entropy

MDPI

*Article*

# InfoMat: Leveraging Information Theory to Visualize and Understand Sequential Data †

**Dor Tsur * and Haim Permuter**

School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Be'er Sheva 8410501, Israel; haimp@bgu.ac.il

* Correspondence: dortz@post.bgu.ac.il

† Part of this work was presented at the International Symposium on Information Theory (ISIT), Athens, Greece, 7–12 July 2024.

**Abstract:** Despite the widespread use of information measures in analyzing probabilistic systems, effective visualization tools for understanding complex dependencies in sequential data are scarce. In this work, we introduce the information matrix (InfoMat), a novel and intuitive matrix representation of information transfer in sequential systems. InfoMat provides a structured visual perspective on mutual information decompositions, enabling the discovery of new relationships between sequential information measures and enhancing interpretability in time series data analytics. We demonstrate how InfoMat captures key sequential information measures, such as directed information and transfer entropy. To facilitate its application in real-world datasets, we propose both an efficient Gaussian mutual information estimator and a neural InfoMat estimator based on masked autoregressive flows to model more complex dependencies. These estimators make InfoMat a valuable tool for uncovering hidden patterns in data analytics applications, encompassing neuroscience, finance, communication systems, and machine learning. We further illustrate the utility of InfoMat in visualizing information flow in real-world sequential physiological data analysis and in visualizing information flow in communication channels under various coding schemes. By mapping visual patterns in InfoMat to various modes of dependence structures, we provide a data-driven framework for analyzing causal relationships and temporal interactions. InfoMat thus serves as both a theoretical and empirical tool for data-driven decision making, bridging the gap between information theory and applied data analytics.

**Keywords:** data analysis; data visualization; directed information; information matrix; information conservation; mutual information; transfer entropy



Academic Editor: Kichun Lee

Received: 23 February 2025
Revised: 21 March 2025
Accepted: 24 March 2025
Published: 28 March 2025

**Citation:** Tsur, D.; Permuter, H. InfoMat: Leveraging Information Theory to Visualize and Understand Sequential Data. *Entropy* **2025**, *27*, 357. https://doi.org/10.3390/e27040357



## 1. Introduction

Information theory plays a key role in the analysis of dynamics in stochastic systems [1–4]. Through the lens of information theory, one can study the temporal evolution of dependence in a sequential system, which is often interpreted as the exchange of information between its interacting components. For example, in communication channels, directed information quantifies the information flow from the encoder to the channel [5,6]. Consequently, the feedback capacity of communication channels, which is characterized by the optimization of average directed information, can be interpreted as the maximum amount of information flow from the transmitter to the channel [7]. Another pertinent example is neuroscience [8], where information theory is widely used for the analysis

and processing of collected data. For example, transfer entropy [9] is used to infer functional connectivity and to analyze information flow from recorded data, such as EEG and fMRI [10]. Beyond communications and neuroscience, information theory was shown useful for various fields of sequential analysis, encompassing control [11–13], reinforcement learning [14,15], causal inference [16–18], and various machine learning tasks [19–21].

In contrast to its wide use to infer and quantify relations in time series data, information theory fails to offer simple visualization tools to demonstrate its merits, and existing visualization techniques are not applicable to such settings. A common example of dependence visualization is the cross-correlation matrix [22]. For a given random vector pair $(X, Y) \in \mathbb{R}^{d_x+d_y}$, the correlation matrix is an $\mathbb{R}_{d_x \times d_y}$-valued matrix, whose $(i, j)$th entry is the correlation between $(X_i, Y_j)$. While being a powerful tool for simple exploratory analysis, the performance of the correlation matrix heavily relies on the data domain and structure of the joint distribution. Specifically, it provides a complete characterization of dependence only when the vectors' entries are univariate Gaussian sequences with linear relations. Furthermore, the correlation matrix does not generalize to the conditional setting, which is crucial for visualization in time series analysis.

A common visualization of information theoretic relationships is the Venn diagram [4], which serves as a visualization of mutual information, through decomposition into joint and conditional entropies. Finally, the information diagram [23], which is based on the Taylor diagram [24], explores the interplay between entropies and mutual information in a geometric fashion by mapping dependence into an angle between the marginal entropy vectors. However, this method is limited to a random pair. Beyond the visualization of information measures, information theory is widely used to evaluate other visualization and rendering techniques [25]. In this work, we attempt to close this gap, and propose a new visualization method.

*Contributions*

In this work, we propose the information matrix (InfoMat), a novel visualization tool of the information transfer in dynamic stochastic systems. Given two stochastic sequences of length $m$, the InfoMat arranges the conditional mutual information terms that describe the evolution of dependence in an $m \times m$ matrix. By arranging the $m^2$ conditional information terms that describe the temporal processes' evolution in an $m \times m$ matrix, various dependence patterns emerge in the InfoMat though visual patterns. We show how the InfoMat captures popular sequential information measures such as the directed information and transfer entropy through linear operations. We then demonstrate the InfoMat utilities for both theoretical and practical methods.

For theoretical purposes, the InfoMat provides a visual representation of existing information theoretic conservation laws and decompositions, while revealing new relationships. The relationships are proven by characterizing different subsets of the matrix with corresponding information measures. Additionally, the InfoMat serves as a practical visualization tool for arbitrary sequential data. Using a heatmap representation, we can link various dependence structures in the data with visual patterns. We propose a Gaussian mutual information estimator of the InfoMat that relies on the calculation of sample covariance matrices and analyze its theoretical guarantees, while empirically demonstrating the power of the resulting visualization tool. When a Gaussian estimator is insufficient, we develop a neural InfoMat estimator, which is based on masked normalizing flows (MAFs), which expand the class of distributions captured by the InfoMat. We demonstrate the estimated InfoMat to visualize the power of optimal coding schemes in communication channels with memory and for the visualization of information flows in real-world datasets.

The rest of this paper is organized as follows. Section 2 presents required background and preliminaries, followed by Section 3, which presents the InfoMat and demonstrates its capabilities for the visualization of theoretical information-theoretic conservation laws. Then, Section 4 discusses the Gaussian estimator of the InfoMat, while Section 5 explores neural estimation. Finally, Section 6 demonstrates the utility of the estimated InfoMat for visualizing dependence structures in sequential data, and Section 7 provides concluding remarks and discusses future work.

## 2. Background and Preliminaries

### *2.1. Notation*

Sets are denoted by calligraphic letters, e.g., $\mathcal{X}$. When $\mathcal{X}$ is finite, we use $|\mathcal{X}|$ for its cardinality. For any $n \in \mathbb{N}$, $\mathcal{X}^n$ is the $n$-fold Cartesian product of $\mathcal{X}$, while $x^n = (x_1, \ldots, x_n)$ denotes an element of $\mathcal{X}^n$. For $i, j \in \mathbb{Z}$ with $i \leq j$, we use the shorthand $x_i^j \triangleq (x_i, \ldots, x_j)$; the subscript is omitted when $i = 1$. Expectations are denoted by $\mathbb{E}$. When $\mathcal{X}$ is countable, we use $p$ for the PMF associated with the probability measure $P$. Random variables are denoted by upper-case letters, e.g., $X$. The Kullback–Leibler (KL) divergence between $P$ and $Q$, with $P \ll Q$, is $\mathsf{D}_{\mathsf{KL}}(P\|Q) \triangleq \mathbb{E}_P\big[\log \frac{\mathrm{d}P}{\mathrm{d}Q}\big]$. The mutual information between $(X, Y) \sim P_{XY}$ is $I(X;Y) \triangleq \mathsf{D}_{\mathsf{KL}}(P_{XY}\|P_X \otimes P_Y)$, where $P_X$ and $P_Y$ are the marginals of $P_{XY}$. The entropy of a discrete random variable $X \sim P$ is $H(X) \triangleq -\mathbb{E}[\log p(X)]$.

### *2.2. Sequential Information Measures*

Consider a pair of jointly distributed sequences of length $m$, $(X^m, Y^m) \sim P_{X^m,Y^m}$. The causal nature of sequential communication systems impedes mutual information from properly describing sequential information flows, as it decomposes into non-causal conditional mutual information terms, i.e., [26]

$$I(X^m;Y^m) = \sum_{i=1}^{m} I(X_i;Y^m|X^{i-1}),$$

To this end, several causal adaptations of mutual information to time series data were developed in the literature. The first is directed information [5], which was originally developed to characterize information rates in communication channels with feedback. The directed information *from* $X^m$ *to* $Y^m$ is given by

$$I(X^m \to Y^m) \triangleq \sum_{i=1}^{m} I(X^i;Y_i|Y^{i-1}), \tag{1}$$

where the directed information in the opposite direction $I(Y^m \to X^m)$ is defined symmetrically. The second information measure, which gained popularity in neuroscience and physics is transfer entropy [9,27,28]. For parameters $(m, k, l)$, the transfer entropy is given by

$$T_m^{X\to Y}(k,l) \triangleq I(X_{m-k}^{m-1};Y_m|Y_{m-l}^{m-1}), \tag{2}$$

Transfer entropy and directed information follow various decompositions and information conservation laws. We will further discuss them through the lenses of the proposed InfoMat in the upcoming section.

In physics, time series measures—such as approximate entropy [29], sample entropy [30], permutation entropy [31], and dispersion entropy [32]—are widely used to quantify the local complexity and irregularity of sequential data. While these scalar metrics provide valuable insights into local dynamical properties, they are conceptually distinct from directed information and transfer entropy, as they are not based on quantification through conditional mutual information.

### *2.3. Information Decomposition and Conservation*

In stochastic systems with memory, the temporal evolution of the dependence between the system's elements can be viewed through the lenses of information exchange. This exchange of information can be captured within conversation laws, which quantify the total amount of information flow in a given system. Specifically, for a system with $m$ time steps and two sources $(X^m, Y^m) \sim P_{X^m,Y^m}$, Massey [33] proved the following law of information conservation

$$I(X^m; Y^m) = I(X^m \to Y^m) + I(\mathsf{D} \circ Y^m \to X^m) \tag{3}$$

where $(\mathsf{D}^k \circ X^m)$ is a left concatenation of $k$ 'dummy' deterministic symbols with $X_{k+1}^m$ and $(\mathsf{D} \circ X^m) = (\mathsf{D}^1 \circ X^m)$. The authors of [34] propose a modification of (3) that distinguishes between past and present effects, given by

$$I(X^m; Y^m) = I(\mathsf{D} \circ X^m \to Y^m) + I(\mathsf{D} \circ Y^m \to X^m) + I_{\mathsf{inst}}(X^m, Y^m), \tag{4}$$

where $I_{\mathsf{inst}}(X^m, Y^m) \triangleq \sum_{i=1}^n I(X_i; Y_i | X^{i-1}, Y^{i-1})$ is the *instantaneous* information, which measures the symmetric dependence between $X^m$ and $Y^m$, given a shared history. Information conservation has been shown to be useful for the analysis of causal information flows in neural spike trains and financial data [35,36].

## 3. Information Matrix

Fix $m \in \mathbb{N}$. We are interested in characterizing the interaction between two stochastic sequences (the sequences can be thought of as an $m$-step sample from some underlying joint stochastic process) of length $m$ in a visually meaningful manner. Denote the sequences with $(X^m, Y^m) \sim P_{X^m,Y^m}$. Our emphasis is on the information-theoretic description of this interaction, which can be seen as the evolution of dependence between the interacting components over time. This characterization can alternatively be viewed as a transfer of information. The entire dependence structure between $X^m$ and $Y^m$ is captured by the $m$-fold mutual information, given by the following chain rule [4]

$$I(X^m; Y^m) = \sum_{i=1}^{m} \sum_{j=1}^{m} I(X_i; Y_j | X^{i-1}, Y^{j-1}). \tag{5}$$

The above decomposition implies that, along $m$ steps, the interaction is characterized with $m^2$ conditional mutual information terms. This acts as our motivation for the following definition and this work.

We define the InfoMat as the following $m \times m$ matrix:

$$\mathrm{I}^{X,Y} \in \mathbb{R}_{\geq 0}^{m \times m}, \quad \mathrm{I}_{i,j}^{X,Y} \triangleq I(X_i; Y_j | X^{i-1}, Y^{j-1}). \tag{6}$$

The InfoMat captures all the information transfer within the two-user system $(X^m, Y^m)$ as (5) implies that

$$I(X^m; Y^m) = \sum_{i,j} \mathrm{I}_{i,j}^{X,Y} = \mathbf{1}_m \mathrm{I}^{X,Y} \mathbf{1}_m^\intercal,$$

where $\mathbf{1}_m \in \mathbb{R}^m$ is an $m$-length vector of ones and the inequality follows from the chain rule for mutual information [4].

As we further show, the InfoMat allows for the interpretations of the chain rule (5), and other information decomposition laws of $I(X^m; Y^m)$ as grouping the entries of $\mathrm{I}^{X,Y}$ into meaningful subsets that sum up to $I(X^m; Y^m)$. These decompositions are often termed

information conservation laws [33]. We will show that such laws, whose proofs are usually technical and algebraic, can be easily visualized by coloring subsets of the entries of $\mathrm{I}^{X,Y}$.

### *3.1. Visualizing Sequential Information Measures*

We begin by demonstrating that the aforementioned sequential information measures can be recovered through the summation of elements of the InfoMat. Directed information (1) is given by the sum of a triangular sub-matrix, i.e.,

$$I(X^m \to Y^m) = \mathbf{1}^\mathsf{T} \begin{pmatrix} \mathrm{I}^{X,Y}_{1,1} & \mathrm{I}^{X,Y}_{1,2} & \cdots & \mathrm{I}^{X,Y}_{1,m} \\ 0 & \mathrm{I}^{X,Y}_{2,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathrm{I}^{X,Y}_{m-1,m} \\ 0 & \cdots & 0 & \mathrm{I}^{X,Y}_{m,m} \end{pmatrix} \mathbf{1}. \tag{7}$$

Due to its direction sensitivity, directed information in each direction is associated with a certain triangular sub-matrix. Specifically, the direction $X^m \to Y^m$ corresponds to the upper triangular part of $\mathrm{I}^{X,Y}$, while the lower triangular part represents the direction $Y^m \to X^m$.

It is also useful to define the time-delayed directed information, which describes the causal flow of information under a time delay at the transmitting node. For a delay of $k < m$, the $k$-time-delayed directed information is given by $I(\mathsf{D}^k \circ X^m \to Y^m) \triangleq \sum_{i=1}^{m} I(X^{i-k}; Y_i | Y^{i-1})$. In $\mathrm{I}^{X,Y}$, a $k$-delay of directed information corresponds to a right shift of $k$ indices in the $X^m \to Y^m$ direction, and a down shift in the $Y^m \to X^m$ direction.

Next, the transfer entropy term $T^{X\to Y}_{i+1}(i,i) = I(X^i; Y_{i+1} | Y^i)$, which quantifies the causal effect of $X^i$ on $Y_{i+1}$ given $Y^i$ is given by the sum over a column of length $i$ in row $i+1$. For example, we have

$$T^{X\to Y}_3(2,2) = \mathbf{1}^\mathsf{T} \begin{pmatrix} 0 & \cdots & \mathrm{I}^{X,Y}_{1,3} & \cdots & 0 \\ \vdots & 0 & \mathrm{I}^{X,Y}_{2,3} & \ddots & \vdots \\ \vdots & \ddots & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 \end{pmatrix} \mathbf{1}. \tag{8}$$

Note that a transfer entropy does not include terms on the InfoMat diagonal. This stresses that transfer entropy focuses on strict past influence on present interaction, which is a key distinction from directed information [37]. The relation between mentioned information measures and the patterns in $\mathrm{I}^{X,Y}$ are summarized in Table 1.

**Table 1.** Visual shapes of dependence patterns in $\mathrm{I}^{X,Y}$.

| Information measure | Visual pattern in $\mathrm{I}^{X,Y}$ |
|---|---|
| $I(X^m \to Y^m)$ | Upper triangular **with diagonal** |
| $I(\mathsf{D}^k \circ X^m \to Y^m)$ | Upper triangular, side $(m-k)$ |
| $T^{X\to Y}_{i+1}(i,i)$ | Col. in row $i+1$ with length $i$ |

### *3.2. Capturing Information Conservation Laws*

Having identified the relation between information measures and their patterns in $\mathrm{I}^{X,Y}$, we can visualize the aforementioned laws of information conservation [33,34]. Such rules are often derived via algebraic manipulation of information measures and therefore may

lack intuition and may fail to provide a deeper understanding of the underlying interaction. Recall that Massey's information conservation law is given by (3)

$$I(X^m;Y^m) = I(X^m \to Y^m) + I(\mathsf{D}\circ Y^m \to X^m) \tag{9}$$

Following the identification of DI with triangular submatrices of $\mathrm{I}^{X,Y}$, (3) follows by coloring index subsets and summing over each color group.

$$I(X^m;Y^m) = \mathbf{1}^{\mathsf{T}} \begin{pmatrix} \mathrm{I}^{X,Y}_{1,1} & \mathrm{I}^{X,Y}_{1,2} & \cdots & \mathrm{I}^{X,Y}_{1,m} \\ \mathrm{I}^{X,Y}_{2,1} & \mathrm{I}^{X,Y}_{2,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathrm{I}^{X,Y}_{m-1,m} \\ \mathrm{I}^{X,Y}_{m,1} & \cdots & \mathrm{I}^{X,Y}_{m,m-1} & \mathrm{I}^{X,Y}_{m,m} \end{pmatrix} \mathbf{1}. \tag{10}$$

Considering the finer decomposition using instantaneous information (4), we note that $I_{\mathsf{inst}}(X^n,Y^n) = \mathsf{Trace}(\mathrm{I}^{X,Y})$. We can therefore identify this decomposition by excluding the diagonal from the upper sub-triangle.

$$I(X^m;Y^m) = \mathbf{1}^{\mathsf{T}} \begin{pmatrix} \mathrm{I}^{X,Y}_{1,1} & \mathrm{I}^{X,Y}_{1,2} & \cdots & \mathrm{I}^{X,Y}_{1,m} \\ \mathrm{I}^{X,Y}_{2,1} & \mathrm{I}^{X,Y}_{2,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathrm{I}^{X,Y}_{m-1,m} \\ \mathrm{I}^{X,Y}_{m,1} & \cdots & \mathrm{I}^{X,Y}_{m,m-1} & \mathrm{I}^{X,Y}_{m,m} \end{pmatrix} \mathbf{1}. \tag{11}$$

Next, we will leverage the constructed relations between InfoMat patterns and sequential information measures to derive new information-theoretic decompositions and formulas.

### *3.3. Developing New Information-Theoretic Relations*

Beyond the visualization of existing relationships, the simplicity of the InfoMat visualization allows us to develop new meaningful information-theoretic equivalences. We begin with the following proposition that relates the two information measures of interest.

**Proposition 1** (Transfer entropic decomposition of directed information). *For* $(X^m,Y^m) \sim P_{X^m,Y^m}$ *and* $1 \le k \le m$, *we have*

$$I(\mathsf{D}^k \circ X^m \to Y^m) = \sum_{i=1}^{m-k} T^{X\to Y}_{i+1}(i,i).$$

**Proof.** We provide the proof for $k=1$. Recall that directed information corresponds to the upper triangular part of $\mathrm{I}^{X,Y}$. We note that $T^{X\to Y}_{i+1}(i,i)$ corresponds to a column in $\mathrm{I}^{X,Y}$ that begins in the $(i+1)$th column and has length $i$. Thus, we color the triangular part (directed information) as follows

$$\begin{pmatrix} \mathrm{I}^{X,Y}_{1,2} & \mathrm{I}^{X,Y}_{1,3} & \cdots & \mathrm{I}^{X,Y}_{1,m} \\ & \mathrm{I}^{X,Y}_{2,3} & \ddots & \vdots \\ & & \ddots & \vdots \\ & & & \mathrm{I}^{X,Y}_{m-1,m} \end{pmatrix}$$

The relation then follows by summing over the upper triangular part and dividing the sum into the corresponding rows, i.e.,

$$I(\mathsf{D}\circ X^n \to Y^n) = T^{X\to Y}_2(1,1) + T^{X\to Y}_3(2,2) + \cdots + T^{X\to Y}_m(m-1,m-1)$$

The proof similarly extends to $k > 1$ with similar steps. □

The proof demonstrates the simplicity of InfoMat, as it boils down to identifying the aforementioned information measures as entry subsets within $\mathrm{I}^{X,Y}$. Equipped with Proposition 1, we can provide an information conservation rule in terms of transfer entropy terms.

**Proposition 2** (Conservation of transfer entropy)**.** *Let* $(X^m, Y^m) \sim P_{X^m,Y^m}$. *Then*

$$I(X^m;Y^m) = \sum_{i=1}^{m-1} T_{i+1}^{X\to Y}(i,i) + T_{i+1}^{Y\to X}(i,i) + I_{\mathsf{inst}}(X^m, Y^m).$$

Finally, we propose a recursive decomposition of directed information via transfer entropy.

**Proposition 3** (Directed information chain rule)**.** *Let* $(X^m, Y^m) \sim P_{X^m,Y^m}$. *Then*

$$I(\mathsf{D}^k \circ X^m \to Y^m) = I(\mathsf{D}^{k+1} \circ X^m \to Y^m) + \sum_{i=1}^{m-k} I(X_{i-k}; Y_i | Y^{i-1}).$$

The proofs of Propositions 2 and 3 follow by arguments and tools similar to Proposition 1 and are given in Appendix A.1 for completeness. The new derived information-theoretic conservation laws and decompositions elucidate the relationships between mutual information, directed information, and transfer entropy. Strengthening these connections is crucial, as each measure has distinct tools and applications, thereby enhancing their cross-utilization potential and potentially bridging various applications. While these proposed relations can indeed be derived through existing algebraic manipulations of mutual information, the technical complexity may make these deductions less apparent.

### *3.4. Relating Dependency Structures and Visual Patterns in the InfoMat*

We now show that, beyond the derivation of various information decomposition laws, the InfoMat can be used to identify temporal patterns in the data. To this end, we relate various dependence structures to corresponding visual patterns in the InfoMat. We later demonstrate those relations on data via various estimates of the InfoMat (Section 6). To this end, we assume in this section that the sequence $(X^m, Y^m)$ is an $m$-fold projection of some jointly stationary stochastic process defined over $\mathcal{X} \times \mathcal{Y}$. As a first example, we note that when the joint process is independent and identically distributed (i.i.d.), every conditional mutual information with $i \neq j$ vanishes. The corresponding InfoMat takes the form $\mathrm{I}^{X,Y} = I(X;Y)\mathrm{I}_m$ with $\mathrm{I}_m$ being the $m$-dimensional identity matrix.

Next, we focus on a specific case of interest, when the joint process is jointly Markov with some order $k$. In this case, we know that, for $\mathrm{I}_{i,j}^{X,Y}$, when both $i$ and $j$ are larger than $k$, we obtain a similar value of conditional mutual information due to joint stationarity. This implies that within the square block within $\mathrm{I}^{X,Y}$ that is determined by the indices $\{(i,j)|i > k, j > k\}$, we have a Toeplitz structure (which is a matrix whose constant along its diagonals.). Furthermore, when $|i-j| > k$, due to the joint Markov nature of the process, we have $\mathrm{I}_{i,j}^{X,Y} = 0$. This implies a banded structure of the InfoMat, which is determined by a $k$-width 'Markov band' outside the main diagonal. This is useful for InfoMat estimation. As we further elaborate in the next sections, estimating the InfoMat may be generally computationally expensive, as we are required to estimate the $m^2$ conditional mutual information terms. However, when the joint process is a stationary Markov of order $k$, the number of distinct conditional mutual information terms reduces to $O(km)$. We demonstrate this relation between dependence structures and visual patterns on a simple

example in Figure 1. The discussion readily extends to asymmetric Markov orders for the $X$ and $Y$ processes.

Finally, we note that, as initially observed in Section 3.1, various areas in the InfoMat correspond to different directions of information. Specifically, values in the upper triangular correspond to information flow in the direction $X^m \to Y^m$, while information flow in the opposite direction is represented by the lower triangular. Therefore, the trace of $\mathrm{I}^{X,Y}$ represented the instantaneous exchange of information, also quantifying its operator norm. We believe that further relations and linear algebraic can be unveiled, with task and setting specific structures. In Section 6, we demonstrate these relations on InfoMat estimates from data.

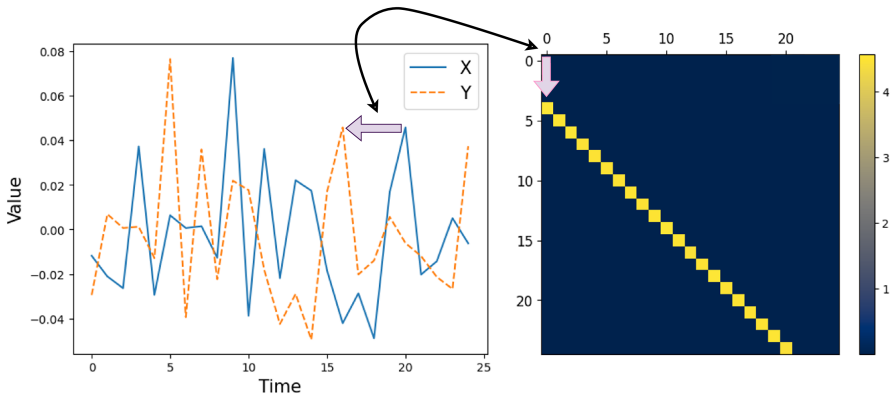



**Figure 1.** Visualizing temporal dependencies in the InfoMat. We take a simple Gaussian process, where $X_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and $Y_t = X_{t+4} + Z_t$ with $Z_t \overset{i.i.d.}{\sim} \mathcal{N}(0,0.1)$. In this case, the data show that (**left**) $Y_t$ follows $X_t$ with a delay of 4 time-steps. This is also visible in the InfoMat by a shift of 4 indices of the InfoMat Heatmap representation (**right**). The acquisition of the InfoMat Heatmap representation is explained in Section 6.

## 4. InfoMat Estimation via Gaussian Mutual Information

Beyond its theoretical merit as a proof visualization tool, we argue that the InfoMat is also effective for the visualization and analysis of dependence structures in time series data. However, the underlying data distribution is often unknown. Even if it is known, the corresponding conditional mutual information terms may not be given in closed form. To this end, to utilize the InfoMat as a visualization tool in real data setting, an estimator is required. In this section, we propose an approximation of $\mathrm{I}^{X,Y}$ from samples of the joint distribution $P_{X^m,Y^m}$.

Estimating $\mathrm{I}^{X,Y}$ boils down to the estimation of $m^2$ conditional mutual information terms. Without assumptions on the data distribution, the complexity and performance of mutual information estimators tend to deteriorate with the length of the conditioned joint history. That is, the bigger $(i,j)$ are, the more samples are required and the worse the performance of the single $\mathrm{I}^{X,Y}$ entry is expected to be. To that end, we begin by proposing a data-efficient estimation of mutual information, focusing on an approximation that follows from the Gaussian mutual information term, which is given in closed form. In this case, estimating the entries of $\mathrm{I}^{X,Y}$ boils down to the estimation of the corresponding covariance matrices, whose guarantees are well studied, and for which we can use estimators with parametric error rates. We begin by constructing and analyzing the proposed estimator. Then, we analyze its theoretical performance.

### 4.1. Proposed Estimator

Let $(X^n, Y^n) \sim P_{X^n,Y^n}$ be a given dataset from which we want to estimate $\mathrm{I}^{X,Y}$. For simplicity, we assume that all variables have zero mean. We begin by using the following representation of conditional mutual information.

**Lemma 1.** *Let $(X^m, Y^m) \sim P_{X^m,Y^m}$, and $1 \le i, j \le m$. Then,*

$$I(X_i; Y_j | X^{i-1}, Y^{j-1}) = H(X^i, Y^{j-1}) + H(X^{i-1}, Y^j) - H(X^{i-1}, Y^{j-1}) - H(X^i, Y^j). \tag{12}$$

*If $(X^m, Y^m)$ are jointly Gaussian, then*

$$I(X_i; Y_j | X^{i-1}, Y^{j-1}) = \mathrm{I}^{X,Y}_{G,i,j} =\triangleq \frac{1}{2} \log \frac{\left|\mathrm{K}_{X^i,Y^{j-1}}\right| \left|\mathrm{K}_{X^{i-1},Y^j}\right|}{\left|\mathrm{K}_{X^{i-1},Y^{j-1}}\right| \left|\mathrm{K}_{X^i,Y^j}\right|}, \tag{13}$$

*where $\mathrm{K}_Z$ is the covariance matrix of $Z \sim P_Z$, and $|\mathrm{K}_Z|$ representing its determinant.*

To estimate $\mathrm{I}^{X,Y}_{i,j}$ from a dataset $(x^n, y^n)$ using the Gaussian estimator (13), we estimate the corresponding sample covariance matrices $\hat{\mathrm{K}}_{i,j} \triangleq \hat{\mathrm{K}}_{X^i,Y^j}$, and plug those into (13). We denote the Gaussian estimator of $\mathrm{I}^{X,Y}_{i,j}$ with $\hat{\mathrm{I}}^{X,Y}_{G,i,j}(x^n, y^n)$. Finally, a Gaussian estimator of $\mathrm{I}^{X,Y}$ is an $m \times m$ matrix whose $(i, j)$ entry is given by $\hat{\mathrm{I}}^{X,Y}_{G,i,j}$. The Gaussian mutual information estimation procedure is summarized in Algorithm 1.

**Algorithm 1** Gaussian InfoMat Estimation

**input:** Data $(x^n, y^n)$, matrix length $m$
**output:** Gaussian estimate of $\mathrm{I}^{X,Y}$

Initialize $\hat{\mathrm{I}}^{X,Y}_{G,i,j} = 0$ for $(i, j) \in (1, \ldots, m) \times (1, \ldots, m)$

**for** $(i, j)$ in $(1, \ldots, m) \times (1, \ldots, m)$ **do**

Divide $(x^n, y^n)$ into datasets

$$((x^{i-1}, y^{j-1})_l, (x^i, y^{j-1})_l, (x^{i-1}, y^j)_l, (x^i, y^j)_l)_{l=1}^N$$

Calculate sample covariance matrices

Calculate $\hat{\mathrm{I}}^{X,Y}_{G,i,j}$ via (13).

**return** Estimated InfoMat.

The proposed Gaussian estimator for $\mathrm{I}^{X,Y}$ relies on the estimation of covariance matrices. Therefore, it inherits its guarantees from those of log determinants of sample covariance matrices. We assume that the underlying data-generating process is stationary, and obtain the dataset for the estimation of $\mathrm{I}^{X,Y}_{i,j}$ by dividing $(x^n, y^n)$ into the corresponding $i \times j$ sequences. We thus have the following

**Proposition 4** (Gaussian estimator performance guarantees)**.** *Let $(X^n, Y^n)$ be a sequence of jointly Gaussian random vectors over $\mathbb{R}^{d_x+d_y}$ and let $d = \max(d_x, d_y)$. Then*

1. *Bias:* $\lim_{n\to\infty} \mathbb{E}\left[\hat{\mathrm{I}}^{X,Y}_{G,i,j}\right] = \mathrm{I}_{G,i,j}$.
2. *Variance:* $\lim_{n\to\infty} \mathsf{Var}(\hat{\mathrm{I}}^{X,Y}_{G,i,j}) = O(\frac{dm^2}{n}) = O(\frac{1}{n})$.

The proof follows from the analysis in [38,39], and the dependence on $m^2$ follows from the division of the dataset $(x^n, y^n)$ into the corresponding subsequences.

We note that, as $m$ grows, the performance of the Gaussian estimator deteriorates, as the corresponding conditional mutual information term considers higher dimensional variables. To this end, we propose an alternative dataset acquisition approach—divide $(x^n, y^n)$ into $n - m$ samples such that the $l$th subsequence for $infomat_{i,j}$ is given by $(x_l^{l+i}, y_l^{l+j})$. This provides us with a significantly larger effective dataset when $n$ is not bigger than $m$ by orders of magnitude. However, the resulting sampled sequences are no longer i.i.d., and therefore, the estimator's guarantees no longer hold. We refer to such a dataset as a "correlated dataset", and use it for the visualization of real-world data when data availability is low.

While estimating $\mathrm{I}^{X,Y}_{G,i,j}$ is a simpler task, a Gaussian approximation can capture only partial information when the data distribution is far from a joint Gaussian. We propose an upper bound on the error of using the Gaussian approximation.

**Proposition 5.** *Let $(X, Y, Z) \sim P_{X,Y,Z}$ and let $(P_{X_G,Y_G,Z_G})$ be the corresponding Gaussian joint distribution with the same moments as $(P_{Z,Y,Z})$. We have the following bound:*

$$\begin{aligned} |I(X;Y|Z) - I(X_G;Y_G|Z_G)| \leq \\ \mathsf{D_{KL}}(P_{X,Y|Z} \| P_{X_G,Y_G|Z} | P_Z) - \mathsf{D_{KL}}(P_{X|Z} \otimes P_{Y|Z} \| P_{X_G|Z} \otimes P_{Y_G|Z} | P_Z) \\ + \max_{z \in \mathcal{Z}} \mathsf{D_{KL}}(P_{X_G,Y_G|Z_G=z} \| P_{X_G|Z_G=z} \otimes P_{Y_G|Z_g=z}) d_{\mathsf{TV}}(p_Z, p_{Z_G}) \end{aligned} \tag{14}$$

The proof of Proposition 5 is given in Appendix A.2.

### *4.2. InfoMat Estimation for Discrete Datasets*

When the data domain is discrete, i.e., $X^m$ and $Y^m$ are drawn from some finite sets $\mathcal{X}$ and $\mathcal{Y}$, respectively, the Gaussian mutual information estimator is no longer valid. In such a case, we propose a plug-in estimator of mutual information. The estimator relies on the entropic factorization in Lemma 1, followed by a standard plug-in estimator for each entropy term. For completeness, we provide more details on the plug-in estimation methodology and demonstrate its application to the InfoMat in Appendix A.3. In the proposed applications, the plug-in estimator had demonstrated satisfactory results. However, its complexity grows exponentially with the size of conditioned history. In these situations, context tree weighting methodologies [36] can be utilized, to which our approach seamlessly extends.

## 5. Beyond Gaussian—Neural Estimation

Despite its simplicity and data efficiency, the Gaussian mutual information approximation can only fully capture the dependence structure under strong assumptions, which are often violated [40]. To this end, we propose a conditional mutual information estimator that does not require joint Gaussianity. The algorithm relies on the concept of neural estimation [41–44], which utilizes a variational formula of the measure of interest and optimizes it with neural networks. With the purpose of maintaining the simplicity of the Gaussian method, we turn to a recent scheme that leverages normalizing flows. We utilize the recently proposed diffeomorphic conditional mutual information estimator [39] that leverages a type of network optimization scheme termed MAFs [45]. MAF-based estimators of mutual information map the jointly distributed pair into a corresponding Gaussian pair, such that the learned mapping is a diffeomorphism (which is a differentiable invertible map with a differentiable inverse). In what follows, we introduce the proposed estimator and discuss its performance.

### 5.1. Masked Autoregressive Flows

This section provides a high-level description of MAFs. For an in-depth discussion, we refer the reader to [39,45]. Consider a pair $(X, Y)$ that has a conditional distribution $P_{XY|Z}$ for some random variables $Z$, such that $P_{XY|Z=z}$ is absolutely continuous for any $z \in \mathcal{Z}$. The employed estimator consists of two stages and relies on obtaining a diffeomorphism that maps $(X, Y)$ into a Gaussian pair $(X', Y')$. The second stage consists of calculation of the $I(X'; Y'|Z)$ which has a simple form and will be a proxy for $I(X; Y|Z)$.

To learn a parametric diffeomorphism, we optimize an MAF. MAFs map samples from a (usually simple) base distribution $p(U)$ to samples from an arbitrary target distribution $p(X)$, assuming both are absolutely continuous and defined on $\mathcal{U} \subseteq \mathbb{R}^d$ and $\mathcal{X} \subseteq \mathbb{R}^d$ for some finite $d \in \mathbb{N}$. We denote the MAF with $T_\theta$. MAFs are designed such that their Jacobian is a triangular matrix. This implies that its determinant is simply given by the product of its diagonal. This property is crucial for the design as we are interested in representing the likelihood of the target distribution as a function of the likelihood of the base distribution and the partial derivatives along the parametric map $p_\theta$. For example, when $p(U) = \mathsf{Unif}[0,1]^d$, the parametric likelihood of $X$ under $T_\theta$ is given by

$$\log p_\theta(x) = \log(d) + \sum_{i=1}^{d} \log \left| \frac{\partial u_i}{\partial x_i} \right|, \tag{15}$$

where $\left| \frac{\partial u_i}{\partial x_i} \right|$ is the Jacobian of $T_\theta$. The map $T_\theta$ can be realized by neural networks with masked weight matrices and monotone activations [45]. Training an MAF consists of maximum likelihood optimization. Therefore, it is trained using minibatch gradient methods with (15) serving as the loss for $\theta$. This scheme readily adapts to conditional distributions $p_\theta(x|z)$. The generalization considers a conditioner model $g_{\theta'}(z)$ with parameters $\theta'$, whose purpose is to transfer the relevant information in $Z$ about $X$ into $T_\theta$. The conditional MAF is trained similarly to the unconditional MAF, with the slight change that now $x_i = f_\theta(u) + g_{\theta'}(z)$. Note that the conditioner model need not to be a diffeomorphism.

### 5.2. Proposed Estimator

Equipped with an MAF model, we can discuss the proposed diffeomorphic conditional mutual information estimator from [39]. We demonstrate the method on an unconditional mutual information term, and then discuss the required modification to introduce conditioning. Recall that our goal is to map the pair $(X, Y)$ into a Gaussian pair $(X', Y')$. We break this task into a concatenation of two stages. First, we will map $(X, Y)$ into a uniform distribution over $[0,1]^{d_x+d_y}$ by learning a MAF. Then, the uniform distribution will be mapped into a Gaussian distribution using the inverse of the Gaussian cumulative distribution function (CDF) This method is often referred to as generalized inverse transform sampling [46].

To map $(X, Y)$ into a pair of uniform random variables, the learned MAF maps each variable into a mixture of Gaussian CDFs. Specifically, for both $X$ and $Y$, we learn a mapping $\tau$ : as follows

$$\tau(x_i, h_i) = \sum_{j=1}^{k} w_{i,j}(h_i) \Phi(x_i; \mu_{i,j}(h_i), \sigma^2_{i,j}(h_i)),$$

where $\Phi$ is the Gaussian CDF with mean $\mu$ and variance $\sigma^2$, $(w_{i,j})_{i,j}$ are the model parameters, and $h_i$ is some parametric autoregressive summary map, i.e., $h_i = h_i(x_{<i})$. As explained in [39], the reasoning behind this design choice is that $\partial\tau/\partial x_i$ is a Gaussian mixture model, which is known to be a universal approximator of probability distributions. Finally, the decomposition of $\tau$ and the inverse Gaussian CDF yields a parametric model that maps an arbitrary random variable into a Gaussian variable. Thus, when applied to $X$ and $Y$, we result with a pair of Gaussian variables $X'$ and $Y'$.

Having mapped $(X, Y)$, which are distributed according to the conditional distribution $P_{XY|Z}$ into the Gaussian pair, we can use the following result on conditional mutual information invariance.

**Proposition 6** (Conditional mutual information invariance)**.** *Let* $(X, Y, Z) \sim P_{X,Y,Z}$ *and denote by* $(X', Y') = (f_\theta(X, Z), g_\phi(Y, Z))$*, where* $f_\theta$ *and* $g_\phi$ *are conditional diffeomorphisms. Then,*

$$I(X; Y|Z) = I(X'; Y'|Z) \tag{16}$$

Proposition 6 is a slight modification of ([39] (Lemma 2)). The existence of optimal MAFs is guaranteed by the universal approximation properties of normalizing flows [47]. Finally, the Gaussian mutual information term is calculated from sample covariance matrices, as elaborated in Section 4. The MAF-based scheme is depicted in Figure 2, and the algorithm steps are summarized in Algorithm 2.

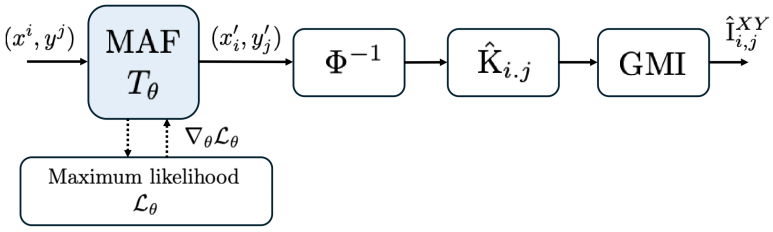



**Figure 2.** Neural estimation model. Dashed line represents maximum-likelihood (ML) training phase, while the filled lines account for the inference (mutual information calculation) phase.

To estimate $\mathrm{I}^{X,Y}$ from a given dataset, we apply Algorithm 2 to each coordinate pair $(i, j)$. The data are split in a similar fashion to Algorithm 1, but due to the parametric nature of the estimator, we optimize using iterative minibatch-gradient descent. In the training phase, we optimize all DMI models in parallel through the optimization of the corresponding maximum-likelihood loss (15) for a fixed number of epochs. When the training concludes, for each entry, $\mathrm{I}^{X,Y}_{i,j}$, we feed the corresponding dataset through the optimized MAFs and estimate the sample covariance matrices, from which we calculate the mutual information term. Neural network optimization is considerably expensive. However, ref. [39] shows that the DMI outperforms existing conditional mutual information estimators in terms of sample requirements. The proposed method boils down to the optimization of $m^2$ MAF models, which may be computationally expensive. We believe that this complexity can be alleviated by incorporating recurrent architectures or attention models. However, this investigation is out of the scope of this work.

**Algorithm 2** Neural InfoMat Estimation

**input:** Data $(x^n, y^n)$, matrix length $m$, number of epochs $N$.
**output:** Neural estimate of $\mathrm{I}^{X,Y}$

Initialize MAF parameters $\theta_{i,j}$ for $(i,j) \in (1,\ldots,m) \times (1,\ldots,m)$

**for** $(i,j)$ in $(1,\ldots,m) \times (1,\ldots,m)$ **do**

Divide $(x^n, y^n)$ into datasets

$$((x^{i-1}, y^{j-1})_l, (x^i, y^{j-1})_l, (x^{i-1}, y^j)_l, (x^i, y^j)_l)_{l=1}^N$$

Optimize $T_{\theta_{i,j}}$ for $N$ epoch via maximum likelihood (15)

Calculate $\hat{\mathrm{I}}^{X,Y}_{G,i,j}$ via (13) on $T_{\theta_{i,j}}$ to the sample set.

**return** Estimated InfoMat.

**Remark 1** (Performance of normalizing flows)**.** *MAFs are a recent promising method for the estimation of (conditional) mutual information using the expressive power of neural network, and were shown to outperform previous methodologies in an array of experiments [39]. However, using the MAF method separately on X and Y yields a pair of Gaussian variables, which are not guaranteed to be jointly Gaussian [48]. Consequently, the estimated mutual information (through the Gaussian formula) is, in general, a lower bound of $I(X;Y)$. Nonetheless, the proposed method showed good empirical performance in considered tasks, as we show in Section 6.*

**Remark 2** (Computational Complexity and Scalability)**.** *Estimating the $m \times m$ InfoMat, where each entry represents a conditional mutual information term, poses significant computational challenges. In the neural estimation approach, each of the $m^2$ entries is estimated via a separate neural network, leading to an overall computational complexity of $O(Tm^2)$, where $T$ is the complexity of estimating a single conditional mutual information term. This quickly becomes prohibitive as $m$ increases, both in terms of training time and memory usage. This complexity can potentially be reduced by incorporating weight or state sharing through recurrent architectures, which we aim to explore in future work. The current implementation offers two estimation options: the first is the neural estimator, which provides a more accurate estimation when ample data are available, albeit with increased computational overhead. The second is the Gaussian mutual information Estimator, which relies on covariance matrix estimation and offers a more computationally efficient alternative that performs well in low-data regimes or for moderate values of $m$.*

## 6. Visualization of Information Transfer

In this section, we demonstrate the utility of the InfoMat as a visualization tool for sequential data. We demonstrate how one can use the relations between information measures and InfoMat entry subsets (Section 3) to deduce temporal interactions in sequential data. We show that the InfoMat provides a more informative mode of information compared to existing measures, and propose measurements that can be coupled with the InfoMat estimate to deduce relationships in the data. Throughout, we adopt the interpretation that higher directed information in a certain direction implies a higher causal effect [36,43]. We analyze the InfoMat through its heatmap representation. Specifically, the heatmap $X$-axis corresponds to the $X$-process time index and $Y$-axis corresponds to the $Y$-process time index, as per the InfoMat definition (6). An implementation of considered experiments can be found at https://github.com/DorTsur/infomat (accessed on 22 March 2025).

### *6.1. Synthetic Data—Gaussian Processes*

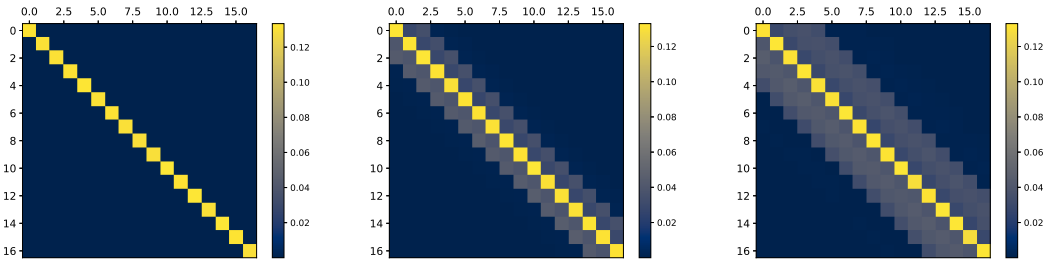
We begin with a sequential Gaussian process, which allows us to clearly present the relations between various dependence structures and the resulting InfoMat structure. Specifically, we consider the following joint Gaussian autoregressive (AR) process

$$\begin{aligned} X_t &= \sum_{k=0}^{\bar{k}_x} \alpha_k^X X_{t-k} + \alpha_k^Y Y_{i-k} + N_t^X, \quad t \in \mathbb{N} \\ Y_t &= \sum_{k=0}^{\bar{k}_y} \beta_k^X X_{t-k} + \beta_k^Y Y_{t-k} + N_t^Y, \end{aligned} \tag{17}$$

where $N_t^Y$ and $N_t^Y$ are samples of a centered i.i.d. Gaussian processes with covariance matrices $\mathrm{K}_{N_x}$ and $\mathrm{K}_{N_y}$, respectively. By controlling the values for the AR model parameters $(\bar{k}_x, \bar{k}_y, \alpha_k^X, \alpha_k^Y, \beta_k^X, \beta_k^Y)_{k=1}^m$, we induce various dependence structures on the sequence $(X^m, Y^m)$, which we then visualize via $\mathrm{I}^{X,Y}$. All visualizations in this subsection are obtained via Algorithm 1, with $n \approx 10^5$ samples, which as. We assume that the samples are given from the stationary distribution by omitting the first $\max(\bar{k}_x, \bar{k}_y)$ samples.).

We begin with a simple i.i.d. setting by taking $\beta_k^X = \gamma$ for $\gamma \in (0, 1)$ and nullifying the rest of the parameters. In this case, $X_i \perp\!\!\!\perp Y_j$ for $i \neq j$. Thus, all shared information is instantaneous, implying that we should expect a diagonal InfoMat. As seen in Figure 3a, this is indeed the case. The corresponding InfoMat captures the dependence structure, resulting in a diagonal matrix.

Next, denote $\bar{x}_x = \bar{k}_y = \bar{k}$. We increase the values of AR weights for $\bar{k} > 0$, inducing dependence in the history of the joint process. We consider a symmetric dependence structure and set $\alpha_k^X = \alpha_k^Y = \beta_k^X = \beta_k^Y = 0.3$ for $j \in \{0, \ldots, \bar{k}\}$ considering several values of $\bar{k}$. The value of $\bar{k}$ can be interpreted as how far into the past the dependence of the present terms extends. As shown in Figure 3b,c, the larger $\bar{k}$ is, the bigger the 'information band' around the diagonal, whose width depends on the value of $\bar{k}$. Consequently, we may deduce that the farther away we reside on the off-diagonal, the farther in history we observe dependence.

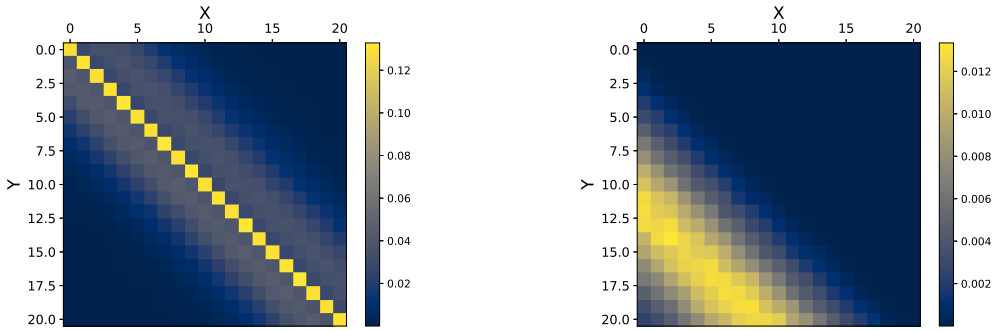



(**a**) (**b**) (**c**)

**Figure 3.** Visualization of history dependence in ARMA Gaussian process with various parameter settings. The bigger the value of $\bar{k}$ is, the bigger the information band around the instantaneous information, represented by the diagonal. (**a**) Gaussian i.i.d. ($\bar{k} = 0$). (**b**) Gaussian AR, $\bar{k} = 2$. (**c**) Gaussian AR, $\bar{k} = 4$.

Finally, we demonstrate how more complex temporal structures can be captured via the InfoMat. Specifically, we take two cases in which the ARMA parameters are time-varying, leading to nontrivial temporal relations between the processes. The first considers that ARMA parameters vary over time, i.e., the amount of parameters which are not nullified depends on the $(i, j)$ value. As seen in Figure 4a, this results in a decay in the aforementioned 'dependence band'. Second, we demonstrate a case of pure unidirectional information transfer in a single direction by introducing a time delay in the parameters

and a significant difference between their values. As shown in Figure 4b, the information transfer in dense in the lower triangular, which correspond to the DI term $I(Y^m \to X^m)$ (see Section 3). We conclude that the InfoMat successfully captures nontrivial temporal dependence structures.



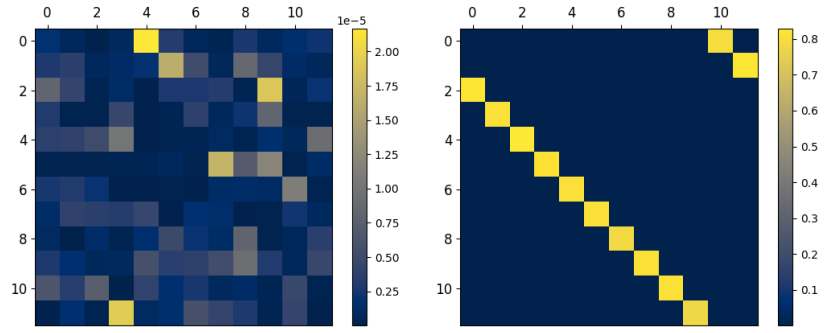

(**a**) (**b**)

**Figure 4.** Visualization of complex dependence structures in the Gaussian AR setting. (**a**) Decaying Gaussian AR, $\gamma = 0.5$. (**b**) One-sided information transfer.

### *6.2. Expressiveness of Neural Estimation*

As previously discussed, the efficient InfoMat estimation through Gaussian conditional mutual information estimation comes at the cost of a mismatch when the relations are, e.g., nonlinear. Herein, we demonstrate the utility of neural estimation (Section 5) to the InfoMat capabilities by comparing its performance with the Gaussian mutual information estimator. We take an i.i.d. jointly Gaussian dataset with correlation coefficient $\rho = 0.9$. We then apply a cyclic shift of $T < m$ to the samples $Y^m$ within each $m$-length sequence, effectively resulting in a time-shift dependence structure. We introduce nonlinearity by considering the mapping $X_i \mapsto \log(X_i)$ and $Y_i \mapsto Y_i^3$. Such mappings are invertible and therefore the overall mutual information should be preserved.

As visualized in Figure 5, the neural estimator successfully recovers the correct structure in the data, while the Gaussian mutual information estimator fails to provide a meaningful visualization in the given setting. As the mappings are invertible, the resulting mutual information is 0.83 [nats] on nonzero entries, which are approximately the corresponding values in the neural estimator of $\mathrm{I}^{X,Y}$. However, this accuracy comes at the cost of training $m^2$ neural nets, which is significantly slower than calculating sample covariance matrices. Consequently, we argue that one should consider neural estimation when the functional nature of the data are complex, and consider the Gaussian estimator when $m$ is large, or when simple, initial results may be of need.




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

**Figure 5.** Estimated InfoMat under nonlinearities and cyclic shift. (**a**) Gaussian mutual information. (**b**) Neural estimator.

### *6.3. Visualizing Information Flow in Physiological Data*

Directed information and transfer entropy have been previously addressed as measures of causal effect between two interacting processes [36,49]. To that end, they have been used to quantify and compare flows of information in stochastic systems. We demonstrate this paradigm with the InfoMat, while showing its applicability to visualize dynamics in real-world datasets. We consider the Apnea dataset from Santa Fe Time Series Competition (https://physionet.org/content/santa-fe/1.0.0/ (accessed on 22 March 2025)) [50,51]. The Apnea dataset is common benchmark for the evaluation of transfer entropy estimation. It is a sequential dataset which consists of measures of heart rate and chest volume (representing respiration force). The Apnea dataset was previously addressed in the literature as a case study to understand the relation between sequential information measures and causal effect. We estimate the InfoMat on the Apnea dataset, denoting the interacting processes at hand being $(X_t)_{t\in\mathbb{N}} = \mathsf{Heart}$ and $(Y)_{t\in\mathbb{N}} = \mathsf{Breath}$. The visualization of the estimated InfoMat is given in Figure 6.

It was shown in [37,49] that the transfer entropy in the direction $\mathsf{Breath} \to \mathsf{Heart}$ is higher than the transfer entropy measures in the other direction. By calculating the directed information in each direction, we recover the same conclusion on the relationship in the data. Specifically, we have

$$\hat{I}(\mathrm{D}\circ X^n \to Y^n) = 0.14 < 0.41 = \hat{I}(\mathrm{D}\circ Y^n \to X^n), \quad \hat{I}_{inst}(X^n, Y^n) = 0.034.$$

This calculation implies that the causal effect in the $\mathsf{Breath} \to \mathsf{Heart}$ direction is the prominent one, which is in agreement with the previous literature. This conclusion provides another validation of the InfoMat method consistency.

Beyond its agreement with known previous results, the InfoMat provides a more informative observation of the exchange of control between the two measures. For example, we observe that most of the information transfer occurs in the first upper subdiagonal, which corresponds to $I(X_{i-1}; Y_i | X^{i-2}, Y^{i-1})$, i.e., most information is transferred from the previous time step to the next one. In the opposite direction, we have a significantly smaller information transfer. Surprisingly, information is transferred from the steps further in the process past, i.e., the effect is from $Y_{t-4}^{t-2}$ to $X_t$. These results further motivate the use of the InfoMat as a visual tool for the task of exploratory data analysis, we apply it to real-world data.

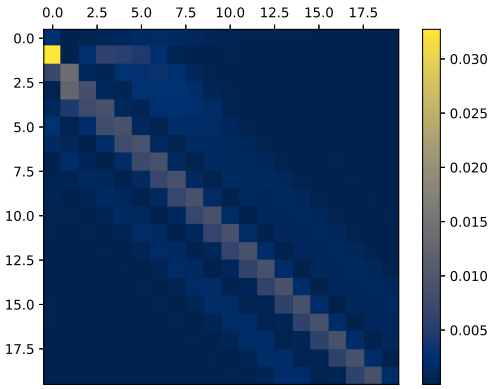



**Figure 6.** Physiological data. Larger effect in observed in the 'breath'→'heart' direction.

### *6.4. Visualizing Coding Schemes Effect*

As a final application, we demonstrate an application of the InfoMat to analyze information flow in digital communication schemes. In this case, $X^m$ and $Y^m$ represent the input and output of some communication channel with memory. The joint distribution of

$(X^m, Y^m)$ is determined by the channel transition kernel input distribution. The purpose of the encoder, which generates the sequence $X^m$ according to some causally conditional law $P(X^m \| Y^{m-1}) \triangleq \prod_{i=1}^{m} P(X_i | Y^{i-1} X^{i-1})$ [26], is to maximize the downstream directed information. Formally, under mild assumptions, the *feedback* capacity is characterized by the following optimization problem [7]

$$\mathsf{C} = \lim_{n \to \infty} \sup_{P(X^n \| Y^{n-1})} \frac{1}{n} I(X^n \to Y^n). \tag{18}$$

where $\mathsf{C}$ is termed the 'feedback capacity' of the communication channel, which is determined by the causal conditional law $P(Y^m \| X^m) \triangleq \prod_{i=1}^{m} P(Y_i | Y^{i-1} X^i)$. Solving the channel capacity optimization (18) provides the user with the value of maxima achievable rate of reliable communications, coupled with a coding scheme that achieves this rate under block length asymptotics. In what follows, we visualize the InfoMat for several channels, under both a channel oblivious coding scheme and the capacity-achieving coding scheme. As the data for this application are discrete, we utilize a simple plug-in estimator for the considered conditional mutual information terms. For completeness, we provide a characterization and analysis of the plug-in entropy estimator in Appendix A.3.

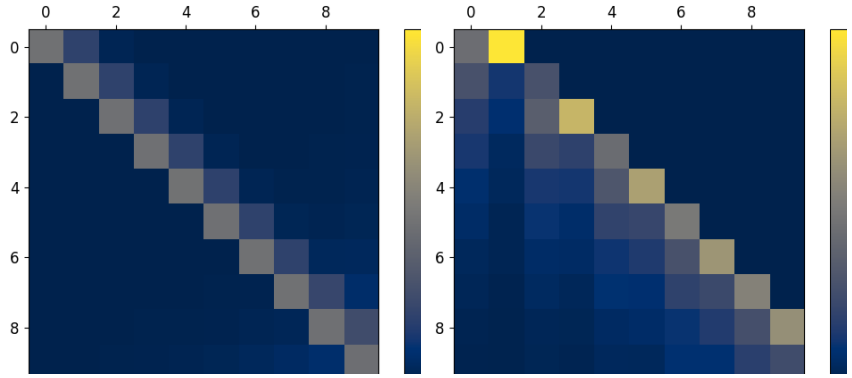
#### 6.4.1. Ising Channel

As a first example, we consider the Ising channel [52], which is a popular example of a communication channel with memory, which adheres to the famous Ising model. In this case, the channel law is defined according to the relation

$$Y_t = \begin{cases} X_t, & \text{w.p.} \quad 0.5 \\ X_{t-1}, & \text{w.p.} \quad 0.5 \end{cases}$$

The feedback capacity-achieving coding scheme was obtained in [53] by representing the Ising channel is a finite state channel [7], which allowed for a dynamic programming approximation of the corresponding optimization.

We estimate the InfoMat in the Ising channel under two coding schemes. The first, which we refer to as the oblivious scheme, sends $X^m \overset{i.i.d.}{\sim} \mathrm{Ber}(1/2)$ independently of the channel outputs. (Figure 7a). The second generates $X^m$ according to [53] (Figure 7b). We note that the oblivious scheme generates an InfoMat with most of its information in the main diagonal and a small residue in the off-diagonal. The diagonal information follows from the i.i.d. scheme, and it is constant along all time steps. The nonzero off-diagonal entries are due to event $Y_t = X_{t-1}$ which injects memory through the channel transition kernel. When we consider the feedback capacity-achieving scheme from [53], we result with a non-constant pattern of information flow. Specifically, most of the information is sent along the off diagonal, i.e., most of the information is sent through the effect of past channel inputs. Additionally, we note that the amount of conveyed information is non-constant. This is a result of the underlying finite-state machine that defines the evolution of $X^m$ according to past inputs, outputs and states ([53] Figure 5).




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

**Figure 7.** Visualization of information transfer in the Ising channel under various coding schemes. (**a**) Oblivious coding scheme. (**b**) Optimal coding scheme.

Finally, we calculate the normalized directed information for each scheme, as it serves as the proxy for the information rate in the channel. We have

$$\frac{1}{m}\hat{I}_{\text{i.i.d.}}(X^m \to Y^m) \approx 0.45, \quad \frac{1}{m}\hat{I}_{\text{opt}}(X^m \to Y^m) \approx 0.56$$

We note that, as expected, the capacity-achieving scheme yields a significantly greater normalized directed information. The resulting quantity is very close to the theoretical capacity values $(0.575)$. We conjecture that the mismatch results from the plug-in estimation error. Finally, we note that, in contrast to the i.i.d. scheme, the optimal scheme generates information in the direction $Y \to X$ as well, which quantifies the usage of feedback in the scheme.

#### 6.4.2. Trapdoor Channel

Next, we visualize the effect of the coding strategy on the transfer of information in the Trapdoor channel. The trapdoor channel is an example of a binary finite-state channel whose state and output evolve according to the relation

$$Y_t = \begin{cases} \mathsf{Z}_{1/2}(X_t), & \text{if } S_{t-1} = 0 \\ \mathsf{S}_{1/2}(X_t), & \text{otherwise} \end{cases}, \quad S_t = S_{t-1} \oplus X_t \oplus Y_t,$$

where $\mathsf{Z}_{1/2}$ and $\mathsf{S}_{1/2}$ denote the Z- and S-channels with a probability of $1/2$. The output of a $\mathsf{Z}_p$ $(\mathsf{S}_p)$ channel equals its input when the latter is 0 (1) and distributes according to $\mathsf{Ber}(p)$ otherwise. These channels are fundamental and have been extensively investigated in the literature [4,54]. Again, we consider the channel oblivious coding scheme and the optimal coding scheme from [7], as can be seen in Figure 8. We note that the highest amount of conveyed information is in the beginning of transmission. Notably, the information transfer under the optimal Trapdoor scheme is more uniform than the optimal Ising coding scheme. In this case, the estimated normalized directed information is

$$\frac{1}{m}\hat{I}_{\text{i.i.d.}}(X^m \to Y^m) \approx 0.441, \quad \frac{1}{m}\hat{I}_{\text{opt}}(X^m \to Y^m) \approx 0.663$$

Again, we note that the optimal coding strategy induces information transfer in the backward direction $Y \to X$ due to the incorporation of feedback into the input distribution.

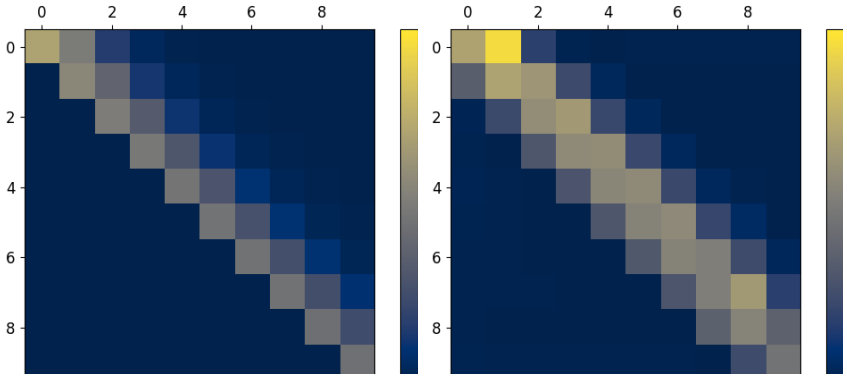



**Figure 8.** Visualization of information transfer in the Trapdoor channel under various coding schemes. (**a**) Oblivious coding scheme. (**b**) Optimal coding scheme.

## 7. Conclusions

In this work, we developed the InfoMat, a matrix representation of information exchange. We showed the utility of the InfoMat for the visual proofs of information conservation laws via matrix coloring arguments, which allowed us to expand the existing set of decompositions for information measures. Then, we proposed several estimators of the InfoMat, which were studied both theoretically and empirically. Equipped with the InfoMat estimators, we presented several applications of the InfoMat as a visualization tool to analyze the dependence structures and information transfer in sequential datasets in various settings. For future work, we aimed to develop a computationally efficient neural estimator of the InfoMat using weight sharing, sequential architectures [55], and slicing techniques [56]. Given this work's simplicity and the popularity of information measures, the InfoMat can serve as an effective tool for data exploration in sequential data analysis pipelines. Furthermore, we believe that the InfoMat can be highly useful for a myriad of contemporary research fields that involve time series. Such fields encompass empowerment [14], which characterizes robust sequential decision-making via information theory, and causal inference [16], in which information theory has been shown to be beneficial. Finally, we aim to investigate the multivariate extensions of the InfoMat, as it is central to contemporary setting. This extension can be obtained by constructing higher-dimensional information tensors that capture conditional mutual information among multiple signals. while this extension can uncover rich patterns of inter-dependencies in complex datasets, it also introduces new challenges in computational scalability and visualization clarity.

**Author Contributions:** Conceptualization, methodology, software, formal analysis, investigation and project administration: D.T.; Writing: D.T. and H.P.; Supervision: H.P.; Funding acquisition: H.P. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by Israel Science Foundation (ISF), grant numbers 3211/23 and 899/21.

**Institutional Review Board Statement:** Not applicable.

**Data Availability Statement:** All data and experiments are available at the publig GirHub repository at https://github.com/DorTsur/infomat (accessed on 22 March 2025).

**Acknowledgments:** This article is a revised and expanded version of a paper [57], which was presented at the International Symposium on Information Theory (ISIT), Athens, Greece, 7–12 July 2024.

**Conflicts of Interest:** The authors declare no conflicts of interest.

# Appendix A

## *Appendix A.1. Proofs*

### Appendix A.1.1. Proof of Proposition 2

The proof utilizes the observation of Proposition 1. We decompose mutual information, which is given by the entire matrix, into the upper and lower sub-triangulars (excluding the main diagonal), and the main diagonal. As noted in the main text, $I_{\mathsf{inst}}(X^n, Y^n)$ corresponds to the main diagonal (black), $T_{i+1}^{X\to Y}(i,i)$ corresponds to a sub-column and $T_{i+1}^{Y\to X}(i,i)$ corresponds to a sub-row. We therefore have

$$\begin{pmatrix} \mathrm{I}_{1,1}^{X,Y} & \mathrm{I}_{1,2}^{X,Y} & \mathrm{I}_{1,3}^{X,Y} & \dots & \mathrm{I}_{1,n}^{X,Y} \\ & \mathrm{I}_{2,2}^{X,Y} & \mathrm{I}_{2,3}^{X,Y} & \ddots & \vdots \\ & & \mathrm{I}_{3,3}^{X,Y} & \ddots & \vdots \\ & & & \ddots & \mathrm{I}_{n-1,n}^{X,Y} \\ & & & & \mathrm{I}_{n,n}^{X,Y} \end{pmatrix} = \begin{pmatrix} \mathrm{I}_{1,1}^{X,Y} & \mathrm{I}_{1,2}^{X,Y} & \mathrm{I}_{1,3}^{X,Y} & \dots & \mathrm{I}_{1,n}^{X,Y} \\ & \mathrm{I}_{2,2}^{X,Y} & \mathrm{I}_{2,3}^{X,Y} & \ddots & \vdots \\ & & \mathrm{I}_{3,3}^{X,Y} & \ddots & \vdots \\ & & & \ddots & \mathrm{I}_{n-1,n}^{X,Y} \\ & & & & \mathrm{I}_{n,n}^{X,Y} \end{pmatrix} \tag{A1}$$

$$\begin{pmatrix} \mathrm{I}_{1,1}^{X,Y} & \mathrm{I}_{1,2}^{X,Y} & \mathrm{I}_{1,3}^{X,Y} & \mathrm{I}_{1,4}^{X,Y} & \dots & \mathrm{I}_{1,n}^{X,Y} \\ \mathrm{I}_{2,1}^{X,Y} & \mathrm{I}_{2,2}^{X,Y} & \mathrm{I}_{2,3}^{X,Y} & \mathrm{I}_{2,4}^{X,Y} & \ddots & \vdots \\ \mathrm{I}_{3,1}^{X,Y} & \mathrm{I}_{3,2}^{X,Y} & \mathrm{I}_{3,3}^{X,Y} & \mathrm{I}_{3,4}^{X,Y} & \ddots & \vdots \\ \mathrm{I}_{4,1}^{X,Y} & \mathrm{I}_{4,2}^{X,Y} & \mathrm{I}_{4,3}^{X,Y} & \mathrm{I}_{4,4}^{X,Y} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \mathrm{I}_{n-1,n}^{X,Y} \\ \mathrm{I}_{n,1}^{X,Y} & \dots & \dots & \dots & \mathrm{I}_{n,n-1}^{X,Y} & \mathrm{I}_{n,n}^{X,Y} \end{pmatrix} \tag{A2}$$

$$\begin{aligned} &I(X^n;Y^n) \\ &= (T_2^{X\to Y}(1,1) + T_3^{X\to Y}(2,2) + \cdots + T_n^{X\to Y}(n-1,n-1)) \\ &+ I_{\mathsf{inst}}(X^n, Y^n) \\ &+ (T_2^{Y\to X}(1,1) + T_3^{Y\to X}(2,2) + \cdots + T_n^{Y\to X}(n-1,n-1)) \end{aligned}$$

### Appendix A.1.2. Proof of Proposition 3

The relation follows from noting that a delayed directed information term $I(D^{k+1} \circ X^n \to Y^n)$ corresponds to a subtriangular element, which forms the one-step reduced directed information element $I(D^k \circ X^n \to Y^n)$ when combined with the appropriate sub-diagonal, which, in turn, correspond to a 'delayed' instantaneous mutual information term. For example, when $k = 0$, it is given by (A1), which we decompose by means of coloring and elements-gathering as

$$I(X^n \to Y^n) = I_{\mathsf{inst}}(X^n, Y^n) + I(\mathsf{D} \circ X^n \to Y^n)$$

### Appendix A.1.3. Proof of Proposition 4

Let $Z^n$ be a set of $n$ samples from $\mathcal{N}(0, \mathrm{K}_Z)$, defined on $\mathbb{R}^d$.

**Bias:** According to [38], the bias of $\log|\mathrm{K}_Z|$ is given by

$$\tau_{n,d} \triangleq \sum_{k=1}^{d} \left( \psi\left(\frac{n-k+1}{2}\right) \log\frac{n}{2} \right),$$

where $\psi(z)$ is the digamma function, which is known to asymptotically behave as $\psi(z) \approx \log(z) - \frac{z}{2}$. Thus, the bias $\tau_{n,d}$ term behaves as $1/n$, vanishing at $n \to \infty$.

**Variance:** We have the following:

$$\begin{aligned}\mathsf{Var}(\hat{\mathrm{I}}^{X,Y}_{G,i,j}) \leq \mathsf{Var}(\log|\hat{\mathrm{K}}_{X^i,Y^{j-1}}|) + \mathsf{Var}(\log|\hat{\mathrm{K}}_{X^{i-1},Y^j}|) \\ + \mathsf{Var}(\log|\hat{\mathrm{K}}_{X^{i-1},Y^{j-1}}|) + \mathsf{Var}(\log|\hat{\mathrm{K}}_{X^i,Y^j}|).\end{aligned}$$

Thus, the estimator variance is governed by the variance of the log-determinant estimator. Following the analysis in the proof of ([39] Lemma 6),

$$\mathsf{Var}(\log|\mathrm{K}_Z|) \xrightarrow{L} O\left(\frac{d_Z}{n}\right)$$

In our case, the dimension of $\mathrm{K}_{X^i,Y^j}$ is $d_x i + d_y j$, which is upper bounded by $2dm$, and the number of samples for the estimation of $\mathrm{K}_{X^i,Y^j}$ is $n/\min(i,j)$, which is lower bounded by $n/m$. As these bounds hold for all four sample covariance matrices, we have

$$\mathsf{Var}\left(\hat{\mathrm{I}}^{X,Y}_{G,i,j}\right) = O\left(\frac{dm}{n/m}\right) = O\left(\frac{dm^2}{n}\right),$$

which is sharp in $n$.

Appendix A.1.4. Proof of Lemma 1

Let $(X^m, Y^m) \sim P_{X^m,Y^m}$. The entropy decomposition of conditional mutual information (12) follows from the following steps

$$\begin{aligned}&I(X_i;Y_j|X^{i-1},Y^{j-1}) \\ &= H(X_i|X^{i-1},Y^{j-1}) + H(Y_j|X^{i-1},Y^{j-1}) \\ &\quad - H(X_i,Y_j|X^{i-1},Y^{j-1}) \\ &= H(X^i,Y^{j-1}) - H(X^{i-1},Y^{j-1}) + H(X^{i-1},Y^j) \\ &\quad - H(X^{i-1},Y^{j-1}) - (H(X^i,Y^j) - H(X^{i-1},Y^{j-1})) \\ &= H(X^i,Y^{j-1}) + H(X^{i-1},Y^j) \\ &\quad - H(X^{i-1},Y^{j-1}) - H(X^i,Y^j).\end{aligned}$$

The formula for the Gaussian case (13) follows using the definition of multivariate Gaussian entropy [4].

*Appendix A.2. Analysis of the Gap Between Mutual Information and Gaussian Mutual Information*

We provide an upper bound on the error of employing the Gaussian mutual information term instead of $I(X;Y|Z)$. We analyze the gap for arbitrary $(i,j)$ and for simplicity we denote $X_i = X$, $Y_j = Y$ and $(X^{i-1},Y^{j-1}) = Z$. Thus $\mathrm{I}^{X,Y}_{i,j} = I(X;Y|Z)$. We denote the with $(X_G,Y_G,Z_G)$ the jointly Gaussian triplet whose first and second moments are similar to those of $(X,Y,Z)$. Consequently, the Gaussian conditional mutual information term corresponds to $I(X_G;Y_G|Z_G)$. We utilize the following result from [58]

$$\begin{aligned}&I(X;Y) - I(X_G;Y_G) \\ &= \mathsf{D}_{\mathsf{KL}}(P_{X,Y}\|P_{X_G,Y_G}) - \mathsf{D}_{\mathsf{KL}}(P_X \otimes P_Y\|P_{X_G} \otimes P_{Y_G}).\end{aligned}$$

We have

$$\begin{aligned} I(X;Y|Z) - I(X_G;Y_G|Z_G) \\ \leq \underbrace{I(X;Y|Z) - I(X_G;Y_G|Z)}_{\triangleq \Delta_G} \\ + \underbrace{I(X_G;Y_G|Z) - I(X_G;Y_G|Z_G)}_{\triangleq \Delta_Z}. \end{aligned}$$

We analyze each error term separately. For $\Delta_G$, we have

$$\begin{aligned} \Delta_G &= \int_{\mathcal{Z}} (I(X;Y|Z=z) - I(X_G;Y_G|Z=z)) p_Z(z)\, \mathrm{d}z \\ &= \int_{\mathcal{Z}} \Big( \mathsf{D}_{\mathsf{KL}}(P_{X,Y|Z=z} \| P_{X_G,Y_G|Z=z}) \\ &- \mathsf{D}_{\mathsf{KL}}(P_{X|Z=z} \otimes P_{Y|Z=z} \| P_{X_G|Z=z} \otimes P_{Y_G|Z=z}) \Big) p_Z(z)\, \mathrm{d}z \\ &= \mathsf{D}_{\mathsf{KL}}(P_{X,Y|Z} \| P_{X_G,Y_G|Z} | P_Z) \\ &- \mathsf{D}_{\mathsf{KL}}(P_{X|Z} \otimes P_{Y|Z} \| P_{X_G|Z} \otimes P_{Y_G|Z} | P_Z). \end{aligned}$$

To bound the second error, we insert an additional term, which considers the KL term conditioned on $Z_G$, while integrated with respect to $P_Z$, given by

$$\mathbb{E}_Z \Big[ \mathsf{D}_{\mathsf{KL}}(P_{X_G,Y_G|Z_G} \| P_{Z_G|Z_G} \otimes P_{Y_G|Z_G} | P_{Z_G}) \Big].$$

We therefore have

$$\begin{aligned} \Delta_Z &= \int_{\mathcal{Z}} \mathsf{D}_{\mathsf{KL}}(P_{X_G,Y_G|Z=z} \| P_{X_G|Z=z} \otimes P_{Y_G|Z=z}) p_z(z)\, \mathrm{d}z \\ &- \int_{\mathcal{Z}} \mathsf{D}_{\mathsf{KL}}(P_{X_G,Y_G|Z_G=z} \| P_{X_G|Z_G=z} \otimes P_{Y_G|Z_g=z}) p_{z_G}(z)\, \mathrm{d}z \\ &= \int_{\mathcal{Z}} \mathsf{D}_{\mathsf{KL}}(P_{X_G,Y_G|Z=z} \| P_{X_G|Z=z} \otimes P_{Y_G|Z=z}) p_z(z) \\ &- \int_{\mathcal{Z}} \mathsf{D}_{\mathsf{KL}}(P_{X_G,Y_G|Z_G=z} \| P_{X_G|Z_G=z} \otimes P_{Y_G|Z_G=z}) p_z(z) \\ &+ \int_{\mathcal{Z}} \mathsf{D}_{\mathsf{KL}}(P_{X_G,Y_G|Z_G=z} \| P_{X_G|Z_G=z} \otimes P_{Y_G|Z_G=z}) p_z(z) \\ &- \int_{\mathcal{Z}} \mathsf{D}_{\mathsf{KL}}(P_{X_G,Y_G|Z_G=z} \| P_{X_G|Z_G=z} \otimes P_{Y_G|Z_g=z}) p_{z_G}(z)\, \mathrm{d}z \\ &= \mathsf{D}_{\mathsf{KL}}(P_{X_G,Y_G|Z} \| P_{X_G,Y_G|Z_G} | P_Z) \\ &+ \mathsf{D}_{\mathsf{KL}}(P_{X_G|Z} \otimes P_{Y_G|Z} \| P_{X_G|Z_G} \otimes P_{Y_G|Z_G} | P_Z) \\ &= \int_{\mathcal{Z}} \mathsf{D}_{\mathsf{KL}}(P_{X_G,Y_G|Z_G=z} \| P_{X_G|Z_G=z} \otimes P_{Y_G|Z_g=z}) \\ &\cdot (p_Z(z) - p_{z_G}(z))\, \mathrm{d}z, \end{aligned}$$

where the last term can be upper bounded by

$$\begin{aligned} &= \int_{\mathcal{Z}} \mathsf{D}_{\mathsf{KL}}(P_{X_G,Y_G|Z_G=z} \| P_{X_G|Z_G=z} \otimes P_{Y_G|Z_g=z}) \\ &\qquad \cdot (p_Z(z) - p_{z_G}(z))\, \mathrm{d}z \\ &\leq \max_{z \in \mathcal{Z}} \mathsf{D}_{\mathsf{KL}}(P_{X_G,Y_G|Z_G=z} \| P_{X_G|Z_G=z} \otimes P_{Y_G|Z_g=z}) \\ &\qquad \cdot d_{\mathsf{TV}}(p_Z, p_{z_G}). \end{aligned}$$

We note that the resulting upper bound extends the result from [58]. We also note that, as desired, $(\Delta_G + \Delta_Z) \to 0$ when $P_{X,Y,Z} \to P_{X_G,Y_G,Z_G}$. This concludes the proof.

*Appendix A.3. Additional Information on Plug-In Entropy Estimator*

This section mainly follows the presentation and analysis from [59]. We begin by describing the plug-in entropy estimator for a general variable $X \sim P_X$ with finite alphabet $|\mathcal{X}| < \infty$. Then, we build upon this description to describe the employed entropy estimation method in the InfoMat setting.

Let $X^n \overset{i.i.d.}{\sim} P_X$ and w.l.o.g. let $\mathcal{X} = [0, 1, \ldots, k]$. The plug-in estimator of $P_X$ from $X^n$ is given by frequency counting, i.e.,

$$\hat{P}_n(i) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}_{X_j=i}.$$

Consequently, the plug-in estimator of $H(X)$ is given by $\hat{H}_n(X^n) = \mathbb{E}_{\hat{P}_n}[-\log \hat{P}_n]$. Thus, to estimate the entropy of $X$ from a sample $X^n$, we estimate the empirical distribution of $X$, and then plug it into the entropy expectation. The plug-in entropy estimator has well-established theoretical guarantees [59,60] and is very simple to implement. To mention a few, $\hat{H}_n(X^n)$ is a consistent estimator of $H(X)$, both in the $P_X$-a.s. and $L_2$ sense, its variance decays with $\log n/n^2$, it has strong concentration properties, and the estimation error asymptotically behaves as a centered Gaussian distribution.

Following the construction of the plug-in entropy estimator, given a sample of $n$ samples $D^n_{i,j} \triangleq (X^i_{(l)}, Y^j_{(l)})^n_{l=1}$, we utilize the conditional mutual information entropy decomposition (12), to result with

$$\hat{I}_n(D^n_{i,j}) = \hat{H}_n(D^n_{i,j-1}) + \hat{H}_n(D^n_{i-1,j}) - \hat{H}_n(D^n_{i-1,j-1}) - \hat{H}_n(D^n_{i,j}).$$

As conditional mutual information is a linear combination of entropy terms, it benefits from many of the entropy guarantees, and its error is governed by the joint entropy term, and is linear in the alphabet of the joint variable (as the plug-in proofs summarize the errors over the various alphabet terms). Specifically, in our case, the error of $\mathrm{I}^{X,Y}_{i,j}$ is linear in $|\mathcal{X}|^i|\mathcal{Y}|^j$, which can easily 'explode' for large matrices. Other existing estimators, such as context tree weighting estimators [36], have better time and space complexity, but at the cost of a more complex algorithm that still suffers from the same exponential dependence on the alphabet size.

## References


1. Battiti, R. Using mutual information for selecting features in supervised neural net learning. *IEEE Trans. Neural Netw.* **1994**, *5*, 537–550.
2. Viola, P.; Wells, W.M., III. Alignment by maximization of mutual information. *Int. J. Comput. Vis.* **1997**, *24*, 137–154.
3. Tishby, N.; Pereira, F.C.; Bialek, W. The information bottleneck method. *arXiv* **2000**, arXiv:physics/0004057.
4. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley: New York, NY, USA, 2006.
5. Massey, J. Causality, feedback and directed information. In Proceedings of the International Symposium on Information Theory and Its Applications (ISITA-90), Waikiki, Hawaii, 27–30 November 1990; pp. 303–305.
6. Permuter, H.H.; Kim, Y.H.; Weissman, T. Interpretations of directed information in portfolio theory, data compression, and hypothesis testing. *IEEE Trans. Inf. Theory* **2011**, *57*, 3248–3259.
7. Permuter, H.; Cuff, P.; Van Roy, B.; Weissman, T. Capacity of the trapdoor channel with feedback. *IEEE Trans. Inf. Theory* **2008**, *54*, 3150–3165.
8. Wibral, M.; Vicente, R.; Lizier, J.T. *Directed Information Measures in Neuroscience*; Springer: Berlin/Heidelberg, Germany, 2014; Volume 724.
9. Schreiber, T. Measuring information transfer. *Phys. Rev. Lett.* **2000**, *85*, 461. [PubMed]

10. Wibral, M.; Vicente, R.; Lindner, M. Transfer entropy in neuroscience. In *Directed Information Measures in Neuroscience*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 3–36.
11. Tanaka, T.; Esfahani, P.M.; Mitter, S.K. LQG control with minimum directed information: Semidefinite programming approach. *IEEE Trans. Autom. Control* **2017**, *63*, 37–52.
12. Tiomkin, S.; Tishby, N. A unified bellman equation for causal information and value in markov decision processes. *arXiv* **2017**, arXiv:1703.01585.
13. Sabag, O.; Tian, P.; Kostina, V.; Hassibi, B. Reducing the LQG Cost with Minimal Communication. *IEEE Trans. Autom. Control* **2022**.
14. Klyubin, A.S.; Polani, D.; Nehaniv, C.L. Empowerment: A universal agent-centric measure of control. In Proceedings of the 2005 IEEE Congress on Evolutionary Computation, Edinburgh, UK, 2–5 September 2005; Volume 1, pp. 128–135.
15. Salge, C.; Glackin, C.; Polani, D. Empowerment—An introduction. In *Guided Self-Organization: Inception*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 67–114.
16. Raginsky, M. Directed information and pearl's causal calculus. In Proceedings of the 2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA, 28–30 September 2011; pp. 958–965.
17. Shu, Y.; Zhao, J. Data-driven causal inference based on a modified transfer entropy. *Comput. Chem. Eng.* **2013**, *57*, 173–180.
18. Wieczorek, A.; Roth, V. Information theoretic causal effect quantification. *Entropy* **2019**, *21*, 975. [CrossRef]
19. Zhou, Y.; Spanos, C.J. Causal meets submodular: Subset selection with directed information. *Adv. Neural Inf. Process. Syst.* **2016**, *29*. Available online: https://proceedings.neurips.cc/paper_files/paper/2016/file/81ca0262c82e712e50c580c032d99b60-Paper.pdf (accessed on 20 March 2025).
20. Kalajdzievski, D.; Mao, X.; Fortier-Poisson, P.; Lajoie, G.; Richards, B. Transfer Entropy Bottleneck: Learning Sequence to Sequence Information Transfer. *arXiv* **2022**, arXiv:2211.16607.
21. Bonetti, P.; Metelli, A.M.; Restelli, M. Causal Feature Selection via Transfer Entropy. *arXiv* **2023**, arXiv:2310.11059.
22. Friendly, M. Corrgrams: Exploratory displays for correlation matrices. *Am. Stat.* **2002**, *56*, 316–324.
23. Correa, C.D.; Lindstrom, P. The mutual information diagram for uncertainty visualization. *Int. J. Uncertain. Quantif.* **2013**, *3*, 187–201. [CrossRef]
24. Taylor, K.E. Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res. Atmos.* **2001**, *106*, 7183–7192.
25. Chen, M.; Feixas, M.; Viola, I.; Bardera, A.; Shen, H.W.; Sbert, M. *Information Theory Tools for Visualization*; CRC Press: Boca Raton, FL, USA, 2016.
26. Kramer, G. *Directed Information for Channels with Feedback*; Citeseer: Princeton, NJ, USA, 1998.
27. Vicente, R.; Wibral, M.; Lindner, M.; Pipa, G. Transfer entropy—A model-free measure of effective connectivity for the neurosciences. *J. Comput. Neurosci.* **2011**, *30*, 45–67.
28. Bossomaier, T.; Barnett, L.; Harré, M.; Lizier, J.T.; Bossomaier, T.; Barnett, L.; Harré, M.; Lizier, J.T. *Transfer Entropy*; Springer: Berlin/Heidelberg, Germany, 2016.
29. Pincus, S.M. Approximate entropy as a measure of system complexity. *Proc. Natl. Acad. Sci. USA* **1991**, *88*, 2297–2301. [CrossRef]
30. Richman, J.S.; Moorman, J.R. Physiological time-series analysis using approximate entropy and sample entropy. *Am. J. Physiol.-Heart Circ. Physiol.* **2000**, *278*, H2039–H2049. [CrossRef]
31. Bandt, C.; Pompe, B. Permutation entropy: A natural complexity measure for time series. *Phys. Rev. Lett.* **2002**, *88*, 174102. [CrossRef]
32. Rostaghi, M.; Azami, H. Dispersion entropy: A measure for time-series analysis. *IEEE Signal Process. Lett.* **2016**, *23*, 610–614. [CrossRef]
33. Massey, J.L.; Massey, P.C. Conservation of mutual and directed information. In Proceedings of the Proceedings. International Symposium on Information Theory, 2005. ISIT 2005, Adelaide, SA, Australia, 4–9 September 2005; pp. 157–158.
34. Amblard, P.O.; Michel, O.J. On directed information theory and Granger causality graphs. *J. Comput. Neurosci.* **2011**, *30*, 7–16. [CrossRef]
35. Quinn, C.J.; Coleman, T.P.; Kiyavash, N.; Hatsopoulos, N.G. Estimating the directed information to infer causal relationships in ensemble neural spike train recordings. *J. Comput. Neurosci.* **2011**, *30*, 17–44. [CrossRef] [PubMed]
36. Jiao, J.; Permuter, H.H.; Zhao, L.; Kim, Y.H.; Weissman, T. Universal estimation of directed information. *IEEE Trans. Inf. Theory* **2013**, *59*, 6220–6242. [CrossRef]
37. Luxembourg, O.; Tsur, D.; Permuter, H. TREET: TRansfer Entropy Estimation via Transformer. *arXiv* **2024**, arXiv:2402.06919.
38. Cai, T.T.; Liang, T.; Zhou, H.H. Law of log determinant of sample covariance matrix and optimal estimation of differential entropy for high-dimensional Gaussian distributions. *J. Multivar. Anal.* **2015**, *137*, 161–172. [CrossRef]
39. Duong, B.; Nguyen, T. Diffeomorphic information neural estimation. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023 ; Volume 37, pp. 7468–7475.
40. Knief, U.; Forstmeier, W. Violating the normality assumption may be the lesser of two evils. *Behav. Res. Methods* **2021**, *53*, 2576–2590. [CrossRef]

41. Belghazi, M.I.; Baratin, A.; Rajeshwar, S.; Ozair, S.; Bengio, Y.; Courville, A.; Hjelm, D. Mutual information neural estimation. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 531–540.
42. Sreekumar, S.; Goldfeld, Z. Neural estimation of statistical divergences. *J. Mach. Learn. Res.* **2022**, *23*, 1–75.
43. Tsur, D.; Aharoni, Z.; Goldfeld, Z.; Permuter, H. Neural estimation and optimization of directed information over continuous spaces. *IEEE Trans. Inf. Theory* **2023**, *69*, 4777–4798. [CrossRef]
44. Tsur, D.; Aharoni, Z.; Goldfeld, Z.; Permuter, H. Data-driven optimization of directed information over discrete alphabets. *IEEE Trans. Inf. Theory* **2023**, *70*, 1652–1670. [CrossRef]
45. Papamakarios, G.; Pavlakou, T.; Murray, I. Masked autoregressive flow for density estimation. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. Available online: https://proceedings.neurips.cc/paper_files/paper/2017/file/6c1da886822c67822bcf3679d04369fa-Paper.pdf (accessed on 20 March 2025).
46. Rosenblatt, M. Remarks on a multivariate transformation. *Ann. Math. Stat.* **1952**, *23*, 470–472.
47. Papamakarios, G.; Nalisnick, E.; Rezende, D.J.; Mohamed, S.; Lakshminarayanan, B. Normalizing flows for probabilistic modeling and inference. *J. Mach. Learn. Res.* **2021**, *22*, 2617–2680.
48. Butakov, I.; Tolmachev, A.; Malanchuk, S.; Neopryatnaya, A.; Frolov, A.; Andreev, K. Mutual Information Estimation via Normalizing Flows. *arXiv* **2024**, arXiv:2403.02187.
49. Bossomaier, T.R.J.; Barnett, L.C.; Harré, M.S.; Lizier, J.T. *An Introduction to Transfer Entropy*; Springer: Berlin/Heidelberg, Germany, 2016.
50. Rigney, D.; Goldberger, A.; Ocasio, W.; Ichimaru, Y.; Moody, G.; Mark, R. Multi-channel physiological data: Description and analysis. In *Time Series Prediction: Forecasting the Future and Understanding the Past*; Weigend, A., Gershenfeld, N., Eds.; Addison-Wesley: Reading, MA, USA, 1993; pp. 105–129.
51. Goldberger, A.; Amaral, L.; Glass, L.; Hausdorff, J.; Ivanov, P.; Mark, R.; Stanley, H. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* **2000**, *101*, e215–e220. [PubMed]
52. Ising, E. Beitrag zur Theorie des Ferro-und Paramagnetismus. Ph.D. Thesis, Grefe & Tiedemann, Hamburg, Germany, 1924.
53. Elishco, O.; Permuter, H. Capacity and coding for the Ising channel with feedback. *IEEE Trans. Inf. Theory* **2014**, *60*, 5138–5149.
54. Tallini, L.G.; Al-Bassam, S.; Bose, B. On the capacity and codes for the Z-channel. In Proceedings of the IEEE International Symposium on Information Theory, Lausanne, Switzerland, 30 June–5 July 2002; p. 422.
55. Vaswani, A. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. Available online: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (accessed on 20 March 2025).
56. Tsur, D.; Goldfeld, Z.; Greenewald, K. Max-sliced mutual information. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 80338–80351.
57. Tsur, D.; Permuter, H. InfoMat: A Tool for the Analysis and Visualization Sequential Information Transfer. *arXiv* **2024**, arXiv:2405.16463.
58. Goldfeld, Z.; Greenewald, K.; Nuradha, T.; Reeves, G. *k*-Sliced Mutual Information: A Quantitative Study of Scalability with Dimension. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 15982–15995.
59. Antos, A.; Kontoyiannis, I. Convergence properties of functional estimates for discrete distributions. *Random Struct. Algorithms* **2001**, *19*, 163–193.
60. Basharin, G.P. On a statistical estimate for the entropy of a sequence of independent random variables. *Theory Probab. Its Appl.* **1959**, *4*, 333–336.