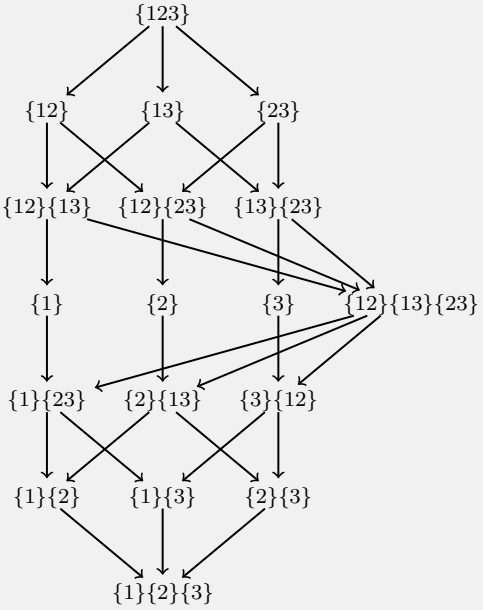
Pointwise Partial Information Decomposition Using Specificity and Ambiguity Lattices

IPCS Satellite @ CCS17 Cancun

Conor Finn (presenting)
Mikhail Prokopenko
Joseph Lizier

September 14, 2017



Unique, redundant and synergistic information

Consider three random variables S_1, S_2 and T

- Aim: predict T using S_1 and S_2
- Several types of information:
  1. **Unique information** $U(T : S_1 \setminus S_2)$
  2. **Redundant information** $R(T : S_1, S_2)$
  3. **Synergistic information** $C(T : S_1, S_2)$

UNQ

p	s_1	s_2	t
1/4	0	0	0
1/4	0	1	0
1/4	1	0	1
1/4	1	1	1

RDN

p	s_1	s_2	t
1/2	0	0	0
1/2	1	1	1

XOR

p	s_1	s_2	t
1/4	0	0	0
1/4	0	1	1
1/4	1	0	1
1/4	1	1	0

Information decomposition

In general, all types of information are present

- Mutual information captures

$$I(T;S_1) = R(T:S_1,S_2) + U(T:S_1 \backslash S_2)$$
$$I(T;S_2) = R(T:S_1,S_2) + U(T:S_2 \backslash S_1)$$

AND

p	s_1	s_2	t
1/4	0	0	0
1/4	0	1	1
1/4	1	0	1
1/4	1	1	1

- Joint mutual information captures

$$I(T;S_1S_2) = R(T:S_1,S_2) + U(T:S_1 \backslash S_2) + U(T:S_2 \backslash S_1) + C(T:S_1,S_2)$$

- Three equations with four unknowns

Partial Information Decomposition

- Axiomatic framework extending this decomposition to arbitrary number of source

Axioms (PID)

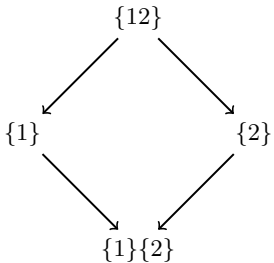
(1) *Symmetry*: $R(T : S_1, \ldots, S_n)$ is invariant under permutations of the S_i's

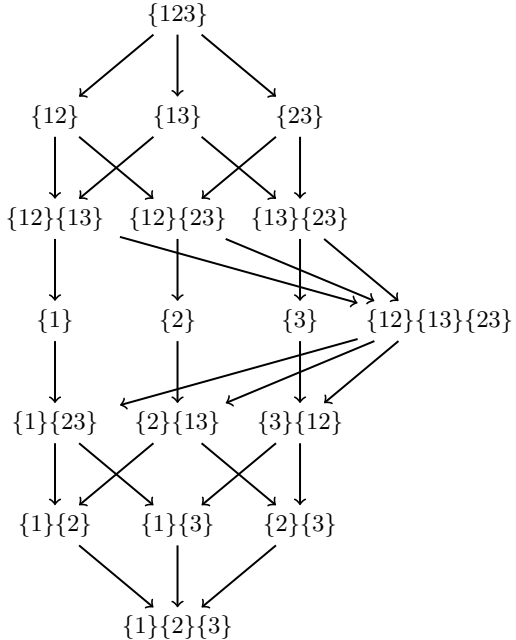
(2) *Monotonicity*: $R(T : S_1, \ldots, S_n) \leq R(T : S_1; \ldots; S_{n-1})$

(3) *Self-redundancy*: $R(T : S_i) = I(T; S_i)$

- Yields a **redundancy lattice**
- Still no accepted, compatible definition of unique, redundant, or synergistic information

Redundancy lattice





Pointwise information theory

- From four postulates, Fano (1961) derived the **pointwise** mutual information

$$i(x;y) = \log \frac{p(x,y)}{p(x)p(y)} \gtrless 0$$

- Corollaries: (average) mutual information, pointwise entropy and (Shannon) entropy

Pointwise information decomposition

- Pointwise decomposition for each realisation

$$
\begin{aligned}
i(t; s_1) &= r(t : s_1, s_2) + u(t : s_1 \setminus s_2) \\
i(t; s_2) &= r(t : s_1, s_2) + u(t : s_2 \setminus s_1) \\
i(t; s_1 s_2) &= r(t : s_1, s_2) + u(t : s_1 \setminus s_2) + u(t : s_2 \setminus s_1) + c(t : s_1, s_2)
\end{aligned}
$$

Pointwise information decomposition

- Pointwise decomposition for each realisation

$$
\begin{aligned}
i(t; s_1) &= r(t : s_1, s_2) + u(t : s_1 \setminus s_2) \\
i(t; s_2) &= r(t : s_1, s_2) + u(t : s_2 \setminus s_1) \\
i(t; s_1 s_2) &= r(t : s_1, s_2) + u(t : s_1 \setminus s_2) + u(t : s_2 \setminus s_1) + c(t : s_1, s_2)
\end{aligned}
$$

- Should be able to take the expectation over all realisations

$$
\begin{aligned}
R(T : S_1, S_2) &= \big\langle r(t : s_1, s_2) \big\rangle & U(T : S_1 \setminus S_2) &= \big\langle u(t : s_1 \setminus s_2) \big\rangle \\
C(T : S_1, S_2) &= \big\langle c(t : s_1, s_2) \big\rangle & U(T : S_2 \setminus S_1) &= \big\langle u(t : s_2 \setminus s_1) \big\rangle
\end{aligned}
$$

Pointwise information decomposition

- Pointwise decomposition for each realisation

$$
\begin{aligned}
i(t; s_1) &= r(t : s_1, s_2) + u(t : s_1 \setminus s_2) \\
i(t; s_2) &= r(t : s_1, s_2) + u(t : s_2 \setminus s_1) \\
i(t; s_1 s_2) &= r(t : s_1, s_2) + u(t : s_1 \setminus s_2) + u(t : s_2 \setminus s_1) + c(t : s_1, s_2)
\end{aligned}
$$

- Should be able to take the expectation over all realisations

$$
\begin{aligned}
R(T : S_1, S_2) &= \big\langle r(t : s_1, s_2) \big\rangle & U(T : S_1 \setminus S_2) &= \big\langle u(t : s_1 \setminus s_2) \big\rangle \\
C(T : S_1, S_2) &= \big\langle c(t : s_1, s_2) \big\rangle & U(T : S_2 \setminus S_1) &= \big\langle u(t : s_2 \setminus s_1) \big\rangle
\end{aligned}
$$

- This should recover the (average) information decomposition

$$
\begin{aligned}
I(T; S_1) &= R(T : S_1, S_2) + U(T : S_1 \setminus S_2) \\
I(T; S_2) &= R(T : S_1, S_2) + U(T : S_2 \setminus S_1) \\
I(T; S_1 S_2) &= R(T : S_1, S_2) + U(T : S_1 \setminus S_2) + U(T : S_2 \setminus S_1) + C(T : S_1, S_2)
\end{aligned}
$$

Motivation: PwUnq

- Consider PwUnq from Finn et al. (2017b)

p	s_1	s_2	t	
1/4	0	1	1	
1/4	1	0	1	
1/4	0	2	2	
1/4	2	0	2	
Expected values				

Motivation: PwUnq

- Consider PwUnq from Finn et al. (2017b)

p	s_1	s_2	t	$i(t;s_1)$	$i(t;s_2)$	$i(t;s_1s_2)$
1/4	0	1	1	0	1	1
1/4	1	0	1	1	0	1
1/4	0	2	2	0	1	1
1/4	2	0	2	1	0	1
Expected values				1/2	1/2	1

Motivation: PwUnq

- Consider PwUnq from Finn et al. (2017b)

p	s_1	s_2	t	$i(t;s_1)$	$i(t;s_2)$	$i(t;s_1s_2)$	r
1/4	0	1	1	0	1	1	0
1/4	1	0	1	1	0	1	0
1/4	0	2	2	0	1	1	0
1/4	2	0	2	1	0	1	0
Expected values				1/2	1/2	1	0

Motivation: PwUnq

- Consider PwUnq from Finn et al. (2017b)

p	s_1	s_2	t	$i(t;s_1)$	$i(t;s_2)$	$i(t;s_1s_2)$	r	u_1	u_2	c
1/4	0	1	1	0	1	1	0	0	1	0
1/4	1	0	1	1	0	1	0	1	0	0
1/4	0	2	2	0	1	1	0	0	1	0
1/4	2	0	2	1	0	1	0	1	0	0
Expected values				1/2	1/2	1	0	1/2	1/2	0

Motivation: PwUnq

- Consider PwUnq from Finn et al. (2017b)

p	s_1	s_2	t	$i(t;s_1)$	$i(t;s_2)$	$i(t;s_1s_2)$	r	u_1	u_2	c
1/4	0	1	1	0	1	1	0	0	1	0
1/4	1	0	1	1	0	1	0	1	0	0
1/4	0	2	2	0	1	1	0	0	1	0
1/4	2	0	2	1	0	1	0	1	0	0
Expected values				1/2	1/2	1	0	1/2	1/2	0

- According to $I_{\min}$ Williams and Beer (2010), $\widetilde{UI}$ of Bertschinger et al. (2014), S_{VK} of Griffith and Koch (2014) and I_{red} of Harder et al. (2013)

$$R = \langle r \rangle = 1/2 \text{ bit} \neq 0 \text{ bit}$$

Pointwise Partial Information Decomposition

Axioms (PPID)

(1) *Symmetry*: $r(t : s_1, \ldots, s_n)$ is invariant under permutations of the s_i's

(2) *Monotonicity*: $r(t : s_1, \ldots, s_n) \leq r(t : s_1; \ldots; s_{n-1})$

(3) *Self-redundancy*: $r(t : s_i) = i(t; s_i)$

- Problems:
  1. Pointwise mutual information is not non-negative
  2. Still no clear definition of redundant information

What is pointwise information?

- The surprise of the posterior compared to the surprise of the prior

Prior $p(t) \longrightarrow p(t|s_1)$ Posterior

What is pointwise information?

- The surprise of the posterior compared to the surprise of the prior

Prior $p(t) \longrightarrow p(t|s_1)$ Posterior

- Finn et al. (2017a)—this change is ultimately derived from **exclusions**

What is pointwise information?

- The surprise of the posterior compared to the surprise of the prior

  Prior $\quad p(t) \longrightarrow p(t|s_1) \quad$ Posterior

- Finn et al. (2017a)—this change is ultimately derived from **exclusions**

$P(T, S_1)$

t	1/8	s_1
	3/8	$s_1{}^{\mathsf{c}}$
t^{c}	1/4	s_1
	1/4	$s_1{}^{\mathsf{c}}$

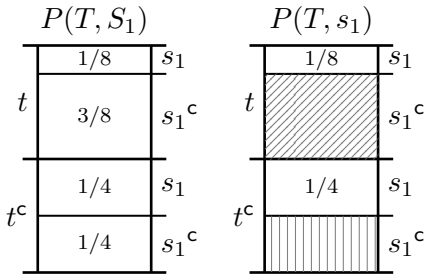
where $t^{\mathsf{c}} = \{\mathcal{T} \backslash t\}$ and $s_1{}^{\mathsf{c}} = \{\mathcal{S}_1 \backslash s_1\}$

What is pointwise information?

- The surprise of the posterior compared to the surprise of the prior

Prior $p(t) \longrightarrow p(t|s_1)$ Posterior

- Finn et al. (2017a)—this change is ultimately derived from **exclusions**



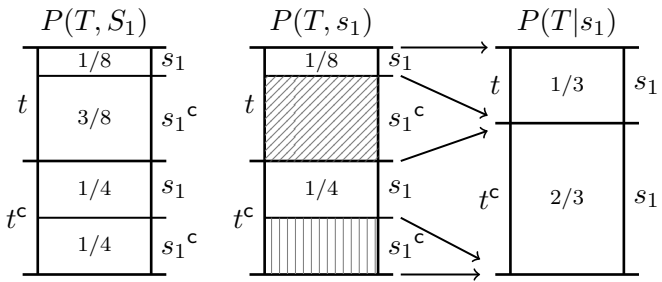
where $t^c = \{\mathcal{T} \backslash t\}$ and $s_1{}^c = \{\mathcal{S}_1 \backslash s_1\}$

What is pointwise information?

- The surprise of the posterior compared to the surprise of the prior

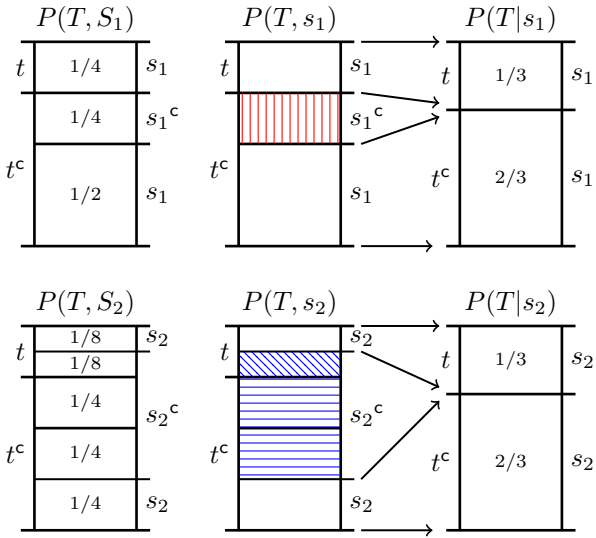
Prior $\quad p(t) \longrightarrow p(t|s_1) \quad$ Posterior

- Finn et al. (2017a)—this change is ultimately derived from **exclusions**



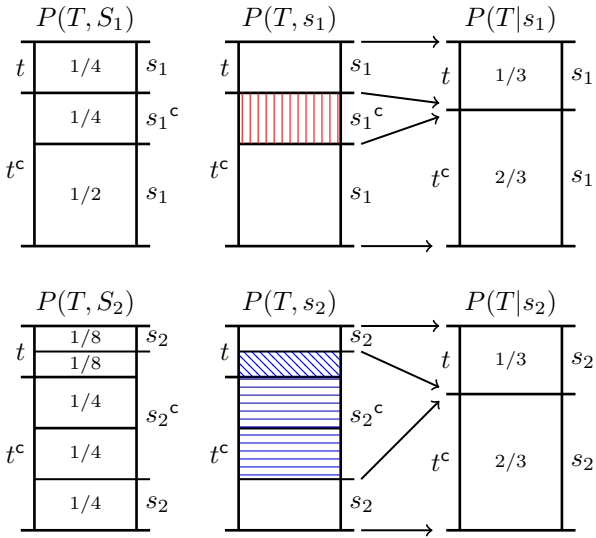
where $t^c = \{\mathcal{T} \backslash t\}$ and $s_1{}^c = \{\mathcal{S}_1 \backslash s_1\}$

Motivation



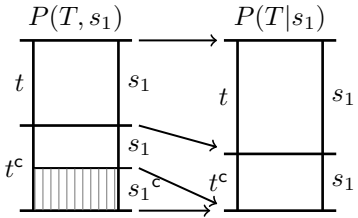
- The exclusions differ, but yet $i(t; s_1) = i(t; s_2) = 4/3$ bit

Motivation



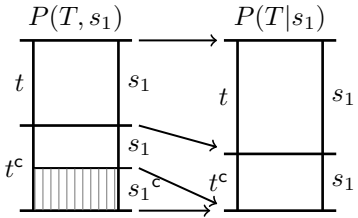
- The exclusions differ, but yet
  $$i(t; s_1) = i(t; s_2) = 4/3 \text{ bit}$$
- Pointwise MI is not injective
- Same info $\leftrightarrow$ same exclusions

Two types of exclusions

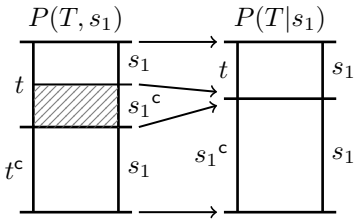


Purely informative exclusion

Two types of exclusions

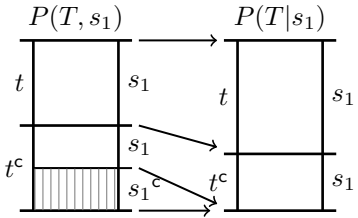


Purely informative exclusion

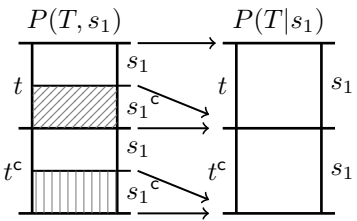


Purely misinformative exclusions

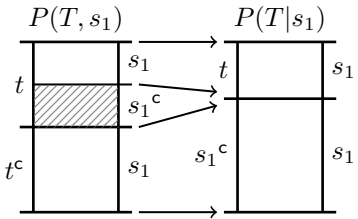
Two types of exclusions



Purely informative exclusion



General case

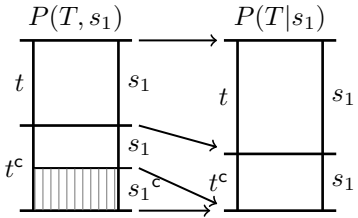


Purely misinformative exclusions

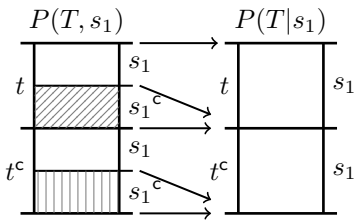
- Idea: split the pointwise MI into two components

$$i(s \to t) = i^+(s \to t) - i^-(s \to t)$$

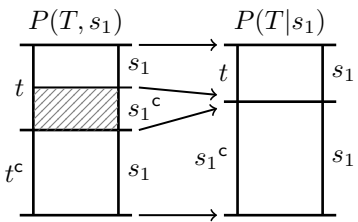
Two types of exclusions



Purely informative exclusion



General case



Purely misinformative exclusions

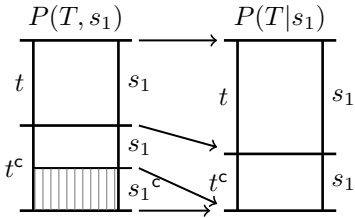
- Idea: split the pointwise MI into two components

$$i(s \to t) = i^+(s \to t) - i^-(s \to t)$$

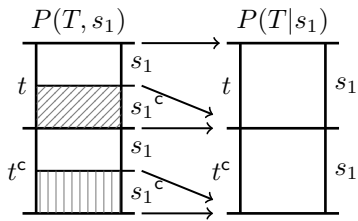
- In Finn et al. (2017a) we proved that

$$i^+(s_1 \to t) = h(s_1)$$
$$i^-(s_1 \to t) = h(s_1|t)$$

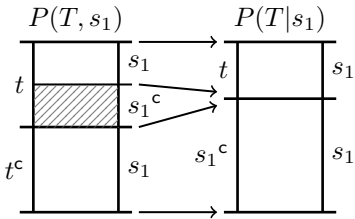
Two types of exclusions



Purely informative exclusion



General case



Purely misinformative exclusions

- Idea: split the pointwise MI into two components

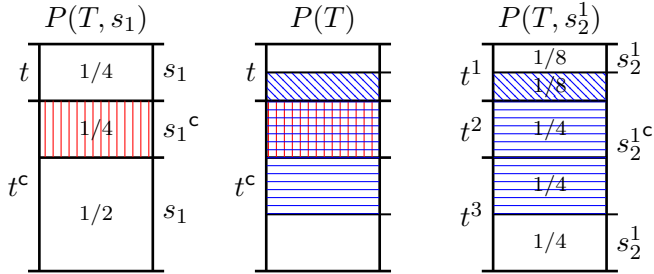
$$i(s \to t) = i^+(s \to t) - i^-(s \to t)$$

- In Finn et al. (2017a) we proved that

(**Specificity**) $\quad i^+(s_1 \to t) = h(s_1)$

(**Ambiguity**) $\quad i^-(s_1 \to t) = h(s_1|t)$

Specificity and ambiguity decomposition



$$i(s_1 \to t) = i(s_2 \to t) = \log \tfrac{4}{3} \text{ bit}$$

$$i_+(s_1 \to t) = \log \tfrac{4}{3} \text{ bit}, \qquad i_-(s_1 \to t) = 0 \text{ bit}$$

$$i_+(s_2 \to t) = \log \tfrac{8}{3} \text{ bit}, \qquad i_-(s_2 \to t) = 1 \text{ bit}$$

PPID using Specificity and Ambiguity

Axioms (PPID using Specificity and Ambiguity)

(1) *Symmetry*: $r^{\pm}(t : s_1, \ldots, s_n)$ is invariant under permutations of the s_i's

(2) *Monotonicity*: $r^{\pm}(t : s_1, \ldots, s_n) \leq r^{\pm}(t : s_1; \ldots; s_{n-1})$

(3) *Self-redundancy*: $r^{\pm}(t : s_i) = i^{\pm}(t; s_i)$

- Yields two redundancy lattices: the specificity and ambiguity lattices
- No longer have the non-negativity problem
- Still need a measure of redundant information on each lattice

Operational definition of unique information

Consider a horse race T and two gamblers with side wire's S_1 and S_2 respectively

- Gambler with wire S_1 will have the expected return
$$\Delta W = I(T; S_1) = \mathrm{E}\big[i(t^k, s_1^k)\big]$$
- Pointwise return on the k-th race Δw is give by $i(t^k, s_1^k)$
- Difference between pointwise return $\implies$ difference unique information
- Leads us to define the redundant specificity and redundant ambiguity
$$r^+_{\mathsf{min}}(s_1, \ldots, s_k \to t) = \min_{s_j} h(s_j) \qquad r^-_{\mathsf{min}}(s_1, \ldots, s_k \to t) = \min_{s_j} h(s_j|t)$$
- No identity rule—but justifiable operationally

Example: PwUnq

p	s_1	s_2	t	i_1^+	i_1^-	i_2^+	i_2^-	i_{12}^+	i_{12}^-	r^+	u_1^+	u_2^+	c^+	r^-	u_1^-	u_2^-	c^-
1/4	0	1	1	1	1	2	1	2	1	1	0	1	0	1	0	0	0
1/4	1	0	1	2	1	1	1	2	1	1	1	0	0	1	0	0	0
1/4	0	2	2	1	1	2	1	2	1	1	0	1	0	1	0	0	0
1/4	2	0	2	2	1	1	1	2	1	1	1	0	0	1	0	0	0
Expected values				3/2	1	3/2	1	2	1	1	1/2	1/2	0	1	0	0	0

- Recombining the average specificities and average ambiguities yields the PID

$$R(T : S_1, S_2) = 1 - 1 = 0 \text{ bit} \qquad U(T : S_1 \setminus S_2) = 1/2 - 0 = 1/2 \text{ bit}$$
$$C(T : S_1, S_2) = 0 - 0 = 0 \text{ bit} \qquad U(T : S_2 \setminus S_1) = 1/2 - 0 = 1/2 \text{ bit}$$

- Matches the PPID suggested earlier

Example: XOR

p	s_1	s_2	t	i_1^+	i_1^-	i_2^+	i_2^-	i_{12}^+	i_{12}^-	r^+	u_1^+	u_2^+	c^+	r^-	u_1^-	u_2^-	c^-
1/4	0	0	0	1	1	1	1	2	1	1	0	0	1	1	0	0	0
1/4	0	1	1	1	1	1	1	2	1	1	0	0	1	1	0	0	0
1/4	1	0	1	1	1	1	1	2	1	1	0	0	1	1	0	0	0
1/4	1	1	0	1	1	1	1	2	1	1	0	0	1	1	0	0	0
Expected values				1	1	1	1	2	1	1	0	0	1	1	0	0	0

- Recombining the average specificities and average ambiguities yields the PID

$$R(T : S_1, S_2) = 1 - 1 = 0 \text{ bit} \qquad U(T : S_1 \setminus S_2) = 0 - 0 = 0 \text{ bit}$$
$$C(T : S_1, S_2) = 1 - 0 = 1 \text{ bit} \qquad U(T : S_2 \setminus S_1) = 0 - 0 = 0 \text{ bit}$$

- Identifies redundancy due to shared knowledge from Bertschinger et al. (2013)

Finally...

Has a target chain rule!

References

Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, and Jürgen Jost. Shared informationnew insights and problems in decomposing information in complex systems. In *Proceedings of the European Conference on Complex Systems 2012*, pages 251–269. Springer, 2013.

Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014.

Robert Fano. *Transmission of Information*. The MIT Press, 1961.

Conor Finn, Mikhail Prokopenko, and Joseph T. Lizier. Decomposing pointwise information into directed positive and negative components. 2017a. To appear.

Conor Finn, Mikhail Prokopenko, and Joseph T. Lizier. Pointwise partial information decomposition using the specificity and ambiguity lattices. 2017b. To appear.

Virgil Griffith and Christof Koch. Quantifying synergistic mutual information. In Mikhail Prokopenko, editor, *Guided Self-Organization: Inception*, volume 9 of *Emergence, Complexity and Computation*, pages 159–190. Springer Berlin Heidelberg, 2014. ISBN 978-3-642-53733-2.

Malte Harder, Christoph Salge, and Daniel Polani. Bivariate measure of redundant information. *Physical Review E*, 87 (1):012130, 2013.

Robin AA Ince. Measuring multivariate redundant information with pointwise common change in surprisal. *Entropy*, 19 (7):318, 2017.

Claude E Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:623–656, 1948.

Paul L Williams and Randall D Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.

Comparison to Other Decompositions and Measures

- Approach is most similar to Ince (2017) but differs in how the non-positivity is dealt with
- Similar to I_{min} of Williams and Beer (2010) but now fully pointwise
- Axiom 4 is similar to Assumption $(**)$ of Bertschinger et al. (2014), i.e. measure $\widetilde{UI}$
- This also makes it similar to S_{VK} of Griffith and Koch (2014)
- Like other measures, there is no target monotonicity, i.e. don not have that

$$R_{\text{min}}\big(S_1, S_2 \to T_1\big) \leq R_{\text{min}}\big(S_1, S_2 \to T_1, T_2\big)$$

- But unlike other measures, there is a target chain rule

$$R_{\text{min}}\big(S_1, S_2 \to T_1, T_2\big) = R_{\text{min}}\big(S_1, S_2 \to T_1\big) + R_{\text{min}}\big(S_1, S_2 \to T_2 | T_1\big)$$

Example: IMPRDN

p	s_1	s_2	t	i_1^+	i_1^-	i_2^+	i_2^-	i_{12}^+	i_{12}^-	r^+	u_1^+	u_2^+	c^+	r^-	u_1^-	u_2^-	c^-
1/2	0	0	0	1	0	$\lg 8/5$	0	1	0	$\lg 8/5$	$\lg 5/4$	0	0	0	0	0	0
3/8	1	1	1	1	0	$\lg 8/3$	$\lg 4/3$	$\lg 8/3$	$\lg 4/3$	1	0	$\lg 4/3$	0	0	0	$\lg 4/3$	0
1/8	1	0	1	1	0	$\lg 8/5$	2	3	2	$\lg 8/5$	$\lg 5/4$	0	2	0	0	2	0
Expected				1	0	0.954	0.406	1.406	0.406	0.799	0.201	0.156	0.250	0	0	0.406	0

- Recombining the average specificities and average ambiguities yields the PID

  $R(T : S_1, S_2) = 0.799 - 0 = 0.799$ bit $\quad U(T : S_1 \setminus S_2) = 0.201 - 0 = 0.201$ bit

  $C(T : S_1, S_2) = 0.25 - 0.25 = 1$ bit $\quad U(T : S_2 \setminus S_1) = 0.156 - 0.406 = -0.25$ bit

- May be negative unique information on average if a source is uniquely misinformative

Example: TwoBitCopy

p	s_1	s_2	t	i_1^+	i_1^-	i_2^+	i_2^-	i_{12}^+	i_{12}^-	r^+	u_1^+	u_2^+	c^+	r^-	u_1^-	u_2^-	c^-
1/4	0	0	00	1	0	1	0	2	0	1	0	0	1	0	0	0	0
1/4	0	1	01	1	0	1	0	2	0	1	0	0	1	0	0	0	0
1/4	1	0	10	1	0	1	0	2	0	1	0	0	1	0	0	0	0
1/4	1	1	11	1	0	1	0	2	0	1	0	0	1	0	0	0	0
Expected values				1	0	1	0	2	0	1	0	0	1	0	0	0	0

- Recombining the average specificities and average ambiguities yields the PID

$$R(T : S_1, S_2) = 1 - 0 = 1 \text{ bit} \qquad U(T : S_1 \setminus S_2) = 0 - 0 = 0 \text{ bit}$$
$$C(T : S_1, S_2) = 1 - 0 = 1 \text{ bit} \qquad U(T : S_2 \setminus S_1) = 0 - 0 = 0 \text{ bit}$$

- Result is the same as it is for I_{min} of Williams and Beer (2010)
- The measure and decomposition does not possess the identity property
  - Does mean that we can use this decomposition for more than 3 variables

Example: TwoBitCopy Horse Race

p	s_1	s_2	t	i_1^+	i_1^-	i_2^+	i_2^-	i_{12}^+	i_{12}^-	r^+	u_1^+	u_2^+	c^+	r^-	u_1^-	u_2^-	c^-
1/4	b	r	a	1	0	1	0	2	0	1	0	0	1	0	0	0	0
1/4	b	g	b	1	0	1	0	2	0	1	0	0	1	0	0	0	0
1/4	w	r	c	1	0	1	0	2	0	1	0	0	1	0	0	0	0
1/4	w	g	d	1	0	1	0	2	0	1	0	0	1	0	0	0	0
Expected values				1	0	1	0	2	0	1	0	0	1	0	0	0	0

- Recombining the average specificities and average ambiguities yields the PID

$$R(T : S_1, S_2) = 1 - 0 = 1 \text{ bit} \qquad U(T : S_1 \backslash S_2) = 0 - 0 = 0 \text{ bit}$$
$$C(T : S_1, S_2) = 1 - 0 = 1 \text{ bit} \qquad U(T : S_2 \backslash S_1) = 0 - 0 = 0 \text{ bit}$$

- Result is the same as it is for I_{min} of Williams and Beer (2010)
- The measure and decomposition does not possess the identity property
  - Does mean that we can use this decomposition for more than 3 variables