Decomposing multivariate information

Nathaniel Virgo[1] and Daniel Polani[2]

[1]Earth-Life Science Institute (ELSI), Tokyo, Japan

[2]University of Hertfordshire, Hatfield, UK

Pairwise relationships between random variables are well understood in information theory, but there are a number of important open questions on how to understand and quantify the relationships between three or more random variables. The intricacies of the multivariable case were lucidly highlighted in [8]. James and Crutchfield [4] give another useful framing; they raise the following questions: how much of the information in a three-variable system is in the form of three-way interactions, as opposed to pair-wise ones? More generally, how can we decompose the joint entropy of n variables to properly understand all of the k-way interactions (with $1 \leq k \leq n$) and their relationships?

We argue that the first of these questions is precisely answered by the method of Amari [1], developed also by [5, 6], which we extend in order to to answer the second question. For n random variables, Amari constructs a hierarchy of n probability distributions, each one embodying only the correlations between up to k random variables. For $k = 1$ this is the product distribution, $p_1(x_1, \dots, x_n) = p(x_1) \dots p(x_n)$, and for $k = n$ it is the full joint distribution. Using information geometry he shows that that the total correlation (also called multiinformation), $\sum_j H(X_j) - H(X_1, \dots, X_n)$, decomposes into a sum of Kullback-Leibler divergences between the intermediate distributions, forming an *orthogonal decomposition*.

We extend Amari's hierarchy into a lattice. This gives not one but a family of orthogonal decompositions, expressing the correlations between subsets of the variables in a more fine-grained way. The intermediate distributions may be readily computed using an iterative scaling algorithm. Applying this method to the example joint distributions given in [4], we show that it gives the numerical values the authors argued for intuitively. Finally, we comment on the relationship between this framework and the "partial information decomposition" proposed in [8].

We start with a joint distribution P over n random variables, $X_1, \dots, X_n$, which we call the *primary variables*. Sets of primary variables form *composite random variables*. We denote these by concatenating the primary variable names, e.g. X_1X_2. Sets of composite variables are our main object of interest; we term these *structures*. A structure is defined as a set of sets of the primary variables. We will only be concerned with structures in which (i) each primary variable appears at least once; and (ii) no composite variable in the structure is a subset of another. We denote structures by joining their members with $\langle \cdot \rangle$, e.g. $X_1X_2{\cdot}X_3$. For each structure $\mathcal{U}$ we will form a probability distribution $P_\mathcal{U}$, over the same sample space as P which captures precisely the correlations that follow the structure. It has the same marginal distribution as P for each composite variable $U \in \mathcal{U}$, i.e. $P(U) = P_\mathcal{U}(U)$, but lacks all other correlations. We define this by $P_\mathcal{U} := \operatorname{argmin}_Q D(Q \| P_0)$, subject to the constraint that $q(u) = p(u)$ for each u an outcome of $U \in \mathcal{U}$ and with P_0 the product distribution $p_o(X_1 \dots X_n) = p(X_1) \dots p(X_n)$. We write $\mathcal{U}$ in place of $P_\mathcal{U}$ when this does not create ambiguity.

A partial order can be defined on the set of structures. For structures $\mathcal{U}, \mathcal{V}$ we say that $\mathcal{U} \leq \mathcal{V}$ iff $\forall U \in \mathcal{U} : \exists V \in \mathcal{V} : U \subseteq V$. This is closely related to, but different from, the partial order defined in [8]. Our first result is that, if $\mathcal{U} \leq \mathcal{V} \leq \mathcal{W}$ according to this partial order, then $D(\mathcal{W} \| \mathcal{U}) = D(\mathcal{V} \| \mathcal{U}) + D(\mathcal{W} \| \mathcal{V})$. This follows from the results of [1], with each chain in the resulting lattice being an e-flat hierarchical structure in the terminology of that paper. In contrast to [8] and related approaches, here non-negative information quantities are associated with the *edges* rather than the nodes of this lattice, since they arise as Kullback-Leibler divergences *between* the distributions associated with the nodes.

Any individual structure may be seen as a hypergraph, with the primary variables as vertices and the structure's composite variables as edges. Our second result is that if this hypergraph is acyclic (in the

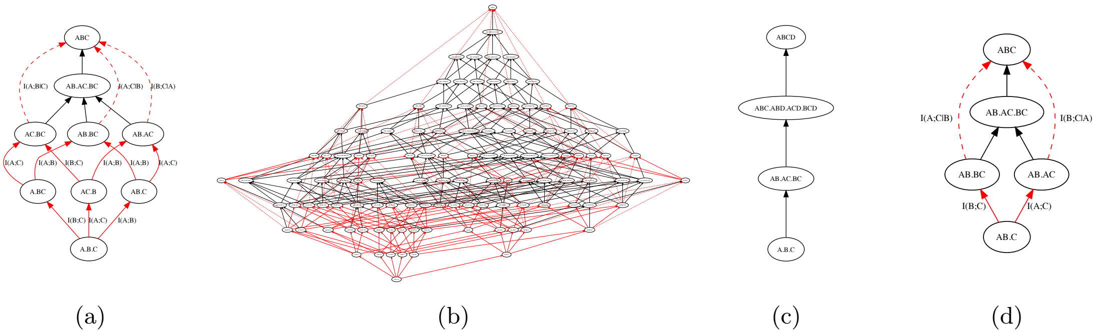

Figure 1: (a,b) Our lattices for $n = 3$ and $n = 4$. Solid lines show the Hasse diagram for the full lattice. Superimposed on this in red is the Hasse diagram for the lattice restricted to acyclic structures. (Dashed — red — lines indicate edges that are present in the acyclic, but not the full diagram) Each edge has an amount of information associated with it, given by the Kullback-Leibler divergence between the nodes; the orthogonality property means that these compose additively. Red edges can be expressed in terms of regular mutual informations (possibly conditional), whereas black edges include new terms that decompose multivariate interactions more finely. (c) Amari's hierarchy for $n = 4$, which is a subset of our lattice, under the same ordering. (d) a sublattice of (a), showing the decomposition of $I(AB; C)$.

sense of α-acyclicity in database theory, e.g. [3]) then $P_{\mathcal{U}}$ may be found analytically; and for two acyclic structures $\mathcal{U} \leq \mathcal{V}$, we can express $D(\mathcal{V}\|\mathcal{U})$ as a sum of conditional mutual information terms between the composite variables. Because of this, we can see that some of the edges in our lattice are traditional (conditional) mutual information terms, while others are new quantities that generalise Amari's hierarchy, splitting up multivariate interactions into more fine-grained terms. (See Fig. 1)

This allows us to give decompositions of any (conditional) mutual information term. For example, Fig. 1d shows the sublattice of nodes between $AB{\cdot}C$ and ABC. We have that $D(ABC\|AB{\cdot}C) = I(AB; C)$, so we can read this diagram as saying (for example) that $I(AB; C) = I(A; C) + D(AB{\cdot}AC{\cdot}BC\|AB{\cdot}AC) + D(ABC\|AB{\cdot}AC{\cdot}BC)$. This may be interpreted as the mutual information between A and C, plus the additional information in the pairwise correlations between B and C, plus the information in triadic interactions between all three variables.

Compare this to the approach to multivariate information from [8] (see also [2]), who also sought a decomposition of $I(AB; C)$. However, they specifically interpreted A and B as predictors of C, rather than allowing for a possibly more symmetric treatment of the variables, as in e.g. [7]. Our framework decomposes $I(AB; C)$ in a different way. While it does not give a non-negative decomposition into the same four terms, it has the advantage that any mutual information, in any number of variables, may be decomposed in a similar way. The connection to Amari's hierarchy means that the decomposition can be interpreted explicitly in terms of multivariate interactions, addressing the questions raised in [4].

References

[1] S. Amari. Information geometry on hierarchy of probability distributions. *IEEE Trans. Information Theory*, 2001.

[2] N. Bertschinger, J. Rauh, E. Olbrich, J. Jost, and N. Ay. Quantifying Unique Information. *Entropy*, 16(4):2161–2183, 2014.

[3] R. Fagin. Degrees of acyclicity for hypergraphs and relational database schemes. *Journal of the ACM*, 30(3):514–550, 1983.

[4] R. G. James and J. P. Crutchfield. Multivariate Dependence beyond Shannon Information. *Entropy*, 19(10):531, 2017.

[5] T. Kahle, E. Olbrich, J. Jost, and N. Ay. Complexity Measures from Interaction Structures. *arXiv.org*, (2):517, 2008.

[6] P. Perrone and N. Ay. Hierarchical Quantification of Synergy in Channels. *Frontiers in Robotics and AI*, 2:1701, 2016.

[7] F. Rosas, V. Ntranos, C. Ellison, S. Pollin, and M. Verhelst. Understanding Interdependency Through Complex Information Sharing. *Entropy*, 18(2):38, 2016.

[8] P. L. Williams and R. D. Beer. Nonnegative Decomposition of Multivariate Information. *arXiv.org*, 2010.

Acknowledgement. This work was supported by the ELSI Origins Network (EON), which is supported by a grant from the John Templeton Foundation.