Information Decomposition Based On Common Information

Pradeep Kr. Banerjee (pradeep@mis.mpg.de)
Max Planck Institute for Mathematics in the Sciences, Inselstr. 22, Leipzig, Germany

MAX-PLANCK-GESELLSCHAFT

Partial Information (PI) Decomposition

Consider three random variables (RVs) X_1, X_2 and Y taking values in finite alphabets $\mathcal{X}_1$, $\mathcal{X}_2$ and $\mathcal{Y}$ resp. The total mutual information that a pair of *predictor* RVs (X_1,X_2) convey about a *target* RV Y can have aspects of

- redundant information – conveyed *identically* by both X_1 and X_2, denoted $I_\cap(\{X_1,X_2\};Y)$,
- unique information – conveyed *exclusively* by either X_1 or X_2, denoted resp., $UI(\{X_1\};Y)$ and $UI(\{X_2\};Y)$,
- synergistic information – conveyed *jointly* by X_1 and X_2 that is not available from either alone, denoted $SI(\{X_1X_2\};Y)$.

The equations governing such a partial information (PI) decomposition are:

$$I(X_1X_2;Y) = \underbrace{I_\cap(\{X_1,X_2\};Y)}_{\text{redundant}} + \underbrace{SI(\{X_1X_2\};Y)}_{\text{synergistic}} + \underbrace{UI(\{X_1\};Y)}_{\text{unique } (X_1 \text{ wrt } X_2)} + \underbrace{UI(\{X_2\};Y)}_{\text{unique } (X_2 \text{ wrt } X_1)}$$

$$I(X_i;Y) = I_\cap(\{X_1,X_2\};Y) + UI(\{X_i\};Y), i = 1,2 \qquad (1)$$

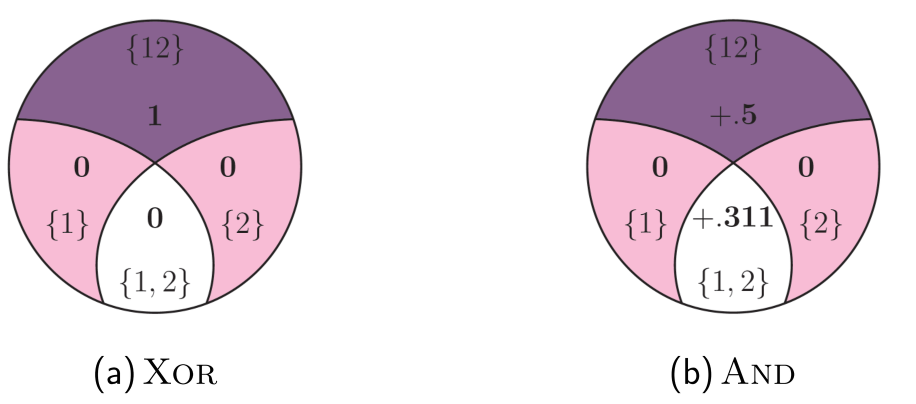

Figure 1: PI-diagrams showing the decomposition of $I(X_1X_2;Y)$ for some canonical examples. {1,2} denotes the redundant information $I_\cap(\{X_1,X_2\};Y)$; {1} and {2} denote, resp. $UI(\{X_1\};Y)$ and $UI(\{X_2\};Y)$; {12} denotes $SI(\{X_1X_2\};Y)$. X_1 and X_2 are binary, independent and uniformly distributed.
(a) $Y = \mathrm{XOR}(X_1,X_2)$ and the pmf $p_{X_1X_2Y}$ is such that $p_{(000)} = p_{(011)} = p_{(101)} = p_{(110)} = \frac{1}{4}$. The joint RV X_1X_2 fully specifies Y, i.e., $I(X_1X_2;Y) = 1$ whereas the singletons X_1 and X_2 specify nothing, i.e., $I(X_i;Y) = 0$, $i = 1,2$. Xor is an instance of a purely synergistic mechanism.
(b) $Y = \mathrm{AND}(X_1,X_2)$ and $p_{X_1X_2Y}$ is such that $p_{(000)} = p_{(010)} = p_{(100)} = p_{(111)} = \frac{1}{4}$. Note $X_1 \perp X_2$; however if either $X_1 = 0$ or $X_2 = 0$, then both X_1 and X_2 can exclude the possibility of $Y = 1$ with probability of agreement one; $I_\cap(\{X_1,X_2\};Y) \geq 0$ [2]-[6]. There is no unique information since the marginal distributions of the pairs (X_1,Y) and (X_2,Y) are identical and $\mathcal{X}_1 = \mathcal{X}_2$ [4].

The problem:

- Define a measure of redundant information, $I_\cap$ that yields a nonnegative decomposition of $I(X_1X_2;Y)$ per (1).
- Explore the relationship between redundant information and the more familiar notions of common information due to Gács-Körner and Wyner [8].

Earlier work: PI lattice [1]; Information-geometric approaches [2],[3]; Operational interpretation of unique information [4]-[6]; Common information-based measures [7].

Desirable properties of $I_\cap$

(S) Weak symmetry: $I_\cap(\{X_1,X_2\};Y)$ is invariant under reordering of the X_i's.
(I) Self-redundancy: $I_\cap(\{X_1\};Y) = I(X_1;Y)$.
(M) Monotonicity: $I_\cap(\{X_1,X_2\};Y) \leq I_\cap(\{X_1\};Y)$ with equality if $X_1 \subseteq X_2$.
(LN) Local Nonnegativity: For a given measure $I_\cap$, the derived partial information functions UI and SI are nonnegative.

Bounds on $I_\cap$

- Coinformation

$$I_{Co}(X_1;X_2;Y) := I(X_1;X_2) - I(X_1;X_2|Y) = I_\cap(\{X_1,X_2\};Y) - SI(\{X_1X_2\};Y)$$

- If $X_1 - X_2 - Y$, then $I_\cap(\{X_1,X_2\};Y) = I(X_1;Y)$
- If $X_2 - X_1 - Y$, then $I_\cap(\{X_1,X_2\};Y) = I(X_2;Y)$
- If $X_1 - X_2 - Y$ and $X_2 - X_1 - Y$, then $I_\cap(\{X_1,X_2\} = I(X_1X_2;Y)$
- If $X_1 \perp Y$ or $X_2 \perp Y$, then $I_\cap(\{X_1,X_2\};Y) = 0$
- If $X_1 - Y - X_2$, then $I_\cap(\{X_1,X_2\};Y) \geq I(X_1;X_2)$
- If $X_1 \perp X_2$ and $X_1 - Y - X_2$, then $I_{Co}(X_1;X_2;Y) = 0$

Common Information

Suppose $X = (X',Q)$ and $Y = (Y',Q)$ where X',Y',Q are independent. Intuitively, the *common RV* of X and Y (denoted $X \wedge Y$) is Q and a natural measure of *common information* (CI) of X and Y is $H(Q)$. Can extend this to arbitrary (X,Y) in a couple of ways [8], see Fig. 2(a):

- [Gács-Körner] Find the "largest" RV Q that is determined by X alone as well as by Y alone (w.p. 1); exploit the combinatorial structure of the distribution p_{XY}.

$$C_{GK}(X;Y) := \max_{\substack{p_{Q|XY}:\\ H(Q|X)=H(Q|Y)=0}} H(Q) = \max_{\substack{p_{Q|XY}:\\ Q-X-Y,\ Q-Y-X}} I(XY;Q),\ |\mathcal{Q}| \leq |\mathcal{X}||\mathcal{Y}| + 2$$

- [Wyner] Find the "smallest" RV Q such that conditioned on Q there is no residual mutual information.

$$C_W(X;Y) := \min_{\substack{p_{Q|XY}:\\ X-Q-Y}} I(XY;Q),\ |\mathcal{Q}| \leq |\mathcal{X}||\mathcal{Y}| + 2$$

- $C_{GK}(X;Y) \leq I(X;Y) \leq C_W(X;Y)$ with equality iff there exists a pmf $p_{Q|XY}$ such that the Markov chains $X - Q - Y$, $Q - X - Y$, $Q - Y - X$ hold [8].

Common Information-based Measures of $I_\cap$

Three candidate measures to assess how well the redundancy that X_1 and X_2 share about Y can be captured by a RV:

$$I_\cap^{GK}(\{X_1,X_2\};Y) := \max_{\substack{p_{Q|X_1X_2Y}:\\ H(Q|X_1)=H(Q|X_2)=0}} I(Q;Y) = I(X_1 \wedge X_2;Y) \qquad (2)$$

$$I_\cap^{W}(\{X_1,X_2\};Y) := \min_{\substack{p_{Q|X_1X_2Y}:\\ X_i-Q-Y,\ i=1,2}} I(Q;Y) \qquad (3)$$

$$I_\cap(\{X_1,X_2\};Y) := \max_{\substack{p_{Q|X_1X_2Y}:\\ Q-X_i-Y,\ i=1,2}} I(Q;Y) \qquad (4)$$

where $|\mathcal{Q}| \leq |\mathcal{X}_1||\mathcal{X}_2||\mathcal{Y}| + 2$.

- $I_\cap^{GK}$: maximum mutual information $I(Q:Y)$ that some RV Q conveys about Y, subject to Q being a function of each of the X_i's, $i = 1,2$; $I_\cap^{GK}$ violates **(LN)** since the supermodularity law does not hold for the Gács-Körner CI in general [7].
- $I_\cap^{W}$: monotonically nondecreasing in the number of X_i's, i.e., $I_\cap^{W}$ violates **(M)**.
- $I_\cap$: if Q specifies the optimal redundant RV, then conditioning on any predictor X_i, $i = 1,2$, should remove all the redundant information about Y [7]; $I_\cap$ violates **(LN)**:
  - If $X_1 \perp X_2$, then $I_\cap(\{X_1,X_2\};Y) = 0$; see Fig. 2(b).
  - If $X_1 - Y - X_2$, then $I_\cap(\{X_1,X_2\};Y) \leq I(X_1;X_2)$. The derived PI function $SI(\{X_1X_2\};Y) \leq 0$; see Fig. 2(c).
  - Let $\mathcal{Y} = \mathcal{X}_1 \times \mathcal{X}_2$ and $Y = X_1X_2$. Then $I_\cap(\{X_1,X_2\};Y) = C_{GK}(X_1;X_2) \leq I(X_1;X_2)$.

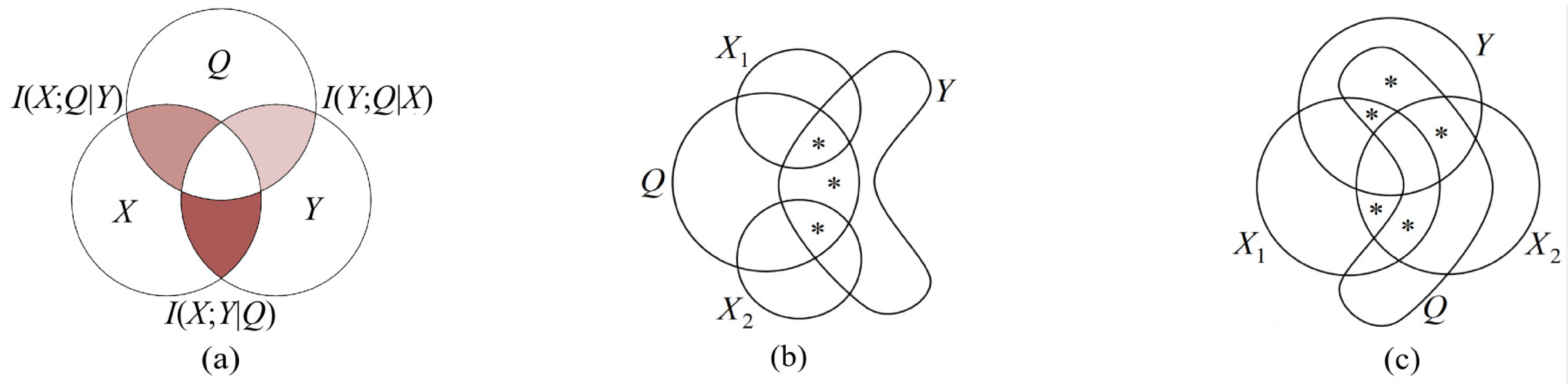

Figure 2: For finite RVs, there is a one-to-one correspondence between Shannon's information measures (I-measure) and a signed measure μ^* over sets. (a) The generic I-diagram for RVs X,Y, and Q. Given a RV X, we use X to also label the corresponding set in the I-diagram. (b) Denote the I-Measure of RVs (Q,X_1,X_2,Y) by μ^*. The atoms on which μ^* vanishes when the Markov chains $Q - X_i - Y$, $i = 1,2$ hold and $X_1 \perp X_2$ are marked by an asterisk; $\mu^*(Q \cap Y) = 0$. (c) The atoms on which μ^* vanishes when the Markov chains $Q - X_i - Y$, $i = 1,2$ and $X_1 - Y - X_2$ hold are marked by an asterisk; $\mu^*(X_1 \cap X_2 \cap Y) = \mu^*(X_1 \cap X_2) \geq 0$ and $\mu^*(Q \cap Y) \leq \mu^*(X_1 \cap X_2)$. *Note*: The I-diagrams in (b) and (c) are valid diagrams since the sets Q,X_1,X_2,Y intersect each other generically and the region representing the set Q splits each atom into two smaller ones.

- Conclusion: For independent predictor RVs when any nonvanishing redundancy can be attributed solely to a mechanistic dependence between the target and the predictors, common information-based measures of redundant information cannot induce a nonnegative PI decomposition.

References

[1] P. L. Williams and R. D. Beer (2010), Nonnegative decomposition of multivariate information, *arXiv:1004.2515*.
[2] M. Harder, C. Salge, and D. Polani, "Bivariate measure of redundant information," *Physical Review E*, vol. 87, no. 1, p. 012130, 2013.
[3] N. Bertschinger, J. Rauh, E. Olbrich, and J. Jost, "Shared information–New insights and problems in decomposing information in complex systems," in *Proc. European Conf. on Complex Systems 2012 (ECCS'12)*, 2012, pp. 251-269.
[4] N. Bertschinger, J. Rauh, E. Olbrich, J. Jost, and N. Ay, "Quantifying unique information," *Entropy*, vol. 16, no. 4, pp. 2161-2183, 2014.
[5] J. Rauh, N. Bertschinger, E. Olbrich, and J. Jost, "Reconsidering unique information: Towards a multivariate information decomposition," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT 2014)*, 2014, pp. 2232-2236.
[6] V. Griffith and C. Koch, "Quantifying synergistic mutual information," in *Guided Self-Organization: Inception*, Springer, 2014, vol. 9, Emergence, Complexity and Computation Series, pp. 159-190.
[7] V. Griffith, T. Ho (2015), Quantifying redundant information in predicting a target random variable, *Entropy*, vol. 17, no. 7, pp. 4644-4653.
[8] R. Ahlswede and J. Körner (1974), "On common information and related characteristics of correlated information sources," in *7th Prague Conference on Information Theory*.