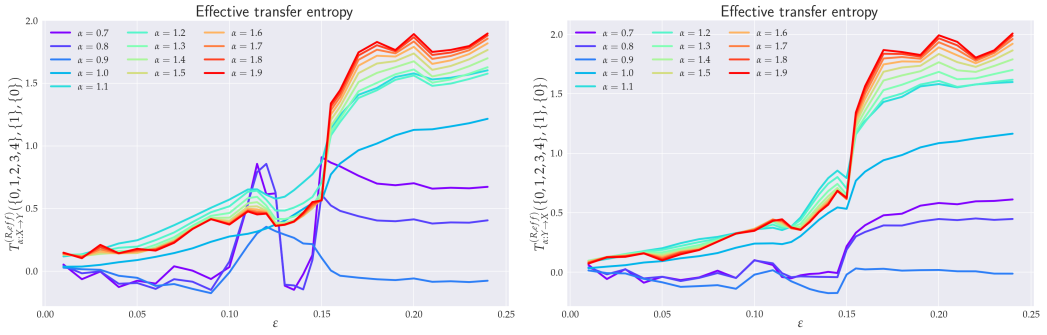

Figure 10. Effective RTE between x_3 and y_3 directions. From left to right: $T^{R,\text{effective}}_{\alpha,x_3\to y_3}(\{0,1,2,3,4\},\{1\},\{0\})$ and $T^{R,\text{effective}}_{\alpha,y_3\to x_3}(\{0,1,2,3,4\},\{1\},\{0\})$. Note a sudden increase in entropy transfer from the master to slave system at $\varepsilon=0.12$ (i.e., threshold to transient behavior) for $\alpha<1$. RTE is measured in nats.

What is particularly noticeable is a sudden increase in entropy transfer from the master to slave system at $\varepsilon=0.12$ (i.e., at the threshold to transient behavior) for $\alpha<1$. No comparable increase is observed from slave to master. This, might be explained as an influx of information needed to organize the chaotically correlated regime that exists prior the (correlated) transient regime (cf. x_i–y_i projections in Figures 5 and 8). It should also be noticed that ordinary Shannonian TE ($\alpha=1$) is completely blind to such an information transfer.

As for the transient region, we can observe that the effective RTE has qualitatively very similar behavior to the effective RTE between x_1 and y_1, namely a distinct decrease in the information transfer for $\alpha<1$ and an increase for $\alpha>1$. This again reveals a measure concentration. In this case, the orbit occupation density concentrates around the y_1-y_2 plane of the slave systems, cf. projections depicted in Figure 8. The situation abruptly changes at the synchronization threshold $\varepsilon=0.15$ after which the effective RTE approaches for each α a fixed asymptotic value that turns out to be the same for both $T^{R,\text{effective}}_{\alpha,x_3\to y_3}$ and $T^{R,\text{effective}}_{\alpha,y_3\to x_3}$.

6.3. Effective RTE for the Full System

In general, for a reliable inference, it is desirable that the conditioning variable in the definition or RTE (10) contains all relevant information about future values of the system or processes generating this variable in the uncoupled case. So, it should be a full three-dimensional vector X or Y in the case of RS. To this end, we display in Figure 7 the effective RTE for the full six-dimensional RS with information transfers in both $X\to Y$

and $Y \to X$ directions. Corresponding plots are depicted for different coupling strengths ε, different order αs, and different memories.

In particular, we can see that the information flow in the transient region starts after a brief decrease at around $\varepsilon \sim 0.12$ and sharply increases (in both directions) for $\alpha \gtrsim 1.2$. This implies that there is an increase in the correlating activity in between regions with higher occupation densities in both REs. The behavior depicted in Figure 8 can help us to better understand this situation. In particular, we see that in the transient region the $\{Y\}$ system reshapes its orbit occupation density so that the ensuing measure concentrates more around its peak while its tail parts are thinner. In fact, Figure 8 also shows that this measure concentration increases until almost $\varepsilon \sim 0.15$. The measure concentration behavior is reflected by the decrease of the RTE for $\alpha \lesssim 1$, i.e., decreasing information transfers between tail parts. This situation is even more pronounced when more memory is included in the effective RTEs, cf. both right pictures in Figure 7.

At the synchronization threshold $\varepsilon = 0.15$, the information flow abruptly changes for all αs, with a particularly strong increase for $\alpha \lesssim 1$. This indicates that the orbit occupation density of the $\{Y\}$ system abruptly reshapes by lowering the measure concentrated around its peak and broadening it in tails, so that the tail parts may also enter the full synchronization regime.

Let us finally comment on the issue of bidirectional information flown for single-component RTEs. By envisioning the discretized versions of RSs, (51) and (52), one can see that RTE from the slave to the master system (e.g., between the x_3 and y_3 direction) cannot easily be zero. This is because $H_\alpha(X_{3,t_{n+1}}|X_{3,n}^{(k)}, Y_{3,n}^{(l)})$ in the relation (10) is not simply $H_\alpha(X_{3,t_{n+1}}|X_{3,n}^{(k)})$. Note that due to the nonlinear nature of the coupled RSs, $y_3(t_n)$ depends both on $y_1(t_n)$ and $y_1(t_{n-1})$ (via the third equation in (52)), while $y_1(t_n)$ depends on $x_1(t_n)$ and $x_1(t_{n-1})$ (via the first equation in (52)); finally, $x_1(t_n)$ depends on $x_3(t_n)$ and $x_3(t_{n-1})$ and also $x_2(t_n)$ and $x_2(t_{n-1})$ (via the first equation in (51)); hence, $y_3(t_n)$ depends not only on $x_3(t_n)$, $x_3(t_{n-1})$, $x_3(t_{n-2})$ and $x_3(t_{n-3})$ but also on historical values of x_2. In this way, $H_\alpha(X_{3,t_{n+1}}|X_{3,n}^{(k)}, Y_{3,n}^{(l)}(\mathbf{X}))$ is not simply $H_\alpha(X_{3,t_{n+1}}|X_{3,n}^{(k)})$, as other components beyond $X_{3,n}$ are also needed. Consequently, when single-component RTEs for RS are computed, we inevitably find a non-zero information transfer from the slave to the master system. The latter is not so much a problem of k and l but rather the fact that we did not account for all relevant components (we simply missed some information).

It is true that for a reliable inference, in general, it would be desirable to obtain a zero value in the uncoupled direction $Y \to X$. This should be attained by proper conditioning—the conditioning variable should contain full information about future values of the system or processes generating this variable in the uncoupled case. So, it should be a three-dimensional vector $\mathbf{X}$ or $\mathbf{Y}$ for RS. Here, we computed effective RTE for the full six-dimensional system (vectors $\mathbf{X}$ and $\mathbf{Y}$). From Figure 7, we can see that $T_{\alpha,Y\to X}^{R,\,\text{effective}}$ in the uncoupled direction stays at the zero value (particularly for larger values of α) up to close to the synchronization threshold ($\varepsilon = 0.12$), while $T_{\alpha,X\to Y}^{R,\,\text{effective}}$ is distinctly positive there. So, RTE is a good *causal measure* only if the conditioning has a sufficient dimension (in our case, 3); otherwise, it can be viewed only as a measure of dependence.

6.4. Balance of Effective RTE

In order to quantify the difference between coupled ($X \to Y$) and uncoupled ($Y \to X$) information flow directions, we depict in Figure 9 the balance of effective RTEs between $T_{\alpha,X\to Y}^{R,\,\text{effective}}$ and $T_{\alpha,Y\to X}^{R,\,\text{effective}}$ for two different situations. Let us first concentrate on the balance of effective RTE $T_{\alpha,x_1\to y_1}^{R,\,\text{balance, effective}}(\{0,1\},\{1\},\{0\})$. There, we can clearly see that before the synchronization threshold ("topological phase transition"), i.e., for $\varepsilon \lesssim 0.12$, we have $T_{\alpha,x_1\to y_1}^{R,\,\text{effective}} > T_{\alpha,y_1\to x_1}^{R,\,\text{effective}}$, which indicates the correct direction of coupling. The fact that for $\alpha > 1.6$ and $\varepsilon \lesssim 0.04$ one has $T_{\alpha,x_1\to y_1}^{R,\,\text{balance, effective}}(\{0,1\},\{1\},\{0\}) < 0$ can be attributed to smaller reliability of the estimator in this region, cf. Figure 11 for estimation of ensuing the standard deviations. We can also observe that the synchronization threshold

$T^{R,\,\text{balance, effective}}_{\alpha,x_1\to y_1}(\{0,1\},\{1\},\{0\})$ changes sign and slowly return back to positive values in the fully synchronized regime. Similar behavior was reported in [22] for Shannon's TE. Moreover, in this transient region, the effective RTEs have the same values irrespective of α, or, in other words, information transfer is the same across all sectors of the underlying probability distributions. This is akin to the behavior, which, in statistical physics, is typically associated with phase transitions—except for the fact that now we have a critical line rather than a critical point. However, as we already mentioned in the previous two paragraphs, this degeneracy is only spurious and will be removed by considering either the effective RTE for the full (six-dimensional) RS or longer memory.

After $\varepsilon \sim 0.15$, the approach to full synchronization proceeds at slightly different rates for different αs. This can equivalently be restated as saying that different parts of the underlying distributions enter synchronization differently. The dependence of the balance of effective RTE for the full (six-dimensional) system is shown on the right in Figure 9. Here, the behavior is less reliable for larger values of α ($\alpha \gtrsim 1.2$) and for smaller αs ($\alpha \lesssim 0.8$), cf. Figure 11. In the region of reliable αs, the behavior is qualitatively similar to that of $T^{R,\,\text{balance, effective}}_{\alpha,x_1\to y_1}(\{0,1\},\{1\},\{0\})$. On the other hand, apart from the region of a transient synchronization, we clearly have $T^{R,\,\text{effective}}_{\alpha,X\to Y} > T^{R,\,\text{effective}}_{\alpha,Y\to X}$, which implies the correct direction of coupling. The approach to full synchronization is also easily recognized—the RTEs saturate to constant values (i.e., information transfer is ε-independent) and both $T^{R,\,\text{effective}}_{\alpha,X\to Y}$ and $T^{R,\,\text{effective}}_{\alpha,Y\to X}$ start to approach each other. In this respect, RTEs with lower αs enter the synchronization regime slower than RTEs with larger αs. In other words, events described by the tail parts of the distributions $p(x_{n+1}|x_n^{(k)})$ and $p(x_{n+1}|x_n^{(k)},y_n^{(l)})$ (corresponding to $\alpha < 1$) will fully synchronize at higher values of ε than corresponding events described by central parts ($\alpha > 1$).

In passing, we might notice that since both $T^{R,\,\text{effective}}_{\alpha,X\to Y}$ and $T^{R,\,\text{effective}}_{\alpha,Y\to X}$ approach each other in the fully synchronized state, both the $\{X\}$ and $\{Y\}$ systems have to have the same underlying distributions (due to the reconstruction theorem for REs [21,34]) and, hence, they are indistinguishable, as one would expect.

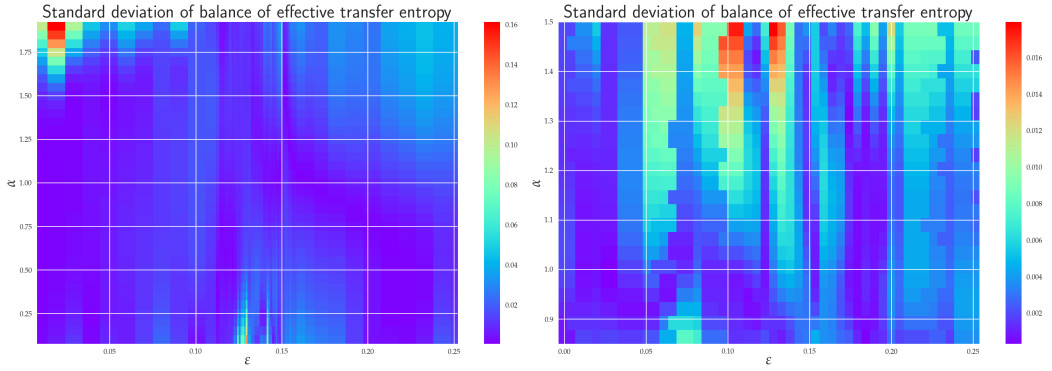

Figure 11. Dependence of standard deviation of the balance of effective RTEs $T^{R,\,\text{balance, effective}}_{\alpha,x_1\to y_1}(\{0,1\},\{1\},\{0\})$ (**left**) and $T^{R,\,\text{balance, effective}}_{\alpha,X\to Y}(\{0\},\{1\},\{0\})$ (**right**).

7. Discussion and Conclusions

7.1. Theoretical Results

How one discerns 'cause' from 'effect' is the main question in many scientific areas. The seminal contribution of Wiener and Granger led to the so-called Granger causality principle and time series analysis method for inference of causality from experimental data. The traditional Granger causality method is based on linear autoregressive processes. However, nonlinear complex systems cannot be well-described by linear autoregressive models and require appropriate generalizations of the Granger causality method. One successful generalization stems from information theory, using a form of conditional mutual information, also known as transfer entropy. Shannon entropy-based TE has become a

standard tool used for inferring causality from time series in all areas of science (including finance, climatology, neuroscience, etc.).

In this paper, instead of the Shannon entropy, we employed yet another information quantity, namely Rényi entropy. The ensuing RTE has the principal advantage that it is based on *a bona fide* information measure. In this way, one has a clear quantifier of the conveyed directional information (measured in bits or nats). Consequently, statements, such as: "the conveyed directional information from a tail part of the distribution is comparable with information from central part of distribution" or "information transfer is small/large (or good/bad)" are meaningful. RE is a measurable quantity; in principle, it can be measured directly (similar to Clausius entropy or Shannon entropy) without invoking the concept of the underlying distribution. This is because RE has an operational meaning given by various coding theorems. In practice, this is how RE is measured, e.g., in quantum optics (or more generally quantum information theory) [50,55]. In a conventional time series, one does not proceed this way because coding theorems (such as the Campbell coding theorem [44]) are difficult to implement for a large number of data.

As a proof of principle, we tested the concept of RTE on two unidirectionally coupled Rössler systems. The idea was to illustrate how the RTE can deal with such issues as synchronization and, more generally, causality in systems that are complex enough and yet amenable to a numerical analysis. Coupled RS is one of a handful of (simple) coupled chaotic systems that have been studied in the literature by means of Shannon's TE. This point is particularly important because we needed a gauge to which we could compare our results (and to which our results should reduce for $\alpha = 1$). Despite the earlier applications of the RTE in bivariate (mostly financial) time series, many questions remained unanswered about how to properly qualify and quantify the results obtained. Here, we went 'some way' toward this goal.

First, we showed that the concept of the Granger causality is exactly equivalent to the RTE for Gaussian processes, which may, in turn, be used as a test of Gaussianity. This is because RTEs are in the Gaussian framework all the same, and, hence, the results should be α-independent. On the other hand, since the efficiency and robustness of RTE estimators crucially hinge on the parameter α employed, it might be (in many cases) easier to follow the information-theoretic route to Granger causality (provided the Gaussian framework is justified).

Second, we demonstrated that the equivalence between the Granger causality and RTE can also be established for certain heavy-tailed processes—for instance, for soft α-Gaussian processes. In particular, in this latter case, one could clearly see the connection between Granger causality, Rényi's parameter α, and the heavy-tail power.

7.2. Numerical Analysis of RTE for Rössler Systems

In order to estimate the RTE, we employed the ℓ-nearest-neighbor entropy estimator of Leonenko et al. [26]. The latter is not only suitable for RE evaluation but it can also be easily numerically implemented to RTEs so that these can be computed almost in real time, which is relevant, e.g., in finance, regarding various risk-aversion decisions. Spurious effects caused by the finite size of the data set were taken into account by working with effective RTEs.

In order to gain further insight into the practical applicability and efficiency of the RTE, we tested it on two unidirectionally coupled Rössler systems—the master and slave system. To have a clear idea about what to expect, we first looked at the phenomenology of the coupled RSs by means of simple numerical simulations (presented in Figure 5). This was also accompanied by comparisons with Lyapunov exponents computed in references [18,22] and reproduced in Figure 4. In particular, we could clearly observe how the RSs synchronized with the increasing value of coupling strength. In this connection, we also identified critical values of coupling strengths at which *thresholds to transient behavior* (or the "topological phase transition") and the *threshold to full synchronization* occurred.

More specifically, we were particularly interested in the transient region between chaotic correlation regimes and full synchronization, which had not as yet been discussed in the literature. To gain a better understanding of this region, we employed in the range $\varepsilon \in [0.1, 0.15]$ a higher frequency sampling, namely 0.001, in contrast to the standard 0.01 one used for other εs. The threshold to transient behavior was identified at the scale $\varepsilon = 0.12$ where the positive LE crossed to negative values and where the projection on the x_1-y_1 and x_2-y_2 planes underwent topology changes (cf. Figure 5). From the point of view of RTEs, this threshold behavior was reflected in peaking the information flow in various directions. The increase in the effective RTE between x_1 and y_1 (in both directions) for $\alpha > 1$ was pronounced, in particular, which reflected the increase in orbit occupation density around the peak in the y_1-y_2 plane in the slave system. Even more marked was the high peak in information flow from x_3 to y_3 for $\alpha < 1$ (see Figure 10), which described an influx of information needed to "organize" chaotic correlations that existed between the x_3 and y_3 directions prior to $\varepsilon \lesssim 0.12$. Furthermore, the RTE was especially instrumental in understanding the measure concentration phenomenon in the transient regime. Finally, after a sharp "first-order-type" transition at the threshold of synchronization, the effective RTEs slowly approached their asymptotic values (distinct for each α) in the synchronized state. In addition, in the synchronized state, both $T^{R,\,\text{effective}}_{\alpha,X\to Y}$ and $T^{R,\,\text{effective}}_{\alpha,Y\to X}$ approached each other, which reveals that both $\{X\}$ and $\{Y\}$ systems have the same underlying distributions and, hence, they are indistinguishable.

As for the causality issue, we observed that the RTE is a good *causal measure* only if the conditioning has a sufficient dimension (in our case 3); otherwise, it is merely a *measure of dependence*. By employing effective RTE for the full system, we could reliably infer the coupling direction but only until $\varepsilon \lesssim 0.12$, i.e., until the threshold to transient behavior. After this value, the RSs started to synchronize, first partially (in the transient regime) and then fully $\varepsilon = 0.15$. In fact, the full synchronization started when the information flows from all sectors of underlying distributions (i.e., for all αs) began to be (almost) ε independent and when $T^{R,\,\text{balance, effective}}_{\alpha,X\to Y}$ approached zero— so there was a one-to-one relation between the states of the systems and the time series of the $\{X\}$ system could be predicted from the time series $\{Y\}$ system, and vice versa; hence, one could not make any statement about the coupling direction.

We should also reemphasize that the standard deviation of the RTE importantly depends on α, cf. Equation (11). For instance, the balance effective RTE for the full system is around the transient region quite reliably described by $0.8 \lesssim \alpha \lesssim 1.25$, though the minimal noise value is not attained at $\alpha = 1$ (Shannon transfer entropy) but at $\alpha = 1.16$. Clearly, the α-dependence of fluctuations is generally dynamics-dependent, and in many interesting real-world processes, it is simply more reliable to utilize non-Shannonian TEs.

7.3. Conclusions

In this paper, we discussed the Rényi transfer entropy and its role in the inference of causal relations between two systems, i.e., in the identification of the driving and driven systems from the experimental time series. On the theoretical side, our focus was on understanding the connection between RTE and Granger causality. In particular, we proved that the Granger causality is entirely equivalent to the RTE for Gaussian processes. This generalizes the classic result of Barnett et al. [61] that is valid for Shannon's TE. Furthermore, we have also shown how the Granger causality and the RTE are related in the case of heavy-tailed (namely α-Gaussian) processes. These results allow one to bridge the gap between autoregressive and Rényi entropy-based information-theoretic approaches.

On the experimental side, we illustrated the inner workings of the RTE by analyzing RTE between the synthetic time series generated from two unidirectionally coupled Rössler systems that are known to undergo synchronization. The route to synchronization was scrutinized by considering the effective RTE (and other derived concepts) between various master–slave components as well as between the full master and slave systems. We observed that with the effective RTE one could clearly identify a transient synchronization

region (in the coupling strength), i.e., the regime between chaotic (master–slave) correlations and the synchronization threshold. In the transient region, the effective RTE allowed inferring the measure concentration for the orbit occupation density. It is noteworthy to mention that the latter cannot be deduced from Shannon's TE alone.

We also saw that the direction of coupling and, hence, causality, could be reliably inferred only for coupling strengths $\varepsilon < 0.12$ (the onset of the transient regime), i.e., when two RSs were coupled, but not yet fully. This is in agreement with earlier observations, cf., e.g., reference [22]. As soon as the RSs were synchronized, they produced identical time series; hence, there is no way to infer the correct causality relation solely from the measured data.

We conclude with a general observation—a clear conceptual advantage of information-theoretic measures in general, and RTE in particular, as compared to the standard Granger causality, are sensitive to nonlinear signal properties, as they do not rely on linear regression models. On the other hand, a clear limitation of RTEs, in comparison to the Granger causality, is that they are—by their very formulation—restricted to bivariate situations (though multivariate generalization is possible, it substantially increases dimensionality in the estimation problem, which might be hard to solve with a limited amount of available data). In addition, the RTEs often require substantially more data than regression methods.

Author Contributions: Conceptualization, P.J.; formal analysis, H.L. and Z.T.; methodology, P.J., H.L. and Z.T.; validation, H.L. and Z.T.; software design, data structures, computer calculation, and visualization, H.L.; writing—original draft, P.J.; writing—review and editing, P.J., H.L. and Z.T. All authors have read and agreed to the published version of the manuscript.

Funding: P.J. and H.L. were supported by the Czech Science Foundation, grant no. 19-16066S and Z.T. by the Jubiläumsfonds der Österreichischen Nationalbank Project 18696. This work was also in part supported by the U.S. Army RDECOM—Atlantic Grant No. W911NF-17-1-0108.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We thank Milan Paluš for the helpful comments and discussions and for providing us with source code for Figure 4.

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

Abbreviations

The following abbreviations are used in this manuscript:

RE	Rényi entropy
TE	transfer entropy
RTE	Rényi transfer entropy
PDF	probability density function
ITE	information-theoretic entropy
RS	Rössler system
KSE	Kolmogorov–Sinai entropy rate
LE	Lyapunov exponent

Appendix A

Here, we provide a brief technical exposition of the RE estimator employed.

Finding good estimators for the RE is an open research area. The estimators for the Shannon entropy based on ℓ-nearest-neighbor in one-dimensional spaces were studied in statistics almost 60 years ago by Dobrushin [79] and Vašíček [80]. One disadvantage of these estimators is that they cannot easily be generalized to higher-dimensional spaces, so they are inapplicable to the TE calculations. Nowadays, there are many usable frameworks—

most of them, of course, in the Shannonian setting (see reference [2] for a recent review). However, it is important to stress that the naive estimation of TE by partitioning of the state space is problematic [19] and that such estimators frequently fail to converge to the correct result [81]. In practice, more sophisticated techniques, such as kernel [82] or ℓ–nearest-neighbor estimators [83,84], need to be utilized. However, the latter techniques may bring about their own assumptions about the empirical distributions of the data (see [81] for a discussion about the issues involved).

In our work, we used the ℓ-nearest-neighbor entropy estimator for higher-dimensional spaces introduced by Leonenko et al. [26]. This estimator is suitable for RE and it can be effectively adapted and implemented by using formulas from the above-mentioned papers. In particular, the approach is based on an estimator of the RE from a finite sequence of N points that is defined as

$$\widehat{H}_{N,\ell,\alpha} = \begin{cases} \alpha \neq 1 & \log_B((N-1)\cdot V_m) + \frac{1}{1-\alpha}\left[\log_B \frac{\Gamma(\ell)}{\Gamma(\ell+1-\alpha)}\right. \\ & \left. + \log_B\left(\frac{1}{N}\sum_{i=1}^{N}\left(\rho_\ell^{(i)}\right)^{m(1-\alpha)}\right)\right] \\ \alpha = 1 & \log_B((N-1)\cdot \exp(-\psi(\ell))\cdot V_m) \\ & + \frac{m}{N}\sum_{i=1}^{N}\log_B\left(\rho_\ell^{(i)}\right) \end{cases} . \tag{A1}$$

Here, $\Gamma(x)$ is Euler's gamma function, $\psi(x) = -\Gamma'(x)/\Gamma(x)$ is the (negative) digamma function, $m = \dim X_t$ is the dimension of the data set space X_t, and $\rho_\ell^{(i)}$ is the distance from the data i to the ℓ-th nearest data counterpart using a metric in the space X_t. Moreover, V_m is the size of the ball in space X_t defined via the same metric. Finally, $\log_B$ is the logarithm with base B (we typically use $B = e$). In our computations, we employed the Euclidean metric, which has $V_m = \pi^{\frac{m}{2}}/\Gamma\left(\frac{m}{2}+1\right)$. Note that the estimator basically depends on N, i.e., the number of data in a data set and on ℓ, i.e., the rank of the nearest-neighbor used.

The advantages of the estimator (A1) in contrast to the standard histogram method are:

- It has relative accuracy for a small data set;
- It has applicability for high-dimensional data;
- The set estimators provide statistics for the estimation.

We should also note that, in contrast to other RE estimators, such as *fixed-ball* estimator [2], the estimator (A1) is not confined to any specific ranges of α values, though the efficiency of the estimator is, of course, α-dependent. We comment more on this point in Section 6. On the other hand, the disadvantage of this method involves the computational complexity of the algorithm and the complicated data container.

To calculate RTE and the related quantities (48)–(50), we apply the estimator Equation (A1). Ensuing estimators to (47)–(50)—let us call them generically $\mathcal{X}$—become dependent on ℓ (i.e., the nearest-neighbor rank). We exploit this feature and define the mean value $\overline{\mathcal{X}}$ and standard deviation $\sigma_\mathcal{X}$ with the Bessel correction, respectively, as

$$\overline{\mathcal{X}} = \frac{\sum_{\ell=n_{min}}^{n_{max}} \mathcal{X}_\ell}{n_{max} - n_{min} + 1}, \tag{A2}$$

$$\sigma_\mathcal{X} = \sqrt{\frac{\sum_{\ell=1}^{n}\left(\mathcal{X}_\ell - \overline{\mathcal{X}}\right)^2}{n_{max} - n_{min}}}. \tag{A3}$$

Here, n_{max} and n_{min} are the highest and the lowest orders of the nearest data counterparts, respectively. Theoretically, we should use $n_{max} = M$, where M stands for the number of samples, but such a setup would require an enormous amount of computer memory to hold the distances.

In our calculations, we used $n_{max} = 50$, which turned out to be a good compromise between accuracy and computer time. On the other hand, for n_{min}, we were a little bit restricted by the fact that n_{min} influenced the interval of convergence of the estimator for various α (cf. discussion and proof in [26]). For instance, for $\ell = 1$, the estimator converged

in the interval $\alpha \in [0, 1 + \frac{1}{2\dim(X_t)}]$, while for $\ell > 1$, one had $\alpha \in [0, \frac{\ell+1}{2}]$. For our particular purpose, it will suffice to set $n_{min} = 5$, so that the interval of convergence will be $\alpha \in [0, 3]$. This will fully suit our needs.

Appendix B

Here, we provide the heat maps for the relevant figures from the main text. These depict standard deviations (A2) and their dependencies on both α and ε.

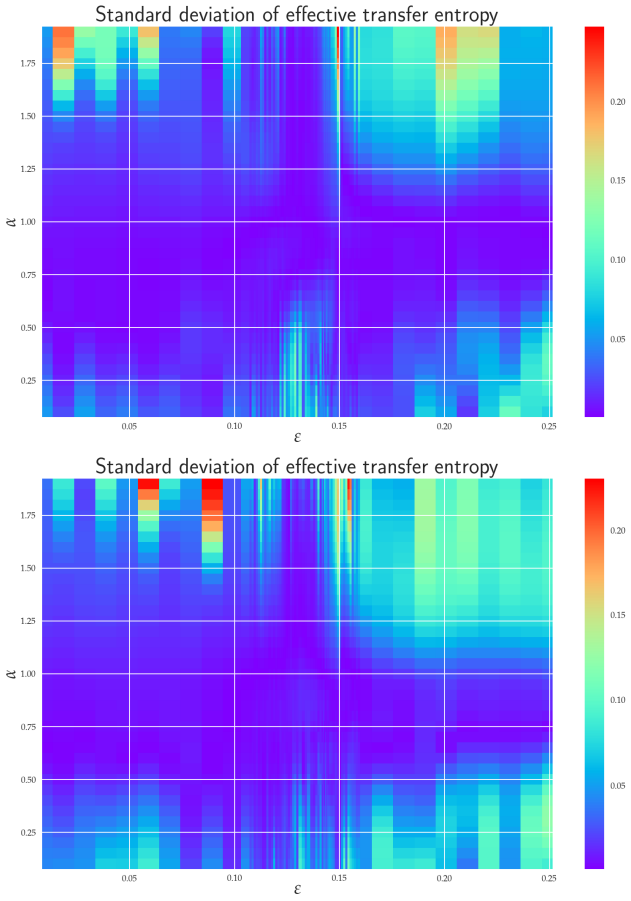

Figure A1. Standard deviation of the effective RTE between x_1 and y_1 $T^{R,\,\text{effective}}_{\alpha,x_1\to y_1}(\{0,1,2,3,4,5,6\},\{1\},\{0\})$, and $T^{R,\,\text{effective}}_{\alpha,y_1\to x_1}(\{0,1,2,3,4,5,6\},\{1\},\{0\})$.

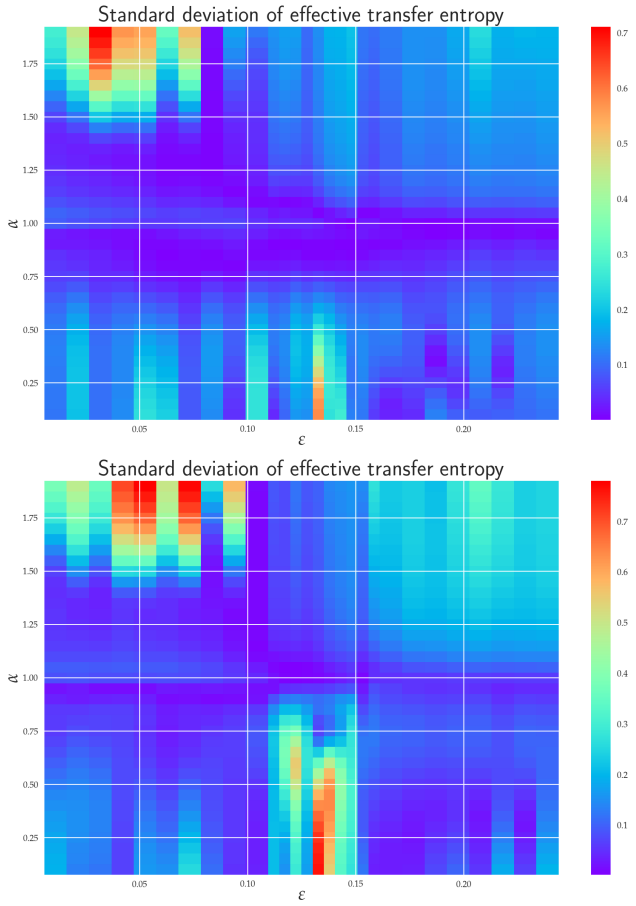

Figure A2. Standard deviation of the effective RTE between x_3 and y_3 for $T^{R,\text{effective}}_{\alpha,x_3\to y_3}(\{0,1,2,3,4\},\{1\},\{0\})$, and $T^{R,\text{effective}}_{\alpha,y_3\to x_3}(\{0,1,2,3,4\},\{1\},\{0\})$.

References

1. Schreiber, T. Interdisciplinary application of nonlinear time series methods. *Phys. Rep.* **1999**, *308*, 1–64. [CrossRef]
2. Kantz, H.; Schreiber, T. *Nonlinear Time Series Analysis*; Cambridge University Press: Cambridge, UK, 2010.
3. Pecora, L.M.; Carroll, T.L. Synchronization in chaotic systems. *Phys. Rev. Lett.* **1990**, *64*, 821–824. [CrossRef] [PubMed]
4. Boccaletti, S.; Kurths, J.; Osipov, G.; Valladares, D.L.; Zhou, C.S. The synchronization of chaotic systems. *Phys. Rep.* **2002**, *366*, 1–101. [CrossRef]
5. Quiroga, R.Q.; Arnhold, J.; Grassberger, P. Learning driver-response relationships from synchronization patterns. *Phys. Rev.* **2000**, *E61*, 5142–5148. [CrossRef]
6. Nawrath, J.; Romano, M.C.; Thiel, M.; Kiss, I.Z.; Wickramasinghe, M.; Timmer, J.; Kurths, J.; Schelter, B. Distinguishing Direct from Indirect Interactions in Oscillatory Networks with Multiple Time Scales. *Phys. Rev. Lett.* **2010**, *104*, 038701. [CrossRef]
7. Sugihara, G.; May, R.; Ye, H.; Hsieh, C.; Deyle, E.; Fogarty, M.; Munch, S. Detecting causality in complex ecosystems. *Science* **2012**, *338*, 496–500. [CrossRef]
8. Feldhoff, J.H.; Donner, R.V.; Donges, J.F.; Marwan, N.; Kurths, J. Geometric detection of coupling directions by means of inter-system recurrence networks. *Phys. Lett.* **2012**, *A376*, 3504–3513. [CrossRef]
9. Wiener, N. *Modern Mathematics for Engineers*; Beckenbach, E.F., Ed.; McGraw-Hill: New York, NY, USA, 1956.
10. Granger, C.W.J. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica* **1969**, *37*, 424–438. [CrossRef]
11. Ancona, N.; Marinazzo, D.; Stramaglia, S. Radial basis function approach to nonlinear Granger causality of time series. *Phys. Rev.* **2004**, *R70*, 056221. [CrossRef]
12. Chen, Y.; Rangarajan, G.; Feng, J.; Ding, M. Analyzing multiple nonlinear time series with extended Granger causality. *Phys. Lett.* **2004**, *A324*, 26–35. [CrossRef]
13. Wismüller, A.; Souza, A.M.D.; Vosoughi, M.A.; Abidin, A.Z. Large-scale nonlinear Granger causality for inferring directed dependence from short multivariate time-series data. *Sci. Rep.* **2021**, *11*, 7817. [CrossRef]
14. Zou, Y.; Romano, M.; Thiel, M.; Marwan, N.; Kurths, J. Inferring indirect coupling by means of recurrences. *Int. J. Bifurc. Chaos* **2011**, *21*, 1099–1111. [CrossRef]

15. Donner, R.V.; Small, M.; Donges, J.F.; Marwan, N.; Zou, Y.; Xiang, R.; Kurths, J. Recurrence-based time series analysis by means of complex network methods. *Int. J. Bifurc. Chaos* **2011**, *21*, 1019–1046. [CrossRef]
16. Romano, M.; Thiel, M.; Kurths, J.; Grebogi, C. Estimation of the direction of the coupling by conditional probabilities of recurrence. *Phys. Rev.* **2007**, *E76*, 036211. [CrossRef]
17. Vejmelka, M.; Paluš, M. Inferring the directionality of coupling with conditional mutual information. *Phys. Rev.* **2008**, *77*, 026214. [CrossRef] [PubMed]
18. Paluš, M.; Krakovská, A.; Jakubík, J.; Chvosteková, M. Causality, dynamical systems and the arrow of time. *Chaos* **2018**, *28*, 075307. [CrossRef] [PubMed]
19. Schreiber, T. Measuring Information Transfer. *Phys. Rev. Lett.* **2000**, *85*, 461–464. [CrossRef]
20. Marschinski, R.; Kantz, H. Analysing the Information Flow Between Financial Time Series. *Eur. Phys. J. B* **2002**, *30*, 275–281. [CrossRef]
21. Jizba, P.; Kleinert, H.; Shefaat, M. Rényi's information transfer between financial time series. *Physica A* **2012**, *391*, 2971–2989. [CrossRef]
22. Paluš, M.; Vejmelka, M. Directionality of coupling from bivariate time series: How to avoid false causalities and missed connections. *Phys. Rev.* **2007**, *75*, 056211. [CrossRef]
23. Runge, J.; Heitzig, J.; Petoukhov, V.; Kurths, J. Escaping the Curse of Dimensionality in Estimating Multivariate Transfer Entropy. *Phys. Rev. Lett.* **2012**, *108*, 258701. [CrossRef] [PubMed]
24. Faes, L.; Kugiumtzis, D.; Nollo, G.; Jurysta, F.; Marinazzo, D. Estimating the decomposition of predictive information in multivariate systems. *Phys. Rev.* **2015**, *91*, 032904. [CrossRef] [PubMed]
25. Sun, J.; Taylor, D.; Bollt, E.M. Causal Network Inference by Optimal Causation Entropy. *SIAM J. Appl. Dyn. Syst.* **2015**, *14*, 73–106. [CrossRef]
26. Leonenko, N.; Pronzato, L.; Savani, V. A class of Rényi information estimators for multidimensional densities. *Ann. Stat.* **2008**, *36*, 2153–2182; Correction in *Ann. Stat.* **2010**, *38*, 3837–3838. [CrossRef]
27. Lungarella, M.; Pitti, A.; Kuniyoshi, Y. Information transfer at multiple scales. *Phys. Rev.* **2007**, *76*, 056117. [CrossRef]
28. Faes, L.; Nollo, G.; Stramaglia, S.; Marinazzo, D. Multiscale Granger causality. *Phys. Rev.* **2017**, *76*, 042150. [CrossRef]
29. Paluš, M. Multiscale Atmospheric Dynamics: Cross-Frequency Phase-Amplitude Coupling in the Air Temperature. *Phys. Rev. Lett.* **2014**, *112*, 078702. [CrossRef]
30. Tsallis, C. *Introduction to Nonextensive Statistical Mechanics: Approaching a Complex World*; Springer: New York, NY, USA, 2009.
31. Thurner, S.; Hanel, R.; Klimek, P. *Introduction to the Theory of Complex Systems*; Oxford University Press: London, UK, 2018.
32. Rössler, O.E. An equation for continuous chaos. *Phys. Lett.* **1976**, *57*, 397–398. [CrossRef]
33. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *5*, 379–423, 623–656. [CrossRef]
34. Jizba, P.; Arimitsu, T. The world according to Rényi: Thermodynamics of multifractal systems. *Ann. Phys.* **2004**, *312*, 17–59. [CrossRef]
35. Burg, J.P. The Relationship Between Maximum Entropy Spectra In addition, Maximum Likelihood Spectra. *Geophysics* **1972**, *37*, 375–376. [CrossRef]
36. Tsallis, C. Possible generalization of Boltzmann-Gibbs statistics. *J. Stat. Phys.* **1988**, *52*, 479–487. [CrossRef]
37. Havrda, J.; Charvát, F. Quantification Method of Classification Processes: Concept of Structural a-Entropy. *Kybernetika* **1967**, *3*, 30–35.
38. Frank, T.; Daffertshofer, A. Exact time-dependent solutions of the Rényi Fokker–Planck equation and the Fokker–Planck equations related to the entropies proposed by Sharma and Mittal. *Physica A* **2000**, *285*, 352–366. [CrossRef]
39. Sharma, B.D.; Mitter, J.; Mohan, M. On measures of "useful" information. *Inf. Control* **1978**, *39*, 323–336. [CrossRef]
40. Jizba, P.; Korbel, J. On q-non-extensive statistics with non-Tsallisian entropy. *Physica A* **2016**, *444*, 808–827. [CrossRef]
41. Vos, G. Generalized additivity in unitary conformal field theories. *Nucl. Phys. B* **2015**, *899*, 91–111. [CrossRef]
42. Rényi, A. *Probability Theory*; North-Holland: Amsterdam, The Netherlands, 1970.
43. Rényi, A. *Selected Papers of Alfréd Rényi*, 2nd ed.; Akademia Kiado: Budapest, Hungary, 1976.
44. Campbell, L.L. A coding theorem and Rényi's entropy. *Inf. Control* **1965**, *8*, 423–429. [CrossRef]
45. Csiszár, I. Generalized cutoff rates and Rényi's information measures. *IEEE Trans. Inform. Theory* **1995**, *26*, 26–34. [CrossRef]
46. Csiszár, I.; Shields, P.C. *Information and Statistics: A Tutorial*; Publishers Inc.: Boston, MA, USA, 2004.
47. Aczél, J.; Darótzy, Z. *Measure of Information and Their Characterizations*; Academic Press: New York, NY, USA, 1975.
48. Halsey, T.C.; Jensen, M.H.; Kadanoff, L.P.; Procaccia, I.; Schraiman, B.I. Fractal measures and their singularities: The characterization of strange sets. *Phys. Rev.* **1986**, *A33*, 1141–1151. [CrossRef]
49. Mandelbrot, B.B. *Fractals: Form, Chance and Dimension*; W. H. Freeman: San Francisco, CA, USA, 1977.
50. Bengtsson, I.; Życzkowski, K. *Geometry of Quantum States. An Introduction to Quantum Entanglement*; Cambridge University Press: Cambridge, UK, 2006.
51. Jizba, P.; Korbel, J. Maximum Entropy Principle in Statistical Inference: Case for Non-Shannonian Entropies. *Phys. Rev. Lett.* **2019**, *122*, 120601. [CrossRef] [PubMed]
52. Jizba, P.; Korbel, J. When Shannon and Khinchin meet Shore and Johnson: Equivalence of information theory and statistical inference axiomatics. *Phys. Rev.* **2020**, *E101*, 042126. [CrossRef] [PubMed]
53. Lesche, B. Instabilities of Rényi entropies. *J. Stat. Phys.* **1982**, *27*, 419–422. [CrossRef]

54. Rényi, A. On measures of entropy and information. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 20–30 June 1961; pp. 547–561.
55. Jizba, P.; Ma, Y.; Hayes, A.; Dunningham, J.A. One-parameter class of uncertainty relations based on entropy power. *Phys. Rev. E* **2016**, *93*, 060104(R). [CrossRef]
56. Hentschel, H.G.E.; Procaccia, I. The infinite number of generalized dimensions of fractals and strange attractors. *Physica D* **1983**, *8*, 435–444. [CrossRef]
57. Harte, D. *Multifractals Theory and Applications*; Chapman and Hall: New York, NY, USA, 2019.
58. Latora, V.; Baranger, M. Kolmogorov–Sinai Entropy Rate versus Physical Entropy. *Phys. Rev. Lett.* **1999**, *82*, 520–523. [CrossRef]
59. Jizba, P.; Korbel, J. On the Uniqueness Theorem for Pseudo-Additive Entropies. *Entropy* **2017**, *19*, 605. [CrossRef]
60. Geweke, J. Measurement of Linear Dependence and Feedback between Multiple Time Series. *J. Am. Stat. Assoc.* **1982**, *77*, 304–313. [CrossRef]
61. Barnett, L.; Barrett, A.B.; Seth, A.K. Granger Causality and Transfer Entropy are Equivalent for Gaussian Variables. *Phys. Rev. Lett.* **2009**, *103*, 238701. [CrossRef]
62. Jizba, P.; Dunningham, J.A.; Joo, J. Role of information theoretic uncertainty relations in quantum theory. *Ann. Phys.* **2015**, *355*, 87–114. [CrossRef]
63. Seth, A.K. A MATLAB toolbox for Granger causal connectivity analysis. *J. Neurosci. Methods* **2010**, *186*, 262–273. [CrossRef] [PubMed]
64. Jizba, P.; Korbel, J. Multifractal Diffusion Entropy Analysis: Optimal Bin Width of Probability Histograms. *Physica A* **2014**, *413*, 438–458. [CrossRef]
65. Kečkić, J.D.; Vasić, P.M. Some inequalities for the gamma function. *Publ. De L'Institut Mathématique* **1971**, *11*, 107–114.
66. Fisher, R.A.; Yates, F. *Statistical Tables for Biological, Agricultural and Medical Research*, 3rd ed.; Oliver & Boyd: Edinburgh, UK, 1963.
67. Matsumoto, M.; Nishimura, T. Mersenne Twister: A 623-Dimensionally Equidistributed Uniform Pseudo-Random Number Generator. *ACM Trans. Model. Comput. Simul.* **1998**, *8*, 3–30. [CrossRef]
68. Theiler, J.; Eubank, S.; Longtin, A.; Galdrikian, B.; Farmer, J.D. Testing for nonlinearity in time series: The method of surrogate data. *Physica D* **1992**, *58*, 77–94. [CrossRef]
69. Schreiber, T.; Schmitz, A. Improved Surrogate Data for Nonlinearity Tests. *Phys. Rev. Lett.* **1996**, *77*, 635–638. [CrossRef]
70. Schreiber, T.; Schmitz, A. Surrogate time series. *Physica D* **2000**, *142*, 346–382. [CrossRef]
71. Paluš, M., Linked by Dynamics: Wavelet-Based Mutual Information Rate as a Connectivity Measure and Scale-Specific Networks. In *Advances in Nonlinear Geosciences*; Springer International Publishing: Cham, Switzerland, 2018; pp. 427–463.
72. Rosenblum, M.G.; Pikovsky, A.; Kurths, J. Phase Synchronization of Chaotic Oscillators. *Phys. Rev. Lett.* **1996**, *76*, 1804–1807. [CrossRef]
73. Cheng, A.L.; Chen, Y.Y. Analyzing the synchronization of Rössler systems—When trigger-and-reinject is equally important as the spiral motion. *Phys. Lett.* **2017**, *381*, 3641–3651. [CrossRef]
74. Rössler, O.E. Different Types of Chaos in Two Simple Differential Equations. *Z. Naturforsch.* **1976**, *31*, 1664–1670. [CrossRef]
75. Virtanen, P.; Gommers, R.; Oliphant, T.; Travis, E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261–272. [CrossRef] [PubMed]
76. Hunter, J.D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95. [CrossRef]
77. Harris, C.R.; Millman, K.J.; van der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; et al. Array programming with NumPy. *Nature* **2020**, *585*, 357–362. [CrossRef]
78. Use Branch `tranfer_entropy`. Available online: https://github.com/jajcayn/pyclits (accessed on 16 March 2022).
79. Dobrushin, R.L. A simplified method of experimentally evaluating the entropy of a stationary sequence. *Teor. Veroyatnostei I Ee Primen.* **1958**, *3*, 462–464. [CrossRef]
80. Vašíček, O. A test for normality based on sample entropy. *J. Roy. Stat. Soc. Ser. B Methodol.* **1976**, *38*, 54–59.
81. Kaiser, A.; Schreiber, T. Information transfer in continuous processes. *Physica D* **2002**, *166*, 43–62. [CrossRef]
82. Silverman, B.W. *Density Estimation for Statistics and Data Analysis*; Chapman & Hall: London, UK, 1986.
83. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. *Phys. Rev.* **2004**, *69*, 066138. [CrossRef]
84. Frenzel, S.; Pompe, B. Partial Mutual Information for Coupling Analysis of Multivariate Time Series. *Phys. Rev. Lett.* **2007**, *99*, 204101. [CrossRef]