Entropy **2015**, *17*, 3501-3517; doi:10.3390/e17053501

OPEN ACCESS
entropy
ISSN 1099-4300
www.mdpi.com/journal/entropy

Article

Information Decomposition and Synergy

Eckehard Olbrich [1,*], Nils Bertschinger [2] and Johannes Rauh [3]

[1] Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, 04103 Leipzig, Germany
[2] Frankfurt Institute for Advanced Studies, Ruth-Moufang-Straße 1, 60438 Frankfurt am Main, Germany; E-Mail: bertschinger@fias.uni-frankfurt.de
[3] Institute of Algebraic Geometry, Leibniz Universität Hannover, Welfengarten 1, 30167 Hannover, Germany; E-Mail: rauh@math.uni-hannover.de

* Author to whom correspondence should be addressed; E-Mail: olbrich@mis.mpg.de; Tel.: +49-341-9959-568.

Academic Editor: Rick Quax

Received: 26 March 2015 / Accepted: 19 May 2015 / Published: 22 May 2015

Abstract: Recently, a series of papers addressed the problem of decomposing the information of two random variables into shared information, unique information and synergistic information. Several measures were proposed, although still no consensus has been reached. Here, we compare these proposals with an older approach to define synergistic information based on the projections on exponential families containing only up to k-th order interactions. We show that these measures are not compatible with a decomposition into unique, shared and synergistic information if one requires that all terms are always non-negative (local positivity). We illustrate the difference between the two measures for multivariate Gaussians.

Keywords: Shannon information; mutual information; information decomposition; shared information; synergy

PACS classifications: 89.70.Cf; 89.75.-k

MSC classifications: 94A15; 94A17

---

1. Introduction

Studying a complex system usually involves figuring out how different parts of the system interact with each other. If two processes, described by random variables X and Y, interact with each other to bring about a third one, S, it is natural to ask for the contribution of the single processes. We might distinguish unique contributions of X and Y from redundant ones. Additionally, there might be a component that can be produced only by X and Y acting together: this is what we will call synergy in the following. Attempts to measure synergy were already undertaken in several fields. When investigating neural codes, S is the stimulus, and one asks how the information about the stimulus is encoded in neural representations X and Y [1]. When studying gene regulation in systems biology, S could be the target gene, and one might ask for synergy between transcription factors X and Y [2]. For the behavior of an autonomous system S, one could ask to which extent it is influenced by its own state history X or the environment Y [3].

Williams and Beer proposed the partial information lattice as a framework to achieve such an information decomposition starting from the redundant part, *i.e.*, the shared information. It is based on a list of axioms that any reasonable measure for shared information should fulfill [4]. The lattice alone, however, does not determine the actual values of the different components, but just the structure of the decomposition. In the bivariate case, there are four functions (redundancy, synergy and unique information of X and Y, respectively), related by three linear conditions. Thus, to complete the theory, it suffices to provide a definition for one of these functions. In [4], Williams and Beer also proposed a measure $I_{\min}$ for shared information. This measure $I_{\min}$ was, however, criticized as unintuitive [5,6], and several alternatives were proposed [7,8], but only for the bivariate case so far.

In this paper, we do not want to propose another measure. Instead, we want to relate the recent work on information decomposition to work on information decompositions based on projections on exponential families containing only up to k-th order interactions [2,9–11]. We focus on the synergy aspect and compare both approaches for two instructive examples: the AND gate and multivariate Gaussian distributions. We start with reviewing the construction of the partial information lattice by Williams and Beer [4] and discussing the terms for the bivariate case in more detail. In particular, we show how synergy appears in this framework and how it is related to other information measures. In Section 2.2, we recall the exponential families of k-th-order interactions and the corresponding projections and how they can be used to decompose information. In Section 3, we provide the definitions of specific synergy measures, on the one hand side, in the frame work of the partial information lattice, and on the other side, in the framework of interaction spaces, and discuss their properties. In Section 4, we compare the two measures for specific examples and conclude the paper by discussing the significance of the difference between the two measures for analyzing complex systems.

2. Information Decomposition

Let $X_1, \ldots, X_n, S$ be random variables. We are mostly interested in two settings. In the discrete setting, all random variables have finite state spaces. In the Gaussian setting, all random variables have continuous state spaces, and their joint distribution is a multivariate Gaussian.

For discrete random variables, information-theoretic quantities, such as entropy and mutual information, are canonically defined. For example, the entropy of a discrete random variable X is given by $H(X) = -\sum_x p(x) \log p(x)$, and the mutual information of two discrete random variables is:

$$MI(X : Y) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$$

The conditional mutual information is defined accordingly as:

$$MI(X : Y|Z) = \sum_{x,y,z} p(x, y|z)p(z) \log \frac{p(x, y|z)}{p(x|z)p(y|z)}$$

For continuous random variables, there is no canonical entropy function. Instead, there is differential entropy, which is computed with respect to some reference measure $\mathrm{d}x$:

$$H(X) = -\int_x p(x) \log p(x) \mathrm{d}x$$

where p now denotes the probability density of X with respect to $\mathrm{d}x$. Taking the Lebesgue measure, the entropy of an m-dimensional Gaussian random vector with covariance matrix Σ_X is given by:

$$H(X) = \frac{1}{2} \log |\Sigma_X| + \frac{1}{2} m \log(2\pi e)$$

where $|\Sigma_X|$ denotes the determinant of Σ_X. This entropy is not invariant under coordinate transformations. In fact, if $A \in \mathbb{R}^{m \times m}$, then the covariance matrix of AX is given by $A\Sigma_X A^t$, and so the entropy of AX is given by:

$$H(AX) = H(X) + \log |A|$$

In contrast, the mutual information of continuous random variables does not depend on the choice of a reference measure. The relation $MI(X : Y) = H(X) + H(Y) - H(X, Y)$ shows that, for Gaussian random vectors with covariance matrices Σ_X, Σ_Y and with a joint multivariate Gaussian distribution with joint covariance matrix $\Sigma_{X,Y}$,

$$MI(X : Y) = \frac{1}{2} \log \frac{|\Sigma_X| \cdot |\Sigma_Y|}{|\Sigma_{X,Y}|}$$

and it is easy to check directly that this is independent of linear transformations of X and Y (of course, here, one should not apply a linear transformation to the total vector (X, Y) that mixes components of X and Y).

2.1. Partial Information Lattice

We want to analyze how the information that $X_1, \ldots, X_n$ have about S is distributed among $X_1, \ldots, X_n$. In Shannon's theory of information, the total amount of information about S contained in $X_1, \ldots, X_n$ is quantified by the mutual information:

$$MI(S : X_1, \ldots, X_n) \tag{1}$$

We are looking for a way to write $MI(S : X_1, \ldots, X_n)$ as a sum of non-negative functions with a good interpretation in terms of how the information is distributed, e.g., redundantly or synergistically, among $X_1, \ldots, X_n$. For example, as we have mentioned in the Introduction and as we will see later, several suggestions have been made to measure the total synergy of $X_1, \ldots, X_n$ in terms of a function $\mathit{Synergy}(S : X_1; \ldots; X_n)$. When starting with such a function, the idea of the information decomposition is to further decompose the difference:

$$MI(S : X_1, \ldots, X_n) - \mathit{Synergy}(S : X_1; \ldots; X_n) \tag{2}$$

as a sum of non-negative functions. The additional advantage of such a complete information decomposition would be to give a better interpretation of the difference (2), apart from the tautological interpretation that it just measures "everything but the synergy." Throughout the paper, we will use the following notation: the left argument of the information quantities, the target variable S, is divided by a colon from the right arguments. The semicolon separates the different arguments on the right side, while comma-separated random variables are treated as a single vector-valued argument.

When looking for such an information decomposition, the first question is what terms to expect. In the case $n = 2$, this may seem quite easy, and it seems to be common sense to expect a decomposition of the form:

$$MI(S : X_1, X_2) = SI(S : X_1; X_2) + UI(S : X_1 \setminus X_2) + UI(S : X_2 \setminus X_1) + CI(S : X_1; X_2) \tag{3}$$

into four terms corresponding to the redundant (or shared) information $SI(S : X_1; X_2)$, the unique information $UI(S : X_1 \setminus X_2)$ and $UI(S : X_2 \setminus X_1)$ of X_1 and X_2, respectively, and the synergistic (or complementary) information $CI(S : X_1; X_2)$.

However, when $n > 2$, it seems less clear in which different ways $X_1, \ldots, X_n$ may interact with each other, combining redundant, unique and synergistic effects.

As a solution, Williams and Beer proposed the partial information framework. We explain the idea only briefly here and refer to [4] for more detailed explanations. The basic idea is to construct such a decomposition purely in terms of a function for shared information $I_\cap(S : X_1; \ldots; X_n)$ that measures the redundant information about S contained in $X_1, \ldots, X_n$. Clearly, such a function should be symmetric in permutations of $X_1, \ldots, X_n$. In a second step, $I_\cap$ is also used to measure the redundant information $I_\cap(S : A_1; \ldots; A_k)$ about S contained in combinations $A_1, \ldots, A_k$ of the original random variables (that is, $A_1, \ldots, A_k$ are random vectors whose components are among $\{X_1, \ldots, X_n\}$). Moreover, Williams and Beer proposed that $I_\cap$ should satisfy the following monotonicity property:

$$I_\cap(S : A_1; \ldots; A_k; A_{k+1}) \leq I_\cap(S : A_1; \ldots; A_k)\ , \text{ with equality if } A_i \subseteq A_{k+1} \text{ for some } i \leq k$$

(where the inclusion $A_i \subseteq A_{k+1}$ means that any component of A_i is also a component of A_{k+1}).

The monotonicity property shows that it suffices to consider the function $I_\cap$ in the case where $A_1, \ldots, A_k$ form an antichain; that is, $A_i \not\subseteq A_j$ for all $i \neq j$. The set of antichains is partially ordered by the relation:

$$(B_1, \ldots, B_l) \preceq (A_1, \ldots, A_k) \quad :\Longleftrightarrow \quad \text{for each } j = 1, \ldots, k, \text{ there exists } i \leq l \text{ with } B_i \subseteq A_j$$

and, again by the monotonicity property, $I_\cap$ is a monotone function with respect to this partial order. This partial order actually makes the set of antichains into a lattice.

If $(B_1, \ldots, B_l) \preceq (A_1, \ldots, A_k)$, then the difference $I_\cap(S : A_1; \ldots; A_k) - I_\cap(S : B_1; \ldots; B_l)$ quantifies the information contained in all A_i, but not contained in some B_l. The idea of Williams and Beer can be summarized by saying that all information can be classified according to within which antichains it is contained. Thus, the third step is to write:

$$I_\cap(S : A_1; \ldots; A_k) = \sum_{(B_1,\ldots,B_l)\preceq(A_1,\ldots,A_k)} I_\partial(S : B_1; \ldots; B_l)$$

where the function I_∂ is uniquely defined as the Möbius transform of $I_\cap$ on the lattice of antichains.

For example, the PI lattices for $n = 2$ and $n = 3$ are given in Figure 1. For $n = 2$, it is easy to make the connection with (3): The partial measures are:

$$\begin{aligned} I_\partial(S : (X_1, X_2)) &= CI(S : X_1; X_2) \\ I_\partial(S : X_1) &= UI(S : X_1 \setminus X_2) \\ I_\partial(S : X_2) &= UI(S : X_2 \setminus X_1) \\ I_\partial(S : X_1; X_2) &= SI(S : X_1; X_2) \end{aligned}$$

and the redundancy measure satisfies:

$$\begin{aligned} I_\cap(S : (X_1, X_2)) = MI(S : X_1, X_2) \quad &= CI(S : X_1; X_2) + UI(S : X_1 \setminus X_2) \\ &\quad + UI(S : X_2 \setminus X_1) + SI(S : X_1; X_2) \\ I_\cap(S : X_1) = MI(S : X_1) \quad &= UI(S : X_1 \setminus X_2) + SI(S : X_1; X_2) \\ I_\cap(S : X_2) = MI(S : X_2) \quad &= UI(S : X_2 \setminus X_1) + SI(S : X_1; X_2) \\ I_\cap(S : X_1; X_2) \quad &= SI(S : X_1; X_2) \end{aligned} \tag{4}$$

From (4) and the chain rule for the mutual information:

$$MI(S : X_1, X_2) = MI(S : X_2) + MI(S : X_1|X_2)$$

follows immediately

$$MI(S : X_1|X_2) = UI(S : X_1 \setminus X_2) + CI(S : X_1; X_2) \tag{5}$$

Even if $I_\cap$ is non-negative (as it should be as an information quantity), it is not immediate that the function I_∂ is also non-negative. This additional requirement was called local positivity in [5].

While the PI lattice is a beautiful framework, so far, there has been no convincing proposal of how the function $I_\cap$ should be defined. There have been some proposals of functions $I_\cap(S : X_1; X_2)$ with up to two arguments, so-called bivariate information decompositions [7,8], but so far, only two general information decompositions are known. Williams and Beer defined a function $I_{\min}$ that satisfies local positivity, but, as mentioned above, it was found to give unintuitive values in many examples [5,6]. In [5], $I_{\min}$ was compared with the function:

$$I_{MMI}(S : A_1; \ldots; A_k) = \min_i MI(S : A_i) \tag{6}$$

which was called minimum mutual information (MMI) in [12] (originally, it was denoted by I_I in [5]). This function has many nice mathematical properties, including local positivity. However, I_{MMI} clearly does not have the right interpretation as measuring the shared information, since I_{MMI} only compares the different amounts of information of S and A_i, without checking whether the measured information is really the "same" information [5]. However, for Gaussian random variables, I_{MMI} might actually lead to a reasonable information decomposition (as discussed in [12] for the case $n = 2$).

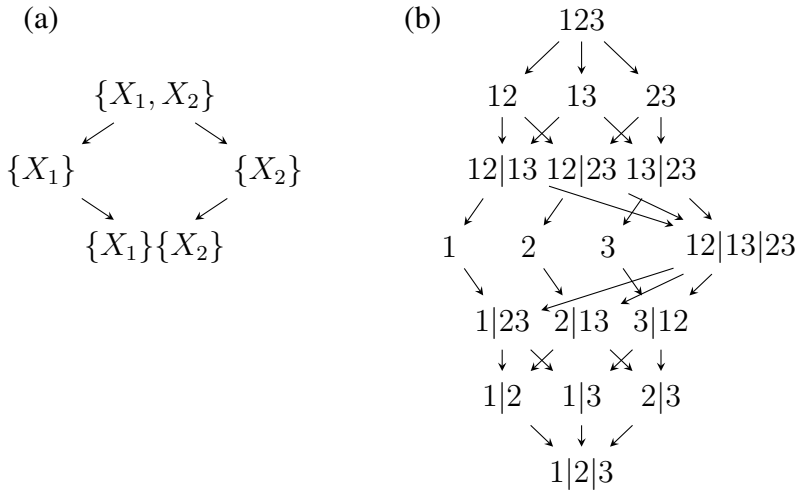

Figure 1. (**a**) The PI lattice for two random variables; (**b**) the PI lattice for $n = 3$. For brevity, every antichain is indicated by juxtaposing the components of its elements, separated by bars $|$. For example, $12|13|23$ stands for the antichain $\{X_1, X_2\}, \{X_1, X_3\}, \{X_2, X_3\}$.

2.2. Interaction Spaces

An alternative approach to quantify synergy comes from the idea that synergy among interacting systems has to do with interactions beyond simple pair interactions. We slightly change the notation and now analyze the interaction of $n + 1$ random variables $X_0, X_1, \ldots, X_n$. Later, we will put $X_0 = S$ in order to compare the setting of interaction spaces with the setting of information decompositions.

For simplicity, we restrict ourselves here to the discrete setting. Let $\binom{X}{k}$ be the set of all subsets $A \subseteq \{X_0, \ldots, X_n\}$ of cardinality $|A| = k$. The exponential family of k-th order interactions $\mathcal{E}^{(k)}$ of random variables $X_0, X_1, \ldots, X_n$ consists of all distributions of the form:

$$p(x_0, \ldots, x_n) = \prod_{A \in \binom{X}{k}} \Psi_A(x_0, \ldots, x_n)$$

where Ψ_A is a strictly positive function that only depends on those x_i with $X_i \in A$. Taking the logarithm, this is equivalent to saying that:

$$p(x_0, \ldots, x_n) = \exp\left(\sum_{A \in \binom{X}{k}} \psi_A(x_0, \ldots, x_n)\right)$$

where, again, each function ψ_A only depends on those x_i with $X_i \in A$. This second representation corresponds to the Gibbs–Boltzmann distribution used in statistical mechanics, and it also explains the name exponential family. Clearly, $\mathcal{E}^{(1)} \subseteq \mathcal{E}^{(2)} \subseteq \ldots \subseteq \mathcal{E}^{(n)} \subseteq \mathcal{E}^{(n+1)}$.

The set $\mathcal{E}^{(k)}$ is not closed (for $k > 0$), in the sense that there are probability distributions outside of $\mathcal{E}^{(k)}$ that can be approximated arbitrarily well by k-th order interaction distributions. Thus, we denote by $\overline{\mathcal{E}^{(k)}}$ the closure of $\mathcal{E}^{(k)}$ (technically speaking, for probability spaces, there are different notions of approximation and of closure, but in the finite discrete case, they all agree; for example, one may take the induced topology by considering a probability distribution as a vector of real numbers). For example, $\overline{\mathcal{E}^{(k)}}$ contains distributions that can be written as products of non-negative functions Ψ_A with zeros. In particular, $\overline{\mathcal{E}^{(n+1)}}$ consists of all possible joint distributions of $X_0, \ldots, X_n$. However, for $1 < k \leq n$, the closure of $\mathcal{E}^{(k)}$ also contains functions that do not factorize at all (see Section 2.3 in [13] and the references therein).

Given an arbitrary joint distribution p of $X_0, \ldots, X_n$, we might ask for the best approximation of p by a k-th order interaction distribution q. It is customary to measure the approximation error in terms of the Kullback–Leibler divergence:

$$D(p\|q) = \sum_{x_0,\ldots,x_n} p(x_0, \ldots, x_n) \log \frac{p(x_0, \ldots, x_n)}{q(x_0, \ldots, x_n)}.$$

There are many relations between the KL divergence and exponential families. We need the following properties:

Proposition 1. (1). *Let $\mathcal{E}$ be an exponential family, and let p be an arbitrary distribution. Then, there is a unique distribution $p_{\mathcal{E}}$ in the closure of $\mathcal{E}$ that best approximates p, in the sense that:*

$$D(p\|p_{\mathcal{E}}) = D(p\|\mathcal{E}) := \inf_{q\in\mathcal{E}} D(p\|q).$$

$p_{\mathcal{E}}$ is called the rI-projection of p to $\mathcal{E}$.
(2). *If $\mathcal{E} \subseteq \mathcal{E}'$ are two exponential families, then:*

$$D(p\|\mathcal{E}) = D(p\|\mathcal{E}') + D(p_{\mathcal{E}'}\|\mathcal{E})$$

See [9,14] for a proof and further properties of exponential families. The second identity is also called the Pythagorean theorem for exponential families.

In the following, we will abbreviate $q^{(k)} := p_{\mathcal{E}^{(k)}}$. For example, $q^{(n+1)} = p$. For $n \geq k > 1$, there is no general formula for $q^{(k)}$. For $k = 1$, one can show that:

$$q^{(1)}(x_0, \ldots, x_n) = p(X_0 = x_0) \cdot p(X_1 = x_1) \cdot \ldots \cdot p(X_n = x_n)$$

Thus, $D(p\|q^{(1)}) = \sum_{i=0}^{n} H(X_i) - H(X_0, \ldots, X_n)$ equals the multi-information [15] (also known as total correlation [16]) of $X_0, \ldots, X_n$. Applying the Pythagorean theorem $n - 1$ times to the hierarchy $\mathcal{E}^{(1)} \subseteq \mathcal{E}^{(2)} \subseteq \ldots \subseteq \mathcal{E}^{(n)}$, it follows that:

$$D(p\|q^{(1)}) = D(p\|q^{(n)}) + D(q^{(n)}\|q^{(n-1)}) + \cdots + D(q^{(2)}\|q^{(1)})$$

This equation decomposes the multi-information into terms corresponding to different interaction orders. This decomposition was introduced in [9] and studied for several examples in [10] or [17] with the single terms called connected information or interaction complexities, respectively. The idea that synergy should capture everything beyond pair interactions motivates us to define:

$$S^{(2)}(X_0; \ldots; X_n) := D(p\|q^{(2)}) = D(p\|q^{(n)}) + D(q^{(n)}\|q^{(n-1)}) + \cdots + D(q^{(3)}\|q^{(2)})$$

as a measure of synergy. In this interpretation, the synergy of $X_0, \ldots, X_n$ is a part of the multi-information of $X_0, \ldots, X_n$. The last sum shows that the hierarchy of interaction families gives a finer decomposition of $S^{(2)}$ into terms that may be interpreted as "synergy of a fixed order". In the case $n = 3$ that we will study later, there is only one term, since $p = q^{(3)}$ in this case. Using the maximum entropy principle behind exponential families [14], the function $S^{(2)}$ can also be expressed as:

$$S^{(2)}(S; X; Y) = \max_{q \in \Delta_p^{(2)}} H_q(S, Y, X) - H(S, Y, X)$$

where:

$$\Delta_p^{(2)} = \{r(x_0, \ldots, x_n) \mid r(x_i, x_j) = p(x_i, x_j) \text{ for all } i, j = 0, \ldots, n\}$$

denotes the set of all joint distributions r of $X_0, \ldots, X_n$ that have the same pair marginals as p.

In contrast, the partial information lattice provides a decomposition of the mutual information and not the multi-information. However, a decomposition of the mutual information $MI(X_0 : X_1, \ldots, X_n)$ can be achieved in a similar spirit as follows. Let $\binom{X}{k}_0$ be the set of all subsets $A \subseteq \{X_0, \ldots, X_n\}$ of cardinality $|A| = k$ that contain X_0, and let $\hat{\mathcal{E}}^{(k)}$ be the set of all probability distributions of the form:

$$p(x_0, \ldots, x_n) = \prod_{A \in \binom{X}{k}_0} \Psi_A(x_0, \ldots, x_n) \cdot \Psi_{[n]}(x_1, \ldots, x_n)$$

where the Ψ_A are as above and where $\Psi_{[n]}$ is a function that only depends on $x_1, \ldots, x_n$. As above, each $\hat{\mathcal{E}}^{(k)}$ is an exponential family.

We will abbreviate $\hat{q}^{(k)} := p_{\hat{\mathcal{E}}^{(k)}}$. Again, for general k, there is no formula for $\hat{q}^{(k)}$, but for $k = 1$, one can show that:

$$\hat{q}^{(1)}(x_0, \ldots, x_n) = p(X_0 = x_0) \cdot p(X_1 = x_1, \ldots, X_n = x_n)$$

Therefore, $D(p\|\hat{q}^{(1)}) = MI(X_0 : X_1, \ldots, X_n)$ Moreover, by the Pythagorean theorem,

$$D(p\|\hat{q}^{(1)}) = D(p\|\hat{q}^{(n)}) + D(\hat{q}^{(n)}\|\hat{q}^{(n-1)}) + \cdots + D(\hat{q}^{(2)}\|\hat{q}^{(1)})$$

Thus, we obtain a decomposition of the mutual information $MI(X_0 : X_1, \ldots, X_n)$.

Again, one can group together all terms except the last term that corresponds to the pair interactions and define:

$$\hat{S}^{(2)}(X_0 : X_1; \ldots; X_n) := D(p\|\hat{q}^{(2)}) = D(p\|\hat{q}^{(n)}) + D(\hat{q}^{(n)}\|\hat{q}^{(n-1)}) + \cdots + D(\hat{q}^{(3)}\|\hat{q}^{(2)})$$

as a measure of synergy. In this interpretation, synergy is a part of the mutual information $MI(S : X_0, \ldots, X_n)$. Using the maximum entropy principle behind exponential families [14], the function $\hat{S}^{(2)}$ can also be expressed as:

$$\hat{S}^{(2)}(S; X; Y) = \max_{q \in \hat{\Delta}_p^{(2)}} H_q(S, Y, X) - H(S, Y, X)$$

where:

$$\hat{\Delta}_p^{(2)} = \{r(x_0, \ldots, x_n) \mid r(x_0, x_i) = p(x_0, x_i) \text{ for all } i = 1, \ldots, n \text{ and } r(x_1, \ldots, r_{x_n}) = p(x_1, \ldots, x_n)\}$$

denotes the set of all joint distributions r of $X_0, \ldots, X_n$ that have the same pair marginals as p and for which, additionally, the marginal distribution for $X_1, \ldots, X_n$ is the same as for p.

While the exponential families $\mathcal{E}^{(k)}$ are symmetric in all random variables $X_0, \ldots, X_n$, in the definition of $\hat{\mathcal{E}}^{(k)}$, the variable X_0 plays a special role. This is reminiscent of the special role of S in the information decomposition framework, when the goal is to decompose the information about S. Thus, also in $\hat{S}^{(2)}$, the variable X_0 is special.

There are some relations between the hierarchies $\mathcal{E}^{(1)} \subseteq \mathcal{E}^{(2)} \subseteq \ldots \subseteq \mathcal{E}^{(n)}$ and $\hat{\mathcal{E}}^{(1)} \subseteq \hat{\mathcal{E}}^{(2)} \subseteq \ldots \subseteq \hat{\mathcal{E}}^{(n)}$.

By definition, $\mathcal{E}^{(i)} \subseteq \hat{\mathcal{E}}^{(i)}$ for $i = 1, \ldots, n$, and thus:

$$D(p\|\hat{q}^{(i)}) = D(p\|\hat{\mathcal{E}}^{(i)}) \geq D(p\|\mathcal{E}^{(i)}) = D(p\|q^{(i)})$$

In particular, $S^{(2)}(X_0; \ldots; X_n) \geq \hat{S}^{(2)}(X_0 : X_1; \ldots; X_n)$. Moreover, $\mathcal{E}^{(n)} = \hat{\mathcal{E}}^{(n)}$, which implies:

$$D(p\|\hat{q}^{(n)}) = D(p\|\hat{\mathcal{E}}^{(n)}) = D(p\|\mathcal{E}^{(n)}) = D(p\|q^{(n)})$$

In particular, for $n = 2$, this shows $S^{(2)}(S; X; Y) = \hat{S}^{(2)}(S : X; Y)$.

The case $n = 2$, $k = 2$ is also the case that we are most interested in later for the following reasons. First, for $n = 2$, the terms in the partial information lattice have an intuitively clear interpretation. Second, while there are not many examples of full information decompositions for $n > 2$, there exist at least two proposals for reasonable measures of shared, unique and complementary information [7,8], which allow a direct comparison with measures based on the decompositions using the interaction spaces.

While the symmetric hierarchy of the families $\mathcal{E}^{(k)}$ is classical, to our best knowledge, the alternative hierarchy of the families $\hat{\mathcal{E}}^{(k)}$ has not been studied before. We do not want to analyze this second hierarchy in detail here, but we just want to demonstrate that the framework of interaction exponential families is flexible enough to give a nice decomposition of mutual information, which can naturally be compared with the information decomposition framework. In this paper, in any case, we only consider cases where $\mathcal{E}^{(k)} = \hat{\mathcal{E}}^{(k)}$.

It is possible to generalize the definitions of the interaction exponential families to continuous random variables, but there are some technical issues to be solved. For example, the corresponding exponential families will be infinite-dimensional. We will not do this here in detail, since we only need the following observation later: any Gaussian distribution can be described by pair-interactions. Therefore, when p is a multivariate normal distribution, then $q^{(2)} = \hat{q}^{(2)} = p$.

3. Measures of Synergy and Their Properties

Synergy or complementary information is very often considered as a core property of complex systems, being strongly related to "emergence" and the idea of the "whole being more than the sum of its parts". In this section, we discuss three approaches to formalize this idea. We first introduce a classical function called WholeMinusSum synergy in [6], which reduces to the interaction information or (up to the sign) co-information when $n = 2$. This function can become negative. It is sensitive to redundancy, as well as synergy, and its sign tells which kind of information dominates. In Section 3.2, we recall the definition of the measure of synergy $\widetilde{CI}$ from [8] that comes from a (bivariate) information

decomposition. In Section 3.3, we compare $\widetilde{CI}$ with the synergy defined from the interaction spaces in Section 2.2.

3.1. WholeMinusSum Synergy

WholeMinusSum synergy is the difference between joint mutual information between explaining variables and the target variables and the sum of the pairwise mutual information. Griffith and Koch [6] trace it back to [18–20]. In the $n = 2$ case, this reduces to:

$$\begin{aligned} S_{WMS}(S:X;Y) &= MI(S:X,Y) - MI(S:X) - MI(S:Y) \\ &= MI(S:Y|X) - MI(S:Y) \\ &= -H(S,Y,X) + H(X,Y) + H(S,X) + H(S,Y) - H(S) - H(X) - H(Y) \\ &= -CoI(S,X,Y) \end{aligned} \tag{7}$$

with $CoI(S,X,Y)$ being the co-information [21] or interaction information [22]. This measure of synergy was used, e.g., in [1] to study synergy in neural population codes. As one can easily see from Equation (4), for any information decomposition, S_{WMS} is the difference between the complementary and the shared information:

$$S_{WMS}(S:X;Y) = CI(S:X;Y) - SI(S:X;Y) \tag{8}$$

Therefore, the WholeMinusSum synergy is a lower bound for the complementary information in the partial information lattice. Obviously it can become also negative, which makes it a deficient measure for synergy. However, it fulfills the condition of strong symmetry, *i.e.*, it is not only invariant with respect to permutation of X and Y, but to permutations of all three arguments.

3.2. Synergy from Unique Information

In [8], it was proposed to use the following function as a measure of synergy:

$$\widetilde{CI}(S:X;Y) = MI(S:X;Y) - \min_{q \in \Delta_p} MI_q(S:X;Y) \tag{9}$$

where:

$$\Delta_p = \{q(s,x,y) \mid q(s,x) = p(s,x) \wedge q(s,y) = p(s,y)\}$$

denotes the set of all joint distributions of S, X, Y that have the same pair marginals as p for the pairs (S,X) and (S,Y). Originally, this function was motivated from considerations about decision problems. The basic idea is that unique information should be observable in the sense that there should be a decision problem in which this unique information is advantageous. One crucial property is the idea that the amount of unique information should only depend on the marginal distributions of the pairs (S,X) and (S,Y), *i.e.*:

$(*)$ The functions $\widetilde{UI}(S:X \setminus Y)$ and $\widetilde{UI}(S:Y \setminus X)$ are constant on Δ_p.

These thoughts lead to a formula for unique information $\widetilde{UI}$, from which formulas for $\widetilde{SI}$ and the above formula for $\widetilde{CI}$ can be derived. Thus, in particular, $\widetilde{CI}$ is part of a (non-negative) bivariate information decomposition. While it is not easy to see directly that $\widetilde{SI}$ is non-negative, it follows right from the definition that $\widetilde{CI}$ is non-negative.

Heuristically, the formula for $\widetilde{CI}$ also encodes the idea that synergy has to do with pair interactions, here in the form of pair marginals. Namely, the joint distribution is compared with all other distributions that have the same marginals for the pairs (S, X) and (S, Y). In Section 3.3, we will see how this is related to the synergy function $S^{(2)}$ coming from the interaction decomposition.

The same measure of synergy was proposed in [6], without any operational justification, and generalized to $n > 2$ variables as follows:

$$\widetilde{CI}(S : X_1; \ldots; X_n) = MI(S : X_1, \ldots, X_n) - \min_{q \in \Delta_p} MI_q(S : X_1, \ldots, X_n)$$

where now:

$$\Delta_p = \{q(s, x_1, \ldots, x_n) \mid q(s, x_i) = p(s, x_i) \text{ for } i = 1, \ldots, n\}$$

3.3. Synergy from Maximum Entropy Arguments

Quantifying synergy using maximum entropy projections on k-th-order interaction spaces can be viewed as a more direct approach of quantifying the extent that "a system is more than the sum of its parts" [11] than the WholeMinusSum (WMS) synergy discussed above. Surprisingly, we are not aware of any publication using this approach to define explicitly a measure of synergy, but the idea seems to be common and is proposed, for instance, in [2]. Consider the joint probability distribution $p(s, x, y)$. Synergy should quantify dependencies among S, Y, X that cannot be explained by pairwise interactions. Therefore, one considers:

$$S^{(2)}(S; X; Y) = D(p \| \mathcal{E}^{(2)})$$

as a measure of synergy.

In [10], $S^{(2)}(S; X; Y)$ was discussed under the name "connected information" $I_C^{(3)}$, but it was not considered as a measure of synergy. Synergy was measured instead by the WMS synergy measure (7).

Comparing $\widetilde{CI}(S : X; Y)$ and $S^{(2)}(S; X; Y)$, we see that:

1. Both quantities are by definition ≥ 0.
2. $S^{(2)}(S; X; Y)$ is symmetric with respect to permutation of all of its arguments, in contrast to $\widetilde{CI}(S : X; Y)$.
3. $S^{(2)}(S; X; Y) \leq \widetilde{CI}(S : X; Y)$, because $\Delta_p^{(2)} \subseteq \Delta_p$ and:

$$S^{(2)}(S; X; Y) = MI(S : X, Y) - \min_{q^{(2)} \in \Delta_p^{(2)}} MI_{q^{(2)}}(S : X, Y)$$

$$\widetilde{CI}(S : X; Y) = MI(S : X, Y) - \min_{q \in \Delta_p} MI_q(S : X, Y)$$

In fact, as shown in [8], any measure CI of complementary information that comes from an information decomposition and that satisfies property $(*)$ must satisfy $\widetilde{CI}(S : X; Y) \leq CI(S : X; Y)$, and thus, the inequality:

$$S^{(2)}(S;X;Y) \leq CI(S:X;Y)$$

also holds in this more general setting.

However, we will show in the next section that if $S^{(2)}(S;X;Y)$ is considered as a synergy measure in the information decomposition [8], one gets negative values for the corresponding shared information, which we will denote by $SI^{(2)}(S;X,Y)$.

4. Examples

4.1. An Instructive Example: AND

Table 1. Joint probabilities for the AND example and corresponding values of selected entropies.

X	Y	S	p
0	0	0	$\frac{1}{4}$
0	1	0	$\frac{1}{4}$
1	0	0	$\frac{1}{4}$
1	1	1	$\frac{1}{4}$

$$\begin{aligned} H(S) &= 2\log 2 - \frac{3}{4}\log 3 \\ H(S,X) &= \frac{3}{2}\log 2 \\ H(S,X,Y) &= 2\log 2 \\ MI(S:X) &= \frac{3}{2}\log 2 - \frac{3}{4}\log 3 \\ MI(S:X|Y) &= \frac{1}{2}\log 2 \end{aligned}$$

Let X and Y be independent binary random variables with $p(0) = p(1) = \frac{1}{2}$ and $S = X \,\text{AND}\, Y$. Because the marginal distributions of the pairs (S,X) and (S,Y) are identical (by symmetry), in this example, there is no unique information [8] (Corollary 8), and therefore, by (5),

$$\widetilde{CI}(S:X;Y) = MI(S:X|Y) = \frac{1}{2}\log 2 = 0.5\,\text{bit}$$

In this example, the co-information is:

$$\begin{aligned} CoI(S;X;Y) &= SI(S:X;Y) - CI(S:X;Y) \\ &= MI(S:X) - MI(S:X|Y) \\ &= \log 2 - \frac{3}{4}\log 3 \approx -0.1887\,\text{bit} \end{aligned} \tag{10}$$

Thus, the shared information is $\widetilde{SI}(S:X;Y) = \frac{3}{2}\log 2 - \frac{3}{4}\log 3 \approx 0.3113\,\text{bit}$ and the WMS synergy is $S_{WMS}(S:X;Y) \approx 0.1887\,\text{bit}$. On the other hand, in the AND case, the joint probability distribution $p(s,x,y)$ is already fully determined by the marginal distributions $p(x,y)$, $p(s,y)$ and $p(s,x)$; that is, $\Delta_p^{(2)} = \{p\}$ (see, e.g., [10]). Therefore,

$$S^{(2)}(S;X;Y) = 0$$

If we now consider $S^{(2)}$ as a measure $CI^{(2)}$ for the complementary information in the information decomposition (4), we see from (10) that the corresponding shared information becomes negative:

$$SI^{(2)} = CoI + CI^{(2)} = -0.1887 \text{ bits} < 0$$

4.2. Gaussian Random Variables: When Should Synergy Vanish?

Let $p(s, x, y)$ be a multivariate Gaussian distribution. As mentioned above, $S^{(2)}(S; X; Y) = 0$.

What about $\widetilde{CI}$? As shown by [12], the result is that one of the two unique pieces of information $\widetilde{UI}$ always vanishes. Let r_{SX} and r_{SY} denote the correlation coefficients between S and X and S and Y, respectively. If $|r_{SX}| \leq |r_{SY}|$, then X has no unique information about S, *i.e.*: $\widetilde{UI}(S : X \setminus Y) = 0$, and therefore, $\widetilde{CI}(S : X; Y) = MI(S : X|Y)$. This was shown in [12] using explicit computations with semi-definite matrices. Here, we give a more conceptual argument involving simple properties of Gaussian random variables and general properties of $\widetilde{UI}$.

For any $\rho \in \mathbb{R}$, let $X_\rho = Y + \rho\epsilon$, where ϵ denotes Gaussian noise, which is independent of X, Y and S. Then, X_ρ is independent of S given Y, and so, $|r_{SX_\rho}| \leq |r_{SY}|$. It is easy to check that r_{SX_ρ} is a continuous function of ρ, with $r_{SX_0} = r_{SY}$ and $r_{SX_\rho} \to 0$ as $\rho \to \infty$. In particular, there exists a value $\rho_0 \in \mathbb{R}$, such that $r_{SX} = r_{SX_{\rho_0}}$. Let $X' = \frac{\sigma_X}{\sigma_{X_{\rho_0}}} X_{\rho_0}$. Then, the pair (X', S) has the same distribution as the pair (X, S) (since X' has the same variance as X and since the two pairs have the same correlation coefficient). Thus, $\widetilde{UI}(S : X \setminus Y) = \widetilde{UI}(S : X' \setminus Y)$. Moreover, since $MI(S : X'|Y) = 0$, it follows from (5) that $\widetilde{UI}(S : X' \setminus Y) = 0$.

In summary, assuming that $|r_{SX}| \leq |r_{SY}|$, we arrive at the following formulas:

$$\begin{aligned}
\widetilde{SI}(S : X; Y) &= MI(S : X) = \frac{1}{2} \log \left(1 - r_{SX}^2\right) \\
\widetilde{UI}(S : X \setminus Y) &= 0 \\
\widetilde{UI}(S : Y \setminus X) &= MI(S : Y) - MI(S : X) = \frac{1}{2} \log \left(\frac{1 - r_{SX}^2}{1 - r_{SY}^2} \right) \\
\widetilde{CI}(S : X; Y) &= MI(S : XY) - MI(S : Y) = MI(S : X|Y) \\
&= \frac{1}{2} \log \left(\frac{(1 - r_{SY}^2)(1 - r_{XY}^2)}{1 - (r_{SX}^2 + r_{SY}^2 + r_{XY}^2) + 2 r_{SX} r_{SY} r_{XY}} \right)
\end{aligned}$$

Thus, for Gaussian random variables, $\widetilde{SI}$ agrees with I_{MMI}. In fact, any information decomposition according to the PI lattice satisfies $SI(S : X; Y) \leq I_{MMI}(S : X; Y)$ [4]. Moreover, any information decomposition that satisfies $(*)$ satisfies $SI(S : X; Y) \geq \widetilde{SI}(S : X; Y)$ (Lemma 3 in [8]), and thus, all such information decompositions agree in the Gaussian case (this was first observed by [12]). In [12], it is shown that this result generalizes to the case where X and Y are Gaussian random vectors. The proof of this result basically shows that the above argument also works in this more general case.

The fact that, for Gaussian distributions, all bivariate information decompositions (that satisfy $(*)$) agree with the I_{MMI} decomposition suggests that the information decomposition based on I_{MMI} may also be sensible for Gaussian distributions for larger values of n.

Here, we do not pursue this line of thought. Instead, we want to provide another interpretation of synergy $CI(S : X; Y)$ in the Gaussian case. Based on the apparent simplicity of Gaussians where all

information measures are obtained from the correlation coefficients, one could be led to the conclusion that there should be no synergy (recall that $S^{(2)}(S;X;Y)$ vanishes). On the other hand, $S_{WMS}(S:X;Y) = CI(S:X;Y) - SI(S:X;Y)$ can be positive for Gaussian variables, and thus, synergy must be positive, as well (see [12]; for a simple example, choose $0 < r_{SX} = r_{SY} = r < 1/\sqrt{2}$ and $r_{XY} = 0$; then $S_{WMS}(S:X;Y) = \frac{1}{2}\log\frac{1-r^2}{1-2r^2} > 0$).

To better understand this situation, we regress S on X and Y, *i.e.*, we write $S = \alpha X + \beta Y + \sigma\epsilon$ for some coefficients α, β and normally distributed noise ϵ that is independent of X and Y. Let us again assume that $|r_{SX}| < |r_{SY}|$. From $CI(S:X;Y) = MI(S:X|Y)$, we see that synergy vanishes if and only if S and X are conditionally independent given Y. Since all distributions are Gaussian and information measures do not depend on the mean values, this condition can be checked by computing the conditional variances $\mathbb{V}\mathrm{ar}[S|X,Y] = \sigma^2$ and $\mathbb{V}\mathrm{ar}[S|Y] = \alpha^2\mathbb{V}\mathrm{ar}[X|Y] + \sigma^2$. We see that these distributions agree, and thus, S is conditionally independent of X given Y if $\mathbb{V}\mathrm{ar}[X|Y] = 0$, *i.e.*, X is a function of Y and effectively the same variable or if $\alpha = 0$. Positive synergy arises whenever X contributes to S with a non-trivial coefficient $\alpha \neq 0$. This is a very reasonable interpretation and shows that the synergy measure $CI(S:X;Y)$ nicely captures the intuition of X and Y acting together to bring about S.

5. Discussion and Conclusions

We think that using maximum entropy projections on k-th-order interaction spaces can be viewed as a direct approach of quantifying the extent that "a system is more than the sum of its parts" [11]. According to this view, synergy requires and manifests itself in the presence of higher-order interactions, which can be quantified using projections on the exponential families of k-th order interactions. While this idea is not new, it has, to our knowledge, not been explicitly formulated as a definition of synergy before. However, the synergy measure $S^{(2)}$ based on the projection on the exponential family of distributions with only pairwise interactions is not compatible with the partial information lattice framework, because it does not yield a non-negative information decomposition, as we have shown in the examples. The reason why we believe that it is important to have a complete non-negative information decomposition is that, in addition to a formula for synergy, it would give us an interpretation of the "remainder" $MI(S:X_1,\ldots,X_n) - \textit{Synergy}$. In the bivariate case, $\widetilde{CI}(S:X;Y)$ provides a synergy measure, which complies with the information decomposition.

One could argue that the vanishing $S^{(2)}$ for multivariate Gaussians reflects their "simplicity" in the sense that they can be transformed into independent sub-processes by a linear transformation. In contrast, this simplicity is reflected in the information decomposition by the fact that one of the unique information always vanishes. Since the WholeMinusSum synergy (or co-information) can be positive for Gaussian distributions, it is not possible to define an information decomposition for Gaussian variables that puts the synergy to zero.

Overall, our results suggest that intuition about synergy should be based on information processing rather than higher-order dependencies. While higher-order dependencies, as captured by the measure $S^{(2)}(S:X;Y)$, are part of the synergy, *i.e.*, $S^{(2)}(S:X;Y) \leq CI(S:X;Y)$, they are not required as demonstrated in our AND example and the case of Gaussian random variables. Especially, the latter

example leads to the intuitive insight that synergy arises when multiple inputs X, Y are processed simultaneously to compute the target S. Interestingly, the nature of this processing is less important and can be rather simple, *i.e.*, the output is literally just "the sum of its inputs". In this sense, we believe that our negative result, regarding the non-negativity of $S^{(2)}(S : X; Y)$, provides important insights into the nature of synergy in the partial information decomposition. It is up to future work to develop a better understanding of the relationship between the presence of higher-order dependencies and synergy.

Acknowledgments

Eckehard Olbrich has received funding from the European Community's Seventh Framework Program under Grant Agreement No. 318723 (Mathematics of Multilevel Anticipatory Complex Systems). Eckehard Olbrich also acknowledges interesting discussions with participants, in particular Peter Grassberger and Ilya Nemenman at the seminar and workshop on Causality, Information Transfer and Dynamical Networks (CIDNET14) in Dresden, Germany, 12 May–20 June 2014, which led to some of the ideas in this paper regarding information decompositions based on maximum entropy arguments. We also thank the organizers of the workshop Information Processing in Complex Systems (IPCS14) in Lucca, Italy, 24 September 2014, for having the opportunity to present and discuss the first version of these ideas.

Author Contributions

The research was initiated by Eckehard Olbrich and carried out by all authors. The manuscript was written by Eckehard Olbrich, Johannes Rauh and Nils Bertschinger. All authors have read and approved the final manuscript.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Schneidman, E.; Bialek, W.; Berry, M.J.I. Synergy, redundancy, and independence in population codes. *J. Neurosci.* **2003**, *23*, 11539–11553.
2. Margolin, A.A.; Wang, K.; Califano, A.; Nemenman, I. Multivariate dependence and genetic networks inference. *IET Syst. Biol.* **2010**, *4*, 428–440.
3. Bertschinger, N.; Olbrich, E.; Ay, N.; Jost, J. Autonomy: An information theoretic perspective. *Biosystems* **2008**, *91*, 331–345.
4. Williams, P.; Beer, R. Nonnegative Decomposition of Multivariate Information. **2010** arXiv:1004.2515v1.
5. Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J. Shared Information—New Insights and Problems in Decomposing Information in Complex Systems. In Proceedings of the European Conference on Complex Systems 2012 (ECCS'12), Brussels, Belgium, 3–7 September 2012 ; pp. 251–269.

6. Griffith, V.; Koch, C. Quantifying Synergistic Mutual Information. In *Guided Self-Organization: Inception*; Prokopenko, M., Ed.; Springer: Berlin/ Heidelberg, Germany, 2014; Volume 9, Emergence, Complexity and Computation Series; pp. 159–190.
7. Harder, M.; Salge, C.; Polani, D. Bivariate measure of redundant information. *Phys. Rev. E* **2013**, *87*, 012130.
8. Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J.; Ay, N. Quantifying unique information. *Entropy* **2014**, *16*, 2161–2183.
9. Amari, S.I. Information geometry on hierarchy of probability distributions. *IEEE Trans. Inf. Theory* **2001**, *47*, 1701–1711.
10. Schneidman, E.; Still, S.; Berry, 2nd, M.J.; Bialek, W. Network information and connected correlations. *Phys. Rev. Lett.* **2003**, *91*, 238701.
11. Ay, N.; Olbrich, E.; Bertschinger, N.; Jost, J. A geometric approach to complexity. *Chaos* **2011**, *21*, 037103.
12. Barrett, A.B. Exploration of synergistic and redundant information sharing in static and dynamical Gaussian systems. *Phys. Rev. E* **2015**, *91*, 052802.
13. Rauh, J.; Kahle, T.; Ay, N. Support sets of exponential families and oriented matroids. *Int. J. Approx. Reason.* **2011**, *52*, 613–626.
14. Csiszár, I.; Shields, P.C. Information Theory and Statistics: A Tutorial. *Found. Trends in Commun. Inf. Theory* **2004**, *1*, 417–528.
15. Studený, M.; Vejnarová, J. The Multiinformation Function as a Tool for Measuring Stochastic Dependence. In *Learning in Graphical Models*; Jordan, M.I., Ed.; Springer: Dordrecht, The Netherlands, 1998; Volume 89, NATO ASI Series; pp. 261–297.
16. Watanabe, S. Information Theoretical Analysis of Multivariate Correlation. *IBM J. Res. Dev.* **1960**, *4*, 66–82.
17. Kahle, T.; Olbrich, E.; Jost, J.; Ay, N. Complexity Measures from Interaction Structures. *Phys. Rev. E* **2009**, *79*, 026201.
18. Gawne, T.J.; Richmond, B.J. How independent are the messages carried by adjacent inferior temporal cortical neurons? *J. Neurosci.* **1993**, *13*, 2758–2771.
19. Gat, I.; Tishby, N. Synergy and Redundancy among Brain Cells of Behaving Monkeys. In *Advances in Neural Information Processing Systems 11*; Kearns, M., Solla, S., Cohn, D., Eds.; MIT Press: Cambridge, MA, USA, 1999; pp. 111–117.
20. Chechik, G.; Globerson, A.; Anderson, M.J.; Young, E.D.; Nelken, I.; Tishby, N. Group Redundancy Measures Reveal Redundancy Reduction in the Auditory Pathway. In *Advances in Neural Information Processing Systems 14*; Dietterich, T., Becker, S., Ghahramani, Z., Eds.; MIT Press: Cambridge, MA, USA, 2002; pp. 173–180.
21. Bell, A.J. The Co-Information Lattice. In Proceedings of the 4th International Workshop on Independent Component Analysis and Blind Signal Separation, Nara, Japan, 1–4 April 2003.

22. McGill, W.J. Multivariate information transmission. *Psychometrika* **1954**, *19*, 97–116.