Measures of Complexity: Information Based

Ryan G. James

Complexity Sciences Center
UC DAVIS
UNIVERSITY OF CALIFORNIA

June 20, 2018

Information-Based Complexity

Why information based measures?

- computable
- estimable
- diverse
- comparable
- interpretable(ish)

Shannon Measures [1, 2, 3]

entropy	$\mathrm{H}[X]$	$= -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x)$
joint entropy	$\mathrm{H}[X, Y]$	$= -\sum_{\substack{x \in \mathcal{X} \\ y \in \mathcal{Y}}} p(x, y) \log_2 p(x, y)$
conditional entropy	$\mathrm{H}[X \vert Y]$	$= \mathrm{H}[X, Y] - \mathrm{H}[Y]$
mutual information	$\mathrm{I}[X : Y]$	$= \mathrm{H}[X] + \mathrm{H}[Y] - \mathrm{H}[X, Y]$
cond. mutual information	$\mathrm{I}[X : Y \vert Z]$	$= \mathrm{H}[X \vert Z] - \mathrm{H}[X \vert Y, Z]$

Algorithmic to Shannon Information Theory

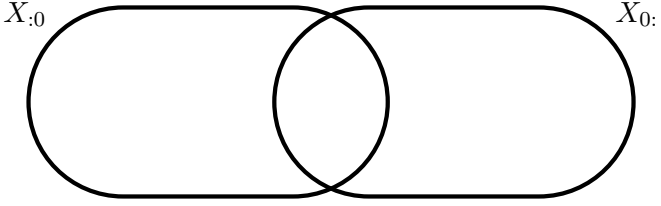
Yesterday it was stated that Kolmogorov Complexity quantifies randomness. Let's make that more concrete. Consider a stationary information source X:

$$\left\langle \frac{\mathrm{K}(x_{0:\ell})}{\ell} \right\rangle \to \frac{\mathrm{H}\left[X_{0:\ell}\right]}{\ell} \to h_\mu = \mathrm{H}\left[X_0 | X_{:0}\right]$$

Excess Entropy [5]

How much does everything that has happened have to do with everything that will?:

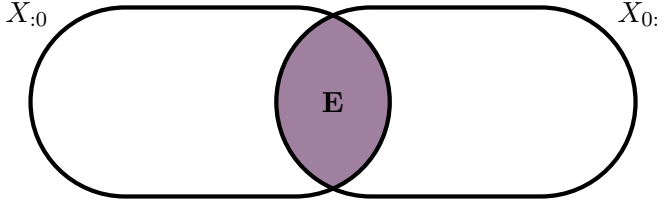
$$\mathbf{E} = \mathrm{I}\left[X_{:0} : X_{0:}\right]$$



Excess Entropy [5]

How much does everything that has happened have to do with everything that will?:

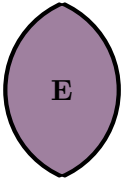
$$\mathbf{E} = \mathrm{I}\left[X_{:0} : X_{0:}\right]$$



Excess Entropy [5]

How much does everything that has happened have to do with everything that will?:

$$\mathbf{E} = \mathrm{I}\left[X_{:0} : X_{0:}\right]$$

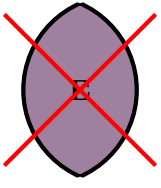


We would like this.

Excess Entropy [5]

How much does everything that has happened have to do with everything that will?:

$$\mathbf{E} = \mathrm{I}\left[X_{:0} : X_{0:}\right]$$

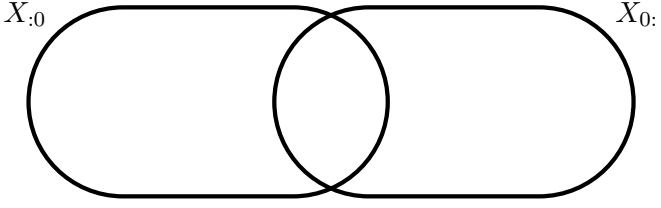


We would like this.
But it doesn't exist [4].

Statistical Complexity [6]

Statistical Complexity is the minimal sufficient statistic of the past about the future:

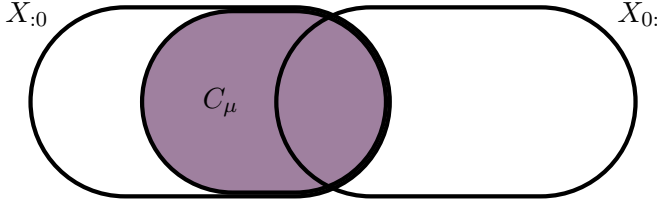
$$C_\mu = X_{:0} \searrow X_{0:}$$



Statistical Complexity [6]

Statistical Complexity is the minimal sufficient statistic of the past about the future:

$$C_\mu = X_{:0} \searrow X_{0:}$$



It is the amount of information about the past one must retain in order to optimally predict the future.

C_μ is Sophistication

Statistical Complexity is the subextensive part of the average Kolmogorov Complexity:

$$\langle \mathrm{K}(x_{0:\ell}) \rangle = \underbrace{\langle \mathrm{soph_c}(x_{0:\ell}) \rangle}_{C_\mu} + \underbrace{\langle \text{the data part} \rangle}_{\ell h_\mu}$$

C_μ is Sophistication

Statistical Complexity is the subextensive part of the average Kolmogorov Complexity:

$$\langle \mathrm{K}(x_{0:\ell}) \rangle = \underbrace{\langle \mathrm{soph_c}(x_{0:\ell}) \rangle}_{C_\mu} + \underbrace{\langle \text{the data part} \rangle}_{\ell h_\mu}$$

$$C_\mu = \mathrm{H}\left[X_{:0} \searrow X_{0:}\right]$$
$$\mathcal{S} \sim X_{:0} \searrow X_{0:}$$

C_μ is Sophistication

Statistical Complexity is the subextensive part of the average Kolmogorov Complexity:

$$\langle \mathrm{K}(x_{0:\ell}) \rangle = \underbrace{\langle \mathrm{soph_c}(x_{0:\ell}) \rangle}_{C_\mu} + \underbrace{\langle \text{the data part} \rangle}_{\ell h_\mu}$$

$$C_\mu = \mathrm{H}\left[X_{:0} \searrow X_{0:}\right]$$
$$\mathcal{S} \sim X_{:0} \searrow X_{0:} \sim \Pr\left(X_{0:} | x_{:0}\right)$$

Building Models from Equivalence Classes

$\mathcal{S} \sim \Pr\left(X_{0:}|x_{:0}\right)$ *partitions* the pasts.

Building Models from Equivalence Classes

$\mathcal{S} \sim \Pr\left(X_{0:}|x_{:0}\right)$ *partitions* the pasts.

But $x_{:0} + x_1 = x_{:1}$ is *another past* due to stationarity.

Building Models from Equivalence Classes

$\mathcal{S} \sim \Pr\left(X_{0:}|x_{:0}\right)$ *partitions* the pasts.

But $x_{:0} + x_1 = x_{:1}$ is *another past* due to stationarity.

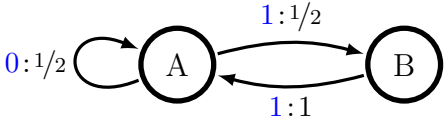
Therefore, the minimal sufficient statistic provides a mapping between pasts!

Building Models from Equivalence Classes

$\mathcal{S} \sim \Pr\left(X_{0:}|x_{:0}\right)$ *partitions* the pasts.

But $x_{:0} + x_1 = x_{:1}$ is *another past* due to stationarity.

Therefore, the minimal sufficient statistic provides a mapping between pasts!



Are $\mathbf{E}$ and C_μ Good Measures of Complexity?

$\mathbf{E}$ and C_μ are both intuitive and interpretable.

But consider long periodic processes:

- no randomness
- large $\mathbf{E}$
- large C_μ

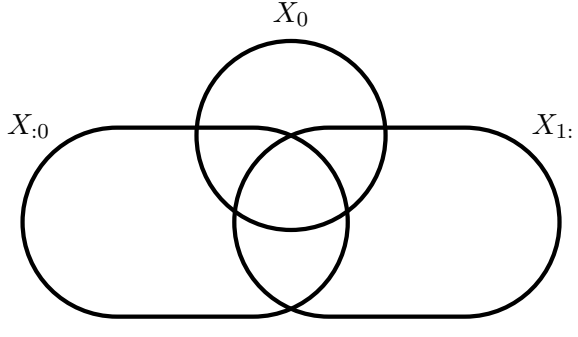
Are $\mathbf{E}$ and C_μ Good Measures of Complexity?

$\mathbf{E}$ and C_μ are both intuitive and interpretable.

But consider long periodic processes:

- no randomness
- large $\mathbf{E}$
- large C_μ
- humpology violation!

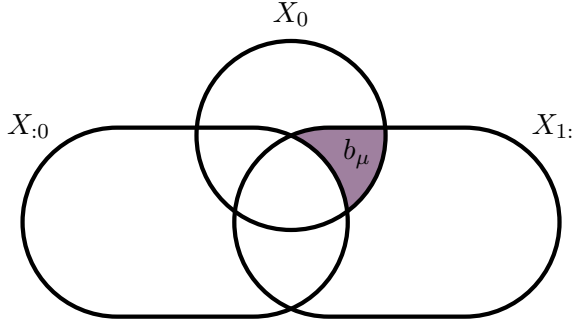
Predictive Information Rate [7, 8]



Predictive Information Rate [7, 8]

How much of the generated randomness is relevant for the future:

$$b_\mu = \mathrm{I}\left[X_0 : X_{0:} | X_{:0}\right]$$



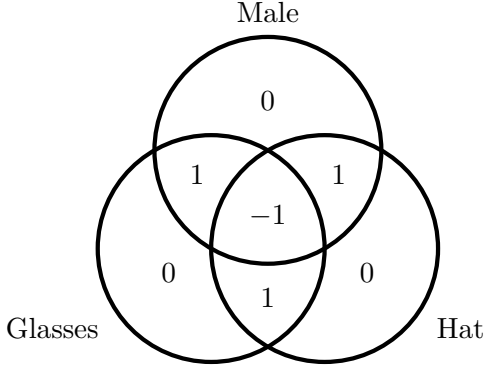
Distribution Exemplar: Definition

Guess Who?

Male	Glasses	Hat	Pr
✗	✗	✗	1/4
✗	✓	✓	1/4
✓	✗	✓	1/4
✓	✓	✗	1/4

Distribution Exemplar: Information

Information Diagrams [2, 1, 3] summarize all Shannon measures of the distribution:



Guess Who?

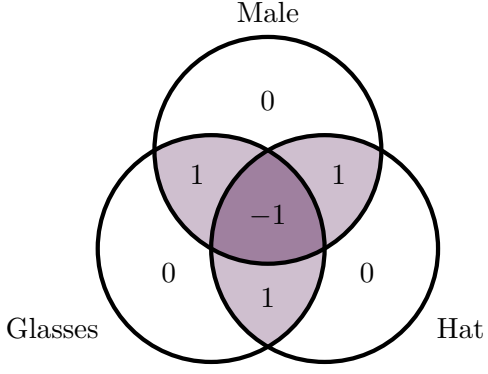
Male	Glasses	Hat	Pr
✗	✗	✗	1/4
✗	✓	✓	1/4
✓	✗	✓	1/4
✓	✓	✗	1/4

Total Correlation [9]

$$\mathrm{I}[X:Y] = \mathrm{H}[X] + \mathrm{H}[Y] - \mathrm{H}[X,Y]$$

$$\Downarrow$$

$$\mathrm{T}[X_0:\ldots:X_n] = \sum_i \mathrm{H}[X_i] - \mathrm{H}[X_0,\ldots,X_n]$$



Guess Who?			
Male	Glasses	Hat	Pr
✗	✗	✗	1/4
✗	✓	✓	1/4
✓	✗	✓	1/4
✓	✓	✗	1/4

$$\mathrm{T}[\text{Male}:\text{Glasses}:\text{Hat}] = 1\,\text{bit}$$

Total Correlation [9]

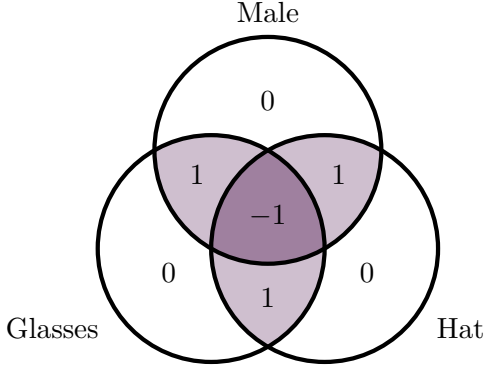
$$\mathrm{I}[X:Y] = \mathrm{D}_{\mathrm{KL}}[p(X,Y) \mid\mid p(X)p(Y)]$$

$$\Downarrow$$

$$\mathrm{T}[X_0:\ldots:X_n] = \mathrm{D}_{\mathrm{KL}}[p(X_0,\ldots,X_n) \mid\mid p(X_0)\ldots p(X_n)]$$

Guess Who?

Male	Glasses	Hat	Pr
✗	✗	✗	1/4
✗	✓	✓	1/4
✓	✗	✓	1/4
✓	✓	✗	1/4



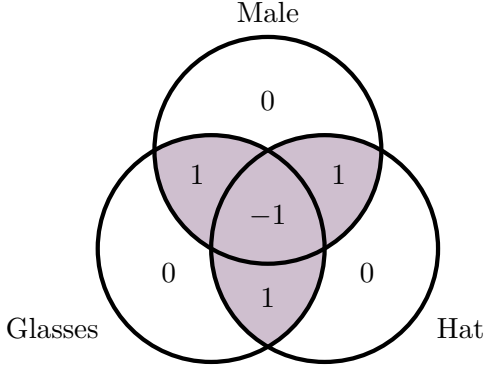
$$\mathrm{T}[\mathrm{Male}:\mathrm{Glasses}:\mathrm{Hat}] = 1\,\mathrm{bit}$$

Dual Total Correlation [10]

$$\mathrm{I}[X:Y] = \mathrm{H}[X,Y] - \mathrm{H}[X|Y] - \mathrm{H}[Y|X]$$

$$\Downarrow$$

$$\mathrm{B}[X_0 : \ldots : X_n] = \mathrm{H}[X_0, \ldots, X_n] - \sum_i \mathrm{H}\left[X_i | X_{[n]\setminus\{i\}}\right]$$



Guess Who?

Male	Glasses	Hat	Pr
✗	✗	✗	1/4
✗	✓	✓	1/4
✓	✗	✓	1/4
✓	✓	✗	1/4

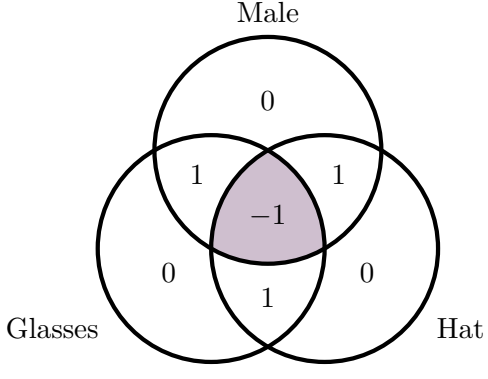
$$\mathrm{B}[\text{Male} : \text{Glasses} : \text{Hat}] = 2\,\text{bit}$$

Coinformation [11]

$$\mathrm{I}[X:Y] = \mathrm{H}[X] + \mathrm{H}[Y] - \mathrm{H}[X,Y]$$

$$\Downarrow$$

$$\mathrm{I}[X_0:\ldots:X_n] = \sum_{S \in \mathcal{P}(\{X_0,\ldots,X_n\})} (-1)^{|S|+1}\,\mathrm{H}[S]$$



Guess Who?			
Male	Glasses	Hat	Pr
✗	✗	✗	1/4
✗	✓	✓	1/4
✓	✗	✓	1/4
✓	✓	✗	1/4

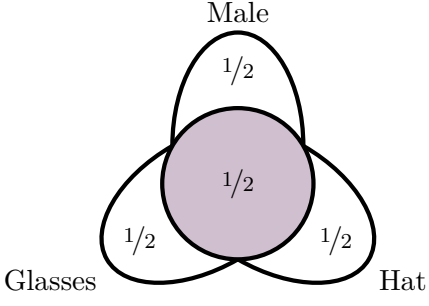
$$\mathrm{I}[\mathrm{Male}:\mathrm{Glasses}:\mathrm{Hat}] = -1\,\mathrm{bit}$$

CAEKL Mutual Information [12]

$$\mathrm{I}[X:Y] = \arg\max_{\gamma} \{\mathrm{H}[X] - \gamma + \mathrm{H}[Y] - \gamma = \mathrm{H}[X,Y] - \gamma\}$$

$$\Downarrow$$

$$\mathrm{J}[X_0 : \ldots : X_n] = \arg\max_{\gamma} \left\{ \sum_i \mathrm{H}[X_i] - \gamma = \mathrm{H}[X_0, \ldots, X_n] - \gamma \right\}$$



Guess Who?

Male	Glasses	Hat	Pr
✗	✗	✗	1/4
✗	✓	✓	1/4
✓	✗	✓	1/4
✓	✓	✗	1/4

$$\mathrm{J}[\mathrm{Male}:\mathrm{Glasses}:\mathrm{Hat}] = 1/2 \text{ bit}$$

These Measure Are Not Enough [13]

Distributions

Dyadic

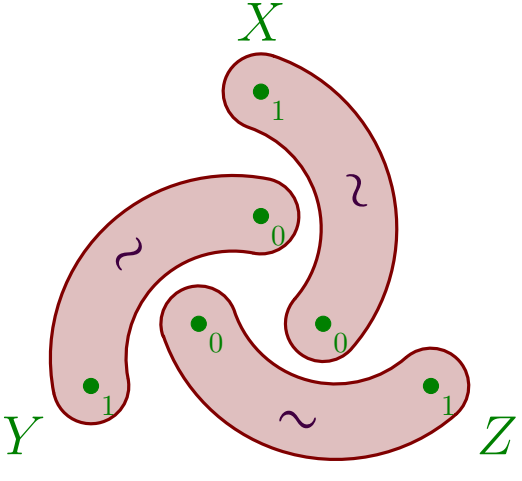
X	Y	Z	
X_0X_1	Y_0Y_1	Z_0Z_1	Pr
0 0	0 0	0 0	1/8
0 0	1 0	0 1	1/8
0 1	0 0	1 0	1/8
0 1	1 0	1 1	1/8
1 0	0 1	0 0	1/8
1 0	1 1	0 1	1/8
1 1	0 1	1 0	1/8
1 1	1 1	1 1	1/8

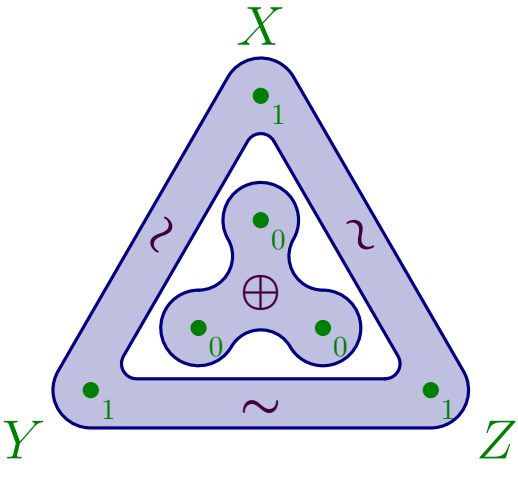
Triadic

X	Y	Z	
X_0X_1	Y_0Y_1	Z_0Z_1	Pr
0 0	0 0	0 0	1/8
0 1	0 1	0 1	1/8
0 0	1 0	1 0	1/8
0 1	1 1	1 1	1/8
1 0	0 0	1 0	1/8
1 1	0 1	1 1	1/8
1 0	1 0	0 0	1/8
1 1	1 1	0 1	1/8

These Measure Are Not Enough [13]

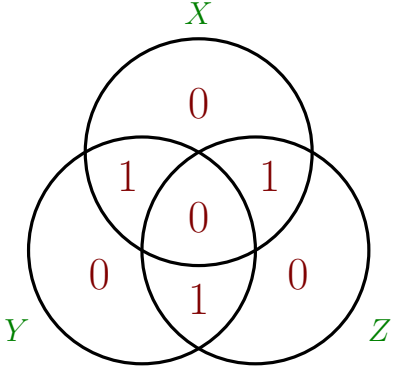
Representations

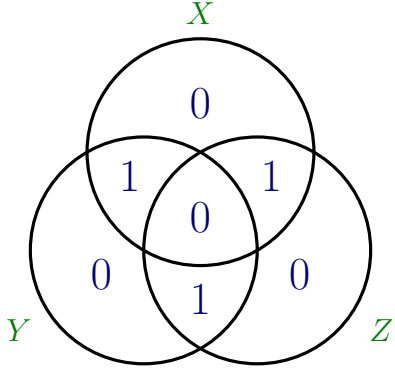




These Measure Are Not Enough [13]

Informations





Decomposing Multivariate Information [14]

Consider a black box with two inputs (X_0, X_1) and one output (Y).
Let's assume that their mutual information can be decomposed into semantically meaningful, non-negative components:

$$\mathrm{I}\left[X_0, X_1 : Y\right] =$$

Decomposing Multivariate Information [14]

Consider a black box with two inputs (X_0, X_1) and one output (Y).
Let's assume that their mutual information can be decomposed into semantically meaningful, non-negative components:

$$\mathrm{I}[X_0, X_1 : Y] = \quad \mathrm{I}_\cap[0 \cdot 1 \to Y] \qquad \textit{redundancy}$$

Decomposing Multivariate Information [14]

Consider a black box with two inputs (X_0, X_1) and one output (Y).
Let's assume that their mutual information can be decomposed into semantically meaningful, non-negative components:

$$\begin{aligned} \mathrm{I}\left[X_0, X_1 : Y\right] = \quad & \mathrm{I}_\cap\left[0 \cdot 1 \to Y\right] && \textit{redundancy} \\ & + \mathrm{I}_\cap\left[0 \to Y\right] && \textit{unique from } X_0 \end{aligned}$$

Decomposing Multivariate Information [14]

Consider a black box with two inputs (X_0, X_1) and one output (Y).
Let's assume that their mutual information can be decomposed into semantically meaningful, non-negative components:

$$\begin{aligned} \mathrm{I}[X_0, X_1 : Y] = \quad & \mathrm{I}_\cap[0 \cdot 1 \to Y] && \textit{redundancy} \\ & + \mathrm{I}_\cap[0 \to Y] && \textit{unique from } X_0 \\ & + \mathrm{I}_\cap[1 \to Y] && \textit{unique from } X_1 \end{aligned}$$

Decomposing Multivariate Information [14]

Consider a black box with two inputs (X_0, X_1) and one output (Y).
Let's assume that their mutual information can be decomposed into semantically meaningful, non-negative components:

$$
\begin{aligned}
\mathrm{I}\left[X_0, X_1 : Y\right] = \quad & \mathrm{I}_\cap\left[0 \cdot 1 \to Y\right] && \textit{redundancy} \\
& + \mathrm{I}_\cap\left[0 \to Y\right] && \textit{unique from } X_0 \\
& + \mathrm{I}_\cap\left[1 \to Y\right] && \textit{unique from } X_1 \\
& + \mathrm{I}_\cap\left[01 \to Y\right] && \textit{synergy}
\end{aligned}
$$

Decomposing Multivariate Information [14]

Consider a black box with two inputs (X_0, X_1) and one output (Y).
Let's assume that their mutual information can be decomposed into semantically meaningful, non-negative components:

$$\begin{aligned} \mathrm{I}[X_0, X_1 : Y] = \quad & \mathrm{I}_\cap[0 \cdot 1 \to Y] && \textit{redundancy} \\ & + \mathrm{I}_\cap[0 \to Y] && \textit{unique from } X_0 \\ & + \mathrm{I}_\cap[1 \to Y] && \textit{unique from } X_1 \\ & + \mathrm{I}_\cap[01 \to Y] && \textit{synergy} \end{aligned}$$

$$\begin{aligned} \mathrm{I}[X_0 : Y] = \quad & \mathrm{I}_\cap[0 \cdot 1 \to Y] && \textit{redundancy} \\ & + \mathrm{I}_\cap[0 \to Y] && \textit{unique from } X_0 \end{aligned}$$

Decomposing Multivariate Information [14]

Consider a black box with two inputs (X_0, X_1) and one output (Y).
Let's assume that their mutual information can be decomposed into semantically meaningful, non-negative components:

$$\begin{aligned} \mathrm{I}[X_0, X_1 : Y] = \quad & \mathrm{I}_\cap[0 \cdot 1 \to Y] && \textit{redundancy} \\ & + \mathrm{I}_\cap[0 \to Y] && \textit{unique from } X_0 \\ & + \mathrm{I}_\cap[1 \to Y] && \textit{unique from } X_1 \\ & + \mathrm{I}_\cap[01 \to Y] && \textit{synergy} \end{aligned}$$

$$\begin{aligned} \mathrm{I}[X_0 : Y] = \quad & \mathrm{I}_\cap[0 \cdot 1 \to Y] && \textit{redundancy} \\ & + \mathrm{I}_\cap[0 \to Y] && \textit{unique from } X_0 \end{aligned}$$

$$\begin{aligned} \mathrm{I}[X_1 : Y] = \quad & \mathrm{I}_\cap[0 \cdot 1 \to Y] && \textit{redundancy} \\ & + \mathrm{I}_\cap[1 \to Y] && \textit{unique from } X_1 \end{aligned}$$

Summary

- **E** is interesting and interpretable, but doesn't correspond to a model
- C_μ is average Sophistication *and* the minimal sufficient statistic of $X_{:0}$ about $X_{0:}$
- b_μ quantifies how much of the generated randomness is "interesting"
- multivariate information measures access different aspects of distributions
- ... but they can be insensitive to qualitatively distinct distributions
- PID is a promising approach but computationally difficult

Other Neat Things

- common informations [4, 15, 16]
- pointwise information
- information bottleneck [17]
- secret key agreement [18]
- dependency decomposition [19]

Calculations

All calculations seen here were computed using the `dit` Python package:

R. G. James, C. J. Ellison, and J. P. Crutchfield. "dit: a Python package for discrete information theory". In: *The Journal of Open Source Software* 3.25 (2018), p. 738

References I

[1] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Second. New York: Wiley-Interscience, 2006, p. 776. ISBN: 0471241954.

[2] Raymond W. Yeung. *Information theory and network coding*. Springer, 2008.

[3] David J.C. MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[4] Peter Gács and János Körner. "Common information is far less than mutual information". In: *Problems of Control and Information Theory* 2.2 (1973), pp. 149–162.

[5] James P Crutchfield and David P Feldman. "Regularities unseen, randomness observed: Levels of entropy convergence". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 13.1 (2003), pp. 25–54.

[6] Cosma Rohilla Shalizi and James P Crutchfield. "Computational mechanics: Pattern and prediction, structure and simplicity". In: *Journal of statistical physics* 104.3-4 (2001), pp. 817–879.

References II

[7] Samer A Abdallah and Mark D Plumbley. "A measure of statistical complexity based on predictive information with application to finite spin systems". In: *Physics Letters A* 376.4 (2012), pp. 275–281.

[8] Ryan G James, Christopher J Ellison, and James P Crutchfield. "Anatomy of a bit: Information in a time series observation". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 21.3 (2011), p. 037109.

[9] S. Watanabe. "Information theoretical analysis of multivariate correlation". In: *IBM Journal of research and development* 4.1 (1960), pp. 66–82.

[10] H. Te Sun. "Multiple mutual informations and multiple interactions in frequency data". In: *Information and Control* 46 (1980), pp. 26–45.

[11] A. J. Bell. "The Co-information Lattice". In: *Proc. Fifth Intl. Workshop on Independent Component Analysis and Blind Signal Separation*. Ed. by S. Makino S. Amari A. Cichocki and N. Murata. Vol. ICA 2003. New York: Springer, 2003, pp. 921–926.

References III

[12] Chung Chan et al. “Multivariate Mutual Information Inspired by Secret-Key Agreement”. In: *Proceedings of the IEEE* 103.10 (2015), pp. 1883–1913.

[13] Ryan G James and James P Crutchfield. “Multivariate dependence beyond Shannon information”. In: *Entropy* 19.10 (2017), p. 531.

[14] Paul L Williams and Randall D Beer. “Nonnegative decomposition of multivariate information”. In: *arXiv preprint arXiv:1004.2515* (2010).

[15] A. D. Wyner. “The common information of two dependent random variables”. In: *Information Theory, IEEE Transactions on* 21.2 (1975), pp. 163–179.

[16] G. R. Kumar, C. T. Li, and A. El Gamal. “Exact common information”. In: *Information Theory (ISIT), 2014 IEEE International Symposium on*. IEEE. 2014, pp. 161–165.

[17] Naftali Tishby, Fernando C Pereira, and William Bialek. “The information bottleneck method”. In: *arXiv preprint physics/0004057* (2000).

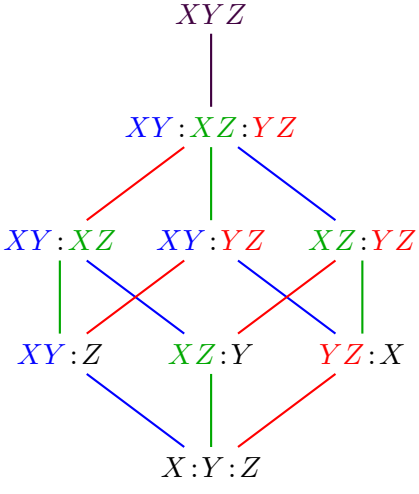
References IV

[18] Ueli Maurer and Stefan Wolf. “The intrinsic conditional mutual information and perfect secrecy”. In: *IEEE international symposium on information theory*. Citeseer. 1997, pp. 88–88.

[19] Ryan G. James, Jeffrey Emenheiser, and James P. Crutchfield. “Unique Information via Dependency Constraints”. In: *arXiv preprint arXiv:1709.06653* (2017).

[20] R. G. James, C. J. Ellison, and J. P. Crutchfield. “dit: a Python package for discrete information theory”. In: *The Journal of Open Source Software* 3.25 (2018), p. 738.

Dependency Decomposition



Dependency Decomposition