



Statistical methods in information theory

gene network reconstruction | epigenetic causal discovery

Soumik Purkayastha

Department of Biostatistics, University of Michigan

2023-12-15

Why information theory?



Digital Communications

Information Theory laid the foundations for digital signal processing and telecommunication protocols, over which all wireless, wireline and satellite networks.



Data Storage

Information Theory is the basis for efficient and compact data encoding and compression, which all digital storage depends on today.



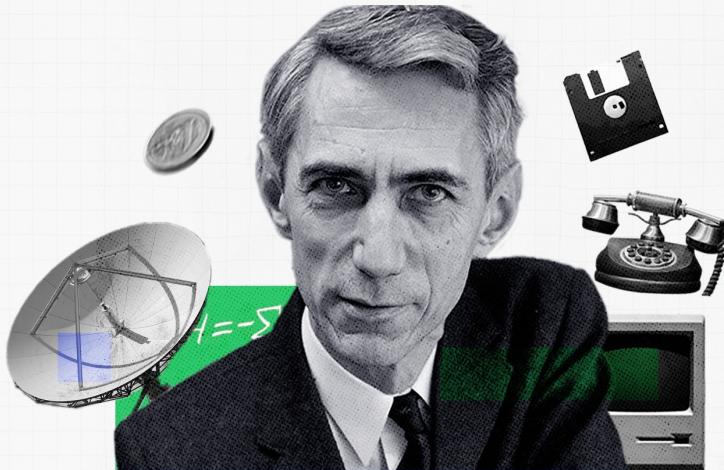
Digital Media

Information Theory defined the principles behind compression algorithms, which allows high-quality media files to be stored and or streamed.



Computing

Binary code – 1s and 0s -- is at the heart of computing systems. Information Theory enabled the processing, storage and retrieval of data in binary form, making modern computing possible.



```
1101 01111011 11011001 11001010 11101000 1001111  
0010 01011111 11010011 10001100 10001101 0111010  
0010 01011011 11000101 10001011 01000011 0100011  
0100 01011011 01001111 10100101 01111001 0011100  
0011 00000100 000001011 10011100 00101000 00010110  
0111 00011000 10011101 01111011 01011010 10001101  
0101 11110001 00110111 00100100 11010110 10101101  
1110 11101111 10100000 10010111 00100001 00010111
```



Internet

Without Shannon, the internet simply would not exist. Information Theory defined the “bit,” the basic unit of digital information on which the internet is built.



AI/ML

Shannon's ideas on information gain and entropy are crucial for AI systems' decision-making processes and in the creation of more accurate AI models.



Cryptography

Information Theory established a framework for secure communication by introducing the concepts of randomness and entropy, which are key to modern cryptographic systems and practices.

“Information Theory Turns 75.” Nokia Bell Labs, www.bell-labs.com/about/history/innovation-stories/information-theory-turns-75/. Accessed 12 Dec. 2023.



Quantifying complex association using mutual information

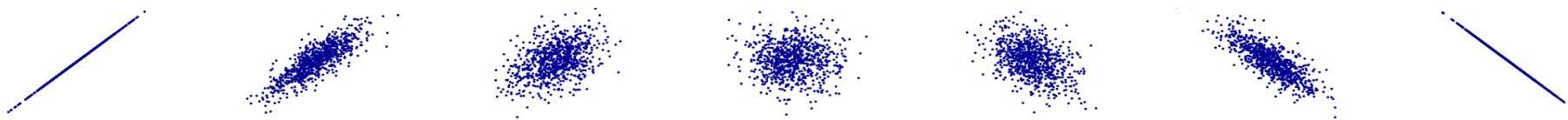
<gene network reconstruction>

Association studies

1. Measuring dependence in data is a key issue in statistics.
2. Popular measures of correlation include:
 1. Moment based: Pearson's r
 2. Rank based: Spearman's ρ , Kendall's τ
 3. Distance based: HHG (*Heller et al. (2013)*)
3. These measures are bad at detecting complex, non-linear relationships.

Heller, Ruth, Yair Heller, and Malka Gorfine. "A consistent multivariate test of association based on ranks of distances." Biometrika 100.2 (2013): 503-510.

Complex dependence

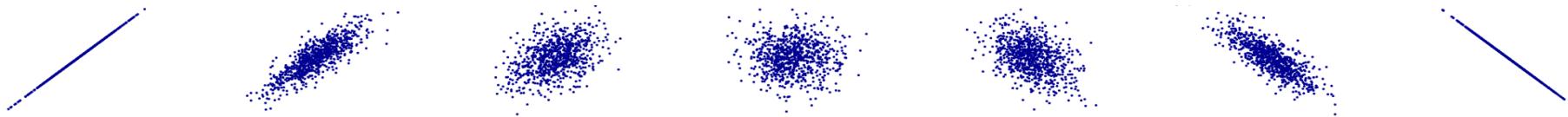


Here are some typical examples of scatterplots for bivariate data

But more complex patterns CAN exist!!



Complex dependence



Pearson, Spearman, Kendall, and HHG:
“These are ‘good’ data to capture association 

Pathological examples:
“These are ‘bad’ data to capture association A horizontal row of seven scatter plots showing data points forming various non-linear and non-monotonic patterns. From left to right: a W-shaped cloud of points; a U-shaped cloud of points; a diamond-shaped cloud of points; a U-shaped cloud of points; an X-shaped cloud of points; a circular cloud of points; and a dense, irregular cloud of points.

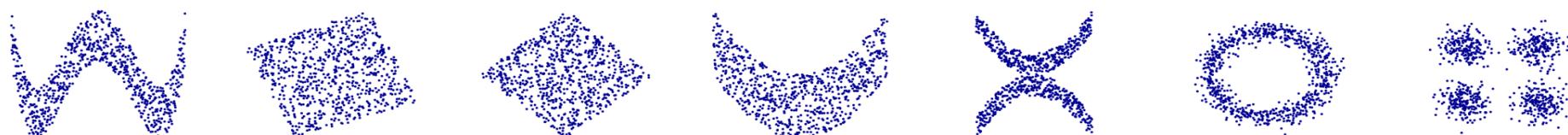
Complex dependence



Previous attempts at understanding association: moment-based, rank-based, and distance-based.

Claude E. Shannon: focused on the *densities* of random variables to propose ***Mutual Information (MI)***.

Using information theory, MI is able to detect association in ALL these cases.



Association in information theory

1. Information theory = probability + statistics + engineering.
2. Theoretical framework for quantifying information in statistical models.
3. Mutual information is a powerful measure of association!

Shannon (1948): Consider random variables X and Y with marginal probability functions p_X and p_Y with joint probability function p_{XY} .

$$MI(X, Y) = E_{XY} \left[\log \left(\frac{p_{XY}(X, Y)}{p_X(X)p_Y(Y)} \right) \right]$$

Shannon, C. E. "A Mathematical Theory of Communication." Bell System Technical Journal 27, no. 3 (1948): 379–423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>.

Properties of $MI(X, Y)$

$$MI(X, Y) = E_{XY} \left[\log \left(\frac{p_{XY}(X, Y)}{p_X(X)p_Y(Y)} \right) \right]$$

Information theory: $MI(X, Y)$ quantifies information exchanged between the two random variables.

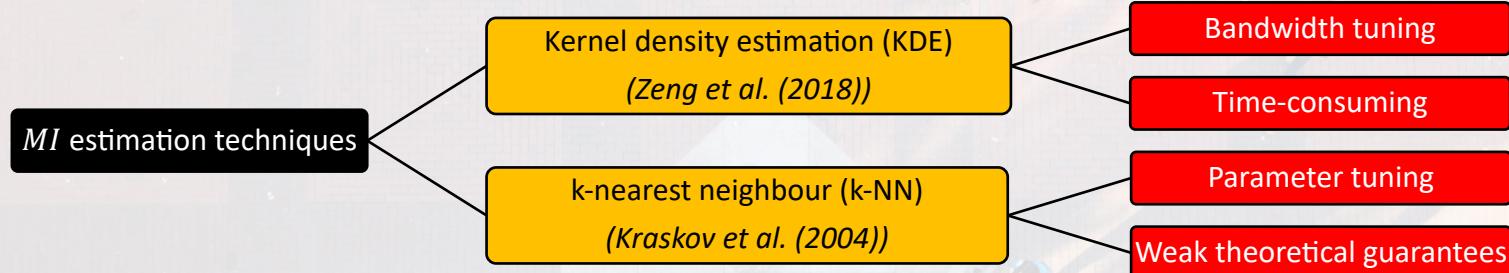
1. Measure of dependence: $MI(X, Y) = 0 \Leftrightarrow X$ and Y are independent.
2. Symmetric and always non-negative (*Kullback and Leibler (1951)*).
3. Measures ‘distance’ between **observed joint distribution** and **hypothetical independence distribution** through likelihood ratio:

$$\log(p_{XY}(X, Y)/p_X(X)p_Y(Y)).$$

Kullback, Solomon, and Richard A. Leibler. "On information and sufficiency." The annals of mathematical statistics 22.1 (1951): 79-86.

Challenge: estimation and inference

1. Estimation of MI is a non-parametric statistical problem.

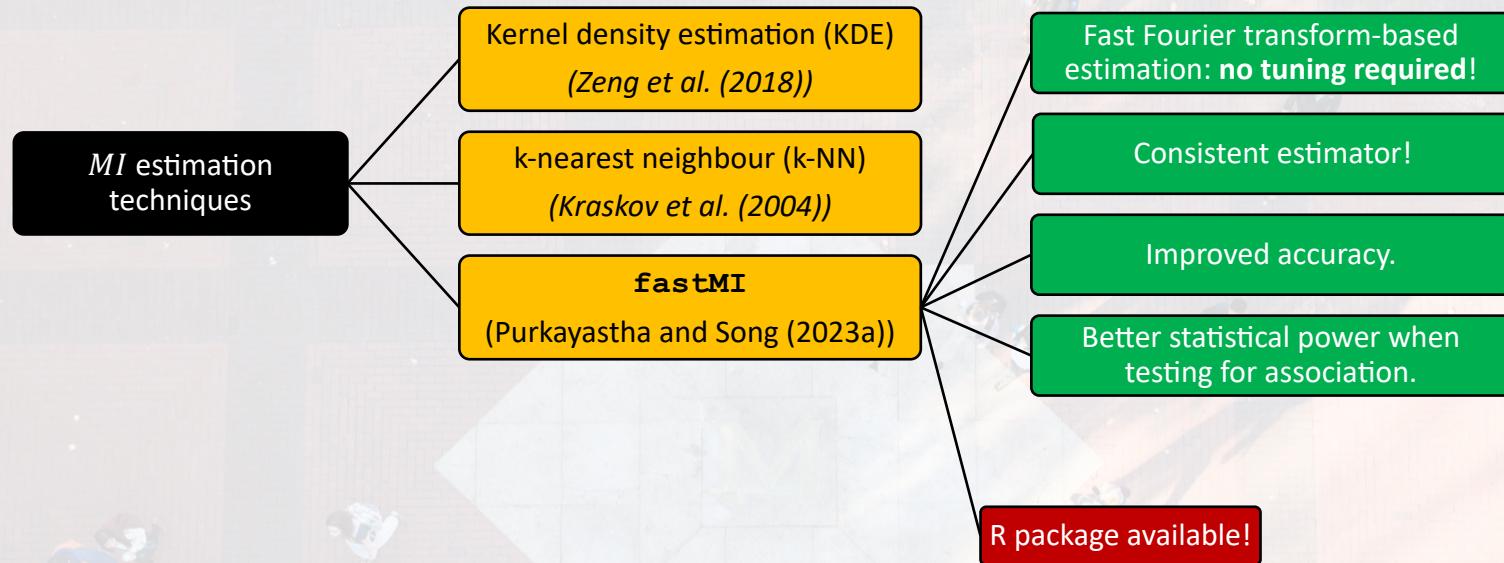


2. Typical challenges involve:
 1. Estimation without incurring predetermined tuning parameters.
 2. Theoretical guarantees on performance: consistency and asymptotic normality.
 3. Establishing an *MI*-based rule to test for independence.

1. Zeng, Xianli, Yingcun Xia, and Howell Tong. "Jackknife approach to the estimation of mutual information." *Proceedings of the National Academy of Sciences* 115.40 (2018): 9956-9961

2. Kraskov, Alexander, Harald Stögbauer, and Peter Grassberger. "Estimating mutual information." *Physical review E* 69.6 (2004): 066138.

New work: **fastMI**



Purkayastha, Soumik, and Peter X.-K. Song. "fastMI: A Fast and Consistent Copula-Based Nonparametric Estimator of Mutual Information." *Journal of Multivariate Analysis*, November 2023, 105270.

fastMI beats current standard!

1. Copula-based: robust to outliers in data.
2. Leverage tuning-free density estimation through Fourier transformation.
3. Simulations: improved accuracy and reduced run-time.
4. Empirical power comparison: fastMI beats standard estimators.



The image shows a screenshot of a journal article from the Journal of Multivariate Analysis. The article is titled "fastMI: A fast and consistent copula-based nonparametric estimator of mutual information". It is authored by Soumik Purkayastha¹, Peter X.-K. Song^{*1}. The article is from the Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA. The page includes sections for Article Info (AMS 2020 subject classifications: primary 62H12, secondary 62G10), Abstract (describing the development of a consistent and powerful estimator for mutual information using Fast Fourier transform-based estimation of the underlying density), and Keywords (Copula, Fast Fourier transformation, Kernel density estimation, Permutation test, Statistical dependence). The journal logo for Elsevier is visible at the top left, and the journal title "Journal of Multivariate Analysis" is at the top right. The journal's homepage URL is provided as well.



Application I

fastMI-based gene network reconstruction using gene expression data

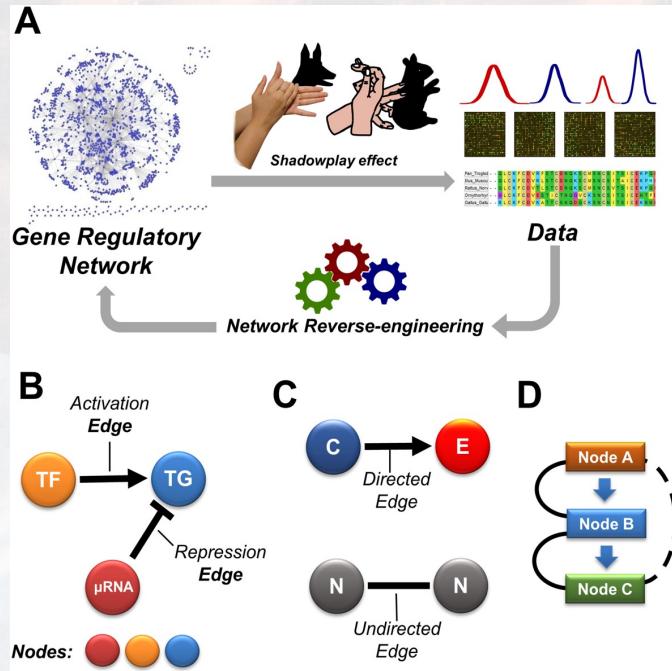
Scientific background

Transcriptional regulation is a fundamental molecular mechanism described by Gene Regulatory Networks (GRNs).

Latent: GRN controls transcriptional regulation in living cells.

Observable: “projection” of GRN that generates gene expression data from qualitative and quantitative empirical measurements.

Problem: Reverse engineer observable data to infer latent structure.



Mercatelli, Daniele, et al. "Gene regulatory network inference resources: A practical overview." *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* 1863.6 (2020): 194430.

Inferring GRNs using MI and PCA

Zhang et al. (2012) first suggested inferring GRNs by employing path consistency algorithm (PCA, *Spirites et al. (2000)*) based on *MI*.

1. MI estimated using Gaussian kernel.
2. Asymptotic behaviour used when inferring gene network is not correct.

Conditional mutual information (*CMI*): measures association between X and Y conditioned on a third variable Z :

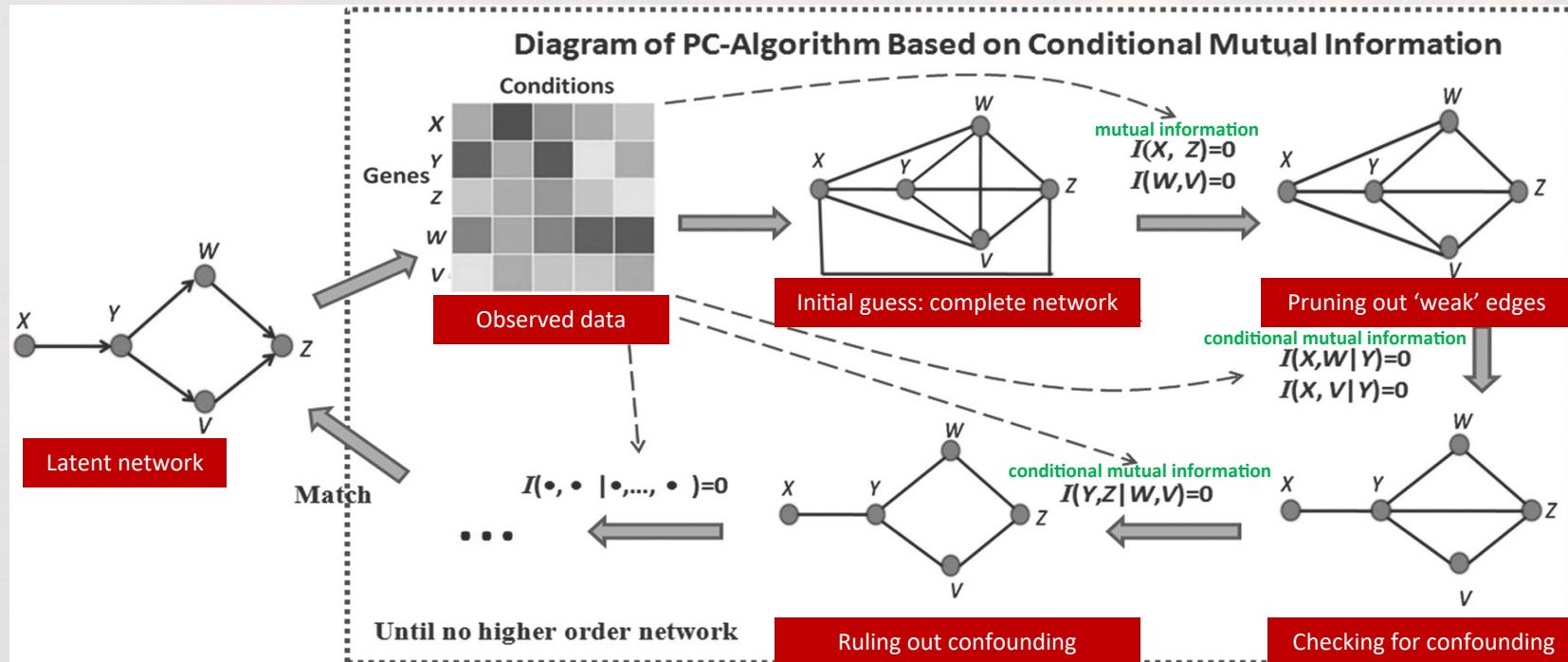
$$CMI(X, Y|Z) = MI(X, Y) - MI(X, Y, Z)$$

$$CMI(X, Y|Z) = 0 \Leftrightarrow X \text{ and } Y \text{ are independent, conditioned on } Z.$$

1. Zhang, Xiujun, Xing-Ming Zhao, Kun He, Le Lu, Yongwei Cao, Jingdong Liu, Jin-Kao Hao, Zhi-Ping Liu, and Luonan Chen. "Inferring Gene Regulatory Networks from Gene Expression Data by Path Consistency Algorithm Based on Conditional Mutual Information." *Bioinformatics* 28, no. 1 (January 1, 2012): 98–104.

2. Spirtes, Peter, Clark N. Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

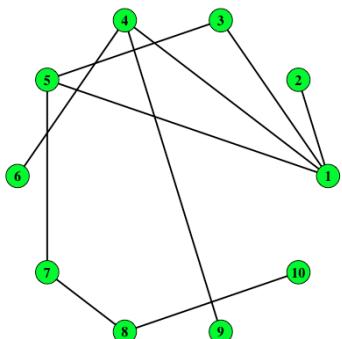
PC algorithm: GRN reconstruction



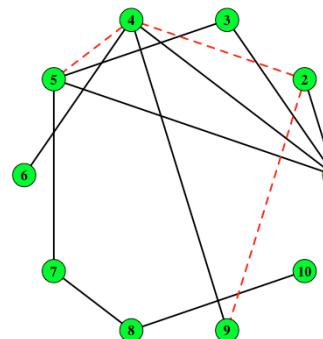
Zhang, Xijun, Xing-Ming Zhao, Kun He, Le Lu, Yongwei Cao, Jingdong Liu, Jin-Kao Hao, Zhi-Ping Liu, and Luonan Chen. "Inferring Gene Regulatory Networks from Gene Expression Data by Path Consistency Algorithm Based on Conditional Mutual Information." *Bioinformatics* 28, no. 1 (January 1, 2012): 98–104.

GRN in DREAM3 dataset

DREAM3: benchmark gene expression data with known underlying GRN. Aim is to predict GRN from *in vivo* and *in silico* gene expression datasets.

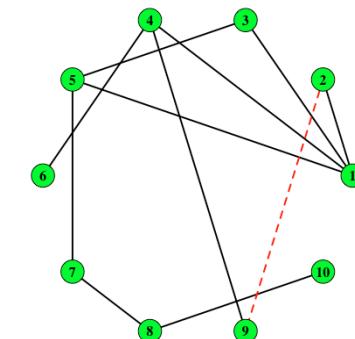


Known GRN in DREAM 3



Inferred GRN in Zhang (2012)

3 false positives

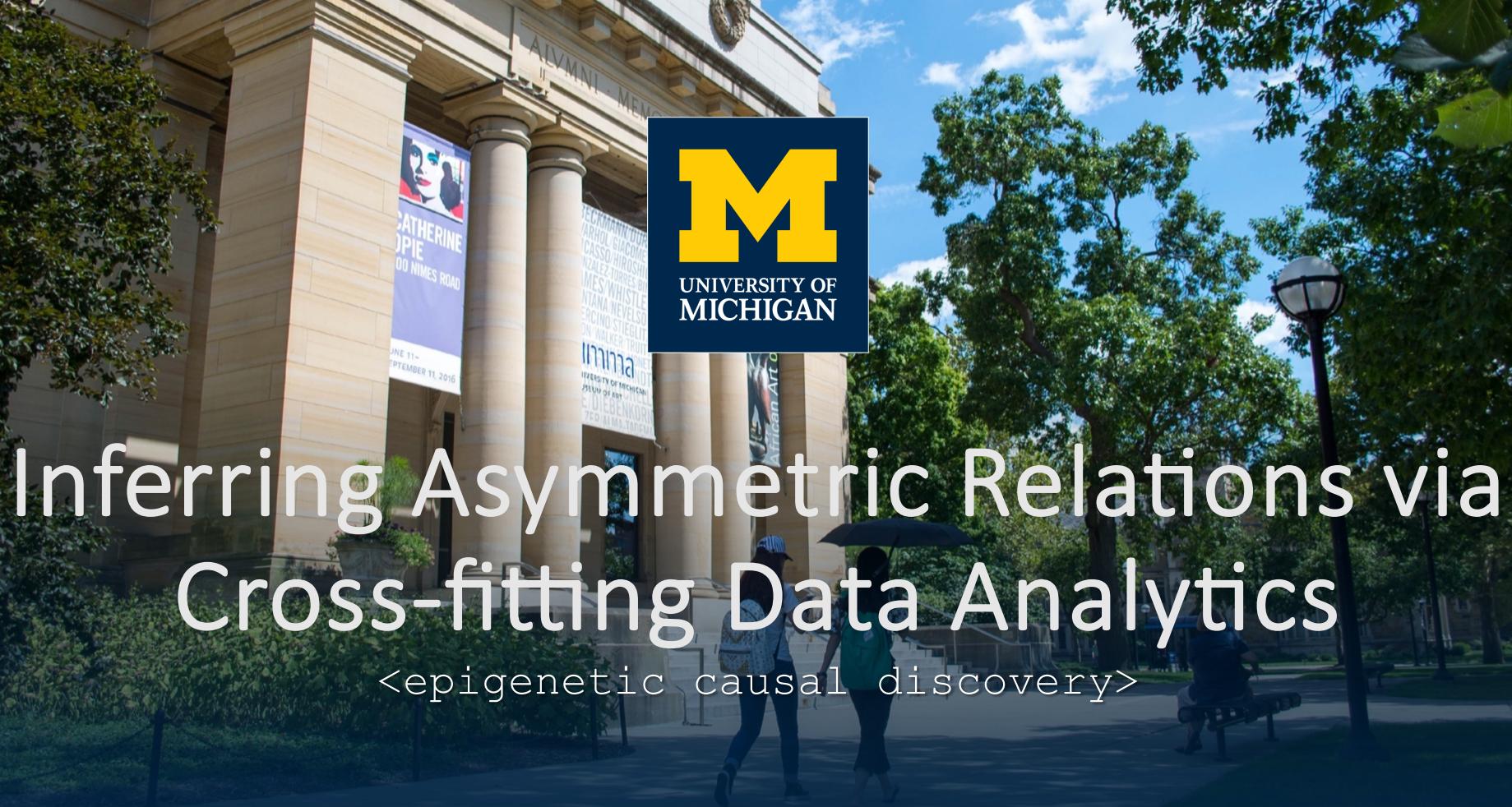


Our inferred GRN

1 false positive

Summary

1. Need for association measures that are:
 1. Able to detect complex association
 2. Have attractive theoretical properties
 3. Exhibit good practical performance (reduced error, improved runtime).
2. fastMI uses tuning-free estimation strategy; is a powerful test of independence.
3. Using MI, we can infer gene networks.

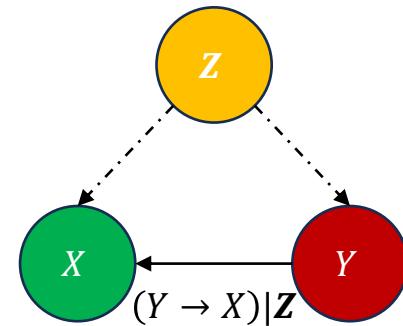
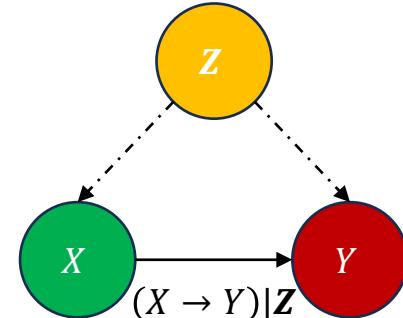


Inferring Asymmetric Relations via Cross-fitting Data Analytics

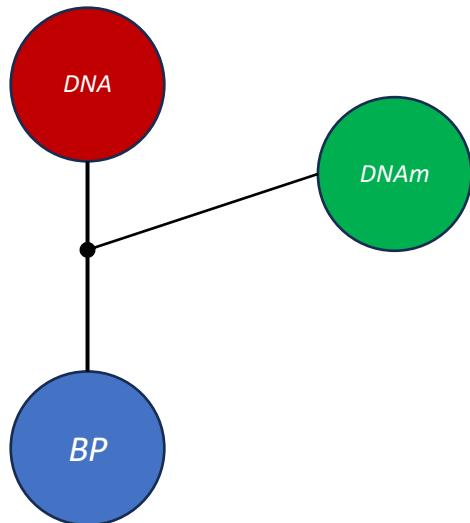
<epigenetic causal discovery>

Studying asymmetry

1. **Problem:** Investigate asymmetry in variables (X, Y) given confounders Z .
2. Motivated by Shannon's seminal work on entropy.
3. Data-splitting and cross-fitting inference technique.



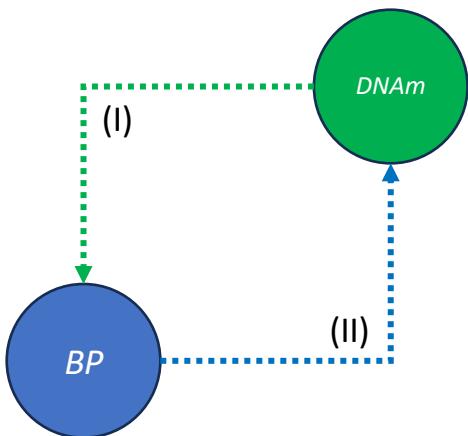
Asymmetry in DNAm and BP.



1. ELEMENT* cohort study
2. DNA associated with blood pressure: **ATP2B1, FGF5, and PRDM8.**
3. DNA expression controlled by methylation DNAm.
4. Epigenetics and cardiovascular function such as blood pressure (BP) are closely related (*Shi et al, 2022*).

*ELEMENT : Early Life Exposures in Mexico to Environmental Toxicants.

Does *DNAm* → *BP* or *BP* → *DNAm*



1. ELEMENT* cohort study
2. DNA associated with blood pressure: ***ATP2B1*, *FGF5*, and *PRDM8*.**
3. DNA expression controlled by methylation *DNAm*.
4. Epigenetics and cardiovascular function such as blood pressure (BP) are closely related (*Shi et al, 2022*).
5. Confirm if (I) ***DNAm* → *BP*** or (II) ***BP* → *DNAm***?

*ELEMENT : Early Life Exposures in Mexico to Environmental Toxicants.

Is there asymmetry in (X, Y) ?

1. Start with simpler case of no confounders.
2. Weak asymmetry framework for studying statistical asymmetries between two variables.
 - Based on fundamental **entropy decomposition equation**.
3. New information theory-based measure $C_{X>Y}$ to quantify asymmetry between X and Y .

Entropy decomposition equation

(X, Y) with joint (copula) density (f_{XY}) c_{XY} and marginal densities f_X and f_Y .

$$\mathbf{H}(\mathbf{X}, \mathbf{Y}) = \textcolor{green}{MI}(\mathbf{X}, \mathbf{Y}) + \mathbf{H}(\mathbf{X}|\mathbf{Y}) + \mathbf{H}(\mathbf{Y}|\mathbf{X})$$

- Joint entropy measures total information in (X, Y)

$$\mathbf{H}(\mathbf{X}, \mathbf{Y}) := \mathbf{E}_{\mathbf{XY}}\{-\log(f_{XY})\}$$

- Marginal entropies $\mathbf{H}(\mathbf{X})$ and $\mathbf{H}(\mathbf{Y})$ defined similarly.

Entropy decomposition equation

(X, Y) with joint (copula) density (f_{XY}) c_{XY} and marginal densities f_X and f_Y .

$$H(X, Y) = MI(X, Y) + H(X|Y) + H(Y|X)$$

- Joint entropy measures total information in (X, Y)

$$H(X, Y) := E_{XY}\{-\log(f_{XY})\}$$

- Marginal entropies $H(X)$ and $H(Y)$ defined similarly.

- MI is copula entropy

$$MI(X, Y) := E_{XY}\{\log(c_{XY})\}$$

- Test for independence

$$MI = 0 \Leftrightarrow X \perp Y.$$

Entropy decomposition equation

(X, Y) with joint (copula) density (f_{XY}) c_{XY} and marginal densities f_X and f_Y .

$$H(X, Y) = MI(X, Y) + H(X|Y) + H(Y|X)$$

- Joint entropy measures total information in (X, Y)

$$H(X, Y) := E_{XY}\{-\log(f_{XY})\}$$

- Marginal entropies $H(X)$ and $H(Y)$ defined similarly.

- MI is copula entropy

$$MI(X, Y) := E_{XY}\{\log(c_{XY})\}$$

- Test for independence

$$MI = 0 \Leftrightarrow X \perp Y.$$

- Conditional entropy $H(Y|X)$ randomness in Y if we know X

$$H(X|Y) := E_{XY}\{-\log(f_{XY}/f_X)\}$$

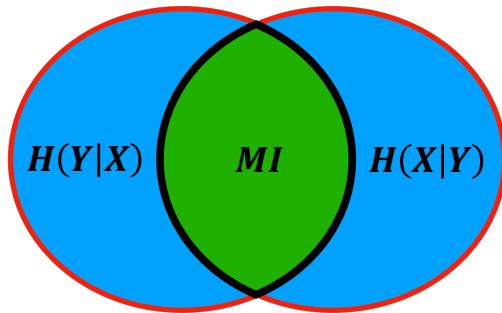
- Asymmetric measure and forms basis of weak asymmetry!

Entropy decomposition equation

(X, Y) with joint (copula) density (f_{XY}) c_{XY} and marginal densities f_X and f_Y .

$$H(X, Y) = MI(X, Y) + H(X|Y) + H(Y|X)$$

Symmetric if $H(X|Y) = H(Y|X)$



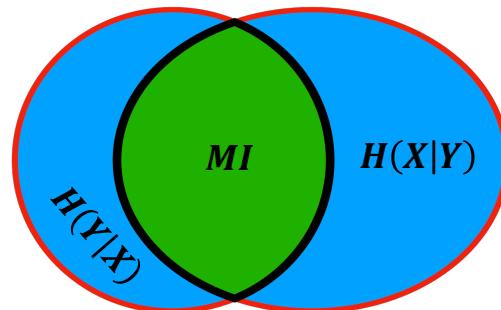
Entropy decomposition equation

(X, Y) with joint (copula) density (f_{XY}) c_{XY} and marginal densities f_X and f_Y .

$$H(X, Y) = MI(X, Y) + H(X|Y) + H(Y|X)$$

“Less uncertainty in Y after conditioning on Y than the converse”

$$X \succ_w Y \text{ if } H(X|Y) > H(Y|X)$$



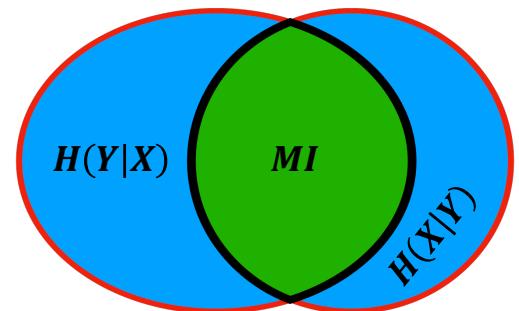
Entropy decomposition equation

(X, Y) with joint (copula) density (f_{XY}) c_{XY} and marginal densities f_X and f_Y .

$$H(X, Y) = MI(X, Y) + H(X|Y) + H(Y|X)$$

"Less uncertainty in X after conditioning on Y than the converse"

$$Y >_w X \text{ if } H(Y|X) > H(X|Y)$$



Weak asymmetry coefficient $C_{X>Y}$

- Comparison reveals potential asymmetry in predictability:

$$H(X|Y) \leq H(Y|X)$$

- “Asymmetric predictability is an imprint of underlying causal direction”
- The weak asymmetry coefficient (WAC)

$$C_{X>Y} = H(X|Y) - H(Y|X) = H(X) - H(Y).$$

- Last equality from chain rule of entropy:

$$H(X|Y) + H(Y) = H(X, Y) = H(Y|X) + H(X)$$

Adjusting for confounders \mathbf{Z}

- Unadjusted weak asymmetry coefficient (WAC)

$$C_{X>Y} = H(X) - H(Y).$$

- Adjusting for confounders \mathbf{Z} by examining conditional WACs
- Case 1: WAC for specific value of $\mathbf{Z} = \mathbf{z}_0$

$$C_{X>Y|Z=z_0} = H(X|\mathbf{Z} = \mathbf{z}_0) - H(Y|\mathbf{Z} = \mathbf{z}_0).$$

- Case 2: Aggregated WAC for all possible values of \mathbf{z}_0

$$C_{X>Y|Z} = H(X|\mathbf{Z}) - H(Y|\mathbf{Z}).$$

Density estimation

- To estimate $H(X)$ and $H(X|\mathbf{Z} = \mathbf{z}_0)$, need plug-in estimators of underlying density.
- Bernacchia et al. (2011): consider the inverse Fourier transform of the density estimator \hat{f}_X , given by

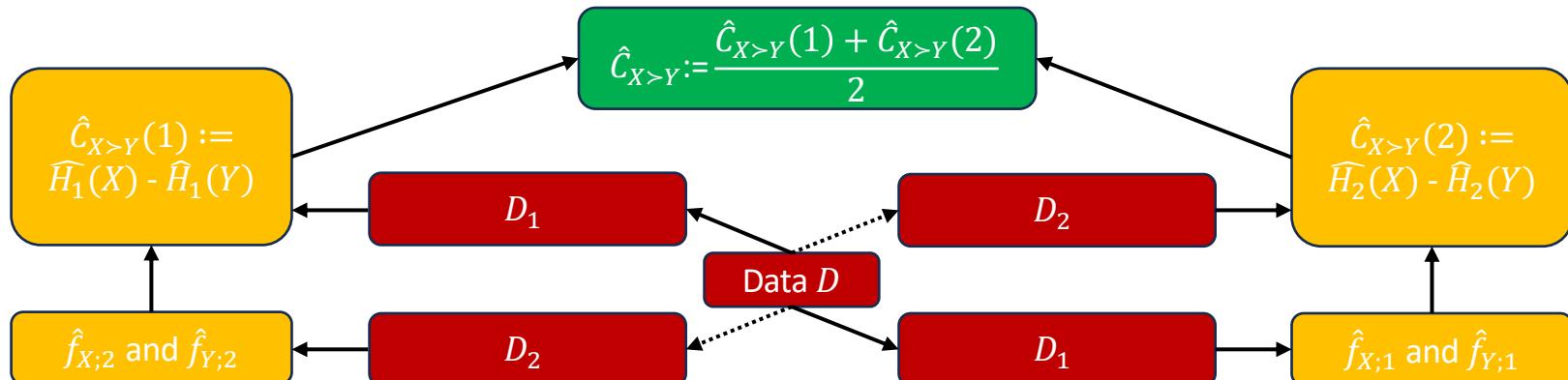
$$\hat{\phi}(t) = \frac{n c(t)}{2(n-1)} \left[1 + \sqrt{1 - \frac{4(n-1)}{|n c(t)|^2}} \right] A_n(t),$$

where A_n is a filter that ensures a stable estimator \hat{f}_X , which is anti-transformed back from $\hat{\phi}$.

- Under some regularity conditions (Purkayastha and Song (2023a)), \hat{f}_X uniformly converges almost surely to \hat{f}_X .

Data-splitting and cross-fitting

- \hat{f}_X and \hat{f}_Y : infinite-dimensional nuisance parameters; using same data for estimating density and inference for $\hat{C}_{X>Y}$ may lead to bias!
- Data-splitting and cross-fitting
 1. Split data D into two halves D_1 and D_2 .
 2. Use D_2 to estimate \hat{f}_X and \hat{f}_Y and use \hat{f}_X and \hat{f}_Y and D_1 to obtain $\hat{C}_{X>Y}(1)$.
 3. Interchange roles of D_1 and D_2 to obtain $\hat{C}_{X>Y}(2)$ to obtain cross-fitted $\hat{C}_{X>Y}$.



Cross-fitting inference

Setup: Given data $\{(x_i, y_i)\}_{i=1}^{2n}$ on (X, Y) , consider data splits

$$D_1 := \{(x_i, y_i)\}_{i=1}^n \text{ and } D_2 := \{(x_{n+i}, y_{n+i})\}_{i=1}^n$$

Assuming the following regularity conditions (Purkayastha and Song (2023b))

1. The regularity conditions in Purkayastha and Song (2023a) hold for both data splits D_1 and D_2 hold when estimating $\hat{C}_{X>Y}(1)$ and $\hat{C}_{X>Y}(2)$.
2. The true underlying densities f_X and f_Y are bounded below and above, we get the limiting distribution

$$\sqrt{n}(\hat{C}_{X>Y} - C_{X>Y}) \xrightarrow{D} N(0, \sigma_C^2) \text{ as } n \rightarrow \infty.$$

where $\sigma_C^2 := V_{XY}[\log(f_X(X)) + \log(f_Y(Y))]$ can be estimated by Monte-Carlo methods given estimated densities \hat{f}_X and \hat{f}_Y .



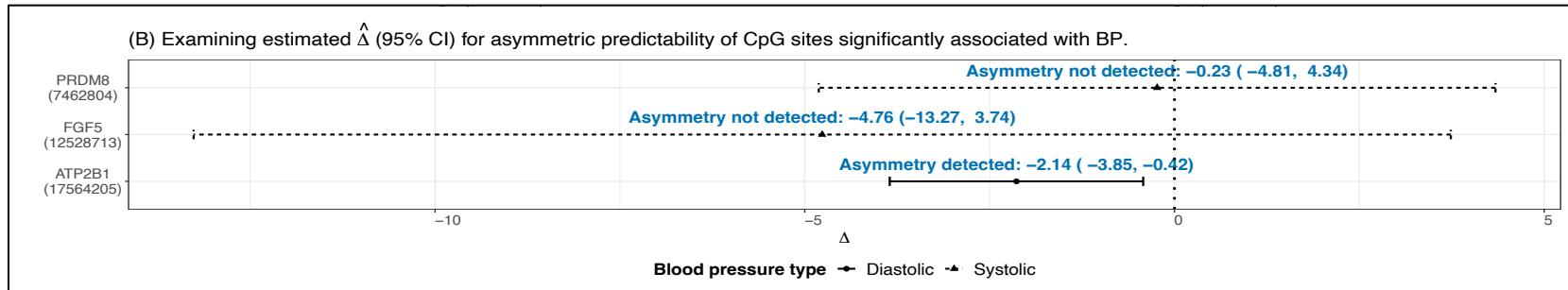
Application II

$\hat{C}_{X>Y}$ based epigenetic causal discovery

Asymmetry in epigenetics

1. Investigate asymmetry between DNA methylation (*DNAm*) and blood pressure (*BP*) variation in ELEMENT dataset ($n = 525$)
2. Three candidate genes: ***ATP2B1***, ***FGF5***, and ***PRDM8*** known for significant association with BP.
3. Using fastMI (Purkayastha and Song (2023a)) detect methylation sites within each gene that are significantly associated with BP after multiple testing correction.
4. Hits:
 1. ***ATP2B1*** has 21 sites; only 1 (***ATP2B1:CG17564205***) is associated with diastolic BP (DBP).
 2. ***FGF5*** also has 21 sites, only 1 (***FGF5:CG12528713***) is associated with systolic BP (SBP).
 3. ***PRDM8*** has 51 sites, only 1 (***PRDM8:CG7462804***) is associated with SBP.
5. Aim: For each of the hits above, investigate if $DNAm \succ_w BP$ or $BP \succ_w DNAm$.

Weak asymmetry: $DBP \succ_w DNAm$



Findings:

1. Sites **FGF5:CG12528713** and **PRDM8:CG7462804** are significantly associated with SBP but do not exhibit asymmetry.
2. Site **ATP2B1:CG17564205** exhibits asymmetry with DBP. We note that $\hat{C}_{DNAm > DBP}$ is negative with its 95% CI not covering zero: we conclude $DBP \succ_w DNAm$.



Thank you for your time!

soumikp@umich.edu

soumikp.github.io