LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
Department of Physics
Dr. Christoph Räth

Master's Thesis

Evaluation and Extension of the Transfer Entropy Calculus for the Measurement of Information Flows Between Futures Time Series During the COVID-19 Pandemic

Max Raphael Mynter
11819675
max.mynter@physik.lmu.de

Processing Period: March 12, 2021 - September 13, 2021
Supervisor: Dr. Christoph Räth

LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
Fakultät für Physik
Dr. Christoph Räth

Masterarbeit

Evaluation und Ausweitung des Transfer Entropie Kalküls zur Quantifizierung von Informationsflüssen zwischen Futures Zeitreihen während der COVID-19 Pandemie

Max Raphael Mynter
11819675
max.mynter@physik.lmu.de

Bearbeitungszeitraum: 12.03.2021 bis 13.09.2021
Betreuung und Prüfung: Dr. Christoph Räth

Acknowledgments

I want to thank my supervisor, Christoph Räth, for giving me the time and space to explore my scientific interests and provide tight guidance when needed. I want to thank Alexander Haluszczynski and Haochun Ma for fruitful discussions and ideas in regular meetings. I sincerely appreciate my friends for their moral (and proofreading) support during my thesis. Lastly, I want to thank my parents for their support throughout my studies.

Contents

1 Introduction

Non-linear relations between variables are ubiquitous in both natural and social systems. For researchers, it is challenging to evaluate these relations if their functional form is unknown. Canonical measures to quantify relations between variables such as the *Bravais Pearson Product Moment Correlation* can only quantify linear relationships. This limitation renders the common practice [9] to classify this measure as non-parametric short-sighted.

Transfer entropy [60] is an information-theoretic measure able to quantify both linear and non-linear relations without assuming a functional form of this relation or the distribution of the underlying samples. Unlike the *Pearson Correlation*, the transfer entropy is asymmetric. Therefore, we can use it to infer the direction of information flow. When we apply this measure to time series data, we can interpret this information flow as akin to a causal relation.
We will develop the theoretical foundation of his measure in the chapter (3).

While the concept of a genuinely non-parametric measure sounds promising, transfer entropy is only well defined for discrete data. Proper discretization seems like an easy fix for this limitation when researchers apply this measure to continuous data. However, previous research [44] found a strong dependence of the measure on the discretization hyperparameters.

In this thesis, we will extensively study the dependence of the transfer entropy measure onto these parameter choices and provide an attempt to develop a more robust transfer entropy calculus. Subsequently, we provide a recipe as a result of these analyses in section (5). We conduct this evaluation with a non-linear model system and compare multiple standard binning methods. Some readers might also find the theoretical discussion of these binning methods in chapter 2 (specifically 2.3) useful for other applications.

Next to the discretization methods, we also cover the influence of sample size and data partition detail and the effect of other standard practices such as gaussian remapping or rescaling. Additionally, we will evaluate the effect of noise on the estimation.

In section (4), we develop a novel probability estimator to incorporate the bin size into probability estimation (and thus transfer entropy calculation). We will also discuss some of its properties. This estimator is evaluated in chapter (5.6).
Another main contribution of this section is the evaluation of the normalizations of transfer entropy in section (5.3). The normalization enables comparability of the measure's value.

Once we finish the evaluation and extension of the transfer entropy method, we will apply it to stock index futures contracts as an example system of non-linearly related time series. Namely, in the chapter (6), we will discuss the transfer entropy relation of index futures before and during the first wave of the COVID-19 pandemic. In that context, we will also evaluate the information flow from online sentiment (captured from the Reddit [6] platform). We will then compare it to the financial time series relations in other periods, namely the dot-com bubble and the US housing crisis. Finally, we will also evaluate the relation of information flow between index futures (that should be zero in arbitrage-free markets [20]) to public policy uncertainty.

We find increases in information flows between index futures on the onset of the COVID-19 response measures. Testing these flows with surrogate data reveals the directionality of these flows. Additionally, we will find the flows to correlate with the economic policy uncertainty significantly.

As such, this thesis provides two main contributions to the scientific literature. First, we

conduct a thorough evaluation of the transfer entropy measure, discuss its caveats and provide novel attempts to control them.
Secondly, we uncover a change in the information dynamics between index futures starting in March 2020 during the COVID-19 pandemic by applying the transfer entropy calculus. We find this increase to be highly correlated with economic policy uncertainty. We, therefore, uncover evidence for the hypothesis that in economically uncertain times, financial markets lose efficiency.

Conventions and Notation Before we begin the discussion of the methods, let us discuss the notations used in this thesis. We will use uppercase letters, for example, X, to refer to a variable or a sampled process. Then, the calligraphic $\mathscr{X}$ refers to a specific empirical sample. Therefore, we will discuss the information-theoretic measures as functions of variables. Nonetheless, we calculate their empirical value from data. Thus the meaning of X and $\mathscr{X}$ is mostly interchangeable as in all cases throughout this thesis. We can only infer X via $\mathscr{X}$ as a proxy.

Within the evaluation of the transfer entropy measure in section (6), we will see plots of the same process with different resolutions. To unclutter these information-dense plots, we will omit legends. The lines (if not noted otherwise) refer to an increasing number of bins starting from a binary partition from bottom to top.

2 Data Analysis Methodology

This chapter discusses the necessary data acquisition, analysis, and processing methodology, which we apply in this thesis. First, we will introduce different data qualities and the applicable univariate measures for sample characterization and distribution estimation. Then, as proper discretization (or coarse-graining) of (quasi-) continuous data is a significant contribution of this thesis, we will introduce traditional binning methods alongside those adapted from statistical criteria and clustering algorithms. We will then proceed to discuss bivariate measures, some of which rely on the chosen quantization. Within this section, we will also introduce transfer entropy with its technical details. However, since transfer entropy is the central bivariate measure of this thesis, we will devote the subsequent chapter (3) to a comprehensive theoretical derivation. The last section of this chapter concerns additional processing techniques (other than discretization) for samples that we apply within this thesis. These methods enable studying the temporal evolution of bivariate measures and quantifying the nonlinear contribution to the variable association.

2.1 Measurement, Time, and Scale Types

A collected sample can carry different amounts of information depending on the acquisition process, how the data is processed, stored, and evaluated. In this first subsection, we will discuss these different qualities.

2.1.1 Measurement Process

Data is a collection of observations. The designation *datum* stems from the Latin for *something given* and describes an information-carrying observation on a population or population sample [1]. Thus the measurement process is the assignment of signs or numerals to observed sample facts according to a specific rule [67]. As such, the measurement $\mathscr{M}$ is a map from the state space of the underlying population or process, Ω, of the to be measured variable $\Omega = \Omega_V$ to the sample $\mathscr{V}$ which is a subset of the possible measurement values S and their corresponding order relation:

$$\begin{aligned} \mathscr{M} :&\Omega_V \to \mathscr{X} \subseteq S \\ &\omega_V \mapsto x + \varepsilon. \end{aligned} \tag{1}$$

The resulting observations can contain a single or multiple variables, $\Omega = \Omega_{V_1,V_2,\ldots}$ in which case the above expression is generalized accordingly. $\varepsilon = \varepsilon_{\text{Systematic}} + \varepsilon_{\text{Noise}}$ is the measurement error. It consists of a systematic contribution that remains constant. Systematic errors may occur due to miscalibration of measurement instruments or errors in the data processing. The noise term is different for each datum. It may root in different factors such as misreading of a scale or variations of the context. We understand the noise term as the realization of a random variable that has a certain probability distribution.

Now, the sample, $\mathscr{V}$, from a source with distribution V is a set of arbitrarily ordered observations $v_j \in \mathscr{V}$. If we take subsequent observations, then the map originates from the time-dependent state space Ω^t and one generates longitudinal data of temporally ordered observations, $\mathscr{V}_t$. Within this thesis, context determines if a sample is time-ordered. Therefore, at times we drop the subscript t to simplify expressions.
If the temporal spacing of observations is small and resultingly the measurement frequency is high enough, one refers to the longitudinal data as time series [25]. However, this distinction is subjective, and no sharp boundary exists.

2.1.2 Scale Qualities

What distinguishes a datum $x \in \mathscr{X} \subseteq S$ from a mere number is the associated meaning. How this numerical value is interpretable depends on the set S of possible values. Resultingly, the sample can be of one of four different scales of measurement: the nominal, ordinal, interval, and ratio scale [76]. We will now introduce these scale types. However, we will neglect a mathematically rigorous discussion as intuitive understanding suffices for our purposes. Interested readers can find it in the appendix (8.1).

Nominal Scale The nominal scale allows for a distinction between values but does not allow ordering. For example, a person's name allows for distinction, but the notion that Alice is in some sense more than Bob carries no meaning. Even if Alice and Bob were soccer players with the Jersey numbers 10 and 2 respectively, the statement would be void.

Ordinal Scale For an ordinal scale the $x \in \mathscr{X} \subseteq S$ can be ordered. It makes sense to say that Alice is in a better (higher) mood than Bob. However, while ordering is possible, there is no meaning in distances between values. We cannot quantify the extent to which Alice is happier than Bob.

Interval Scale For the interval scale, we can order the measured values by magnitude. Their differences exist and can also be ordered. The relationship between the possible values of the variables is positive monotonic, and linear. Thus, differences in values carry meaning. One interval scale can be mapped onto another by a linear transformation [9].

Ratio Scale The ratio scale is distingushed from the interval scale by the existence of a true zero point. Thus, the ratios of values carry meaning. For example $^{2°\mathrm{C}}/_{1°\mathrm{C}} = 2$ but 2°C is not twice as hot as 1°C whereas $^{2\mathrm{K}}/_{1\mathrm{K}} = 2$ is. Squeezing operations on the ratio scale preserve the scale type where as the shift by a constant value (obviously) does not [9].

Qualitative data is of a nominal or ordinal scale. Quantitative data is on the interval or ratio scale, which are thus designated metric scales. If a quantitative variable can take only specific values within an interval, it is called discrete. If it can take any value in the interval, it is continuous.

2.2 Univariate Data Characterization

In this section, we will discuss univariate measures, $S : X \to \mathbb{R}$, to characterize a variable X. Usually, we do not have complete knowledge about X to analytically determine its properties. We then infer these from the empirical sample $\mathscr{X}$. Since $S(X)$ is the true value and $S(\mathscr{X})$ is an approximation, we use the first notation to discuss the measures within this thesis. All statements are straightforwardly adapted to deal with samples instead of variables.

Which measures S can be applied depends on the data scale. The nominal scale has fewer conditions than the ordinal scale, which has fewer than the interval scale. The ratio scale has the tightest axioms. Therefore measures defined on scale types with fewer axioms can be applied to

data on a scale with more if one transforms the sample to another scale type with a mapping

$$\begin{aligned} \varphi^D : \mathscr{X} &\to \mathscr{X}' \\ x &\mapsto x' \end{aligned} \tag{2}$$

that destroys the overhead information. For some measures, φ^D might be the identity map. Generally, however, φ^D is not injective and thus nonlinear and not invertible. In subsequent sections, we will establish that the choice of φ^D is nontrivial. The contribution towards a proper choice is one of the main merits of this thesis.

Measures can be parametric or non-parametric. The latter assume the distribution of the variable to belong to some family of distributions. Therefore obtaining the identifying parameters from the sample allows defining the form of the distributions uniquely. In contrast, non-parametric statistics do not rely on such assumptions.

2.2.1 Measures of Central Tendency

The measures of central tendency include, among others, the arithmetic and geometric means, median, and mode. Whereas the mean calculation requires metric scales, we can apply the median to the sub-metric ordinal scale. The mode works on nominal data. They designate a central value around which the data clusters. If one wants to calculate the mode on interval scale data, for example, a temperature sample with sufficient resolution, one might want to apply a coarse-graining φ^D which clusters the data such that the mode value becomes meaningful. Without φ^D, the values repeat too sparsely (if at all) for the mode to yield an accurate interpretation. While the clustering of data destroys the specific location information of each datum, the aggregated data yields a more robust picture of the most common temperature range in the measuring period.

2.2.2 Measures of Variability

Measures of variability measure the dispersion of the data around this central value. The variance $\sigma^2 = \sum_{i=1}^{n}(x_i - \mu)^2$ measures the quadratic deviation of the data from the mean, μ. The interquartile range, IQR, is another measure of dispersion that is more robust against extreme values [9].

2.2.3 Information Entropy

Information entropy is a univariate measure that we can interpret as a measure for the average information content or surprise encoded in a random variable. On an empirical sample, the statistic reads:

$$H(X) = \sum_{j=1}^{m} p(x_j) \log\left(p\left(x_j\right)\right). \tag{3}$$

where m is the number of discrete values of X and $p(x_j)$ the probability mass of the j^{th} one. This measure as well as possible generalizations and their respective problems are discussed in detail in section (3.2). Because of these problems the measure (3) is applied for continuous data with a suitable discretization (see section 2.3). The development of a criterion to identify such a suitable discretization is one contribution of this thesis and discussed in section (5.8).

2.2.4 Autocorrelation

Autocorrelation is the correlation (see section 2.4.1) of a time series with its time-shifted self,

$$r_x(\tau) = \rho(x_t, x_{t-\tau}) = \frac{\sum_{j=\tau+1}^{n}(x_{j-\tau} - \mu_x)(x_j - \mu_x)}{\sum_{j=1}^{n}(x_j - \mu_x)^2}. \tag{4}$$

τ is the shift of the time series. The value is bound to the $[-1,1]$ interval ranging from perfect anticorrelation to perfect correlation. Thus, 0 indicates no self-association over the time τ. As we want to calculate the autocorrelation for a finite sample $\mathscr{X}$, we are truncating them accordingly. This thesis will evaluate $r_x(\tau)$ to determine how much of the past (τ) we should include when calculating the transfer entropy measure. Autocorrelation is a linear measure and is not capable of capturing non-linear self-dependencies.

2.2.5 Power Spectrum

The power spectrum is the spectral density of a time series signal, the distribution of power over frequency. It is related to the autocorrelation function by the Fourier transform. This relation is known as the Wiener-Khinchin theorem [40]. Therefore the power spectrum $s = s(f)$ as a function of frequency f is calculated as

$$s_x(f) = \sum_{k=-\infty}^{\infty} r_x(\tau) e^{-2\pi\tau f}. \tag{5}$$

In our applications, we only calculate a finite number of terms. As both the Fourier transform and autocorrelation are linear, the power spectrum is a linear property.

2.2.6 Distribution Estimation

In the last section, we established the measurement process $\mathscr{M}$ as a map from the state space to the sample $\Omega \to \mathscr{X}$ (plus an error term). The state measured can be thought of as realizing a random variable from a particular probability distribution. This distribution might vary with time. However, given proper normalization, this corresponds merely to shaping the probability distribution throughout the measurement period for time series data. We are therefore keen to estimate the probability distribution in the state space using the sample. If one knows (or guesses) the family of probability distributions of the variable, this task reduces to the mere estimation of the parameters from the sample.

The Histogram Estimator for Continuous Variables If the distribution family is unknown, we can resort to non-parametric estimation methods. A histogram probably represents the simplest form of non-parametric density estimation [74]. For a continuous variable, it is a mapping from the real line to a finite set of corresponding disjoint class intervals $\{B_i\}_i$ with $|\{B_i\}_i| = m$. The histogram requires the sample to be of metric scale type, and the probability distribution we want to estimate is a probability density function (PDF). The mapping then reads

$$\begin{aligned} &\mathscr{H} : \mathbb{R} \to \{B_i\}_i \\ &x \mapsto B(x) = \begin{cases} B_1 & \text{if } x \leq \sup B_1 \\ B_2 & \text{if } \inf B_2 < x \leq \sup B_2 \\ & \vdots \\ B_m & \text{if } x > \inf B_m \end{cases} \end{aligned} \tag{6}$$

Literature commonly refers to these class intervals as bins. Let $\mathscr{X} = \{x_i\}_i$ be a sample of data from the PDF f with the support $[a,b]$. Then a partition partition into m uniformly sized bins is:

$$B_1 = \left[a, a+\frac{b-a}{m}\right), B_2 = \left[a+\frac{b-a}{m}, a+\frac{2(b-a)}{m}\right), \ldots, B_m = \left[a+\frac{(m-1)(b-a)}{m}, b\right].$$

However, the bins must not necessarily be of uniform size. We will cover several approaches for choosing the optimal number and size(-s) of bins in section (2.3). Additionally, the bounds a,b might be unknown. In that case, we estimate them from the minimum and maximum of the sample.

For a given binning B_j with the respective binwidths h_j we obtain the fraction of data points in the j^{th} bin $n_j/n = \hat{p}_j$, as an estimator for $p_j(x) = \mathbb{1}_{x\in B_j} \cdot \int_{B_j} f(u)du$. Subsequently, the histogram density estimator is defined as

$$\hat{f}_n(x) = \sum_{j=1}^{m} \frac{\hat{p}_j}{h_j} \mathbb{1}_{x\in B_j}. \tag{7}$$

To calculate the bias of the histogram estimator for continuous data, we first calculate its expected value. With F the cumulative density function corresponding to f, the expected value is:

$$\begin{aligned}
\mathbb{E}(\hat{f}_n(x)) &= \mathbb{E}\left(\frac{\hat{p}_j}{h_j}\right) = \frac{1}{h_j}\mathbb{E}\left(\frac{n_j}{n}\right) \\
&= \frac{1}{h_j} \cdot P(x_i \in B_j) \\
&= \frac{1}{h_j} \int_{a+\sum_{k=1}^{j-1} h_k}^{a+\sum_{k=1}^{j} h_k} f(u)du \\
&= \frac{F(a+\sum_{k=1}^{j} h_k) - F(a+\sum_{k=1}^{j-1} h_k)}{a+\sum_{k=1}^{j} h_k - \left(a+\sum_{k=1}^{j-1} h_k\right)} \\
&= f(x^*), \qquad (8)
\end{aligned}$$

for at least one $x^* \in B_j$ in the bin B_j containing the argument of the estimator $x \in B_j$. We obtained the last equality with the mean value theorem of differentiation. If the derivative of the PDF is bounded, $|f'(x)| \leq L$, (see appendix: 8.2.1), and M the width of the biggest bin, we obtain a bound for the bias with

$$\begin{aligned}
\textbf{bias}(\hat{f}_n(x)) = \mathbb{E}(\hat{f}_n(x)) - f(x) &= f(x^*) - f(x) \\
&= f'(x^{**}) \cdot (x^* - x) \\
&\leq LM, \qquad (9)
\end{aligned}$$

where used the mean value again from the first to second. Hence, we learn that smaller (and thus more) bins decrease the bias of the histogram estimator.
To now calculate the variance of $\hat{f}_n(x)$, we set h to be the width of the smallest bin. Further, we observe that we can model the number of data points in the bin B_j with a binomial random variable

$\mathscr{B}(n,p;x)$ with mean np and variance $np(1-p)$. Thus, we obtain:

$$\begin{aligned}\mathbf{var}(\hat{f}_n(x)) &= \mathbf{var}\left(\sum_{j=1}^{m}\frac{\hat{p}_j}{h_j}\mathbb{1}_{x\in B_j}\right)\\ &\leq \frac{1}{h^2n^2}\mathbf{var}\left(\sum_{j=1}^{m}n_j\mathbb{1}_{x\in B_j}\right)\\ &= \frac{1}{h^2n^2}\mathbf{var}\left(n_j\right)\\ &= \frac{P(x\in B_j)(1-P(x\in B_j))}{h^2n}.\\ &\leq \frac{P(x\in B_j)}{h^2n}.\end{aligned}$$

From equation (9), we know that for a bin of width h we have $P(x\in B_j) = p(x^*)h$. We choose h to be the bin with the smallest width. With again $|p(x)| \leq L$, we obtain

$$\mathbf{var}(\hat{f}_n(x)) \leq \frac{p(x^*)h}{h^2n} \leq \frac{L}{hn}. \tag{10}$$

We can conclude that variance decreases with sample size. Additionally, it can increase with shrinking bin sizes (and thus the number of bins).
We, therefore, have to consider a bias-variance-tradeoff when choosing bin widths and their number. If we jointly consider the bounds of the bias in equation (9) and variance in equation (10), we can conclude that the bias-variance-tradeoff is (at least in terms of bounds) optimized for equal-width bins.

The Histogram Estimator for Discrete Variables With regards to the later discussion of the transfer entropy measure, it is important to note that the discussed estimator is defined for continuous data. The division by the h_js in equation (7) serves the normalisation of the integral $\int_{\mathbb{R}}\hat{f}_n(x)dx = 1$ and in general $\sum_{j=1}^{m}\hat{f}_n(x\in B_i) \neq 1$. A histogram of discrete data requires the second normalisation. Subsequently, with the same binning setup the histogram is a map from the discrete set D to the binning:

$$\begin{aligned}&\mathscr{H}_{\text{Discrete}} : \mathbb{R}\supset D \to \{B_i\}_i\\ &x\mapsto B(x) = \begin{cases}B_1 & \text{if } x\leq \sup B_1\\ B_2 & \text{if } \inf B_2 < x \leq \sup B_2\\ & \vdots\\ B_m & \text{if } x > \inf B_m\end{cases}\end{aligned} \tag{11}$$

and the corresponding estimator for the discrete probability mass function is

$$\hat{P}_n(x) = \sum_{j=1}^{m}\frac{n_j}{n}\mathbb{1}_{x\in B_j}. \tag{12}$$

The properties of the discrete histogram estimator can be obtained similar to the continuous case. The expected value is

$$\mathbb{E}\left(\hat{P}_n(x)\right) = \mathbb{E}\left(\frac{n_j}{n}\right) = P\left(x\in B_j\right) = p_j$$

with p_j the probability of $x\in B_j$. Then the bias vanishes as

$$\mathbf{bias}\left(\hat{P}_n(x)\right) = \mathbb{E}\left(\hat{P}_n(x)\right) - P(x) = p_j - p_j = 0.$$

When we again utilize the fact that the probability of n values in B_j follows a binomial variable. With this knowledge we obtain

$$\mathbf{var}\left(\hat{P}_n(x)\right) = \frac{1}{n^2}\mathbf{var}\left(n_j\right) = \frac{p(x \in B_j)\left(1 - P(x \in B_j)\right))}{n^2} \leq \frac{P(x \in B_j)}{n^2} = \frac{p_j}{n^2} \leq \frac{1}{n^2}$$

since the probability mass function is bounded by 1. In the discrete case, the bias vanishes, and the variance decreases faster than for the continuous histogram estimator.

2.3 Data Discretization

In the last section (equation 1), we introduced the discretization method or binning without specifying optimal shape. Regarding the later discussion of transfer entropy ([**?**]), the chosen discretization map is essential. Therefore, this section presents a literature review of existing binning techniques. The author started the review with the methods that are implemented in standard statistic packages such as NumPy [32], or R [55]. He undertook significant efforts to obtain the primary source for binning methods. However, these sources sometimes lack a statistically rigorous justification, heuristic, or even explanation of the method. Warranted by their widespread adoption, we will discuss their performance nonetheless. In some cases, the original publications were not untraceable.
Since we evaluate binning as a data discretization problem, we extend our discussion beyond classical direct binning methods to include approaches to adjacent problems.; namely, discretization based on optimizing some statistical criterion and clustering methods where a subsequent partition of the data range yields bins. The latter two are not standard for histogram calculation. However, many applications of histograms do not venture beyond exploratory univariate data analysis. For these, the influence is negligible. In our case, however, the discretization is integral to the transfer entropy calculation (see section 3). Thus, a comprehensive analysis of methods addressing adjacent problems is warranted.

2.3.1 Direct Binning Methods

There have been numerous methods suggested to find the optimal binning. They take different approaches to fix the number m of equal-width-bins (where the highest and lowest sampled values bound the binned range) or find the width of bins h. However, they share the approach that sample characteristics yield the number or width of bins.

$\sqrt{n}$ Bins A statistical rule of thumb to choose the number of equal-width bins is

$$m = \lceil \sqrt{n} \rceil, \tag{13}$$

where n is the number of data points. Various software employs this easy to compute standard method to create histograms. For example, Excel [13].[1]

Sturges Formula Sturges formula [71], published in 1926, is the oldest attempt to systematically choosing bin widths. His approach was motivated by the notion of an ideal histogram for normal data [62]. The simplest discrete probability mass function (PMF), which for a large number of trials that converges to a Gaussian density, is the binomial distribution, $\mathscr{B}(n, p; j)$, with a probability of success $p = 1/2$. Consider a histogram with k bins labeled by $j = 0, \ldots, k-1$. The resulting binomial PMF,

$$\mathscr{B}(k-1, \frac{1}{2}; j) = \binom{k-1}{j}\left(\frac{1}{2}\right)^{k-1},$$

[1] The author could not obtain any publication within reputable scientific sources.

now yields the number of data points (i.e. the number of successes of $\mathscr{B}$) for the j^{th} bin. Respectively, if we scale $\mathscr{B}(k-1,\frac{1}{2};j)$ with the sample size, $\mathscr{B}(k-1,\frac{1}{2};j) \mapsto n \cdot \mathscr{B}(k-1,\frac{1}{2};j)$, we obtain the respective count for the $B_j{}^{\text{th}}$ bin, n_{B_j}.
By identifying n_{B_j} with the respective binomial coefficients, we obtain the total number of data points

$$n = \sum_{j=0}^{k-1} n_{B_j} = \sum_{j=0}^{k-1} \binom{k-1}{j} = (1+1)^{k-1} = 2^{k-1}$$

where we used the binomial theorem, $(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^n$ for $x = y = 1$.
From this expression we derive Sturges formula for the bin count m:

$$m = \lceil \log_2(n) + 1 \rceil. \tag{14}$$

However, this rationale requires the sample size n to be a power of two (though the error of breaking this assumption diminishes for large n). Additionally, it relies on the assumption of (approximately) normally distributed data that justifies identifying the binomial coefficients with bin counts. Nonetheless, other distribution shapes and resulting bin counts are entirely valid. Resultingly, it is not generalizable to the extent proposed by Sturges [36]. It is, however, still an option or even standard in popular statistical computer packages such as NumPy[32] or R [55].

Rice Rule The rice rule [41] is proposed as an alternative to Sturges' formula and proposes a higher number of bins. It reads [2]

$$m = \lceil 2\sqrt[3]{n} \rceil. \tag{15}$$

Doane's Formula Doane's formula [22] is a modification of Sturges formula aiming to better the results for non-normal data. It therefore incorporates the skewness of the data with the term

$$\log_2 \left(1 + \frac{|g_1|}{\sigma_{g_1}} \right)$$

where g_1 is the third moment (skewness),

$$g_1 = \frac{\sum_i (x_i - \mu)^3}{(\sum_i (x_i - \mu)^2)^{\frac{3}{2}}},$$

and

$$\sigma_{g_1} = \sqrt{\frac{6(n-2)}{(n+1)(n+3)}}$$

is the standard deviation of the third moment. If the data is not skewed, no correction is added to equation (14) as $g_1 = 0$. The full formula reads:

$$m = 1 + \log_2(n) + \log_2 \left(1 + \frac{|g_1|}{\sigma_{g_1}} \right). \tag{16}$$

The more the distribution departs from the symmetric normal, the more bins are added. However, the rate of this process decreases.

[2] Extensive literature research did not yield any systematical study, derivation, or justification of the rule. A proper discussion is even absent in the original publication [41] where the authors state it as the preferred alternative to Sturges' rule.

Scotts Normal Reference Rule Scotts normal reference rule [63] obtains the asymptotic optimal bin width h^* by minimizing the integrated mean squared error of the histogram estimator. This yields,

$$h^* = \frac{6}{\int_{\mathbb{R}} f'(x)^2 dx} \frac{1}{\sqrt[3]{n}}.$$

However, this approach requires knowledge of the underlying PDF $f(x)$ (and its derivative needs to exist and must be square-integrable). Therefore, the Gaussian density is proposed as a reference standard. This yields

$$h = \frac{3.49\hat{\sigma}}{\sqrt[3]{n}} \tag{17}$$

where $\hat{\sigma}$ is the standard deviation of the distribution of the data estimated from the sample.

Freedmann-Diaconis Rule Freedmann and Diaconis rule [27] is a modification of Scotts rule, which replaces the standard deviation dependent scaling factor with the interquartile range, IQR, of the sample $\mathscr{X}$:

$$h = 2\frac{\mathrm{IQR}(\mathscr{X})}{\sqrt[3]{n}}. \tag{18}$$

As the IQR is robust against outliers, they carry less influence in the bin width calculation. However, Freedman and Diaconis' approach performs suboptimally for multimodal densities as the derivation for equation (18) assumes characteristics of the underlying density. For example it requires $\int_{\mathbb{R}} f'(x)^2 dx > 0$. Subsequently, the method performs poorly, for example, for the uniform distribution [39].

Variable Bin Widths Until now, all methods assumed a constant width of bins h. One different approach is however, to choose the bin widths h_j such that the bins are (approximately) equiprobable, $p_j = {}^{n_j}/_{n} \approx {}^{1}/_{m}$.
An optimal number of bins m can is obtained with a criterion introduced by Agostino [17]:

$$m = 4\left(\frac{2n^2}{(\Phi^{-1}(\alpha))^2}\right)^{\frac{1}{5}}, \tag{19}$$

where $\Phi^{-1}(\alpha)$ is the inverse of the cummulative distribution function of the standard normal. α is the one-sided confidence. $\alpha = 0.05$ is a common choice.

2.3.2 Criterion Based Binning

Criterion-based binning methods are imposing a criterion on the binned data set. The bin edges are varied to optimize said criterion. These techniques often utilize a Bayesian approach. We will neglect a thorough discussion of Bayesian inference. Inclined readers can find them in textbooks such as [29]. In short: Bayesian statistics makes inferences about the model parameter θ given the sample $\mathscr{X}$ with Bayes theorem $p(\theta|\mathscr{X}) \propto p(\theta)p(\mathscr{X}|\theta)$. When necessary, we will briefly cover important terms in footnotes.

Knuth's Bins Knuth [39] utilizes a bayesian approach on the piecewise constant histogram model described in section 2.2.6 to optimize the bin width. The likelihood of the sample $\{x_i\}_i = \mathscr{X}$ to take on the specific values $\{x'_i\} = \mathscr{X}'$ follows a multinomial distribution with a different prefactor (which - of course - still integrates to 1). It is calculated as

$$p\left(\mathscr{X} = \mathscr{X}'|\vec{p},m,I\right) = \left(\frac{m}{V}\right)^n \prod_{k=1}^{m} p_k^{n_k} \tag{20}$$

with $\vec{p} = \{p_1, p_2, \ldots, p_{m-1}\}$ the probability mass of the m bins. The n_k are the respective number of data points. The probability of the m^{th} bin, p_m, is given by the normalization condition $p_m = 1 - \sum_{k=1}^{m-1} p_k$. The volume or range of the data V is given by $V = \sum_j h_j$. I is the prior knowledge about the problem. The n_j are the number of data points in the j^{th} bin and $n = \sum_j n_j$.
The author assumes a uniform PMF for the number of bins

$$p(m) = \begin{cases} \frac{1}{C} & \text{if } m \in [1, C] \\ 0 & \text{if } m \notin [1, C] \end{cases} \tag{21}$$

with C the maximum number of bins considered. Next, he chooses the Dirichlet distribution with parameter vector $\vec{\alpha}$ of uniform constituents $\alpha_i = 1/2$ as an non-informative prior. It reads

$$P(\vec{p}|m,I) = \frac{\Gamma\left(\frac{m}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^m} \left(\left(1 - \sum_{k=1}^{m-1} p_k\right) \prod_{k=1}^{m-1} p_k \right)^{-\frac{1}{2}} \tag{22}$$

The Dirichlet distribution is the Jeffreys' prior[3] to the multinomial likelihood [38, 31, 7]. Additionally, it is the conjugate prior[4] to the multinomial likelihood in equation (20). Hence, the posterior is also a Dirichlet distribution.
According to Bayes theorem, this posterior is proportional to the product of priors and the likelihood,

$$p(\vec{p}, m|\mathscr{X}, I) \propto p(\vec{p}|m,I)\, p(m|I)\, p(\mathscr{X}|\vec{p}, m, I).$$

Thus, with equations (20, 21, 22) and some algebra, the author obtains the following posterior:

$$p(\vec{p}, m|\mathscr{X}, I) \propto \left(\frac{m}{V}\right)^n \frac{\Gamma\left(\frac{m}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^m} \left(1 - \sum_{k=1}^{m-1} p_k\right) \prod_{k=1}^{m-1} p_k^{n_k - 1/2}.$$

To obtain the posterior probability in terms of the number of bins m, he integrates over all possible values[5] of $\vec{p}$. He arrives at

$$p(m|\mathscr{X}, I) \propto n \log m + \log\Gamma\left(\frac{m}{2}\right) - m\log\Gamma\left(\frac{1}{2}\right) - \log\Gamma\left(n + \frac{m}{2}\right) + \sum_{k=1}^{m} \log\Gamma\left(n_k + \frac{1}{2}\right). \tag{23}$$

Now the optimal number of bins is found by solving the optimization problem $\hat{m} = \arg\max_m p(m|\mathscr{X}, I)$ for which he provides an algorithm.
Knuth's method was previously applied to transfer entropy calculation [30].

Minimizing Cross-Validation The cross-validation estimator for a general density estimator $\hat{f}$ is defined as

$$\hat{J}(h) = \int_{\mathbb{R}} (f(x))^2\, dx - \frac{2}{n} \sum_{j=1}^{n} \hat{f}_{(-j)}(x_j). \tag{24}$$

Here, $\hat{f}_{(-i)}$ refers to the estimator $\hat{f}$ that is obtained from the sample if x_i were excluded [74]. Thus each term in the sum yields an density estimate for a datum which did not contribute to $\hat{f}$. Now equation (24) can be rewritten in terms of the bin width h:

$$\hat{J}(h) = \frac{2}{h(n-1)} - \frac{n+1}{h(n-1)} \sum_{j=1}^{m} \hat{p}_j^2.$$

Here, we sum over the probability masses p_j of the m bins. Similar to Scott's rule this approach seeks to minimize the cross-validation [68]. Subsequently the solution to the optimization problem $h^* = \arg\min_h \hat{J}(h)$ yields the asymptotical optimal bin width h.

[3] One important property of Jeffreys' prior $p(\vec{\theta})$ is its invariant under coordinate transformations of the parameter vector $\vec{\theta}$. In Knuths' model, the occupied probability volume of the elements in $\vec{p}$ is thus independent of the coordinates.

[4] Distributions are conjugate if both, the prior and posterior are within the same family of distributions; they exhibit the same structure and might only differ in the specific values of parameters [56].

[5] That is all vectors $\vec{p}$ in the $(n-1)$ dimensional probability simplex [10].

Maximum Likelihood of the Cross Validation Chow, Geman, and Wu [16] proposed a different approach based on cross-validation. They define the histogram estimator in terms of the bin width h as

$$\hat{f}_{h,n}(x) = \frac{1}{hn} \sum_{i=-\infty}^{\infty} \mathbb{1}_{[h(i-1),hi)}(x) \left(\sum_{j=1}^{n} \mathbb{1}_{[h(i-1),hi]}(x_i) \right)$$

where $\mathbb{1}_{\mathscr{A}}(x)$ is the indicator function for the set $\mathscr{A}$ that equals unity if $x \in \mathscr{A}$ and vanishes otherwise. The optimal choice of h is now found by maximizing the likelihood-like expression

$$L_h = \prod_{i=1}^{n} \hat{f}^{i}_{h,n-1}(x_i)$$

where $\hat{f}^{i}_{h,n-1}$ is the histogram estimator exclusive of the data point x_i. The choice for the bin width is given by the resulting optimization problem $h^* = \arg\max_h L_h$.

Shimazaki and Shinomoto squared Error Minimization Shimazakis and Shinomotos method [65] was initially introduced for time histograms to display changes in rates of events. One remnant of these intentions is modeling the number of data points in a given bin by a Poisson distribution. However, we can apply this method can also for data independent of time. It functions by minimizing the integrated mean squared error (MISE) for the density estimator $\hat{f}$:

$$\text{MISE} = \frac{1}{b-a} \int_a^b \mathbb{E}\left(\hat{f}(x) - f(x)\right)^2 dx. \tag{25}$$

where $b-a$ is the range of the sampled data. By now partitioning this range equation (25) can be rewritten in terms of the m bins of size h:

$$\text{MISE} = \frac{1}{h} \int_0^h \left(\frac{1}{m} \sum_{j=1}^{m} \mathbb{E}\left(\hat{\Theta}_j - f(x+(j-1)h)\right)^2 \right) dx \tag{26}$$

where $\hat{\Theta}_j = {}^{n_i}/_{nh}$ is the empirical height of the j^{th} bin. We now denote the average over $f(x+(j-1)h)$ as an ensemble average on $[0,h]$ to rewrite and decompose equation (26) to

$$\begin{aligned}
\text{MISE} &= \int_0^h \langle \mathrm{E}\left(\hat{\Theta}(x) - f(x)\right)^2 \rangle \\
&= \langle \mathbb{E}\left(\hat{\Theta} - \Theta\right)^2 \rangle + \frac{1}{h} \int_0^h \langle (f(x) - \Theta)^2 \rangle dx. \\
&= \langle \mathbb{E}\left(\hat{\Theta} - \Theta\right)^2 \rangle + \frac{1}{h} \int_0^h (f(x) - \langle \Theta \rangle)^2 dx - \langle (\Theta - \langle \Theta \rangle)^2 \rangle.
\end{aligned}$$

where we added $0 = \langle \Theta \rangle - \langle \Theta \rangle$ from the second to third line and observe that the outer ensemble average under the integral vanishes. We now subtract the integral term from the MISE to obtain the following cost function:

$$C_n(h) = \langle \mathbb{E}\left(\hat{\Theta} - \Theta\right)^2 \rangle - \langle (\Theta - \langle \Theta \rangle)^2 \rangle. \tag{27}$$

We now find that $\mathbb{E}(\hat{\Theta}) = \Theta$ for an unbiased estimator. Thus

$$\langle \mathbb{E}(\hat{\Theta} - \langle \mathbb{E}\left(\hat{\Theta}\right) \rangle)^2 \rangle = \langle \mathbb{E}\left(\hat{\Theta} - \Theta\right)^2 \rangle + \langle (\Theta - \langle \Theta \rangle)^2 \rangle.$$

Resultingly, we can rewrite equation (27) as

$$C_n(h) = 2\langle \mathbb{E}\left(\hat{\Theta} - \Theta\right)^2 \rangle - \langle \mathbb{E}\left(\hat{\Theta} - \langle \mathbb{E}\left(\hat{\Theta}\right) \rangle\right)^2 \rangle.$$

Now we use the assumption that the probability of data in a bin follows a Poisson distribution. Hence the variance and mean of the number of data points in bin j, n_j, is equal. Thus we get the mean-variance relation $\mathbb{E}\left(\hat{\Theta}-\Theta\right)^2 = {}^{\mathbb{E}(\hat{\Theta})}/_{nh}$.
We then finally obtain

$$C_n(h) = \frac{2}{nh}\mathbb{E}\left(\hat{\Theta}\right) - \langle \mathbb{E}\left(\hat{\Theta} - \langle \mathbb{E}\left(\hat{\Theta}\right)\rangle\right)^2\rangle$$

for the cost function. In this method the solution to the optimization problem $\arg\min_h C_n(h) = h^*$ gives the optimal bin widths, h^*. Therefore, $\mathbb{E}\left(\hat{\Theta}\right)$ is replaced by the the expectation value of the bin count given by the sample partition. Hence, we optimize

$$C_n(h) = \frac{2\bar{n}_j - \sigma_{n_j}}{(nh)^2} \tag{28}$$

where $\bar{n}_j$ is the mean data count of the bins, and σ_{n_j} the respective variance.

Akaike Information Criterion Akaike's information criterion (AIC) [2] scores the goodness of a model considering their different number of parameters. Therefore AIC value has to be minimized for each candidate. The lowest score marks the best choice. AIC can be applied to select histogram bin width [72]. The AIC is a function of the maximum likelihood L and defined as

$$\mathrm{AIC}(L) = -2\ln(L) + 2k, \tag{29}$$

where k is the number of independent parameters of the model. $k = m$ in the case of a histogram with m bins. The likelihood of the histogram is

$$L = \prod_{i=1}^{n} \hat{f}(x_i) = \prod_{j=1}^{m} p_j^{n_j} = L(p),$$

where p_i is the probability mass of the j^{th} bin and n_j the number of data points inside. We now employ the constraint $p_i \geq 0$ and obtain for the bin width h the condition $h \cdot \sum_j p_j = 1 \Rightarrow \sum_j p_j = {}^1/_h$. We then maximize L with $p_j = {}^{n_j}/_{nh}$. Therefore, we must choose m, h to minimize

$$m - \ln\left(\prod_{j=1}^{m}\left(\frac{n_j}{nh}\right)^{n_j}\right) = m + n\ln(n) + n\ln(h) - \sum_{j=1}^{m} n_j \ln(n_j). \tag{30}$$

If the histogram is based on a small sample, the corrected AIC yields better results [75]. To account for a small sample, a first order bias correction is multiplied to the model dimension:

$$\mathrm{AIC}_c(L) = -2\ln(L) + \frac{2k}{{}^n/_{n-k-1}}. \tag{31}$$

Bayesian Information Criterion Based on bayesian reasoning Schwarz [61] proposed to scale the model dimension term (in our case the number or bins m) of the AIC in equation (29) by ${}^{\ln(n)}/_2$. One must therefore minimize

$$\frac{m}{2}\ln(n) + n\ln(n) + n\ln(h) - \sum_{j=1}^{m} n_j \ln(n_j). \tag{32}$$

2.3.3 Clustering Based Binning Methods

Clustering algorithms serve to cluster data into distinct groups which are similar within. Clustering approaches differ from the above methods because they group data bottom-up, whereas binning introduced the top-down bin edges. However, this distinction is not sharp since sample

characteristics such as the IQR, n_j,σ influence the binning. Nevertheless, this influence is to a much lesser extent than in clustering methods. Though the strengths of clustering algorithms are more apparent in higher dimensions, several authors employed them for binning on the number line, for example [48], [50], and [73]. We translate the clusters into bins by utilizing a Voronoi partition [11] of the data range.

k-Means Clustering Generally, k-means is an algorithm that clusters data into k clusters. Researchers usually apply the algorithm for multidimensional feature spaces. For our application, only the one-dimensional case is relevant. To stay consistent with the literature, we will discuss the multidimensional procedure nonetheless. The algorithm in its current form was proposed by Lloyd [43], but I will follow the instructive explanation laid out by Mackay [45] in this chapter.

The first thing needed for the k-means algorithm is a distance measure. We will use the simple squared euclidean distance:

$$d(\vec{x},\vec{y}) = \sum_i (x_i - y_i)^2 .$$

The K clusters are parametrized by their respective means m_k, the centers of the clusters. In the beginning, we somehow initialize them. If possible, one adjusts the according to existing knowledge in order to value to speed up convergence. Alternatively, we apply random initialization. Then follows the *assignment* step, where each data point x_i is assigned to the cluster K with the nearest mean $d(m_k,x_i) \leq d(m_{k'},x_i)$. in case of $d(m_k,x_i) = d(m_{k'},x_i)$ the data is assigned to the cluster with fewer attached data. Should that be equal, too, one could resort to random assignment between the cluster candidates. After the assignment the *update* step follows. This means that the means m_k are recalculated

$$m_k(x) = \frac{\sum_n r_k^n x^{(n)}}{R_k}. \tag{33}$$

Here the $r_k^{(n)}$ are the cluster assignment function where $R_k = \sum_n r_k^{(n)}$ and $r_k(x) = 1,0$ depending on whether m_k is the closest mean to x or not. Now the reassigment and subsequent recalculation of the means are repeated until the assignments do not change anymore. We subsequently apply a Voronoi partition to generate bin edges from the cluster assignment to the data.

Expectation Maximization Clustering with Gaussian Mixture Model The expectation-maximization algorithm is based on fitting a family of densities $g(x|\Phi)$ to the data $x \in \mathscr{X}$ where Φ are the parameters of the densities [21]. We now iteratively repeat two steps for the algorithm: the expectation and the maximization step applied with the model functions. Generally we consider a probability $P(\mathscr{X},\mathscr{Z}|\Phi)$ of the measured data $\mathscr{X}$ and the unknown values $\mathscr{Z}$ (the cluster assignment). Since Φ is unknown, it may be initialized randomly or by guessing. $\mathscr{Z}$ is also unknown. Hence the maximization of the marginal probability $P(\mathscr{X}|\Phi)$ serves as a proxy. Now in the t^{th} expectation step one calculates

$$Q(\Phi|\Phi^{(t-1)}) = \mathbb{E}_{\mathscr{Z}|\mathscr{X},\Phi^{(t-1)}}\left(P(\mathscr{X}|\Phi)\right). \tag{34}$$

With the maximization step we subsequently calculate $Q(\Phi^{(t)}|\Phi^{(t-1)}) = \arg\max_\Phi Q(\Phi|\Phi^{(t-1)})$. We repeat these two steps until conversion. This model is similar to k-means in the sense that the fitted density parameters are iteratively updated. However, it differs since we are only calculating probabilities here, whereas, in k-means, cluster affiliation is a definite criterion.
In the scope of this thesis, we will apply Gaussian densities because of expected white noise on data and the existing implementation in python [52]. As a result, we obtain the final binning by a Voronoi partition between adjacent points assigned to different clusters.

DBSCAN Clustering The acronym DBSCAN stands for the density-based spatial clustering of applications with noise [24]. The algorithm relies on the following definitions: (1) the data point p is a core point if at least `minPts` points are within an ε neighborhood. (2) a point q is said to be *directly reachable* from p if it is within the distance ε. It is (3) *reachable* if there is a chain of subsequently direct reachable points from p to q. (4) We classify points that are not reachable as outliers.

The algorithm does not require a specified number of clusters. Nonetheless, it depends on the parameters `minPts` and ε. Then the core points are determined. We then group core points within a cluster. Non-core points are assigned to clusters if they are reachable from the cluster points. We classify leftover points as outliers. Concerning our binning method, we apply a Voronoi partition to bin the different clusters. Doing this, we regard Outliers as a separate cluster.

Mean Shift Clustering The mean shift algorithm was first proposed in 1975 [28]. It is a mode-seeking algorithm that works by shifting the means of clusters using a kernel function $K(x)$ [14]. In this thesis, we will apply a flat kernel for clustering. For the algorithms iterated steps, we compute the mean of the data for a sample subset $\mathscr{S} \in \mathscr{X}$, where $\mathscr{S}$ is a scaled (hyper-) sphere of the same dimension as $\mathscr{X}$. Now the mean is calculated as

$$m(x) = \frac{\sum_{s_i \in \mathscr{S}} K(s_i - x) s_i}{\sum_{s_i \in \mathscr{S}} K(s_i - x)}. \tag{35}$$

In the subsequent mean shift step we change $\mathscr{S}$ to be centered at $m(x)$. We repeat these two steps until convergence. To apply this mode-seeking algorithm in clustering, initialize the mode-seeking step at every data point in the sample, and then, the means are simultaneously updated. Naturally, upon convergence, some data points might be assigned to multiple subsets. To achieve a unique classification, we assign data points to the cluster which contains the most points. The diameter of the neighborhoods $\mathscr{S}$ is a tuning parameter of the model commonly referred to as bandwidth.

Agglomerative Hierarchical Clustering Agglomerative hierarchical clustering [19] is a bottom-up hierarchical clustering method. It starts with every data point assigned to its cluster. Subsequently, we merge the two clusters that minimally increase the total in-cluster-linkage distance in each iteration step. We can apply several norms to determine the linkage distance. In this thesis, we used Euclidean distance. The algorithm terminates if it attains a preset number of clusters.

2.4 Bivariate Relationships between Samples

If one obtained two or more time series, one is interested in their pairwise relationship. In this chapter, we will cover the bivariate measures applied in this thesis. These measures differ in the type of relationships between variables they can capture.
We can understand the relationship between two variables as the following mapping Φ from one variable V_1 to another V_2:

$$\begin{aligned} \Phi : V_1 &\to V_2 \\ x_1 &\mapsto x_2. \end{aligned}$$

The empirical measurements will contain an additional noise contribution that subsumes random or systematic measurement errors and influences onto V_2, which originate outside the studied bivariate system.
The map Φ can be linear or non linear. In the first case it fulfils additivity, $\Phi(u+v) = \Phi(u) + \Phi(v)$ and homogeneity, $\Phi(c \cdot u) = c \cdot \Phi(u)$. If Φ does not fulfil these properties, the relationship between

V_1 and V_2 is non linear.

2.4.1 Linear Measures

Linear measures quantify the relation between two time-ordered samples. Thus, they implicitly assume a linear relationship between the variables. However, this assumption includes the notion that small changes in one variable correspond to small samples in the other. This circumstance is different for non-linear measures that we will discuss in the subsequent section.

Product-Moment Correlation The *Product-Moment Correlation* or *Bravais-Pearson-Correlation* is a standard measure to quantify the association between samples of variables. It is a symmetric and linear dependency measure. We can calculate the correlation between variables $\mathscr{X}$ and $\mathscr{Y}$ as

$$\rho_{xy} = \frac{\sum_{j=1}^{n}(x_j - \mu_x)\cdot(y_j - \mu_y)}{\sqrt{\sum_{j=1}^{n}(x_j - \mu_x)^2}\sqrt{\sum_{j=1}^{n}(y_j - \mu_y)^2}} \tag{36}$$

for samples of length n and μ_X, μ_Y the means of the sample of variables X and Y respectively. As such, ρ_{XY} is the ratio of the covariance by the product of the constituent variables standard deviations. Therefore it requires metric scales. The measure is normalized to the interval $[-1,1]$. Positive values indicate a directional temporal evolution of the two variables, whereas negative values indicate anti-directional evolution. In essence, ρ is a non-parametric measure and only requires the variance and covariance of the variables to exist, which is true for finite datasets. However, the measure only exhaustively captures the association if a normal distribution of X and Y is assumed. It does not capture non-linear relationships [9].

Spearmans Rho Dissimilar to the Product-Moment-Correlation *Speaman's* ρ is defined for ordinal scaled variables as well. Though in literature also referred to with the letter ρ, we will use P for Spearman's ρ to distinguish both quantities. P is the Pearson correlation the rank variables r_X, r_Y of the samples $\mathscr{X}$ and $\mathscr{Y}$. Therefore it is symmetric and linear as well. When we apply P to metric data, the ranking is a particular choice of the information destroying mapping φ^D (see equation 2). P reads

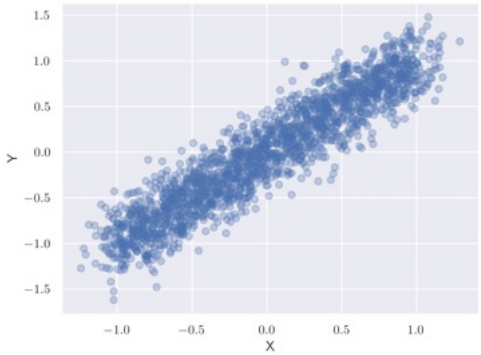
$$\mathrm{P} = \rho^{\mathrm{Pearson}}_{r_X, r_Y} = \frac{\mathrm{cov}\,(r_X, r_Y)}{\sigma_{r_X}\sigma_{r_Y}} = 1 - \frac{\sum_{j=1}^{n} d_j}{n(n^2-1)}, \tag{37}$$

where $d_j = r_X(x_j) - r_Y(y_j)$ is the rank difference. The last equality only holds if the ranks are integers.

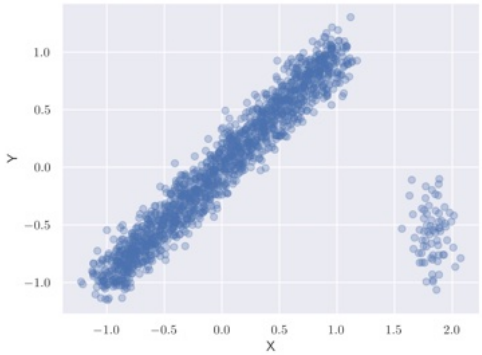
In figure (2.1) we see the values of ρ and P for different datasets. In (2.1a), we see that both measures are consistent for the linear association. However, ρ is more sensitive to outliers. Both measures cannot quantify the nonlinear properties of data. That is evident in figure (2.1c). Though, if the variables are monotonically but nonlinearly related, P exhibits a higher association than ρ due to the rank mapping.

2.4.2 Non-Linear Measures

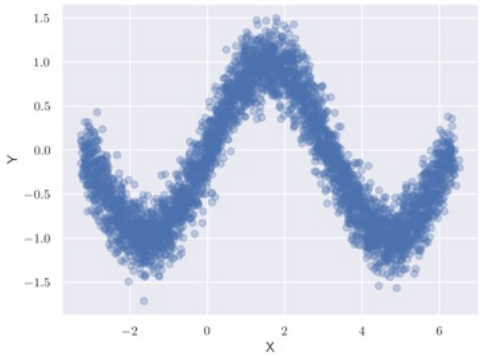
The nonlinear measures we utilize in this thesis originated in information theory. We will conduct a more comprehensive derivation of the measures in section (3), particularly (3.3). These measures are non-parametric and capture linear and nonlinear associations as they quantify the information shared between variables. The first measure we are introducing is mutual information.



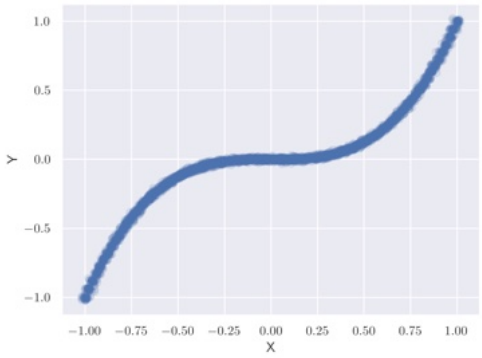
(a) Linear relation: $\rho = .94$, P $= .94$



(b) Linear relation with outliers: $\rho = .69$, P $= .81$



(c) Non-linear sine relation: $\rho = 0.00$, P $= 0.00$



(d) Non–linear cubic relation: $\rho = .92$, P $= .99$

Figure 2.1: Values for Pearsons ρ and Spearmans P for different datasets. For linear relations both measures produce equal results (2.1a). The Sperman correlation is less influenced by outliers (2.1b). Both measures fail to correctly quantify non-linear relations (2.1c and 2.1d).

Mutual Information Mutual information, MI, is a symmetric measure to quantify the amount of information about a variable X that is encoded in variable Y. MI is given by

$$MI(X,Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X,Y). \tag{38}$$

MI is a symmetric measure and quantifies the amount of information one knows about one constituent variable of the measure if one knows the other one. In this thesis, we will not comprehensively cover MI. However, we will discuss the measure since we can understand transfer entropy, the primary method in this thesis, as MI conditioned onto the variables pasts.

Transfer Entropy Other than the previously discussed measures, transfer entropy, TE, explicitly requires time series data. It was initially introduced by Schreiber [60] as an asymmetric measure capable of distinguishing drive and response in coupled time series systems. As stated above, we can understand TE as conditional mutual information of the influenced time series Y_t and the lagged time series $X_{t-1:t-K}$ conditioned on the history of the influenced time series $Y_{t-1:t-L}$. Therefore, we can rewrite TE in terms of entropies. The resulting formula for empirical measurements $x_t \in \mathscr{X}_t$, $y_t \in \mathscr{Y}_t$ reads

$$TE_{X\to Y} = \sum_{t=1}^{T} p(y_t, y_{t-L:t-1}, x_{t-K:t-1}) \log\left(\frac{p(y_t|y_{t-L:t-1}, x_{t-K:t-1})}{p(y_t|y_{t-L:t-1})}\right) \tag{39}$$

$$= H(Y_t|Y_{t-L:t-1}) - H(Y_t|Y_{t-L:t-1}, X_{t-K:t-1}), \tag{40}$$

where p is the PMF of the variables. Usually, p is unknown and estimated from the data, for example, by methods introduced in section (2.2.6). One part of this thesis is the development and evaluation of an adapted estimator. This method is derived in in section (4) and evaluated in section (5.7).
K and L are the lags of the time-series X_t, Y_t (or their samples, $\mathscr{X}_t, \mathscr{Y}_t$, respectively). These lags determine how much of the past the TE s into account for the TE calculation. Often one chooses $K = L$.

Non-Linearity Quantification To quantify the amount of non-linear information flowing between X and Y, we must deduct contributions from the linear association. For the non-linear measure η, we define the η-non linearity measure as

$$\eta_{\mathrm{NL}}(X,Y) = \min\left(\frac{\eta(X,Y) - \langle\tilde{\eta}(X,Y)\rangle}{\eta(X,Y)}, 0\right). \tag{41}$$

We have $\eta = \mathrm{TE}$. The tilde refers to the value of the measure η for surrogate data, that is, the preprocessed data which preserves linear but destroys nonlinear association between the variables X, Y (see section 2.5.3). The bar refers to the mean over several iterations as random nonlinearities may occur in the surrogatization process. For this thesis we calculate η_{NL} with 25 surrogate samples.

2.5 Data Pre-Processing

In order to calculate the measures introduced above, we may preprocess the data with the methods discussed in this chapter.

2.5.1 Rank-Ordered Remapping

The rank-ordered remapping algorithm maps a source Y time series onto a target time series X with a specific distribution. We perform a rank-ordered remapping by applying three steps:

1. obtain an ordered source distribution either by random sampling or utilizing an empirical distribution of equal length to the target.

2. apply a rank mapping onto both time series. If the values are not unique, we add slight noise to the variables to implement a random order of duplicate values.

3. reorder the source time series such that the rank time series of X, Y are identical.

Now the source time series has similar evolution in time as the target but follows the specified distribution. A remapping of a nonlinear target time series with a gaussian source destroys static nonlinearities caused by nonlinear measurement functions $\mathscr{M}$ if subsequently the reordered source is used. We can, therefore, purely test for dynamic nonlinearities.

2.5.2 Rescaling

Rescaling is a remapping (shifting, squeezing, or stretching) of the data onto a specified interval that preserves the relative shape of the distribution. This approach allows to pass data into the same range and thus easily apply identical discretization or other operations to different data. Rescaling the sample $\mathscr{X}$ onto a range R is conducted by extracting the rescaled value x' of the value x for every $x \in \mathscr{X}$ in the sample via

$$x' = \frac{x - \min(\mathscr{X})}{\max(\mathscr{X}) - \min(\mathscr{X})} \cdot (\max(R) - \min(R)) + \min(R). \tag{42}$$

2.5.3 Surrogates

Data surrogatization aims to destroy a data set's non-linear properties while preserving all other (linear) ones, namely, the data distribution, autocorrelation function, and power spectrum.

Bootstrap Surrogates Bootstrap surrogatization is the simple act of shuffling the time series sample. Therefore the real-space distribution is trivially conserved. However, autocorrelation and, therefore, by the Wiener-Khinchin-theorem, the power spectrum is destroyed.

Fourier Transform Surrogates Fourier Transform (FT-) surrogates are a standard surrogatization method. They adapt the data by phase randomization within the frequency domain. To generate them, one applies the following steps [58]:

1. We map the sample $\mathscr{X}$ of the variable X into the frequency domain with a FT. The linear properties are stored within the amplitude, non-linear ones within the phases.

2. The algorithm destroys the non-linear properties by adding uniform random numbers drawn from the phases from the $[0, 2\pi]$ interval to the phases.

3. Subsequently, we apply the inverse FT to the adapted data in the frequency domain and obtain surrogate data within real space.

We repeat and subsequently average the surrogatization process in order to obtain stable results.
This surrogatization method preserves the power spectrum and, therefore, by the Wiener-Khinchin-theorem also the autocorrelation. In contrast, phase randomization destroys the non-linear properties. However, the method does not reproduce the time domain distribution.
One can address this drawback by employing Amplitude Adjusted FT Surrogates. These differ from FT surrogated by initially applying a remapping onto a gaussian distribution, the subsequent execution of the FT surrogatization algorithm, and a final remapping onto the initial distribution. However, the final remapping has an unknown effect on the phase randomization and whitens the

power spectrum. One may address the whitening by Iterative Amplitude Adjusted FT surrogatization, which stores the Fourier coefficients after an initial FT of the sample, subsequently shuffles the time domain data, performs a FT of which the coefficients are replaced by the initially stored to force the desired power spectrum. Then an inverse FT is applied, and the result is remapped onto the initial distribution to force the initial sample PDF. This process is iterated from the shuffling step until it achieves convergence or a maximum iteration number. While this process preserves autocorrelation, power spectrum, and time-domain distribution, the remapping effect on the phase randomization is unknown. Both adapted FT surrogatization methods might introduce artificial nonlinearities through the backdoor. Therefore we will only apply simple FT surrogates within this thesis.

2.5.4 Sliding Windows

We calculate the introduced measures for the whole time series. However, we assume the distributions of the variables to be time-dependent (section 2.1.1). To study the temporal evolution of these measures, we partition the whole sample $\mathscr{X}$ into time slices $\mathscr{X}_{t-t_0:t}$ of size(s) t_0. These slices can be of a fixed or variable size. The latter can occur if we want to study the measure on a fixed time interval (e.g., daily), but the number of data points within a day varies. The whole procedure yields a new time series of the measure of interest with downsampled time resolution. For fixed-size sliding windows, we start the measure calculation for the values with indices $[0,t_0]$, then for $[1,t_0+1]$ and so on. The resulting measure time series is of length $n-t_0$. Thus subsequent windows largely overlap, and the resulting time series provides a detailed picture of the measure evolution. One may apply bigger step sizes for computation-intensive measures for large samples. However, the step sizes remained $\leq t_0$. In the context of this thesis, where we do not mention step size, we utilized unit steps.

2.6 Data Post-Processing

After calculating all the measures, we want to evaluate the results further and apply data post-processing methods from network theory. Of course, this field and its methods go far beyond what we will discuss, but we will limit the discussion to the aspects applied in this thesis.

2.6.1 Networks

A *network* is a tuple containing two sets. One set contains the nodes or vertices of the network, the other the links or edges that connect the nodes [4]. These edges can have a weight associated with them. Additionally, they can be undirected if the direction of a link is symmetric or directed if the weight of the edge connecting node i and j is different from the one connecting j to i, $w_{ij} \neq w_{ji}$. Given the asymmetry of the TE measure, we will use directed networks in our analysis.

A multitude of network statistics can be calculated. However, we are mainly interested in network cohesion and thus will only utilize the link density:

$$\langle L \rangle = \frac{\sum_{w_{ij}}}{N \cdot (N-1)} \tag{43}$$

with $w_{ij} \in [0,1]$ the normalized weights between the N nodes of the network. As such, $\langle L \rangle$ quantifies the fraction of actual by possible connections within the network.

3 Development of the Transfer Entropy Measure

This chapter will cover the fundamental underpinnings of information theory to derive the transfer entropy measure. Therefore we will start with the information entropy and attempts to proper generalizations to continuous variables. Subsequently, we will introduce the Kullback-Leibler divergence as a measure to quantify (dis-) similarity of probability mass functions. We will then use this measure to derive mutual information and transfer entropy.

3.1 Information Entropy

The information entropy was introduced by Claude Shannon [64] as a measure for the information contained within a set of measurements $\mathscr{X} = \{x_1, \ldots, x_n\}$. Each value has an associated probability mass or density, depending on whether we sample measurements from discrete or continuous variables. We write $p(x_i) = p_i$. Therefore, every value has an associated information content or surprisal

$$I(E_i) = \log_a(p_i). \tag{44}$$

The base a of the logarithm determines the units we measure information in. It is often set to $a = 2$, corresponding to bit units. Having $a = 10$ would correspond to digits and $a = e$ to nats.

For discrete variables, the Shannon Entropy is a measure that fulfills the following desiderata. The quantity

1. is continuous in the p_i,
2. if the events or measurements occur with uniform probability, $p_i = 1/n \; \forall p_i$, the measure is a monotonic increasing function of their number, and
3. it should the events be split into successive events the quantity is a weighted sum of individual entropies.

The only measure satisfying these three conditions is

$$H(X) = -K \cdot \sum_{i=1}^{n} p_i \log_a p_i. \tag{45}$$

with $K > 0$, a constant. We can interpret it as the average surprise of realizing a certain x_i from the random variable X in the measurement. Shannon chose K to be unity, and we will continue with this convention for now but later adapt it for normalization purposes. Additionally, we will drop the base of the logarithm a in the notation for simplicity and use $a = e$ throughout this thesis.

When we index X with i and Y with j, the information entropy can be generalized to a joint probability for joint distributions as

$$H(X,Y,\ldots) = -\sum_{i,j,\ldots} p_{i,j,\ldots} \log(p_{i,j,\ldots}). \tag{46}$$

It follows that $H(X,Y) \leq H(X) + H(Y)$.
Similarly, we obtain conditional entropy

$$H(X|Y) = -\sum_{i,j} p_{ij} \log(p_{i|j}). \tag{47}$$

Therefore, $H(X,Y) = H(Y) + H(X|Y)$ and

$$H(X) + H(Y) \geq H(X,Y) = H(Y) + H(X|Y) \tag{48}$$

$$\Rightarrow H(X) \geq H(X|Y). \tag{49}$$

These measures quantify the joint or conditional surprisal of the measured variable.

3.2 Entropy Estimation of (Quasi-) Continuous Distributions

There are several attempts to generalize the above definition (45) to continuous variables. However, these generalizations exhibit specific problems, which we will cover in this section.

3.2.1 The Differential Entropy

A naive generalization of the entropy to continuous variables is the differential entropy of the probability density function (PDF), $f(x)$ of a random variable:

$$h[f] = -\int_{\mathbb{R}} f(x)\log f(x)dx. \tag{50}$$

It is obtained by simply substituting the summation in equation (45) with integrals (and using $K = 1$). However, this generalization lacks certain properties of the discrete Shannon entropy.

Negativity For instance, the differential entropy of an exponential random variable with PDF

$$\mathrm{Exp}(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

is calculated as

$$h[\mathrm{Exp}(\lambda)] = 1 - \ln(\lambda)$$

which becomes negative for $\lambda > e$. When we understand entropy as an average of surprisal within a set of events, it is unclear how we can interpret a negative differential entropy since it is an average of the suprisals of the events in a distribution. These surprisals have a lower bound of 0.

The Continuum Limit of the Shannon Entropy Another problem of the differential entropy is that it is not a valid extension of the Shannon Entropy [37], because of the following circumstance. If one discretizes a continuous random variable on the support $[a,\, b]$ via binning with bins of equal[6] size Δ, the mean value theorem yields that there is an x_i in each bin $[a+(i-1)\Delta,\, a+i\Delta]$ such that

$$f(x_i)\Delta = \int_{a+(i-1)\Delta}^{a+i\Delta} f(x)dx.$$

We can now associate a probability $p_i = f(x_i)\Delta$ to each bin on the support of $f(x)$. We additionally know that $f(x) \geq 0$, since f is a PDF. If we let terms with $f(x_i) = 0$ vanish, we can define the following extension of equation (45):

$$H^{\mathrm{Binned}}[f] \equiv -\sum_i f(x_i)\Delta\log\left(f(x_i)\Delta\right),$$

where we sum over all bins i. When we now let $\Delta \to 0$:

$$\begin{aligned} H^{\mathrm{Binned}}[f] &\to -\log(\Delta) - \int_{\mathbb{R}} f(x)\log f(x)dx \\ &\neq h[f]. \end{aligned} \tag{51}$$

[6] We use equal size bins here for simplicity. The result is the same for differently sized bins as long as these bins approach the limit of zero width equally fast. The equivalence can be shown by setting Δ to the smallest and biggest bin and observing that the resulting expression must respectively be greater and smaller than the limit of equation (52).

Here we have additionally the term $-\log(\Delta) \to -\infty$ in the limit of $\Delta \to 0$. We can therefore conclude that the differential entropy is not the continuous limit of the Shannon Entropy. However, both quantities are connected via the relation:

$$h[f] = \lim_{\Delta \to 0} (H^{\Delta} + \log(\Delta)). \tag{52}$$

Opposed to the expression in equation (51), the formula (52) remains finite in the continuum limit, $\Delta \to 0$.

3.2.2 The Limiting Density of Discrete Points

There are, however, further problems with both the differential entropy and the modification introduced above. Namely, they are not invariant under parameter changes. Jaynes [37] utilizes a formalism labeled the limiting density of discrete points to derive an invariant expression.

Suppose, we sample increasingly more points x into our dataset, $\mathrm{D} = \{x_i\}_i$, with $|\mathrm{D}| = n$. In the limit $n \to \infty$ the density of points then converges to the measure function $m(x)$. That is

$$\lim_{n \to \infty} \frac{1}{n} |\{x | x \in \mathrm{D} \cap [a,b]\}| = \int_a^b m(x) dx$$

If this limit is sufficiently well behaved, differences of adjacent data points $x_{i+1} - x_i$ behave as

$$\lim_{n \to \infty} n(x_{i+1} - x_i) = \frac{1}{m(x_i)} \tag{53}$$

and the discrete probabilities p_i of (binned) data points in D will go over into a continuous PDF $f(x)$ with

$$f(x_i) = \frac{p_i}{(x_{i+1} - x_i)}.$$

When we rearrange the terms and use equation (53), we obtain

$$p_i \to \frac{f(x_i)}{n m(x_i)}$$

as a continuum limit of the respective bin probabilities. We now use equation (53) again to obtain the differential

$$\lim_{n \to \infty} \frac{1}{n m(x)} = \lim_{n \to \infty} x_{i+1} - x_i = dx$$

and find the Shannon Entropy (45) continuum limit of

$$H^{\mathrm{Discrete}}[f] \to -\int_{\mathbb{R}} f(x) \log\left(\frac{f(x)}{n m(x)}\right) dx.$$

This expression contains the divergent term $\log n \to \infty$. However, if we subtract this term, we obtain the finite expression:

$$H^{\mathrm{Continuous}}[f] \equiv \lim_{n \to \infty} H^{\mathrm{Discrete}}[f] - \log n = -\int_{\mathbb{R}} f(x) \log\left(\frac{f(x)}{m(x)}\right) dx. \tag{54}$$

If we now perform a parameter change of the PDF and the measure function $m(x)$,

$$f_y(y) dy = f(x) dx$$
$$m_y(y) dy = m(x) dx,$$

then equation (54) transforms to

$$H^{\text{Continuous}}[f] = -\int_{\mathbb{R}} f_y(y) \log\left(\frac{f(y)}{m_y(y)}\right) dy. \tag{55}$$

From this expression, we can understand why it was necessary to introduce the measure function $m(x)$. The quantity keeps the entropy invariant. If we do not include it, the parameter change will introduce additional factors within the logarithm argument.

Still, the expression (55) depends on the PDF of the variable, which is usually unknown.

3.3 From Relative Entropy to Transfer Entropy

It is the purpose of data characterizing measures to quantify and compare data characteristics. The relative entropy or Kullback-Leibler (KL) divergence is one such measure in information theory. It quantifies the dissimilarity of two distributions of random variables X and Y with probability masses $P_{X,j}$ and $P_{Y,j}$ for the j^{th} event:

$$\text{KL}(X||Y) = \sum_{j \in X,Y} p_{X,j} \log\left(\frac{p_{X,j}}{p_{Y,j}}\right). \tag{56}$$

However, it is important to note that even if it is sometimes called a distance, the KL does not fulfill the properties of a metric. Namely, it is not symmetric in its arguments: $\text{KL}(X||Y) \neq \text{KL}(Y||X)$.
KL can be expressed in terms of entropies: $KL(X||Y) = H(X,Y) - H(X)$.
To incorporate a conditionality of X, Y onto some fact Z, we can extend the above measure as

$$KL_{X,Y|Z}(X||Y) = \sum_{j,k,z \in X,Y,Z} p_{X,j,k,z} \log\left(\frac{p_{X,j,k|z}}{p_{Y,j,k|z}}\right). \tag{57}$$

Mutual Information We can now derive the mutual information as a particular case of the KL divergence. MI is the relative entropy of the joint probability distribution of two variables X and Y to the product of their marginal probabilities:

$$\text{MI}(X,Y) = KL(P_{X,Y}(X,Y)||P_X(X)P_Y(Y)). \tag{58}$$

Therefore, MI quantifies the KL-distance of the joint process from the joint process with independence assumed (in which case the joint probability density is the product of the marginals). The measure is symmetric, and like the product-moment correlation, it only quantifies the association between variables without inferring directional information. Since the MI measure is a special case of the KL divergence, we can also express MI in terms of entropies as

$$MI(X,Y) = H(x) - H(X|Y) = H(X) + H(Y) - H(X,Y). \tag{59}$$

This expression is the formula stated in the previous section (38).
The MI is non negative since $H(X) + H(Y) \geq H(X,Y)$ [69]. To bound the measure in the $[0,1]$ interval we can normalize in one of two ways,

$$\text{NI}(X,Y) = 1 - \frac{H(X,Y)}{H(X) + H(Y)} \in [0,1]$$
$$\text{NI}'(X,Y) = 1 - \frac{H(X,Y)}{\log(m_X m_Y)} \in [0,1].$$

Here m refers to the number of possible values the variables can take (this is the number of bins for discretized continuous data). These normalizations differ from the one used by Ma [44]. The author followed the convention by [70] that interprets MI as an information-theoretic analog to covariance. Thus, he utilized the normalization

$$\mathrm{NI}^{\rho}(X,Y) = \frac{H(X)+H(Y)-H(X,Y)}{\sqrt{H(X)\cdot H(Y)}}. \tag{60}$$

We justify this differing choice of normalization by the experimental validation with model systems in section (5.3). This justification is one key result of this thesis.
Other, more exotic, normalizations exist [30] [51] [47]. We will, however, neglect those in the scope of this thesis to favor the easier to interpret normalizations above.
Additionally, we will exercise the empirical analysis of normalizations employing model systems only to the transfer entropy. We choose this restriction to keep the scope of this thesis concise. Nonetheless, we shall cover the MI for a comprehensive derivation and theoretical justification of the transfer entropy measure. We do so since the transfer entropy measure is interpretable as a conditional case of the MI.

The MI, as stated here, is an entropy-derived measure and, as such, only defined on nominal scales. Since the formula does not depend on the magnitude of data, $\varphi^D = \mathrm{Id}$ is a possible choice for ordinal and discrete data. To employ this φ^D, we merely have to let go of the connotation of magnitude to the numeric labels. Continuous metric variables, on the other hand, need discretization. Similar to the mode, values in continuous sets repeat too sparsely for the measure to yield accurate results.
Though generalizations of MI to continuous data exist, they exhibit problems. We discuss these problems in the preceding section (3.2). Therefore, we will apply a mapping φ^D to discretize the data and then apply formula (38) or its normalizations. The derivation of an optimal φ^D is one of the main contributions of this thesis.

For time series, this directional information $X \to Y$ is introduced into the measure by applying a lag to one time ordered sample:

$$\mathrm{MI}^{\tau}(Y_t,X_t) = \mathrm{MI}(Y_t,X_{t-\tau}) = \sum_{t=1}^{n} p(y_t,x_{t-\tau}) \log\left(\frac{p(x_t,y_{t-\tau})}{p(y_t)p(x_{t-\tau})}\right). \tag{61}$$

We can adapt this quantity to capture the dynamical structure of the association of the variables. Therefore, we interpret the time series as a Markov process. To do so, we condition the probability masses of the time series onto the past: $p(y_t) \to P(y_t|y_{t-1},\ldots,y_{t-\tau})$.
We now want to control for both self-information as well as information flow between variables. This approach serves to quantify the deviation of the generalized Markov property to sole self influence

$$P(y_t|y_{t-1},\ldots,y_{t-K}) = P(y_t|y_{t-1},\ldots,y_{t-K};x_{t-1},\ldots x_{t-L})). \tag{62}$$

We can achieve this goal by employing the Kullback-Leibler divergence.

Transfer Entropy The resulting relative entropy is called transfer entropy (TE) [60]:

$$TE_{X\to Y} = \sum_{t=1}^{T} p(y_t,y_{t-L:t-1},x_{t-K:t-1}) \log\left(\frac{p(y_t|y_{t-L:t-1},x_{t-K:t-1})}{p(y_t|y_{t-L:t-1})}\right). \tag{63}$$

Here, we explicitly allow for different bounds of temporal influence of the past K,L of the variables X_t,Y_t. We will determine the proper choice of K,L by evaluating the autocorrelation function. We choose the lag for the TE to be the temporal shift τ for which the self-correlation

vanishes.
As for all entropy derived measures, the base of the logarithm determines the units of the quantity. It is important to note that the TE measure is explicitly asymmetric.

As stated here, TE is only defined for discrete signals. This circumstance is especially apparent when we rewrite equation (63) in terms of entropies:

$$TE_{X\to Y} \quad = \quad H(Y_t|Y_{t-L:t-1}) - H(Y_t|Y_{t-L:t-1}, X_{t-K:t-1}). \tag{64}$$

It is ambiguous how to evaluate this expression for continuous variables. The differential entropy does not fulfill non-negativity and is not the continuum limit of the Shannon entropy. However, we cannot apply the suggested corrections (see 3.2.1, 3.2.2) since the PDFs are unknown.
To solve this problem, one could apply density estimation methods such as kernel density estimators. However, these approaches still suffer from finite sample effects and introduce computational overhead. Alternatively, one can apply the histogram estimators discussed in section (2.2.6). This estimator, however, is, in essence, a discretization and coarse-graining or binning, φ^D, into discrete variables (see section 2.3) of the data for which we can then apply the above expression (63).
In that case, the choice of a suitable φ^D is crucial. Therefore, we want to apply a PMF estimator such as displayed in equation (11) as opposed to PDF estimators as in equation (6). We justify this choice because, generally, the probability of the discrete PDF classes only integrate but does not sum to unity, resulting in underestimating the NTE value. In chapter (5.7), we will conduct studies that will show the extreme sensitivity of (normalized) TE values on the applied φ^D. This finding is consistent with results from [30].

When δ quantifies the level of the discretized resolution after applying $\varphi^D = \varphi^D_\delta$, in the $\delta \to 0$ limit this measure is finite and independent of the partition φ^D. However, from equation (63), we can infer that TE explicitly depends on the sample length. Additionally, for quantized continuous variables, TE depends on the level of preserved detail, that is, the number of discrete values in the target set of φ^D. However, in the chapter (5.4), we will determine that this limit requires large samples, and we will evaluate the finite sample, $\delta > 0$ performance of the transfer entropy. To mitigate these sample effects, we seek to normalize TE and therefore express it in terms of entropies as in equation (64).

This yields two methods of normalization. The first is straightfoward. We divide TE by its maximal value, $H(Y_t|Y_{t-l-1:t-1}) = H(Y_t, Y_{t-L:t-1}) - H(Y_{t-L:t-1})$ which results in the expression:

$$H\,NTE_{X\to Y} = 1 - \frac{H(Y_t, Y_{t-L:t-1}, X_{t-L:t-1}) - H(Y_t|Y_{t-L:t-1}, X_{t-K:t-1})}{H(Y_t, Y_{t-L:t-1}) - H(Y_{t-L:t-1})}. \tag{65}$$

The second normalization is by $\log m$ with m the number of values of (the discretized) variable Y.

$$\log m\,NTE_{X\to Y} = \frac{TE_{X\to Y}}{\log m}. \tag{66}$$

This holds since $H(Y_t|Y_{t-l-1:t-1}) \leq H(Y_t) \leq \log m$ (see section 3.1 above). We will refer to these normalisations as H NTE and $\log m$ NTE. The H NTE normalization is commonly utilized in literature. See for example [15], or [23].
These normalizations differ from the ones utilized by [44]. He applied the same normalization as for the mutual information in equation (60). Nonetheless, the evaluation of the normalization utilized by Ma is implicitly covered in the subsequent analysis in section (5.3). This coverage is due to the fact that the model systems univariate entropies are identical. Thus for $L = K$ we have

$$\sqrt{H(X_t|X_{t-K})H(Y_t|Y_{t-L})} \approx \sqrt{H(Y_t|Y_{t-L})^2} = H(X_t|Y_{t-L}).$$

The abovementioned exotic normalizations for mutual information can be adapted for TE interpreted as conditional MI. Nonetheless, we neglect a thorough analysis of them as these normalizations offer no intuitive interpretation and are more complex and thus disfavoured by the principle of Occams razor.

The H NTE normalisation has the problem that for $H(Y_t|Y_{t-l-1:t-1}, X_{t-l-1:t-1}) = 0$ and $H(Y_t|Y_{t-l-1:t-1}) = \varepsilon$ the normalized TE converges to one in the limit of $\varepsilon \to 0$ even though the transferred information approaches zero.
The $\log m$ NTE normalization, on the other hand, yields difficulties in interpretation. It reaches its maximum value not when the possible information transfer $x \to y$ is maximal, but only when additionally $H(Y)$ is at its maximum. While H NTE might feel more intuitive, we will see in the chapter (5.3) that $\log m$ NTE offers advantages when developing a robust TE calculus for finite datasets.

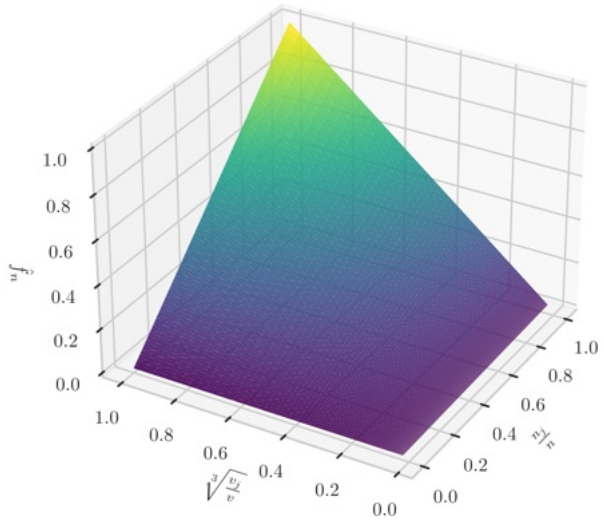


Figure 4.1: The figure displays the probability of a discretized value mapped from a continuous distribution with the adapted estimator. This probability depends linearly on the fraction of volume ($D = 3$) and data points within it.

4 Incorporation of Bin Size into Discretized Probability Estimation

The methods we introduced in section (2.3) are distinguished by the way they partition the data. Many of them utilize fixed-width bins. Others, however, use the values of the different measurements for clustering and subsequent partitioning. In that case, the size of the bins does hold information about the sample. Therefore, in this chapter, we will develop an adapted probability estimator within this chapter. The resulting estimator should incorporate the bin sizes and thus recover the common probability estimator value for equal-sized bins. Additionally, the estimator should perform better than the common estimator. The operationalisation of *better* is non-trivial. We will evaluate the consistency and thus score variance. Since we are dealing with an estimator that should perform a consistent discretization of a continuous variable, it is desirable but no necessity that the continuum limit of the estimator is the variables PDF.

Regarding the novel estimator, we have no justification to either overvalue the contribution of the number of data points within or the size of the bin. Thus we aim for equal contribution. To quantify the relative contribution of the components, we will normalize both terms by the total number of data points and the total range occupied by values. Since the contribution of the volume term is proportional to the exponent of the data dimension, D, we will take the D^{th} root of the volume term. Lastly, we apply a normalization constant such that the individual probabilities sum to unity. The resulting estimator reads:

$$\hat{f}_n(x) = \frac{1}{C} \sum_{j=1}^{m} \frac{n_j}{n} \sqrt[D]{\frac{v_j}{v}} \mathbb{1}_{x \in B_j} \tag{67}$$

where n_j/n is the fraction of data points within the j^{th} bin, $\sqrt[D]{v_j/v}$ is the D-root of the fraction of the total volume of non-empty bins and C is a normalisation constant.
Figure (4.1) shows the value of $\hat{f}_n$ with the fraction of data points and sample volume. We can see that the estimator is linear in both quantities, and they have an equal contribution.
We will now continue to discuss several theoretical properties of this estimator. An evaluation with model systems is discussed in chapter (5.5).

4.1 Expected Value

We will first calculate the expected value of our novel estimator (equation 67) and therefore look a the one-dimensional case, $D = 1$. The volume v_j of the bin containing a given x is fixed. The normalization C is a constant. Thus,

$$\mathbb{E}\left(\hat{f}_n(x)\right) = \mathbb{E}\left(\frac{1}{C}\sum_{j=1}^{m}\frac{n_j}{n}\sqrt[D]{\frac{v_j}{v}}\mathbb{1}_{x\in B_j}\right) \tag{68}$$

$$= \mathbb{E}\left(\frac{1}{C}\frac{n_j}{n}\sqrt[D]{\frac{v_j}{v}}\right) \tag{69}$$

$$\overset{D=1}{=} \mathbb{E}\left(\frac{1}{C}\frac{n_j}{n}\frac{v_j}{v}\right) \tag{70}$$

$$\overset{v_j \text{ fixed}}{=} \frac{v_j}{Cv}P(x \in v_j) \tag{71}$$

$$= \frac{v_j}{Cv}\int_{a+\sum_{k=1}^{j-1}v_k}^{a+\sum_{k=1}^{j}v_k} f(u)\,du \tag{72}$$

$$= \frac{v_j}{Cv}\left(F\left(a+\sum_{k=1}^{j}v_k\right) - F\left(a+\sum_{k=1}^{j-1}v_k\right)\right). \tag{73}$$

We can now define $p_j = \int_{B_j} f(u)du$ and apply the mean value theorem. Subsequently, we obtain

$$\mathbb{E}\left(\hat{f}_n(x)\right) = \frac{v_j}{Cv}p(x^*)\left(a+\sum_{k=1}^{j}v_k - \left(a+\sum_{k=1}^{j-1}v_k\right)\right) \tag{74}$$

$$= \frac{v_j^2 f(x^*)}{Cv} \tag{75}$$

for at least one $x^* \in B_j$. This result is easily generalized to $D > 1$ for which one obtains

$$\mathbb{E}\left(\hat{f}_n(\vec{x})\right) = \frac{\sqrt[D]{v_j^{D+1}}}{Cv}f(\vec{x}^*). \tag{76}$$

We can therefore conclude that the estimator rescales the actual probability of the underlying distribution. Additionally, the magnitude of this scaling depends on the volume v_j of the bin B_j containing x. This circumstance makes it dependent on the binning methodology. The rescaling effect makes $\hat{f}_n$ inferior to the usual histogram estimator. Nonetheless, the explicit dependence on v_j allows for the steering of the rescaling magnitude with the choice of binning. Since we are not interested in absolute values of the entropy, the ability to tune the bias introduced by binning might be worth the tradeoff when we apply the estimator for entropy analysis.

4.2 Bias

We will now proceed to calculate the bias. Therefore we again set $D = 1$. Additionally, we assume the derivative of the PDF to be bounded $|f'(x)| \leq L_1$. Resultingly, $|f(x)| \leq L_2$ (see appendix 8.2.1).

Let L be the larger of L_1, L_2. We further assume M to be the biggest bin. Then

$$\begin{aligned}
\textbf{bias}\left(\hat{f}_n(x)\right) &= \mathbb{E}\left(\hat{f}_n(x)\right) - f(x) && (77)\\
&= \frac{v_j^2}{Cv} f(x^*) - f(x) && (78)\\
&= f'(x^{**})(x^* - x) - f(x^*) + f(x^*)\frac{v_j^2}{Cv}. && (79)\\
&= f'(x^{**})(x^* - x) + f(x^*)\left(\frac{v_j^2}{Cv} - 1\right) && (80)
\end{aligned}$$

Here we added $0 = f(x^*) - f(x^*)$ in the second line and used the mean value theorem. From our assumptions we conclude that $0 \leq f'(x^{**}), f(x^*) \leq L$ and $x^* - x \leq M$ as well as $v_j^2 \leq M^2$. Thus

$$\textbf{bias}\left(\hat{f}_n(x)\right) \leq L\left(\frac{M^2}{Cv} + M - 1\right), \tag{81}$$

which differs from the bound of the bias for the common histogram estimator in equation (9) by the term $L\left({}^{M^2}/_{Cv} - 1\right)$. Therefore, the bound of the adapted estimator bias it is lower than the one of the common histogram if $M < \sqrt{Cv} = \sqrt{\sum_{j=1}^{m} \frac{n_j}{n} v_j}$. This quantity is the root of the sum of all bin volumes weighted by the fraction of the data points contained inside. For bins of equal size $M = v_j$, this condition reduces to

$$M < M\sqrt{\frac{\sum_{j=1}^{m} \frac{n_j}{n}}{M}} \Leftrightarrow 1 < \frac{1}{\sqrt{M}}. \tag{82}$$

Likewise to the ordinary histogram estimator, smaller bins result in a more negligible overall bias. Additionally, for equal-size bins, when condition (82) holds, the bias is smaller than the one of the ordinary histogram estimator.
For higher dimensions the equation (81) generalizes to

$$\textbf{bias}\left(\hat{f}_n(x)\right) \leq L\left(\sqrt[D]{\frac{M^{D+1}}{C^D v}} + M - 1\right). \tag{83}$$

4.3 Variance

We again let M be the biggest bin width and $|f(x)| \leq L$. Hence with $\tilde{C} = \sqrt{C}$

$$\begin{aligned}
\textbf{var}\left(\hat{f}_n(x)\right) &= \textbf{var}\left(\frac{1}{\tilde{C}} \sum_{j=1}^{m} \frac{n_j}{n} \sqrt[D]{\frac{v_j}{v}} \mathbb{1}_{x \in B_j}\right) && (84)\\
&= \frac{1}{Cn^2 v^2} \textbf{var}\left(\sum_{j=1}^{m} n_j v_j \mathbb{1}_{x \in B_j}\right) && (85)\\
&\leq \frac{M^2}{Cn^2 v^2} \textbf{var}\left(\sum_{j=1}^{m} n_j \mathbb{1}_{x \in B_j}\right) && (86)\\
&= \frac{M^2}{Cn^2 v^2} \textbf{var}(n_j). && (87)
\end{aligned}$$

The n_j are a binomial random variable $\mathscr{B}(n,p;x)$ with mean np and variance $np(1-p)$. Therefore

$$\mathbf{var}\left(\hat{f}_n(x)\right) \leq \frac{M^2}{Cn^2v^2} np_j(1-p_j) \tag{88}$$

$$= \left(\frac{M}{v}\right)^2 \frac{p_j(1-p_j)}{Cn} \tag{89}$$

$$\leq \left(\frac{M}{v}\right)^2 \frac{1}{4Cn}. \tag{90}$$

In the last step we used $p_j(1-p_j) \leq 1/4$.
Therefore, we can conclude that the variance decreases with the size of the dataset, with a larger value range, spanned by the sample, and decreasing bin size. This variance behavior is a difference to the variance of the common histogram estimator for probability in equation (10), which increases for smaller bins.
For $D > 1$, this generalizes to

$$\mathbf{var}\left(\hat{f}_n(x)\right) \leq \left(\frac{M}{v}\right)^{\frac{2}{D}} \frac{1}{4Cn}. \tag{91}$$

4.4 Behavior for Equal-Width Bins

For equi-width bins the estimator $\hat{f}_n(x)$ converges to the common discrete probability $\hat{f}_n(x) \to P(x \in B_j) = n_j/N$.
When the volume of bin B_j is constant $v_j = \prod_{k=1}^{D} v_k = v^D$. Thus the normalization constant can be calculated as

$$1 \stackrel{!}{=} \frac{1}{C} \sum_{j=1}^{m} \frac{n_j}{n} \sqrt[D]{\frac{v_j}{v}} \tag{92}$$

$$\Rightarrow C = \sum_{j=1}^{m} \frac{n_j}{n} \sqrt[D]{\frac{v_j}{v}} \tag{93}$$

$$= \frac{v}{\sqrt[D]{V}} \sum_{j=1}^{m} \frac{n_j}{n} \tag{94}$$

$$= \frac{v}{\sqrt[D]{V}}. \tag{95}$$

This gives

$$\hat{f}_n(x) = \frac{1}{C} \sum_{j=1}^{m} \frac{n_j}{n} \sqrt[D]{\frac{v_j}{v}} \mathbb{1}_{x \in B_j} \tag{96}$$

$$= \frac{\sqrt[D]{V}}{v} \frac{n_j}{n} \sqrt[D]{\frac{v^D}{V}} \tag{97}$$

$$= \frac{n_j}{n} = P(x \in B_j). \tag{98}$$

The contribution of the bin width correction to the assigned probability vanishes in the normalization for equal-width bins. Then the adapted estimator reproduces the standard frequentist probability estimator for bins of equal size.

4.5 Behavior for Equal-Frequency Bins

For equal-frequency bins we have a constant number of data points in every bin: $n_j/n = n_0/n$. Thus the normalization becomes

$$1 \stackrel{!}{=} \frac{1}{C} \sum_{j=1}^{m} \frac{n_j}{n} \sqrt[D]{\frac{v_j}{v}} \tag{99}$$

$$\Rightarrow C = \frac{n_0}{n} \sum_{j=1}^{m} \sqrt[D]{\frac{v_j}{V}}. \tag{100}$$

Therefore we have

$$\hat{f}_n(x) = \frac{1}{C} \sum_{j=1}^{m} \frac{n_j}{n} \sqrt[D]{\frac{v_j}{v}} \mathbb{1}_{x \in B_j} \tag{101}$$

$$= \frac{1}{\frac{n_0}{n} \sum_{k=1}^{m} \sqrt[D]{\frac{v_k}{V}}} \frac{n_0}{n} \sqrt[D]{\frac{v_j}{V}} \tag{102}$$

$$= \frac{1}{1 + \sum_{k \neq j}^{m} \sqrt[D]{\frac{v_k}{v_j}}}. \tag{103}$$

Now let v_q be the bin with the largest volume. Then

$$\hat{\hat{f}}_n(x) \geq \frac{1}{1 + (m-1) \sqrt[D]{\frac{v_q}{v_j}}}. \tag{104}$$

If now $v_q = v_j$ we obtain $\hat{\hat{f}}_n(x) \geq 1/m$. If we consider v_q to be the smallest bin, the orientation of the inequality switches and $\hat{\hat{f}}_n(x) \leq 1/m$. Thus the hard criteria of equal probability for each bin with the frequentist notion of discrete probability transforms to a bound with this estimator. It allows for different probability based on bin sizes.

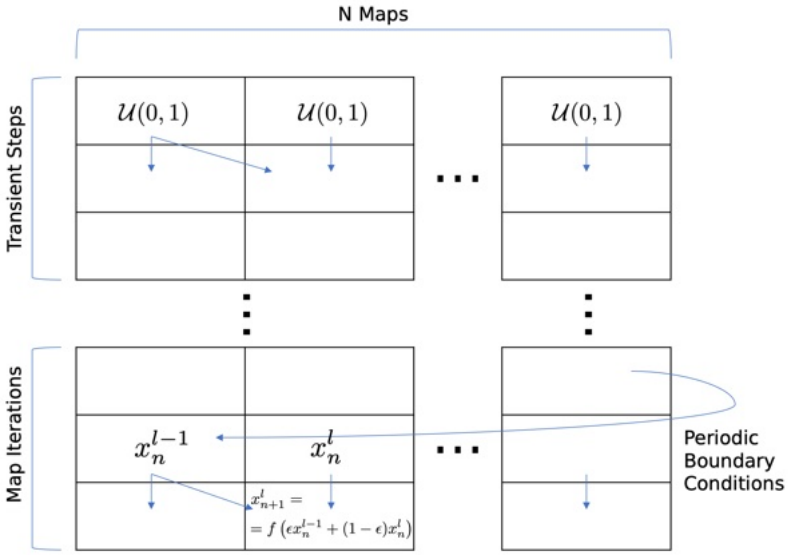


Figure 5.1: Schema of the coupled map lattice setup.

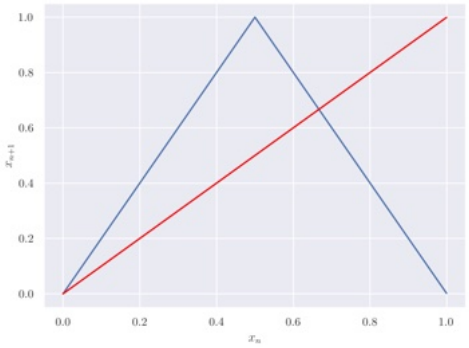
5 Transfer Entropy Estimation

In this section, we will evaluate the performance and caveats of the transfer entropy estimator. Therefore we will use synthetic systems which are similar to the ones utilized by Schreiber in his paper that first presented the measure [60]. We will extend this analysis and apply additional non-linear maps within the framework. Additionally, we will extend the parameter ranges Schreiber studied and evaluate the effects of small sample sizes, noise, range, type of discretization, and the effect of remapping the distribution onto another. We will evaluate the effects of these factors on the different normalizations of the transfer entropy. This chapter aims to provide merit to developing a robust transfer entropy estimator that yields comparable results when applied to different time series pairs and is therefore primarily independent of the just mentioned factors.

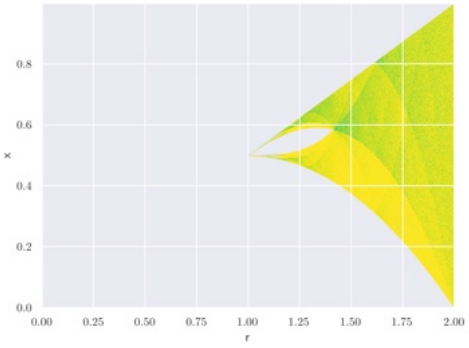
We applied the complete analysis to all maps. However, to spare the reader repetition, we will only discuss the findings of one instructive example map for every effect. First, we will start with the tent map to stay consistent with Schreiber. Subsequently, we will also refer to CML systems based on the other maps.

5.1 Model Systems

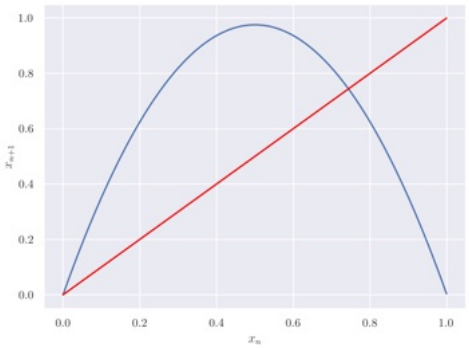
To study the abovementioned characteristics we will utilise coupled map lattices (CMLs). These systems are constructed such that the l values in row $n+1$ depend on row n. To start with, the individual maps are seeded uniformly on the $[0,1]$ interval and subsequently for the next row and map l we have $x_{n+1}^l = f(\varepsilon x_n^{l-1} + (1-\varepsilon)x_n^l)$ with with a non-linear function f and periodic boundary conditions. For statistics, we will generate 100 time series and run the iteration for a 10^5 transient steps. Then we start to record the timeseries. A schema of this lattice structure is displayed in figure (5.1).
Whereas Schreiber only used the tent map, we will additionally study the logistic, Bellows, and exponential map. The last two exhibit a long tail distribution. This is evident from figure (5.2).
To study the behavior of the measures, we will vary the coupling strength ε of the coupled map lattice. Therefore, we will proceed to introduce the maps used for the CML systems. First, however, we will discuss the non-linear functions.



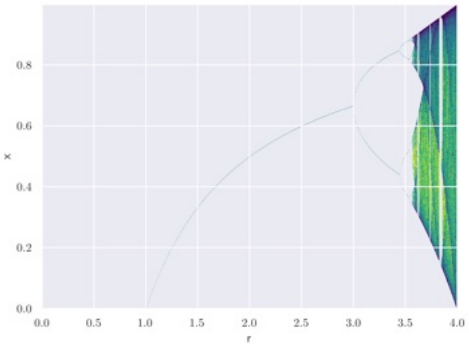
(a) Tent map Cobweb diagram.



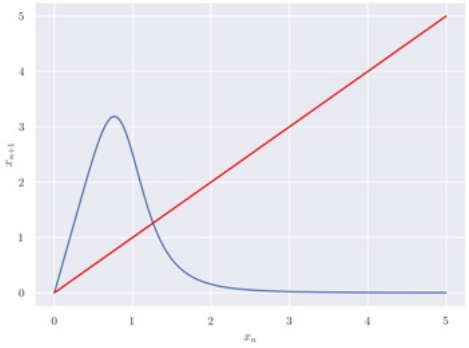
(b) Tent map bifurcation diagram



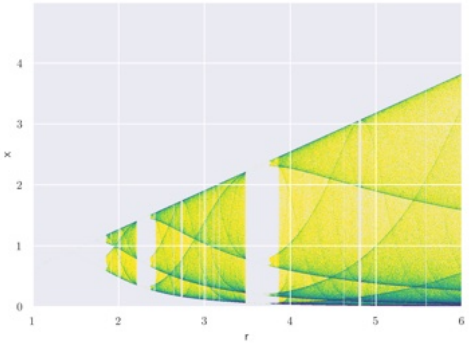
(c) Logistic map Cobweb diagram.



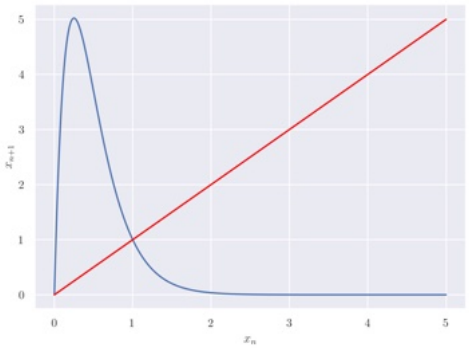
(d) Logistic map bifurcation diagram



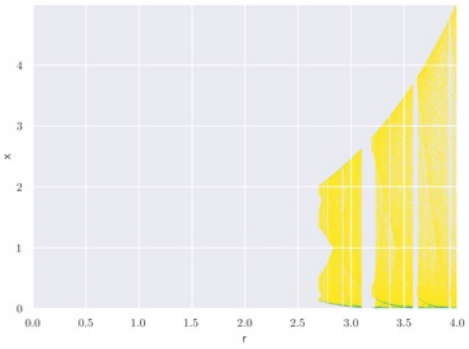
(e) Bellows map Cobweb diagram.



(f) Bellows map bifurcation diagram



(g) Exponential map Cobweb diagram.



(h) Exponential map bifurcation diagram

Figure 5.2: The figures show Cobweb and bifurcation diagrams of the tent, logistic, Bellows, and exponential map. The latter two exhibit power-law distribution characteristics.
The gradient indicates increasing point density in bifurcation diagrams from bright to dark colors.

Tent Map Lattice Schreiber [60] evaluates TE (among others) with a coupled tent map lattice. Specifically, he studies a low ε regime, $\varepsilon \in [0, 0.05]$ and uses a binary partition $\{[0, 0.5), [0.5, 1]\}$. We reproduce and extend this analysis to the whole $\varepsilon \in [0, 1]$ range.
The tent map is defined as

$$\text{Tent Map} : [0,1] \to [0,1]$$
$$f(x) = \begin{cases} rx & \text{if } x < 1/2 \\ r(1-x) & \text{else} \end{cases}. \tag{105}$$

The Cobweb and the bifurcation diagram for the parameter r of the tent map are displayed in the figures (5.2a) and (5.2b) respectively. Schreiber used $r = 2$ in his analysis which proved numerically unstable in our implementation. It converged to a fixed point. Therefore, in this analysis, we used $r = 1.99999$.

For this choice of r the tent map has two unstable fixed points at $x_1^* = 0$ and $x_2^* = r/1+r \approx 2/3$. However, in the CML setup, the fixed points depend on the neighboring values. Thus they are different from the singular case.
Figure (5.4a) shows the empirical density, autocorrelation and power spectral density (PSD) of the tent map CML. The PDF plot instructively shows the repellent nature of the $x^* = 0$. Additionally, we observe diminishing autocorrelation and more weight in the higher frequencies of the PSD, though frequency amplitudes are smaller than for other maps.

Logistic Map Lattice The logistic map is defined as

$$f(x) = rx(1-x) \tag{106}$$

and was devised to model population dynamics. For this analysis we used $r = 4$ which is a non-linear transformation of the $r_{\text{tent}} = 2$ case of the tent map. The map has two unstable fixed points at $x_1^* = 0$ and $x_2^* = (r-1)/r = 3/4$. The bifucation diagram (5.2d) and empirical PDF (5.4b) at $\varepsilon = 0.5$ show the highest density of data points towards the higher edge of the possible data range. This fact is also visible in figure (5.2d).
Interestingly, the common parameter choice $r = 3.9$ exhibits intermittent behavior and subsequently runs into a periodic orbit when applied to coupled lattice systems with $\varepsilon = .5$. Figure (5.3) displays this effect. In the plot, we observe intermittency in the time series, which eventually reaches a steady state. These fixed points for certain ε result in the vanishing TE observed in figure (5.5b).[7]
The logistic map still shows a dynamic of the autocorrelation function, the autocorrelation for the longest temporal distance from 0 of all maps studied in this section. Therefore, it also exhibits the sharpest peak in the power spectral density (see figure 5.4b).

Bellows Map Lattice The Bellows map [57] is defined as

$$f(x) = \frac{rx}{1-x^b}. \tag{107}$$

We chose $b = 6$ and $r = 5$ for the ε-coupled lattice. The Bellows map is not a common system used in complex systems research. We nonetheless incorporate it into our analysis since it exhibits a fat-tailed distribution with no finite maximum value. This is also evident from the cobweb diagram (5.2e) and the empirical distribution of the CML at $\varepsilon = 0.5$ in figure (5.4c). Fat-tailed distributions are of interest since asset return distributions often exhibit similar features.
With our parameters, the Bellows map has one unstable fixed point at $x_1^* = 0$.[8]

[7] The study of the dynamics of the coupled lattice is certainly interesting. Nonetheless, it is beyond our application for transfer entropy and thus out of the scope of this thesis. We, therefore, neglect further discussion at this point.

[8] Would we evaluate the Bellows map on the complex plane, there is another unstable fixpoint at $x_2^* = \sqrt[b]{1-r} = i\sqrt[6]{5}$.

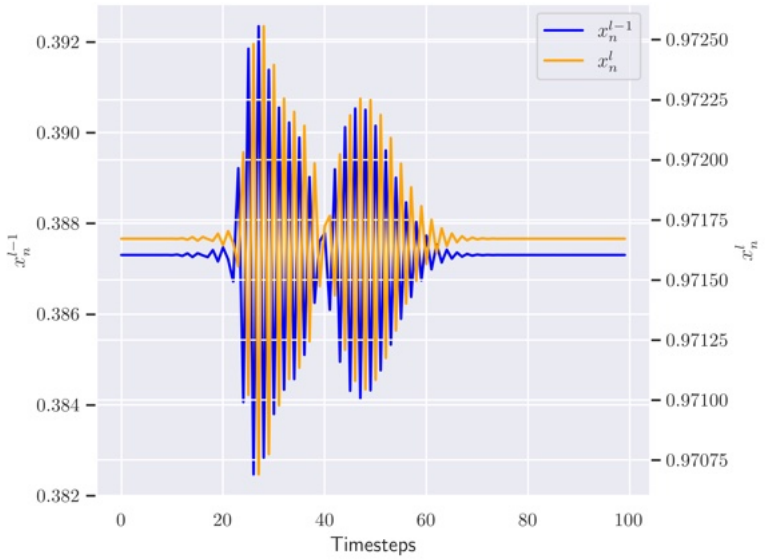


Figure 5.3: This figure shows the intermittency and convergence to a fixed point of the exponential map lattice at $\varepsilon = 0.857143$. Note the different scales of the two y-axes.

Exponential Map Lattice The exponential map is defined as

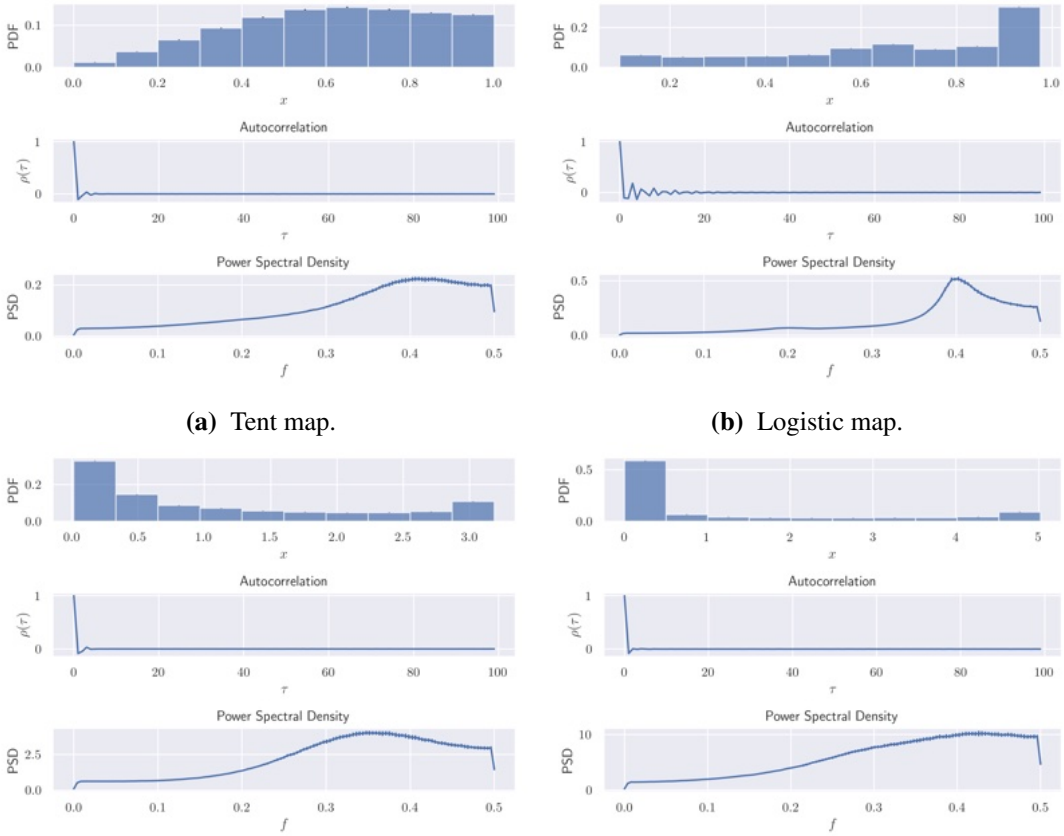
$$f(x) = x \cdot \exp\left(r(1-x)\right). \tag{108}$$

In our analysis we used $r = 4$. There are other forms of the exponential map. However, all forms are identical with a change of variables. As the Bellows map, the exponential map is unbounded with a long tail. With our parameter choice, the map has two unstable fixed points at $x_1^* = 0$ and $x_2^* = 1$. Figure (5.4d) shows the empirical PDF for $\varepsilon = 0.5$ with the highest density closer to zero. It has the highest peak within the PSD of all example maps we study in this thesis. It exhibits no autocorrelation.

5.2 Transfer Entropy by ε in CML Systems

We now study the information flow by evaluating the transfer entropy values between the rows of the CMLs. TE is calculated to quantify the information flow between every column and its left neighbour, $\mathrm{TE} = TE_{x_n^{l-1} \to x_n^l}$ (again with periodic boundary conditions). The transient period ensures an identical empirical distribution of the processes. As a result, the information flows between the CML columns are identically distributed. We see the vanishing errors of probability density, power spectrum, and autocorrelation in figure (5.4) that provide evidence for this result. We can, therefore, aggregate the measurement results of each CML column to provide statistics.
Even though the autocorrelation function might suggest a lag > 1 for the exponential map, the self-correlation decline is so steep that we take $K = L = 1$. Thus we keep consistency among our analysis and with Schreiber.
Now, figure (5.5) displays the variation of Schreiber's TE measure in the CML systems with the coupling strength ε. Generally, we see the expected increase in TE with increasing ε. However, there are exceptions to this trend. These exceptions appear when the CML system reaches a fixed point, periodic orbit or exhibits intermittency for the given parameter configuration.

Additionally, the figure (5.5) is an example of the dependence of the measured TE value on the partition of the data. We obtain the lowest (blue) line in the plots (5.5) by using a binary partition. Each higher line used one more class for discretization up until 30 bins for the highest



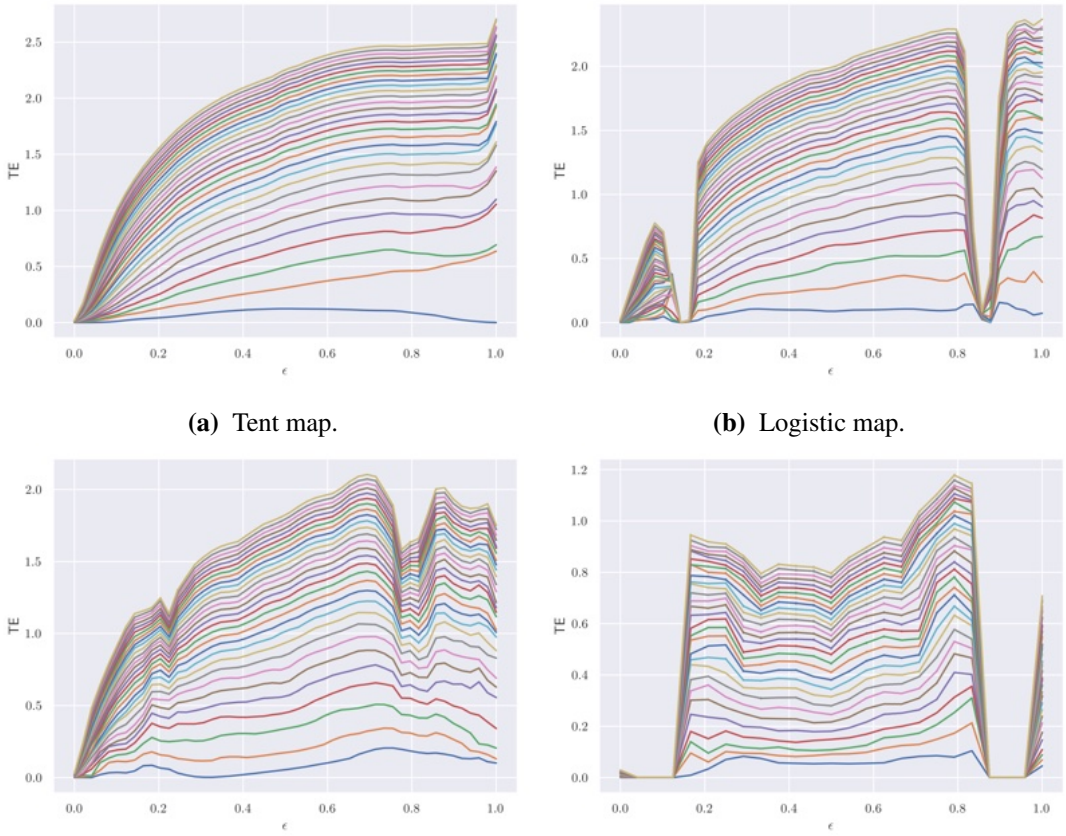
(a) Tent map.

(b) Logistic map.

(c) Bellows map.

(d) Exponential map.

Figure 5.4: The figure shows the empirical density (PDF), autocorrelation ($\rho(\tau)$), and power spectral density (PSD) of the four studied processes at $\varepsilon = 0.5$. Errors are taken over the CML columns but are mostly so small that they are not visible.



(a) Tent map.

(b) Logistic map.

(c) Bellows map.

(d) Exponential map.

Figure 5.5: The figure displays the CML systems' transfer entropy (TE) by $\varepsilon \in [0,1]$ of the four maps studied. The different lines correspond to the TE by ε variation with a different level of detail. The lowest blue line was obtained via a binary partition, and the upmost line was discretized into 30 classes. Generally, a finer partition allows for more information flow to be measured. This effect results in the stronger expression of features the variation of information flow with ε for finer partitions.

line. Since we calculated the measures with $10^5 \gg 30$ data points, the obtained TE values do not suffer from small sample effects.
We can observe two effects depending on the level of detail of the partition. First, a rough resolution is not able to resolve all the underlying dynamics. Examples of this are the exponential or Bellows map (figures 5.5c and 5.5d) which at some point show a decline of TE with increasing ε. The increased influencing relation between two adjacent CML columns stays hidden behind the partitioning of values.
The second effect is the increase of TE with the level of detail of the partition. These increases are intuitive in the sense that (in the absence of finite sample effects) a finer partition uncovers more information about the underlying process. The maximum possible entropy of a variable is $\log m$ when m is the number of different possible values. Resultingly, more information contained within a partitioned time series also allows for more information to flow between two time series. Again, this finding underlines the need for a proper normalization to uncover the relationship of the time series generating processes and not evaluate any skew in our result caused by length-of-sample effects.

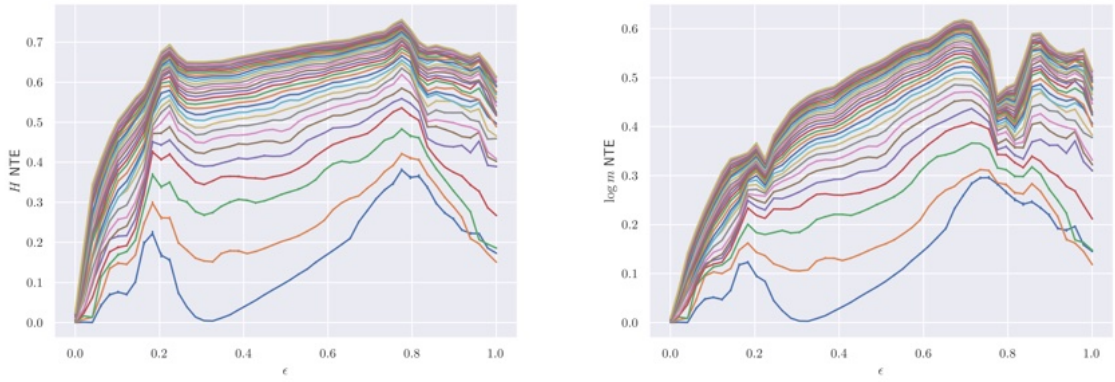
5.3 Normalization Dependence

We have discussed two normalizations in chapter (3.3) that we want to evaluate in this section. The first normalisation is the division of the TE value by the entropy of the influenced variable conditioned on its own past, $1 - {}^{\mathrm{TE}}/_{H(Y_t|Y_{t-L})}$. This normalization is intuitive as $H(Y|Y_{Y_t|Y_{t-L}})$ is the maximum of the TE and achieved if $H(Y_t|Y_{t-L:t-1}, X_{t-K:t-1}) = 0$ and the past of X fully explains Y. Therefore, the TE equals 0 when there is no change to the (joint and conditional) entropies of Y and its past whether or not we consider the past of X.
The maximum entropy of a variable with m different values, $H_{\mathrm{Max}} = \log m$, yields the second normalization candidate $\log m$ NTE = ${}^{\mathrm{TE}}/_{\log m}$. This measure is more difficult to interpret than the previous normalization. This circumstance becomes apparent since the measure's value is higher when one of two conditions is met (and the other constant): first, a higher flow of information, or, second, more information content within the influenced variable. Vice versa, we attain a low value if there is low information flow.
In chapter (4) we also introduced a normalization which relies on the interpretation of MI as an information theoretic analogue to the product moment correlation. NTE = ${}^{\mathrm{TE}_{X\to Y}}/\sqrt{H(X)H(Y)}$. We can omit separate study of said normalization since in our CML systems X and Y are identical processes and thus $H(X) = H(Y)$. For $H(Y_t) = H(Y_t|Y_{t-L1})$ this is identical to the first normalization we are studying. If $H(Y_t) > H(Y_t|Y_{t-L})$ then the resulting NTE does not cover the $[0,1]$ range as NTE $\in [0,u]$ with $0 \leq u < 1$.

Therefore, we will only compare the H and $\log m$ NTE normalizations. Figure (5.6) shows plots of these normalizations for the Bellows map. The plot shows the ε dependence of the Bellows map CML of 100 time series of 10^5 values. As intended, the normalizations shrink the TE values (see figure 5.5c). Therefore, it becomes apparent that the normalization squeezes the value differences of the NTE values obtained with a higher bin count more than in the rough partition regime. Two circumstances cause this effect. First, a more granular partition allows higher information storage, and second, the normalization is non-linear.
However, we see that the H NTE normalization skews the TE variation with ε to the point where we attain local maxima instead of local minima at $\varepsilon \approx 0.2$ and $\varepsilon \approx 0.8$. These maxima appear since the entropy conditioned on its past (with lag $K = L = 1$) responds more than the joint entropy to the ε specific CML dynamics (fixed points, periodic orbit, or intermittency). For these cases, H is lower than for the surrounding ε-values. The corresponding NTE values are therefore higher than in the ε neighborhood.
The $\log m$ normalisations shape stays closer to the unbounded TE variation (see figure 5.5c). Additionally, the $\log m$ normalization is a sample independent constant for a fixed resolution and thus does not add to the error of the resulting NTE value. This effect is evident in figure (5.7), where the $\log m$ NTE mostly shows the smallest variance. This effect is more pronounced in the small sample regimen than for sufficient data points.
Another drawback of the entropy normalization is its small sample behavior. For example, in figure (5.8), we can observe a washout of fine structure in the ε dependence of the H NTE measure on ε for the finer partitions, especially for intermediate values.

5.4 Sample Size Dependence

TE estimation relies on density estimation of the underlying processes. For some terms, this density estimation is trivariate as it depends on the distribution of Y_t, Y_{t-L}, X_{t-K}. Therefore, the number of classes for the probability estimation is cubic regarding the partition of X and Y. That is why we need sufficient sample sizes to obtain a robust estimate of the probability densities and TE. The lower bound of the sample size is of particular interest for our application since we want to study the temporal evolution of TE employing sliding windows (see section 2.5.4). Naturally,



(a) H NTE, $\varepsilon \in [0,1]$.

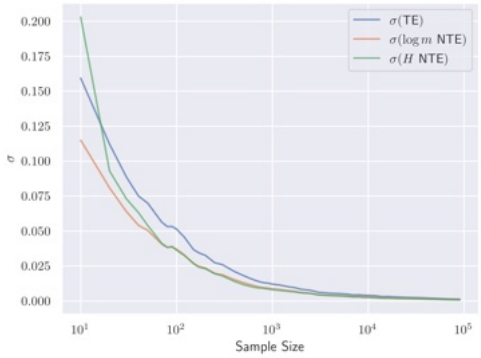
(b) $\log m$ NTE, $\varepsilon \in [0,1]$.

Figure 5.6: The plots show normalized TE values of the ε coupled Bellows map lattice for two different normalizations. We averaged values over 100 column realizations of 10^5 values with diminishing errors almost covered by the linewidth. The lines correspond to 2 to 30 bins for NTE calculation. The H normalization introduces a greater deviation from the unbounded TE variation with ε 5.5c) than the $\log m$ normalization.

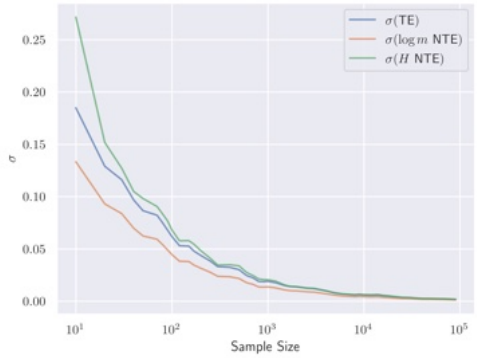
we would prefer a finer resolution of this temporal evolution. However, for a finite dataset, this means smaller sample sizes for each TE calculation time step.

Figure (5.7) shows the decrease of the error of the (N)TE values by sample size for a partition of four equal-sized bins at $\varepsilon = 0.5$ of the four CML systems.[9] We see that for increasing sample size, all errors converge to zero. For the most part, the error of the $\log m$ NTE is the smallest in absolute terms. It is trivial that the error of the $\log m$ normalized measure is smaller than the unnormalized TE since its value is divided by a constant for normalization. For the H normalization, the denominator contributes to the error of the measure. Therefore, the TE and H NTE error difference shows a more complex dynamic and is different for the different detail levels of the partition. In the four bin case, the difference of the H NTE and TE errors changes sign at certain points for the tent and Bellows map. For the most part, however, the H NTE exhibits the largest errors. The error magnitude is notable since the TE is larger than the normalized measures and thus has a smaller relative error. The finer partition shows a different dynamic. Here, for low sample sizes, H NTE shows a significantly larger error, which decreases to be soon lower than the TE error. Nonetheless, it is never significantly smaller than the other normalization. Generally for the $\log m$ NTE and sample sizes above 100 data points $\sigma < .05$. Since the NTE values are bounded, this corresponds to an error of less than 5%.
Since we now established how the error of the measures scales with sample size, we will now proceed to discuss the error dynamics in our CML systems for a small sample by varying ε. In the previous section (5.3 we established that the H normalization skews the TE variation with ε more than the $\log m$ normalization and generally yields higher NTE values. We observe that the magnitude of these effects increases with decreasing sample size. In figure (5.8), we see that for the same number of bins, the H NTE value is higher at the same ε when calculated for the small sample than for the large sample case displayed in figure (5.5c).
Additionally, the initial increase with ε at small ε values is steeper. On top of this, we can observe a washout of the H NTE variation features with ε, especially for finer partitions. These effects are also present within the other model systems. None of these small sample drawbacks is present for the $\log m$ NTE normalization.

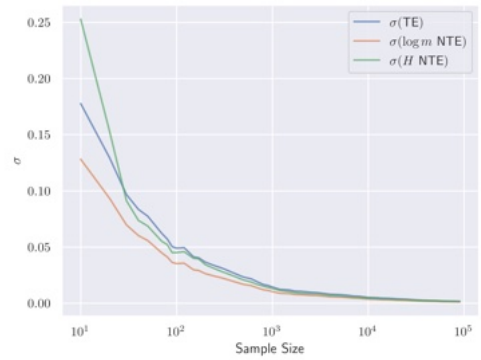
[9] The reader can observe the ε dependence of the CML dynamics for the partition into four bins at the lowest green lines in figure (5.5).



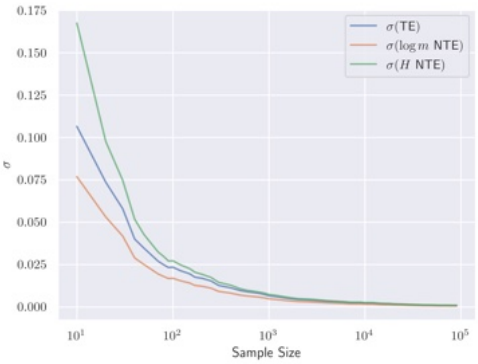
(a) Tent Map, 4 Bins.



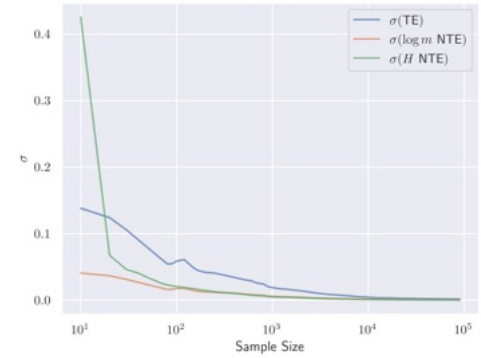
(b) Logistic Map, 4 Bins.



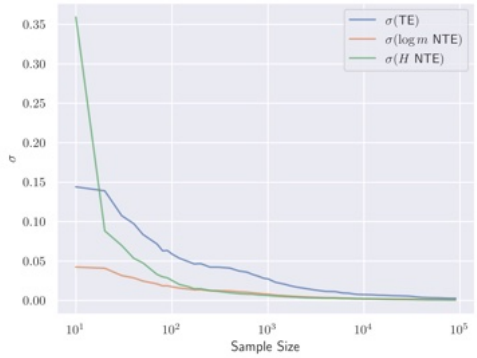
(c) Bellows Map, 4 Bins.



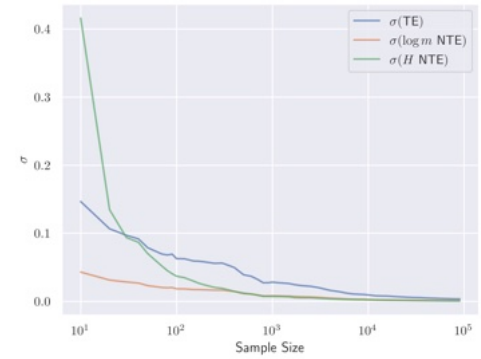
(d) Exponential Map, 4 Bins.



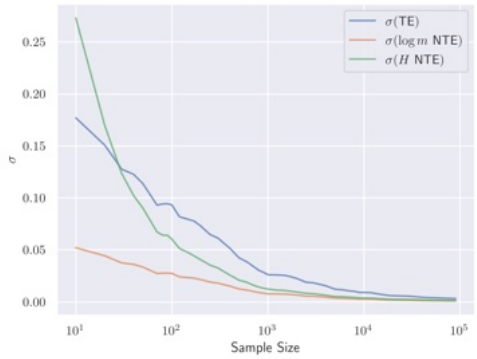
(e) Tent Map, 30 Bins.



(f) Logistic Map, 30 Bins.

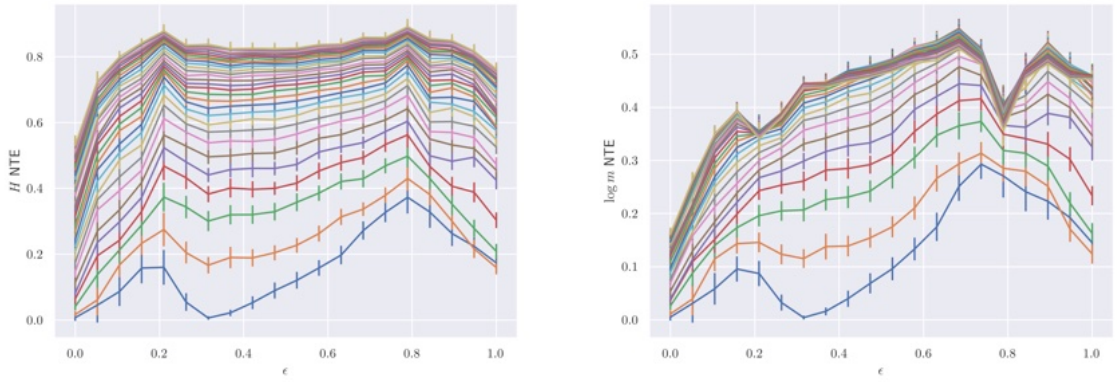


(g) Bellows Map, 30 Bins.



(h) Exponential Map, 30 Bins.

Figure 5.7: The figure displays σ by sample size for the four maps with 4 (5.7a, 5.7b, 5.7c, 5.7d) and 30 (5.7e, 5.7f, 5.7g, 5.7h) bins. More bins show a steeper decrease of errors with increasing sample size.



(a) H NTE, $\varepsilon \in [0,1]$.

(b) $\log m$ NTE, $\varepsilon \in [0,1]$.

Figure 5.8: The figure shows the small sample (300 data points) variant of the Bellows map setup in figure (5.6). We see that the H NTE measure washes out certain features of the information flow variation with ε. The same features are preserved with the $\log m$ NTE normalization.

Consistent with the prior discussion above, we see an increase in error due to the small sample size. We can also observe the non-vanishing NTE for both normalizations for the uncoupled CML ($\varepsilon = 0$). This effect, however, is more pronounced for the H NTE normalization. Regarding sample size, we can conclude that the $\log m$ normalization for an NTE measure is far more robust regarding small sample sizes than the H normalization.

5.5 Partition Detail Dependence

As evident from the figure (5.5), the TE value depends on the level of detail of the partition. However, the level of detail that is possible to resolve depends on the sample size. This section will discuss how the detail and type of partition influence the final (N)TE value. We will first discuss the combined effects of resolution and sample size, for which we utilize uniform-sized binning spanning the whole value range. In the subsequent section, we investigate the influence of differing binning methods that either fix the number of bins or allow for unequal bin sizes. The third type of binning method, which merely yields the appropriate number of bins, is implicitly covered with this first discussion for which we scan the number of bins.

Now that we have established the dependence of the (N)TE estimation on the sample size and normalization in the previous chapter, we will cover the influence of the partition itself. We have already discovered that the TE value depends on the level of detail of the partition (see 5.5). Additionally, the level of detail that is resolvable to a satisfactory level depends on the sample size. If one chooses a too fine partition for the number of data points, the resulting density will not reflect the proper PDF. Therefore, we will now discuss the interaction of these two effects and look at the interplay of partition detail and sample size.

The figures (5.9) show a surface plot of the (N)TE value dependent on the sample size and partition detail (of equal-sized bins) as well as the bivariate view of the relations at the surface edges. We can observe a significant difference in the surface shape between the H NTE and $\log m$ NTE variation. Whereas the latter exhibits a local maximum and subsequent decrease of the measure with increasing level of detail of the partition, the H NTE normalization at first increases steeply with the number of bins with a subsequent significant slope decline. Arguably, however, the most striking difference is the lack of variation of the H NTE variation with the sample size.

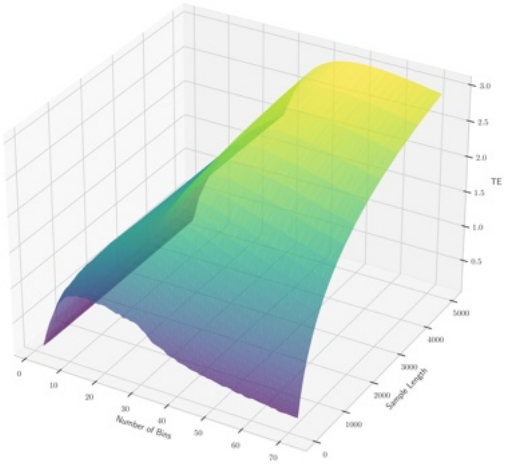
At first glance, this might seem desirable. After all, the purpose of normalization is to normalize for the variation of some parameters such as sample size. However, the H NTE of the logistic map converges to an upper bound value close to one. This decrease is because of the decrease in entropy with a finer partition and more evenly distributed probabilities of each class of the discretized PMF as they only contain fewer data points (or are empty and do not contribute the entropy). Said effect is more prominent for the numerator in equation (65) since the number of classes scales cubic with the partition detail for the trivariate joint and conditional entropy but only squares for the joint and conditional entropy in the denominator.
Additionally, the H NTE normalization yields higher values than the unbounded TE in the small detail regime. This odd effect of the normalized value being greater than the unbounded value is induced by the inability of the rough binning to resolve the dynamics of the process resulting in low entropy in the denominator and subsequently higher H NTE values.

Contrary to this, the $\log m$ normalization variation by sample size and partition detail recovers a rescaled shape of the unbounded TE. As such, the normalization fails to make the measure independent of either of these variables. Nonetheless, it introduces an upper bound to the measure. The TE and $\log m$ NTE variations exhibit a local maximum by partition detail. This extreme value occurs because of the interplay of two counteracting effects. First, an increase in partition detail uncovers more information within the constituent time series, increasing possible information flow.
On the other hand, finite sample effects and desaturation of bins result in a decrease in the final value. Since this desaturation occurs at a finer partition level for larger samples, we observe an increase in the measured value with the sample size. Naturally, this effect is far more pronounced at finer partitions. Moreover, since the probability of a data point falling into a particular bin decreases with an increasing number of equal-sized bins, the slope of this increase of ($\log m$ N)TE with the sample size decreases. We will exploit these characteristics to introduce a binning criterion in section (5.8).

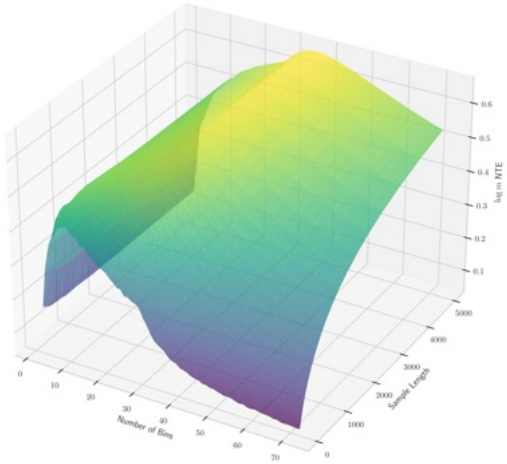
5.6 Discretization Method Dependence

In the last section, we have established the variation of (N)TE with the level of detail of the partition. To accomplish this, we utilized a binning with uniform width buckets. However, bins of unequal size are easily implemented and might better reflect the underlying structure of the data without increasing the level of detail. The ability to resolve more detail of the underlying process without increasing the resolution would prove useful if we consider the finite sample effects that demand a certain roughness of the partition. Additionally, several methods yield criteria for a "correct" number of bins. We established all these methods in detail in the section (2.3). They are additionally presented in the table (5.1). Of these methods, the EMGMM bins, a clustering-based method, were excluded from thorough analysis because of severe numerical instabilities and underperformance of the method compared to others in instances where the algorithm did converge.

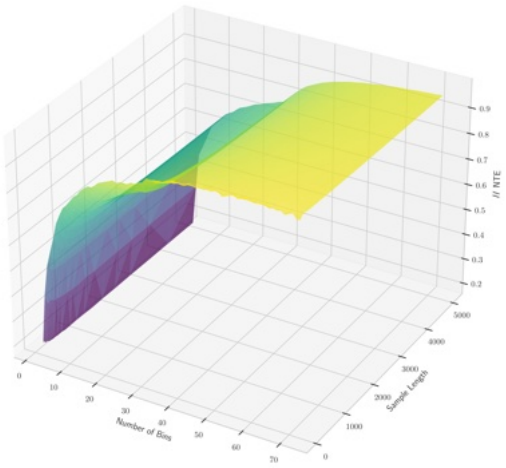
Scoring the performance of the binning methods is problematic because it is not trivial to devise a scoring criterion. The underlying processes of our CML systems are continuous values, yet (N)TE needs discretization. We cannot use the proximity of the estimated PDF to the true PDF for two reasons: first, the PDF of a time series in a CML system (or general) system might be unknown. Second, we cannot use the normalization by bin width necessary for PDF convergence because the integral over all values does not converge to one. Instead, we have to use a PMF, which is only normalized by the number of data points. This PMF, however, does not represent or converge to the actual density of the variable since it is continuous and not discrete.



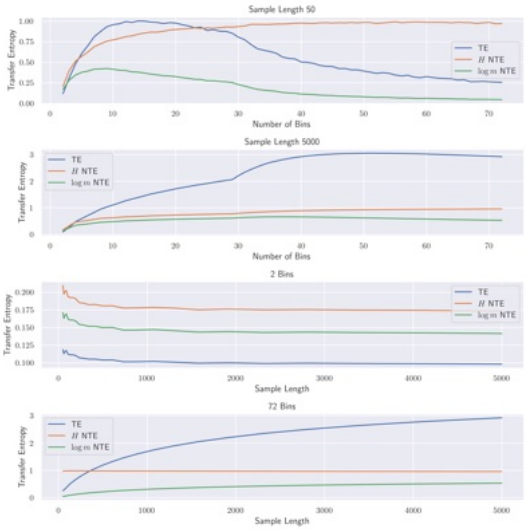
(a) TE.



(b) $\log m$ NTE.



(c) H NTE.



(d) Variation of (N)TE with bins and sample fixed respectively.

Figure 5.9: The variation of the three (N)TE values with the sample size and partition detail for the logistic map. Figure (5.9d) shows the side view of the surfaces. We see that the $\log m$ NTE measure recovers the shape of the unnormalized quantity. While the absence of variation with the sample length of the H NTE measure seems desirable, it occurs at a value close to one that the quantity loses meaning. This large information flow is not a feature of the process but just induced by the normalization.

Nr.	Binning Method	Outputs	Inputs
1	Uniform Width	h	Sample Range, m
2	$\sqrt{n}$	m	Sample Size
3	Sturges	m	Sample Size
4	Rice	m	Sample Size
5	Doane	m	Sample Size and Moments
6	Scott	h	Sample Size and Moments
7	Freedman-Diaconis	h	Sample Size, Interquartile Range
8	Knuth	m	Sample Size, m
9	k-Means	h_j	Sample, m
10	DBSCAN	h_j	Sample
11	Mean Shift	m, h_j	Sample
12	Uniform Probability	h_j	Sample, m
13	Agglomerative Hierarchical	h_j	Sample, m
14	Minimum Cross Validation	h	Sample, Sample Size and Range
15	Maximum Cross Likelihood	h	Sample, Sample Size and Range
16	Shimazaki-Shinomoto	h	Sample, Sample Size, Range, and Moments
17	Akaike	m, h	Sample Size and Range, m
18	Small Sample Akaike	m, h	Sample Size and Range, m
19	Bayesian Information Criterion	m	Sample Size and Range
20	Agostino Uniform Probability	m, h_j	Sample
21	EMGMM (Excluded)	h_j	Sample, m

Table 5.1: Overview of inputs and outputs of the binning methods. The numbering in this table serves as a reference for the subsequent plots. h refers to the bin width. If unequal bin sizes are possible this is indicated with the subscript j. m refers to the number of bins

Since we cannot score against some true value, we score the estimator consistency in the subsequent analysis. Namely, we compare the variance of (N)TE values over different realizations of the same process. The result is visible in the figures (5.10) and (5.11). For now, we are only interested in the left (blue) part of each violin plot. The orange part represents the result for the adapted probability estimator within the (N)TE calculation. We will discuss its performance in a subsequent section.

The violin plots show the empirical probability density for the values on the vertical axis of the sample of (N)TEs for 100 realizations of the same process. The displayed PDF is inferred with kernel density estimation. Unfortunately, this approach sometimes results in nonzero probability for subzero (N)TE values despite the (N)TE being ≥ 0, a common drawback of the KDE method. Within the violin plots, the white dot and black bar represent the mean and interquartile range of the sample. The numerical identifiers correspond to the binning methods as per table (5.1).

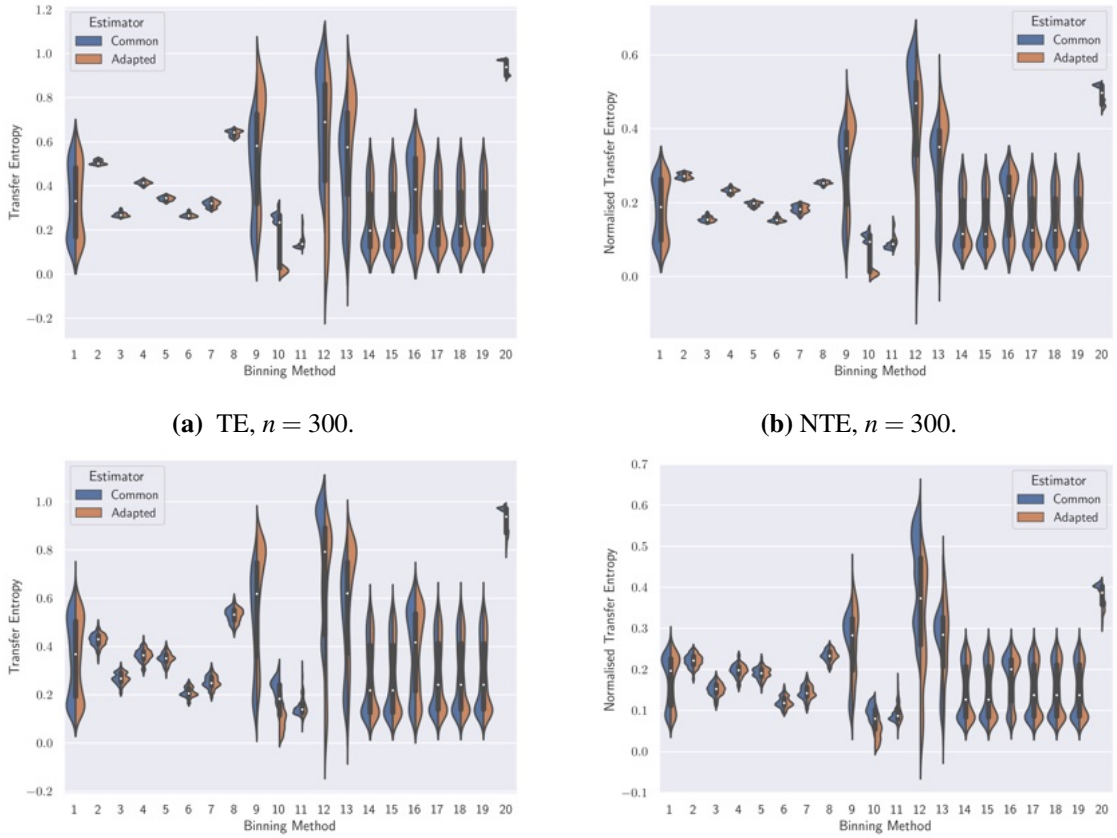
The for our purposes most important effect we can observe in figure (5.10) is that the different binning methods yield vastly different (N)TE results. They have different means and spread. The methods 2 to 8 exhibit the tightest distribution about the (N)TE spectrum. They perform comparably well because they yield a consistent number of bins based on the sample size, which is kept constant in figure (5.10). However, when we compare figure (5.10a) with (5.10c) or (5.10b) with (5.10d) we see that the value of these methods shifts as well. This shift is also apparent when we look at the spread of the uniform width bins (method number 1). As opposed to methods 2 to 8, which yield a certain number of uniform width bins based on sample characteristics, for method 1, the number of bins were varied, which shows a significant spread about the (N)TE value range. As such, uniform width bins yield a (N)TE measure that is a function of sample size, even when the number of bins is set with a criterion.

Let us now discuss the clustering-based methods 9 to 11, and 13. We can see that these methods fail to provide a consistent estimation as they exhibit a big spread of (N)TE values over the value range. Additionally, the PDF of the measure's value is not sharply peaked. The underperformance of these methods might be due to the nature of the data. As visible in figure (5.4), the data does not exhibit distinct clustered but merely value ranges of higher and lower density. Hence, clustering-method-based bins might still prove helpful in other applications where the data itself is more clustered. In our applications, however, this is not the case.

Next, we will discuss the uniform probability binning methods 12 and 20. The difference between the methods is that the Agostinos bins (number 20) provide a criterion for the number of bins, whereas these were varied for method number 12. Naturally, the latter shows a larger variance about the possible value range. However, the same reasons as above apply. Agostino's bins are a function of sample size. This is evident when we compare figure (5.10a) with (5.10c) or (5.10b) with (5.10d) between which the (N)TE value of Agostinos bins shift.

Lastly, we want to discuss the binning methods that calculate the number of bins with a criterion. These are the methods 14 to 19. Already in figure (5.10), it is apparent that these methods show a spread over the possible value range. When we study figure (5.11), we can infer how this spread comes about. The (N)TE value increases with the number of bins. Additionally, the spread of the (N)TE values increases with the number of bins.

We can therefore conclude that we could not identify a binning method that consistently outperforms the other ones. The methods that fix the number of bins show the lowest variance and spread, but the resulting (N)TE value remains a function of sample characteristics - for example, the sample size - that do not correspond to attributes of the underlying sample generating processes. As of now, it is apparent that the methods that allow for unequal bin sizes show the most



(a) TE, $n = 300$.

(b) NTE, $n = 300$.

(c) TE, $n = 1500$

(d) NTE, $n = 1500$

Figure 5.10: This figure shows violin plots of the TE and $\log m$ NTE values for the ordinary and adapted discrete probability estimator by different binning methods for two different sample sizes of the exponential map. The binning methods are labeled according to the table (5.1). Vertically, we see the empirical PDF of the (N)TE values. (N)TE is non-negative. The non-zero densities for (N)TE< 0 with some binning methods are an artifact of the kernel density estimator used for plotting.
We can see that the final (N)TE value severely depends on the shape, count, and method to generate bins. This dependence is true for both the standard and adapted probability estimators.

extensive spread. This spread might be because we neglect available information encoded in the bin sizes. Therefore, in the next section, we will evaluate (N)TE calculation with the adapted estimator that incorporates bin size into PMF (and notably not PDF) estimation.

5.7 Evaluation of Discretization with Incorporated Partition Width

So far, we have only evaluated partitions of equal width. In section (4), we proposed a novel discretizing probability estimator that incorporates not only the number of data points but additionally the bin size. The results of the (N)TE calculations with the adapted PMF estimation are displayed in the orange densities in the plots (5.10) and (5.11). These figures show an empirical verification that the estimator recovers the common PMF-based (N)TE values for uniform width bins. Therefore we will limit the subsequent study onto binning methods that allow for unequal bin sizes. These are the methods with h_j as an output in table (5.1).

Nonetheless, we again have to pose the question of how we should evaluate performance. In an earlier study included in the appendix (appendix section 8.3), we scored the estimators' performance on samples drawn from know distributions against the true PDF values. However, data from this approach was not the result of dynamic system interaction, and as such, we were only able to test only PDF estimation and not (N)TE in this framework. Therefore, this section will stay consistent with the above analysis and score our CML systems' performance. This approach, again, is difficult as we do not know the true underlying PDF of the process. We will therefore evaluate consistency. We will conduct this evaluation by utilizing probing points at certain x values. Then we evaluate the variance of the inferred probability mass at that point. The result is displayed in figure (5.12). The first two rows show the probing for the common and adapted probability estimator, respectively. The third row shows their difference. The first and second columns show the edge cases for the number of bins at 2 and 100 bins, whereas the last column shows the values over all binning methods. Each subplot in figure (5.12) shows the probing results for all of our CML systems. The different systems are plotted adjacently but separated by vertical blue lines. We utilized 100 probing points that were evenly spaced over the data range.

We can see that for a higher number of bins, most of the methods yield consistent results. The exceptions are the DBSCAN and mean shift methods. Nonetheless, when taken overall bins, all methods with both PMF estimators show nonzero variance. This variance is highest for the DBSCAN method. We can also see that performance varies by the particular CML analyzed.

To evaluate the performance of the adapted PMF estimator against the common one, the last row in figure (5.12) is most instructive. Here we see the differences of variances, $\sigma(\text{Adapted}) - \sigma(\text{Common})$, between the estimation methods at the probing points. Since variances are ≥ 0 a value > 0 indicated better performance of the common estimator, a value < 0 the opposite and vanishing difference signifies equal performance.

The results of this probing evaluation are at odds with the prior results that compared the estimated PMF to the actual probability density. We see both under and overperformance of the adapted estimator when compared to the standard probability density estimation. All in all, the differences seem reasonably moderate. An exception to this is the exponential map CML with the uniform probability estimator. Here we see a substantial underperformance of the adapted estimator on almost the whole value range of the exponential map. When we evaluate the performance overall bin numbers, we see that the overperformance of the adapted estimator, if it occurs, tends to be at the edge of the value ranges of the respective maps.

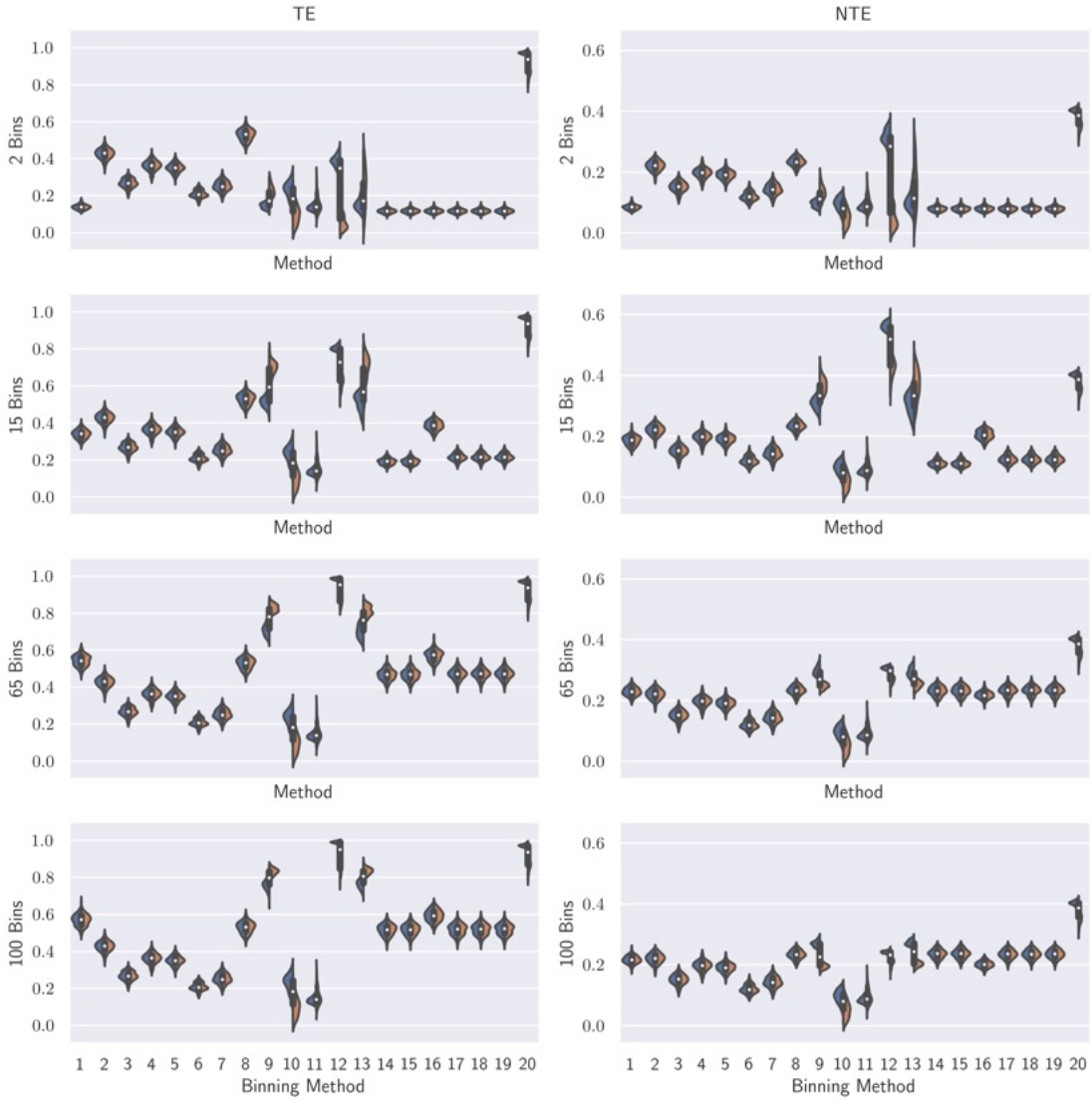


Figure 5.11: The figure shows the same setup as in figure (5.10) for $n = 300$. It differs in the aspect that we controlled the number of bins (if applicable for the method, i.e., m is a method input). Other samples were taken over different numbers of bins in the range of 2 to 100 in each plot. We see a slight convergence for larger bin counts.

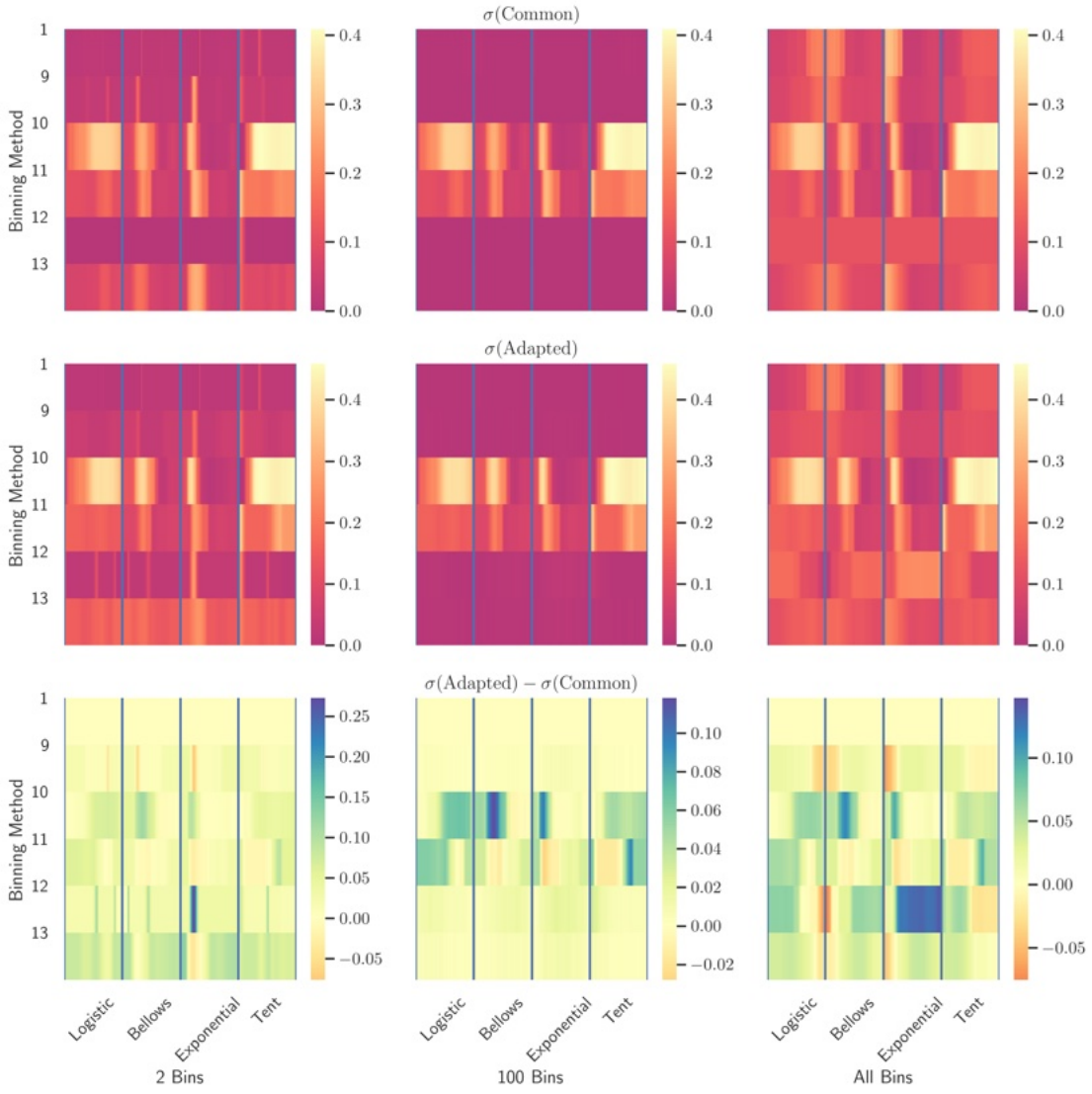
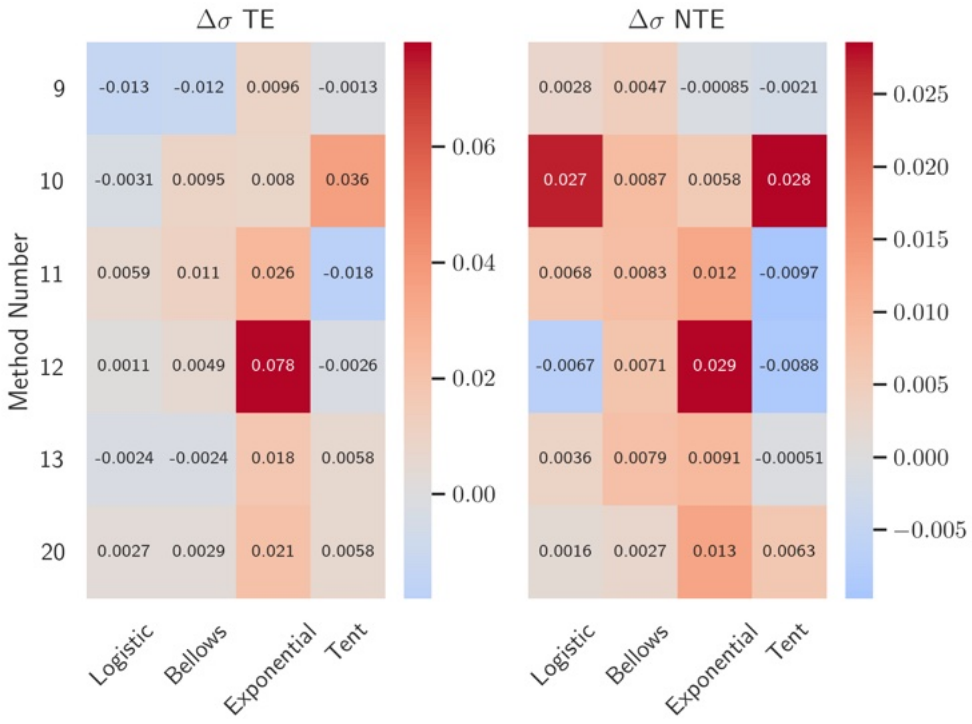
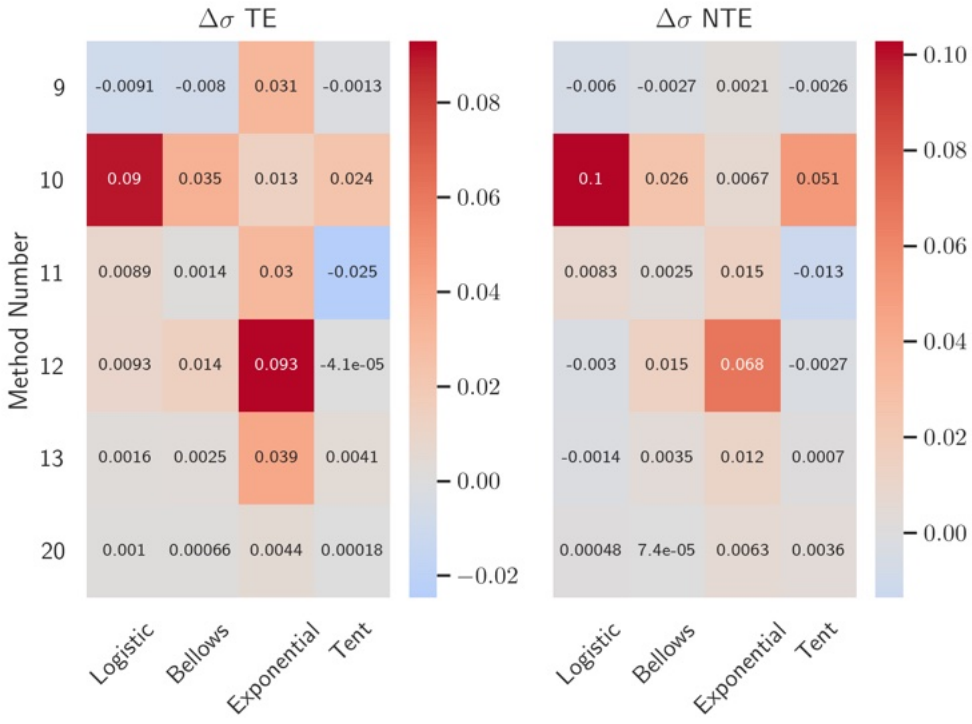


Figure 5.12: This figure shows the variance of the $n = 1000$-CML process empirical PDF of all maps at 100 probing points. These points are uniformly distributed over the respective CML processes value range for all binnings (right) and the extreme 2 (left) and 100 (middle) bins case. All binning methods which do not have h_j as output are omitted. The first row shows the variance of the common and the second of the adapted probability estimator, and the last shows the variance differences. We find that both methods perform better in specific arrangements. However, the standard estimator outperforms the adapted more often than not.



(a) $n = 300$



(b) $n = 1500$.

Figure 5.13: The heatmaps show the values of $\Delta(\text{N})\text{TE} = \sigma((\text{N})\text{TE}_{\text{Adapted}}) - \sigma((\text{N})\text{TE}_{\text{Common}})$ for the binning methods that allow for different bin sizes and thus exhibit differences with the estimator. Methods are enumerated as per table (5.1). We evaluate this difference for two sample sizes, $n = 300$ and $n = 1500$. Red values display the outperformance of the adapted estimator by the common one. The values are close to zero. Stronger deviations indicate a performance plus of the common estimator.

We are, of course, not only interested in the variance of the estimators concerning PMF estimation but also in the resulting (N)TE values. These are visible in the heatmaps in figure (5.13). Here again, we evaluate the standard deviation of the measure over multiple realizations of the CML system. Instead of evaluating probing points, we calculate the (N)TE measure directly. As with the probing approach for the probability estimation, we find both under- and overperformance of the adapted estimator compared to the common one. Most differences are fairly close to zero, though there are some outliers. When evaluating the different sample sizes, we see that the differences are, for the most part, more pronounced for smaller samples.

We conclude that these findings do not warrant any recommendation to utilize an adapted estimator that incorporates the bin size into PMF estimation. The approach of this thesis to include information in the bin width for PMF estimation did not prove fruitful. There are instances where this approach yields a smaller variance of results in both PMF estimation and subsequent (N)TE calculation. Nonetheless, we could not identify any systematic logic about when they occur.
Additionally, performance differences are generally minor in magnitude. Hence, it is advisable to use the ordinary PMF estimator, which has fewer hyperparameters as only the bin count and not the bin size contributes to the PMF estimation. Other approaches to extract and utilize this information might exist. Nonetheless, for the remainder of this thesis, we will utilize the standard PMF estimator.

5.8 Maximum Information Transfer Criterion for Partition Resolution

The above analysis showed no superior binning method. Nonetheless, we know that different binnings yield a different (N)TE value. Nevertheless, we have no criterion to choose a "true" (N)TE value since we need discretized values of an underlying continuous variable. A remaining goal, however, is to make at least the estimation consistent. Therefore we will use the two counteracting effects we first described in section (5.5) that govern the resulting (N)TE value. On the one hand, these effects are the (N)TE increase due to a more detailed partition and the ability to resolve more information for finer resolution, and on the other hand, the decrease of (N)TE due to the finite sample effects. These effects yield filling levels for bins that do not match the underlying density if too many bins are utilized for a given sample size.

We can take advantage of these effects by scanning the number of bins. Resultingly we find a binning for which the (N)TE exhibits a local maximum since the information increase is steepest in the low-resolution regimen, whereas small sample effects dominate for finer partitions. This relation becomes apparent for the TE and $\log m$ NTE measures in figure (5.14). The small sample desaturation of the data with increasing resolution causes the H NTE to converge to 1 as the measure's constituent joint and conditional entropies reach their maximum. Therefore, at least for the TE and $\log m$ NTE measure, we can apply a maximum transfer entropy criterion and utilize the binnings for the variables X and Y for $\mathrm{TE}_{X\to Y}$ and $\log m$ $\mathrm{NTE}_{X\to Y}$ that maximize the quantities. In doing so, we incorporate both the increase caused by increased resolution and the decrease caused by desaturation by finite sample effects. At the maximum, these effects have a similar magnitude. We can infer the same effect when comparing the evolution of the $\log m$ NTE value with the number of bins. The fit (red line) is calculated on the $\log m$ NTE evolution from 2 bins to its maximum with the formula for a (N)TE increase solely driven by the increase in resolution, $\propto a + b \log x$. We can see that the finite sample effects, therefore, drive the decrease of the measure. Additionally, the empirical NTE value increase is steeper than the solely resolution-driven increase, which suggests the contribution of the actual underlying information flow that is detected.
This method still depends on the sample size as larger samples allow for a finer resolution and subsequently more detected information flow and desaturation at a higher number of bins.

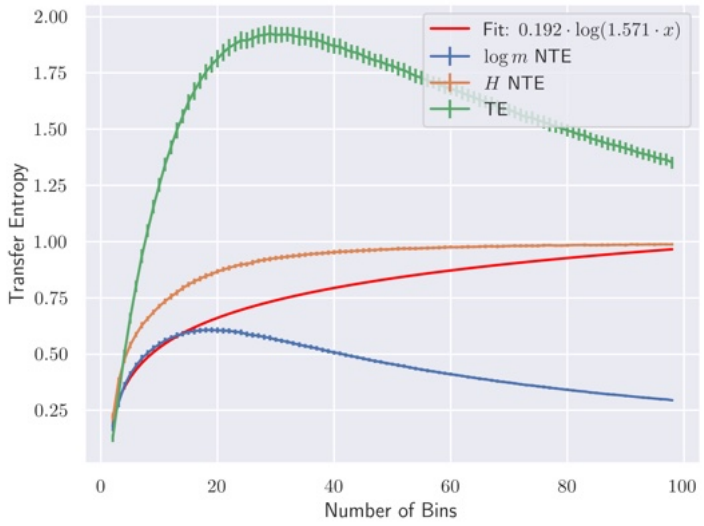


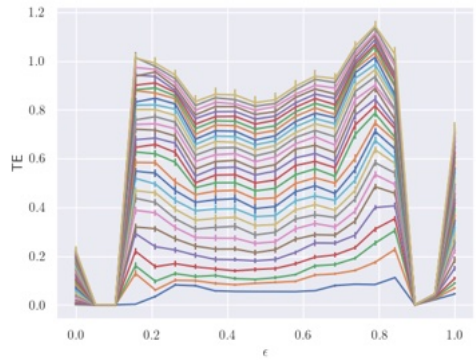
Figure 5.14: This figure shows that TE declines with increasing bin counts due to finite sample effects. This effect is preserved with the $\log m$ normalization but not for H NTE. The fit is the only binning driven TE increase by increasing resolution fitted to the $\log m$ NTE until its maximum. This visualizes the $\log m$ NTE decrease induced by finite sample effects.

However, this fact is valid for the transfer entropy measure in general.

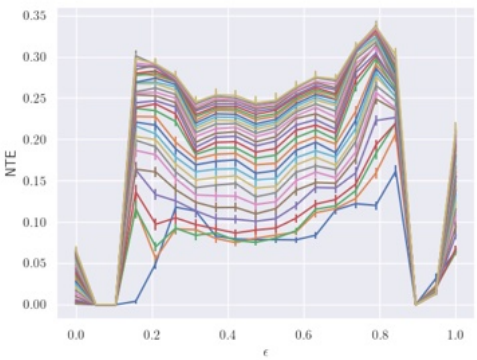
Ideally, we would prefer a measure that does not incorporate sample effects without correspondence to the underlying process. However, we assume that these finite sample effects are similar in shape and magnitude for all samples, and thus, we incorporate a consistent bias into our model and keep the desired comparability of the quantities. Still, larger samples can yield higher TE as we then resolve more information. We can mitigate this effect by utilizing the $\log m$ NTE. Therefore, we have established the superiority of the $\log m$ normalization for the transfer entropy, at least for our purposes. The measure allows for a maximum point criterion that enables consistency and, therefore, comparability of the measure between different samples or processes. Additionally. as discussed above, the variation of the $\log m$ normalized measure stays closer to the variation of the free TE. It also has the lowest absolute errors for a given sample size. Therefore in the subsequent section, we will drop the discussion of the H NTE normalized quantity. If we refer to NTE in the subsequent sections, we refer to the $\log m$ NTE measure.

5.9 Influence of Gaussian Remapping

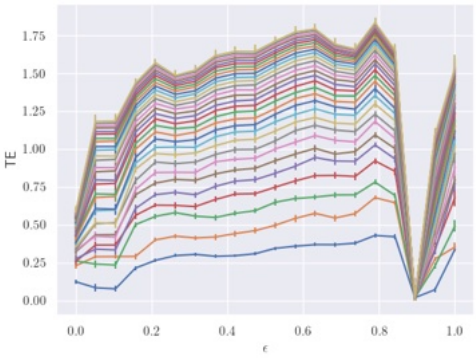
In the preceding sections, we have encountered the problem that we generally do not know the analytical PDF of the constituent processes for the (N)TE calculation. This knowledge gap be circumvented by applying a remapping of the sample onto a known (in our case a Gaussian) distribution as described in section (2.5.1).
We will now evaluate the remapping effect on the information flow. The result for the exponential map is displayed in figure (5.15). The remapping severely changes the variation of the exponential CML with the coupling strength ε. For a constant partition detail, the information flow increases when the remapping is applied. The non-linear normalisation then yields the more pronounced squeeze of the normalised measure when comparing (5.15c) to (5.15d) than for (5.15a) to (5.15b). Additionally, the errors are larger for the remapped data (N)TE measure.
A skewing of the calculated (N)TE is undesirable. Therefore, the trade-off to know the shape of



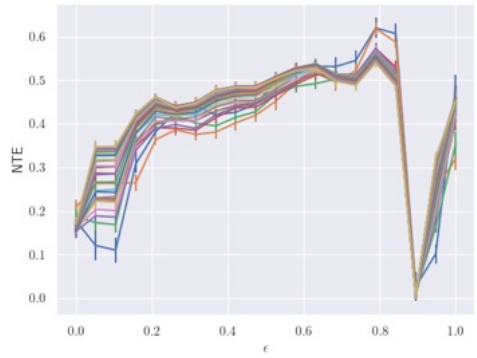
(a) TE.



(b) NTE.



(c) TE of remapped data.



(d) NTE of remapped data..

Figure 5.15: This figure shows the effects of the Gaussian remapping onto the variation of the information flow with ε for the exponential CML. We find a severe skew of this variation and increased errors. This skew indicates that the Gaussian remapping of data significantly alters the results of (N)TE calculation.

the underlying process is not worth the cost, and we will restrain from applying the method.

5.10 Effects of Data Rescaling

Intuitively, the rescaling of the data should not affect the final (N)TE value as we shift and scale all parameters uniformly. The study within our CML systems indicates these assumptions to be correct. Nonetheless, if we rescale the data to very small or large values, numerical inaccuracies might weigh more. For the data ranges we applied in this thesis, this effect proved negligible. We will apply rescaling in the subsequent sections about noise dependence and surrogatization.

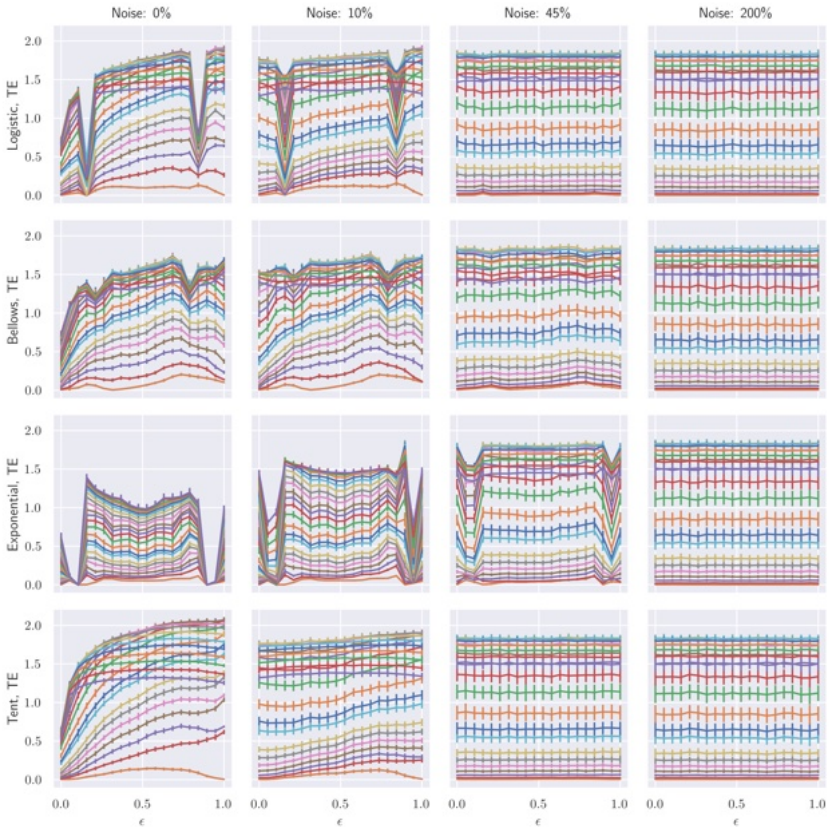
5.11 Noise Dependence

One of our (N)TE calculus applications is onto financial time series. These data mostly have a low signal-to-noise ratio [20]. Therefore it is instructive to evaluate the effects of noise onto (N)TE calculation. To accomplish this feat, we rescaled the data onto the $[0, 1]$ interval. Then we add standard normal white noise that we scale with factors between 0.1 and 2. The results are plotted in figure (5.16). There we can observe multiple effects. First, we find the intuitive notion confirmed that added noise increases the (N)TE value error. Additionally, noise washes out the underlying structure of (N)TE variation with the coupling strength ε. This effect is more prominent for finer than for less detailed resolution. Nonetheless, less detailed partitions can still resolve some of the underlying structure. However, if we add large amplitude noise, all processes yield a (within errors) constant evolution of (N)TE with ε. In this high noise regime, the final value of the quantity is purely noise-driven. We can also observe a similar effect for the higher resolution binning. This effect is even present with zero noise where the resolution has a finite sample effect induced maximum that causes lower resolution (N)TE evolutions to cross the ones with higher resolution. For high noise amplitudes, these resolutions yield (N)TE values constant over ε variation at this point.
Another noise-induced effect is the non-zero (N)TE at ε values where the unperturbed CML runs into intermittency, periodic orbits, or fixed points and thus exhibits vanishing (N)TE. Due to the added noise, here, the (N)TE values are greater than zero.

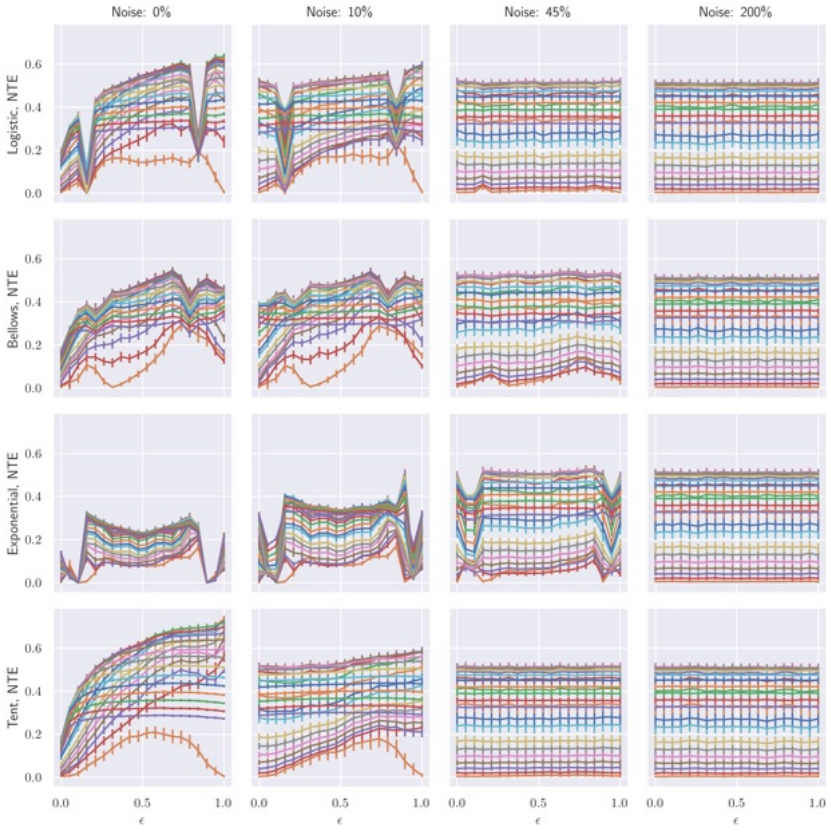
We can conclude that noise washes out the (N)TE values. Lower resolution proves more robust than higher resolution for this application. Sometimes lower resolution variation lines intersect those with higher resolution. We utilize this effect in the (N)TE maximizing binning criterion.

5.12 Surrogatization

We evaluated two surrogatization methods: shuffling surrogates and Fourier transform surrogates. Figure (5.17) shows the effect of this surrogatization onto the data for the Bellows map example. We observe that shuffling surrogatization fully destroys the variation of the (N)TE measure with ε. The rightmost column of the plot that shows the Δ(N)TE $=$ (N)TE(Raw) $-$ (N)TE(Shuffled) looks almost like the variation with ε of the raw data. The low ε regime is an exception, however. Here shuffling introduces higher (N)TE resulting in a negative Δ(N)TE value. We must keep that circumstance in mind when evaluating time series with low information flow against surrogatized data.
Shuffling surrogates (figure 5.17a) result in the variation with ε of the (N)TE values to straighten out. Nonetheless, the resulting relation still shows some features, such as an increase and decrease at low and high ε values and two local minima, which are present for finer partitions. The system entering periodic orbits causes these minima. This orbit allows for fewer different values and, therefore, desaturation occurs earlier for the increase in resolution.
The Fourier transform surrogates (figure 5.17b) exhibit different behavior as they keep the linear

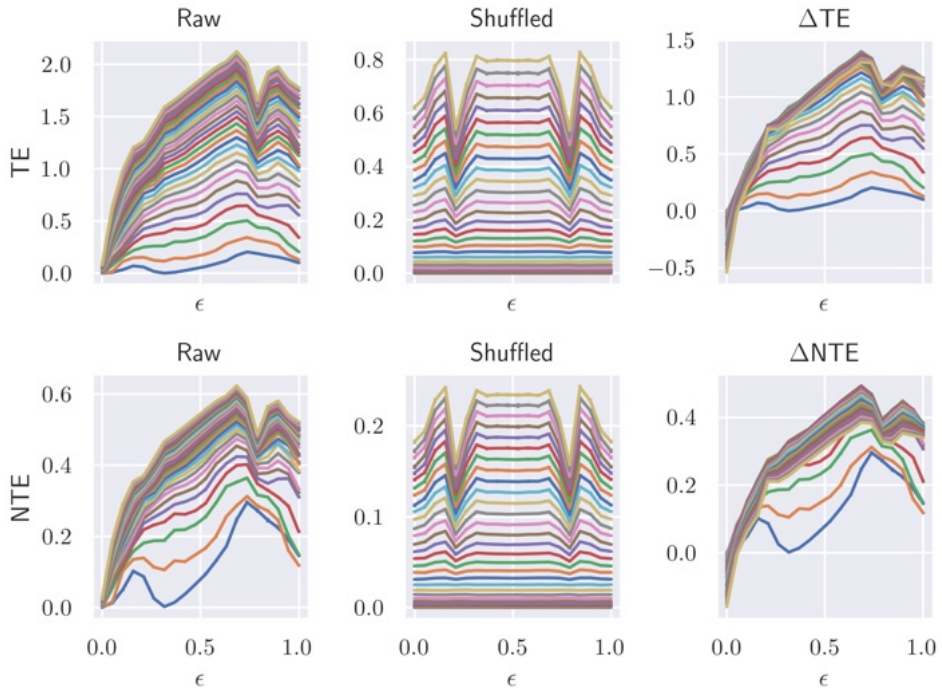


(a) TE.

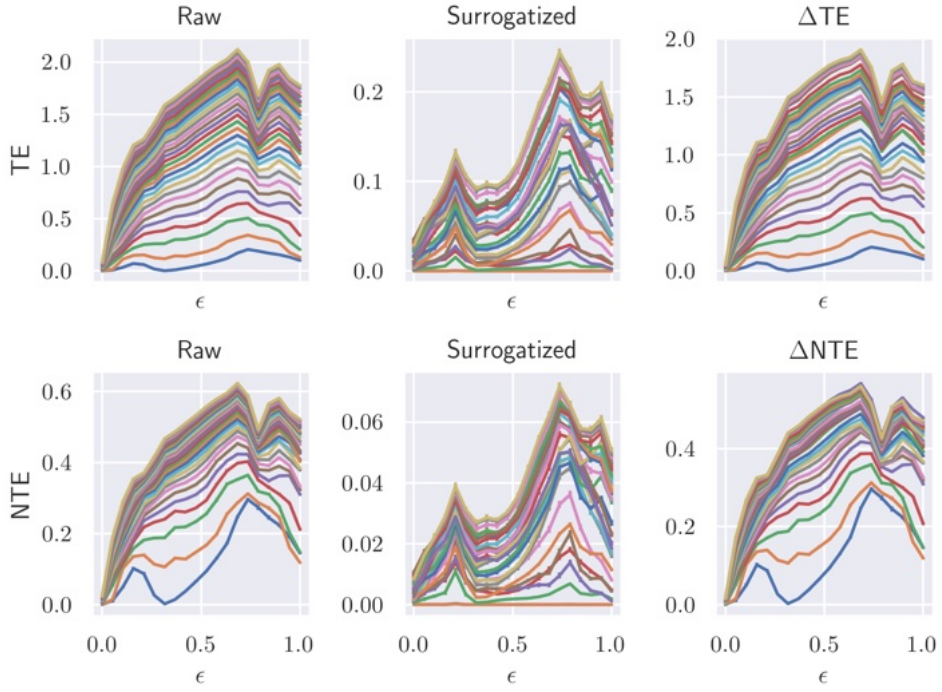


(b) NTE.

Figure 5.16: These figures show the variation of the information flows with ε if we add scaled standard normal white noise (in % of signal range) for the data rescaled to $[0,1]$. We find an increase of the error with an increasing noise level. This increase coincides a washout of signal. The lines correspond to a geometrically spaced subset of 2 to 100 bins on a dataset of 300 points. We have evaluated the TE (5.16a) and $\log m$ NTE (5.16b).

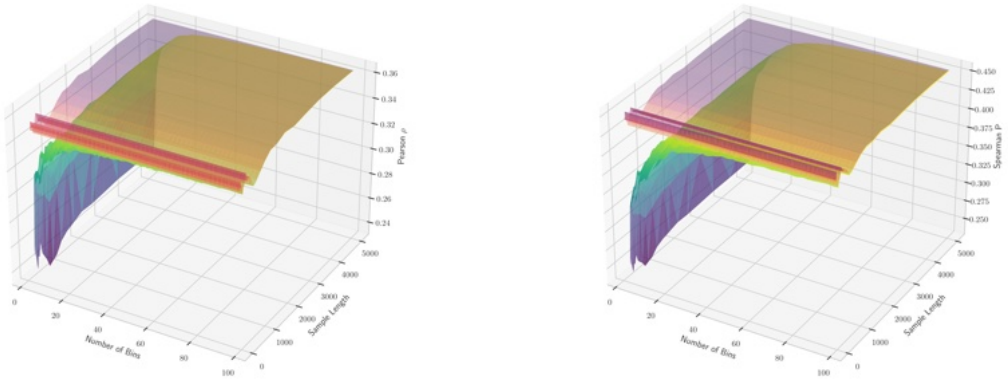


(a) Shuffling Surrogates



(b) FT Surrogates.

Figure 5.17: This figure shows the (N)TE for the raw (left column) and shuffling-surrogatized (middle column) Bellows map CML. Within the rightmost column we see Δ(N)TE = (N)TE(Raw) − (N)TE(Surrogatized or Shuffled) . We find that this quanitity exhibits the same evolution as the one of the raw data-set. This indicates that the measured information flows are not stochastic in nature but reflect characteristics of the underlyin process.



(a) Pearson ρ). **(b)** Spearman P .

Figure 5.18: This figure displays surfaces of the variation of Pearson ρ and Spearman P with binning and sample size (green) as well as for the unbinned data (red). We see that for an increasingly fine partition the measure values covnerge to the value calculated from the unbinned data.

relation and destroy non-linear relation between the variables. Therefore, the ε variation of the (N)TE measure on the surrogatized data is still feature-rich. Nonetheless, the amplitude of the residual information flows is far smaller such that the Δ(N)TE variation preserves shape at a lower magnitude. FT surrogatization has a caveat. The method broadens the spectrum of the data (even into ranges outside the image of the underlying map). Therefore one has to apply a rescaling onto the original data range of the surrogatized dataset. Only then can we use the same partition for (N)TE calculation of both samples while ensuring comparability. We can also test the FT surrogatized data against shuffling surrogates to detect non-linear relations significantly larger than zero using equation (41).
All in all, this analysis suggests that we can utilize the shuffling to test for non-zero information flow between the variables and the FT surrogates to test for non-zero information flow via non-linear relations. In this thesis, if the Δ(N)TE values are smaller than zero, it is assumed to be consistent with zero as in this case, any information flow in the unprocessed sample is minor than information transfer between random time series and thus considered negligible.

5.13 Comparison with linear Measures

As the last step to evaluate the (N)TE calculus, we compare it to the linear measures Pearson ρ and Spearman P. In doing so, we also apply the shuffling and Fourier transform surrogatization. Results are displayed in figure (5.19). Figure (5.18) shows the variation of the two linear measures with sample size and partition detail as well as for the unbinned data.

We will first start with discussing the linear measures before we compare them to (N)TE. Once we employ a sufficient sample size and partition detail, the measure generally becomes independent of these sample characteristics. This property is very desirable as it reflects the underlying relations. In opposition to the (N)TE measure, ρ and P can process the raw data without a scale change. The low opacity red surface reflects this in the plots in figure (5.18). We can see that a low detail binning destroys so much information that the true ρ or P values are severely underestimated. This effect becomes more prominent the smaller the sample size becomes. Another effect is the overestimation of the measure for small sample sizes. This overestimation might be due to

the overfitting of the small sample, meaning that the measure interprets random characteristics as a signal. The effect of this randomness only dissipates for larger sample sizes.
In figure (5.19), we recover the finding that the linear methods show less dependence on the partition detail. The variation with ε for both ρ and P almost coincides with the unbinned (dashed) variation line. The variation is different for the (N)TE measure that shows finite sample desaturation effects for the finer partition.
A related effect also occurs for the linear measures on the FT surrogatized data: low detail partitions that allow for only a few data values yield very high ρ or P values, respectively. All in all, the measures still infer a linear relation from the residual data, consistent with the FT surrogatization that only destroys non-linear relations.
For the (N)TE measure, the binning of the data introduces measured information flow even for randomly shuffled data, though the variation with ε is void of features. This voidness is induced by the uniform random assignment of the data points of a fixed sample on a finite data range to the different bins, giving rise to some probability differences conditioned on the influencing variables past. For the linear measures shuffling of the data destroys any relation. However, it is notable that for the (N)TE measure, the residual information flow after shuffling is still greater than the information flow between FT surrogatized data.
We can conclude that surrogatization and data shuffling destroy the (N)TE variation features with ε. Other than the linear models, however, there is still residual information flow inferred. Its amount depends on the partition detail level.

5.14 Summary of the Method Evaluation

Now, let us review the main findings from the (N)TE measure study with the four non-linear CML systems. The first main observation is that the amount of information flow that can be quantified depends on the detail of the partition. A finer resolution allows for the detection of more information flow.
Regarding the normalization, we found that the division by the entropy of the influenced variable conditioned on its past, H NTE, underperforms the $\log m$ NTE normalization as the dependence of the normalizing entropy on the sample skews the measure. Nonetheless, this normalization might prove helpful in other applications if researchers can adequately address its drawbacks. Then the interpretation of the H NTE is more intuitive than that of the $\log m$ NTE.
Nonetheless, in this thesis, we will apply the latter normalization since we work with small samples that then greatly suffer from the skew of the H NTE normalization.
Additionally, we found that sample sizes of at least 100 points are required to have an error of $< 5\%$ for the (N)TE measure. Of course, bigger sample sizes are preferable.
Regarding partition detail, we reconfirmed the notion that smaller samples result in larger errors of the measures. Additionally, smaller samples cannot resolve some underlying information flow dynamics, which can be inferred by a flattened out variation of (N)TE with ε that shows fewer features.
We also evaluated several different binning methods but found no method to outperform others consistently. The smallest standard deviation for (N)TE values was for methods that yield a criterion for the number of equal-sized bins. However, for these methods, the final value now depends on the sample size. A sample attribute that does not relate to the dynamics of the underlying system that we studied.
Regarding the evaluation of binning methods, we also studied the behavior of an alternative probability estimator that incorporated the bin size as an information-carrying parameter. Trivially. for equal-sized bins, its value was equal to the standard probability measure.
For unequal bin sizes, it sometimes showed lower and sometimes higher variance than the ordinary estimator. This result is deemed unsatisfactory for further use of the adapted probability estimator, and we will proceed with applying the (N)TE measure in this thesis with the standard

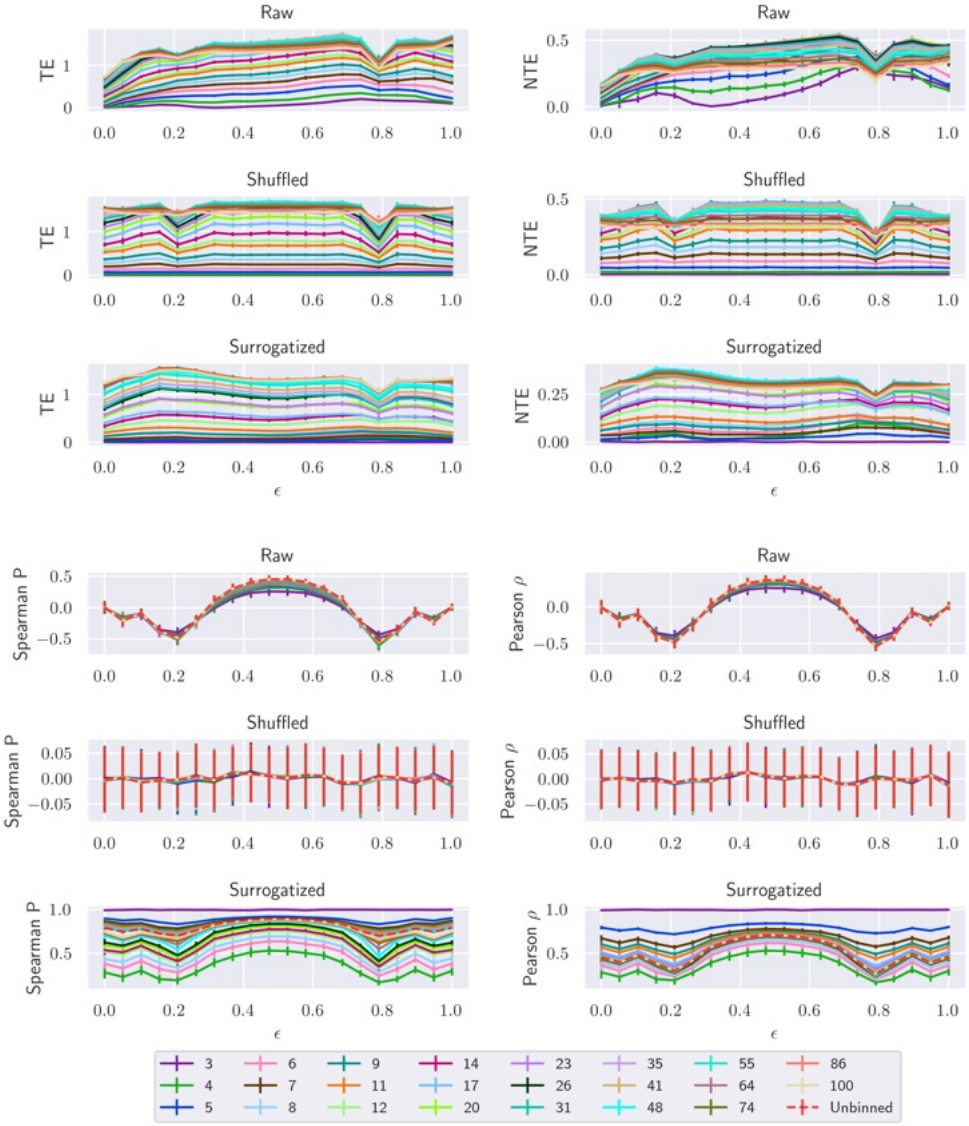


Figure 5.19: This figure shows the effects of data surrogatization. Shuffling destroys all relations. Surrogatization with Fourier transformation and phase randomization only destroys the non-linear relations.The first method flattens the information flow evolution ε. The second method shows a spread of the data (that persists rescaling). This effect yields high (N)TE values for rough partitions.

probability estimator.
As these two approaches did not prove helpful to derive a "proper" binning method, we utilized the two counteracting effects of increasing information flow and the finite-sample washout of information flow with the partition detail. These effects cause a local maximum for the $\log m$ NTE measure. We choose the binning that maximizes $\log m$ NTE. We incorporate some contribution of information flow that is not caused by the underlying process but by finite sample effects. However, we assume this contribution to be equal for all samples, and thus, we incorporate a consistent bias into the model and, therefore, still retain relative comparability, which is the desired property of normalization. This criterion cannot be used with the H NTE normalization as the normalizing entropy normalizes for the desaturating effect.
Additionally, we evaluated the effects of data processing. We found no changes in (N)TE values by uniform rescaling of the data and all relevant sample parameters.
We utilized rescaling to introduce and study the dependence of the (N)TE quantity on the application of scaled white noise. These experiments confirmed the notion that more noise washes out the features of the (N)TE variation with the coupling strength ε of the CML system. The resulting (N)TE value then is primarily dependent on the partition detail.
Surrogatization is a method that destroys the non-linear relation between time series while retaining the linear ones. This method results in a spread of the sample range. If one wants to calculate and compare (N)TE values pre and post surrogatization, rescaling the surrogatized data on the original data range is required. This way, a consistent binning can be used, and the quantities remain comparable, with only the non-linear properties of the sample being subject to change.
Lastly, we compared the (N)TE measures with the linear measures Pearson ρ and Spearman P. We found that the linear measures are less sensitive to sample size variation or partition detail and do not suffer from quantifying residual relation of shuffled data. Nonetheless, (N)TE can distinguish linear and non-linear relations via the application of surrogates. A feat of which the linear methods are not capable.

Recipe for (N)TE calculus This study leaves us with the following recipe we will apply in NTE calculation in the subsequent chapters:

- Apply autocorrelation function to determine lag for (N)TE calculation
- Scan the equal-sized bin number of the constituent time series that maximize the (N)TE.
- Sliding window analysis of sufficiently sized windows to obtain the temporal evolution of the (N)TE over time.
- For comparability, sample sizes should be roughly equal.
- Apply FT surrogatization to each timestep and deduct this linear information contribution from the (N)TE to quantify non-linear relation. Repeat this sufficiently often to obtain a statistically valid result.
- If applicable, utilize linear measures to quantify the linear relation.
- If applicable, further processing of the resulting (N)TE time series.

Limitations This recipie has the following limitations. First, the value is still dependent on sample size as larger samples allow for a finer resolution until desaturation occurs. This enables the detection of more information flow and a higher transfer entropy. This effect is persistent even with the $\log m$ normalisation as a finer partition uncovers more information flow within the same

data.
Additionally, the method rests on the assumption that the finite sample effects are introducing a consistent bias that is equal for all samples. This recipe has the following limitations. First, the value is still dependent on sample size as larger samples allow for a finer resolution until desaturation occurs. Thus, a larger sample detects more information flow and higher transfer entropy. The actual information flow bounds the measure's value. However, in insufficient sample size regimens, the amount of information flow underestimation is a function of sample size. Even with the $\log m$ normalization, this effect is persistent as a finer partition uncovers more information flow within the same data.
Additionally, the method assumes that the finite sample effects are introducing a consistent bias that is equal for all samples.

6 Information Flows Between Futures

We will now apply our (N)TE calculus to a non-linearly related time series system. Therefore, we will evaluate the information flow between future index returns during the (onset of) the COVID-19 pandemic. To ensure comparability of the results, we will utilize the $\log m$ normalized measure and from now on use NTE and TE interchangeably to refer to it.

This analysis is explorative because we have no rigorous underlying theory that yields a hypothesized functional relation that is testable. This circumstance is precisely why the TE as a non-parametric measure is a good measure for this application. Nonetheless, this investigation is motivated by the *a priori* hypothesis that we expect a change in the information flow caused by the turmoil and general insecurity caused by the global pandemic. We expect public discourse and opinion to drive this change in information flow. To test this second-order hypothesis, we will employ the economic policy uncertainty index (EPU) and the sentiment in an online discussion in the Reddit forum. The first serves as a proxy for public discourse, whereas the latter is a proxy for the private sentiment.
Additionally, we will qualitatively compare the change in information flow during the COVID-19 onset with the changes in information flow during other crises, namely the dot-com bubble, the $2007/08$ housing crisis, and subsequent Lehmann Brothers bankruptcy. We will commence this analysis by introducing the data we utilize.

6.1 Data

In section (5.4), we have established the data needs of the TE measure. Therefore the commonly available data with the daily resolution is insufficient for our purposes. With this level of detail, one could infer a value for TEy for the whole time frame. We, however, are interested in the temporal evolution of TE and therefore want to employ sliding windows (see section 2.5.4). This interest is why we need intraday trading data. We need sufficient data to robustly resolve information flow on a sufficient granularity to uncover its temporal changes. This data is more difficult (and more expensive) to procure. Nonetheless, we were able to acquire some intraday trading data for future indices. Sadly, it was not possible to get the data of the same indices for all periods. Nonetheless, there is still residual informative value in comparing these quantities, even though they stem from evaluating differing underlying values.
We will start by introducing the data used for the study of the COVID-19 pandemic.

6.1.1 COVID-19 Onset Futures Intraday Data

To analyze the information flow, we use the intraday data of index and commodity futures. These futures are contracts to transfer ownership of a specified item (goods, stocks, currency, etc.) at a specified time in the future for a specified price. Market makers such as an exchange facilitate these future contracts. Therefore, the trading parties do not necessarily need to know each other.
A rational (in the sense of value-maximizing) actor who buys a futures contract expects the value of the underlying item to increase. This is called a *long* position. If the buyer's expectation is met and the price of the item exceeds the price of the futures contract, the price difference is profit as the buyer still acquires the underlying value for the predetermined price. Vice versa, a seller of a futures contract, is holding a *short* position and expects the value of the underlying item to decrease. If this expectation is met, the seller takes a profit since (s)he can acquire the item for a lower price than what the buyer of the futures contract will pay.

We use these futures as a proxy for the sentiment of market actors. This approach is motivated by the fact that expectations about the situation's future evolution rather than past events drive the

futures prices.
Additionally, we will convert the raw price data into logarithmic returns:

$$r_i = \log\left(\frac{p_i}{p_{i-1}}\right) \tag{109}$$

for the prices p_i. The conversion into returns converts the time series into quantities that show self relative fluctuations and thus enables comparability of indices with prices in different orders of magnitude. The logarithmization is a bijective non-linear transformation that now associates the sign of r_j with the relative motion ($-$ indicating a decrease and $+$ an increase). Additionally, they ease computation since upsampling of the data is done by summing over the log returns in the respective period.

Let us now proceed to introduce the set of futures we incorporate into our analysis. Though the names of these quantities sometimes seem unintuitive, we utilize the ticker symbols of the respective futures throughout this analysis. We evaluate futures of the following indices:

- VG1. Euro Stoxx 50. The underlying index includes 50 large, publicly traded companies headquatered within the euro zone.
- ES1. S&P500. The underlying index includes 500 of the largest US-american corporations. Their contribution to the index is weighted by market capitalization.
- HI1. Hang Seng Index. The Hang Seng includes 50 of the largest companies traded on the Hong Kong stock exchange. Their contribution to the index value is weighted by market capitalization.
- NK1. Nikkei Index. The Nikkei index includes 225 of the largest companies publicly traded on the Tokyo stock exchange.
- CO1. The CO1 is a future on the crude oil commodity.

To evaluate the temporal evolution of the NTE between these quantities, we employ sliding windows covering one day. The results are daily NTE values between the intraday data of different futures time series. The sample sizes are similar so that our entropy maximizing approach yields comparable results.

Autocorrelation In order to determine the proper time lack for NTE calculation, we evaluate the autocorrelation of the time series. The result is displayed in figure (6.1). Except for the Reddit sentiment data, autocorrelation indicates no self-association of the variables. The quantity decreases steeply and fluctuates around 0 for time deltas $\tau > 0$. The Reddit data shows a consistent self correlation at around $\rho(\tau) = 0.1$ for $\tau > 0$. Therefore, we choose time lags of $K = L = 1$ for all futures time series and $K = L = 2$ for the Reddit data.

6.1.2 COVID-19 Onset Reddit Data

To evaluate the influence of public perception about the COVID pandemic and the according to sentiment onto information flow, we acquired a dataset from Reddit. Reddit is an online forum where users can create posts and comment about many topics in communities called subreddits. These communities are commonly referred to as *r/<community name>* mimicking the web address of the respective subreddits. In addition to posting and commenting, users can vote on the posts and comments. They then see the posts on the algorithmically curated newsfeed or visit and explore a particular subreddit on the main page.

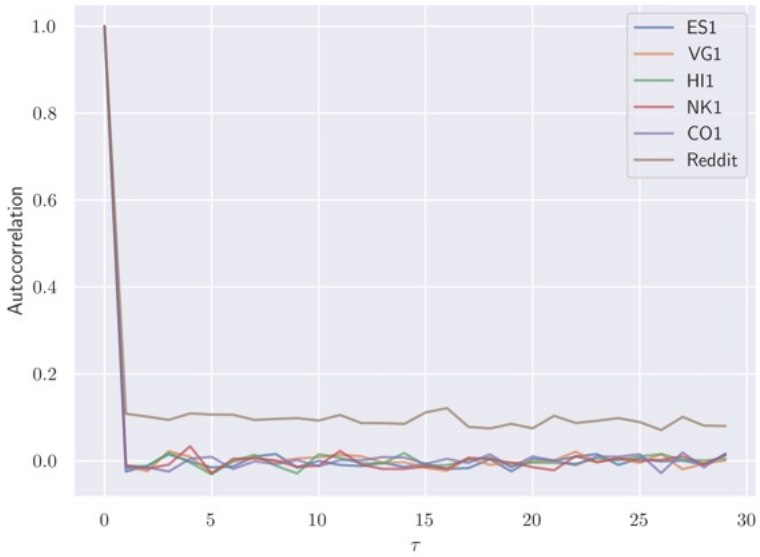
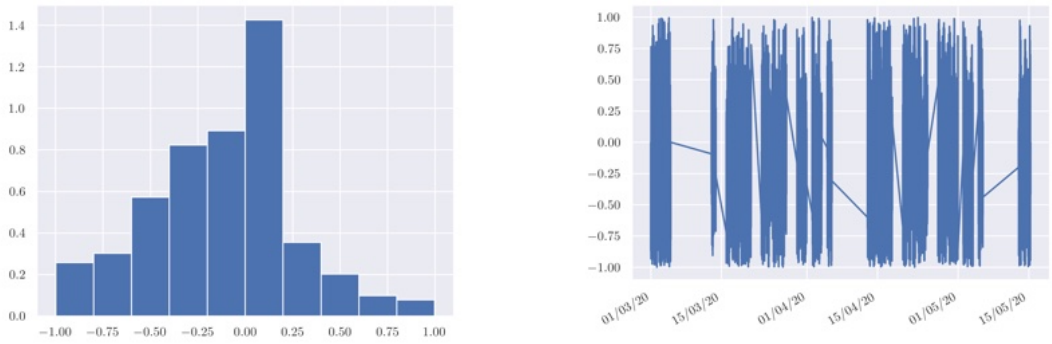


Figure 6.1: The autocorrelation of all indices included into our analysis. indicates that we can utilize $K = L = 1$ for all futures in our data set. For the reddit data we include two time steps into the past as the autocorrelation is higher.

The data was acquired using the API to the pushshift Reddit dataset [6]. This API, however, limits the number of requests per time. For the context of this thesis, this is why it was only feasible to gather a limited amount of data. The pandemic itself, and certainly governmental response, has become a politicized and partisan topic. We, therefore, chose to acquire data from the subreddit *r/politics*, one of the largest and most political discussion forums on the internet [46]. Nonetheless, subreddits and (online) communities, in general, might exhibit strong partisan bias that even might change over time [33]. However, we chose to sample from *r/politics*with the hope that its large and ideological diverse userbase mitigates some of these polarisation effects. This method was chosen over the alternative of hand-selecting smaller communities with a known political bias to aggregate a dataset with diverse opinions. We included every post or comment posted between the 29$^{\text{th}}$ of February 2020 to the 31$^{\text{st}}$ of May 2020 into the dataset when the contribution contained *covid* or similar keywords.

At this point, we have to note that this approach still has incorporated biases. First, internet access and the likelihood of a person to post an opinion online are likely to be a function of the person's socio-demographic background, economic position, political leaning and therefore skews our sample. Additionally, we are only sampling from the *r/politics* subreddit and might lose contribution from political fringe groups to our sample. Additionally, it is unclear whether the opinions posted online even reflect the actual opinion of the posting individual, as polarising statements tend to get more attention and are thus more visible. On top of that, we filtered the posts only for a limited number of COVID-19 related keywords. Thus, we did not include posts that used unrecognized alternative terms or wrong spelling. It is possible that the occurrence of these alternative signifiers of the pandemic and related measures is related to a poster's situation and thus the corresponding stance towards the pandemic response. However, incorporating more terms was not feasible because of the API request restriction. Nonetheless, despite these limitations, the analysis of the Reddit data can provide valuable preliminary insights.

We will now proceed to discuss how the data from *r/politics* was processed to enable the analysis employing transfer entropy. The data gathered from the pushshift API are the posts or



(a) Empirical PDF .

(b) Time evolution.

Figure 6.2: Time evolution and empirical PDF of COVID-19 related Reddit posts. There are timeframes in the data where other topics were dominating the discussion and hence only insufficient COVID related data could be retrieved.

comments content as well as the sum of the up and downvotes (where $+1$ corresponds to an up-and -1 to a downvote).
Subsequently, we applied a sentiment analysis using the Natural Language Toolkit [8] to assign a sentiment to all the coronavirus-related posts and comments in our dataset. This method returns a probability of a specific text to be either positive, neutral, or negative and a compound value $\in [-1, 1]$ where -1 indicates maximum negative and $+1$ maximum positive sentiment. We use this compound value for our analysis. To create matching time series of Reddit sentiment, we downsample the individual time-stamped post sentiments to a mean sentiment of all posts within a period to match the index values.

Sadly, we could not acquire a full dataset spanning the period with the available data because of API request limitations. The final dataset contains stretches without data that were zero-padded for analysis. The time series of Reddit sentiment before the padding operation is shown in figure (6.2b). Figure (6.2a) shows an empirical PDF of the sentiment. We observe a skew towards negative opinions about the pandemic.

6.1.3 Economic Policy Uncertainty

Another proxy to measure public sentiment is the Economic Policy Uncertainty (EPU) index [3]. The authors infer the index from newspaper coverage frequency. Articles are counted if they are published in one of ten leading US newspapers and contain the triplet: "economy", or "economic"; "uncertain", or "uncertainty"; and at least one of "congress", "deficit", "Federal Reserve", "regulation","legislation", or "White House". For the uncertainty in other parts of the world, the translations or names of the respective country-specific institutions were for country-specific newspapers. The evolution of the EPU index in the three-time periods covered in the subsequent analysis is shown in figure (6.3)
The time resolution of the EPU index is too low to apply the TE measure. Therefore, we only evaluate the relation of the EPU to the other quantities with the *Bravais-Pearson-Correlation*.

6.1.4 Dot-com and Housing Crisis Data

Whereas the main focus of this section is the evaluation of the time series data of futures during the onset of the COVID19 pandemic, we are evaluating other quantities for qualitative conclusions.

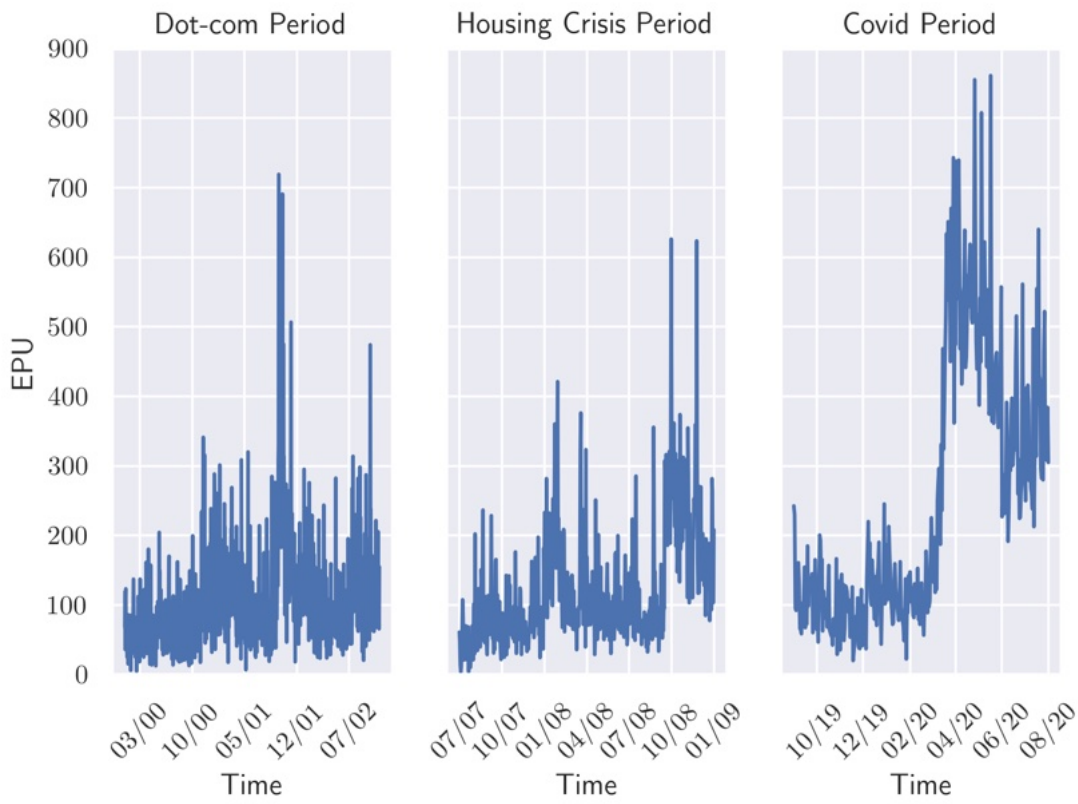


Figure 6.3: This plot shows the temporal evolution of the EPU index in the three periods that are evaluated in this section. We see that the COVID-19 EPU structure differs from the other two because of the sudden spike that not relaxes to a baseline value but remains at an elevated level.

Namely, we are studying the US housing bubble and the Dotcom bubble. Sadly, it was not possible to acquire sufficiently detailed data for the futures of the indices studied for the COVID-19 pandemic. Nonetheless, we were able to gather data for some of the underlying indices:

- STOXX50E. The Stoxx Europe 50 index is similar to the Euro Stoxx 50 index discussed above. The difference to that index is that the index includes European companies from outside the euro zone. Additionally, we are evaluating the index itself and not a derivative.
- TOPX. The TOPX index is an index containing 1111 Japanese stocks with a focus on electronics, financial services, and other modern industries. Again, we are evaluating the index itself and not a derivative.
- SPX. SPX is the ticker symbol for the S&P 500 discussed above. Again, we are evaluating the index itself and not derivatives.

The fact that these are stock return time series and not futures yields the limitation that we will compare different quantities of a different type. Thus, we can understand any conclusions drawn from this comparison merely to indicate a possible difference or similarity.

6.2 Crises Timelines

This section will briefly cover the timeline of significant events during the three crisis periods analyzed in this thesis. The central crisis we are covering is the (at the time of writing still ongoing) COVID-19 pandemic. Thus, our data roughly covers the first wave of the pandemic.

6.2.1 COVID-19 Pandemic

The COVID-19 pandemic started around November 2019 in Wuhan, Hubei, China, by zoonotic transfer (likely from bats) [42]. The onset of symptoms of the first officially recognized SARS-CoV-2 case occurred on the 1st of December 2019 [35]. Antibody studies suggest that the virus was present in the United States by December 2019 [5]. The first case recorded by authorities occurred on the 20th of January 2020. The first confirmed case within Germany occurred on the 27th of January 2020 [66], in Hong Kong on the 23th of January 2020 [53], and in Japan on the 16th of January 2020 [49].
A typical response to the outbreaks was lockdowns. That is the closure of business and leisure activities. The scale and particular rules of the lockdowns vary by and even within countries.
The Federal Republic of Germany issued the first lockdown order on the 16th of March 2020 to take effect six days later [26]. Within the US, the lockdown decisions were up to the individual states. The earlies lockdown started in California on the 19th of March 2020 [12].
Hong Kong was fairly unscathed by the first waves of the pandemic and only imposed the first partial lockdown of COVID-19 hotspot areas on the 10th of December 2020 [54]. In Japan, the government does not have the authority to issue a lockdown order or penalize non-compliance. Nonetheless, the advice of government agencies to self-isolate has largely been followed [59].

6.2.2 US Housing Crisis and Lehmann Bros Bankruptcy

In the years preceding the subprime and subsequent financial crisis, housing prices in the US steadily rose. This increase is why buying homes on credit was incentivized, and credits were given to people who could not afford them while accepting the to-be-purchased property with its assumed increase in value as a security. These credits were coined subprime loans. Starting in 2006, property prices began to stagnate, and interest rates increased. This affected borrowers who were holding loans with adjustable interest rates. Sometimes, these borrowers opted out of the loan, leaving the bank with the property as the house's value was lower than the loan plus interest.

Hence, there was an incentive for borrowers to default. This dynamic sat in a spiral of declining housing prices and more and more defaults.

The housing crisis spilled over into the financial sector because bundles of mortgages were sold as securities to investment firms. These securities also lost value rapidly as more and more constituent loans defaulted. On April 2, 2007, New Century Financial, the United States' largest subprime mortgage lender, filed for bankruptcy. Starting on the 18th of September 2007, the Federal Reserve Bank started cutting interest rates and agreed to lend money directly to Wall Street firms and not only commercial banks while accepting the mortgage-backed securities as collateral. The economic downturn continues, and about a year later on the 6th of September 2008, the US Treasury announced to take over the struggling mortgage giants Freddie Mac and Fannie Mae that jointly owned more than five trillion USD in mortgages. Thereby, it provides up to 200 billion USD to the firms to enable them to finance mortgages for banks and other home lenders.
Later that month, Merrill Lynch was acquired by Bank of America in order to establish trust insolvency of the institution that was damaged by severe losses due to exposure with subprime mortgages. However, a similar deal fell through with the Lehman Brothers investment bank. It filed for bankruptcy on the 15th of September 2008, and roughly 25 thousand employees of the firm were laid off. This bankruptcy was contrary to the assumed *too big to fail* for large investment firms and led to a large amount of distrust regarding money lending in the financial sector. The resulting loss of trust inhibited the flow of capital to the producing sector and damaged the economic development around the world for the upcoming years.

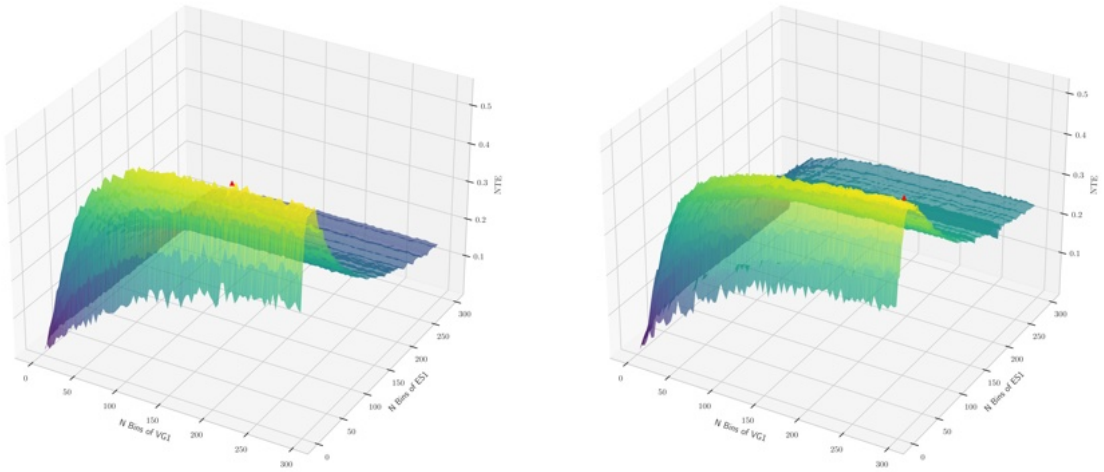
6.2.3 Dot-com Bubble

The introduction of the internet to consumers in the 1990s allowed the internet economy to emerge. Speculative venture financing firms backed many new business models. These venture capitalists rely on a later sale of their stake in a company to realize a return on their investment. Sometimes, this is facilitated with an initial public offering of the companies stock on a stock exchange. Until the dawn of the new millennium, these internet stocks were hyped and rose in value.

Nonetheless, some of them had to close shop as they could not generate a sustainable business model. As a result, the stock valuations began to plummet. The NASDAQ index ends the year 2000 at 2470.52 points - 52% lower than its peak in March of the same year at 5132.52 points.

6.3 Transfer Entropy During the COVID-19 Pandemic

In this section, we will evaluate the TE between the futures during the COVID-19 pandemic. Therefore, we start with evaluating whether the entropy maximizing binning criterion devised in the CML model systems yields promising results when applied to real-world data. Figure (6.4) shows example surfaces of the TE value by bin variation of the relation between the VG1 and ES1 futures on the 24th of March 2020. There we do see that the partition detail variation of the measure constituent variables does yield a maximum. However, the desaturation is far more prominent for the ES1 variable than for the VG1 variable. The TE value decreases from a peak towards a constant value for an increased resolution of the ES1 future time series. For the VG1 future time series, the value of the TE fluctuates less, and there is no apparent decline within the scanned bin range. Therefore, we will evaluate these maximum binning pairs individually for all sliding window batches of the time series data.
The temporal evolution of the NTE with daily resolution at the onset of the first wave of the COVID-19 pandemic is displayed in figure (6.5). We see that the NTE for all directional index



(a) $\mathrm{NTE}_{\mathrm{ES1}\to\mathrm{VG1}}$. **(b)** $\mathrm{NTE}_{\mathrm{VG1}\to\mathrm{ES1}}$.

Figure 6.4: This plot shows the NTE value surface generated by the variation of the binning of the constituent time series ES1 and VG1. Maxima of the surfaces are indicated by the red triangle. Within the displayed bin ranges the desaturation effect of TE is far more prominent for the ES1 variable than for a finer resolution of the VG1 variable. While the latter exhibits a peak, it is far less prominent.

pairs fluctuates about a mean value. Nonetheless, there are extreme values, mostly towards lower values.

A striking feature is the increase of all NTEs starting at about March 2020. Before we evaluate this change in dynamics, we will cover qualitative differences in the pre-outbreak period.

During that time, the ES1 future has the most significant NTE contribution from the Hang Seng, HI1, future and the crude oil, CO1, future at a value close to 0.4. The other two - VG1 and NK1 - futures exhibit less information transfer with NTE values fluctuating at about 0.2. These NTE values increase for all futures above 0.4 at the onset of cases. In Europe, this occurred at the beginning of March 2020. The general insecurity might explain this increase regarding possible lockdown measures and the economic response toward the impending COVID crisis. The NTE values decrease for the information flow from the VG1 and NK1 in April 2020 but remain higher for the other two futures. During this last period, the information flow from the HI1 time series exhibits two atypical values. These coincide (at least partially) with an atypical minimum of the information flow from the VG1 future.

Notably, the information flow from all other evaluated indices towards the ES1 value converges for the peak information transfer of all indices in March 2020. The same effect occurs for all index pairs. Nonetheless, it is the most pronounced with the ES1 as the influenced variable.

Generally, we can observe that the crude oil and Hang Seng futures have a more significant influence throughout the period. We can speculate this to be caused by Hong Kong's proximity to China, the outbreak epicenter, and the fact that investors might understand the Oil future as a proxy for global trade. However, we must note that the first wave was reasonably mild in Hong Kong and that the government implemented no lockdown measures. In the US, lockdown orders went into effect later than in Germany, for example, and there were no lockdown measures in Japan. This fact can be an indicator of the lower NTE value from these regions.

This interpretation, however, does not explain why we calculated the different levels of NTE before the first cases.

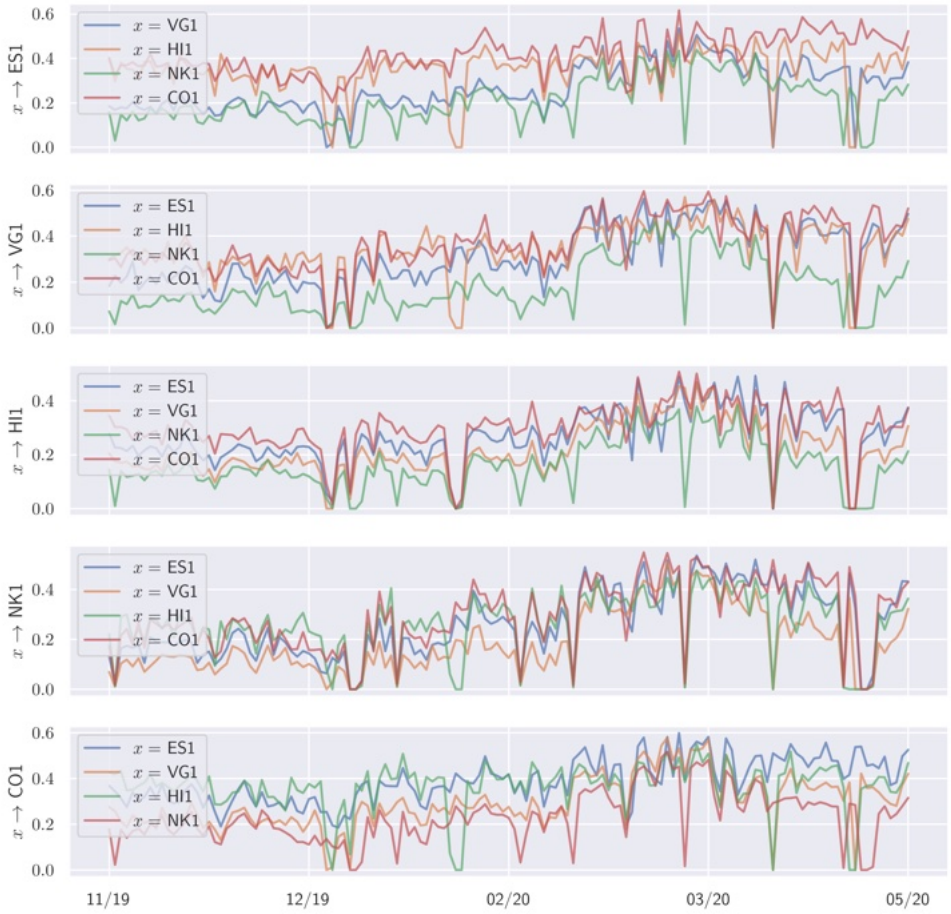


Figure 6.5: This plot shows the temporal evolution of the information flow towards the variable indicated on the y-axis. We see an increase in information flow in March 2020 consistently for most curves.

Let us now proceed to discuss VG1 as an influenced variable. Again HI1 and CO1 are exerting the greatest influence on the evaluated futures. However, the difference in information transfer between these influencing variables is smaller than for the previous case. Generally, the level of information transfer onto the VG1 future is slightly lower. The time series all exhibit the same increase at the beginning of March 2020. In this period, the magnitudes of the NTE of all futures pairs attain a similar magnitude as in the previously discussed ES1 case. However, in April 2020, we only see a decline of information flow from the NK1 future, whereas the other time series pairs yield an elevated NTE compared to the 2019 data.
While this data is not sufficient to draw definitive conclusions, the temporal evolution of the NTE from the other futures might be explained with the same reasons as above. Europe's slight lead might explain the additional influence of the ES1 future with the implementation of pandemic response measures that influenced US decisions.

Generally, the NTE information flow towards the HI1 future is smaller than the two previously discussed cases. It shows the same temporal evolution, but the information transfer before the increase in March 2020 is between 0.15 and 0.3. The peak of the increase in March is at 0.45. The information flow from all other indices subsequently decreases but remains higher than before the increase in the 0.2 to 0.4 range.
Throughout the evaluated period, the amount of information flow from lowest to highest originated from NK1, VG1, ES1, and CO1. Hence, again, the oil futures exhibit the most information transfer. The general lower level of the information transfer might be due to the geospatial proximity (at least compared to the US, Europe, or Japan) of Hong Kong to Wuhan. This proximity could indicate investors are less likely to be looking in other parts of the world for pandemic responses or its effect on the economy as these are expected to lag behind the Hang Seng Index in Hong Kong.

Let us now proceed to evaluate the information transfer towards the Japanese NK1 future. Like with all other variables, we see a considerable increase in information flow in March 2020. The absolute NTE values, however, are comparable to the previously discussed case of the HI1 futures. Primarily, the VG1 (S&P 500 derived) future shows a comparably small influence on the NK1 index below a value of 0.2 until the March increase. The other three influencing variables fluctuate around the 0.2 value, and their temporal evolution intersects. In March, all NTE values peak at around 0.4.
This temporal evolution and relative magnitude are surprising since the Japanese economy heavily relies on imports and exports. Additionally, to stay consistent with the reasoning above, we would expect a heightened influence of the HI1 futures time series. Maybe other mechanisms, unknown to the author, such as cultural preferences, trust in government responses, or similar methods, can explain this deviating evolution.

Lastly, we discuss the influence of the four index-derived futures on the oil future, CO1. The temporal evolutions follow the same familiar pattern and exhibit an increase in March 2020. Additionally, except for the March 2020 peak, the HI1 and ES1 futures consistently show a higher information flow towards the CO1 evolution than the NK1and VG1 indices. The values of the former fluctuate around 0.4, the ones of the latter two around 0.2. All NTE values peak at around 0.6 in March 2020. The subsequent decline is more prominent for the NK1 and VG1 futures than the HI1 and ES1 futures. All declines saturate at an elevated level compared to the values before March 2020.
Again, this co-evolution is striking since the US and Japan are also severely involved in global trade. Whereas one could theorize that the lower influence of the VG1 future is due to the large domestic market and oil reserves of the US, the same is not true for Japan.

The most striking feature of all the co-evolution of NTE is the increase of the measure in March 2020. Whereas we can only speculate for the causes of the different information flows, the increase in information transfer indicates a departure from market efficiency, which results in arbitrage opportunities [20].
It is important to note that all values contain contributions from small sample desaturation. Thus, they are comparable, but our approach does not conclude that another explains $x\%$ of one future time series.

Non-Linear Contribution We will now proceed to discuss the same measure evolution but calculated with FT surrogatized datasets. The values displayed in figure (6.6) are calculated by means of equation (41).
The first striking difference to the unprocessed data is that while non-linear information flow between the variables to some extent persists in the March 2020 period, it vanishes for some pairings in the period before that. Afterward, it is significantly reduced. Unsurprisingly, we also find a reduced magnitude of the persisting signals.
This effect is especially prominent with ES1 and VG1 as influenced variables. For these two, there is still significant non-linear information flow from the CO1 and HI1 futures. Except for the March 2020 period, the NK1 and VG1 or NK1 and ES1 futures contribution, respectively, is consistent with zero information flow. We now compare the non-linear information flow from the CO1 and HI1 to the ES1 and VG1 futures to the ES1 information flows in the unprocessed dataset. From this comparison, we can infer from the magnitude of the NTE values in March 2020 that about two-thirds of the information dynamics are due to non-linear relations.
Regarding the HI1 future, only the crude oil time series has a significant non-zero contribution outside the March 2020 increase. Nevertheless, even in that period, the non-linear information flow from the ES1 and VG1 time series remains lower than 0.15. The NK1 non-linear information flow flattens at about NTE $= 0.2$ whereas the CO1 non-linear information flow reaches 0.3.
The picture is similar to the NK1 futures as the influenced variable. Here, VG1 and ES1 future only yield information flow consistent with zero outside the March 2020 period and even within it is never higher than 0.1 with one exception at the beginning of March 2020 where the non-linear flow of information from ES1 to NK1 attains a value of about NTE $= 0.2$.
When we evaluate the non-linear flow of information towards the crude oil, CO1, futures, we find vanishing flow from the NK1 futures outside the Mach 2020 period. There, also the ES1 and VG1 information transfer values stay below 0.1. The information flow from the Hang Seng $HI1$ future fluctuates around that value. An increase in information flow from this index is barely noticeable. It is nonetheless for the information flow from NK1, ES1, and VG1. The last one exhibits a sharp peak at the beginning of March.[10]

From this analysis of non-linear information flows, we can conclude that flow from crude oil futures influences all the other indices. Additionally, there is non-linear information flow from the HI1 index futures to the ES1 and VG1 futures. The NK1 future seems to be largely unaffected by fluctuations in the other index futures (but not from crude oil futures).
This independence allows the speculative inference that oil prices are causally at an earlier point within the global economic evolution. If one follows this reasoning, however, the same must be valid for the HI1 index. There is no theory (that the author knows of) that suggests this empirical finding for the latter.

Bootstrap Measure Comparison We will now evaluate the information flow against the flow of shuffled time series, where all linear and non-linear information flows are purely by chance. The figure (6.7) then shows the results according to equation (41). The most striking difference to

[10] This peak does not coincide with April 20, 2020, the time where the COVID-19 pandemic drove oil prices negative.

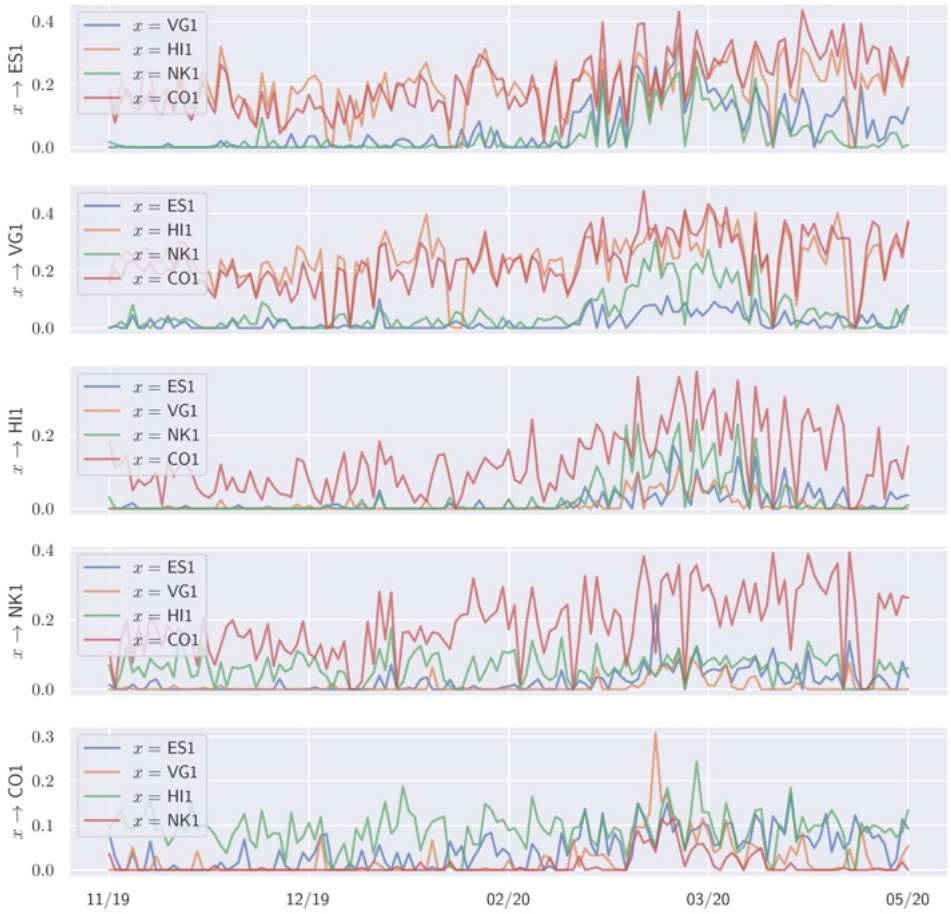


Figure 6.6: The figure displays the information flowing through non-linear relations between the futures. We see that few information flows toward the CO1 future but it is associated with all other variables in some way.

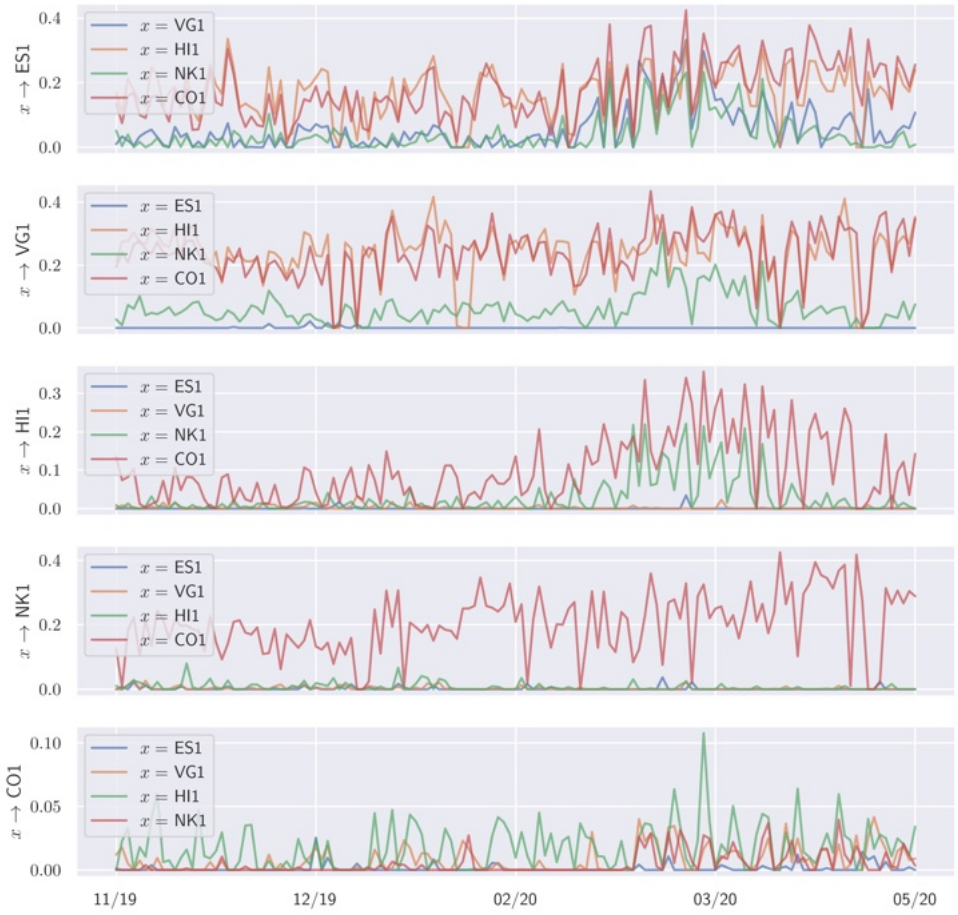


Figure 6.7: This figure shows the information flow between the index futures when we only evaluate the parts significantly different from 0. We can see that the influence onto the CO1 future is far smaller than its influence on other futures.

the previous results from the figure (6.5) are the vanishing information flows from NK1 and VG1 to ES1, from NK1 and ES1 to VG1, from VG1, NK1, and ES1 to HI1, and from ES1 and VG1 to NK1 at the pandemic onset up until the March 2020 increase. There the information flow from all to all futures is significantly different from zero.
Nonetheless, there remain great differences in magnitude. For example, the information flow from the NK1 and VG1 futures to the ES1 future peaks at about 0.2. It peaks at 0.4 from HI1 and CO1. With VG1 as the influenced variable, all information flows are smaller in magnitude. The flow from ES1 does not exceed 0.1. A similar fact is true for NK1 as the influenced variable. Here, only CO1 is significantly larger than 0.2 for extended periods. The information flow from the other indices, for the most part, does not exceed 0.1.

Across all the index futures, the information flow from the crude oil future is the largest. With ES1 and VG1 as the influenced variables, the information flow from CO1 is similar to that from HI1. Notably, ES1 exerts no influence onto VG1 when corrected for the stochastic contribution by bootstrap shuffling.

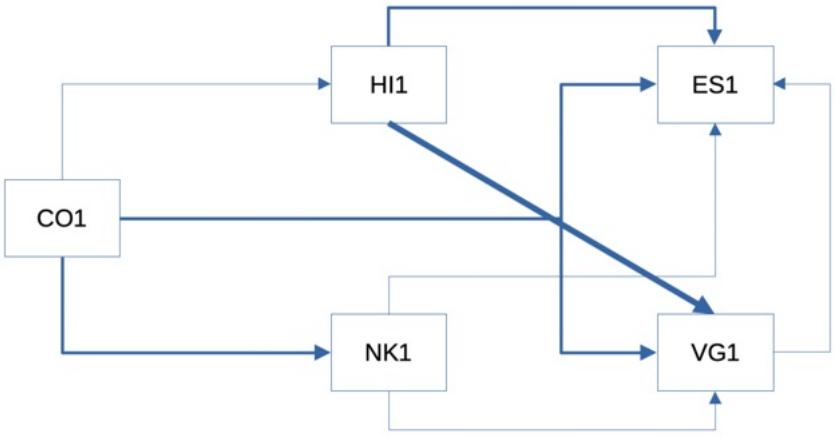
With the Hang Seng futures, HI1, as the dependent variable, information flow from CO1 is the only one significantly larger than 0 outside the March 2020 period. Within that timeframe, the flow from NK1 is also larger than zero and peaks at NTE ≈ 0.2. The flow from ES1 and VG1 remains consistent with zero.
When we evaluate information flow towards the NK1 future time series, we only observe information flow from the crude oil futures. Other information flows are consistent with zero. Interestingly, while there is a slight increase in the information flow starting March 2020, the increase and subsequent decrease are far less pronounced than for other variable pairs or in comparison to the unprocessed sample results.
Lastly, if we evaluate the information flow to the CO1 index, we find little indication towards this peak. The normalized information flow is with one exception NTE $= 0.1$ and mostly even smaller then 0.05. While the March 2020 period still exhibits the most nonzero information flows onto the CO1 independent variable, these are still very small and show less dominant peaks than the other variables. This result is surprising since the analysis of the unprocessed data did not indicate this structural difference in information flows.
We can draw the following conclusion from the above analyses: considerable information flows from the CO1 future to all other analyzed indices. NK1 only receives information flow from the oil futures. The Hang Seng future, HI1, is influenced by the NK1 future and the oil future. The VG1 index receives information from all indices except the ES1 future, and all other variables influence the ES1 future.

Non-Linear Bootstrap Measure Comparison Lastly, we evaluate the residual information flow when we bootstrap shuffling surrogates from the FT surrogatized data. The result is the amount of non-linear information flow that is not due to stochastic effects. It is displayed in figure (6.8). Strikingly we see that the non-linear information flow from the index futures towards the CO1 time series is consistent with zero throughout the whole analysis period (with one minor exception from NK1 with NTE $= 0.002$ in April 2020). CO1 is the only time series that exhibits non-linear information flow towards the NK1 time series. Interestingly, there is no apparent increase in information flow around the March 2020 period. While there are higher non-linear information transfer values, this effect is not pronounced and later than the increase we observed in the previous set-ups.
This temporal evolution is different from the HI1 future as the dependent variable. Here the non-linear influence is concentrated in the later period starting in March 2020. The information flow from the CO1 future is larger in magnitude than the one from the NK1 future. The former peaks at NTE $= 0.3$ and the latter below NTE $= 0.1$. There is no nonzero non-linear information flow from the other variables.
Both VG1 and ES1 receive non-linear information flows from the CO1 and HI1 futures of similar magnitude. This flow did only subtly increase in March 2020. There are only minor additional non-linear information flows from the NK1 to the VG1 future and from the NK1 and VG1 future to the ES1 future. These flows, however, happen in the March 2020 period.

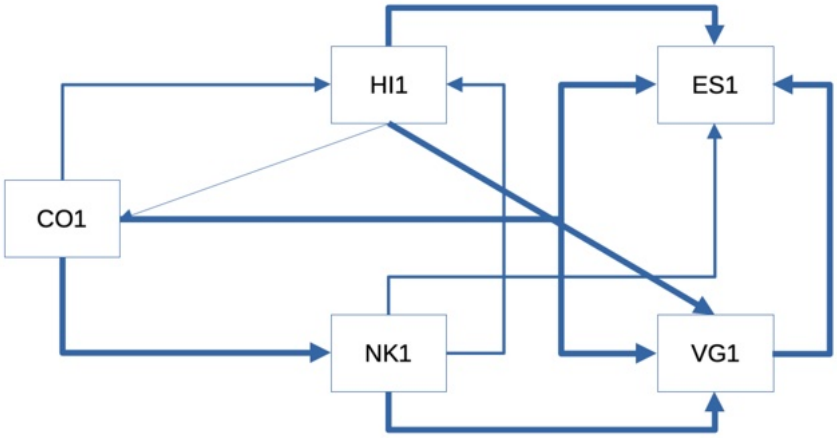
Summary of Information Flows Between Futures We can summarise these findings as follows. For the unprocessed data, we find information flows between all the time series. These information flows increase starting March 2020. However, CO1 is (with one exception) still the origin and not the receiver of information flows. Significantly higher information flow also occurs from the NK1 to the HI1 futures. Figure (6.9) displays these information flow changes .

Subsequent analysis with FT surrogates reveals significant amounts of these information flows,

Figure 6.8: This figure shows the residual information flow after processing the sample with Shuffle and FT surrogatization to infer the non-linear information flow between index futures that is significantly different from zero. We again find the foundational influence of the CO1 future.



(a) Information flows before March 2020



(b) Information flows in March 2020

Figure 6.9: This figure displays the information flows before (6.9a) and during (6.9b) the March 2020 period where the first lockdown measures were discussed and implemented. A thin line indicates an information flow of NTE ≈ 0.1, the medium thickness indicates a flow of NTE ≈ 0.2, and the thick lines indicate flow that breaks NTE $= 0.3$. When we compare the two plots, we see that in March 2020, more information flows along established connections, and new channels are formed. The whole network is more tightly knit.

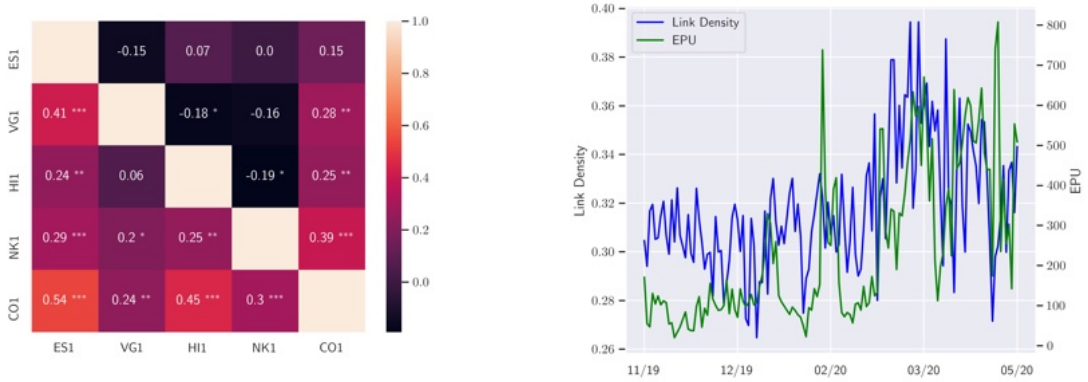
especially towards the ES1 and VG1 futures, to be non-linear. This composition is different for the HI1 future that only attains non-linear information flow from the oil future, which consistently receives no non-linear information contribution more significant than stochastic noise from any other of the analyzed futures time series. However, there is residual non-zero information flow if we consider the linear component. The non-linear information flow exhibits the March 2020 increase with a reduced amplitude or not at all compared to the total contribution.

6.3.1 Economic Policy Uncertainty Information Flows

We have speculated that the increase in information flow stems from the insecurity of market participants regarding the development under the changing dynamics caused by the pandemic. They might refer to the development of other indices to decide to buy or sell rather than sticking to their previous methods.
To test this hypothesis, we will utilize the EPU index (green curve in figure 6.10b). Figure (6.10a) shows the linear correlation of the policy uncertainty with the NTE value. Each tile in the heatmap displays the value of ρ between the EPU index and the information transfer from the row to the column variable. The stars indicate the significance level ($\alpha = 0.05, 0.01, 0.001$).
We find that information flow from the CO1 index is positively associated with policy uncertainty. These information flows are also the strongest for each influenced variable respectively. All the outgoing information flows of the NK1 are also positively correlated with policy uncertainty. All these relations are significant with an α error equal to or less than 5%.
Contrary to these findings, the outbound information flows of the HI1 future toward the NK1 future are negatively correlated. The same is true for information flows from the VG1 to the HI1 index. Both of these negative correlations are significant on a 5% level. Their cause is unclear. Maybe closed borders could have had an effect. Nonetheless, this finding is surprising since the information flows in the qualitative and heuristic analysis above seemed similar to the other ones. The same is true for the information flow towards the CO1 future. Whereas we found these to be small and to some extent consistent with zero, the correlation of information flow with economic policy uncertainty for all others except the ES1 future is significant at the 1% level and ≥ 0.25.
To further explore these connections, we model the daily information flows between the variables as a directed graph and obtain the temporal evolution of the link density. The result is shown in figure (6.10b) together with the temporal evolution of the EPU index on the right y-axis. The two quantities are correlated with $\rho = 0.46$ at a confidence level of $p = 1.47 \cdot 10^{-8}$.
We can therefore conclude that the amount of information flow between variables increases with economic policy uncertainty. This flow indicates the loss of efficiency in the market [20] and may create arbitrage opportunities. Nonetheless, some links exhibit a significant negative correlation of information flow with policy uncertainty.

6.3.2 Online Discourse Sentiment Information Flows

We will now evaluate online discourse about the COVID-19 pandemic as a proxy for public sentiment and insecurities during the pandemic. The results are displayed in figure (6.11). It is important to note that given the difficulties in data acquisition, we only evaluate the timeframe of increased information flow in between the futures time series starting in March 2020.
The upper left plot (6.11a) displays the information flow from all the log-return time series of the futures towards the Reddit sentiment data in the upper panel and the opposite direction in the lower panel. The magnitude of the flows is higher than what we have previously found for the information flows in between time series. Interestingly, these information flows seem to be reasonably symmetric.
Interestingly, this symmetry is broken when we evaluate the flows against shuffled data with the formula (41). Figure (6.11b) shows the results. We find that while the information flow towards



(a) Correlation, ρ, of the TE between futures and the EPU.

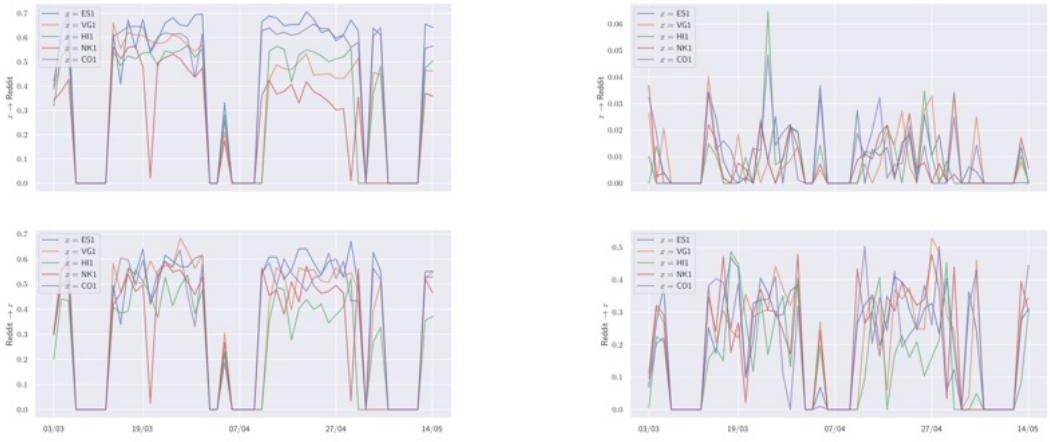
(b) Link Density of Network and EPU.

Figure 6.10: These plots show relation of the information flows with the EPU index. On the left (6.10a), we see the asymmetric TE values for flows between variables. The stars indicate the significance level, $\alpha = 0.05, 0.01, 0.001$. The right (6.10b) plot shows the evolution of the EPU index and the link density. We find a strong correlation with $\rho = 0.46$ at $p = 1.47 \cdot 10^{-8}$. The key takeaway of these figures is, that economical uncertainty is correlated with higher information flow between futures time series.

the Reddit sentiment gets low NTE ≤ 0.07, it peaks almost an order of magnitude higher for the information flow from the Reddit data. The information flows from and to the different indices seem similar when we only compare them within the respective directionality. We find a similar picture when we only evaluate non-linear data. The residual information flows from FT surrogate time series to the sentiment dataset peaks at about 0.2. Interestingly the non-linear flow from NK1 and VG1 is even lower than from the other time series, especially in the latter part of the analyzed period.
This difference in magnitudes is not the case when we evaluate it the other way around. The non-linear information flows towards the Reddit data peak at 0.5, and there seems to be no qualitative difference between the flows from different origin futures.
Lastly, we evaluate the non-linear flow against shuffled FT surrogatized data. These are the non-linear information flows likely not caused by statistical fluctuations. There we find non-linear information flow towards the sentiment data only of diminished size. Nonetheless, there are still significant non-linear flows from the Reddit data. These fluctuate wildly up to a peak of NTE $= 0.4$. We can therefore conclude that about 10% of the fluctuations of the index futures is linearly related to the sentiment data. On the other hand, ES1, VG1, and HI1 receive almost solely linear information flows.
Thus within the analyzed time frame, models exist that operationalize sentiment data to infer future prices of the index futures.

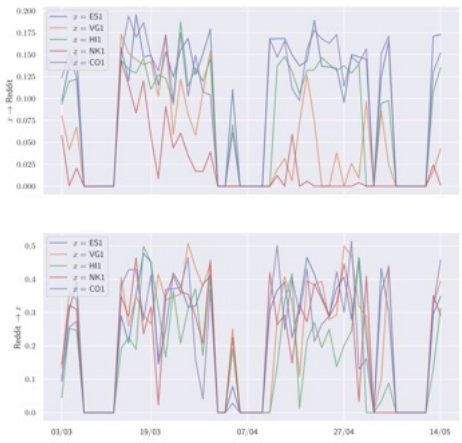
6.4 Comparison of COVID-19 Information Flow Variation to Other Crises

After analyzing the relation between the index futures and their relation to economic policy uncertainty and public sentiment, we will proceed to evaluate the dot-com bubble and 2008 financial crisis to compare information flows. Sadly, this analysis will be of limited informative value since we could only obtain index-level data in the accuracy necessary to compute NTE values. These, however, are no explicit instruments operationalizing future expectations about economic developments as futures do. Therefore, we also briefly cover the information flows between indices during the time of the COVID-19 pandemic. All these information flows are

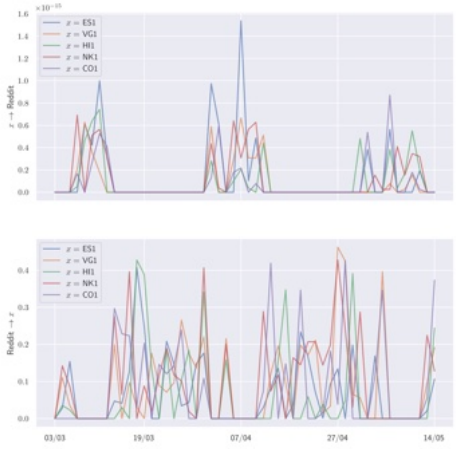


(a) Unprocessed data

(b) Shuffled data



(c) FT surrogates



(d) FT and shuffling surrogates

Figure 6.11: This figure shows the information flow from and to the sentiment in COVID-19 related online discussions. We find significant linear and non-linear contribution. These significant flows are far greater for the information flow from the discussions than for the reverse direction.

shown in figure (6.12).

First of all, we find that the levels of information flow between the indices exhibit far smaller amplitudes. They are about one order of magnitude smaller than the information flows between the futures. Additionally, in the March 2020 period (where we found an increase for the futures' transfer entropy) we detect a minimum for the information flow between the indices. Furthermore, the information transfer seems to saturate at certain thresholds as several stretches of almost constant TE are present. This saturation is not an artifact of leaky data but a feature of the underlying data.[11] These patterns are constant when held against the information flows between shuffled datasets.

We see similar behavior for the housing crisis. Here the information flow between indices drops mid-September 2008, coinciding with the Lehman bankruptcy. These developments persist when we utilize shuffled data.

We do not find such a point for the dot-com bubble. This finding is consistent with the fact that this crisis unfolded over a more extended period and did not have stand-out events associated with it.

We can summarize that we find similar patterns in the information flow between index time series in the Covid-19 and 2008 housing crises. However, it would be a leap of faith to infer that this indicates a similar pattern of the information flows between futures. In that sense, we sadly cannot utilize this index level data reliably for analysis and comparison of the information flows between futures in the different crisis periods. This inability is because the information dynamics between these financial instruments are too different.

6.5 Conclusion of the Application of Transfer Entropy to the COVID-19 Futures

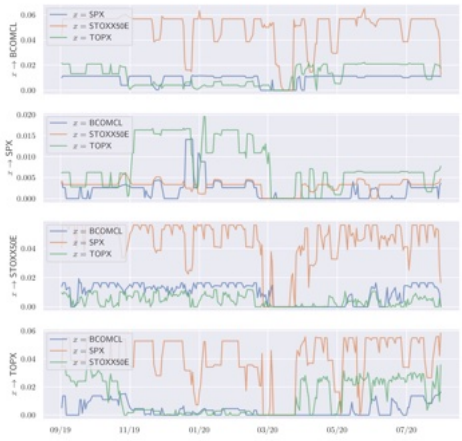
Let us now draw a conclusion about the information dynamics between futures in the onset of the COVID-19 crisis quantified using transfer entropy. The most striking feature of the data we found was the increase in information transfer starting in March 2020. That coincides with the public discussion and subsequent implementation of lockdown measures. Interestingly, there is a directionality with information flowing the CO1 oil future towards all other variables, which received much less information flow. Thus, we identified a directionality of information flows. This directionality is shown in figure (6.9).

We find that a similar dynamic is true when we evaluate only the non-linear dynamics, and when we compare these flows against stochastic information flows from shuffled datasets, this feature of the system is even more pronounced. This flow underlines the paramount influence oil price has on the expectation of economic (or at least price) developments. Additionally, the existence of these information flows are an indicator of the inefficiency of the market, which in turn enables the opportunity for arbitrage.

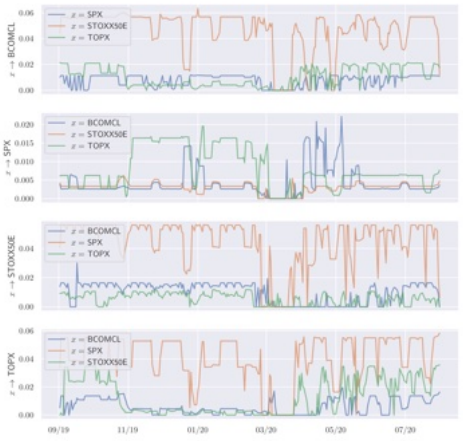
With the economic policy uncertainty analysis, we found that the magnitude of the information flows (for the most part) and thus the market inefficiency is linearly related to the economic policy uncertainty. This relation indicates that the times of crises coincide with times of market inefficiency. This inefficiency then singles out the uniqueness of the COVID-19 crises, which exhibited the highest levels of EPU (see figure 6.3). We did not find an indication for these relations on the index level data. These samples did not prove helpful in comparing different time frames.

Lastly, we also found significant information flow from online sentiment towards the future. Interestingly, this flow is unidirectional because more information flows from the sentiment to the futures prices than vice versa. This directionality indicates the usefulness of operationalizing sentiment in order to learn about future returns of future prices.

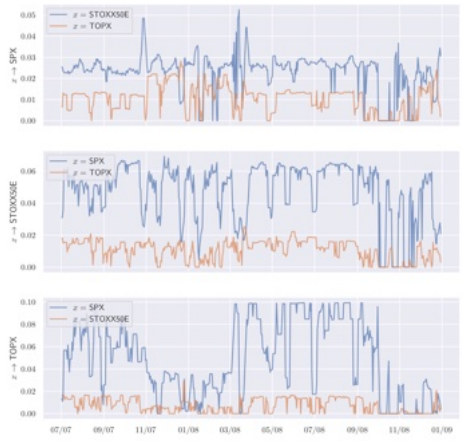
[11] Missing values in the data have been replaced and would result in vanishing TE.



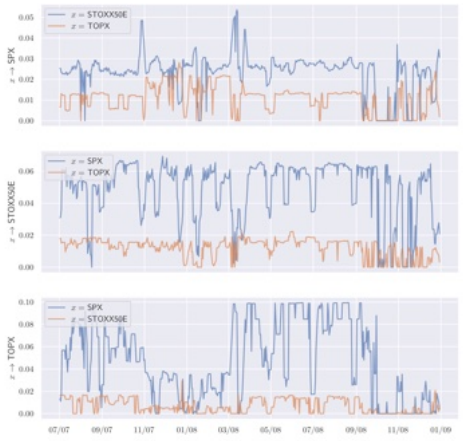
(a) Covid



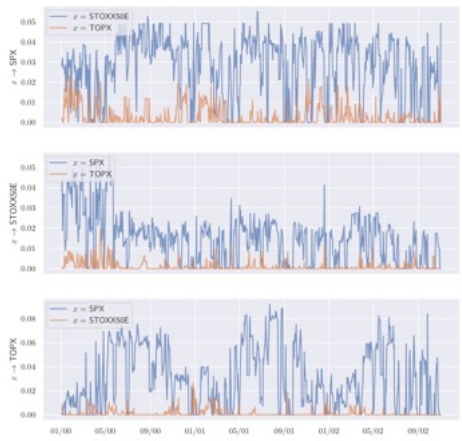
(b) Covid against shuffling surrogates.



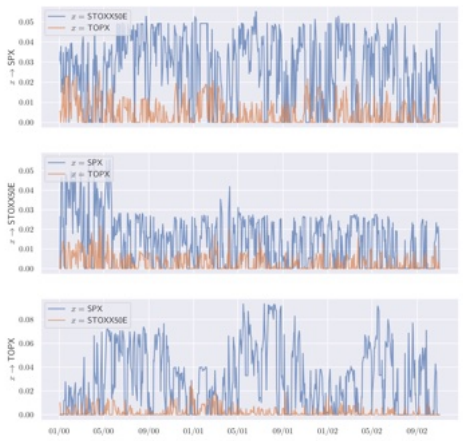
(c) Housing Crisis



(d) Housing Crisis against shuffling surrogates



(e) Dot-com



(f) Dot-com agains shuffling surrogates

Figure 6.12: Transfer entropy between stock indices in the three time periods.We find the structure of these flows to be such different from flows between index futures, that a comparison is can only yield intuition not insight.

6.6 Methodological Evaluation and Limitations

While the results of this analysis are undoubtedly interesting, our methodology suffers some drawbacks. The major limitation of the entropy maximizing approach of this thesis is that the uniformity of the finite sample desaturation remains a postulated assumption. Additionally, reasoning suggests that the increasing contribution from an increased resolution and the decreasing contribution from the finite sample effects are equal at the point of maximum entropy. This assumption, however, might also be false. Should there be systematic effects, our approach would yield systematic over-or underestimation of actual information flow.

Nonetheless, the increase of transfer entropy we identified for the March 2020 period remains consistent throughout different partition details. The last drawback of our approach is that we conducted an explorative analysis not driven by theory. Therefore, we were able to speculate about the causes of the difference in and directions of information flow, but we did provide or tried to verify an existing explanatory framework.

7 Conclusion

This thesis aimed to evaluate, develop, and apply transfer entropy as a non-parametric measure to detect directed association (or "causality") between continuous variables. Therefore, in chapter (2), we first developed different scale types and found that the main problem of TE for our application was in the mismatch of scale types. Based on these insights we discussed several binning methods, φ^D, in chapter (2.3) that mapped the data sampled from an unknown continuous distribution onto discrete values on which we can apply the information-theoretic measure. Some of these binning methods allow for bins of different sizes. We hypothesized that then the bin size carries information about the sample and developed and adapted discrete probability estimator in the chapter (4) that takes the bin volume in the $\mathbb{R}^n$ into account.
To compare the binning methods and evaluate this adapted estimator, we employed coupled map lattice systems in the chapter (5). This analysis is an extension of the studies conducted by Schreiber [60] who introduced the TE measure. We scanned the full range of the coupling strength, $\varepsilon \in [0,1]$ between the lattice rows and evaluated the Bellows, exponential, and logistic map additionally to the tent map considered by Schreiber.

We utilized these synthetic systems to quantify the dependence of the measure on various adaptations and sample characteristics. The inciting finding was the severe dependence of the measure on the partition details. This fact renders comparisons of the measure useless. Therefore, we evaluated two different but commonly used normalizations (see chapter 3), the normalization by $H(Y_t|Y_{t-L})$ the entropy of the influenced variable conditioned on its own past called H NTE and the $\log m$ NTE that is the normalization by the number of categories, m, of the influenced variable. The number $\log m$ is the maximum value of $H(Y_t|Y_{t-L})$. In section (5.3), we found that while the H NTE offers a more intuitive interpretation, it exhibits severe drawbacks when compared to the $\log m$ NTE measure. Most importantly, the normalization, $H(Y_t|Y_{t-L})$ is itself sample dependent which skews the variation of information flow with the coupling strength ε compared to the unnormalized measure. This skew is at times so severe that local minima become local maxima for the normalized quantity. Additionally, H NTE suffers a stronger washout of ε variation features than the $\log m$ NTE measure. Additionally, it also exhibits stronger skew for low ε values. These findings are of special importance given the widespread use of the H NTE normalization. Nonetheless, it offers an easier interpretation. Hence, if one can properly address its drawbacks with large enough samples and common signal shape and discretization, researchers might still opt for H NTE. We, however, decided to use the $\log m$ NTE measure for further analyses.

Our subsequent analysis regarded the sample size. Here we found that (N)TE needs large data sets. This need is due to the joint entropy term, which is cubic in the number of categories of the discretization. In our applications, we found that the error of the measure was smaller than 5% for more than 100 data points.
Additionally, we evaluated the dependence of the measure on the discretization method in the chapter (5.6). We found that no method consistently outperformed the other ones. Subsequently, we evaluated the adapted measure with a probability estimator that incorporates bin size. It is not easy to quantify performance when the underlying form of the probability is not known. Additionally, we face the problem that we need a discretized probability mass, not an estimator for the probability density as the later one integrates, but not sums to unity. We decided to quantify performance by consistency, and therefore scored by minimal standard deviation, σ. The new estimator showed improved performance for some, and decreased performance for other binning methods. For equal-sized bins, the adapted estimator recovered the common one.

Therefore, we could not justify the application of this adapted estimator, but introduced

another criterion to enable robustness for the (N)TE measure in section (5.8). Namely, we utilized the counteracting effects of increasing information transfer with increasing resolution. When we resolve more information, more flow can be detected. Contrary to this, finite sample desaturation in the measure computation results in decreasing information transfer for increasing resolution. Therefore, the measure must exhibit a local maximum. We chose this value as the final (N)TE value with which we then determine the "proper" binning. This approach only works for equal-sized bins as other methods converge to maximum information transfer. The same is true for the H NTE normalization. This measure shows the same desaturation effects such that H NTE converges to one and does not exhibit a finite NTE maximum that carries information.
The caveat of this method is that it rests on the unvalidated assumption that the bias introduced by the small sample effect is the same for all samples. It is thus still a function of sample size. This effect is also present for the unnormalized measure and is controllable by keeping the compared sample sizes constant.

In section (5.9), we then evaluated the effects of rank-ordered remapping onto a gaussian. With the method, one can transform a sample with an unknown distribution towards a target distribution. However, the method severely altered the information flows between variables and we will, therefore, not apply it. Intuitively, however, simple rescaling of the data does not affect transfer entropy if applied uniformly.
We also analyzed the effects of noise on the information flows in section (5.11). We found that for decreasing signal-to-noise ratio, the information flow becomes more and more a function of the partition detail. Additionally, errors increase.
Subsequently, in section (5.12), we evaluated data surrogatization with Fourier transforms, by shuffling, and their combination. We validated that we can utilize the former to test for flows of information by non-linear relations and the latter for flows that are significantly higher than noise-induced information flow. The methods are combined to attain a value for non-noise induced information flows via non-linear relations.

When comparing the TE to linear association measures, we find that it exhibits a stronger dependence on the sample characteristics. Nonetheless, it provides valuable insights and does not assume a functional form of the relation.
We concluded chapter (5) with a TE calculus recipe we then apply to COVID-19 data. The analysis of the futures during the first wave of the COVID-19 pandemic in the chapter (6) was exploratory and not driven by theory. This approach also represents limitations to the result, which we cannot contextualize within the framework of a theory to interpret the results. Nonetheless, our investigation was motivated by the assumption that the information flows during and before the pandemic are different. We validated this assumption with findings from the data. Information flow increases during the March 2020 period where the first lockdown measures were discussed and implemented. With shuffling and Fourier transform surrogates, we also identified a hierarchical structure of information flow. The origin of our analyzed time series was the crude oil future, CO1. Downstream follow the NK1, HI1, VG1, and ES1 index futures. We also found that the linear and non-linear associations between variables differ. For example, the information flow from VG1 to ES1 mainly utilizes linear channels.
We further found that economic policy uncertainty positively correlates with higher information flow in between the variables. This correlation indicates that in times of crisis, the market loses efficiency, and arbitrage opportunities occur.

Additionally, we found that the online sentiment in the Reddit forum regarding the COVID-19 pandemic yields information to all futures. There is only negligible flow in the reverse direction. This finding indicates that there is some relation between online sentiment and index future returns.

Furthermore, we analyzed the information flow between indices during the COVID period and the US housing and dotcom crises. However, the information dynamics between indices and index futures exhibit a different structure that we can infer no further insight.

This thesis shows two main results; one methodological and one regarding information flows between index futures. Regarding methodology, transfer entropy is a truly non-parametric measure that enables the analysis of relations of time series without specifying their functional form. This property is beneficial in exploratory analysis. We can evaluate the kind these relations by employing surrogatization methods. Nonetheless, the measure is less robust and more prone to misestimation due to finite data effects, sample characteristics, and - if applicable - the discretization of data. We developed an information transfer maximizing approach that allows for comparability of the information flows with similar sample sizes and normalizations. However, this method lays on the assumption of the uniformity of the finite sample-induced bias.
At this point, it remains to be evaluated by individual researchers whether further analysis and robustification of the model are worthwhile, or if the measure is only applied for exploratory purposes and does not require overblown rigor.
An essential point within the methodological evaluation is, that one must be careful when comparing transfer entropy values, as many hyperparameters, for example sample size, discretization method, resolution, normalization, etc., need to be carefully chosen.

Regarding the analyzed futures time series, we found a hierarchical structure of information flow with first the crude oil future that influences all other indices, second the Asian indices that influence the western ones, and third, downstream of the information flows the western VG1 and ES1 index futures. A schema of these relations is presented in figure (6.9). The amount of information flow increases during the time where the first COVID-19 counter-measures are discussed and applied. This finding suggests incomplete information and thus market inefficiencies and arbitrage opportunities. Even so, these findings need to be compared to other crises to establish whether we evaluated a peculiarity or a systemic relation.

To summarize: this thesis makes two scientific contributions. First, we thoroughly evaluated the transfer entropy measure, identified its strengths and caveats concerning sample size and partitions. To control those, we introduced entropy maximizing binning at fixed sample sizes to ensure comparability. Additionally, we evaluated different normalizations of the measure and recommended using the normalization by the logarithm of the m discrete categories of a variable.
Secondly, we found an increase in information flow between futures time series in March 2020 when the pandemic response measures began. Generally, we found increased information flow during times of economic uncertainty. This finding indicates that markets become increasingly inefficient with rising uncertainty. This inefficiency generates arbitrage opportunities.

8 Appendix

8.1 Mathematical Discussion of Scale Qualities

Nominal Scale If $S = (\{a,b,\dots\},\sim)$ is a set of elements with the equivalence relation $\sim$, the sample scale is nominal. The equivalence relation fulfils

1. reflexivity, $\forall s \in S : s \sim s$,
2. symmetry $s \sim t \Rightarrow t \sim s$, and
3. transitivity $\forall s,t,u \in S : s \sim t \wedge t \sim u \Rightarrow s \sim u$.

Ordinal Scale If $S = (\{a,b,\dots\},\preceq)$ is a totally ordered set with the order relation $\preceq$, S fulfils [18]

1. reflexivity, $\forall s \in S : s \preceq s$,
2. antisymmetry, $\forall s,t \in S : s \preceq t \wedge t \preceq s \Rightarrow s = t$,
3. transitivity $\forall s,t,u \in S : s \preceq t \vee t \preceq u \Rightarrow s \preceq u$, and
4. comparability $\forall s,t \in S : s \preceq t \vee t \preceq s$.

Interval Scale This is different for the interval scale with $S = (\{a,b,\dots\} \times \{a,b,\dots\},\preceq)$. Here, the cartesian product of the two constituent sets of S is the distance between the values. The $st \in \{a,b,\dots\} \times \{a,b,\dots\}$ fulfil the axioms of the total ordered set. Additionally,

1. $\forall s,t,u,v \in S : st \preceq uv \Rightarrow vu \preceq ts$,
2. $\forall s,t,u,v,a,b \in S : st \preceq uv \wedge ta \preceq vb \Rightarrow st \preceq ub$,
3. $\forall s,t,u.v \in S \preceq uv \preceq tt : \exists v_s, v_t \in S : uv \sim v_s s \sim v_t t$, and
4. $\forall s,t \in S, t \preceq s : \exists n \in \mathbb{N}, ns \preceq t$, the archimedian axiom.

Ratio Scale For the ratio scale $S = (\{a,b,\dots\},\circ,\preceq)$ with the axioms of the ordered set for $\preceq$ as well as

1. $\forall s,t,u \in S, s \circ (t \circ u) \sim (s \circ t) \circ u$, associativity,
2. $\forall s,t,u \in S : s \preceq t \Rightarrow s \circ u \preceq t \circ u \wedge u \circ s \preceq u \circ t$, monotony, and
3. the archimedian axiom (see above).

8.2 Supplementary Proofs

8.2.1 If f' is Bounded, the PDF f is Bounded

Theorem If $f : \mathbb{R} \mapsto \mathbb{R}_+$ is a PDF, its derivative exists and $f'(x) \leq L$ almost everywhere, then $f(x)$ is bounded.

Proof We will prove this by contraposition. Therefore we assume without loss of generality the rightsided limit $\lim_{x \downarrow k} f(x) = \infty$. Since f is a PDF its lower bound is zero. We know that f' does not exist in k since $\lim_{x \to k} {}^{(f(x)-f(k))}/_{(x-k)}$ diverges. However, by assumption f' exists almost everywhere. Thus it exists on $\mathscr{S} = [k-\varepsilon, k)$ for some ε with $0 < \varepsilon < \infty$.
Now let L be the upper bound of $f'(x)$ on $\mathscr{S}$. Therefore with the fundamental theorem of calculus

$$\lim_{x \downarrow k} f(x) - f(k-\varepsilon) = \lim_{x \downarrow k} \int_{k-\varepsilon}^{x} f'(u)du \leq \lim_{x \downarrow k} L \int_{k-\varepsilon}^{x} du = \lim_{x \downarrow k} L \cdot (x - (k-\varepsilon)) = L\varepsilon. \tag{110}$$

We can rearrange this inequality to

$$L \cdot \varepsilon + f(k-\varepsilon) \geq \lim_{x \downarrow k} f(x) = \infty. \tag{111}$$

Now $f(k-\varepsilon)$ is a probability density and thus bounded by one. ε is finite per definition. Therefore L must be not finite. This proves the theorem.

The proof is similar to the proof provided by [34].

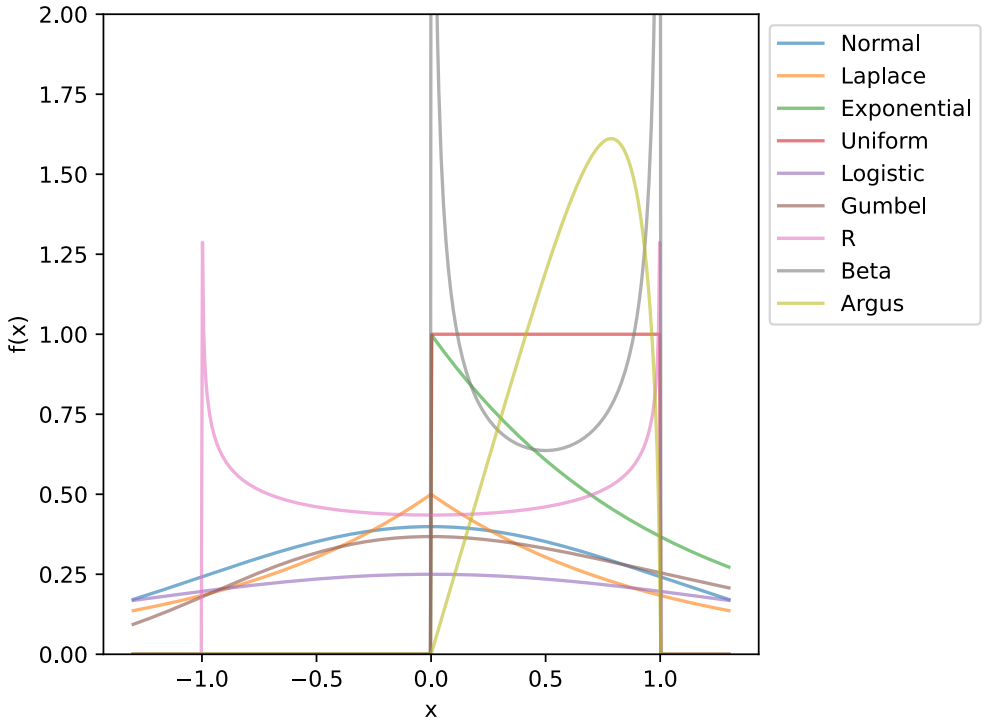


Figure 8.1: This figure displays the PDFs with which we evaluated entropy consistency. We selected them to form an ensemble of differently shaped distributions.

8.3 Adapted Probability Estimator Tested Against Analytic Distributions

The first evaluation of the new probability estimator (section 4) for the transfer entropy measure was conducted in a different framework. Instead of evaluating the transfer entropy consistency, we evaluated the discretizing probability estimator with the entropy. We scored its standard deviation for samples of different sizes against the entropies with the standard probability estimator - both normalized by their respective mean entropy to account for differences in magnitude. The following functions were used:

- Normal Distribution: $f(x) = \exp\left(x^2/2\right)/\sqrt{2\pi}$,
- Laplace Distribution $f(x) = \exp(-|x|)/2$,
- Exponential Distribution $f(x) = \exp(-x)$,
- Uniform Distribution $f(x) = \mathbb{1}_{x\in[0,1]}$,
- Logistic Distribution $f(x) = \exp(-x)/(1+\exp(-x))^2$,
- Gumbel Distribution $f(x) = \exp(-\exp(-x))$,
- R Distribution $f(x) = (1-x^2)^{c/2-1}/B(1/2,c/2)$ with $c = 1.6$ and B the beta distribution,
- Beta Distribution $f(x) = \Gamma(1)x^{-1/2}(1-x)^{-1/2}/\Gamma(1/2)^2$ with Γ the gamma function,
- and the Argus distribution $f(x) = 1/\sqrt{2\pi}\Psi(1)\cdot x\sqrt{1-x^2}\exp\left((1-x^2)/2\right)$, where $\Psi(\chi) = \Phi(\chi) - \chi\phi(\chi) - 1/2$ with Φ,ϕ being the cumulative density function and PDF of the normal distribution.

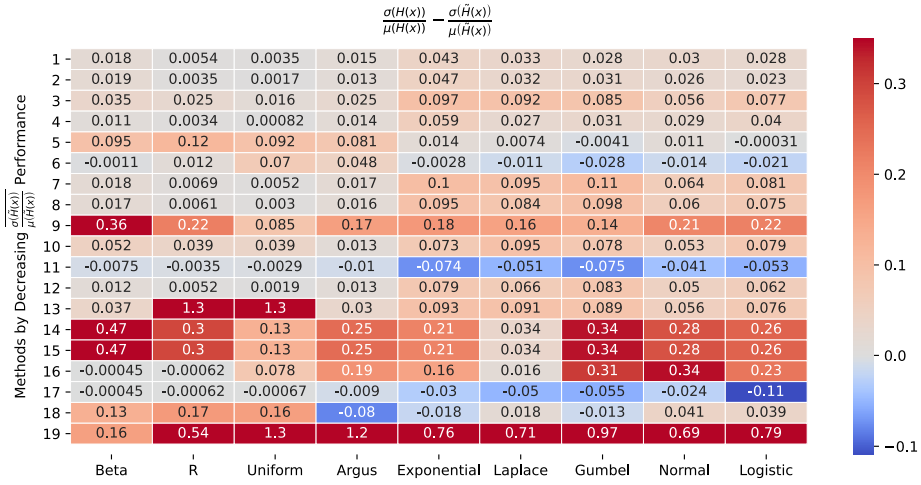


Figure 8.2: This heatmap shows the differences of the relative (to the mean) fluctuations of the entropies calculated with the common and adapted estimator. The numbers refer to the order presented in this chapter (8.3).

The PDFs of these functions are displayed in figure (8.1). We calculated the deviation as the mean of the PDF-PMF deviation at 1000 equispaced probing points over (part of) the value range of the PDF. The results are displayed in figure (8.1). The numbering is according to the performance of the method in the evaluation set up:

1. K-Means
2. Agglomerative Hierarchical Clustering
3. Minimizing Cross Validation
4. Expectation Maximisation with Gaussian Mixture Model
5. Freedman-Diaconis Rule
6. Scotts Rule
7. Square Root N Bins
8. Rice Rule
9. Knuths Method
10. Doanes Bins
11. Agostinos Uniform Probability
12. Sturges Bins
13. Shimazakis Bins
14. Akaike Information Criterion 15. Small Sample Akaike Information Criterion
15. Bayesian Information Criterion
16. Maximising Cross-Validation Likelihood
17. Mean Shift

18. DBSCAN

According to this evaluation setup, we see that the adapted estimator, for the most part, performs better in terms of smaller variance normalized by the mean with all methods. This performance increase for univariate entropies did not translate to the multivariate entropies employed for transfer entropy calculation in the setup above in the thesis body.
This setup, however, yields more insight into the measure calculation within dynamic systems. Unfortunately, the sampling from known distributions here does not allow for the modeling of causal relations.

Bibliography

References

[1] *Data*, pages 149–150. Springer New York, New York, NY, 2008. ISBN 978-0-387-32833-1. doi: 10.1007/978-0-387-32833-1_96. URL https://doi.org/10.1007/978-0-387-32833-1_96.

[2] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[3] Scott R Baker, Nicholas Bloom, and Steven J Davis. Measuring economic policy uncertainty. *The quarterly journal of economics*, 131(4):1593–1636, 2016.

[4] Albert-László Barabási. Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1987):20120375, 2013.

[5] Sridhar V Basavaraju, Monica E Patton, Kacie Grimm, Mohammed Ata Ur Rasheed, Sandra Lester, Lisa Mills, Megan Stumpf, Brandi Freeman, Azaibi Tamin, Jennifer Harcourt, et al. Serologic testing of us blood donations to identify severe acute respiratory syndrome coronavirus 2 (sars-cov-2)–reactive antibodies: December 2019–january 2020. *Clinical Infectious Diseases*, 72(12):1004–1009, 2021.

[6] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839, 2020.

[7] James Berger and Jose Bernardo. Ordered group reference priors with application to the multinomial problem. *Biometrika*, 79:25–37, 03 1992. doi: 10.1093/biomet/79.1.25.

[8] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[9] Jürgen Bortz and Christof Schuster. *Statistik für Human-und Sozialwissenschaftler: Limitierte Sonderausgabe*. Springer-Verlag, 2011.

[10] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[11] Peter A Burrough, Rachael McDonnell, Rachael A McDonnell, and Christopher D Lloyd. *Principles of geographical information systems*. Oxford university press, 2015.

[12] Jennifer Calfas, Margherita Stancati, and CW Yap. California orders lockdown for state's 40 million residents. *Wall Street Journal*, 2020.

[13] Colin Cameron. Excel 2007: Histogram, 2009. URL http://cameron.econ.ucdavis.edu/excel/ex11histogram.html. Accessed: February 10, 2022.

[14] Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17(8):790–799, 1995. URL http://dblp.uni-trier.de/db/journals/pami/pami17.html#Cheng95.

[15] Heeyoul Choi. Localization and regularization of normalized transfer entropy. *Neurocomputing*, 139:408–414, 2014. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2014.02.016. URL https://www.sciencedirect.com/science/article/pii/S0925231214004020.

[16] Y-S Chow, S Geman, L-D Wu, et al. Consistent cross-validated density estimation. *The Annals of Statistics*, 11(1):25–38, 1983.

[17] Ralph B. D'Agostino and Michael A. Stephens. *Goodness-of-Fit Techniques*. Marcel Dekker, 1986. ISBN 0824787056,9780824787059.

[18] Brian A Davey and Hilary A Priestley. *Introduction to lattices and order*. Cambridge university press, 2002.

[19] William HE Day and Herbert Edelsbrunner. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification*, 1(1):7–24, 1984.

[20] Marcos Lopez De Prado. *Advances in financial machine learning*. John Wiley & Sons, 2018.

[21] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[22] David P. Doane. Aesthetic frequency classification. *American Statistician*, 30:181–183, 1976.

[23] Ping Duan, Fan Yang, Tongwen Chen, and Sirish L. Shah. Direct causality detection via the transfer entropy approach. *IEEE Transactions on Control Systems Technology*, 21(6): 2052–2066, 2013. doi: 10.1109/TCST.2012.2233476.

[24] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. of 2nd International Conference on Knowledge Discovery and*, pages 226–231, 1996.

[25] Hua Fang. *Data Analytics for Longitudinal Biomedical Data*, pages 257–260. Springer International Publishing, Cham, 2020. ISBN 978-3-319-78262-1. doi: 10.1007/978-3-319-78262-1_153. URL https://doi.org/10.1007/978-3-319-78262-1_153.

[26] France24. Germany outlines plan for scaling back coronavirus lockdown. https://www.france24.com/en/20200406-germany-outlines-plan-for-scaling-back-coronavirus-lockdown, April 2020.

[27] David Freedman and Persi Diaconis. On the histogram as a density estimator: L2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 57(4):453–476, 1981. doi: 10.1007/BF01025868. URL https://doi.org/10.1007/BF01025868.

[28] Keinosuke Fukunaga and Larry Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on information theory*, 21 (1):32–40, 1975.

[29] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

[30] Deniz Gencaga, Kevin H Knuth, and William B Rossow. A recipe for the estimation of information flow in a dynamical system. *Entropy*, 17(1):438–470, 2015.

[31] George C. Tiao George E. P. Box. *Bayesian Inference in Statistical Analysis (Wiley Classics Library)*. Wiley-Interscience, 1973. ISBN 0201006227,9780201006223.

[32] Charles R. Harris et al. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

[33] Valentin Hofmann, Janet B Pierrehumbert, and Hinrich Schütze. Modeling ideological agenda setting and framing in polarized online groups with graph neural networks and structured sparsity. *arXiv preprint arXiv:2104.08829*, 2021.

[34] Ben (https://stats.stackexchange.com/users/173082/ben). If a probability density function (pdf) has bounded derivative, is the pdf itself bounded? Cross Validated. URL `https://stats.stackexchange.com/q/400527`. URL:https://stats.stackexchange.com/q/400527 (version: 2019-04-04).

[35] Chaolin Huang, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, et al. Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *The lancet*, 395(10223):497–506, 2020.

[36] Rob Hyndman. The problem with Sturges' rule for constructing histograms. Technical Report: `https://robjhyndman.com/papers/sturges.pdf`, 1995. Accessed: February 10, 2022.

[37] Edwin Thompson Jaynes. Information theory and statistical mechanics in statistical physics. *Brandies Lectures in Math Physics 3, 181*, 216, 1963.

[38] Harold Jeffreys. *Theory of probability, 3rd Edition*. Oxford Classic Texts in the Physical Sciences. Clarendon Press, 3 edition, 2003. ISBN 0198503687,9780198503682.

[39] Kevin H Knuth. Optimal data-based binning for histograms and histogram-based probability density models. *Digital Signal Processing*, 95:102581, 2019.

[40] Ryogo Kubo, Morikazu Toda, and Natsuki Hashitsume. *Statistical physics II: nonequilibrium statistical mechanics*, volume 31. Springer Science & Business Media, 2012.

[41] David M. Lane. Online statistics education: A multimedia course of study. Project Leader: David M. Lane, Rice University. `http://onlinestatbook.com`, 1997. Accessed: February 10, 2022.

[42] Xingguang Li, Junjie Zai, Qiang Zhao, Qing Nie, Yi Li, Brian T Foley, and Antoine Chaillon. Evolutionary history, potential intermediate animal host, and cross-species analyses of sars-cov-2. *Journal of medical virology*, 92(6):602–611, 2020.

[43] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[44] Haochun Ma. Causality in time series systems. Master's thesis, Ludwig-Maximilians-Universität, 2020.

[45] David JC MacKay and David JC Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[46] Nahema Marchal. The polarizing potential of intergroup affect in online political discussions: Evidence from reddit r/politics. *Available at SSRN*, 2020.

[47] Massimo Materassi, Giuseppe Consolini, Nathan Smith, and Rossana De Marco. Information theory analysis of cascading process in a synthetic model of fluid turbulence. *Entropy*, 16 (3):1272–1286, 2014.

[48] Hai Nguyen, Nhi Yen, Huong Hoang Luong, Nga Hong, and Hiep Xuan. Binning approach based on classical clustering for type 2 diabetes diagnosis. *International Journal of Advanced Computer Science and Applications*, 11, 01 2020. doi: 10.14569/IJACSA.2020.0110379.

[49] World Health Organization. Covid-19 - japan - (ex-china). https://www.who.int/emergencies/disease-outbreak-news/item/2020-DON236, January 2020.

[50] Mohammed Ouali, Rabii Gharbaoui, and Elmehdi Aitnouri. Benchmarking taxonomy for 1d clustering algorithms. *7th International Workshop on Systems, Signal Processing and their Applications, WoSSPA 2011*, 05 2011. doi: 10.1109/WOSSPA.2011.5931437.

[51] Angeliki Papana, Catherine Kyrtsou, Dimitris Kugiumtzis, and Cees Diks. Simulation study of direct causality measures in multivariate time series. *Entropy*, 15(7):2635–2661, 2013.

[52] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct): 2825–2830, 2011.

[53] South China Morning Post. China coronavirus: death toll almost doubles in one day as hong kong reports its first two cases. https://www.scmp.com/news/hong-kong/health-environment/article/3047193/china-coronavirus-first-case-confirmed-hong-kong, January 2020.

[54] Candice Chau (Hong Kong Free Press). Hong kong plans partial lockdowns for covid-19 hotspots and more tests, as number of new infections surges. https://hongkongfp.com/2020/12/08/hong-kong-plans-partial-lockdowns-for-covid-19-hotpots-and-more-tests-as-number-of-new-infections-surges/, December 2020.

[55] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL `http://www.R-project.org/`.

[56] Howard Raiffa and Robert Schlaifer. *Applied Statistical Decision Theory*. Wiley Classics Library. Wiley, reprint edition, 2000. ISBN 9780471383499, 047138349X.

[57] S Rajasekar, Alexandre Wagemakers, and MAF Sanjuan. Vibrational resonance in biological nonlinear maps. *Communications in Nonlinear Science and Numerical Simulation*, 17(8): 3435–3445, 2012.

[58] Christoph Räth, M Gliozzi, IE Papadakis, and W Brinkmann. Revisiting algorithms for generating surrogate time series. *Physical review letters*, 109(14):144101, 2012.

[59] Reuters. Tokyo governor urges people to stay indoors over weekend as virus cases spike. https://www.japantimes.co.jp/news/2020/03/25/national/science-health/tokyo-logs-40-coronavirus-cases/, March 2020.

[60] Thomas Schreiber. Measuring information transfer. *Phys. Rev. Lett.*, 85:461–464, Jul 2000. doi: 10.1103/PhysRevLett.85.461. URL `https://link.aps.org/doi/10.1103/PhysRevLett.85.461`.

[61] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978. ISSN 00905364. URL `http://www.jstor.org/stable/2958889`.

[62] David Scott and Stephan Sain. Multi-dimensional density estimation. *Data Mining and Data Visualization*, 24, 01 2005.

[63] David W. Scott. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, 12 1979. ISSN 0006-3444. doi: 10.1093/biomet/66.3.605. URL `https://doi.org/10.1093/biomet/66.3.605`.

[64] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

[65] Hideaki Shimazaki and Shigeru Shinomoto. A method for selecting the bin size of a time histogram. *Neural computation*, 19:1503–27, 07 2007. doi: 10.1162/neco.2007.19.6.1503.

[66] Spiegel. Bayerische Behörden bestätigen ersten Fall in Deutschland, January 2020. URL `https://www.spiegel.de/wissenschaft/medizin/corona-virus-erster-fall-in-deutschland-bestaetigt-a-19843b8d-8694-451f-baf7-0189d3356f99`.

[67] Stanley Smith Stevens. On the theory of scales of measurement. *Science*, 103(2684):677–680, June 1946.

[68] Charles J Stone. An asymptotically optimal histogram selection rule. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer*, volume 2, pages 513–520. Wadsworth, 1984.

[69] Ruslan L. Stratonovich. *Theory of Information and its Value*. Springer Nature Switzerland, 2020.

[70] Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617, 2002.

[71] Herbert A. Sturges. The choice of a class interval. *Journal of the American Statistical Association*, 21(153):65–66, 1926. doi: 10.1080/01621459.1926.10502161. URL `https://doi.org/10.1080/01621459.1926.10502161`.

[72] Charles C Taylor. Akaike's information criterion and the histogram. *Biometrika*, 74(3): 636–639, 1987.

[73] Haizhou Wang and Mingzhou Song. Ckmeans.1d.dp: Optimal k-means clustering in one dimension by dynamic programming. *The R Journal*, 3:29–33, 12 2011. doi: 10.32614/RJ-2011-015.

[74] Larry Wasserman. *All of Nonparametric Statistics*. Springer texts in statistics. Springer, 2006. ISBN 9780387251455,0387251456.

[75] Ernst Wit, Edwin van den Heuvel, and Jan-Willem Romeijn. 'all models are wrong...': an introduction to model uncertainty. *Statistica Neerlandica*, 66(3):217–236, 2012.

[76] Karl L. Wuensch. *Scales of Measurement*, pages 1283–1285. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-04898-2. doi: 10.1007/978-3-642-04898-2_67. URL `https://doi.org/10.1007/978-3-642-04898-2_67`.

Erklärung zur Selbstständigkeit

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbstständig angefertigt, alle Zitate als solche kenntlich gemacht sowie alle benutzten Quellen und Hilfsmittel angegeben habe.

Berkeley, CA, USA, 10.09.2021 . . . . . .