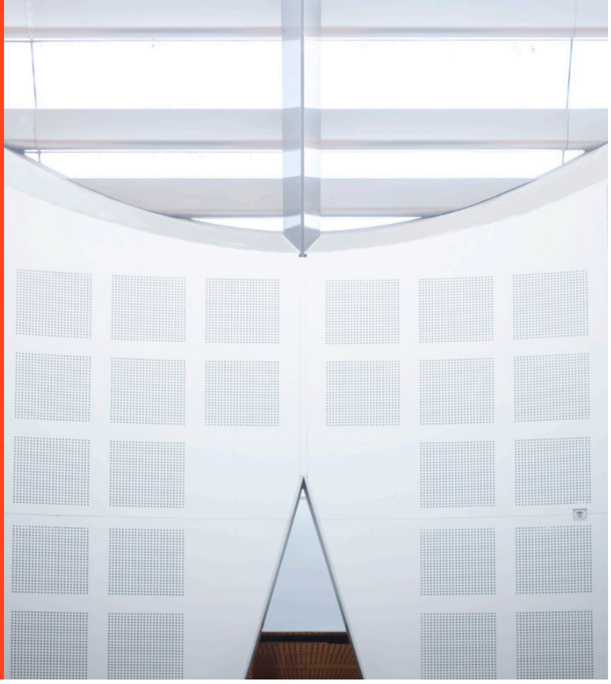
Pointwise Partial Information Decomposition Using Specificity and Ambiguity Lattices

4th Int. Elec. Conf. on Entropy and Its Applications (ECEA-4)

Conor Finn
Mikhail Prokopenko
Joseph T. Lizier (presenting)

November 10, 2017

THE UNIVERSITY OF SYDNEY



Information decomposition

Goal: identify redundant, unique and complementary (synergistic) components of MI

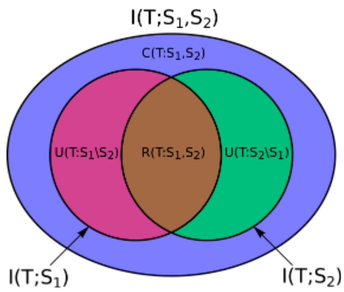
Consider three random variables S_1, S_2 and T

- Aim: predict T using S_1 and S_2
- Mutual information captures

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

- Joint mutual information captures

$$I(T;S_1;S_2) = R(T:S_1;S_2) + U(T:S_1 \setminus S_2) + U(T:S_2 \setminus S_1) + C(T:S_1;S_2)$$



Information decomposition: examples

1. Unique information $U(T:S_1 \setminus S_2)$
2. Redundant information $R(T:S_1;S_2)$
3. Synergistic information $C(T:S_1;S_2)$
4. In general, all types of information are present

UNQ

p	s_1	s_2	t
½	0	0	0
½	1	0	1

RDN

p	s_1	s_2	t
½	0	0	0
½	1	1	1

XOR

p	s_1	s_2	t
¼	0	0	0
¼	0	1	1
¼	1	0	1
¼	1	1	0

AND

p	s_1	s_2	t
¼	0	0	0
¼	0	1	0
¼	1	0	0
¼	1	1	1

Information decomposition

Goal: identify redundant, unique and complementary (synergistic) components of MI

We have 3 equations in 4 unknowns!

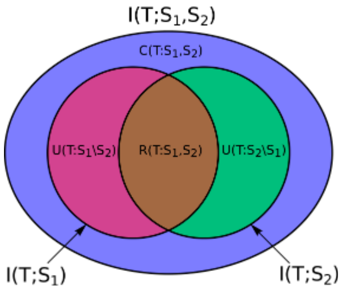
Consider three random variables S_1, S_2 and T

- Aim: predict T using S_1 and S_2
- Mutual information captures

$$I(T; S_1) = R(T : S_1; S_2) + U(T : S_1 n S_2)$$
$$I(T; S_2) = R(T : S_1; S_2) + U(T : S_2 n S_1)$$

- Joint mutual information captures

$$I(T; S_1; S_2) = R(T : S_1; S_2) + U(T : S_1 n S_2) + U(T : S_2 n S_1) + C(T : S_1; S_2)$$



Partial Information Decomposition (Williams and Beer, 2010)

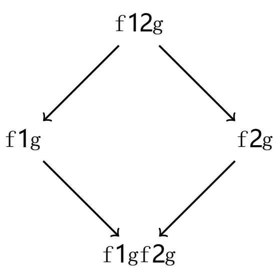
- Axiomatic framework extending this decomposition to arbitrary number of sources

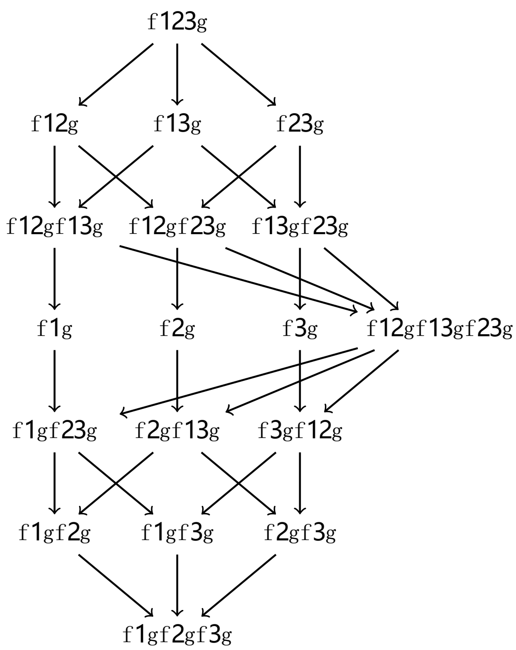
Axioms (PID)

(1) Symmetry: $R(T : S_1; :::; S_n)$ is invariant under permutations of the S_i's
(2) Monotonicity: $R(T : S_1; :::; S_n) \square R(T : S_1; :::; S_{n \square 1})$
(3) Self-redundancy: $R(T : S_i) = I(T; S_i)$

- Yields a redundancy lattice

Redundancy lattice





Partial Information Decomposition (Williams and Beer, 2010)

- 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

Pointwise information theory

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

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

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

Pointwise information decomposition

- Pointwise decomposition for each realisation

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

- Should be able to take the expectation over all realisations

$$\begin{aligned} R(T : S_1; S_2) &= \mathbb{E}\left[r(t : s_1; s_2)\right] \qquad & U(T : S_1 \setminus S_2) &= \mathbb{E}\left[u(t : s_1 \setminus s_2)\right] \\ C(T : S_1; S_2) &= \mathbb{E}\left[c(t : s_1; s_2)\right] \qquad & U(T : S_2 \setminus S_1) &= \mathbb{E}\left[u(t : s_2 \setminus s_1)\right] \end{aligned}$$

- 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:

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
¼	0	1	1	0	1	1	0	0	1	0
¼	1	0	1	1	0	1	0	1	0	0
¼	0	2	2	0	1	1	0	0	1	0
¼	2	0	2	1	0	1	0	1	0	0
Expected values				½	½	1	0	½	½	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: $r(t : s_i) = i(t; s_i)$

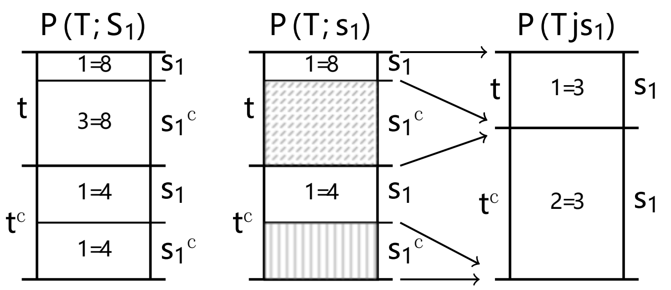
- Problems:
  1. Pointwise mutual information is not non-negative
  2. Still no clear definition of redundant information

What is pointwise information $i(t; s_1)$?

- The surprise of the posterior compared to the surprise of the prior for event t, s_1

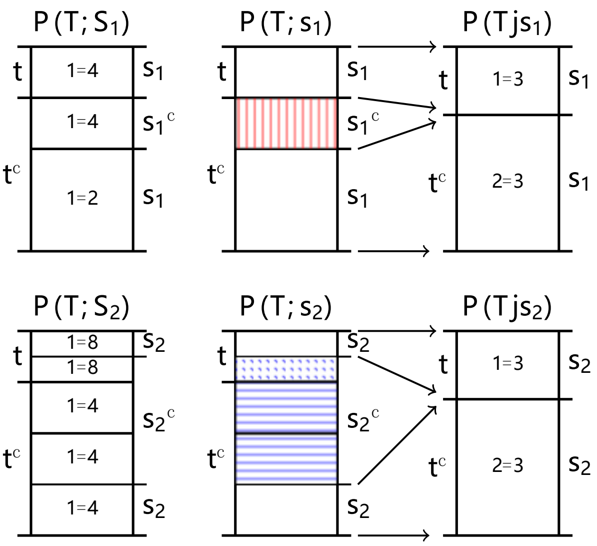
$$\text{Prior} \quad p(t) \longrightarrow p(t|s_1) \quad \text{Posterior}$$

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



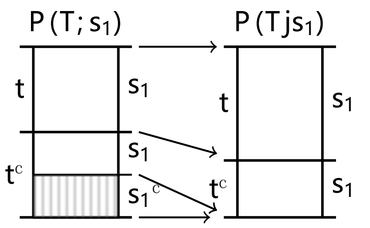
where $t^c = \{T \setminus t\}$ and $s_1{}^c = \{S_1 \setminus s_1\}$

Motivation

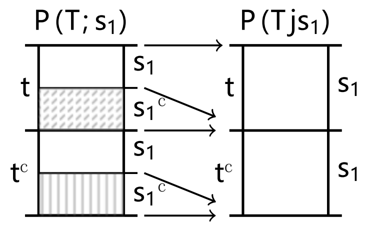


- The exclusions differ, but yet $i(t; s_1) = i(t; s_2) = 4/3$ bit
- Pointwise MI is not injective
- Same info $\Leftrightarrow$ same exclusions

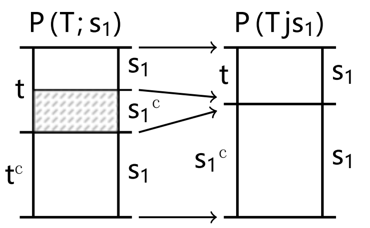
Two types of exclusions



Purely informative exclusion



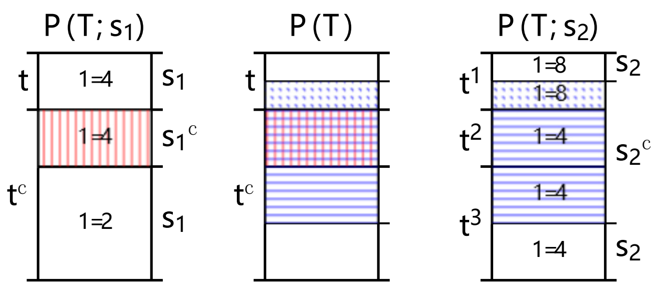
General case



Purely misinformative exclusions

- Idea: split the pointwise MI into two components
  $$i(s \rightarrow t) = i^{+}(s \rightarrow t) - i^{-}(s \rightarrow t)$$
- In Finn et al. (2017) we proved that
  (Specificity) $i^{+}(s_1 \rightarrow t) = h(s_1)$
  (Ambiguity) $i^{-}(s_1 \rightarrow t) = h(s_1|t)$

Specificity and ambiguity decomposition



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

$i_{+}(s_1 \rightarrow t) = \log \frac{4}{3}$ bit; $\quad i_{-}(s_1 \rightarrow t) = 0$ bit

$i_{+}(s_2 \rightarrow t) = \log \frac{8}{3}$ bit; $\quad i_{-}(s_2 \rightarrow t) = 1$ 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

PPID using Specificity and Ambiguity

Additional Axiom (PPID using Specificity and Ambiguity)

(4) Pointwise event space: $r^{\pm}(t : s_1; \ldots; s_n)$ depend only on the size of informative and misinformative exclusions

- Akin to Assumption ([illegible]) of Bertschinger et al. (2014) → cannot infer complementary info (synergy) without full joint distribution.
- Leads us to define the redundant specificity and redundant ambiguity

$$r^{+}_{\min}(s_1; \ldots; s_k \to t) = \min_{s_j} h(s_j) \qquad r^{-}_{\min}(s_1; \ldots; s_k \to t) = \min_{s_j} h(s_j | t)$$

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^-
¼	0	1	1	1	1	2	1	2	1	1	0	1	0	1	0	0	0
¼	1	0	1	2	1	1	1	2	1	1	1	0	0	1	0	0	0
¼	0	2	2	1	1	2	1	2	1	1	0	1	0	1	0	0	0
¼	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	½	½	0	1	0	0	0

- Recombining the average specificities and average ambiguities yields the PID

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

- 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. (2014)

Further properties

- Can be interpreted in terms of Kelly gambling (e.g. difference in unique information is difference in winnings of two players)
- Is generalisable beyond two sources
- Has a target chain rule (in net):

$$r_{\min}^{-}\left(s_1; s_2 \to t_1; t_2\right) = r_{\min}^{-}\left(s_1; s_2 \to t_1\right) + r_{\min}^{-}\left(s_1; s_2 \to t_2 | t_1\right)$$

- Provides consistent decomposition of two-bit copy $I(\{S_1; S_2\}; S_1; S_2)$ via target chain rule
- Many echoes of intuition behind previous measures, e.g. (Bertschinger et al., 2014; Ince, 2017; Williams and Beer, 2010), but is fully pointwise and component-wise.

References

- Entropy special issue on Information Decomposition: http://bit.ly/infodecomp-2017
- Conor Finn, Mikhail Prokopenko, and Joseph T. Lizier. Pointwise Partial Information Decomposition Using Specificity and Ambiguity Lattices 2017. To appear.

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. 2017. 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 Ince. Measuring multivariate redundant information with pointwise common change in surprisal. Entropy, 19(7): 318, 2017. doi: 10.3390/e19070318.
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.