Illum Information

Ravi Kiran Raman, Haizi Yu, and Lav R. Varshney
Coordinated Science Laboratory
University of Illinois at Urbana-Champaign

***Abstract*—Shannon's mutual information measures the degree of mutual dependence between two random variables. Two related information functionals have also been developed in the literature: multiinformation, a multivariate extension of mutual information; and lautum information, the Csiszár conjugate of mutual information. In this work, we define *illum information*, the multivariate extension of lautum information and the Csiszár conjugate of multiinformation. We provide operational interpretations of this functional, including in the problem of independence testing of a set of random variables. Further, we also provide informational characterizations of illum information such as the data processing inequality and the chain rule for distributions on tree-structured graphical models. Finally, as illustrative examples, we compute the illum information for Ising models and Gauss-Markov random fields.**

I. Introduction

Shannon's mutual information between any two random variables X and Y, and more recently its Csiszár conjugate, lautum information [1], have been defined respectively as:

$$I(X;Y) = D(p_{X,Y}\|p_X p_Y), \text{ and} \tag{1}$$

$$L(X;Y) = D(p_X p_Y\|p_{X,Y}). \tag{2}$$

For any convex function f, the Csiszár f-divergence functional [2] corresponding to any two discrete distributions $p = \{p_1, \ldots, p_m\}$ and $q = \{q_1, \ldots, q_m\}$ is defined as

$$D_f(p\|q) = \mathbb{E}_q\left[f\left(\frac{p(X)}{q(X)}\right)\right] = \sum_{i=1}^{m} q_i f\left(\frac{p_i}{q_i}\right). \tag{3}$$

Let $f^*(x) = xf\left(\frac{1}{x}\right)$ be a convex function. Then f^* is the Csiszár conjugate of the function $f(x)$, and

$$D_{f^*}(p\|q) = D_f(q\|p), \tag{4}$$

for any two distributions p and q.

Both mutual information and lautum information have several operational and informational characterizations spanning a variety of domains such as hypothesis testing, communication, and image registration [1].

A natural multivariate extension of Shannon mutual information, the *multiinformation* [3], is defined among random variables X_1^n as:

$$I(X_1;\ldots;X_n) = D(P_{X_1,\ldots,X_n}\|P_{X_1}P_{X_2}\cdots P_{X_n}). \tag{5}$$

The multiinformation satisfies several operational and informational properties, as shown in diverse application areas such as psychology [4], [5], machine learning [3], [6], image processing [7]–[9], cybernetics [10], [11], neuroscience [12], and multiterminal communication [13]–[15].

The purpose of this paper is to provide some operational characterizations and useful properties for an alternative measure of dependence where the roles of the joint and product-of-marginal distributions are reversed. We define the *illum information*[1] among random variables X_1^n as

$$L(X_1;X_2;\ldots;X_n) \triangleq D(P_{X_1}P_{X_2}\cdots P_{X_n}\|P_{X_1,X_2,\ldots,X_n}). \tag{6}$$

Illum information is the Csiszár conjugate of multiinformation, just as lautum information is the Csiszár conjugate of Shannon mutual information.

We also consider the sum of multiinformation and illum information, which we refer to as *sum information*:

$$S(X_1;\ldots;X_n) = I(X_1;\ldots;X_n) + L(X_1;\ldots;X_n). \tag{7}$$

II. Operational Characterizations

In this section, we introduce a few settings where illum information has operational significance.

A. Independence Testing

Consider the independence testing problem defined as

$$\begin{cases} H_0: & (X_1,\ldots,X_m) \sim p \\ H_1: & (X_1,\ldots,X_m) \sim p_1 \otimes \cdots \otimes p_m, \end{cases} \tag{8}$$

where p_i is the marginal distribution of X_i corresponding to p. That is, the null hypothesis corresponds to the case where the components of the random vector are dependent, drawn according to the joint distribution p. On the other hand, the alternate hypothesis corresponds to the case where the component X_i is drawn independently according to the marginal distribution p_i.

For ease, let $Y = (X_1,\ldots,X_m)$, and let $Y_1,\ldots,Y_n$ drawn independently and identically according to the underlying hypothesis as defined by (8). Let

$$\alpha = \mathbb{P}\left[\text{Decide } \{Y_1,\ldots,Y_n\} \in H_1 | H_0\right],$$
$$\beta = \mathbb{P}\left[\text{Decide } \{Y_1,\ldots,Y_n\} \in H_0 | H_1\right].$$

This work was supported in part by Air Force STTR Contract FA8650-16-M-1819 and in part by the IBM-Illinois Center for Cognitive Computing Systems Research (C3SR), a research collaboration as part of the IBM Cognitive Horizons Network.

[1] *Illum* ("that" in Latin) is the reverse spelling of *multi*, if we do not cross our *t*s. We leave that, and working through the appropriate Radon-Nikodym derivatives to the interested reader, especially since this paper is largely restricted to discrete alphabets.

From [16], we have

$$d(\alpha\|1-\beta) \leq nI(X_1;\ldots;X_m), \tag{9}$$
$$d(\beta\|1-\alpha) \leq nL(X_1;\ldots;X_m), \tag{10}$$

where $d(a\|b) = a\log\left(\frac{a}{b}\right) + (1-a)\log\left(\frac{1-a}{1-b}\right)$, $a, b \in (0,1)$. It may be noted that (9) and (10) yield upper bounds on the receiver operating characteristic (ROC) for the independence testing problem.

Let $L(X_1;\ldots;X_m)$ and $I(X_1;\ldots;X_m) < \infty$. In the asymptotic setting, Stein's lemma [17] gives an estimate of the minimum error exponents. In particular, for the best hypothesis test such that $\alpha < \delta$,

$$\lim_{n\to\infty} \frac{1}{n}\log(\beta) = -I(X_1;\ldots;X_m), \tag{11}$$

and similarly, for the best hypothesis test such that $\beta < \delta$,

$$\lim_{n\to\infty} \frac{1}{n}\log(\alpha) = -L(X_1;\ldots;X_m). \tag{12}$$

That is, we note that the Type I and Type II error exponents of independence testing are given by the illum information and multiinformation respectively.

In the Bayesian setting, let π_0 and $\pi_1 = 1-\pi_0$ be the prior probabilities of hypotheses H_0 and H_1 respectively. Then, the log-likelihood ratio of samples $Y_1,\ldots,Y_n$ is

$$\ell_n(Y_1,\ldots,Y_n) = \log\left(\frac{\pi_1}{\pi_0}\right) + \sum_{i=1}^{n}\log\left(\frac{\mathbb{P}\left[Y_i|H_1\right]}{\mathbb{P}\left[Y_i|H_0\right]}\right).$$

Then, from the law of large numbers, we have

$$\frac{1}{n}\ell_n(Y_1,\ldots,Y_n) \overset{a.s.}{\to} \begin{cases} I(X_1;\ldots;X_m), & \text{if } H_0 \\ -L(X_1;\ldots;X_m), & \text{if } H_1 \end{cases} \tag{13}$$

if the information values are finite.

Under the sequential testing framework, the sequential probability ratio test (SPRT) tracks the log-likelihood ratio which is given by

$$S_n = \sum_{i=1}^{n}\log\frac{\mathbb{P}\left[Y_i|H_1\right]}{\mathbb{P}\left[Y_i|H_0\right]},$$

and declares one of $\{H_0, H_1\}$ once the sum crosses an appropriately chosen threshold. Let D be the mean drift of S_n. Then,

$$D = \begin{cases} -I(X_1;\ldots;X_m), & \text{if } H_0 \\ L(X_1;\ldots;X_m), & \text{if } H_1. \end{cases}$$

Consequently, Wald's approximation indicates that the expected sample size, N, required for independence testing with Type I and Type II error levels α and β, is given by

$$N \approx \begin{cases} d(\alpha\|1-\beta)/I(X_1;\ldots;X_m), & \text{if } H_0 \\ d(\beta\|1-\alpha)/L(X_1;\ldots;X_m), & \text{if } H_1. \end{cases}$$

B. A Better Functional for Clustering?

Information-based clustering mechanisms have been defined recently to separate random variables into clusters which have minimal inter-cluster dependence [18]–[20]. These formulations use multivariate information functionals such as partition information and multiinformation to perform clustering, especially in universal settings. However, one practical issue in implementing such algorithms is that these information functionals are upper-bounded by entropy terms. In practice, these quantities could potentially be arbitrarily small, thereby making the clustering process very difficult.

However, we note that the illum information has no such generic upper bound in terms of entropy. Let $\mathbf{X} \sim p$, where the joint distribution p is a product over clusters in a partition P of $[n]$. That is, $p(\mathbf{X}) = \prod_{C\in P} p_C(X_C)$. Then for any partition $P' = \{C'_1,\ldots,C'_{|P'|}\}$,

$$L(X_{C'_1};\ldots;X_{C'_{|P'|}}) \geq 0, \tag{14}$$

with equality if and only if $P' \succeq P$.

For any partition $P = \{C_1,\ldots,C_{|P|}\}$, define $I^P(\mathbf{X}) = I(X_{C_1};\ldots;X_{C_{|P|}})$ and $L^P(\mathbf{X}) = L(X_{C_1};\ldots;X_{C_{|P|}})$. Then, the correct clustering of the given set of random variables, P^*, minimizes $I^P(\mathbf{X}) + L^P(\mathbf{X})$ over all partitions P. Additionally, due to non-negativity of the information functionals, it is easier to resolve between two possible partitions.

Hence we claim that clustering using the sum information functional may be more robust for universal clustering than multiinformation or partition information.

III. Informational Characterizations

Now we discuss some of the formal properties of illum information.

A. Basic Properties

Since illum information is the multivariate extension of lautum information and the Csiszár conjugate of multiinformation, several informational features extend naturally. Some such properties are the following.

1) *Non-negativity:* $L(X_1;\ldots;X_n) \geq 0$ with equality if and only if $\mathbb{P}\left[X_1,\ldots,X_n\right] = \prod_{i=1}^{n}\mathbb{P}\left[X_i\right]$. This follows directly from the fact that illum information is a relative entropy.
2) *Monotonicity:* For any $n > m \geq 2$, $L(X_1;\ldots;X_n) \geq L(X_1;\ldots;X_m)$. This follows from the chain rule and non-negativity of relative entropy as discussed later (19).
3) *Data Processing Inequality:* If $X_1 \leftrightarrow X_2 \leftrightarrow \cdots \leftrightarrow X_n$ forms a Markov chain, then the data processing inequality of lautum information [1] extends to illum information as $L(X_1;\ldots;X_{n-1}) \geq L(X_1;\ldots;X_{n-2};X_n)$. The data processing inequality also extends to tree-structured Bayesian networks.
4) *Convexity:* Directly extending the results from [1], the illum information is
   a) a concave function of $\mathbb{P}\left[X_i\right]$ for any $i \in [n]$, for a given $\mathbb{P}\left[X_{\setminus i}|X_i\right]$, and

b) a convex function of $\mathbb{P}\left[X_i|X_{\setminus i}\right]$ for any $i \in [n]$, for a given $\mathbb{P}\left[X_{\setminus i}\right]$.

5) *Invariance under bijection:* Let $f : \mathcal{X}^n \to \mathcal{Y}^n$ be a bijective mapping and let $(Y_1, \ldots, Y_n) = f(X_1, \ldots, X_n)$. The illum information is invariant to such bijective transformations, i.e.,

$$L(X_1, \ldots, X_n) = L(Y_1, \ldots, Y_n).$$

6) *Lower and upper bounds:* Let the variational information for random variables $X_1, \ldots, X_n$ be defined as

$$V(X_1, \ldots, X_n) = D_f(p_{X_1} \ldots p_{X_n} \| p_{X_1,\ldots,X_n}),$$

for the convex function $f(x) = \frac{1}{2}|x-1|$. Using Pinsker's [21] and reverse Pinsker's [22] inequalities, the illum information can be bounded in terms of the variational information as

$$L(X_1; \ldots; X_n) \geq \frac{\log_2 e}{2} V^2(X_1, \ldots, X_n), \tag{15}$$

$$L(X_1; \ldots; X_n) \leq \frac{\log_2 e}{p_{\min}} V^2(X_1, \ldots, X_n), \tag{16}$$

where $p_{\min} = \min_{x_1,\ldots,x_n} p(x_1, \ldots, x_n)$, if $p_{\min} > 0$.

B. Chain Rules

Not all informational characterizations of multiinformation extend to illum information. In particular, mutiinformation satisfies the chain rule given by

$$I(X_1; \ldots; X_n) = \sum_{i=2}^{n} I(X^{i-1}; X_i). \tag{17}$$

Let $\mathbf{X}$ be a random vector drawn according to a Bayesian network where A_i is the set of parents of node i. Then, the multiinformation decomposes as

$$I(X_1; \ldots; X_n) = \sum_{i=1}^{n} I(X_i; X_{A_i}).$$

In particular, if $P = \{C_1, \ldots, C_k\}$ is a partition of $[n]$, then

$$I(X_1; \ldots; X_n) \geq I(X_{C_1}; \ldots; X_{C_k}).$$

However, such decompositions do not necessarily hold for the case of illum information. For instance, consider $(X, Y, Z) \in \{0, 1\}^3$ drawn as follows. Let $X \sim \text{Bern}(1/2)$,

$$Y \sim \begin{cases} \text{Bern}(\epsilon), & \text{if } X = 0 \\ \text{Bern}(\bar{\epsilon}), & \text{if } X = 1, \end{cases} \quad \text{and } Z \sim \begin{cases} \text{Bern}(\epsilon), & \text{if } X = Y \\ \text{Bern}(\bar{\epsilon}), & \text{if } X \neq Y, \end{cases}$$

for $\bar{\epsilon} = 1 - \epsilon$. Then, for $\epsilon < 1/2$,

$$L(X; Y; Z) > L(X; Y) + L(X, Y; Z).$$

On the other hand, for the distribution p given in the following table,

(X,Y,Z)	$p(X,Y,Z)$	(X,Y,Z)	$p(X,Y,Z)$
$(0,0,0)$	0.04	$(1,0,0)$	0.34
$(0,0,1)$	0.29	$(1,0,1)$	0.12
$(0,1,0)$	0.01	$(1,1,0)$	0.06
$(0,1,1)$	0.11	$(1,1,1)$	0.03

we get that

$$L(X; Y; Z) < L(X; Y) + L(X, Y; Z).$$

In fact, $L(X; Y; Z) < L(X, Y; Z)$, i.e., clustering random variables does not necessarily decrease illum information as it does for multiinformation. Hence in general, the illum information does not satisfy a chain rule of the form of (17).

However, the chain rule does work for tree-structured Bayesian networks. In particular, if the n-dimensional random vector, $\mathbf{X}$ is distributed according to a tree-structured Bayesian network such that the parent of node i is A_i, then

$$L(X_1; \ldots; X_n) = \sum_{i=1}^{n} L(X_i; X_{A_i}). \tag{18}$$

Let $\mathbf{X} \sim p$ and let p_i be the marginal distribution of X_i. Let $q(\mathbf{X}) = \prod_{i=1}^{n} p_i(X_i)$. Using the chain rule of relative entropy, we have

$$L(X_1; \ldots; X_n) = \sum_{i=2}^{n} D\left(p_i(X_i) \middle\| p(X_i|X^{i-1}) \middle| q(X^{i-1})\right). \tag{19}$$

C. Distribution Approximation

Analytically approximating (as opposed to sampling) a target distribution is common in variational inference, where the true yet intractable posterior is to be approximated by a distribution that is easy to handle. This is typically done by restricting the distributions under consideration to a class that has certain properties, e.g. assuming that the class of distributions factorize in a particular way or that they all have a specific parametric form such as Gaussian. As a consequence, the approximation problem reduces to finding a distribution in the restricted class that best approximates the target distribution, i.e., a *projection* of the target distribution onto the restricted class.

Consider the problem of projecting a probability distribution p onto a set S of probability distributions. We define the projection of p onto the set S as the "closest" distribution to p among all distributions in S, where "closeness" between two distributions is measured in the following two ways.

$$\mathcal{P}_S(p) = \arg\min_{p' \in S} D(p\|p') \tag{20}$$

$$\mathcal{P}'_S(p) = \arg\min_{p' \in S} D(p'\|p) \tag{21}$$

where we have taken the two forms of relative entropy. In general, $\mathcal{P}_S(p) \neq \mathcal{P}'_S(p)$, due to the asymmetry of relative entropy. However it is trivial to see that for any $p \in S$, we have $\mathcal{P}_S(p) = \mathcal{P}'_S(p) = p$, which shows that both $\mathcal{P}_S(\cdot)$ and $\mathcal{P}'_S(\cdot)$ are indeed projections (idempotent).

Now we consider a special class of distributions, which all factorize over a given directed acyclic graph (Bayesian network). Formally, let G be a Bayesian network over random variables $X_1, \ldots, X_n$, and S_G be the set of distributions that factorize over G, then one can show that

$$\mathcal{P}_{S_G}(p) = \prod_i p_{i|A_i},$$

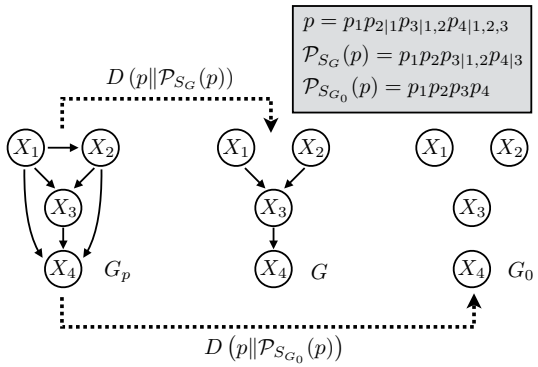

Fig. 1. Project a distribution p onto S_G and S_{G_0}, respectively. Note that: $G_p \supset G \supset G_0$, in which the numbers of independence conditions are getting larger and larger.

where inferred from p, $p_{i|A_i}$ is the conditional distribution of X_i given its parents in G. In an extreme case, let G_0 be the Bayesian network containing no edges (every node is independent of the others), then

$$\mathcal{P}_{S_{G_0}}(p) = \prod_i p_i,$$

i.e. the product of marginals.

In addition, one can show that,

$$I(X_1; \ldots; X_n) = D\left(p \| \mathcal{P}_{S_{G_0}}(p)\right) \tag{22}$$

$$= D\left(p \| \mathcal{P}_{S_G}(p)\right) + \sum_{i=1}^{n} I(X_i; X_{A_i}), \tag{23}$$

where A_i denotes the parents of X_i in G. Note that both terms in (23) are nonnegative, which implies that $I(X_1; \ldots; X_n) \geq D\left(p \| \mathcal{P}_{S_G}(p)\right)$, where equality holds if and only if $G = G_0$; and $I(X_1; \ldots; X_n) \geq \sum_{i=1}^{n} I(X_i; X_{A_i})$, where equality holds if and only if $p \in S_G$. This additive projection property of multiinformation is depicted in Figure 1.

On the other hand, such projection properties do not hold for illum information, when the other projection operator $\mathcal{P}'_S(\cdot)$ is adopted. To see this, let us consider the specific example of projecting a distribution onto the set of distributions that are product-of-marginals, i.e. the mean-field approximation of a distribution into the product distribution. Specifically, let S_{G_0} be the set of all distributions of the form $q(\mathbf{X}) = \prod_{i=1}^{n} q_i(X_i)$. Then, the mean-field approximation,

$$q^\star = \mathcal{P}'_{S_{G_0}}(p) = \arg \min_{q \in S_{G_0}} D(q \| p), \tag{24}$$

is given by the recursive formulation:

$$q_i^\star(x_i) = \exp\left(\mathbb{E}_{q^\star}\left[\log p(X^{i-1}, x_i, X_{i+1}^n)\right] - \lambda_i\right), \tag{25}$$

where λ_i is the log-partition function.

IV. Examples

In this section, we provide some exemplary computations of illum information.

A. Exponential Family

Consider an n-dimensional exponential family of distributions, $\{p_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \mathbb{R}^n\}$, of the form

$$p_{\boldsymbol{\theta}}(\mathbf{X}) = h(\mathbf{X}) \exp\left\{\boldsymbol{\theta}^T T(\mathbf{X}) - A(\boldsymbol{\theta})\right\}, \ \mathbf{X} \in \mathcal{X}^n, \tag{26}$$

where $\boldsymbol{\theta}$ is the vector of parameters, $T : \mathcal{X}^n \to \mathbb{R}^n$ is the function of sufficient statistics, $h : \mathcal{X}^n \to \mathbb{R}$, and $A(\boldsymbol{\theta})$ is the log-partition function. Let $\mathbf{X} \sim p_{\boldsymbol{\theta}}$ be the n-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_n)$. Let $p_i(\cdot)$ be the marginal distribution of X_i and $q(\mathbf{X}) = \prod_{i \in [n]} p_i(X_i)$.

The multiinformation of $p_{\boldsymbol{\theta}}$ is given by

$$I(X_1; \ldots; X_n) = \sum_{i=1}^{n} H(X_i) - A(\boldsymbol{\theta}) - \mathbb{E}_{p_{\boldsymbol{\theta}}}\left[\log h(\mathbf{X})\right] + \boldsymbol{\theta}^T \nabla A(\boldsymbol{\theta}), \tag{27}$$

where $\nabla A(\boldsymbol{\theta})$ is the gradient of the log-partition function. This follows from the fact that $\mathbb{E}_{p_{\boldsymbol{\theta}}}\left[T(\mathbf{X})\right] = \nabla A(\boldsymbol{\theta})$.

On the other hand, the illum information is given by

$$L(X_1; \ldots; X_n) = A(\boldsymbol{\theta}) - \sum_{i=1}^{n} H(X_i) + \mathbb{E}_q\left[\log h(\mathbf{X})\right] - \boldsymbol{\theta}^T \mathbb{E}_q\left[T(\mathbf{X})\right]. \tag{28}$$

Consequently, we note that the sum information

$$I(\mathbf{X}) + L(\mathbf{X}) = \boldsymbol{\theta}^T\left[\nabla A(\boldsymbol{\theta}) - \mathbb{E}_q\left[T(\mathbf{X})\right]\right] + \left[\mathbb{E}_q\left[\log h(\mathbf{X})\right] - \mathbb{E}_{p_{\boldsymbol{\theta}}}\left[\log h(\mathbf{X})\right]\right]. \tag{29}$$

In particular, consider an n-dimensional jointly Gaussian random vector, $\mathbf{X} \sim \mathcal{N}(0, \Sigma)$. Consider the eigendecomposition of the covariance matrix as

$$\Sigma = \sum_{i=1}^{n} \lambda_i u_i u_i^T,$$

where $\{\lambda_i, i \in [n]\}$ are the eigenvalues and $\{u_i, i \in [n]\}$ are the corresponding orthonormal eigenvectors. Without loss of generality, let $\lambda_i > 0$ for all i. Then,

$$I(X_1; \ldots; X_n) = \frac{1}{2} \sum_{i=1}^{n} \log\left(\frac{\sigma_i^2}{\lambda_i}\right), \tag{30}$$

$$L(X_1; \ldots; X_n) = \frac{1}{2} \sum_{i=1}^{n} \left[\frac{u_i^T \hat{\Sigma} u_i}{\lambda_i} - \log\left(\frac{\sigma_i^2}{\lambda_i}\right) - 1\right], \tag{31}$$

where $\hat{\Sigma}$ is the diagonal matrix of variance values.

Let $Y_i = X_i / \sigma_i$ and let $\tilde{\Sigma}$ be the covariance matrix of the normalized Gaussian random variables such that $\tilde{\Sigma} = \sum_{i=1}^{n} \tilde{\lambda}_i \tilde{u}_i \tilde{u}_i^T$ is the orthonormal eigendecomposition. Since

information is invariant to bijective transformations, we have $L(\mathbf{X}) = L(\mathbf{Y})$ and $I(\mathbf{X}) = I(\mathbf{Y})$, where

$$I(Y_1; \ldots; Y_n) = \frac{1}{2} \sum_{i=1}^{n} \log \left(\frac{1}{\tilde{\lambda}_i} \right), \tag{32}$$

$$L(Y_1; \ldots; Y_n) = \frac{1}{2} \sum_{i=1}^{n} \left[\frac{1}{\tilde{\lambda}_i} - \log \left(\frac{1}{\tilde{\lambda}_i} \right) - 1 \right], \tag{33}$$

$$I(\mathbf{Y}) + L(\mathbf{Y}) = \frac{1}{2} \sum_{i=1}^{n} \left[\frac{1}{\tilde{\lambda}_i} - 1 \right], \tag{34}$$

$$L(\mathbf{Y}) - I(\mathbf{Y}) = \sum_{i=1}^{n} \frac{1}{2\tilde{\lambda}_i} - \frac{1}{2} - \log \left(\frac{1}{\tilde{\lambda}_i} \right). \tag{35}$$

For $n = 2$, $L(\mathbf{X}) \geq I(\mathbf{X})$ [1]. However the result does not extend for $n > 2$. For instance, consider the 3-dimensional jointly Gaussian vector with covariance matrix

$$\Sigma = \begin{bmatrix} 1 & 0.25 & 0.25 \\ 0.25 & 1 & -0.25 \\ 0.25 & -0.25 & 1 \end{bmatrix}. \tag{36}$$

Then, $L(\mathbf{X}) - I(\mathbf{X}) = -0.0032$.

B. Pairwise Markov Random Fields

Consider an undirected graph $G = (V, E)$ and the pairwise Markov random field (MRF) defined on G, parametrized by the node potential functions $\{\psi_i(\cdot), i \in V\}$ and edge potential functions $\{\psi_{ij}(\cdot), (i,j) \in E\}$, given by

$$p_G(\mathbf{X}) = \exp \left(\sum_{i \in V} \psi_i(X_i) + \sum_{(i,j) \in E} \psi_{ij}(X_i, X_j) - A(\boldsymbol{\psi}) \right), \tag{37}$$

where $A(\boldsymbol{\psi})$ is the log-partition function. Again, let p_i be the marginal distribution of X_i and $q(\mathbf{X}) = \prod_{i \in V} p_i(X_i)$.

Then, we have

$$\begin{aligned} L(\mathbf{X}) = A(\boldsymbol{\psi}) &- \sum_{i \in V} H(X_i) - \sum_{i \in V} \mathbb{E}\left[\psi_i(X_i)\right] \\ &- \sum_{(i,j) \in E} \mathbb{E}_q\left[\psi_{ij}(X_i, X_j)\right], \end{aligned} \tag{38}$$

$$\begin{aligned} I(\mathbf{X}) = \sum_{i \in V} H(X_i) - A(\boldsymbol{\psi}) &+ \sum_{i \in V} \mathbb{E}\left[\psi_i(X_i)\right] \\ &+ \sum_{(i,j) \in E} \mathbb{E}_{p_G}\left[\psi_{ij}(X_i, X_j)\right]. \end{aligned} \tag{39}$$

This in turn indicates that the sum information

$$L(\mathbf{X}) + I(\mathbf{X}) = \sum_{(i,j) \in E} \mathbb{E}_{p_G}\left[\psi_{ij}(X_i, X_j)\right] - \mathbb{E}_q\left[\psi_{ij}(X_i, X_j)\right], \tag{40}$$

which is equivalent to the cumulative potential difference across edges owing to independence. Note that the sum information is independent of node potentials and the partition function. This indicates that the "effective information" or the symmetric distance from independence is quantified entirely by the edge effects of the MRF. Additionally it may be noted that this sum information may be estimated easily from data, given the edge potentials.

In particular, let us consider the Ising model defined on a graph $G = (V, E)$, with parameter set $\{\theta_i, i \in V\} \cup \{\theta_{ij}, (i,j) \in E\}$. For an Ising model, $\mathbf{X} \in \{-1, +1\}^{|V|}$ and the potentials are defined as

$$\psi_i(X_i) = \theta_i X_i, \text{ and } \psi_{ij}(X_i, X_j) = \theta_{ij} X_i X_j.$$

Let the log-partition function be $A(\boldsymbol{\theta})$.

Then,

$$L(\mathbf{X}) = A(\boldsymbol{\theta}) - \sum_{i \in V} \left(H(X_i) + \theta_i \bar{X}_i \right) - \sum_{(i,j) \in E} \theta_{ij} \bar{X}_i \bar{X}_j, \tag{41}$$

$$I(\mathbf{X}) = \sum_{i \in V} \left(H(X_i) + \theta_i \bar{X}_i \right) - A(\boldsymbol{\theta}) + \sum_{(i,j) \in E} \theta_{ij} \mathbb{E}\left[X_i X_j\right]. \tag{42}$$

It may be noted here that the information functionals depend only on the mean and entropy of the nodes, and the correlation across the edges. In particular,

$$L(\mathbf{X}) + I(\mathbf{X}) = \sum_{(i,j) \in E} \theta_{ij} C_{ij}, \tag{43}$$

where $C_{ij} = \mathbb{E}\left[(X_i - \bar{X}_i)(X_j - \bar{X}_j)\right]$ is the covariance corresponding to edge (i,j). This indicates that the sum information is effectively the sum of edge covariances weighted by the edge potential parameters. Using the Cauchy-Schwarz inequality and the non-negativity of multiinformation, we note that for generic Ising models,

$$L(\mathbf{X}) \leq \sum_{(i,j) \in E} \theta_{ij} \sigma_i \sigma_j \leq \sum_{(i,j) \in E} \theta_{ij}, \tag{44}$$

where σ_i^2 is the variance of X_i. That is, the illum information is bounded in terms of the variance values for Ising models, and more loosely by just the edge potential weights. Note that this upper bound holds for the multiinformation as well.

In addition, result (40), suggests a simple thought experiment. Let the graph $G = (V, E)$ represent a network of friends who are out to vote for a dinner restaurant from a list $\mathcal{X}$. Let us additionally assume that the self choice is reflected in node potentials $\{\psi_i(\cdot), i \in V\}$, and that homophily additionally increases the likelihood of similar responses among friends through edge potentials $\{\psi_{ij}(\cdot), (i,j) \in E\}$ of the form

$$\psi_{ij}(X_i, X_j) = \mathbf{1}\{X_i = X_j\}, \quad \text{for all } (i,j) \in E,$$

where $\mathbf{1}\{\cdot\}$ is the indicator function.

Friendships are strained when any two friends vote for different restaurants. To this end, a fair cost function to consider would be the cumulative strain reflected by

$$C(\mathbf{X}) = \sum_{(i,j) \in E} \mathbf{1}\{X_i \neq X_j\}.$$

From (40) and the non-negativity of information, we observe that

$$\mathbb{E}_{p_G}\left[C(\mathbf{X})\right] \leq \mathbb{E}_q\left[C(\mathbf{X})\right], \tag{45}$$

which indicates that the effective strain on a social group that colludes in taking a decision is less than the strain on a group with matching marginals, but that answers independently. This indicates the need for discussion in a social group to achieve cohesive decision-making.

V. Conclusion

In this paper, we defined illum information, the multivariate extension of lautum information and showed that it is the Csiszár conjugate of multiinformation. We gave some operational and informational characterizations of this functional, and also highlighted settings such as density estimation and the chain rule where the illum information differs from the multiinformation. Finally we computed the illum information and sum information for some examples such as pairwise Markov random fields and exponential distributions.

References

[1] D. P. Palomar and S. Verdú, "Lautum information," *IEEE Trans. Inf. Theory*, vol. 54, no. 3, pp. 964–975, Mar. 2008.

[2] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Stud. Sci. Math. Hung.*, vol. 2, pp. 299–318, 1967.

[3] M. Studený and J. Vejnarová, "The multiinformation function as a tool for measuring stochastic dependence," in *Learning in Graphical Models*, M. I. Jordan, Ed. Dodrecht: Kluwer Academic Publishers, 1998, pp. 261–297.

[4] W. J. McGill, "Multivariate information transmission," *IRE Trans. Inf. Theory*, vol. IT-4, no. 4, pp. 93–111, Sep. 1954.

[5] D. M. Fass, "Human sensitivity to mutual information," Ph.D. dissertation, Rutgers, The State University of New Jersey, New Brunswick, Jan. 2006.

[6] A. Jakulin, "Machine learning based on attribute interactions," Ph.D. dissertation, University of Ljubljana, Slovenia, Jun. 2005.

[7] J. L. Boes and C. R. Meyer, "Multi-variate mutual information for registration," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI'99*, ser. Lecture Notes in Computer Science, C. Taylor and A. Colchester, Eds. Berlin: Springer, 1999, vol. 1679, pp. 606–612.

[8] Y.-M. Zhu, "Volume image registration by cross-entropy optimization," *IEEE Trans. Med. Imag.*, vol. 21, no. 2, pp. 174–180, Feb. 2002.

[9] L. Zollei, "A unified information theoretic framework for pair- and group-wise registration of medical images," Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, Jan. 2006.

[10] W. R. Ashby, "Measuring the internal informational exchange in a system," *Cybernetica*, vol. 8, no. 1, pp. 5–22, 1965.

[11] R. C. Conant, "Laws of information which govern systems," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-6, no. 4, pp. 240–255, Apr. 1976.

[12] G. Chechik, M. J. Anderson, O. Bar-Yosef, E. D. Young, N. Tishby, and I. Nelken, "Reduction of information redundancy in the ascending auditory pathway," *Neuron*, vol. 51, no. 3, pp. 359–368, Aug. 2006.

[13] Y.-S. Liu and B. L. Hughes, "A new universal random bound for the multiple-access channel," *IEEE Trans. Inf. Theory*, vol. 42, no. 2, pp. 376–386, Mar. 1996.

[14] H. Wang and P. Viswanath, "Vector Gaussian multiple description with individual and central receivers," *IEEE Trans. Inf. Theory*, vol. 53, no. 6, pp. 2133–2153, Jun. 2007.

[15] V. Misra, V. K. Goyal, and L. R. Varshney, "Distributed scalar quantization for computing: High-resolution analysis and extensions," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 5298–5325, Aug. 2011.

[16] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Stat.*, vol. 22, no. 1, pp. 79–86, 1951.

[17] H. Chernoff, "Large-sample theory: Parametric case," *Ann. Math. Stat.*, vol. 27, no. 1, pp. 1–22, Mar. 1956.

[18] C. Chan, A. Al-Bashabsheh, J. B. Ebrahimi, T. Kaced, and T. Liu, "Multivariate mutual information inspired by secret-key agreement," *Proc. IEEE*, vol. 103, no. 10, pp. 1883–1913, Oct. 2015.

[19] K. Nagano, Y. Kawahara, and S. Iwata, "Minimum average cost clustering," in *Advances in Neural Information Processing Systems 23*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Eds. MIT Press, 2010, pp. 1759–1767.

[20] R. K. Raman and L. R. Varshney, "Universal joint image clustering and registration using partition information," arXiv:1701.02776, Jan. 2017.

[21] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*. San Francisco: Holden-Day, 1964.

[22] I. Csiszár and Z. Talata, "Context tree estimation for not necessarily finite memory processes, via BIC and MDL," *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 1007–1016, Mar. 2006.