A SURVEY OF MULTIVARIATE MUTUAL INFORMATION

Ying-xin Wu
School of Data Science
University of Science and Technology of China
wuyxin@mail.ustc.edu.cn

ABSTRACT

Multivariate mutual information (MMI) provides a theoretic framework for characterizing higher-order interactions of variables in complex systems. In this work, we elaborately summarize three classes of MMI measures from various aspects, *i.e.,* entropic, partial information decomposition and higher dimensional topology. We further investigate the generalization of the aforementioned information measures, following the work of *Rahimzamani et al.*. Specifically, Graph Divergence Measure (GDM) is formulated for i) defining a general family of MMI measures through distinct graphical structures ii) being consistent in the general data manifold of the random variables when estimating their co-information. To approximate the GDM value, *Rahimzamani et al.*proposed an estimator via a coupling trick that directly estimates the multivariate information measures using the Radon-Nikodym derivative. Experimentally, we show that the estimator well generalizes the MMI measures and accurately approximates the theory values under different settings. In the light of this work, a family of MMI measures through the general graphical formulation can be explored in the future, for discovering multifaceted interaction in information theory.

1 INTRODUCTION

Shannon's Mutual information (MI) and its generalizations play a vital role in machine learning, statistical estimation and inference in a variety of scenarios. To quantify the dependency of any pair of random variables X and Y, a simple measure is Pearson's correlation coefficient, which can only reveal linear dependencies and is not well defined for zero-valued signals (Ball et al., 2017). In contrast, MI is an information theoretic synchrony (*i.e.,* synergy) measure, which depends on the joint probability distribution of X and Y (could be multi-dimensional variables). It has beautiful mathematical properties (*e.g.,* non-negativity, symmetry) and equals zero if and only if X and Y are statistically independent (Kieffer, 1994).

In a more complex system, it is important to consider the high-order interdependencies through multivariate mutual information (MMI) of different variables. Studies on brain coding, for example, have found that neurons can convey redundant, complementary, or synergistic information — the latter referring to neurons that are uninformative individually but informative when analyzed collectively (Elad Schneidman & II, 2003). However, challenges come when a third variable Z engages in — the analogous properties can not generalise beyond two variables since the multivariate mutual information between three random variables is not non-negative (Floyd, 1964). Moreover, the computational cost of MMI is often intractable for small applications (Reing et al., 2020).

Large efforts have been put in uncovering the high-order synergistic relationships of variables. A promising approach, called the Partial Information Decomposition (PID) (Williams & Beer, 2010; Wibral et al., 2015), considers the mutual information within a privileged target variable T and the remaining predictor variables $S = \{S_1, S_2, \ldots, S_n\}$, and then investigates how the information carried by T is distributed over S. The foundations of PID involve finding the redundant information and unique information of predictor variables. The measure of PID is never negative and always support a clear interpretation for the system (Williams & Beer, 2010). However, PID is limited in practice due to the super-exponential increase of terms for large systems (Rosas et al., 2019). A class of pairwise measures of temporal synchrony, *e.g.,* transfer entropy (Schreiber, 2000), has also been widely studied. Regardless of the hierarchical scenario of PID, multiple mutual information (Multi-

pleMI)[1] (McGill, 1954; Han, 1980), is defined as the alternating sum/difference of permutations of the random variables' joint entropies, which may be negative. For another line, Watanabe (1960b); Timme et al. (2014) extended the pairwise MI into a scalar measure of shared information content named total correlation (TC), with guaranteed positivity. Following TC, dual total correlation (DTC) (Watanabe, 1960a) also provides a suitable multivariate information measure.

Although these metrics bring great success in the literature, some researchers have raised questions in terms of their interpretability in large scale (Rosas et al., 2019), sensibility (Ball et al., 2017; Chan et al., 2018), redundancy measure (Ince, 2017; Gutknecht et al., 2020; Reing et al., 2020) and data manifold (Rahimzamani et al., 2018). Very recently, Baudot (2021) generalizes the minima of k multivariate interaction information with k Brunnian links, thus accounting for group interactions beyond pair interaction to catch the essence of many multi-agent interactions. These methods have greatly promoted the prosperity of applications built upon MMI, including feature selection (Doquire & Verleysen, 2012), machine learning (Doquire & Verleysen, 2013; Dilpazir et al., 2015; Chen et al., 2019), interaction/relation discovery (Ding et al., 2016; Pham et al., 2012; Chan et al., 2016), security (Gierlichs et al., 2009) and explainability (Gao et al., 2019; Taverniers et al., 2020; Jung & Nardelli, 2020; Hsieh et al., 2021).

The rest of this paper is arranged as follows: In Section 2, we revisit some basic concepts used in the MMI field. In Section 3, we summaries some prevailing MMI metrics from three aspects: Entropic, partial information decomposition and topology (*e.g.,* Brunnian links). In Section 4, we aim to unify the aforementioned MMI metrics, following the work of Rahimzamani et al. (2018). In that work, they developed a general graph divergence measure (GDM) for any directed acyclic graph, *i.e.,* Bayesian Network (Pearl, 1989), thus generalizing the multivariate information given the graphical structure of the related variables. In Section 5, we validate the effectiveness of the proposed estimator, which offers great interpretability into the real-world applications with complex set of variables.

In a nutshell, the contributions of this work are

- We extensively summarize the recent researches of MMI from three aspects, which is the first work to introduce MMI via the proposed taxonomy, to the best of the authors' knowledge.
- Plus, we follow the previous works which generalize the MMI measures to the general GDM. Specifically, we prove that partial information decomposition can be well generalized to the GDM category, uncovering a distinct perspective of redundant information through graph structures.
- Finally, we empirically validate the proposed framework through three carefully designed numerical experiments. Also, we propose future directions and offer insights into the MMI measure for characterizing higher-order interactions of variables in complex systems.

2 Information Definitions

Entropy. We consider discrete variables here. The multivariate entropy (*aka.* joint-entropy) (Shannon, 1948) for k random variables $(X_1, \ldots, X_k) \sim \mathbb{P}_{(X_1,\ldots,X_k)}$ is defined as

$$H(X_1; \ldots; X_n) = -\sum_{x_i \in A_i}^{\prod_1^n |A_i|} p(x_1, \ldots, x_n) \ln p(x_1, \ldots, x_n) \tag{1}$$

where each A_i corresponds to the alphabet of X_i. For infinite discrete variables whose domain is represented as Ω, we sample a finite alphabet from Ω, *i.e.,* $|A_i| \leq \Omega$. Clearly, $H(\cdot)$ is determinant given (i) the sample space of Ω (ii) the probability distribution $\mathbb{P}$ (Baudot, 2021).

Similary, Conditional Entropy is defined as the same form of Equation 1 with the traversed variables the marginal random variables of cardinality $(n-k)$, while conditioning on the other k observed variables.

Mutual Information (MI). For two variables X and Y, we have their mutual information as

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \geq 0 \tag{2}$$

[1] To distinguish it from multivariate mutual information.

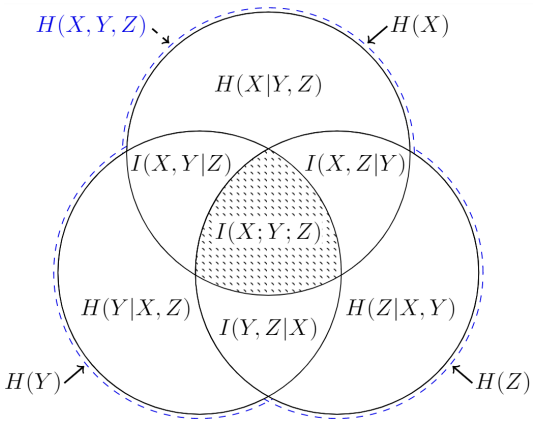


Figure 1: Mutual information of three variables. Figure from Finn & Lizier (2020).

which can be derived from the inequality $H(X)+H(Y) \geq H(X,Y) \geq H(X), H(Y) \geq 0$ and the equality holds only when X and Y are independent. It is worth mentioning that when it goes to a similar form of three variables,

$$I(X;Y;Z) = H(X) + H(Y) + H(Z) - H(X,Z) - H(X,Y) - H(Y,Z) + H(X,Y,Z), \quad (3)$$

the non-negativity no longer holds, *e.g.,* $X \sim Bernoulli(1/2)$, $Y \sim Bernoulli(1/2)$, $Z = \mathrm{XOR}(X,Y)$.

Kullback-Liebler divergence. Kullback-Liebler divergence (MacKay, 2003) is a measure of how distribution P is different from the reference probability distribution Q on the probability space Ω.

$$D_{\Omega}\left(P||Q\right) = \sum_{x} p(x) \ln \frac{q(x)}{p(x)} \quad (4)$$

3 High-order Synergistic Metrics

In this section, we introduce three classes of works from a taxonomic perspective: Works that propose the entropic multivariate mutual information metrics; works that target the synergistic relations by partial information decomposition; works that develop the multivariate information problem based on the field of algebraic topology.

3.1 Entropic Measure

A measure is entropic if it is defined as a function of subset entropies $H\left(X_{sub}\right)$ for any $X_{sub} \subseteq \{X_1, \ldots, X_n\}$. We summarize some prevailing metrics as follows.

Multiple Mutual Information. Multiple mutual information (McGill, 1954) (MultipleMI) is an extension of Equation 3 for more variables, *i.e.,*

$$\mathrm{MultipleMI}\left(X_1;\ldots;X_n\right) = \sum_{k=1}^{n}(-1)^{k+1} \sum_{\substack{M \subseteq \{1,\ldots,n\} \\ |M|=k}} H\left(X_{M_1}, X_{M_2}, \ldots, X_{M_k}\right) \quad (5)$$

The defects of MultipleMI are (i) negativity, which can be proved by induction, (2) misleading people by depicting the Multivariate Mutual Information (MMI) of three or more variables using a Venn diagram, which does not have any substantive meaning (Finn & Lizier, 2020). Specifically, the multivariate mutual information $I(X;Y;Z)$ cannot be accurately represented using an area, as represented by the shadow in Figure 1.

Total Correlation Watanabe (1960b) defined the Total Correlation (TC) as

$$\mathrm{TC}\left(X_1;\ldots;X_n\right)=H\left(X_1\right)+H\left(X_2\right)+\cdots+H\left(X_n\right)-H\left(X_1,X_2,\ldots,X_n\right). \tag{6}$$

The metrics' assured positivity (with a zero value only in the situation of perfect independence among variables) makes it a more appealing quantity over MultipleMI (McGill, 1954; Han, 1980).

Dual Total Correlation Originating from the lattice theoretic duality of information measures by Han (1975), Dual total correlation can be defined as

$$\begin{aligned}\mathrm{DTC}(X)&=H(X)-\sum_{i=1}^{n}H\left(X_i\mid X_{/i}\right)\\&=\sum_{i=1}^{n}H\left(X_{/i}\right)-(n-1)H(X),\end{aligned} \tag{7}$$

where $X_{/i}$ denotes $\{X_k|k\in\{1,\ldots,i-1,i+1,\ldots,n\}\}$, *i.e.,* the marginal random variable of cardinality $(n-1)$ that excludes X_i (Reing et al., 2020). Some desirable properties (Han, 1978; 1975; Reing et al., 2020) of both total correlation and its dual can be summarized as

- ***Non-negative***: A fine metrics of distance which is always ≥ 0.
- ***Symmetric***: A measure is symmetric if remains invariant under any permutations of random variables.
- ***Correlative***: A measure is correlative if it equals to zeros if $X_1,\ldots,X_n$ are mutually independent.
- ***Entropic***: A measure is entropic if it is defined as a function of subset entropies $H\left(X_{sub}\right)$ for any $X_{sub}\subseteq\{X_1,\ldots,X_n\}$.

Normalized Mutual Information The normalized multi-information (Ball et al., 2017) (NMI) is motivated by the Theorem 1, the proof of which is a straightforward calculation of Equation 5 and Equation 6. What implies by the theorem is that TC is unable to convergence when the number of variables (nodes) goes to infinite.

Theorem 1. *The MultipleMI of the same random variable X returns the self information $H(X)$ of that random variable (just as pairwise mutual information of two copies of X does). However, TC returns $(n-1)H(X)$ when evaluated on a variable set consisting of n copies of X :*

$$\begin{aligned}\mathrm{MultipleMI}(\underbrace{X;\ldots;X}_{n})&=H(X)\\\mathrm{TC}(\underbrace{X;\ldots;X}_{n})&=(n-1)H(X).\end{aligned}$$

NMI is thus to addresses the divergence issue in the MMI measured by TC. Consequently, the NMI acting on n copies of X is simply the entropy of X. Note that "normalization" in NMI refers to rescaling to account for inflation that would otherwise occur with increasing variable size (Ball et al., 2017).

$$\begin{aligned}\mathrm{NMI}\left(X_1;\ldots;X_n\right)&=\frac{1}{n-1}\mathrm{TC}\left(X_1;\ldots;X_n\right)\\&=\frac{1}{n-1}\left[\sum_{i=1}^{n}H\left(X_i\right)-H\left(X_1,\ldots,X_n\right)\right]\end{aligned}$$

Theorem 2. *Given two (non-empty) sets of random variables $X_1,\ldots,X_m$ and $Y_1,\ldots,Y_n$ with NMI values:*

$$\begin{aligned}\mathrm{NMI}\left(X_1;\ldots;X_m\right)&=I_X\\\mathrm{NMI}\left(Y_1;\ldots;Y_n\right)&=I_Y\end{aligned}$$

then the NMI between all the random variables satisfies the relation:

$$\begin{aligned}&\frac{m-1}{m+n-1}I_X+\frac{n-1}{m+n-1}I_Y\leq\mathrm{NMI}\left(X_1;\ldots;X_m;Y_1;\ldots;Y_n\right)\leq\\&\begin{cases}\frac{1}{m+n-1}\left((m-1)I_X+(n-1)I_Y+H\left(X_1,\ldots,X_m\right)\right) & H\left(Y_1,\ldots,Y_n\right)\geq H\left(X_1,\ldots,X_m\right)\\\frac{1}{m+n-1}\left((m-1)I_X+(n-1)I_Y+H\left(Y_1,\ldots,Y_n\right)\right) & H\left(X_1,\ldots,X_m\right)\geq H\left(Y_1,\ldots,Y_n\right)\end{cases}\end{aligned} \tag{8}$$

Proof. For lower bound, we have

$$\begin{aligned}
&\mathrm{NMI}\left(X_1;\ldots;X_m;Y_1;\ldots;Y_n\right)\\
&=\frac{1}{m+n-1}\left(\sum_{i=1}^{m}H\left(X_i\right)+\sum_{i=1}^{n}H\left(Y_i\right)-H\left(X_1,\ldots,X_m,Y_1,\ldots,Y_n\right)\right)\\
&\geq\frac{1}{m+n-1}\left(\sum_{i=1}^{m}H\left(X_i\right)+\sum_{i=1}^{n}H\left(Y_i\right)-H\left(X_1,\ldots,X_m\right)-H\left(Y_1,\ldots,Y_n\right)\right)\\
&=\frac{1}{m+n-1}\left((m-1)I_X+(n-1)I_Y\right)
\end{aligned}\tag{9}$$

where the inequality is hold because the joint entropy is less than the sum of the individual entropies. And for upper bound, due to the monotony of entropy, *i.e.,* $H\left(X,Y\right)\geq\max(H\left(X\right),H\left(Y\right))$, where $H\left(X\right)$ is short for $H(X_1,\ldots,X_m,Y_1,\ldots,Y_m)$. Thus,

$$\mathrm{NMI}\left(X_1;\ldots;X_m;Y_1;\ldots;Y_n\right)\leq\frac{1}{m+n-1}\left(\sum_{i=1}^{m}H\left(X_i\right)+\sum_{i=1}^{n}H\left(Y_i\right)-\max(H(X),H(Y))\right)\tag{10}$$

The first equality holds when the two set of variables are independent, while the second equality holds when $H(X|Y)$ (or $H(Y|X)$ if $H(X)>H(Y)$) is zero. □

In the context of neural network activity, the lower boundary property states that the combined NMI of two multi-node networks is always larger than or equal to the weighted sum of the networks' individual NMIs. And the upper bound says that the NMI of two networks is always less than or equal to the lower bound of NMI plus the totality-normalized self-information. This theorem well validate the robustness of NMI with respect to changes in network size (Ball et al., 2017).

Multivariate Mutual Information. Short for MultivarMI, Chan et al. (2015b) defined it as follows: let $\Lambda(\mathcal{X})$ be the collection of all possible partitions of $\mathcal{P}$ which split $\mathcal{X}$ into at least two non-empty disjoint subsets. For any partition $\mathcal{P}\in\Lambda(\mathcal{X})$, the product distribution $\Pi_{C\in\mathcal{P}}\mathbb{P}_{X_C}$ specifies an independence relation, *i.e.,* , X_C 's are treated as agglomerated random variables and are mutually independent (Rahimzamani et al., 2018). Given a particular partition, an information measure $I_{\mathcal{P}}(X)$ is defined as:

$$I_{\mathcal{P}}(X)=\frac{1}{\mathcal{P}-1}D\left(\mathbb{P}_X\|\Pi_{C\in\mathcal{P}}\mathbb{P}_{X_C}\right)\tag{11}$$

which can also be explicitly expressed as $I_{\mathcal{P}}\left(X\right)=I\left(X_{C_1};\ldots;X_{C_k}\right)$, for $\mathcal{P}=\{C_1,\ldots,C_k\}$. Then, MultivarMI is defined as:

$$\mathrm{MultivarMI}(X)=\min_{\mathcal{P}\in\Pi(\mathcal{X})}I_{\mathcal{P}}(X)\tag{12}$$

It is clear that Shannon's mutual information $I\left(X_1;X_2\right)=\mathrm{MultivarMI}(X=(X_1,X_2))$ is the special case when $\mathcal{P}$ is a bipartition. This definition well considers the group synergistic property, while the time complexity largely depends on the size of $\Lambda(\mathcal{X})$.

3.2 Partial Information Decomposition

As shown in Figure 2 (a), the goal of partial information decomposition (Williams & Beer, 2010; Wibral et al., 2015) (PID) is to break down the multivariate mutual information contained in a set of source variables $S_1,\ldots,S_n$ about a target variable T into basic bits, or "atoms of information". Each atom describes a unique way in which the sources might hold information about the target. Next we introduce the core of PID, following the analytical paradigm in Gutknecht et al. (2020).

Q1: What do the atoms of information mean?

For brevity, we refer to source variables and collections by their indices (*e.g.,* $\{1\}$ and $\{1,2\}$ are corresponding to $\{S_1\}$ and $\{S_1,S_2\}$ respectively). As we mentioned, multiple information sources combined can provide some information that is not contained in any individual source. Such synergistic information has motivated us to think the information atoms from different collections of

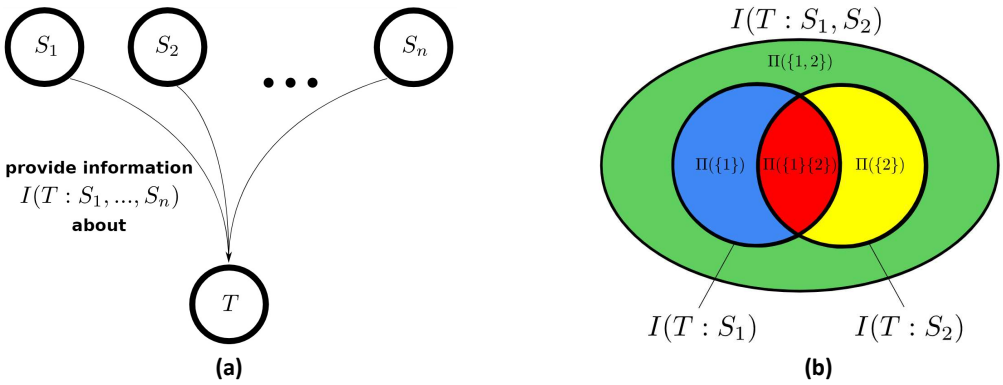


Figure 2: (a) The problem setting of PID, where T is a target variable and $S_1, \ldots, S_n$ are the remaining predictor variables. (b) Information diagram for the case of two information sources. $I(T : S_1)$ and $I(T : S_2)$ represent the mutual information provided by the two sources about the target, respectively. $I(T : S_1, S_2)$ is the joint mutual information of S_1 and S_2 *w.r.t.* the target. Information atoms are differ in colors. Figures from Gutknecht et al. (2020).

Table 1: Parthood distribution of $\Pi(\{1\}, \{2\})$ and $\Pi(\{1, 2\})$.

	$\{\}$	$\{1\}$	$\{2\}$	$\{1, 2\}$
$\Pi(\{1\}\{2\})$	0	1	1	1
$\Pi(\{1, 2\})$	0	0	0	1

source variables. We denote a collection as C_i ($1 \leq i \leq 2^n$), which is a subset of the source variables. In order to characterize an information atom Π, we have to ask for each collection C_i: Is Π part of the information provided by C_i?

To answer it, we start by inspecting these collections and their atoms. Intuitively, three requisites should be satisfied: (i) No information atom is contained in an empty collection. (ii) All information atoms are contained in the collection with full set of source variables. (iii) If an information atom exists in the collection C_i, it won't vanish in any C_j *s.t.* $C_i \subseteq C_j$. More formally, we have the following definition (Gutknecht et al., 2020)

Definition 1. ***A parthood distribution*** *is any function* $f : \{1, \ldots, n\} \to \{0, 1\}$ *such that*
1. $f(\{\}) = 0$ *("There is no information in the empty set")*
2. $f(\{1, \ldots, n\}) = 1$ *("All information is in the full set")*
3. (Monotonicity) For any two collections of source indices C_i, C_j *: If* $C_j \supseteq C_i$*, then* $f(C_i) = 1 \Rightarrow f(C_j) = 1$.

The parthood distribution is exactly the characterization of each atom of information, describing whether or not the atom is part of the information provided by the different possible collections of sources. More formally, the atom $\Pi(f)$ with parthood distribution f is part of the information provided by all collections of sources C_i for which $f(C_i) = 1$, and is not part of the information provided by collections for which $f(C_i) = 0$. In Figure 2 (b), for example, the unique information in S_1 *w.r.t.* T is $\Pi(\{1\})$ (similar for S_2), $\Pi(\{1\}\{2\})$ is the shared information and $\Pi(\{1, 2\})$ is the synergistic information. Note that $\Pi(\{1\}\{2\})$ is different from $\Pi(\{1, 2\})$, their corresponding parthood distribution is shown in Table 1.

Q2: How many atoms of information are there?

As information atom and parthood distribution satisfiy the bijective relation, the answer should be straightforward, *i.e.,* the number of the valid parthood distributions. The most restrictive constraint is the Monotonicity condition, which is related to the Dedekind numbers, a sequence of the numbers of monotonic Boolean functions of n-bits. To take the 1) and 2) conditions in the definition of the parthood distribution into consideration, it is clear that the number of valid parthood distributions equals to the Dedekind numbers minus 2 (exclude $(0, \ldots, 0)$ and $(1, \ldots, 1)$). It turns out that the number of the parthood distributions is a super-exponentially growing sequence of numbers (equals to 7828352 when n is 6), which has greatly limited PID in practice.

Q3: How many bits of information does each atom provide?

Take the example in Figure 2 (b), we have the following equations:

$$\begin{aligned}\Pi(\{1\}\{2\}) + \Pi(\{1\}) + \Pi(\{2\}) + \Pi(\{1,2\}) &= I(T:S_1,S_2)\\ \Pi(\{1\}\{2\}) + \Pi(\{1\}) &= I(T:S_1)\\ \Pi(\{1\}\{2\}) + \Pi(\{2\}) &= I(T:S_2)\end{aligned} \tag{13}$$

which can be generalized as

$$I(T:C) = \sum_{f(C)=1} \Pi(f) \tag{14}$$

As we have four unknowns but only three equations, next is to find a way to come up with the appropriate number of additional equations. Williams & Beer (2010) utilized the concept of redundant information to introduce additional constraints. Specifically, they defined $I_\cap(T:C_1,\ldots,C_m)$ as the information shared by those collection, which consists of all information atoms that are part of the information provided by each C_i $(1 \le i \le m)$, *i.e.,*

$$I_\cap(T:C_1,\ldots,C_m) = \sum_{f(C_i)=1,\forall 1\le i\le m} \Pi(f) \tag{15}$$

The redundant information can be depicted into a lattice structure (we omit this part for brevity), based on which we conclude the the redundant information associated with a parthood distribution f can always be expressed as

$$I_\cap(T:f) = \sum_{g\sqsubseteq f} \Pi(g) \tag{16}$$

where $g \sqsubseteq f \Leftrightarrow (g(C)=1 \Rightarrow f(C)=1, \forall C \subseteq \{1,\ldots,n\})$. Now that we have one equation corresponds to one parthood distribution f, we now obtain as many equations as unknowns. So far, the problem of determining the sizes of the information atoms has been shifted to the problem of coming up with a feasible definition of redundant information $I_\cap(T:f)$. By "feasible", we indicated that $I_\cap(T:f)$ should satisfied the following properties

- ***Symmetry***: $I_\cap\left(T:C_1,\ldots,C_m\right) = I_\cap\left(T:C_{\sigma(1)},\ldots,C_{\sigma(m)}\right)$ for any permutation σ.
- ***Idempotency***: If $C_i = C_j$ for $i \ne j$, then $I_\cap\left(T:C_1,\ldots,C_m\right) = I_\cap\left(T:C_1,\ldots,C_{i-1},C_{i+1},\ldots,C_m\right)$.
- ***Invariance under superset removal / addition***: If $C_i \supset C_j$ for $i \ne j$, then $I_\cap\left(T:C_1,\ldots,C_m\right) = I_\cap\left(T:C_1,\ldots,C_{i-1},C_{i+1},\ldots,C_m\right)$.
- ***Self-redundancy***: $I_\cap(T:C) = I(T:C)$

Other researcher (Griffith et al., 2014; Bertschinger et al., 2012) have also proposed further properties and axioms for measures of redundancy.

Q4: How to define the redundant information function?

There are various definitions of the redundant information function the literature. Williams & Beer (2010) denoted the redundancy as I_{min} and derived as the average minimum specific information over the sources variable.

$$I_{\min}\left(T;C_1,\ldots,C_m\right) = \sum_t p(t) \min_{C_i} I\left(T=t;C_i\right) \tag{17}$$

where $I(T=t;C) = \sum_c p(c \mid t)\left[\log_2 \frac{1}{p(t)} - \log_2 \frac{1}{p(t|c)}\right]$. $I_{\min}$ is non-negative and satisfies the feasible properties. However, as $I_{\min}$ could be positive even if source variables are independent (there should be no overlap), $I_{\min}$ can overestimate the redundancy value (Ince, 2017). For other methods, both Griffith & Koch (2012) and Bertschinger et al. (2014) obtained the redundancy as the maximum of the co-information over all distributions that preserve the source-target marginals. And Gutknecht et al. (2020) performed the entire partial information decomposition on the pointwise level and then defining pointwise redundancy in terms of logical statements. Also, such definitions could vary in different contexts, *e.g.,* neuroscience, electronics (Wibral et al., 2015; Ince, 2017).

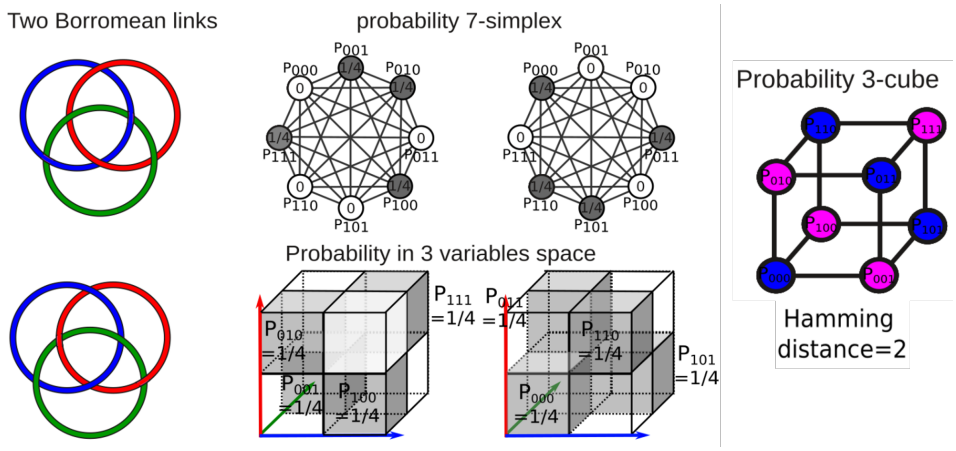


Figure 3: Brunnian 3-links of information. Left shows the two Brunnian links (mirror images). Middle is the two probability laws in the probability 7-simplex for 3 binary random variables, and below the representation of the corresponding configuration in 3 dimensional data space. Right is the graph covering in the 3-cube of the two configurations. Figure from Baudot et al. (2019).

3.3 Information Links

To account for group interactions beyond pair interaction and catch the essence of many multi-agent interactions, researchers have introduced Brunnian links to investigate the principles of more complex structures. A Brunnian link is a nontrivial link that degrades to a set of trivial unlinked circles when any one component is removed (See left of Figure 3 or Figure 4 for better perception). They have drawn great attention in the filed of topology for their beauty and simplicity, and give a clear demonstration of what is an entirely collective characteristic. Specifically, Baudot et al. (2019) suggested that the minima of the 3-way multivariate mutual information corresponds to Brunnian links, where they defined J_k, the k-mutual Information (*aka.* the k-interaction information).

$$J_k = J\left(X_1;\ldots;X_k\right) = \sum_{i=1}^{k}(-1)^{k-i}\sum_{\substack{M\subseteq\{1,\ldots,k\}\\|M|=i}} H\left(X_{M_1},X_{M_2},\ldots,X_{M_i}\right) \tag{18}$$

Note that it is slightly different from MultipleMI. And one interesting property is that, their direct sums give the multivariate entropy, *e.g.,* $H\left(X_1,X_2,X_3\right) = J\left(X_1\right)+J\left(X_2\right)+J\left(X_3\right)+J\left(X_1;X_2\right)+J\left(X_1;X_3\right)+J\left(X_2;X_3\right)+J\left(X_1;X_2;X_3\right)$. Further, Baudot (2021) proposed Theorem 3[2].

Theorem 3. ***k-links of information***. *The absolute minimum of J_k, equal to -1, is attained only in the two cases j-independent j-uplets of unbiased variables for all $1<j<k$ with atomic probabilities $p\left(x_1,\ldots,x_k\right)=1/2^{k-1}$ or $p\left(x_1,\ldots,x_k\right)=0$ such that no associated vertex of the associated k -hypercube covering graph of $p\left(x_1,\ldots,x_k\right)=1/2^{k-1}$ connects a vertex of $p\left(x_1,\ldots,x_k\right)=0$ and conversely. These cases correspond to the two k - Brunnian links, the right one and the left one.*

Thus they generalizes the minima of k multivariate interaction information with k Brunnian links in the binary variable case. Let us first examine its correctness in the low dimension. In Figure 3, when $k=3$, Theorem 3 states that $p_{000}=1/4, p_{001}=0, p_{010}=0, p_{011}=1/4, p_{100}=0, p_{101}=1/4, p_{110}=1/4, p_{111}=0$, or $p_{000}=0, p_{001}=1/4, p_{010}=1/4, p_{011}=0, p_{100}=1/4, p_{101}=0, p_{110}=0, p_{111}=1/4$ are the only two cases to achieve the global minima of J_3.

Proof.

$$p_{000}=a, p_{010}=b, p_{100}=c, p_{110}=d,$$
$$p_{001}=e, p_{011}=f, p_{101}=g, p_{111}=h$$

[2]The proof of it relies on a weak concavity theorem D (Baudot & Bennequin, 2015) which has not been provided yet.

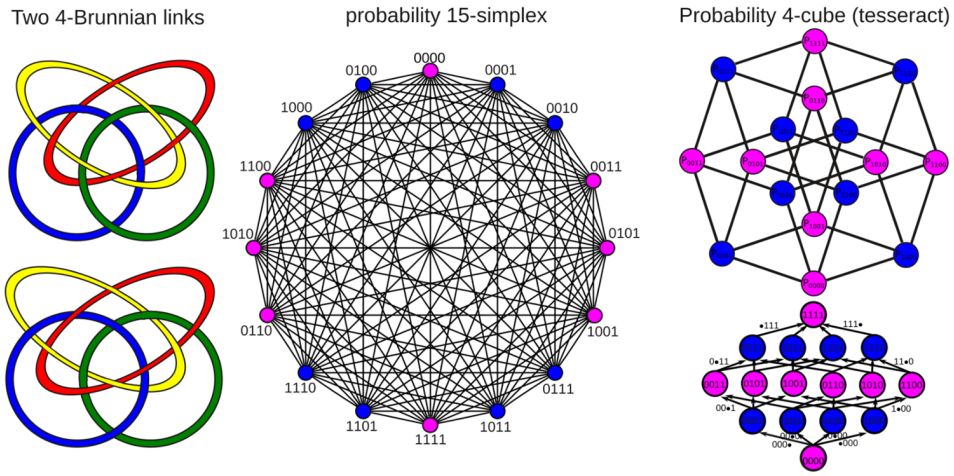


Figure 4: Brunnian 4-links of information. Left shows the two Brunnian links (mirror images). Middle is the two probability laws in the probability 15-simplex for 4 binary random variables. Right is the graph covering in the 4-cube of the two configurations, called the tesserac. Figure from Baudot (2021).

Now we consider all the constraint imposed by independence. The maximum entropy (or 1-independence) of single variable gives $C_4^2 = 6$ equations:

$$\begin{aligned} &a+b+e+f=1/2; \quad a+c+e+g=1/2; \quad a+d+e+h=1/2 \\ &b+c+f+g=1/2; \quad b+d+f+h=1/2; \quad c+d+g+h=1/2 \end{aligned} \tag{19}$$

The 2-independence of (pair of) variables gives the $C_4^1 = 4$ equations:

$$a+e=1/4; \quad b+f=1/4; \quad c+g=1/4; \quad d+h=1/4 \tag{20}$$

According to Theorem D (Baudot & Bennequin, 2015), J_3 is super-harmonic (weakly concave), thus the minima of J should happen on the boundary of the probability simplex, resultng in one additional constraint that for example a is either 0 or $1/4$. Analogously, when $k = 4$, we can prove that the minima of J_4 is attained when $p_{0000} = 0, p_{0001} = 1/8, p_{0010} = 1/8, p_{0011} = 0, p_{0101} = 0, p_{1001} = 0, p_{0111} = 1/8, p_{1011} = 1/8, p_{1111} = 0, p_{1101} = 1/8, p_{1110} = 1/8, p_{0110} = 0, p_{1010} = 0, p_{1100} = 0, p_{1000} = 1/8, p_{0100} = 1/8$, **or** $p_{0000} = 1/8, p_{0001} = 0, p_{0010} = 0, p_{0011} = 1/8, p_{0101} = 1/8, p_{1001} = 1/8, p_{0111} = 0, p_{1011} = 0, p_{1111} = 1/8, p_{1101} = 0, p_{1110} = 0, p_{0110} = 1/8, p_{1010} = 1/8, p_{1100} = 1/8, p_{1000} = 0, p_{0100} = 0$. □

Finally, we rearrange the proof of Theorem 3 provided in Baudot (2021).

Proof. When k is odd, as the minima of J_k are the maxima of $\mathrm{MultipleMI}_k$ (we denote it as I_k for short). We have $J_k \geq -\min\left(H\left(X_1\right) \ldots, H\left(X_k\right)\right)$ and $I_k \leq \min\left(H\left(X_1\right), \ldots, H\left(X_k\right)\right)$. Also, we have

$$I_k = \sum_{i=1}^{k-1} (-1)^{i-1} \left(\begin{array}{c} i \\ k \end{array} \right) \cdot i + (-1)^{k-1} H\left(X_1, \ldots, X_k\right) \Rightarrow I_k = k - H\left(X_1, \ldots, X_k\right)$$

This is obtained by the independent relations of the variables and the fact that all $H(X_i)$ should be maximal. Further, as

$$\begin{aligned} &I\left(X_1; \ldots; X_k\right) \leq \min\left(H\left(X_1\right), \ldots, H\left(X_k\right)\right) \leq \max\left(H\left(X_i\right)\right) = 1 \\ &\Rightarrow H\left(X_1, \ldots, X_k\right) = k-1, I_k = 1, J_k = -1 \end{aligned}$$

It is a minima of J_k because it saturates the bound $J_k \geq -\min\left(H\left(X_i\right)\right) \geq -1$. Since J_k is weakly concave, we have one additional constraint that for example a is either 0 or $1/2^{k-1}$. It gives 2 systems of equations that are hence fully determined and we have 2 solutions. The proof is likewise when k is even. □

The intuition that we can sense here is how the Brunnian links construct the absolute collective relations (when any one variable (circle) is removed, the others no longer linked) of variables. What's more, as Brunnian links provides probabilistic analogs, it inspires us to depict the multivariate mutual information by the probabilistic simplex. However, generalizing such topology structure to complex systems is not easy, since it is hard to introduce noise variables in the current formulation.

3.4 Discussion

Besides measuring the information amount, the information theory broadly includes, *e.g.,* information transmission and information representation. In the rate-distortion problem, for example, the goal is to encode a source in such a way that the code length is minimized while the average distortion is kept constant. In cost-capacity problem, a cost is assigned to each symbol of the channel alphabet. The task is to minimize the ambiguity at the receiver under a constraint on the average cost.

The connections of these problems and MMI are natural when there are multiple source signals or when the independence assumptions do not hold. Take average information (Gilad-Bachrach et al., 2003) for instance, a $\left(2^{nR}, n\right)$ rate information code consists of: an encoding function,

$$f_n : \mathcal{X}^n \longrightarrow \left\{1, 2, \cdots, 2^{nR}\right\}$$

and a coding function,

$$g_n : \left\{1, 2, \cdots, 2^{nR}\right\} \longrightarrow \hat{\mathcal{X}}^n$$

Thus Y-information associated with the $\left(2^{nR}, n\right)$ is defined as

$$Y_{\text{info}}\left(f_n, g_n\right) = \frac{1}{n} \sum_{i=1}^{n} I\left(\left(g_n \circ f_n\left(X^n\right)\right)_i ; \left(Y^n\right)_i\right) \tag{21}$$

where the information of the i's element is calculated with respect to the distribution defined by the code for this coordinate as follows:

$$\bar{p}_i(x, \hat{x}) = \sum_{x^n : \left(x^n\right)_i = x \wedge \left(g_n \circ f_n\left(x^n\right)\right)_i = \hat{x}} p\left(x^n\right) \tag{22}$$

If there exists a sequence of rate information codes $\left(f_n, g_n\right)$ with asymptotic rate R and asymptotic Y-information larger than or equal to I_Y, i.e. with $\lim_{n\to\infty} Y_{\text{info}}\left(f_n, g_n\right) \geq I_Y$, we call $\left(R, I_Y\right)$ as the a achievable rate information pair. However, the independent generation of $\{\hat{\mathcal{X}}^n\}$ underlying the formulation may not hold true, where the MMI measure should be brought in for estimating a more compact higher bound of I_Y. Thus, we can rewrite Equation 21 as

$$Y_{\text{info}}\left(f_n, g_n\right) = \frac{1}{\omega} \text{MMI}\left(\{g_n \circ f_n\left(X^n\right)_i\}_{i=1}^n, \{\left(Y^n\right)_i\}_{i=1}^n\right) \tag{23}$$

where ω denotes the normalization constant. While such extension is less explored in the literature, it is vital for real applications like (Chan et al., 2015a) which motivates the multiterminal secret-key capacity.

4 General Co-Information Measure via Graphical Model

Through this part, we represent the random variables as $X := (X_1, \ldots, X_d)$, which is a d-dimensional vector respects to the distribution $\mathbb{P}_X$. The N observed samples drawn from the distribution are denoted by $x^{(1)}, x^{(2)}, \ldots, x^{(N)}$, *i.e.,* $x^{(i)} = \{x_1^{(i)}, \ldots, x_d^{(i)}\}$ is the i-th observed sample. When estimating the multivariate mutual information (*aka.* co-information) between variable, it is natural for one to associate this with a holistic graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ where the nodes denote variables and edges denote the interactions of variables. Two assumptions/reasons are:

- The (partial) graph structure of the direct interaction of variables tends to be explicit in practice. For example, ***wet*** is the common cause of ***sprinkler*** and ***rain***, thus we have ***sprinkler*** $\rightarrow$ ***wet*** $\leftarrow$ ***rain***, implying that (***sprinkler*** $\not\perp\!\!\!\perp$ ***rain*** $\mid$ ***wet***), which is consistent with the observed distribution.

- While the observed variables $\mathcal{N}$ is fixed in the formulation of the co-information measure, it is possible to generalize different co-information assumptions and definitions under distinct $\mathcal{E}$.

The second assumption acts as the main motivation of this section, under which the Graph Divergence Measure (GDM) is proposed in Section 4.2, following the work of Rahimzamani et al. (2018). In Section 4.3, several aforementioned multivariate measures are generalized as the special cases of GDM. And the detailed estimation algorithm is discussed in Section 4.4. In Section 4.1, we first briefly introduce the used graph concepts.

4.1 Graphical Model of Variables

Definition 2. ***Bayesian networks.*** *A directed acyclic graph (DAG) representing causal or temporal relationships, and the core is that, knowing the values of other preceding variables is redundant once we know the values pa_j of the parent set PA_j, which is also called Markovian Parents.*

Definition 3. ***Markovian Parents.*** *Let $V = \{X_1, \ldots, X_d\}$ be an ordered set of variables, and let $P(v)$ be the joint probability distribution on these variables. A set of variables PA_j is said to be Markovian parents of X_j if PA_j is a minimal set of predecessors of X_j that renders X_j independent of all its other predecessors. In other words, PA_j is any subset of $\{X_1, \ldots, X_{j-1}\}$ satisfying*

$$\mathbb{P}_X\left(x_j \mid pa_j\right) = \mathbb{P}_X\left(x_j \mid x_1, \ldots, x_{j-1}\right) \tag{24}$$

The way we construct a Bayesian network is by recursion with the assistance of some screening conditions. We therefore conclude that a necessary condition for a DAG $\mathcal{G}$ to be a Bayesian network of probability distribution $\mathbb{P}_X$ is for $\mathbb{P}_X$ to admit the product decomposition dictated by $\mathcal{G}$, as

$$\mathbb{P}_X\left(x_1, \ldots, x_n\right) = \prod_{l=1}^{d} \mathbb{P}_X\left(x_l \mid pa_l\right) \tag{25}$$

Definition 4. ***Markov Compatibility.*** *If a probability function $\mathbb{P}_X$ admits the factorization of Equation 25 relative to DGA $\mathcal{G}$, we say that $\mathcal{G}$ represents $\mathbb{P}_X$, that $\mathcal{G}$ and $\mathbb{P}_X$ are compatible, or that $\mathbb{P}_X$ is Markov relative to $\mathcal{G}$.*

It is a necessary and sufficient condition for a DAG $\mathcal{G}$ to explain a body of empirical data represented by $\mathbb{P}_X$, that is, to describe a stochastic process capable of generating $\mathbb{P}_X$ (Pearl, 2000). However, it could be hard to fulfill due to the high time complexity to search for a compatible $\mathcal{G}$. Thus, we define

$$\overline{\mathbb{P}}_X := \prod_{l=1}^{d} \mathbb{P}_{X_l|pa(X_l)} \tag{26}$$

as the natural measure given a graph $\mathcal{G}$ and any distribution $\mathbb{P}_X$, where $\mathbb{P}_{X_l|\mathrm{pa}(X_l)}$ is short for $\mathbb{P}_X(x_l \mid pa_l)$. Since $\mathbb{P}_{S|X\setminus S}$ is well-defined for any subset of variables $S \subset X$. Therefore if $S = X \setminus \mathrm{pa}\left(X_l\right)$, then $\mathbb{P}_{X\setminus \mathrm{pa}(X_l)|\mathrm{pa}(X_l)}$ is well-defined for any $l \in \{1, \ldots, d\}$. By marginalizing over $X\setminus\mathrm{pa} + (X_l) \equiv (X_l, \mathrm{pa}\left(X_l\right))$ we see that $\mathbb{P}_{X_l|\mathrm{pa}(X_l)}$ and thus the distribution $\overline{\mathbb{P}}_X$ is well-defined (Rahimzamani et al., 2018).

4.2 Graph Divergence Measure

A divergence or contrast function is a function that determines the "distance" between two probability distributions on a statistical manifold, *e.g.,* Exponential divergence and Kullback–Leibler divergence (MacKay, 2003). Specifically, we will focus only on the KL Divergence (*cf.* Equation 4) as being the distance metric, *i.e.,* $D(\cdot\|\cdot) = D_{KL}(\cdot\|\cdot)$.

We first consider the case that $\mathbb{P}_X$ is absolutely continuous with respect to $\overline{\mathbb{P}}_X$ and hence the Radon-Nikodym derivative[3] $d\mathbb{P}_X/d\overline{\mathbb{P}}_X$ exists. Thus, for the random variables X and a Bayesian Network

[3]Suppose we have a measurable space (X, Σ) on which two σ -finite measures, μ and ν, are defined. If $\nu \ll \mu$ (i.e. ν is absolutely continuous with respect to μ) then there exists a Σ-measurable function $f : X \rightarrow$

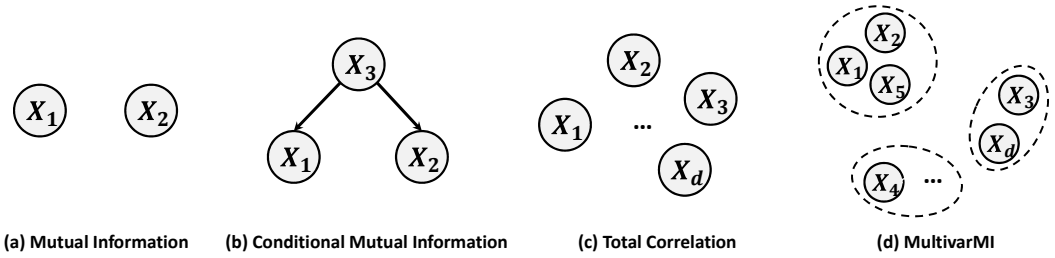


Figure 5: Graphical models for general co-information measure.

$\mathcal{G}$, the Graph Divergence Measure (GDM) defined by Rahimzamani et al. (2018) is:

$$\mathbb{GDM}(X,\mathcal{G}) = D\left(\mathbb{P}_X \| \overline{\mathbb{P}}_X\right) = \int_{\mathcal{X}} \log \frac{d\mathbb{P}_X}{d\overline{\mathbb{P}}_X} d\mathbb{P}_X \tag{27}$$

where they implicitly assume that $\log\left(d\mathbb{P}_X / d\overline{\mathbb{P}}_X\right)$ is measurable and integrable with respect to the measure $\mathbb{P}_X$. Since $\mathbb{GDM}(X,\mathcal{G}) = 0$ if and only if the data distribution is compatible with the graphical model, the GDM may be viewed as a metric of incompatibility with the provided graphical model structure.

4.3 Reduction Examples

We now show that Equation 27 is the generalization of some well-known information (Rahimzamani et al., 2018), given specific random variables X and Bayesian Network $\mathcal{G}$. The illustrations of graphical models are shown in Figure 5.

Mutual Information (MI). $X = (X_1, X_2)$ and $\mathcal{G}$ has no directed edge between X_1 and X_2. Thus $\overline{\mathbb{P}}_X = \mathbb{P}_{X_1} \cdot \mathbb{P}_{X_2}$, which produces $\mathbb{GDM}(X,\mathcal{G}) = I\left(X_1; X_2\right) = D\left(\mathbb{P}_{X_1X_2} \| \mathbb{P}_{X_1}\mathbb{P}_{X_2}\right)$

Conditional Mutual Information (CMI). We let X = (X_1, X_2, X_3) and constrain $\mathcal{G}$ as $X_1 \leftarrow X_3 \rightarrow X_2$. Since $\overline{\mathbb{P}}_X = \mathbb{P}_{X_3} \cdot \mathbb{P}_{X_2|X_3} \cdot \mathbb{P}_{X_1|X_3}$, *i.e.,* $\mathbb{GDM}(X,\mathcal{G}) = I\left(X_1; X_2 \mid X_3\right) = D\left(\mathbb{P}_{X_1X_2X_3} \| \mathbb{P}_{X_1|X_3}\mathbb{P}_{X_2|X_3}\mathbb{P}_{X_3}\right)$

Total Correlation (TC). When $X = (X_1, \cdots, X_d)$, and $\mathcal{G}$ is the graph with no edges, we recover the total correlation of the random variables $X_1, \ldots, X_d$ since $P_X = \mathbb{P}_{X_1} \ldots \mathbb{P}_{X_d}$, *i.e.,* $\mathbb{GDM}\left(X,\mathcal{G}\right) = TC\left(X_1, \ldots, X_d\right) = D\left(\mathbb{P}_{X_1 \ldots X_d} \| \mathbb{P}_{X_1} \ldots \mathbb{P}_{X_d}\right)$

Multivariate Mutual Information (MultivarMI). The definition of Equation 12 can cast as a functional of our Graph Divergence Measure by choosing for every partition, $\mathcal{P} \in \Lambda(\mathcal{X})$, a DAG $\mathcal{G}_{\mathcal{P}}$ with all X_C 's forming an aggregate node but disconnected from each other and thus inducing a measure $\overline{\mathbb{P}}_X^{\mathcal{P}}$. Thus,

$$I_{\mathcal{P}}(X) = \frac{1}{\mathcal{P}-1} \mathbb{GDM}\left(\mathbb{P}_X \| \overline{\mathbb{P}}_X^{\mathcal{P}}\right) \tag{28}$$

which implies,

$$\text{MultivarMI(X)} = \min_{\mathcal{P} \in \Pi(\mathcal{X})} \mathbb{GDM}\left(\mathbb{P}_X \| \overline{\mathbb{P}}_X^{\mathcal{P}}\right) \tag{29}$$

Partial Information Decomposition (PID). The core of PID (*cf.* Section 3.2) is to estimate the information redundancy $I_\cap(T : f)$, which is associated with a parthood distribution f. While most of the definitions of $I_\cap(T : f)$ are entropic, from which the GDM formulations are easy to

$[0, \infty)$, such that for any measurable set $A \subseteq X$, we have $\nu(A) = \int_A f d\mu$. The function f satisfying the above equality is uniquely defined up to a μ -null set, that is, if g is another function which satisfies the same property, then $f = g$ μ-lmost everywhere. Function f is commonly written $\frac{d\nu}{d\mu}$ and is called the Radon-Nikodym derivative. (From Wikipedia Radon–Nikodym theorem).

be deduced like the aforementioned cases. Take the Equation 17 as the definition of information redundancy, we have

$$
\begin{aligned}
&\sum_{f(C_i)=1,\forall 1\le i\le m} \Pi(f) \\
=&I_\cap(T:C_1,\ldots,C_m) \\
=&\sum_t p(t)\min_{C_i} I(T=t;C_i) \\
=&\sum_t p(t)\min_{C_i}\sum_{c_i} p(c_i\mid t)\left[\log_2\frac{1}{p(t)}-\log_2\frac{1}{p(t\mid c_i)}\right] \\
=&\sum_t p(t)\min_{C_i}\left[\log_2\frac{1}{p(t)}-\sum_{c_i} p(c_i\mid t)\log_2\frac{1}{p(t\mid c_i)}\right] \\
=&\sum_t p(t)\min_{C_i}\left[\log_2\frac{1}{p(t)}-\sum_{c_i} p(c_i\mid t)\log_2\frac{p(c_i)}{p(c_i\mid t)p(t)}\right] \\
=&\sum_t p(t)\min_{C_i}\left[\log_2\frac{1}{p(t)}-\sum_{c_i} p(c_i\mid t)\log_2\frac{p(c_i)}{p(c_i\mid t)}+\sum_{c_i} p(c_i\mid t)\log_2 p(t)\right] \\
=&\sum_t p(t)\min_{C_i}\left[-\sum_{c_i} p(c_i\mid t)\log_2\frac{p(c_i)}{p(c_i\mid t)}\right] \\
=&\sum_t p(t)\min_{C_i} D(\mathbb{P}(C_i|T=t)||\mathbb{P}(C_i)) \\
=&\sum_t p(t)\min_{C_i} \mathbb{GDM}\left(\mathbb{P}_{C_i}\|\overline{\mathbb{P}}_{C_i|T=t}\right) \\
=&\sum_t p(t)\min_{C_i} \mathbb{GDM}\left(C_i,\mathcal{G}\right)=\sum_t \alpha_t \min_{C_i} \mathbb{GDM}\left(C_i,\mathcal{G}\right)
\end{aligned}
\tag{30}
$$

where we have $\alpha_t = p(t)$ and $\mathcal{G} = \{(T \to C_i^j)|j \in C_i\}$. Thus, the information atoms can be solved by a series of linear equations, *i.e.,* for every $f_k, k = 1\ldots,n$:

$$
\begin{aligned}
\sum_{g\sqsubseteq f_1}\Pi(g) &= \sum_t \alpha_t \min_{f_1(C_i)=1,\forall 1\le i\le m} \mathbb{GDM}\left(\mathbb{P}_{C_i}\|\overline{\mathbb{P}}_{C_i|T=t}\right) \\
&\cdots \\
\sum_{g\sqsubseteq f_n}\Pi(g) &= \sum_t \alpha_t \min_{f_n(C_i)=1,\forall 1\le i\le m} \mathbb{GDM}\left(\mathbb{P}_{C_i}\|\overline{\mathbb{P}}_{C_i|T=t}\right)
\end{aligned}
\tag{31}
$$

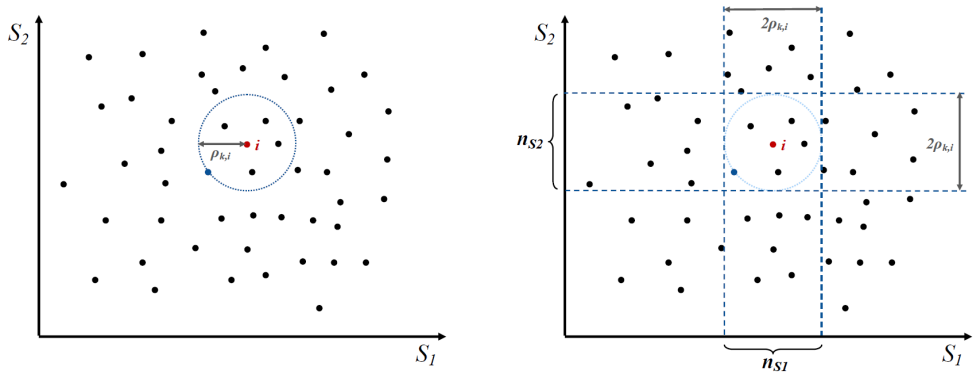
where $\alpha_i^{(j)}, j = 1,\ldots,n$ are the coefficients. The inspiration from such formulation is that we can define the information redundancy by function families of GDM (or their linear combinations), each of which corresponding to different aspects of information redundancy.

Exceptions. There are also some cases that GDM cannot generalize. For example, MultipleMI, as it is not positive in general. However, GDM is general in the sense of the manifold of the KL-divergence measure, through the alterable graph structure $\mathcal{G}$.

4.4 GDM Estimator

Algorithm 1 is the GDM estimator proposed by Rahimzamani et al. (2018), where $\psi(\cdot)$ is the digamma function and $1_{\{\cdot\}}$ is the indicator function, the norm used is ℓ_∞-norm. Let us first look at Equation 32. To prove the consistency of it with the original formulation in Equation 27, Rahimzamani et al. (2018) assume that

- k is set such that $\lim_{N\to\infty} k = \infty$ and $\lim_{N\to\infty} \frac{k\log N}{N} = 0$.



(a) (b)

Figure 6: The method for estimating information theoretic measures. (a) Step 1: Query for the distance to the k th nearest neighbor of each point i in the space $\mathcal{X}$. (b) Step 2 : Inquire for the number of points lying within the $\rho_{k,i}$ -neighborhood of each point i in the subspaces of $\mathcal{X}$ including itself. Figure from Rahimzamani et al. (2018).

Algorithm 1 Estimating Graph Divergence Measure $\mathbb{GDM}(X, \mathcal{G})$

Input: $k \in \mathbb{Z}^+$, Samples: $x^{(1)}, x^{(2)}, \ldots, x^{(N)}$
Input: Bayesian Network: $\mathcal{G}$ on Variables: $\mathcal{X} = (X_1, X_2, \cdots, X_d)$
Output: $\widehat{\mathbb{GDM}}^{(N)}(X, \mathcal{G})$
for $i = 1, \ldots, N$ **do**
# Query
$\rho_{k,i} = \ell_\infty$ -distance to the k th nearest neighbor of $x^{(i)}$ in the space $\mathcal{X}$
Inquire
$\tilde{k}_i = \#$ points within the $\rho_{k,i}$ -neighborhood of $x^{(i)}$ in the space $\mathcal{X}$
$n^{(i)}_{\mathrm{pa}(X_l)} = \#$ points within the $\rho_{k,i}$ -neighborhood of $x^{(i)}$ in the space $\mathcal{X}_{\mathrm{pa}(l)}$
$n^{(i)}_{\mathrm{pa}+(X_l)} = \#$ points within the $\rho_{k,i}$ -neighborhood of $x^{(i)}$ in the space $\mathcal{X}_{\mathrm{pa}+(l)}$
Compute
$\zeta_i = \psi\left(\tilde{k}_i\right) + \sum_{l=1}^{d}\left(1_{\{\mathrm{pa}(X_l)\neq\emptyset\}}\log\left(n^{(i)}_{\mathrm{pa}(X_l)}+1\right) - \log\left(n^{(i)}_{\mathrm{pa}+(X_l)}+1\right)\right)$
end for
return

$$\widehat{\mathbb{GDM}}^{(N)}(X, \mathcal{G}) = \frac{1}{N}\sum_{i=1}^{N}\zeta_i + \left(\sum_{l=1}^{d} 1_{\{\mathrm{pa}(X_l)=\emptyset\}} - 1\right)\log N \tag{32}$$

- The set of discrete points $\{x : P_X(x, 0) > 0\}$ is finite.
- $\int_{\mathcal{X}} |\log f(x)| d\mathbb{P}_X < +\infty$, where $f \equiv d\mathbb{P}_X / d\overline{\mathbb{P}}_X$ is Radon-Nikodym derivative.

Thus, they proposed the following theorems.

Theorem 4. *Under the Assumptions, we have* $\lim_{N\to\infty}\mathbb{E}\left[\widehat{\mathbb{GDM}}^{(N)}(X, \mathcal{G})\right] = \mathbb{GDM}(X, \mathcal{G})$

Theorem 5. *In addition to the Assumptions, assume that we have* $(k_N \log N)^2 / N \to 0$ *as* N *goes to infinity. Then we have* $\lim_{N\to\infty}\mathrm{Var}\left[\widehat{\mathbb{GDM}}^{(N)}(X, \mathcal{G})\right] = 0$.

Their proofs requires the upperbound estimation of $\left|\mathbb{E}\left[\widehat{\mathbb{GDM}}^{(N)}(X, \mathcal{G})\right] - \mathbb{GDM}(X, \mathcal{G})\right|$ and Efron-Stein inequality (*aka.* influence inequality), respectively. Since our main focus is on the

generalization of the co-information measure, we suggest the readers refer to the original paper for details.

Next we investigate the algorithm for the GDM estimation. The process is shown in Figure 6. Intuitively, the estimator estimates the GDM by the resubstitution estimate $\frac{1}{N}\sum_{i=1}^{N}\log\hat{f}\left(x^{(i)}\right)$ in which $\hat{f}\left(x^{(i)}\right)$ is the estimation of Radon-Nikodym derivative at each sample $x^{(i)}$. It is worth mentioning that GDM is able to deal with the circumstances where i) $x^{(i)}$ lies in a region where there is a density. ii) $x^{(i)}$ lies on a point where there is a discrete mass, as discussed in Rahimzamani et al. (2018). According to the type of $x^{(i)}$, we need to estimate the density functions $g(.)$ and the mass functions $h(.)$ in the data manifold of X. However, the existing k th nearest neighbor algorithms will suffer when estimating the mass functions $h(.)$, as $\rho_{n_S,i}$ for such points will be equal to zero for large N. Algorithm 1, however, is designed to approximate both $g(.)$ functions as $\approx \frac{n_S}{N}\frac{1}{\left(\rho_{n_S,i}\right)^{d_S}}$ and $h(.)$ functions as $\approx \frac{n_S}{N}$ dynamically for any subset $S \subseteq X$, which is achieved by setting $\rho_{n_S,i}$ so that all of them cancel out, yielding the estimator as in Equation 32.

5 Experiments

In this section, we conduct several numerical experiments to investigate the Graph Divergence Measure on the co-information of the random variables. The experiments mostly follow Rahimzamani et al. (2018), while we use different settings with analysis from different angles. Additionally, we include more experiments to further validate the correctness of GDM estimator.

Experiment 1: Conditional Mutual Information Estimation by Markov Chain Model.
We simulated a $X \to Z \to Y$ Markov chain model for $\mathbb{GDM}(X,\mathcal{G}) = I\left(X_1;X_2 \mid X_3\right) = D\left(\mathbb{P}_{X_1X_2X_3}\|\mathbb{P}_{X_1|X_3}\mathbb{P}_{X_2|X_3}\mathbb{P}_{X_3}\right)$. Specifically, we let $X \sim \mathrm{Uniform}(0,1)$ which is a random variable sample from the uniform random distribution in $(0,1]$. Then we set $Z = \min(X,\mathbf{a})$ and $Y = \max(Z,\mathbf{b})$, where $\mathbf{a} \neq \mathbf{b}$. Thus, X, Y and Z represent a mixture of continuous and discrete random variables. Theoretically, as X and Y is independent given Z, we have $I(X;Y|Z) = 0$.

We control the value of the hyper-parameter k and number of samples N and output the mean values upon repeating the experiment 5 times. The results are shown in Figure 7a. Clearly, the larger is k or N, the more accurate is the approximation, which validates the necessary of the assumption $\lim_{N\to\infty}\frac{k\log N}{N} = 0$. When $k = 1000$ and $N > 100$, we have the experimental value of CMI$\approx -0.015 \approx 0$, achieving a good approximation when X and Y are conditional independent given Z.

Experiment 2: Total Correlation (TC) for Independent Mixtures.
We estimated the total correlation of three independent random variables X, Y and Z. For each of the variables, *e.g.,* X, we fist generate an auxiliary random variable $\tilde{X} \sim \mathrm{Bern}(0.5)$ then X is generated via

$$X = \begin{cases} \alpha_X & \text{if } \tilde{X} = 0 \\ \sim \mathrm{Uniform}[0,1] & \text{if } \tilde{X} = 1 \end{cases}$$

which means we toss a fair coin, we will fix X at α_X if heads appears, otherwise we will draw X from $\mathrm{Uniform}[0,1]$. Likewise for Y and Z. The theory value of the TC of the three variables X, Y and Z is obviously equal to 0. In the experiment, α_X, α_Y and α_Z are set to $1, 1/2, 1/4$, respectively. We sample N tuples of (X,Y,Z) and estimated the TC by GDM. The results are shown in the Figure 7b. As k becomes large, the trend of the corresponding polyline is more stable. Also, the performance under different settings are comparable when N is large enough, which has distinct fashion as Experiment 1.

Experiment 3: Mixture of AWGN and BSC Channels with Variable Error Probability.
We considered an Additive White Gaussian Noise (AWGN) Channel in parallel with a Binary Symmetric Channel (BSC) where each time only one of them can be activated. The random variable $0 < Z < 1$ controls which channel is activated; *i.e.,* if Z is lower than the threshold β, then the AWGN channel is activated, otherwise the BSC channel will be activated. The AWGN channel is modeled as $Y = X + N$ where $X \sim \mathcal{N}\left(0,\sigma_X^2\right)$ and $N \sim \mathcal{N}\left(0,\sigma_N^2\right)$. BSC channel is modeled as $Y = X \oplus E$, where X and E are two binary random variables $X \sim \mathrm{Bern}(p)$ and $E \sim \mathrm{Bern}(Z)$

denoting the input and the error respectively. Specifically, $E \sim \text{Bern}(Z)$ means that the probability of error in the BSC channel is controlled by the variable Z each time. We let $Z = \min(\alpha, \tilde{Z})$ where $\tilde{Z} \sim \text{Uniform}(0,1)$ is an auxiliary uniform random variable, similar to the previous experiments. The theory value for $I(X:Y \mid Z)$ can be obtained by

$$\begin{aligned} I(X;Y \mid Z) =& \int_{z=0}^{1} I(X;Y \mid Z=z) f_Z(z) dz \\ =& \int_{z=0}^{\beta} I_{\text{AWGN}}(X;Y \mid Z=z) f_Z(z) dz + \int_{z=\beta}^{1} I_{\text{BSC}}(X;Y \mid Z=z) f_Z(z) dz \\ =& \beta I_{\text{AWGN}}(X;Y) + \int_{z=\beta}^{\alpha} I_{\text{BSC}}(X;Y \mid Z=z) dz + (1-\alpha) I_{\text{BSC}}(X;Y \mid Z=\alpha) \\ =& \frac{\beta}{2} \log\left(1 + \frac{\sigma_X^2}{\sigma_N^2}\right) + \int_{z=\beta}^{\alpha} (h(z\bar{p} + p\bar{z}) - h(z)) dz + (1-\alpha)(h(\alpha\bar{p} + p\bar{\alpha}) - h(\alpha)) \end{aligned} \tag{33}$$

where $h(x)$ is the binary entropy function defined as $h(x) = -x \log x - (1-x)\log(1-x)$ for $x \in [0,1]$. We set $p = 0.5, \alpha = 0.3, \beta = 0.2, \sigma_X = 1$ and $\sigma_N = 0.1$. Thus, the theory value of the conditional mutual information is $I(X;Y \mid Z) = 0.53241$. We simulated the system for various number of samples, and obtained the estimated CMI values $\hat{I}_N(X;Y \mid Z)$. The results are shown in Figure 7c, where, again, the GDM estimator performs greatly in approximating the theoretic value.

Through the experiments, we well validate the effectiveness of the GDM estimator in measuring the co-information of random variables. The GDM estimator well approximates the theory values under different settings. Also, the data manifold of GDM is defined in general probability spaces, resulting in a well-defined and general measure regardless of the discrete and continuous (or even mixtures of discrete and continuous) nature of the random variables. Moreover, we can easily construct the measure function families through certain definitions of the graphical structure of the random variables, which is a direction worth being explored in the future.

6 Conclusion

In this work, we roundly survey on the multivariate mutual information measures from various aspects. Inspired by the previous works, we adopt a general paradigm of graph divergence measures and novel estimators, which estimate several generalizations of multivariate mutual information. Through numerical experiments under different settings, this estimator can well approximate the theoretic values, which validates our generalization framework. As for applications, regardless of the high time complexity of computing the graph divergence measure, they are shown to be well-defined in general probability spaces. In the future, we intend to explore the properties of the graph divergence measures as a function family in discovering all-sided co-information of the variables in the complex systems.

References

Kenneth Ball, Christopher Grant, William R. Mundy, and Timothy J. Shafer. A multivariate extension of mutual information for growing neural networks. *Neural Networks*, 95:29–43, 2017. doi: 10.1016/j.neunet.2017.07.009. URL https://doi.org/10.1016/j.neunet.2017.07.009.

Pierre Baudot. On information links. *CoRR*, abs/2103.02002, 2021. URL https://arxiv.org/abs/2103.02002.

Pierre Baudot and Daniel Bennequin. The homological nature of entropy. *Entropy*, 17(5):3253–3318, 2015. doi: 10.3390/e17053253. URL https://doi.org/10.3390/e17053253.

Pierre Baudot, Monica Tapia, Daniel Bennequin, and Jean-Marc Goaillard. Topological information data analysis. *Entropy*, 21(9):869, 2019. doi: 10.3390/e21090869. URL https://doi.org/10.3390/e21090869.

Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, and Jürgen Jost. Shared information – new insights and problems in decomposing information in complex systems. *CoRR*, abs/1210.5902, 2012. URL http://arxiv.org/abs/1210.5902.

Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014. doi: 10.3390/e16042161. URL https://doi.org/10.3390/e16042161.

Chung Chan, Ali Al-Bashabsheh, Javad B. Ebrahimi, Tarik Kaced, and Tie Liu. Multivariate mutual information inspired by secret-key agreement. *Proc. IEEE*, 103(10):1883–1913, 2015a. doi: 10.1109/JPROC.2015.2458316. URL https://doi.org/10.1109/JPROC.2015.2458316.

Chung Chan, Ali Al-Bashabsheh, Javad B. Ebrahimi, Tarik Kaced, and Tie Liu. Multivariate mutual information inspired by secret-key agreement. *Proc. IEEE*, 103(10):1883–1913, 2015b. doi: 10.1109/JPROC.2015.2458316. URL https://doi.org/10.1109/JPROC.2015.2458316.

Chung Chan, Ali Al-Bashabsheh, Qiaoqiao Zhou, Tarik Kaced, and Tie Liu. Info-clustering: A mathematical theory for data clustering. *IEEE Trans. Mol. Biol. Multi Scale Commun.*, 2(1):64–91, 2016. doi: 10.1109/TMBMC.2016.2630054. URL https://doi.org/10.1109/TMBMC.2016.2630054.

Chung Chan, Ali Al-Bashabsheh, and Qiaoqiao Zhou. Change of multivariate mutual information: From local to global. *IEEE Trans. Inf. Theory*, 64(1):57–76, 2018. doi: 10.1109/TIT.2017.2749372. URL https://doi.org/10.1109/TIT.2017.2749372.

Ziliang Chen, Zhanfu Yang, Xiaoxi Wang, Xiaodan Liang, Xiaopeng Yan, Guanbin Li, and Liang Lin. Multivariate-information adversarial ensemble for scalable joint distribution matching. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1112–1121. PMLR, 2019. URL http://proceedings.mlr.press/v97/chen19l.html.

Hammad Dilpazir, Hasan Mahmood, Muhammad Zia, and Hafiz Malik. Face recognition: A multivariate mutual information based approach. In *2nd IEEE International Conference on Cybernetics, CYBCONF 2015, Gdynia, Poland, June 24-26, 2015*, pp. 467–471. IEEE, 2015. doi: 10.1109/CYBConf.2015.7175979. URL https://doi.org/10.1109/CYBConf.2015.7175979.

Yijie Ding, Jijun Tang, and Fei Guo. Predicting protein-protein interactions via multivariate mutual information of protein sequences. *BMC Bioinform.*, 17:398, 2016. doi: 10.1186/s12859-016-1253-9. URL https://doi.org/10.1186/s12859-016-1253-9.

Gauthier Doquire and Michel Verleysen. A comparison of multivariate mutual information estimators for feature selection. In Pedro Latorre Carmona, J. Salvador Sánchez, and Ana L. N. Fred (eds.), *ICPRAM 2012 - Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods, Volume 1, Vilamoura, Algarve, Portugal, 6-8 February, 2012*, pp. 176–185. SciTePress, 2012.

Gauthier Doquire and Michel Verleysen. Mutual information-based feature selection for multilabel classification. *Neurocomputing*, 122:148–155, 2013. doi: 10.1016/j.neucom.2013.06.035. URL https://doi.org/10.1016/j.neucom.2013.06.035.

William Bialek Elad Schneidman and Michael J. Berry II. Synergy, redundancy, and independence in population codes. *Journal of Neuroscience*, 2003. URL https://doi.org/10.1523/JNEUROSCI.23-37-11539.2003.

Conor Finn and Joseph T. Lizier. Generalised measures of multivariate information content. *Entropy*, 22(2):216, 2020. doi: 10.3390/e22020216. URL https://doi.org/10.3390/e22020216.

William B. Floyd. Review of 'information theory and coding' (abramson, n.; 1963). *IEEE Trans. Inf. Theory*, 10(4):392, 1964. doi: 10.1109/TIT.1964.1053709. URL https://doi.org/10.1109/TIT.1964.1053709.

Shuyang Gao, Rob Brekelmans, Greg Ver Steeg, and Aram Galstyan. Auto-encoding total correlation explanation. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1157–1166. PMLR, 2019. URL http://proceedings.mlr.press/v89/gao19a.html.

Benedikt Gierlichs, Lejla Batina, Bart Preneel, and Ingrid Verbauwhede. Revisiting higher-order DPA attacks: Multivariate mutual information analysis. *IACR Cryptol. ePrint Arch.*, 2009:228, 2009. URL http://eprint.iacr.org/2009/228.

Ran Gilad-Bachrach, Amir Navot, and Naftali Tishby. An information theoretic tradeoff between complexity and accuracy. In Bernhard Schölkopf and Manfred K. Warmuth (eds.), *Computational Learning Theory and Kernel Machines, 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003, Proceedings*, volume 2777 of *Lecture Notes in Computer Science*, pp. 595–609. Springer, 2003. doi: 10.1007/978-3-540-45167-9_43. URL https://doi.org/10.1007/978-3-540-45167-9_43.

Virgil Griffith and Christof Koch. Quantifying synergistic mutual information. *CoRR*, abs/1205.4265, 2012. URL http://arxiv.org/abs/1205.4265.

Virgil Griffith, Edwin K. P. Chong, Ryan G. James, Christopher J. Ellison, and James P. Crutchfield. Intersection information based on common randomness. *Entropy*, 16(4):1985–2000, 2014. doi: 10.3390/e16041985. URL https://doi.org/10.3390/e16041985.

Aaron J. Gutknecht, Michael Wibral, and Abdullah Makkeh. Bits and pieces: Understanding information decomposition from part-whole relationships and formal logic. *CoRR*, abs/2008.09535, 2020. URL https://arxiv.org/abs/2008.09535.

Te Sun Han. Linear dependence structure of the entropy space. *Inf. Control.*, 29(4):337–368, 1975. doi: 10.1016/S0019-9958(75)80004-0. URL https://doi.org/10.1016/S0019-9958(75)80004-0.

Te Sun Han. Nonnegative entropy measures of multivariate symmetric correlations. *Inf. Control.*, 36(2):133–156, 1978. doi: 10.1016/S0019-9958(78)90275-9. URL https://doi.org/10.1016/S0019-9958(78)90275-9.

Te Sun Han. Multiple mutual informations and multiple interactions in frequency data. *Inf. Control.*, 46(1):26–45, 1980. doi: 10.1016/S0019-9958(80)90478-7. URL https://doi.org/10.1016/S0019-9958(80)90478-7.

Tsung-Yu Hsieh, Suhang Wang, Yiwei Sun, and Vasant G. Honavar. Explainable multivariate time series classification: A deep neural network which learns to attend to important variables as well as time intervals. In Liane Lewin-Eytan, David Carmel, Elad Yom-Tov, Eugene Agichtein, and Evgeniy Gabrilovich (eds.), *WSDM '21, The Fourteenth ACM International Conference on Web Search and Data Mining, Virtual Event, Israel, March 8-12, 2021*, pp. 607–615. ACM, 2021. doi: 10.1145/3437963.3441815. URL https://doi.org/10.1145/3437963.3441815.

Robin A. A. Ince. Measuring multivariate redundant information with pointwise common change in surprisal. *Entropy*, 19(7):318, 2017. doi: 10.3390/e19070318. URL https://doi.org/10.3390/e19070318.

Alexander Jung and Pedro H. J. Nardelli. An information-theoretic approach to personalized explainable machine learning. *IEEE Signal Process. Lett.*, 27:825–829, 2020. doi: 10.1109/LSP.2020.2993176. URL https://doi.org/10.1109/LSP.2020.2993176.

John Kieffer. Elements of information theory (thomas m. cover and joy a. thomas). *SIAM Rev.*, 36 (3):509–511, 1994. doi: 10.1137/1036124. URL https://doi.org/10.1137/1036124.

David J. C. MacKay. *Information theory, inference, and learning algorithms*. Cambridge University Press, 2003. ISBN 978-0-521-64298-9.

William J. McGill. Multivariate information transmission. *Trans. IRE Prof. Group Inf. Theory*, 4: 93–111, 1954. doi: 10.1109/TIT.1954.1057469. URL https://doi.org/10.1109/TIT.1954.1057469.

Judea Pearl. *Probabilistic reasoning in intelligent systems - networks of plausible inference*. Morgan Kaufmann series in representation and reasoning. Morgan Kaufmann, 1989.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. 2000.

Tho Hoan Pham, Tu Bao Ho, Quynh Diep Nguyen, Dang Hung Tran, and Van Hoang Nguyen. Multivariate mutual information measures for discovering biological networks. In *2012 IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), Ho Chi Minh City, Vietnam, February 27 - March 1, 2012*, pp. 1–6. IEEE, 2012. doi: 10.1109/rivf.2012.6169834. URL https://doi.org/10.1109/rivf.2012.6169834.

Arman Rahimzamani, Himanshu Asnani, Pramod Viswanath, and Sreeram Kannan. Estimators for multivariate information measures in general probability spaces. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 8678–8689, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/c5ab6cebaca97f7171139e4d414ff5a6-Abstract.html.

Kyle Reing, Greg Ver Steeg, and Aram Galstyan. Maximizing multivariate information with error-correcting codes. *IEEE Trans. Inf. Theory*, 66(5):2683–2695, 2020. doi: 10.1109/TIT.2019.2956144. URL https://doi.org/10.1109/TIT.2019.2956144.

Fernando Rosas, Pedro A. M. Mediano, Michael Gastpar, and Henrik J. Jensen. Quantifying high-order interdependencies via multivariate extensions of the mutual information. *CoRR*, abs/1902.11239, 2019. URL http://arxiv.org/abs/1902.11239.

T. Schreiber. Measuring information transfer. *Physical review letters*, 2000.

Claude E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x. URL https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Søren Taverniers, Eric Joseph Hall, Markos A. Katsoulakis, and Daniel M. Tartakovsky. Mutual information for explainable deep learning of multiscale systems. *CoRR*, abs/2009.04570, 2020. URL https://arxiv.org/abs/2009.04570.

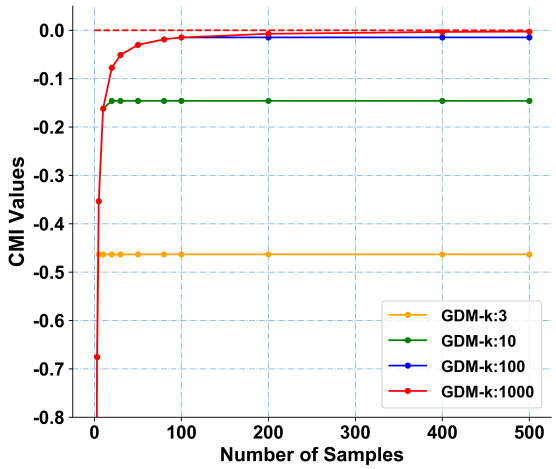
Nicholas M. Timme, Wesley Alford, Benjamin Flecker, and John M. Beggs. Synergy, redundancy, and multivariate information measures: an experimentalist's perspective. *J. Comput. Neurosci.*, 36 (2):119–140, 2014. doi: 10.1007/s10827-013-0458-4. URL https://doi.org/10.1007/s10827-013-0458-4.

Michael Satosi Watanabe. Information theoretical analysis of multivariate correlation. *IBM J. Res. Dev.*, 4(1):66–82, 1960a. doi: 10.1147/rd.41.0066. URL https://doi.org/10.1147/rd.41.0066.

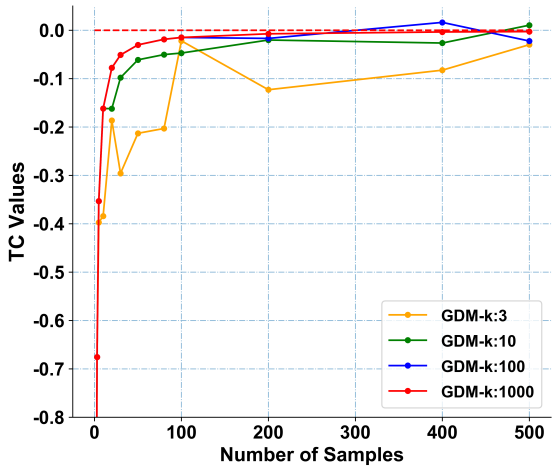
Michael Satosi Watanabe. Information theoretical analysis of multivariate correlation. *IBM J. Res. Dev.*, 4(1):66–82, 1960b. doi: 10.1147/rd.41.0066. URL https://doi.org/10.1147/rd.41.0066.

Michael Wibral, Viola Priesemann, Jim W. Kay, Joseph T. Lizier, and William A. Phillips. Partial information decomposition as a unified approach to the specification of neural goal functions. *CoRR*, abs/1510.00831, 2015. URL http://arxiv.org/abs/1510.00831.

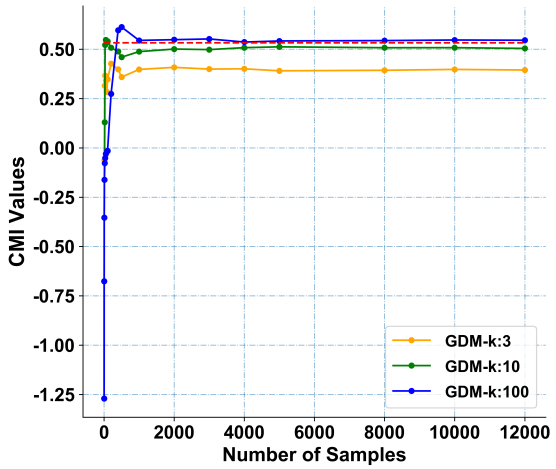
Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information. *CoRR*, abs/1004.2515, 2010. URL http://arxiv.org/abs/1004.2515.



(a) Markov Chain Simulation.



(b) Total Correlation Simulation.



(c) Mixture of AWGN and BSC channels Simulation.

Figure 7: Results of the Experiments.