Multivariate Information Decomposition – Progress, Problems, and Outlook

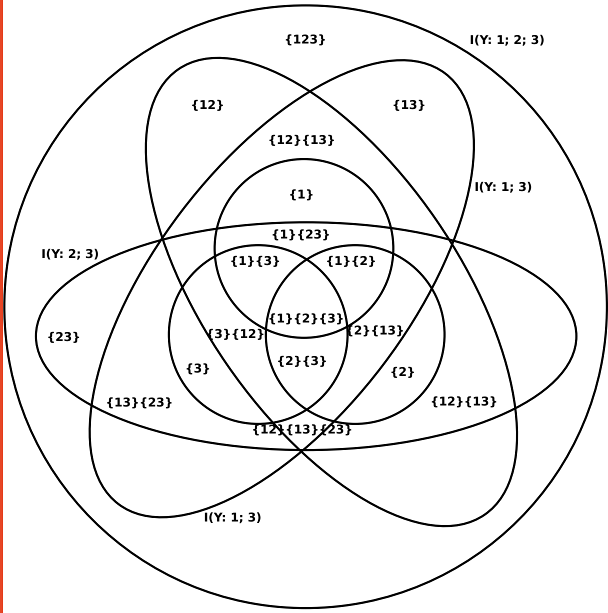
Analytics Group Retreat

Conor Finn

April 27, 2017

THE UNIVERSITY OF SYDNEY

DATA 61 CSIRO



Information Theory

- (Shannon) entropy: expected information in a realisation of a random variable

$$H(X) = \sum_x p(x) \log \frac{1}{p(x)} \geq 0$$

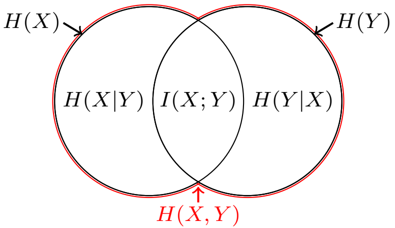
Information Theory

- (Shannon) entropy: expected information in a realisation of a random variable

$$H(X) = \sum_x p(x) \log \tfrac{1}{p(x)} \geq 0$$

- The Shannon inequalities provide means to define non-negative symmetric quantity

$$H(X),\ H(Y) \leq H(X,Y) \leq H(X) + H(Y)$$



$$\begin{aligned} I(X;Y) &:= H(X) + H(Y) - H(X,Y) \geq 0 \\ H(X|Y) &:= H(X,Y) - H(Y) \qquad\qquad\quad\;\, \geq 0 \\ H(Y|X) &:= H(X,Y) - H(X) \qquad\qquad\quad\;\, \geq 0 \end{aligned}$$

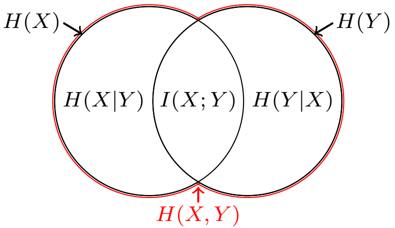
Information Theory

- (Shannon) entropy: expected information in a realisation of a random variable

$$H(X) = \sum_x p(x) \log \tfrac{1}{p(x)} \geq 0$$

- The Shannon inequalities provide means to define non-negative symmetric quantity

$$H(X),\ H(Y) \leq H(X,Y) \leq H(X) + H(Y)$$



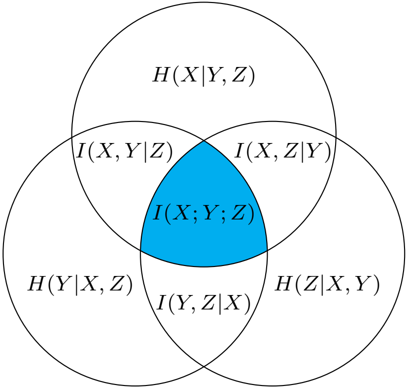
$$\begin{aligned} I(X;Y) &:= H(X) + H(Y) - H(X,Y) \geq 0 \\ H(X|Y) &:= H(X,Y) - H(Y) \geq 0 \\ H(Y|X) &:= H(X,Y) - H(X) \geq 0 \end{aligned}$$

- The mutual information quantifies the interdependency between **two** random variables

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

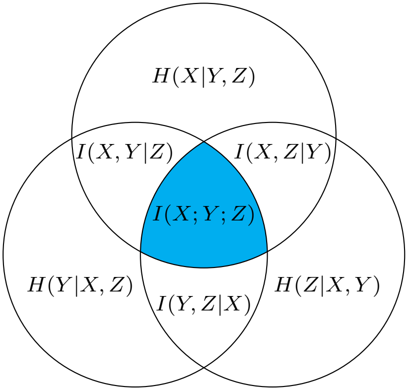
Multivariate Information Theory

Can we quantify the mutual interdependence between three or more random variables?



Multivariate Information Theory

Can we quantify the mutual interdependence between three or more random variables?

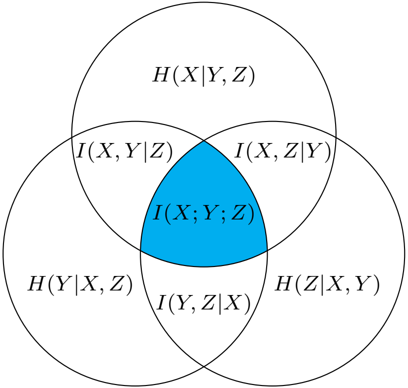


- For two variables we had

$$I(X;Y) \coloneqq H(X) + H(Y) - H(\{X,Y\}) \geq 0.$$

Multivariate Information Theory

Can we quantify the mutual interdependence between three or more random variables?



- For two variables we had

$$I(X;Y) := H(X) + H(Y) - H(\{X,Y\}) \geq 0.$$

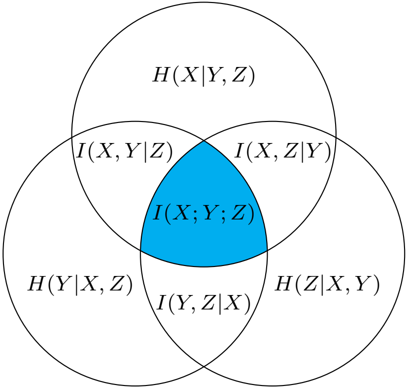
- For three variables we have the co-information

$$I(X;Y;X) := I(X;Y) + I(X;Z) - I(X;\{Y,Z\}).$$

a.k.a. the multivariate mutual information, interaction information, amounts of information

Multivariate Information Theory

Can we quantify the mutual interdependence between three or more random variables?



- For two variables we had

$$I(X;Y) := H(X) + H(Y) - H(\{X,Y\}) \geq 0.$$

- For three variables we have the co-information

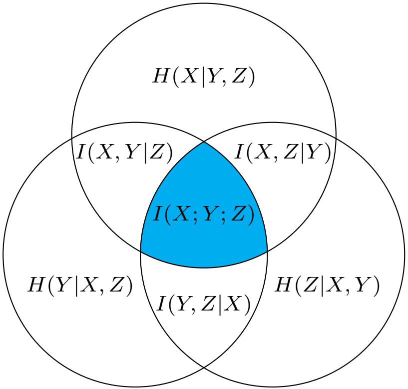
$$I(X;Y;X) := I(X;Y) + I(X;Z) - I(X;\{Y,Z\}).$$

a.k.a. the multivariate mutual information, interaction information, amounts of information

- However, this quantity can be negative!

Multivariate Information Theory

Can we quantify the mutual interdependence between three or more random variables?



- For two variables we had

$$I(X;Y) := H(X) + H(Y) - H(\{X,Y\}) \geq 0.$$

- For three variables we have the co-information

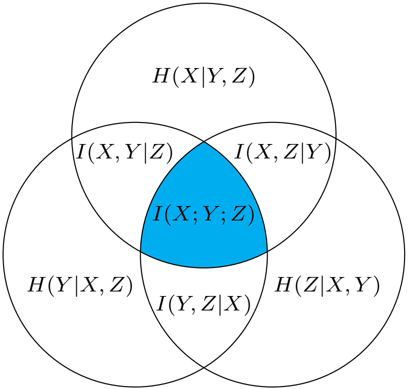
$$I(X;Y;X) := I(X;Y) + I(X;Z) - I(X;\{Y,Z\}).$$

a.k.a. the multivariate mutual information, interaction information, amounts of information

- However, this quantity can be negative!
- What is negative information?

Multivariate Information Theory

Can we quantify the mutual interdependence between three or more random variables?



- For two variables we had

$$I(X;Y) := H(X) + H(Y) - H(\{X,Y\}) \geq 0.$$

- For three variables we have the co-information

$$I(X;Y;X) := I(X;Y) + I(X;Z) - I(X;\{Y,Z\}).$$

a.k.a. the multivariate mutual information, interaction information, amounts of information

- However, this quantity can be negative!
- What is negative information?

- This is because we don't have Shannon inequalities for multivariate information.

Present shortcomings and problems

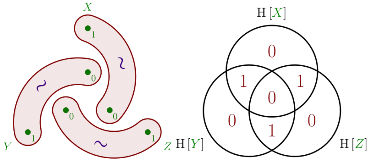
- No well agreed-upon generalisation of the mutual information

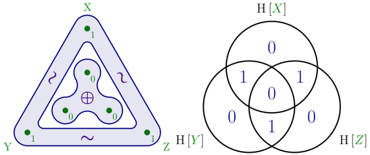
Present shortcomings and problems

- No well agreed-upon generalisation of the mutual information
- Cannot decompose multivariate interdependency
  - Unique, redundant and synergistic information

Present shortcomings and problems

- No well agreed-upon generalisation of the mutual information
- Cannot decompose multivariate interdependency
  - Unique, redundant and synergistic information
- In fact, even worse — it cannot distinguish between systems with vastly different internal interdependency structures





Why bother to solve this problem?

- Neuroscience:
  - Currently, information theory can measure neural information storage and transfer
  - Quantifying information modification requires multivariate information theory

Why bother to solve this problem?

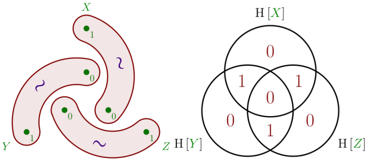
- Neuroscience:
  - Currently, information theory can measure neural information storage and transfer
  - Quantifying information modification requires multivariate information theory
- Feature selection in machine learning:
  - Consider a data set with known heart disease risk factors:
  - Smoker or non-smoker might contribute a large amount of unique information;
  - Obesity and diabetes might be largely redundant;
  - Genetic risks and age might be most important synergistically with other features.

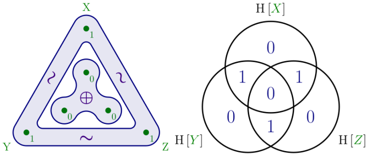
Why bother to solve this problem?

- Neuroscience:
  - Currently, information theory can measure neural information storage and transfer
  - Quantifying information modification requires multivariate information theory
- Feature selection in machine learning:
  - Consider a data set with known heart disease risk factors:
  - Smoker or non-smoker might contribute a large amount of unique information;
  - Obesity and diabetes might be largely redundant;
  - Genetic risks and age might be most important synergistically with other features.
- Lossless compression of structured databases:
  - high-dimensional redundancies need to be removed
  - Shannon's theory is not a very useful for multivariate compression

Present shortcomings and problems

- No well agreed-upon generalisation of the mutual information
- Cannot decompose multivariate interdependency
  - Unique, redundant and synergistic information
- In fact, even worse — it cannot distinguish between systems with vastly different internal interdependency structures





Unique, redundant and synergistic information

Consider three random variables X, Y and Z and suppose we are interested in predicting the value of X from Y and Z

Unique, redundant and synergistic information

Consider three random variables X, Y and Z and suppose we are interested in predicting the value of X from Y and Z

- Unique information: source Z may contain information about X that source Y does not, or vice versa

X	Y	Z	P
0	0	0	1/2
1	0	1	1/2

Unique, redundant and synergistic information

Consider three random variables X, Y and Z and suppose we are interested in predicting the value of X from Y and Z

- Unique information: source Z may contain information about X that source Y does not, or vice versa

X	Y	Z	P
0	0	0	1/2
1	0	1	1/2

- Redundant information: source Y may contain the same information as source Z about X

X	Y	Z	P
0	0	0	1/2
1	1	1	1/2

Unique, redundant and synergistic information

Consider three random variables X, Y and Z and suppose we are interested in predicting the value of X from Y and Z

- Unique information: source Z may contain information about X that source Y does not, or vice versa

X	Y	Z	P
0	0	0	1/2
1	0	1	1/2

- Redundant information: source Y may contain the same information as source Z about X

X	Y	Z	P
0	0	0	1/2
1	1	1	1/2

- Synergistic information: it is possible that neither source Z nor source Y contain information about X but together they do

X	Y	Z	P
0	0	0	1/4
1	0	1	1/4
1	1	0	1/4
0	1	1	1/4

Decomposing bivariate dependency

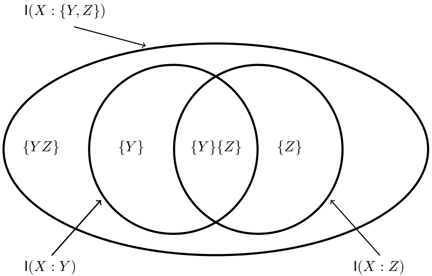
In general, all three types of information are present simultaneously

Decomposing bivariate dependency

In general, all three types of information are present simultaneously

- We seek a meaningful decomposition of $\mathrm{I}(X;\{Y,Z\})$

X	Y	Z	P
0	0	0	1/4
0	0	1	1/4
0	1	0	1/4
1	1	1	1/4

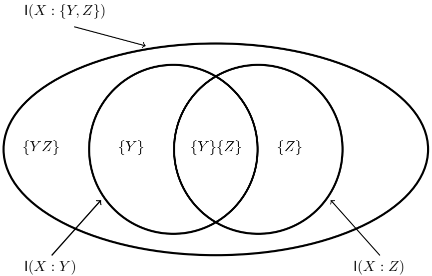


Decomposing bivariate dependency

In general, all three types of information are present simultaneously

- We seek a meaningful decomposition of $\mathrm{I}(X;\{Y,Z\})$

X	Y	Z	P
0	0	0	1/4
0	0	1	1/4
0	1	0	1/4
1	1	1	1/4



- Shannon's information theory insufficient for the decomposition

$$\begin{aligned}\mathrm{CoI}(X;Y;Z) &:= \mathrm{I}(X;Y) + \mathrm{I}(X;Z) - \mathrm{I}(X;\{Y,Z\}) \\ &= \mathrm{RI}(X:Y;Z) - \mathrm{SI}(X:Y;Z)\end{aligned}$$

Partial information decomposition

An axiomatic framework for decomposing multivariate dependence introduced in 2010 by Williams and Beer

Partial information decomposition

An axiomatic framework for decomposing multivariate dependence introduced in 2010 by Williams and Beer

- Principled method for decomposing the multivariate information for an **arbitrary number of variables**
- Derived from axioms a measure of redundancy $I_\cap$ must satisfy

Partial information decomposition

An axiomatic framework for decomposing multivariate dependence introduced in 2010 by Williams and Beer

- Principled method for decomposing the multivariate information for an **arbitrary number of variables**
- Derived from axioms a measure of redundancy $I_\cap$ must satisfy

Axioms

(1) *Symmetry*: $\mathrm{I}_\cap$ is invariant under permutations of the Y_i's
(2) *Self-redundancy*: $\mathrm{I}_\cap(X : Y) = \mathrm{I}(X;Y)$
(3) *Monotonicity*: $\mathrm{I}_\cap(X : Y_1; \ldots ; Y_k) \leq I_\cap(X : Y_1; \ldots ; Y_{k-1})$

Partial information decomposition

An axiomatic framework for decomposing multivariate dependence introduced in 2010 by Williams and Beer

- Principled method for decomposing the multivariate information for an **arbitrary number of variables**

- Derived from axioms a measure of redundancy $I_\cap$ must satisfy

Axioms

(1) *Symmetry*: $\mathrm{I}_\cap$ is invariant under permutations of the Y_i's
(2) *Self-redundancy*: $\mathrm{I}_\cap(X : Y) = \mathrm{I}(X; Y)$
(3) *Monotonicity*: $\mathrm{I}_\cap(X : Y_1; \ldots; Y_k) \leq I_\cap(X : Y_1; \ldots; Y_{k-1})$

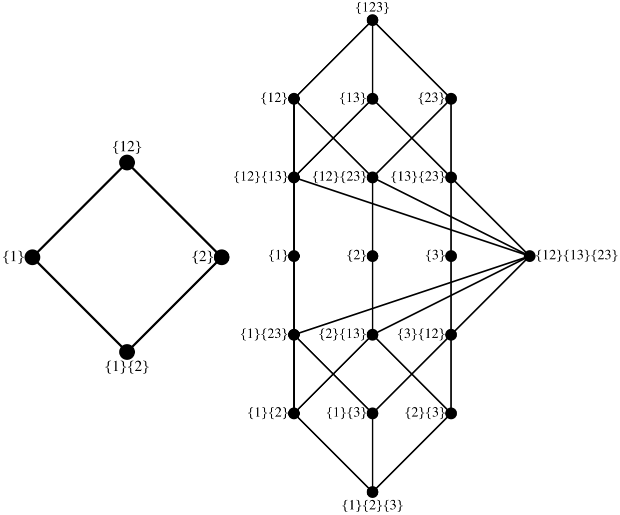
- Based on the intuitive notions from set theory

Partial information decomposition

Provides a structured decomposition of multivariate information (lattice structure)

Partial information decomposition

Provides a structured decomposition of multivariate information (lattice structure)

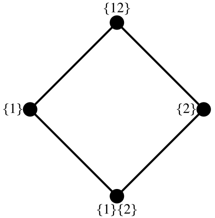


Partial information decomposition

Möbius inversion over the lattice yields partial information atoms

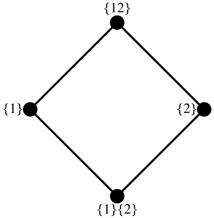
Partial information decomposition

Möbius inversion over the lattice yields partial information atoms



Partial information decomposition

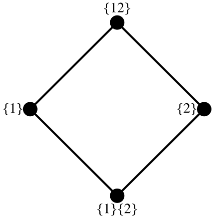
Möbius inversion over the lattice yields partial information atoms



$$\mathrm{RI}(X : 1; 2) = \mathrm{I}_\cap(X : 1; 2)$$

Partial information decomposition

Möbius inversion over the lattice yields partial information atoms

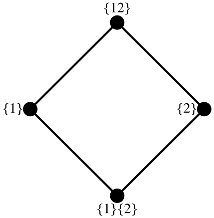


$$\mathrm{UI}(X:1) = \mathrm{I}(X:1) - \mathrm{RI}(X:1;2)$$
$$\mathrm{UI}(X:2) = \mathrm{I}(X:2) - \mathrm{RI}(X:1;2)$$

$$\mathrm{RI}(X:1;2) = \mathrm{I}_{\cap}(X:1;2)$$

Partial information decomposition

Möbius inversion over the lattice yields partial information atoms



$$\mathrm{SI}(X:1;2) = \mathrm{I}(X:\{1,2\}) - \mathrm{RI}(X:1;2) - \mathrm{UI}(X:1) - \mathrm{UI}(X:2)$$

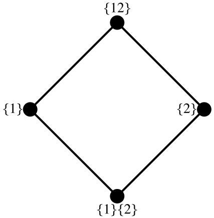
$$\mathrm{UI}(X:1) = \mathrm{I}(X:1) - \mathrm{RI}(X:1;2)$$

$$\mathrm{UI}(X:2) = \mathrm{I}(X:2) - \mathrm{RI}(X:1;2)$$

$$\mathrm{RI}(X:1;2) = \mathrm{I}_{\cap}(X:1;2)$$

Partial information decomposition

Möbius inversion over the lattice yields partial information atoms

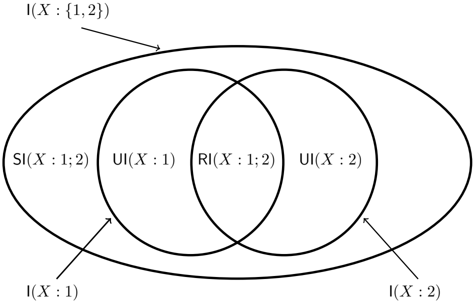


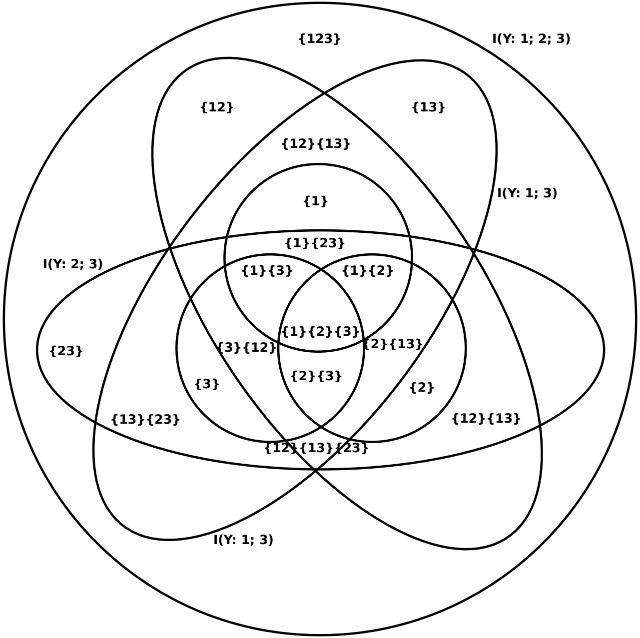
$$\mathrm{SI}(X:1;2) = \mathrm{I}(X:\{1,2\}) - \mathrm{RI}(X:1;2) - \mathrm{UI}(X:1) - \mathrm{UI}(X:2)$$

$$\mathrm{UI}(X:1) = \mathrm{I}(X:1) - \mathrm{RI}(X:1;2)$$

$$\mathrm{UI}(X:2) = \mathrm{I}(X:2) - \mathrm{RI}(X:1;2)$$

$$\mathrm{RI}(X:1;2) = \mathrm{I}_{\cap}(X:1;2)$$




{123}
I(Y: 1; 2; 3)
{12}
{13}
{12}{13}
I(Y: 1; 3)
{1}
{1}{23}
I(Y: 2; 3)
{1}{3}
{1}{2}
{1}{2}{3}
{2}{13}
{3}{12}
{23}
{2}{3}
{3}
{2}
{13}{23}
{12}{13}
{12}{13}{23}
I(Y: 1; 3)

PID is elegant, however...

- These axioms alone do not uniquely specify the form of $I_\cap$!
  - Need to either introduce new axioms to obtain uniqueness
  - Or directly specify a redundancy measure

PID is elegant, however...

- These axioms alone do not uniquely specify the form of $I_\cap$!
  - Need to either introduce new axioms to obtain uniqueness
  - Or directly specify a redundancy measure
- Current three competing redundancy measures—none of which are satisfactory
  - Williams and Beer PID measure I_{min} (same amount, not the same information)
  - Harder et al. I_{red} (difficult to calculate and bivariate only)
  - Bertschinger et al. $\widetilde{\mathrm{UI}}$ (difficult to calculate and bivariate only)

Ongoing research direction

In which direction are we taking our research — our unique edge

Ongoing research direction

In which direction are we taking our research — our unique edge

- We believe that defining a measure which is built from the ground up with a meaningful local intepretation will be fruitful

Ongoing research direction

In which direction are we taking our research — our unique edge

- We believe that defining a measure which is built from the ground up with a meaningful local intepretation will be fruitful

- No obvious reason why multivariate information theory should not be localisable
  - Despite this not many others work on local based approaches

Ongoing research direction

In which direction are we taking our research — our unique edge

- We believe that defining a measure which is built from the ground up with a meaningful local intepretation will be fruitful
- No obvious reason why multivariate information theory should not be localisable
  - Despite this not many others work on local based approaches
- "The problem is $\mathrm{I}_{\mathsf{min}}$ does not distinguish whether sources carry **the same** information or just **the same amount** of information"
  - Going fully local should avoid this issue

Ongoing research direction

In which direction are we taking our research — our unique edge

- We believe that defining a measure which is built from the ground up with a meaningful local intepretation will be fruitful
- No obvious reason why multivariate information theory should not be localisable
  - Despite this not many others work on local based approaches
- "The problem is $\mathrm{I}_{\mathsf{min}}$ does not distinguish whether sources carry **the same** information or just **the same amount** of information"
  - Going fully local should avoid this issue
- Local approach frees up the problem in many ways

Ongoing research direction

In which direction are we taking our research — our unique edge

- We believe that defining a measure which is built from the ground up with a meaningful local intepretation will be fruitful
- No obvious reason why multivariate information theory should not be localisable
  - Despite this not many others work on local based approaches
- "The problem is $\mathrm{I}_{\mathsf{min}}$ does not distinguish whether sources carry **the same** information or just **the same amount** of information"
  - Going fully local should avoid this issue
- Local approach frees up the problem in many ways
- Promising work to be published soon—writing up now!

Questions?

Redundancy measures: $\mathrm{I}_{\mathsf{min}}$

Original measure of redundancy introduced by Williams and Beer

$$\mathrm{I}_{\mathsf{min}}(X : Y_1, \ldots, Y_k) = \sum_{x} p(x) \min_{Y_i} \mathrm{I}(X = x; Y_i)$$

Redundancy measures: $\mathrm{I}_{\mathsf{min}}$

Original measure of redundancy introduced by Williams and Beer

$$\mathrm{I}_{\mathsf{min}}(X : Y_1, \dots, Y_k) = \sum_x p(x) \min_{Y_i} \mathrm{I}(X = x; Y_i)$$

- Semi-local approach: for each $X = x$ the redundant information is the minimum information provided by all of the sources Y_i

Redundancy measures: $\mathrm{I}_{\mathsf{min}}$

Original measure of redundancy introduced by Williams and Beer

$$\mathrm{I}_{\mathsf{min}}(X : Y_1, \ldots, Y_k) = \sum_x p(x) \min_{Y_i} \mathrm{I}(X = x; Y_i)$$

- Semi-local approach: for each $X = x$ the redundant information is the minimum information provided by all of the sources Y_i
- Widely critised after its introduction — two bit copy problem

X	Y	Z	P
00	0	0	1/4
01	0	1	1/4
10	1	0	1/4
01	1	1	1/4

Redundancy measures: $\mathrm{I}_{\mathsf{min}}$

Original measure of redundancy introduced by Williams and Beer

$$\mathrm{I}_{\mathsf{min}}(X : Y_1, \ldots, Y_k) = \sum_x p(x) \min_{Y_i} \mathrm{I}(X = x; Y_i)$$

- Semi-local approach: for each $X = x$ the redundant information is the minimum information provided by all of the sources Y_i
- Widely critised after its introduction — two bit copy problem

X	Y	Z	P
00	0	0	1/4
01	0	1	1/4
10	1	0	1/4
01	1	1	1/4

$$\mathrm{I}_{\mathsf{min}}(X : Y; Z) = 1 \text{ bit}$$

Redundancy measures: $\mathrm{I}_{\mathsf{min}}$

Original measure of redundancy introduced by Williams and Beer

$$\mathrm{I}_{\mathsf{min}}(X : Y_1, \dots, Y_k) = \sum_x p(x) \min_{Y_i} \mathrm{I}(X = x; Y_i)$$

- Semi-local approach: for each $X = x$ the redundant information is the minimum information provided by all of the sources Y_i

- Widely critised after its introduction — two bit copy problem

X	Y	Z	P
00	0	0	1/4
01	0	1	1/4
10	1	0	1/4
01	1	1	1/4

$$\mathrm{I}_{\mathsf{min}}(X : Y; Z) = 1 \text{ bit}$$

- "The problem is $\mathrm{I}_{\mathsf{min}}$ does not distinguish whether sources carry *the same* information or just *the same amount* of information"

Redundancy measures: $\mathrm{I}_{\mathsf{red}}$

Based on information geometry and introduced by Harder et al.

$$\mathrm{I}_{\mathsf{red}}(Z : X; Y) = \min\left\{ \mathrm{I}^{\pi}_{Z}(X \searrow Y)\,,\ \mathrm{I}^{\pi}_{Z}(X \searrow Y)\right\}$$

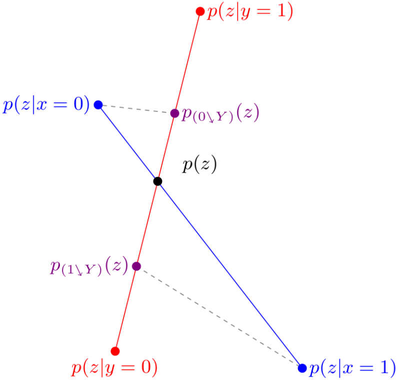
where $\mathrm{I}^{\pi}_{Z}(X \searrow Y)$ is the mutual information between Z and X expressed in terms of the mutual information between Z and Y.

Redundancy measures: I_{red}

Based on information geometry and introduced by Harder et al.

$$\mathrm{I}_{\text{red}}(Z : X; Y) = \min \left\{ \mathrm{I}^{\pi}_{Z}(X \searrow Y)\,,\; \mathrm{I}^{\pi}_{Z}(X \searrow Y) \right\}$$

where $\mathrm{I}^{\pi}_{Z}(X \searrow Y)$ is the mutual information between Z and X expressed in terms of the mutual information between Z and Y.

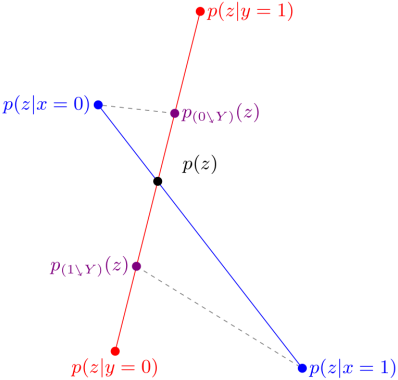


Redundancy measures: I_{red}

Based on information geometry and introduced by Harder et al.

$$\mathrm{I}_{\text{red}}(Z : X; Y) = \min \left\{ \mathrm{I}_Z^{\pi}(X \searrow Y) \,,\ \mathrm{I}_Z^{\pi}(X \searrow Y) \right\}$$

where $\mathrm{I}_Z^{\pi}(X \searrow Y)$ is the mutual information between Z and X expressed in terms of the mutual information between Z and Y.



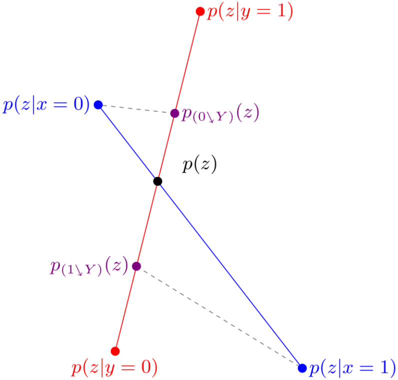
- Only able to quantify bivariate redundancy: multivariate extension highly non-trivial and evaluation is intractable

Redundancy measures: $\mathrm{I}_{\mathsf{red}}$

Based on information geometry and introduced by Harder et al.

$$\mathrm{I}_{\mathsf{red}}(Z : X; Y) = \min\left\{\, \mathrm{I}_Z^{\pi}(X \searrow Y)\,,\ \mathrm{I}_Z^{\pi}(X \searrow Y)\right\}$$

where $\mathrm{I}_Z^{\pi}(X \searrow Y)$ is the mutual information between Z and X expressed in terms of the mutual information between Z and Y.



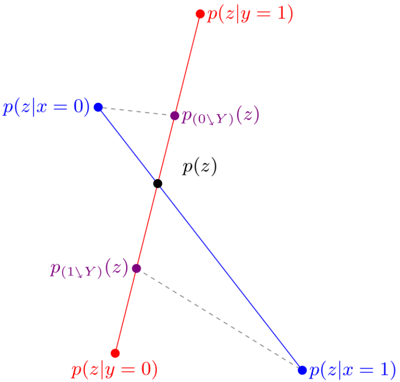
- Only able to quantify bivariate redundancy: multivariate extension highly non-trivial and evaluation is intractable
- Not even clear that it does indeed capture the redundant information

Redundancy measures: I_{red}

Based on information geometry and introduced by Harder et al.

$$\mathrm{I}_{\text{red}}(Z : X; Y) = \min \left\{ \mathrm{I}_Z^{\pi}(X \searrow Y),\ \mathrm{I}_Z^{\pi}(X \searrow Y) \right\}$$

where $\mathrm{I}_Z^{\pi}(X \searrow Y)$ is the mutual information between Z and X expressed in terms of the mutual information between Z and Y.



- Only able to quantify bivariate redundancy: multivariate extension highly non-trivial and evaluation is intractable
- Not even clear that it does indeed capture the redundant information
- No meaningful local intepretation

Unique information measure: $\widetilde{\mathrm{UI}}$

Indroduced by Bertschinger et al. — game-theoretic motivation

- Defining the unique information implicitly defines the redundant information in the partial information decomposition framework

Unique information measure: $\widetilde{\mathrm{UI}}$

Indroduced by Bertschinger et al. — game-theoretic motivation

- Defining the unique information implicitly defines the redundant information in the partial information decomposition framework
- If a source contains unique information then there must be a way to exploit this information in a decision problem

Unique information measure: $\widetilde{\mathrm{UI}}$

Indroduced by Bertschinger et al. — game-theoretic motivation

- Defining the unique information implicitly defines the redundant information in the partial information decomposition framework
- If a source contains unique information then there must be a way to exploit this information in a decision problem
- No unique local intepretation

Unique information measure: $\widetilde{\mathrm{UI}}$

Indroduced by Bertschinger et al. — game-theoretic motivation

- Defining the unique information implicitly defines the redundant information in the partial information decomposition framework
- If a source contains unique information then there must be a way to exploit this information in a decision problem
- No unique local intepretation
- Worse than that

X	Y	Z	P
0	0	0	1/2
1	0	1	1/4
1	1	0	1/4

Unique information measure: $\widetilde{\mathrm{UI}}$

Indroduced by Bertschinger et al. — game-theoretic motivation

- Defining the unique information implicitly defines the redundant information in the partial information decomposition framework
- If a source contains unique information then there must be a way to exploit this information in a decision problem
- No unique local intepretation
- Worse than that

X	Y	Z	P
0	0	0	1/2
1	0	1	1/4
1	1	0	1/4

$$\widetilde{\mathrm{UI}}(X : Y) = \widetilde{\mathrm{UI}}(X : Y) = 0 \text{ bit}$$