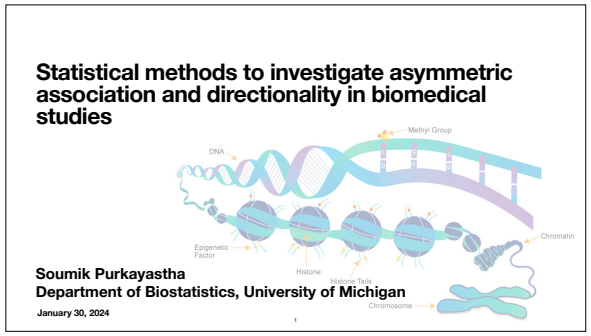
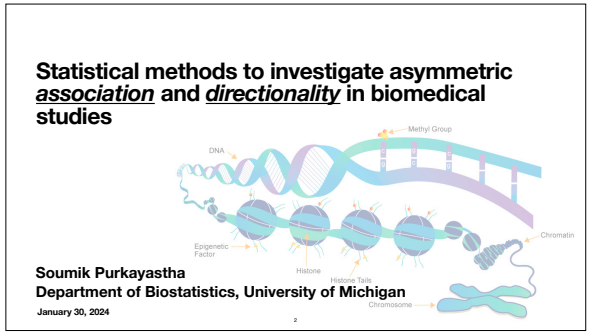
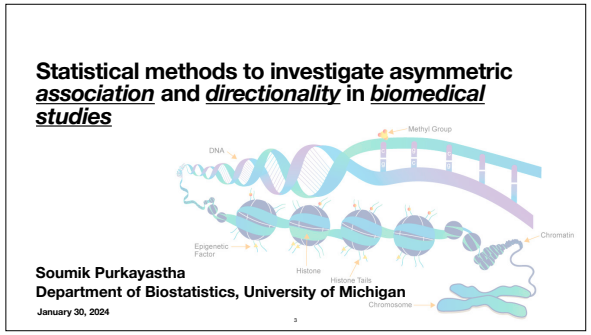


1



2

Today I hope to speak about two key statistical concepts: association and directionality. While the bulk of statistical methods focus on ascertaining the strength of association between features of a dataset, there is a scarcity of methods which investigate directionality. For example, do happier people earn more money, or does earning more money make you happier? Or more formally, when examining the relationship between income and happiness, is there a driving variable?



3

In a biostatistical context, let us focus on *epigenomics*. If asked for an elevator pitch for this talk, I would describe it as a summary of a toolkit that examines the pathway between genes and certain phenotypes of interest, with epigenetic methylation sites serving as intermediaries.

4

Epigenomics

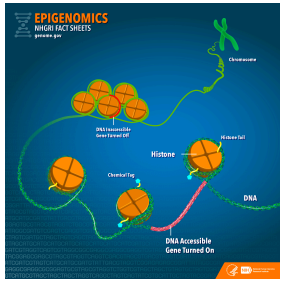
Motivating problem

4

5

Epigenomics

- Genes regulate how to make proteins.
- Epigenomics studies heritable changes in gene expression.
- Impacted by lifestyle, social and environmental determinants.

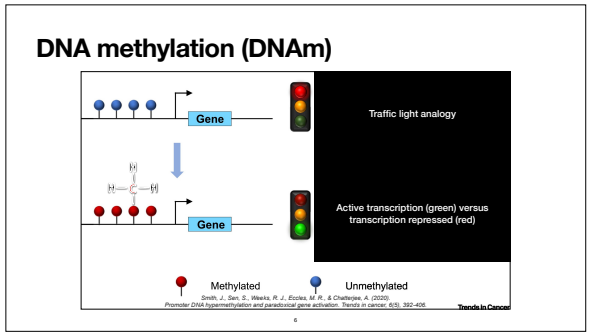


https://www.genome.gov/about-genomics/fact-sheets/Epigenomics-Fact-Sheet

5

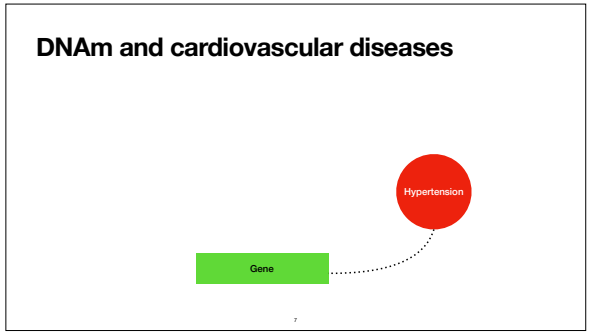
Epigenomics is the study of heritable changes in gene function that do not involve alterations to the underlying DNA sequence. It explores modifications such as DNA methylation and histone modification, providing insights into how environmental factors can influence gene expression and impact cellular function.

Think of the human lifespan as a very long movie.
The cells would be the the artists that make up the movie.
DNA, in turn, would be the script — instructions for all the participants of the movie to perform their roles and the concept of genetics would be like screenwriting.
The concept of epigenetics, then, would be like directing.



6

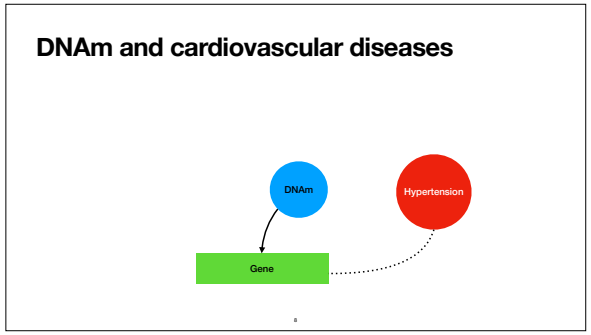
The epigenome directs ‘gene traffic’ based on the level of methylation at a given location in the epigenome.



7

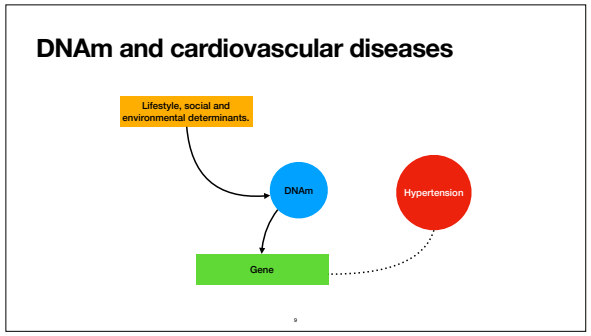
I study methylation in the context of cardiovascular diseases, specifically hypertension and this is what the problem looks like in my head.

Genome wide studies have unearthed association between many genes and hypertension.



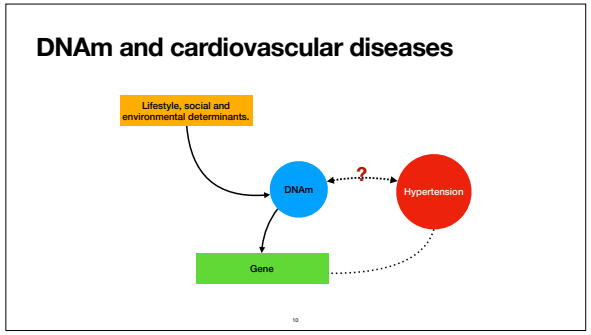
8

We also know of DNA methylation, which plays a regulatory role in gene expression,



9

and is in turn influenced by external stimuli.



10

I explore a certain sense of directionality between changes in a person's epigenome and their cardiovascular profile.

But inferring a certain sense of directionality between changes in a person's epigenome and their cardiovascular profile remains largely unsolved. This would be great not only from a clinical and practical standpoint, but also complement the wide body of causal inference literature in which we assume an underlying structure between exposure and outcome without actually confirming if indeed there is a mechanism that maps the exposure to the outcome.

Epigenetics and blood pressure
Leader or follower?

- Conventional knowledge: DNAm sites are associated with BP.

11

11

So the scientific question is, can I infer a certain sense of directionality between a person's epigenome and their blood pressure?

Conventional knowledge tells us that methylation status and blood pressure are likely to be associated.

Epigenetics and blood pressure
Leader or follower?

- Conventional knowledge: DNAm sites are associated with BP.
- Very few studies establish a direction or ordering between DNAm and BP.

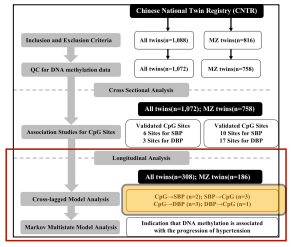
12

12

But the field of research into directionality is very new. However, I am certainly not the first to examine this problem.

Epigenetics and blood pressure
Leader or follower?

- Conventional knowledge: DNAm sites are associated with BP.
- Very few studies establish a direction or ordering between DNAm and BP.
- Hong et al. (2023): directionality from a predictive angle.



13

13

In a recent study published last year, a group working with data from the Chinese National Twin Registry reported directionality in methylation sites and blood pressure from a predictive accuracy perspective. I think this is good news, it's an interesting approach and I like the findings. However, as a statistician I must insist on some kind of uncertainty quantification. Are these point estimates of prediction accuracy even reliable?

Does DNAm drive BP or vice-versa?

What data?

Which genes?

Biological considerations?

14

14

To answer the directionality question, I first establish some basic facts.

Does DNAm drive BP or vice-versa?

Data: Early Life Exposures in Mexico to Environmental Toxicants (ELEMENT) cohort study.

Perng, W., Tang, L., Song, P. X., Tellez-Rojo, M. M., Cantoral, A., & Peterson, K. E. (2019). Metabolomic profiles and development of metabolic risk during the pubertal transition: a prospective study in the ELEMENT Project. Pediatric research, 85(3), 262-268.

15

15

I will study this directionality in the ELEMENT cohort. ELEMENT is short for.

Does DNAm drive BP or vice-versa?

Data: Early Life Exposures in Mexico to Environmental Toxicants (ELEMENT) cohort study.

Perng, W., Tang, L., Song, P. X., Tellez-Rojo, M. M., Cantoral, A., & Peterson, K. E. (2019). Metabolomic profiles and development of metabolic risk during the pubertal transition: a prospective study in the ELEMENT Project. Pediatric research, 85(3), 262-268.

Candidate gene analysis: only examine DNAm for genes that are known to be associated with blood pressure.

ARHGAP42, ATP2B1, FGF5, HSD11B2, KCNK3, PRDM8

16

16

Next, I performed a candidate gene analysis. I focused on six candidate genes which have all been reported to be significantly associated with blood pressure in multiple GWASs.

Does DNAm drive BP or vice-versa?

Data: Early Life Exposures in Mexico to Environmental Toxicants (ELEMENT) cohort study.

Perng, W., Tang, L., Song, P. X., Tellez-Rojo, M. M., Cantoral, A., & Peterson, K. E. (2019). Metabolomic profiles and development of metabolic risk during the pubertal transition: a prospective study in the ELEMENT Project. Pediatric research, 85(3), 262-268.

Candidate gene analysis: only examine DNAm for genes that are known to be associated with blood pressure.

ARHGAP42, ATP2B1, FGF5, HSD11B2, KCNK3, PRDM8

Biological considerations: n = 522 children aged 10 - 18 years, stratified by sex.

17

17

And finally, since the cohort is made up of children in the developmental stages of 10-18 years, it is definitely a good idea to allow for sex-based differences in the relationship between DNA methylation and blood pressure. So I did a sex-stratified analysis.

Does DNAm drive BP or vice-versa?

Data: Early Life Exposures in Mexico to Environmental Toxicants (ELEMENT) cohort study.

Candidate gene analysis: ARHGAP42, ATP2B1, FGF5, HSD11B2, KCNK3, PRDM8

Biological considerations: n = 522 children aged 10 - 18 years, stratified by sex.

18

18

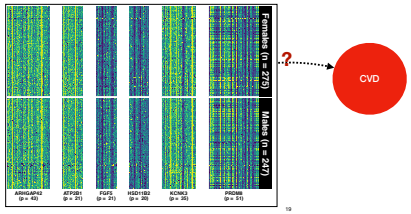
So next, I want to give you an overall view of what the data looks like.

Does DNAm drive BP or vice-versa?

Data: Early Life Exposures in Mexico to Environmental Toxicants (ELEMENT) cohort study.

Candidate gene analysis: ARHGAP42, ATP2B1, FGF5, HSD11B2, KCNK3, PRDM8

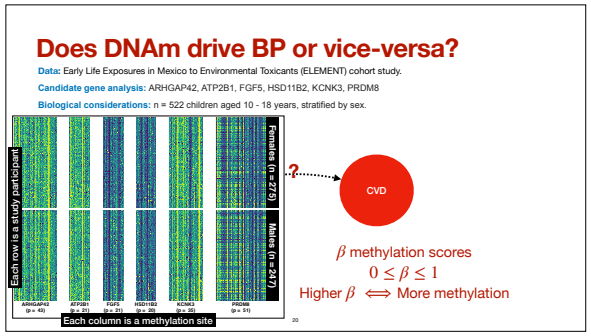
Biological considerations: n = 522 children aged 10 - 18 years, stratified by sex.



19

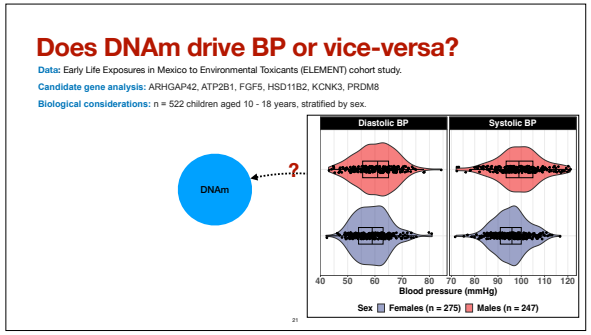
19

First, here is all the methylation data in one plot. IN each of the six facets which correspond to one of the six candidate genes I consider a plot of the beta values which measure methylation status of a given location in a gene. Each row corresponds to a person in the study, and each column corresponds to a location of a methylation site in a given gene.



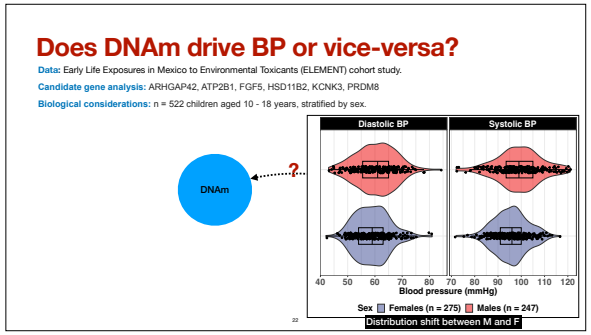
20

First, here is all the methylation data in one plot. IN each of the six facets which correspond to one of the six candidate genes I consider a plot of the beta values which measure methylation status of a given location in a gene. Each row corresponds to a person in the study, and each column corresponds to a location of a methylation site in a given gene.



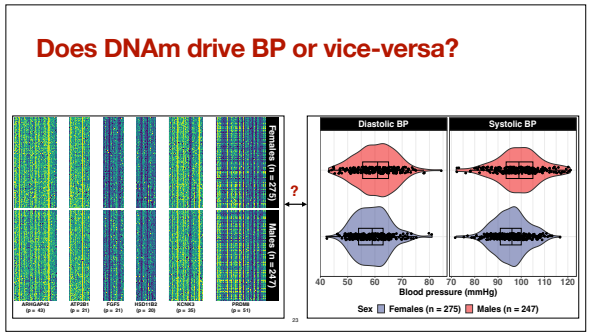
21

Next, here are the sex-stratified violin plots. Already note some evidence that stratifying by sex is a good idea, the boxplots show shifts in distribution from males to females for both systolic and diastolic blood pressure.



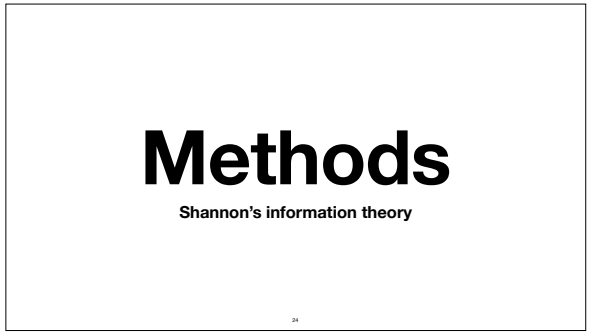
22

Next, here are the sex-stratified violin plots. Already note some evidence that stratifying by sex is a good idea, the boxplots show shifts in distribution from males to females for both systolic and diastolic blood pressure.



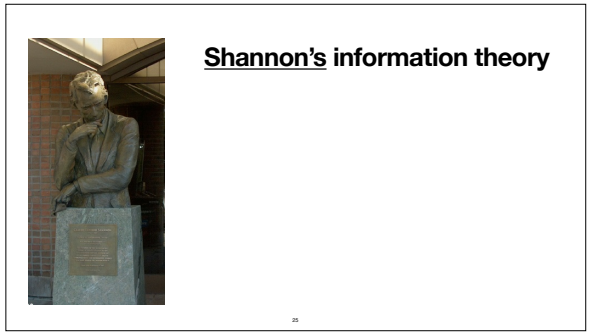
23

Next, here are the sex-stratified violin plots. Already note some evidence that stratifying by sex is a good idea, the boxplots show shifts in distribution from males to females for both systolic and diastolic blood pressure.



24

Now I'll discuss the statistical methods I developed using information theory.



25

The methods focus of my talk will be Shannon's information theory. Here you see a photo of Shannon's statue from the University of Michigan - he went there for his undergraduate degree. Shannon is widely regarded as one of the earliest proponents of information theory.

Shannon's information theory

Association and direction through distributions

$(X, Y) \sim f_{XY} \qquad X \sim f_X \qquad Y \sim f_Y$

26

26

And it turns out that his framework also provides an elegant tool to study association and directionality through distributions.
Let us consider some notation: A bivariate pair XY with joint density and marginal densities.

Shannon's information theory

Association and direction through distributions

$(X, Y) \sim f_{XY} \qquad X \sim f_X \qquad Y \sim f_Y$

Mutual information: measure of association

$$MI(X, Y) = E_{XY}\left[\log\left(\frac{f_{XY}}{f_X f_Y}\right)\right]$$

27

27

One key measure in information theory is mutual information, which is a measure of association and will form the basis of the first half of this talk.

Shannon's information theory

Association and direction through distributions

$(X, Y) \sim f_{XY} \qquad X \sim f_X \qquad Y \sim f_Y$

Mutual information: measure of association

$$MI(X, Y) = E_{XY}\left[\log\left(\frac{f_{XY}}{f_X f_Y}\right)\right]$$

Entropy: measure of randomness

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

28

28

Another key measure is entropy, which is a measure of uncertainty or randomness.
We can have joint entropy, which measures total randomness in bivariate setup.

Shannon's information theory

Association and direction through distributions

$(X, Y) \sim f_{XY} \qquad X \sim f_X \qquad Y \sim f_Y$

Mutual information: measure of association

$$MI(X, Y) = E_{XY}\left[\log\left(\frac{f_{XY}}{f_X f_Y}\right)\right]$$

Entropy: measure of randomness

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

$H(X) = E[-\log(f_X)]$

29

29

Similarly we can define marginal entropies.

Shannon's information theory

Association and direction through distributions

$(X, Y) \sim f_{XY} \qquad X \sim f_X \qquad Y \sim f_Y$

Mutual information: measure of association

$$MI(X, Y) = E_{XY}\left[\log\left(\frac{f_{XY}}{f_X f_Y}\right)\right]$$

Entropy: measure of randomness

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

$H(X) = E[-\log(f_X)]$

$H(X \mid Y) + H(Y) = H(X, Y)$

30

30

And finally, the conditional entropy of X given Y measures the ability of Y to predict X.

The total entropy in a bivariate system can be broken down into a conditional plus marginal entropy.

Shannon's information theory

Association and direction through distributions

$(X, Y) \sim f_{XY} \qquad X \sim f_X \qquad Y \sim f_Y$

Mutual information: measure of association

$$MI(X, Y) = E_{XY}\left[\log\left(\frac{f_{XY}}{f_X f_Y}\right)\right]$$

Entropy: measure of randomness

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

$H(X) = E[-\log(f_X)]$

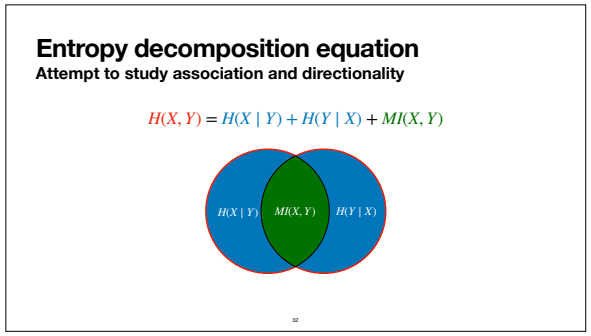
$H(X \mid Y) + H(Y) = H(X, Y)$

Entropy decomposition equation

31

31

And these quantities are all interlinked through the entropy decomposition equation, which may be used to study association and directionality.

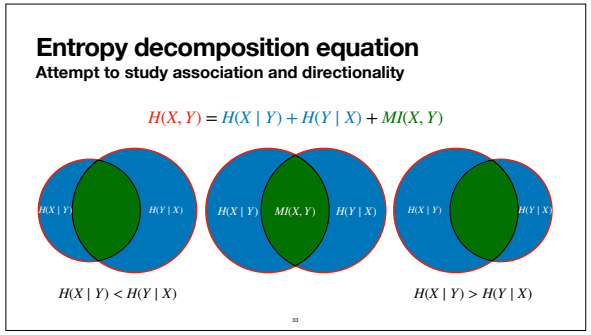


32

I try to use Venn diagrams to visualise this.

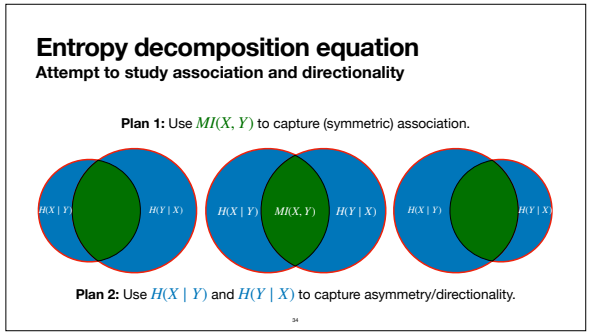
We see that the total joint entropy, measuring total information in the bivariate system can be decomposed into three parts.

Note that by definition, both joint entropy and mutual information are symmetric measures. In set theoretic terms, both union and intersection operators are symmetric as well.



33

However, the two conditional entropy terms may be unequal, thereby reveal a sense of asymmetry or directionality. And that is the central understanding of asymmetry or directionality in my talk. Of course, I will support this claim through a more rigorous framework, but this is my intuition.



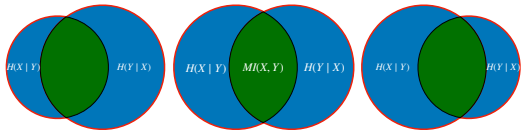
34

So, I have two ideas to share today.

Entropy decomposition equation

Attempt to study association and directionality

Plan 1: Use $MI(X, Y)$ to capture (symmetric) association.



35

35

For the first one, I study mutual information and symmetric association in random variables.

Part I

`fastMI:` **fast and consistent nonparametric estimator of MI**

36

36

MI is a powerful measure of association

MI is self-equitable

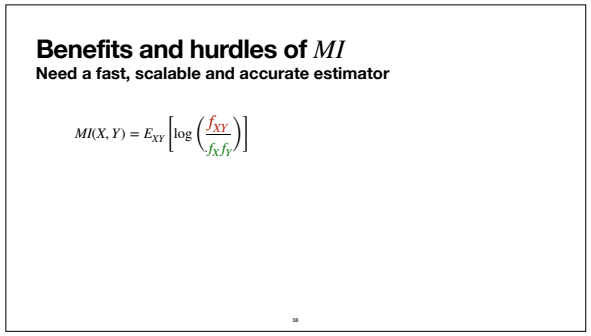
Pearson, Spearman, Kendall:
"These are 'good' data to capture association ⬆"

MI captures association across all patterns ⬆ ⬇

37

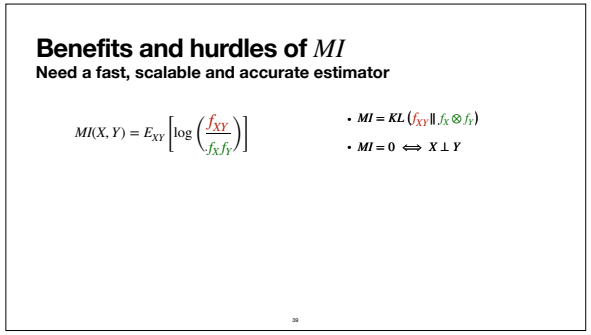
37

The reason I like mutual information so much is that it is a measure of association and not just correlation. Oftentimes we see complex non-linear structures where simple moment based or rank based measures may fail.



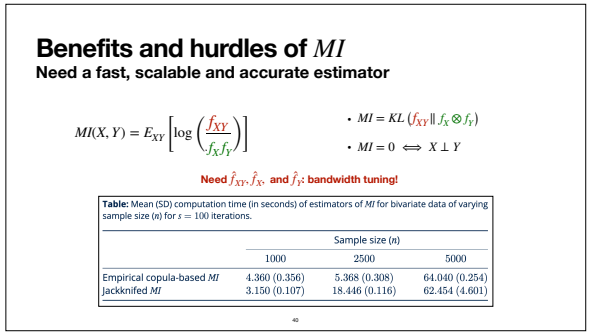
38

However, mutual information measures statistical association and not just correlation.



39

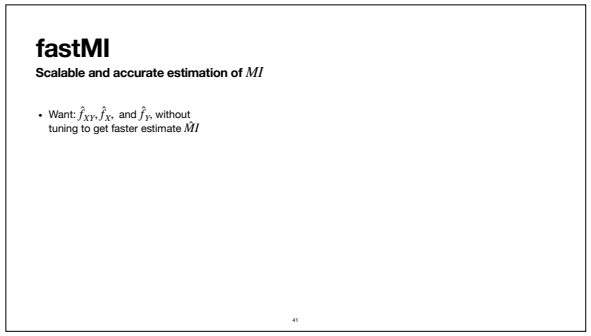
However, mutual information measures statistical association and not just correlation.



40

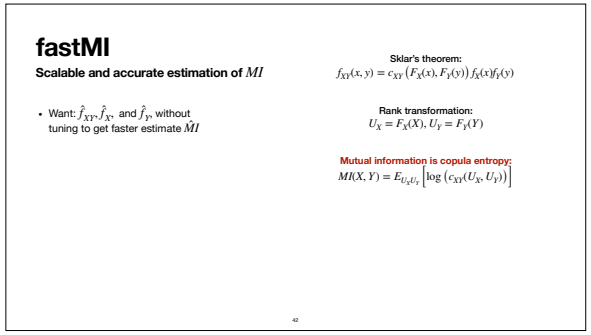
However, one major bottleneck of using mutual information is we need plug-in estimators of underlying density functions and that needs bandwidth tuning.

See this table where I compare run times of two existing mutual information estimators. These are average run times for data sets of different sizes. If this is how long it takes to compute the statistic just once, imagine how intensive it must be to run a simple permutation test. So we definitely need to improve computational efficiency here.



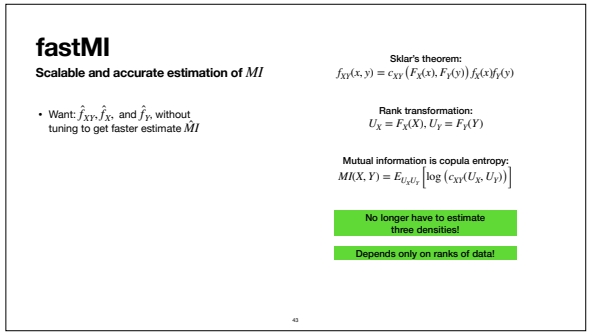
41

Ideally, it would be lovely to not rely on bandwidth tuning at all.



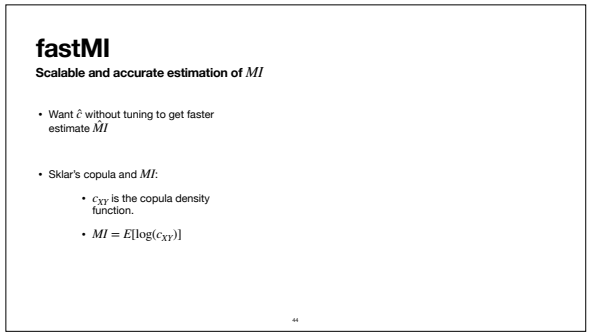
42

But first, to reduce the technical complexity of the problem, we make this very critical observation that mutual information is copula entropy. For a little context, a bivariate density function can expressed as the product of the underlying copula density function and the product of the marginals. This observation is called Sklar's theorem and it invokes rank transformations.



43

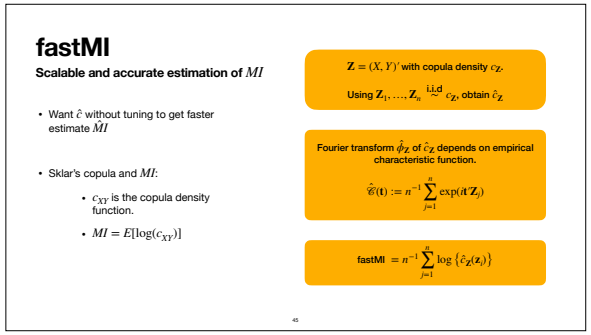
This observation allows us to narrow our focus only to estimating the underlying copula density function in a tuning free manner.



44

And it turns out, using Fourier transformations, we are able to make a dent in the problem.

Further, Fourier transformation-based density estimation can be faster than bandwidth tuning because it operates in the frequency domain, allowing for efficient computation of density estimates. In contrast, bandwidth tuning in traditional methods like kernel density estimation involves optimizing the width of the smoothing kernel, which can be computationally intensive. The Fourier transformation approach provides a more direct and computationally efficient way to estimate densities by leveraging the frequency information of the data.



45

So very briefly, instead of obtaining the kernel density function directly, we attempt to estimate the kernel density estimate in Fourier space. It turns out, that the Fourier transform phi-hat of the estimate c-hat can be expressed as a function of the empirical characteristic function.

So, we apply some smoother to the empirical characteristic function to get our phi-hat, which is antitransformed to get the original c-hat, which serves as our plug-in estimator.

fastMI

Scalable and accurate estimation of MI

- Want $\hat{c}$ without tuning to get faster estimate $\hat{MI}$
- Sklar's copula and MI:
  - c_{XY} is the copula density function.
  - $MI = E[\log(c_{XY})]$
- Use Fourier transformation trick to estimate c_{XY} without tuning.

46

46

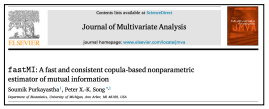
And using these observations we are able to get an estimator that I call fastMI that is much faster than those which are currently in use.

---

fastMI

Scalable and accurate estimation of MI

- Want $\hat{c}$ without tuning to get faster estimate $\hat{MI}$
- Sklar's copula and MI:
  - c_{XY} is the copula density function.
  - $MI = E[\log(c_{XY})]$
- Use Fourier transformation trick to estimate c_{XY} without tuning.



Purkayastha, S., & Song, P. X. K. (2023). fastMI: A fast and consistent copula-based nonparametric estimator of mutual information. Journal of Multivariate Analysis, 105270.

	Sample size (n)		
	1000	2500	5000
Empirical copula-based MI	4.360 (0.356)	5.368 (0.308)	64.040 (0.254)
Jackknifed MI	3.150 (0.107)	18.446 (0.116)	62.454 (4.601)
`fastMI`	**1.199 (0.135)**	**2.964 (0.181)**	**5.952 (0.125)**

47

47

And using these observations we are able to get an estimator that I call fastMI that is much faster than those which are currently in use.

Entropy decomposition equation

Attempt to study association and directionality

Plan 1: Use $MI(X, Y)$ to capture (symmetric) association.

fastMI is an accurate and fast estimator of mutual information.

fastMI enjoys tuning-free estimation of MI

Still no sense of directionality

48

48

For the first one, I study mutual information and symmetric association in random variables.

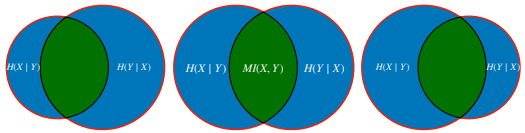
Part II

Entropy reflects underlying asymmetry

49

Entropy decomposition equation

Attempt to study association and directionality



Plan 2: Use $H(X \mid Y)$ and $H(Y \mid X)$ to capture asymmetry/directionality.

50

So instead, we go back to the entropy decomposition equation and attempt the use the entropy terms instead.

Generative exposure mapping (GEM)

Entropy captures directionality!

- Exposure $X \sim f_X$ from an experiment.
- Outcome $Y \sim f_Y$, generated by X being mapped through **bijective** $g(\cdot)$

$$Y = g(X)$$

51

In a framework that I term as the generative exposure mapping, which is an extension of the exposure mapping models used to study causal inference problems in networks.

In GEMs, we have an exposure X with a given density, which is mapped to an outcome Y through a bijective map given by G.

Generative exposure mapping (GEM)
Entropy captures directionality!

- Exposure $X \sim f_X$ from an experiment.
- Outcome $Y \sim f_Y$, generated by X being mapped through **bijective** $g(\cdot)$

$$Y = g(X)$$

- Allow for covariate adjustment:

$$Y = g(X, \mathbf{Z})$$

- Allow for noise contamination:

$$Y = g(X) + \epsilon, \text{ with } X \perp \epsilon.$$

52

52

Of course, this is very restrictive so we must allow for covariate adjustment and also allow for contamination.

So there is a clear sense of direction induced by the generative map linking exposure to outcome.

Generative exposure mapping (GEM)
Entropy captures directionality!

- Exposure $X \sim f_X$ from an experiment.
- Outcome $Y \sim f_Y$, generated by X being mapped through **bijective** $g(\cdot)$

$$Y = g(X)$$

Subject to identifiability constraints:
"GEMs reveals distributional discrepancy between exposure-outcome that are captured using the entropy analytic"

53

53

What is very exciting is that if we impose some identifiability conditions on the GEM, this directionality can be captured using entropy.

Generative exposure mapping (GEM)
Entropy captures directionality!

- Exposure $X \sim f_X$ from an experiment.
- Outcome $Y \sim f_Y$, generated by X being mapped through **bijective** $g(\cdot)$

$$Y = g(X)$$

Subject to identifiability constraints:
"GEMs reveals distributional discrepancy between exposure-outcome that are captured using the entropy analytic"

What identifiability conditions?

54

54

So of course, the question is - what identifiability conditions?

55

Generative exposure mapping (GEM)

Entropy captures directionality!

- Exposure $X \sim f_X$ from an experiment.
- Outcome $Y \sim f_Y$, generated by X being mapped through **bijective** $g(\cdot)$

$$Y = g(X)$$

Subject to identifiability constraints:
"GEMs reveals distributional discrepancy between exposure-outcome that are captured using the entropy analytic"

Impose orthogonality condition on $g(\cdot)$ and f_X

55

Clearly, the distribution of outcome Y is affected by two things, right? The structure of the generative map and the density of exposure X. If I could separate out, or orthogonalise the effects between those two then perhaps my life will be a little easier.

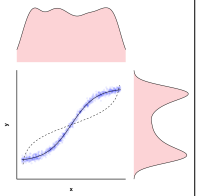
56

Identifiability condition for GEMs

Orthogonality of f_X and g induces discrepancy in f_Y and g^{-1}

$$\int \log\left(|\nabla g(x)|\right) f_X(x)dx = \int \log\left(|\nabla g(x)|\right) dx$$

$$\int \log\left(|\nabla g(x)|\right) f_X(x)dx - \int \log\left(|\nabla g(x)|\right) dx \int f_X(x)dx = 0$$



56

And that is what we try to do here through this identifiability condition.

57

Identifiability condition for GEMs

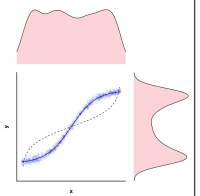
Orthogonality of f_X and g induces discrepancy in f_Y and g^{-1}

$$\int \log\left(|\nabla g(x)|\right) f_X(x)dx = \int \log\left(|\nabla g(x)|\right) dx$$

$X \sim U(0,1)$ satisfies this condition

$U(0,1)$ is maximally random

Flat f_X and bijective g yields spikes in f_Y



57

We note that a uniform density would satisfy this condition. This observation is important from an information theoretic viewpoint since the uniform distribution is maximally random. What I mean by that is for all continuous distributions on a compact space, the uniform distribution has the maximum entropy.

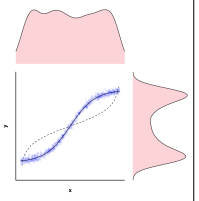
So to illustrate the orthogonality condition, I simulate data from uniform and plug it into a generator g to get the outcome Y. Note that wherever the slopes of g seems to have a role in where the density of Y spikes.

Identifiability condition for GEMs

Orthogonality of f_X and g induces discrepancy in f_Y and g^{-1}

$$\int \log\left(|\nabla g(x)|\right) f_X(x)dx = \int \log\left(|\nabla g(x)|\right) dx$$

$$\int \log\left(|\nabla g^{-1}(y)|\right) f_Y(y)dy \geq \int \log\left(|\nabla g^{-1}(y)|\right) dy$$



58

58

And in fact, if the orthogonality condition holds in one direction, the analogous condition in the opposite direction would not hold.

Identifiability condition for GEMs

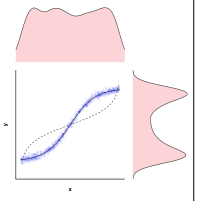
Orthogonality of f_X and g induces discrepancy in f_Y and g^{-1}

$$\int \log\left(|\nabla g(x)|\right) f_X(x)dx = \int \log\left(|\nabla g(x)|\right) dx$$

$$\int \log\left(|\nabla g^{-1}(y)|\right) f_Y(y)dy \geq \int \log\left(|\nabla g^{-1}(y)|\right) dy$$

Easier to retrieve Y from X through g? Or get X from Y through g^{-1}?

Entropy captures this discrepancy.



59

59

So the key question is, is it easier to retrieve Y given X through g or the opposite way? If we impose the identifiability constraint, a disparity is induced between the densities of X and Y, and that discrepancy may be captured using marginal entropies.

Asymmetry in GEMs using entropy

Approve or disprove $X \to Y$

- Hypothesis: $X \to Y$.
- GEM-induced direction $Y = g(X)$.

60

60

So, in a GEM, we have a certain direction induced naturally.

Asymmetry in GEMs using entropy

Approve or disprove $X \to Y$

- Hypothesis: $X \to Y$.
- GEM-induced direction $Y = g(X)$.
- Population: if identifiability conditions hold

$$C_{X \to Y} := H(X) - H(Y) > 0$$

61

61

At a population level, if the identifiability conditions were to hold, the contract of the two marginal entropies would be positive.

Asymmetry in GEMs using entropy

Approve or disprove $X \to Y$

- Hypothesis: $X \to Y$.
- GEM-induced direction $Y = g(X)$.
- Population: if identifiability conditions hold

$$C_{X \to Y} := H(X) - H(Y) > 0$$

- Sample: if $\hat{C}_{X \to Y} > 0$, confirm hypothesis of direction induced by GEM.

62

62

And so, given a sample from a population where we are willing to impose a generative exposure map and apply the identifiability condition, the estimated contrast would be significantly bigger than zero.

Asymmetry in GEMs using entropy

Approve or disprove $X \to Y$

- Hypothesis: $X \to Y$.
- GEM-induced direction $Y = g(X)$.
- Experiment:
  - $X \sim U(0,1)$.
  - Generate $Y = g(X)$.
  - **Simulated behavior of $C_{X \to Y}$?**

63

63

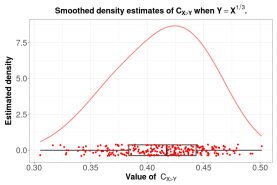
I want to convince you that this is indeed the case. To do so I'll rely on some simulation studies.

In each case, I'll generate data from the uniform distribution and generate data from a specific bijective function and estimate the C statistic repeatedly to get to a sense of the distribution of the C statistic.

Asymmetry in GEMs using entropy

Approve or disprove $X \to Y$

- Hypothesis: $X \to Y$.
- GEM-induced direction $Y = g(X)$.
- Experiment:
  - $X \sim U(0,1)$.
  - Generate $Y = g(X)$.
  - **Simulated behavior of $C_{X>Y}$?**



64

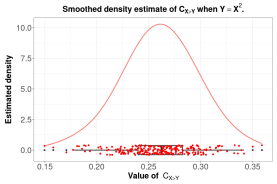
64

Let's start with the case where we take the cube root of X. All simulated values are positive.

Asymmetry in GEMs using entropy

Approve or disprove $X \to Y$

- Hypothesis: $X \to Y$.
- GEM-induced direction $Y = g(X)$.
- Experiment:
  - $X \sim U(0,1)$.
  - Generate $Y = g(X)$.
  - **Simulated behavior of $C_{X>Y}$?**



65

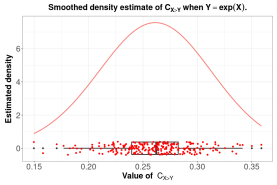
65

Next, for the quadratic function: same comment.

Asymmetry in GEMs using entropy

Approve or disprove $X \to Y$

- Hypothesis: $X \to Y$.
- GEM-induced direction $Y = g(X)$.
- Experiment:
  - $X \sim U(0,1)$.
  - Generate $Y = g(X)$.
  - **Simulated behavior of $C_{X>Y}$?**



66

66

Finally for the exponential function as well - same behaviour.

I want to point out that from the bell shaped curve of the estimated statistic, it is clear that some kind of uncertainty quantification for the statistic is very important.

Asymmetry coefficient $C_{X\to Y}$
Advantages and challenges

Strong asymmetry:

- GEM + identifiability assumptions $\implies C_{X\to Y} > 0$.
- Prove or disprove $X \to Y$ using $\hat{C}_{X\to Y}$

67

67

So now, we have a framework and a statistic to study directionality. Basically we assume a GEM and some identifiability condition and prove or disprove a data generating process using the statistic.

Asymmetry coefficient $C_{X\to Y}$
Advantages and challenges

Strong asymmetry:

- GEM + identifiability assumptions $\implies C_{X\to Y} > 0$.
- Prove or disprove $X \to Y$ using $\hat{C}_{X\to Y}$

Weak asymmetry: what if GEM is absent? What if identifiability conditions don't hold?

- $C_{X\to Y} = H(X) - H(Y) = H(X|Y) - H(Y|X)$
- Better predictor selection using $\hat{C}_{X\to Y}$

68

68

However, if we are unwilling to impose a GEM structure, we can still use the C statistic to compare predictive performance, since the contrast of marginal entropies can be re-written as the contrast of conditional entropies, which measure which is the better predictor among X and Y.

Data splitting and cross fitting
Estimation and inference for $C_{X\to Y}$

Need to estimate $\hat{f}_X$ and $\hat{f}_Y$: **infinite-dimensional nuisance parameters.**

$$\hat{H}(X) = n^{-1}\sum_{j=1}^{n} -\log\left(\hat{f}_X(X_i)\right).$$
$$\uparrow \{X_1, \ldots, X_n\}$$

$\hat{f}_X(X_i)$ and $\hat{f}_X(X_j)$ are not independent!

69

69

So the key challenge is to estimate the underlying density functions. Since the entropy estimator will use the data twice, once to estimate the density and then to estimate the entropy, this kind of double dipping will induce bias in the estimator.

Data splitting and cross fitting

Estimation and inference for $C_{X \to Y}$

Data splitting and **cross-fitting** reduces bias and permits inference.

Split data into two halves.

Use one half to estimate densities.

Use the other half to estimate entropies.

70

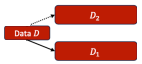
70

To fix this, we use data splitting and also use a cross-fitting approach that permits us to have valid inference and quantify uncertainty.

Data splitting and cross fitting

Estimation and inference for $C_{X \to Y}$

$$D = \{(X_i, Y_i)\}_{i=1}^{n} \cup \{(X_i, Y_i)\}_{i=n+1}^{2n} = D_1 \cup D_2$$



71

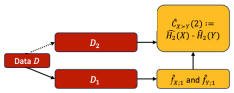
71

I consider the data to have two disjoint halves.

Data splitting and cross fitting

Estimation and inference for $C_{X \to Y}$

$$D = \{(X_i, Y_i)\}_{i=1}^{n} \cup \{(X_i, Y_i)\}_{i=n+1}^{2n} = D_1 \cup D_2$$



72

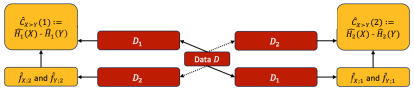
72

Using data from the first half, I estimate the densities. Using the estimated densities and independent data from the second split, I estimate the C statistic.

Data splitting and cross fitting

Estimation and inference for $C_{X\to Y}$

$$D = \{(X_i, Y_i)\}_{i=1}^{n} \cup \{(X_i, Y_i)\}_{i=n+1}^{2n} = D_1 \cup D_2$$



73

73

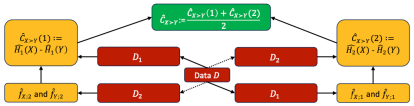
Interchanging the two data splits another C statistic is found

---

Data splitting and cross fitting

Estimation and inference for $C_{X\to Y}$

$$D = \{(X_i, Y_i)\}_{i=1}^{n} \cup \{(X_i, Y_i)\}_{i=n+1}^{2n} = D_1 \cup D_2$$



74

74

Which are pooled together.

Data splitting and cross fitting

Estimation and inference for $C_{X\to Y}$

$$D = \{(X_i, Y_i)\}_{i=1}^{n} \cup \{(X_i, Y_i)\}_{i=n+1}^{2n} = D_1 \cup D_2$$

$\hat{C}_{X\to Y}$ has a **limiting distribution** subject to **regularity conditions**

$$\sqrt{n}\left(\hat{C}_{X\to Y} - C_{X\to Y}\right) \to N(0, \sigma_C^2), \text{ as } n \to \infty,$$

$$\sigma_C^2 = V[\log(f_X(X)) + \log(f_Y(Y))]$$

Estimated by Monte-Carlo methods with estimated $\hat{f}_X$ and $\hat{f}_Y$

75

75

Subject to some regularity conditions this estimator has an asymptotic normal distribution.
This will allow us to quantify uncertainty in the estimated C statistic, which forms a major technical contribution of this work.

Coefficient can tolerate contaminated outcomes!

$H(Y + \sqrt{\sigma}\epsilon) < H(X)$ **for what range of** σ?

- Noise-perturbed GEM (NPGEM)

  $Y^* = g(X) + \sqrt{\sigma}\epsilon$, with $\epsilon \sim N(0,1)$ and $X \perp \epsilon$.

76

76

It's good to know that this method is able to tolerate a certain about of noise, as in the case of Noise perturbed GEMs.

Coefficient can tolerate contaminated outcomes!

$H(Y + \sqrt{\sigma}\epsilon) < H(X)$ **for what range of** σ?

- Noise-perturbed GEM (NPGEM)

  $Y^* = g(X) + \sqrt{\sigma}\epsilon$, with $\epsilon \sim N(0,1)$ and $X \perp \epsilon$.

- Coefficient will work as long as $H(Y^*) \leq H(X)$.

$$I(Y) := E\left[\nabla_y \log(f_Y)\right]^2$$

Non-parametric Fisher information

- Establish "critical value" σ_{CRIT}

$$\sigma \leq \sigma_{CRIT} = \frac{\exp(2C_{X>Y}) - 1}{I(Y)}.$$

77

77

I establish a critical bound or a breakdown point of our detection method, which denotes the contamination level at which our method fails to reflect the true direction.

Coefficient can tolerate contaminated outcomes!

$H(Y + \sqrt{\sigma}\epsilon) < H(X)$ **for what range of** σ?

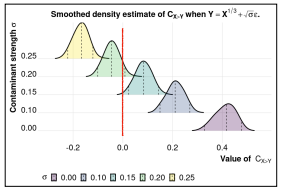
- Noise-perturbed GEM (NPGEM)

  $Y^* = g(X) + \sqrt{\sigma}\epsilon$, with $\epsilon \sim N(0,1)$ and $X \perp \epsilon$.

- Coefficient will work as long as $H(Y^*) \leq H(X)$.

- Establish "critical value" σ_{CRIT}

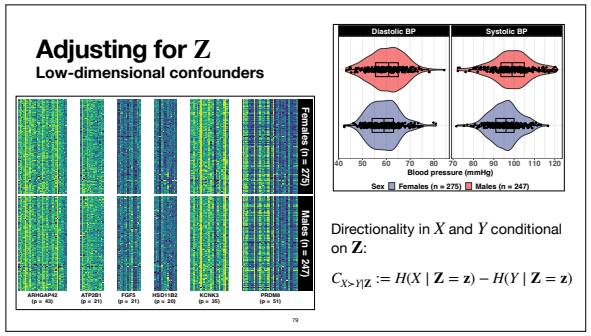
$$\sigma \leq \sigma_{CRIT} = \frac{\exp(2C_{X>Y}) - 1}{I(Y)}.$$



78

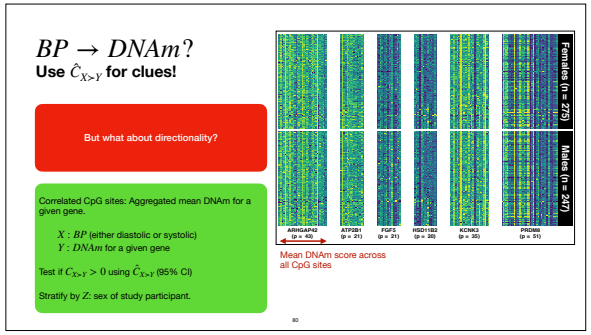
78

Again, I will show you the density plots of simulated C statistics, but this time for various levels of contaminated outcomes. It is not unexpected to see our method is successful only up to a certain level of contamination.



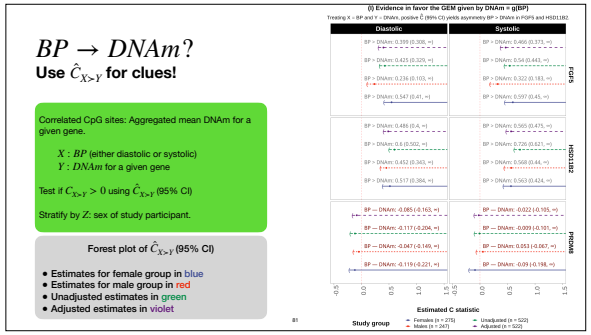
79

To do that we consider strata-specific estimates of our C statistic.



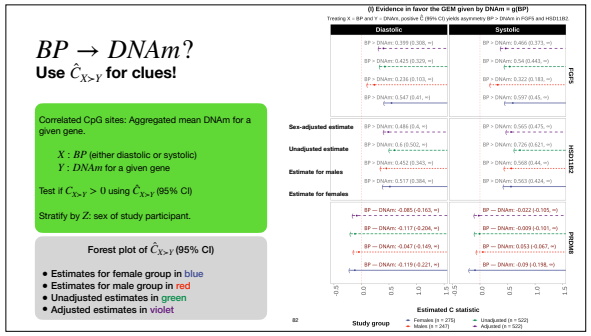
80

But now, in a GEM framework, I want to compute directionality between the two.



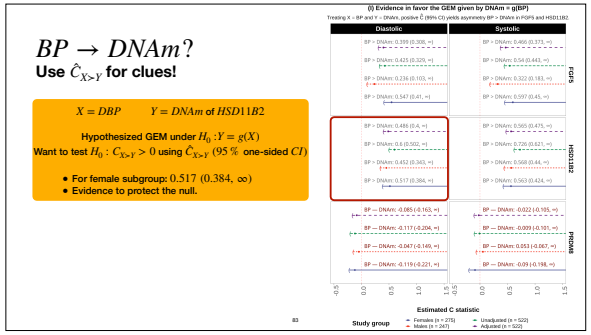
81

I will report sex-stratified and unadjusted c statistics for each of the six candidate genes and both systolic and diastolic blood pressure using a forest plot.

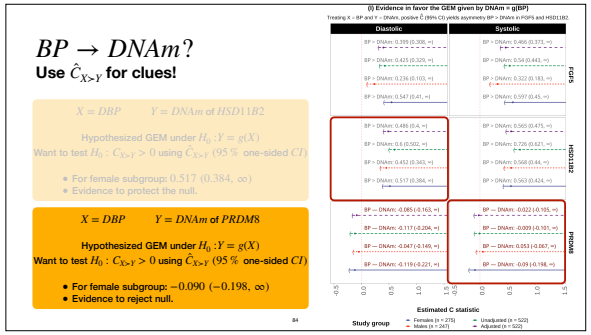


82

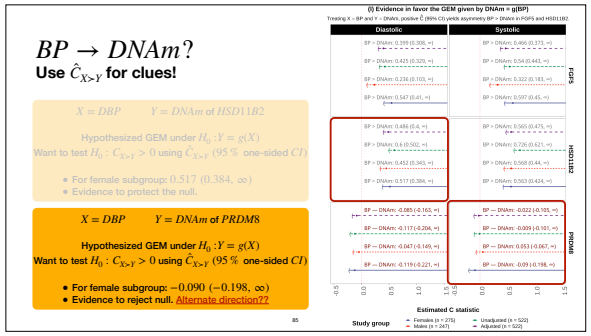
I will report sex-stratified and unadjusted c statistics for each of the six candidate genes and both systolic and diastolic blood pressure using a forest plot.



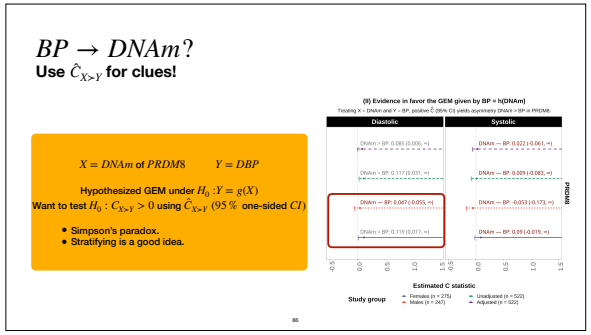
83



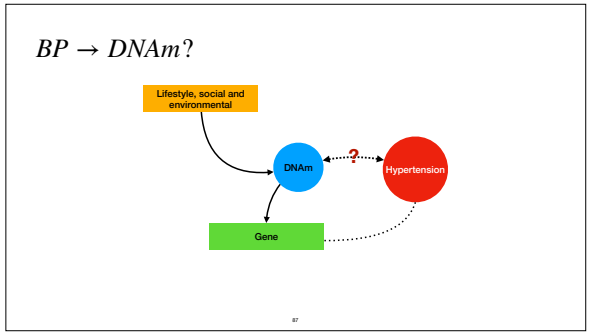
84



85

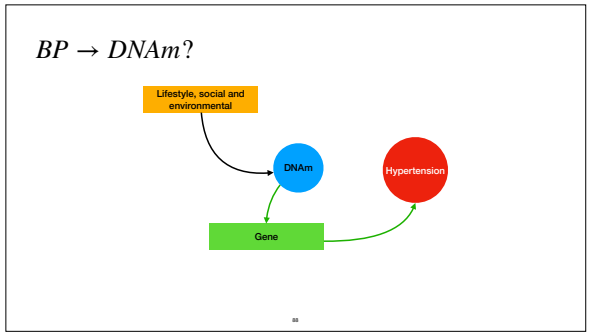


86



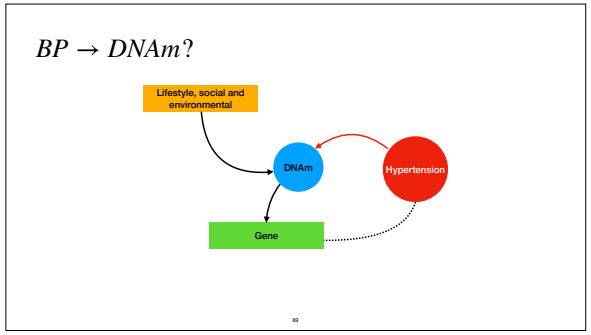
87

So to go back to our diagram,



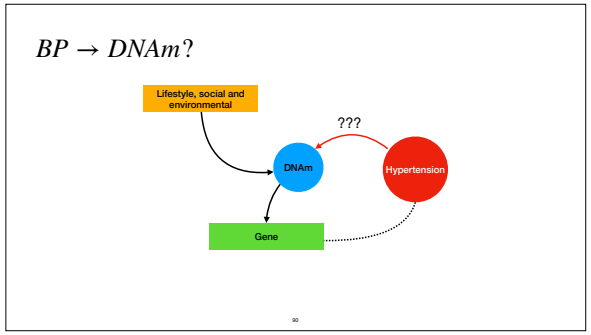
88

The pathway starting from dna methylation to hypertension is one that is probably intuitive. It is in line with our experience of writing snps as predictors of phenotypes, and the understanding that methylation is more like a controlling switch in that equation.



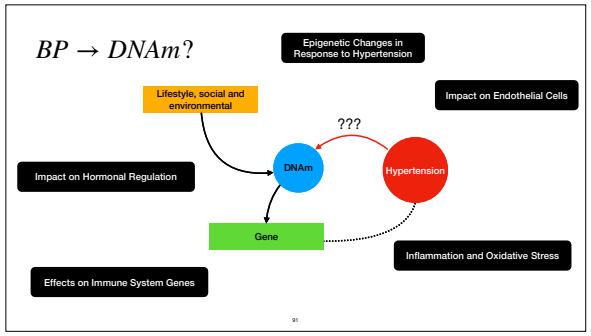
89

But my findings show the other kind of directionality in two genes across the two strata?



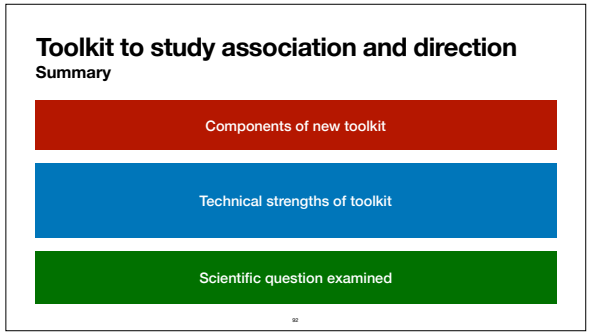
90

So you’re bound to ask what is happenening? Are there biological explanations for this?



91

1. Stress responses and inflammation associated with hypertension can influence the epigenetic landscape.
2. Further, endothelial cells are attached to blood pressure regulation. Changes in blood pressure might affect these cells, influencing DNA methylation patterns.
3. Hypertension is often associated with chronic inflammation and oxidative stress. Both inflammation and oxidative stress can modulate DNA methylation patterns.
4. Hypertension may influence the immune system, and alterations in immune cell DNA methylation patterns have been reported.
5. Hormones such as cortisol, which is released in response to increased blood pressure, can influence DNA methylation. Hypertension, especially if associated with chronic stress, may affect hormonal regulation and subsequently impact epigenetic processes.



92

Toolkit to study association and direction
Summary

- **fastMI to study association.**
- **GEMs and asymmetry coefficient to study directionality.**

Technical strengths of toolkit

Scientific question examined

93

Toolkit to study association and direction
Summary

Components of new toolkit

- **Fast estimation for large n**
- **Reduced estimation error in simulations.**
- **Technical guarantees of data splitting.**

Scientific question examined

94

Toolkit to study association and direction
Summary

Components of new toolkit

Technical strengths of toolkit

- **Directionality in BP variation and epigenetic biomarkers established for the ELEMENT cohort.**

95

Toolkit to study association and direction

Summary

- fastMI to study association.
- GEMs and asymmetry coefficient to study directionality.

- Fast estimation for large n
- Reduced estimation error in simulations.
- Technical guarantees of data splitting.

- Directionality in BP variation and epigenetic biomarkers established for the ELEMENT cohort.

96

96

Future work

97

97

Extend GEM framework to detect third-variable effects

Mediator or collider?

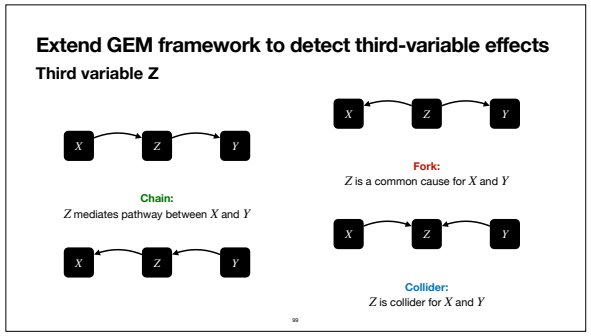
M mediates path from X to Y

C is a collider for X, Y

98

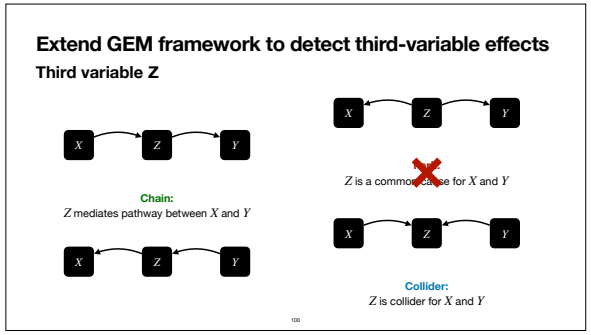
98

I want to explore third-variable structures using GEMs. What I mean by that are mediator and collider models that show various pathways between exposure and outcome.



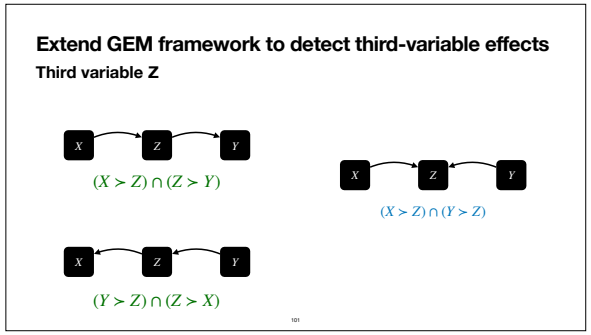
99

I want to explore third-variable structures using GEMs. What I mean by that are mediator and collider models that show various pathways between exposure and outcome.



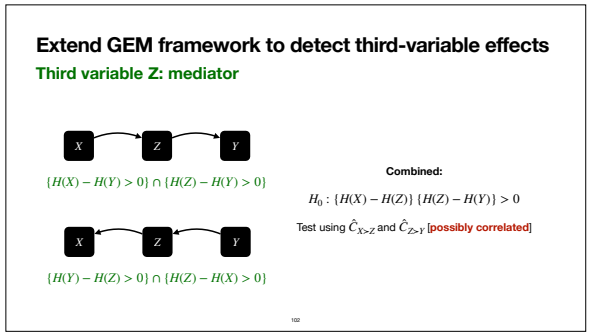
100

I want to explore third-variable structures using GEMs. What I mean by that are mediator and collider models that show various pathways between exposure and outcome.



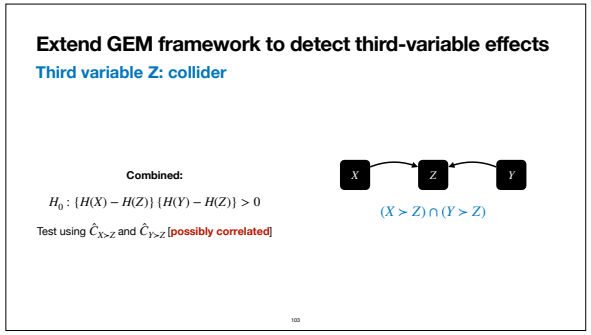
101

I want to explore third-variable structures using GEMs. What I mean by that are mediator and collider models that show various pathways between exposure and outcome.



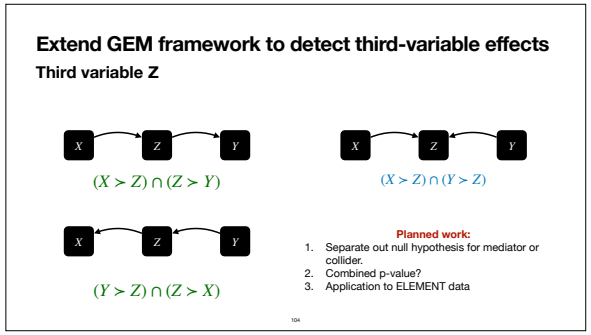
102

I want to explore third-variable structures using GEMs. What I mean by that are mediator and collider models that show various pathways between exposure and outcome.



103

I want to explore third-variable structures using GEMs. What I mean by that are mediator and collider models that show various pathways between exposure and outcome.



104

I want to explore third-variable structures using GEMs. What I mean by that are mediator and collider models that show various pathways between exposure and outcome.

Thanks!

105

Appendix I

Fourier transform-based copula density estimation

Bernacchia, A., & Pigolotti, S. (2011). Self-consistent method for density estimation. Journal of the Royal Statistical Society Series B: Statistical Methodology, 73(3), 407-422.

1. $\mathbf{Z} = (X, Y)'$ with copula density $c_{\mathbf{Z}}$

   Using $\mathbf{Z}_1, \ldots, \mathbf{Z}_n \overset{\text{i.i.d}}{\sim} c_{\mathbf{Z}}$, obtain $\hat{c}_{\mathbf{Z}}$

2. $\hat{c}_{\mathbf{Z}}$ must minimize $MISE = \mathbb{E}\left[\int \left\{\hat{c}_z(\mathbf{z}) - c_z(\mathbf{z})\right\}^2 d\mathbf{z}\right]$

3. $\hat{\phi}_{\mathbf{t}} := \mathscr{F}(\hat{c}_{\mathbf{Z}}) \qquad \phi_{\mathbf{t}} := \mathscr{F}(c_{\mathbf{Z}})$

   $MISE$ in Fourier space $\mathbb{E}\left[\int \left\{\hat{\phi}_{\mathbf{t}}(\mathbf{t}) - \phi_{\mathbf{t}}(\mathbf{t})\right\}^2 d\mathbf{t}\right]$

4. $\hat{\mathscr{C}}(\mathbf{t}) := n^{-1} \sum_{j=1}^{n} \exp(i\mathbf{t}'\mathbf{Z}_j)$

   $\hat{\phi}_{\mathbf{t}}$ depends on empirical characteristic function $\hat{\mathscr{C}}$

5. Antitransform $\hat{\phi}_{\mathbf{t}}$ to get $\hat{c}_{\mathbf{Z}}$

   $\hat{c}_{\mathbf{Z}}(\mathbf{z}) = (2\pi)^{-1} \int \hat{\phi}_{\mathbf{Z}}(\mathbf{t}) \exp(-i\mathbf{t}'\mathbf{z})\, d\mathbf{t}$

6. fastMI $= n^{-1} \sum_{j=1}^{n} \log\left\{\hat{c}_{\mathbf{Z}}(\mathbf{z}_j)\right\}$

106

Appendix II

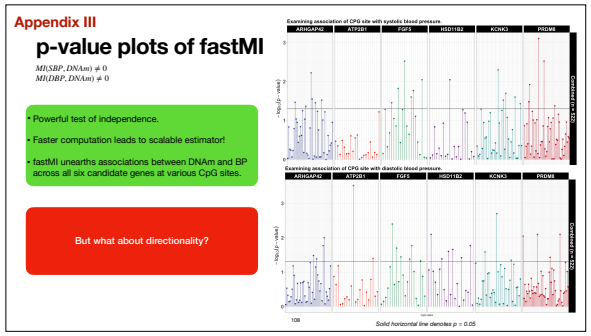
Asymmetry in GEMS reflects underlying directionality

Low-level imprint of Neyman-Rubin causal (NRC) model?

- Implicitly assume $X \to Y$ in NRC.
  - Impact of changing X on Y?
  - Error-based model on Y.
  - SUTVA and random assignment assumptions.

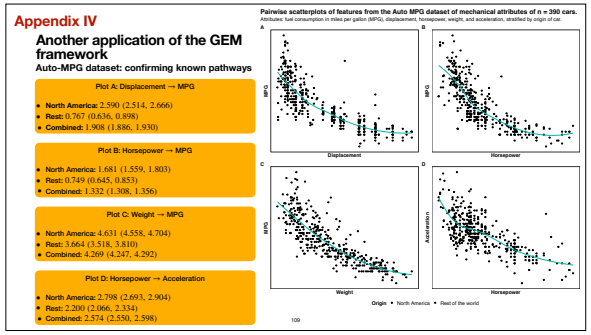
- GEM to approve/disprove $X \to Y$
  - Use Shannon's entropy analytic
  - f_Y depends on f_X and ∇g
  - Identifiability assumption: **orthogonality**.

107

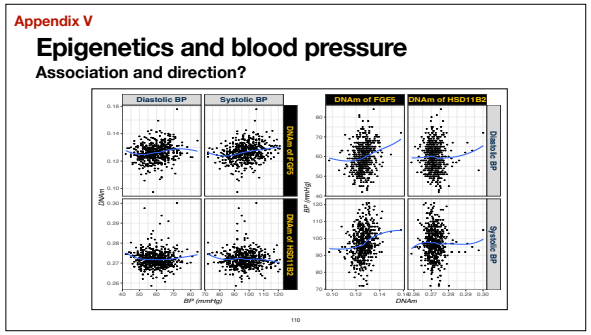


108

Of course there are other considerations such as multiple testing corrections and p-value aggregation to be considered here, but the key point I want to underline is even if there were remarkably strong signals here there would still be no way of telling directionality.



109



110

Appendix VI

Fourier transform-based copula density estimation

(1) Bernacchia, A., & Pigolotti, S. 2011). Self-consistent method for density estimation. Journal of the Royal Statistical Society Series B: Statistical Methodology, 73(3), 407-422.
(2) O'Brien, T. A., Kashinath, K., Cavanaugh, N. R., Collins, W. D., & O'Brien, J. P. (2016). A fast and objective multidimensional kernel density estimation method: fastKDE. Computational Statistics & Data Analysis, 101, 148-160.

$\mathbf{Z} = (X, Y)'$ with copula density $c_{\mathbf{Z}}$. Using $\mathbf{Z}_1, \ldots, \mathbf{Z}_n \overset{\text{i.i.d}}{\sim} c_{\mathbf{Z}}$, obtain $\hat{c}_{\mathbf{Z}}$

Fourier transform $\hat{\phi}_{\mathbf{Z}}$ of $\hat{c}_{\mathbf{Z}}$ depends on empirical characteristic function.

$$\hat{\varphi}(\mathbf{t}) := n^{-1} \sum_{j=1}^{n} \exp(i\mathbf{t}'\mathbf{Z}_j)$$

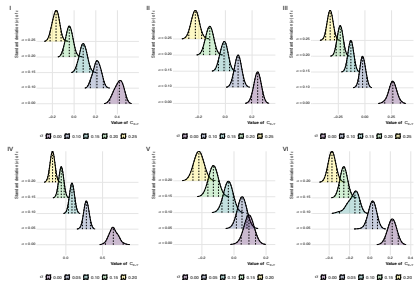
$$\text{fastMI} = n^{-1} \sum_{j=1}^{n} \log\left\{\hat{c}_{\mathbf{Z}}(\mathbf{z}_j)\right\}$$

111

111

Fourier transformation-based density estimation can be faster than bandwidth tuning in certain cases because it operates in the frequency domain, allowing for efficient computation of density estimates. In contrast, bandwidth tuning in traditional methods like kernel density estimation involves optimizing the width of the smoothing kernel, which can be computationally intensive. The Fourier transformation approach provides a more direct and computationally efficient way to estimate densities by leveraging the frequency information of the data.

Appendix VII



112

112

Appendix VIII

Simulation scheme for data-splitting + cross fitting

Identifiability condition doesn't hold but $C_{X>Y} > 0$

1. Generate from f_{XY} with known $C_{X>Y}$.
2. Compute $\hat{C}$ and $\hat{\sigma}_{\hat{C}}$.
3. Repeat steps 1 and 2 $r = 250$ times.

1. SD of simulated $\hat{C}$ gives empirical standard error (ESE).
2. Mean of simulated $\hat{\sigma}_{\hat{C}}$ gives asymptotic standard error (ASE).
3. Using 95% CIs of $\hat{C}$ gives coverage probability (CP).

	Case (I)			Case (II)		
	$n = 250$	$n = 500$	$n = 750$	$n = 250$	$n = 500$	$n = 750$
Absolute bias	0.052	0.049	0.038	0.082	0.062	0.049
ESE	0.081	0.053	0.042	0.087	0.064	0.051
ASE	0.081	0.059	0.049	0.090	0.072	0.060
CP	0.910	0.930	0.960	0.940	0.935	0.955
	Case I: $X \sim LN(5,1); Y \sim N(5,1)$ $C_{X>Y} = H(X) - H(Y) = 5$			Case II: $X \sim E(1); Y \sim Wb(1,3.2)$ $C_{X>Y} = H(X) - H(Y) = 0.213$		

113

113

Appendix IX

fastMI

Scalable and accurate estimation of MI

- Want $\hat{c}$ without tuning to get faster estimate $\hat{MI}$

Improved estimation using probit transformation:
$V_X = \Phi^{-1}(U_X); V_Y = \Phi^{-1}(U_Y)$

$$c_{XY}(U_X, U_Y) := \frac{g\left(\Phi^{-1}(U_X), \Phi^{-1}(U_Y)\right)}{\phi\left(\Phi^{-1}(U_Y)\right)\phi\left(\Phi^{-1}(U_Y)\right)}$$

Estimating c is replaced by estimating g

Fast and tuning free estimation of g

114

114

And it turns out, using a copula-trick and leveraging Fourier transformations, we are able to make a dent in the problem.

First, it turns out that instead of estimating the joint density along with the marginals, it is enough to simply estimate the underlying copula density function, so our technical complexity is greatly reduced.

Further, Fourier transformation-based density estimation can be faster than bandwidth tuning because it operates in the frequency domain, allowing for efficient computation of density estimates. In contrast, bandwidth tuning in traditional methods like kernel density estimation involves optimizing the width of the smoothing kernel, which can be computationally intensive. The Fourier transformation approach provides a more direct and computationally efficient way to estimate densities by leveraging the frequency information of the data.

Appendix X

Simulation I: Estimation accuracy

Compute mean squared error of fastMI when estimating MI

Simulation II: Empirical power

Compute empirical power of fastMI when testing for $H_0 : MI = 0$

Compare with current standard estimations:
empirical copula-based MI and jackknifed MI.

Gaussian, Gumbel and Clayton copula families.
MI can be expressed as function of Kendall's τ.

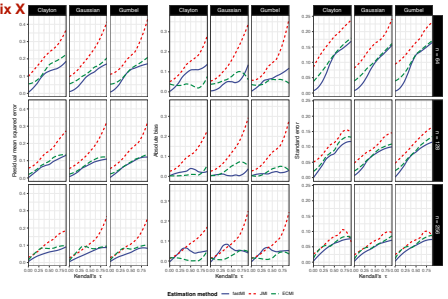
Increasing $\tau \leftrightarrow$ Increasing MI.

115

115

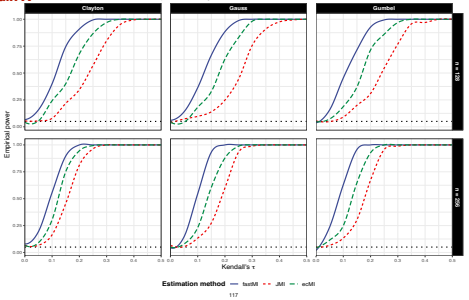
I have two simulation studies to show.

Appendix X



116

Appendix X



117