A REVIEW ON MULTIVARIATE MUTUAL INFORMATION

Sunil Srinivasa
University of Notre Dame

I. Introduction

Typically, mutual information is defined and studied between just two variables. Though the approach to evaluate bivariate mutual information is well established, several problems in multi-user information theory require the knowledge of interaction between more than two variables. Since there exists dependency between the variables, we cannot decipher their relationship without considering all of them at once. The seminal work on the information-theoretic analysis of the interaction between more than two variables (or in other words, multivariate mutual information) was first studied in [1].

If several sources transmit information to a receiver, the bivariate model with certainly fail to discriminate effects due to uncontrolled sources from those due to random variability. We should not confuse the impairments due to system noise with the absence of knowledge of the association between the inputs. Besides, in a practical scenario, we don't know in advance as to how many sources are transmitting information. By employing the multivariate model, we can effectively measure the effects due to the various transmitting sources. It provides a simple method for evaluating and testing dependencies in multidimensional frequency data or contingency tables.

In *section II*, I will summarize the theoretical development and talk about numerous properties of multivariate mutual informations. *Section III* gives an introduction to total multivariate correlation analysis. Asymptotic hypothesis testing is discussed in *section IV* of this report. Some applications of multivariate mutual information are mentioned in *section V*. *Section VI* concludes the report. *Section VII* lists the key references used.

II. Development of the Theory

Consider a single-input single-output channel with a discrete input X and output Y with probability distributions $p_X(x)$ and $p_Y(y)$ respectively. The amount of transmission between X and Y is defined in terms of the individual and joint entropies as

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{1}$$

where entropy (of X in this case) is given by

$$H(X) = -\sum_{x \in \chi} p_X(x) log(p_X(x))$$

Consider now a channel with two inputs U, V and a single output Y. This is commonly known as the two-way channel. The mutual information between the inputs and the output of a two-way channel is written (as an extension of (1)) as

$$I(U,V;Y) = H(U,V) + H(Y) - H(U,V,Y) \tag{2}$$

The introduction of V might affect the relationship between U and Y in several ways. In order to study the effect, the introduction of V has on the single-input single-output channel, we will need to reduce the three-dimensional information to two dimensions. One way to annul the effect of V is by reducing the three dimensional equations to two variables and writing

$$I(U;Y) = H(U) + H(Y) - H(U,Y) \tag{3}$$

Another way in which V could be eliminated is by taking a weighted sum (on the probability of occurrence of that particular value of V) of the mutual information between U and Y for each value of V.

$$\begin{aligned} I_V(U;Y) &= \sum_{v\in\vartheta} p_V(v) I(U;Y|V=v) = I(U;Y|V) \\ &= I(U,V;Y) - I(V;Y) \\ &= \Big(H(U,V) + H(Y) - H(U,V,Y)\Big) - \Big(H(V) + H(Y) - H(V,Y)\Big) \\ &= H(U,V) - H(U,V,Y) - H(V) + H(V,Y) \end{aligned} \tag{4}$$

If V has no effect on the transmission between U and Y, $I(U;Y) = I_V(U;Y)$, and the analysis reduces to that for a single-input single-output channel.
In a general case however, the difference between the two is given by

$$I(U;Y) - I_V(U;Y) = H(U) + H(V) + H(Y) - (H(U,V) + H(V,Y) + H(U,Y)) + H(U,V,Y) \tag{5}$$

McGill [1] defines this term as the mutual interaction between the three variables U, V and Y and is denoted as $I(U;V;Y)$.

The analysis above assumes that the distributions (in probability) of the inputs and output are known. In the case that the exact values of probability distributions are not known, we could use the empirical values of entropy. The multivariate information analysis can be analogously extended to continuously-distributed variables as well.

A. Extension to Multi-dimensional Variables

The definition of mutual information has been extended to a general case (over more than three variables) by Fano [2] and re-formulated in a lattice-theoretic framework by Han [3]. Though each have taken different approaches and the expressions are in terms of different entities (mutual informations in one case and entropies in the other), they can be simplified to be the same.

Fano [2] computes the mutual information between an arbitrary number of events, as an extension to the bivariate case as follows.

$$\begin{aligned} I(X_1;X_2) &= H(X_1) - H(X_1|X_2) \\ &= I(X_1) - I(X_1|X_2) \end{aligned} \tag{6}$$

The second equality is conveniently used since entropy of a variable is its self-information.
Extending to a triple product ensemble

$$I(X_1;X_2;X_3) = I(X_1;X_2) - I(X_1;X_2|X_3) \tag{7}$$

Generalizing over N variables,

$$I(X_1;X_2;\ldots;X_N) = I(X_1;X_2;\ldots;X_{N-1}) - I(X_1;X_2;\ldots;X_{N-1}|X_N) \tag{8}$$

As a side-equation, the self-information of a N-product ensemble can be expanded out in terms of mutual informations between the individual components as

$$I(X_1X_2\ldots X_N) = \sum I(X_i) - \sum I(X_i;X_j) + \ldots\ldots (-1)^n I(X_1;X_2;\ldots;X_N) \tag{9}$$

The summations are taken over all combinations of subscripts.

Han [3] introduced the concept of *difference operator* to describe multiple interactions in frequency data or contingent tables. According to him, the N-information(I_N) is the difference of the entropy function and is given by

$$I_N(X_1; X_2; \ldots; X_N) = \sum_{k=1}^{N} (-1)^{k-1} \sum_{\substack{X \subset (X_1, X_2, \ldots, X_N) \\ |X|=k}} H(X) \tag{10}$$

Expanding out,

$$I_N(X_1; X_2; \ldots X_N) = (H(X_1) + H(X_2) \ldots + H(X_N)) - \ldots + \ldots (-1)^{N-1} H(X_1, X_2, \ldots X_N) \tag{11}$$

B. Properties of Multivariate Mutual Information

- Having defined multivariate mutual information, we will try to give an intuitive meaning to it. It can be interpreted as the gain (or loss) in the information transmitted between a set of variables due additional knowledge of an extra variable. In other words, we can think of I_N as the dependence reduction.

  $$I_N(X^N) = I_{N-1}(X^{N-1}) - I_{N-1}(X^{N-1}|X_N) \tag{12}$$

  where $X^k = (X_1; X_2; \ldots; X_k)$

- A surprising note to make is that, contrary to bivariate information which is always positive, multivariate mutual information can be either positive or negative. This is seen to be possible since the effect of holding one of the variables may increase or decrease dependence between the others. As a trivial case, consider the situation in the trivariate product where variables U and Y are independent when V is not known, but become dependent given V. For this case, $I(U; V; Y)$ is clearly negative. Han [3] has shown that mutual information for multivariate variables need not be always positive, by expanding it in terms of parameters of probability up to the second order.

- From Han's expansion for the N-information (11), it is straightforward to note that multivariate mutual information is completely symmetric with respect to its components. Hence, the N-information can be written out in N possible ways.

- It is easy to see that the multivariate analysis is much more precise compared to the bivariate case. As an example, consider the transmission $(U, V) \rightarrow Y$. In the bivariate information analysis, we would have

  $$H(Y) = H_U(Y) + I(U; Y) \tag{13}$$

  where $H(Y)$ can be interpreted as the uncertainty in the output and $H_U(Y)$ is the residual uncertainty in the output after the information due to input U is accounted for. Working on similar lines for a trivariate analysis, we end up with

  $$H(Y) = H_{UV}(Y) + I(U, V; Y) \tag{14}$$

  Here, $H_{UV}(Y)$ is the residual uncertainty which ends up being the error term. Since conditioning decreases entropy, the analytical error is reduced for the trivariate analysis over the bivariate one. As the dimensionality

of the system increases, we end up with a better estimate of the noise information and hence obtain a better overall transmission. For an N-dimensional system with inputs $(X_1, X_2, \ldots, X_N)$ and output Y,

$$H(Y) = H_{X_1 X_2 \ldots X_N}(Y) + I(X_1, X_2, \ldots, X_N; Y) \tag{15}$$

is used to analyze the multivariate transmission of information.

- A very important property for a multivariate information quantity is the concept of *semi-independence*. A bivariate mutual information is equal to zero if the two variables are independent. For multidimensional variables however, independence is not the necessary condition for no-multivariate interactions . Han [3] discusses the concept of semi-independence as a subtler extension of independence.
  Let $\alpha \in X$ be a subset of the N-dimensional random vector X. Let $r(\alpha)$ be defined as the number of elements of X in the subset α. We call a distribution semi-independent (with respect to α) if

$$\pi_\alpha \equiv \sum_{\phi \leq \gamma \leq \alpha} (-1)^{r(\alpha)-r(\gamma)} Pr^0\{\alpha \cap \bar{\gamma}\} Pr\{\gamma\} = 0 \tag{16}$$

  where $Pr^0\{.\}$ refers to an independent distribution and hence can be expanded as a product of its marginals' probabilities.
  For the bivariate case, not surprisingly, semi-independence essentially leads to independence. This is not the case in general for higher orders. As an illustration, the semi-independence equations for a trivariate product ($\pi_{X_1 \cap X_2 \cap X_3}$=0) reduces to

$$\begin{aligned} 0 =& Pr\{UVY\} - Pr^0\{U\}Pr\{VY\} - Pr^0\{V\}Pr\{UY\} - Pr^0\{Y\}Pr\{UV\} \\ &+ Pr^0\{UV\}Pr\{Y\} + Pr^0\{VY\}Pr\{U\} + Pr^0\{VU\}Pr\{Y\} - Pr^0\{UVY\} \end{aligned} \tag{17}$$

  This in turn implies,

$$Pr\{UVY\} = Pr\{U\}Pr\{VY\} + Pr\{V\}Pr\{UY\} + Pr\{Y\}Pr\{UV\} - 2Pr\{U\}Pr\{V\}Pr\{Y\} \tag{18}$$

  Note that this is much stronger than the "plain" independence conditions. Infact, independence can be thought of as a composite case of semi-independence. For independence, (16) should hold for all $(\alpha \in X$, with $r(\alpha) \geq 2)$.

- The *recursive*(19) and *chaining*(20) properties hold for the multivariate mutual information [5]. They can be respectively stated as follows. Proofs follow immediately on expanding both sides in terms of entropy functions.

$$I_N((X_1, X_2); X_3; \ldots; X_N | X_0) = I_N(X_1; X_3; \ldots; X_N | X_0) + I_N(X_2; X_3; \ldots; X_N | X_0) \tag{19}$$

$$I_N(X_1; X_2; \ldots; X_N | X_0) = I_{N-1}(X_1; X_2; \ldots; X_{N-1} | X_0) - I_{N-1}(X_1; X_2; \ldots; X_{N-1} | X_N, X_0) \tag{20}$$

III. Total Correlation

For correlated sources of information, it will prove useful to study all the possible associations between the system variables. However, (2) accounts for only part of the interaction and is hence incomplete in that respect. By being able to generate all possible components of the multivariate transmission, we can examine and analyze all possible associations. The multivariate correlation model put forth by Watanabe generates all possible combinations of bivariate and higher order mutual informations, thus taking into account all the existing source correlations. Watanabe [4] was the first to discuss the concept of "total correlation" in detail, though this concept had been described earlier by McGill [1]. The definition is as follows.

Assume a set X of random variables $X_1, X_2, \ldots, X_N$, being grouped into subsets $x_1, x_2, \ldots, x_k$. The correlation existing in X is expressed as

$$C(X) = \sum_{i=1}^{k} H(x_i) - H(X) \tag{21}$$

For a given X, C is maximized for $n = k$ (each subset containing only one variable) and the value of C then is defined as *Total Correlation*. According to [4], the total correlation is defined by

$$C(X_1, X_2, \ldots, X_N) = \sum_{i=1}^{N} H(X_i) - H(X_1, X_2, \ldots X_N) \tag{22}$$

A bit of simplification leads to the more-important equation -

$$C(X_1, X_2, \ldots, X_N) = \sum I(X_i; X_j) + \sum I(X_i; X_j; X_k) + \ldots + I(X_1; X_2; \ldots; X_N) \tag{23}$$

The summation is taken over all possible combinations of subscripts. Evidently, (23) is a good method to measure inter-variable dependencies. It can be shown that total correlation is always positive and is nullified if and only if all the sources are independent of each other.

IV. Hypothesis Testing

The maximum likelihood estimators of Shannon's entropy and multiple mutual information are obtained by sampling the system variables and using the resulting empirical values of probabiliy distributions. Based on this, Han [3] shows that the asymptotic likelihood estimators are good statistics for testing semi-independence conditions for the multivariate mutual information. Some of the key results stated in [3] are :

- Let $\alpha \in X, (r(\alpha) \geq 2)$ and $\hat{I}(.)$ be the estimator of I, then $-2m\hat{I}(\alpha)$ converges in distribution asymptotically $(m \to \infty)$ to a χ^2 statistic with $\prod_{X_\alpha \leq a} N_\alpha - 1$ degrees of freedom.
  Furthermore, if $\alpha \neq \beta$, $-2m\hat{I}(\alpha)$ and $-2m\hat{I}(\beta)$ asymptotically approach independent χ^2-distributions.
- $-2m\hat{I}(\alpha)$ converges in distribution to a central χ^2-distribution if and only if

$$\lim_{m \to \infty} (m^{\frac{1}{2}} \pi_\alpha) = 0 \tag{24}$$

Using the above arguments, we can test for the trueness of the zero-hypothesis condition. As an example, consider the case of $\alpha = (X_1 \cap X_2 \cap X_3)$ We can test for the case when the trivariate mutual information is zero.
Under the case of semi-independence, this occurs when
$lim_{m \to \infty}(-2m\hat{I}(X_1; X_2; X_3)) \sim \chi^2$ with $(N_1 - 1)(N_2 - 1)(N_3 - 1)$ degrees of freedom.
If independence is assumed however, the null hypothesis is true when
$lim_{m \to \infty}(-2m\hat{I}(X_1; X_2; X_3)) \sim \chi^2$ with $(N_1 N_2 N_3 - 1) - (N_1 - 1) - (N_2 - 1) - (N_3 - 1)$ degrees of freedom.
A more important problem involves testing suspected information sources wherein multiple quantities can be tested simultaneously and is discussed in detail in [1].

V. Applications of Multiple Mutual Informations

An important application of the multivariate information is in the area of multi-user information theory. [5] deals with calculating capacity regions in a two-way channel and concludes that the trivariate mutual information is an important term to be considered. As mentioned earlier, it can be used to study multi-way interactions in multidimensional frequency data or contingency tables. The usefulness of multivariate mutual information extends beyond the realms of information theory. Ideas derived from it can be used in the area of psychology to test the response of a human being to a certain stimulus and a pre-response and study their associations [1]. Visualization of attribute interactions has provided an insight into the underlying structure of data on a number of domains and has helped in building predictive models [6].

VI. Conclusions

The concept of multivariate mutual information was introduced as an extension to the bivariate case. This measure can be used very effectively and systematically in analyzing discrete experimental data, while the traditional techniques of the analysis of variance cannot be employed because no numerical values are associated with the events. Contrary to the bivariate case, multivariate information can be negative but is much stronger when used for error analysis .The concept of independence is no longer “minimal” for the multidimensional information and can be broken down into finer ones called *semi-independence*. In order to study all possible source interactions, *total correlation* was shown to be an important, yet a tractable measure. The log-likelihood estimators of McGill’s higher-order mutual information asymptotically tend to independent χ^2-square distributions.

VII. References and Bibliography

[1] W. J. McGill. Multivariate Information Transmission. *IEEE Trans. Information Theory*, vol.4(4), pp. 93-111 (1954)

[2] R. M. Fano. ‘Transmission of Information’, pp. 57-59 (1961)

[3] T. S. Han. Multiple Mutual Informations and Multiple Interactions in Frequency Data. *Information and Control*, vol. 46(1), pp. 26-45 (1980)

[4] S. Watanabe. Information Theoretical Analysis of Multivariate Correlation. *IBM Journal of Research and Development*, vol. 4, pp. 66-82 (1960)

[5] A. P. Hekstra, F. M. J Willems. Dependence Balance Bounds for Single-output Two-way Channels. *IEEE Trans. Information Theory*, vol.35(1), pp. 44-53 (1989)

[6] A. Jakulin and I. Bratko. Analyzing Attribute Dependencies. *PKDD of LNAI*, vol. 2838, pp. 229-240 (2003)