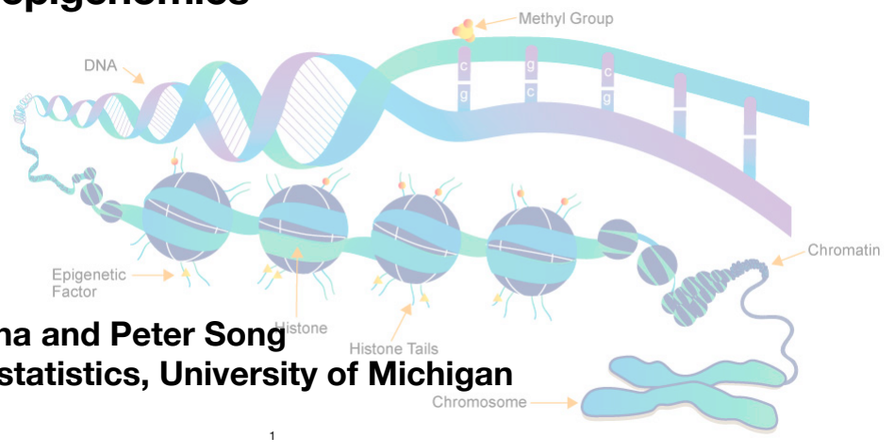
Learning association and directionality

Applications in epigenomics

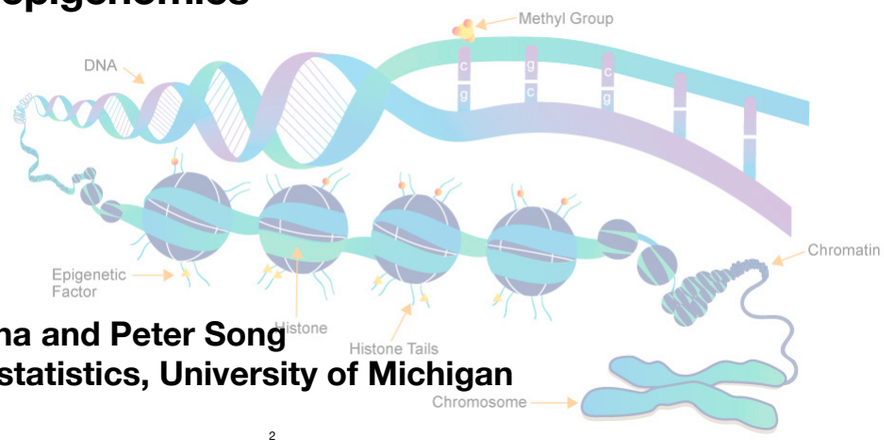


Soumik Purkayastha and Peter Song
Department of Biostatistics, University of Michigan

February 19, 2024

Learning association and directionality

Applications in epigenomics



Soumik Purkayastha and Peter Song
Department of Biostatistics, University of Michigan

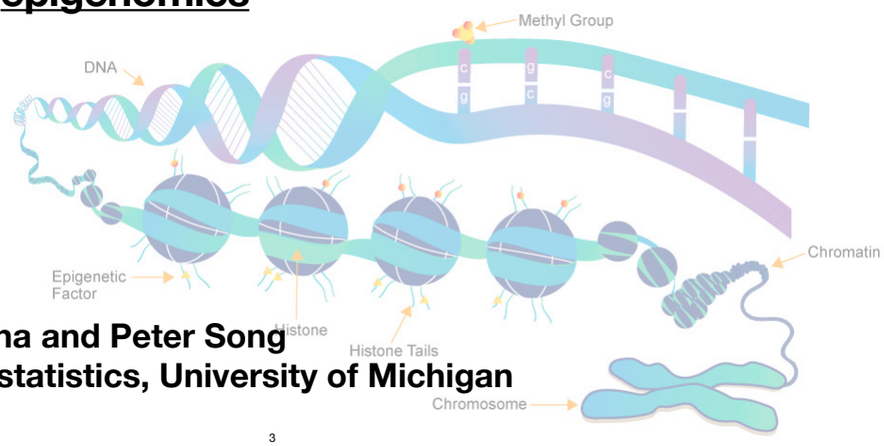
February 19, 2024

Thank you for coming to my talk.

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? Who leads and who follows?

Learning <u>association</u> and <u>directionality</u>

Applications in <u>epigenomics</u>



Soumik Purkayastha and Peter Song
Department of Biostatistics, University of Michigan

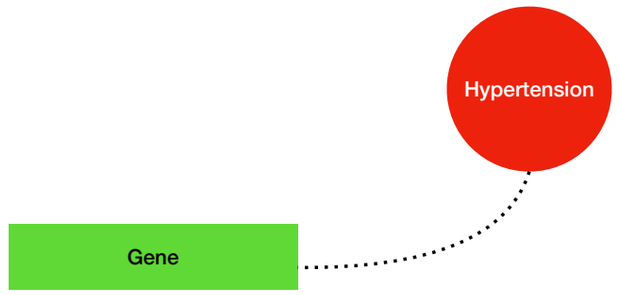
February 19, 2024

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 epigenetic methylation sites, DNA and certain phenotypes of interest, such as a person's cardiovascular profile.**

Epigenomics

Motivating problem

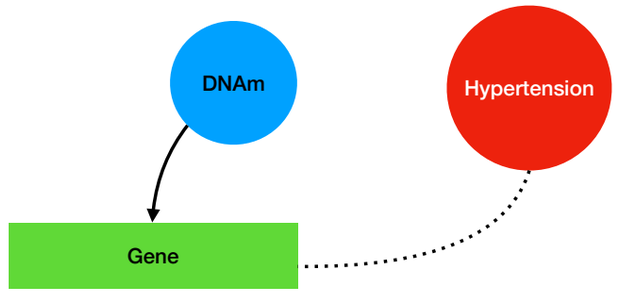
DNAm and cardiovascular diseases



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

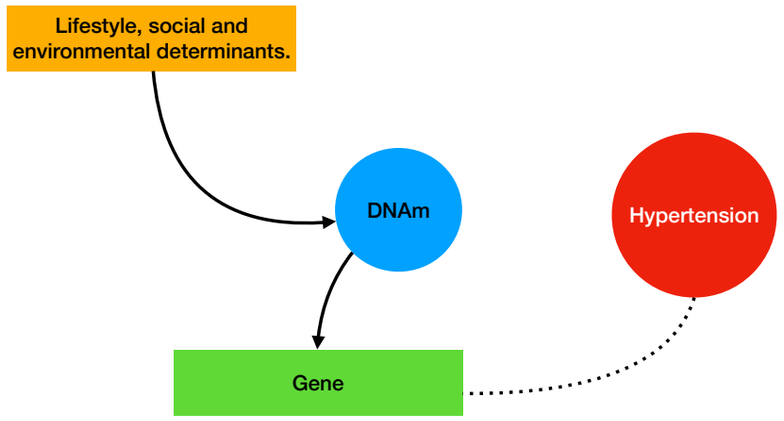
Genome wide studies have unearthed association between many genes and hypertension.

DNAm and cardiovascular diseases



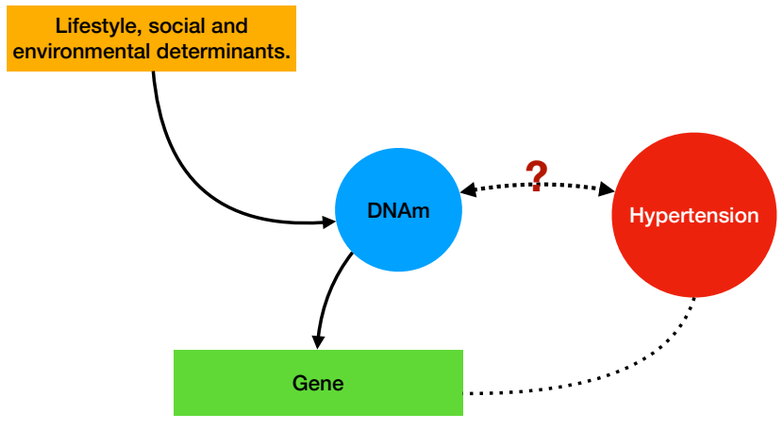
We also know of DNA methylation, which plays a regulatory role in gene expression,

DNAm and cardiovascular diseases



and is in turn influenced by external stimuli.

DNAm and cardiovascular diseases



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.

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.

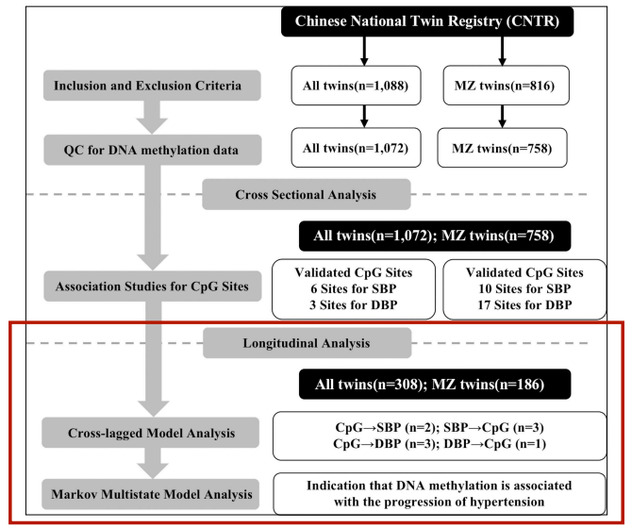
But the field of research into directionality is very new. However, I am certainly not the first to examine this problem.

Epigenetics and blood pressure

Leader or follower?

Hong, Xuanming, et al. "Association between DNA methylation and blood pressure: a 5-year longitudinal twin study." Hypertension 80.1 (2023): 169-181.

- 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.



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?

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.

I will study this directionality in the ELEMENT cohort. ELEMENT is short for….
I must take pause and note that my life was made a whole lot easier because the data was very well organised and I have one of your faculty members and my academic siblings to thank for that. I believe Lu also worked on this dataset during his graduate school days in Michigan.

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

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

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.

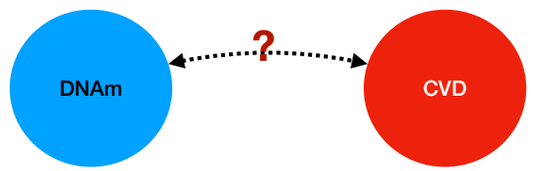
And finally, since the cohort is made up of children in the developmental stages of 10-18 years, it is definitely a good idea to allow for sex-based differences in the relationship between DNA methylation and blood pressure. So I did a sex-stratified analysis.

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.



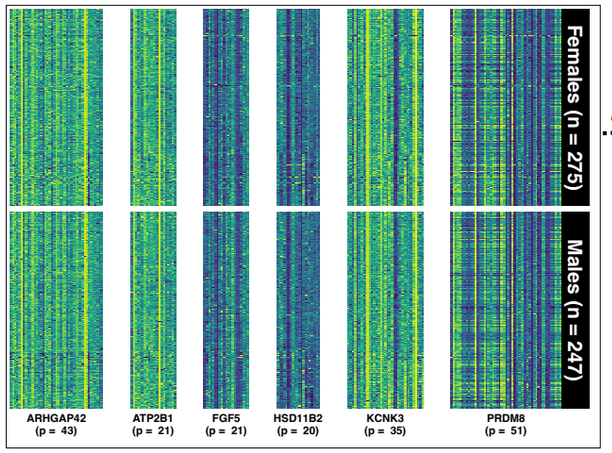
So next, I want to give you an overall view of what the data looks like.

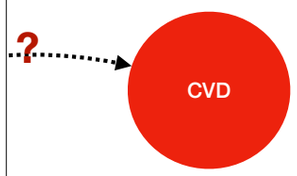
Does DNAm drive BP or vice-versa?

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.





β methylation scores

$0 \leq \beta \leq 1$

Higher β $\iff$ More methylation

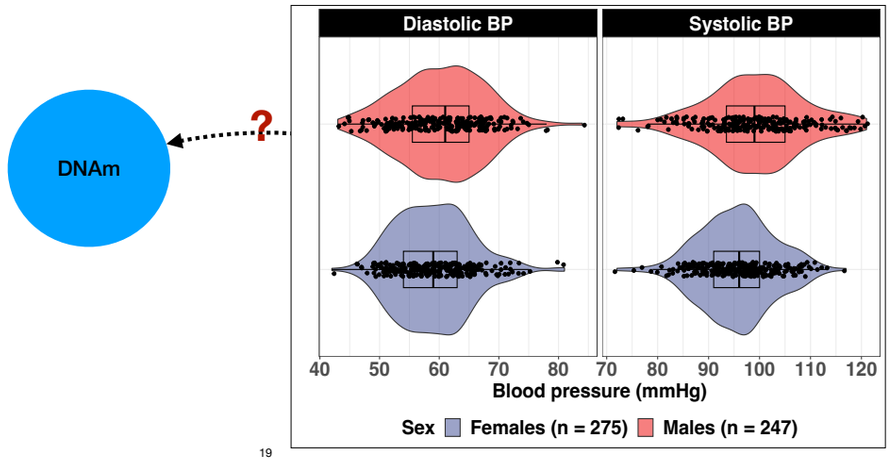
First, here is all the methylation data in one plot. IN each of the six facets which correspond to one of the six candidate genes I consider a plot of the beta values which measure methylation status of a given location in a gene. Each row corresponds to a person in the study, and each column corresponds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

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.



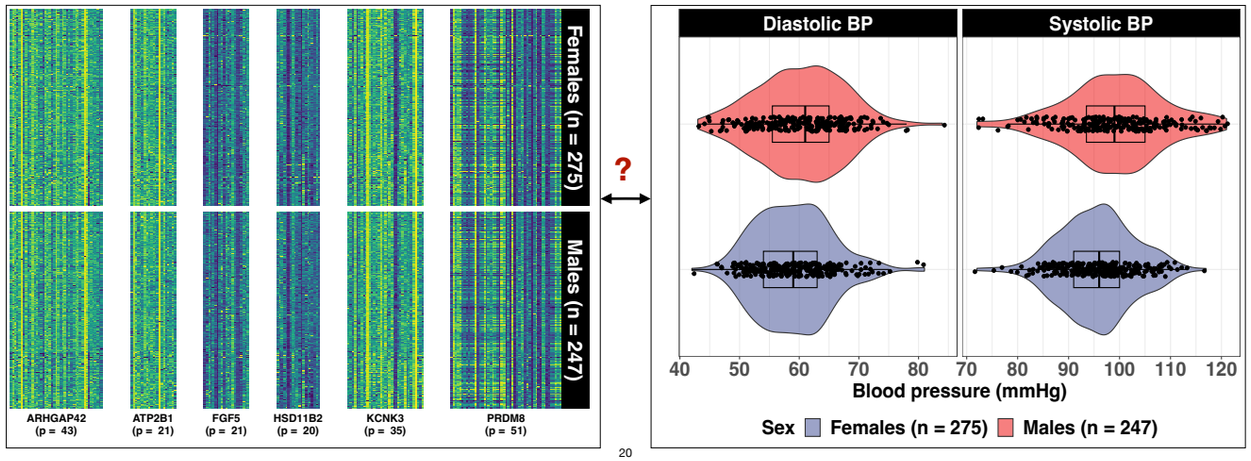
Next, here are the sex-stratified violin plots. Already note some evidence that stratifying by sex is a good idea, the boxplots show shifts in distribution from males to females for both systolic and diastolic blood pressure.

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

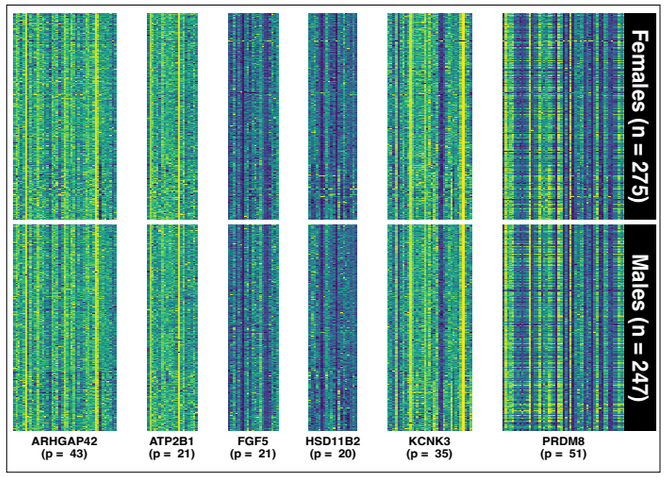
Biological considerations: n = 522 children aged 10 - 18 years, stratified by sex.

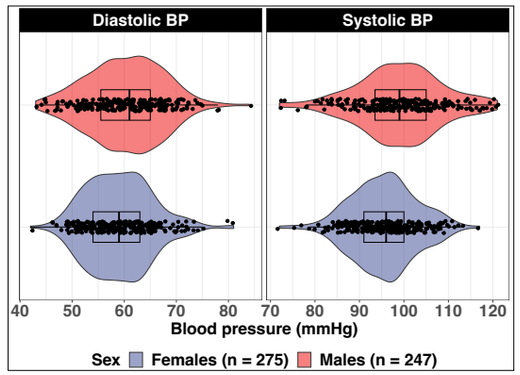


Next, here are the sex-stratified violin plots. Already note some evidence that stratifying by sex is a good idea, the boxplots show shifts in distribution from males to females for both systolic and diastolic blood pressure.

Does DNAm drive BP?

Key challenges





- Many CpG sites.
- Mildly correlated CpG sites.
- Different developmental processes in boys and girls aged 10 - 18.

There are some biological considerations here of course, including a large number of correlated methylation sites and possibly different developmental processes across the male and female subgroups.

Methods

Shannon's information theory

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

Shannon's information theory

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

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

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

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

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 di ributions

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

$MI(X, Y)$ measures strength of association between X and Y

Mutual information: measure of a ociation

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

$MI(X, Y) \geq 0$,
higher values denote 'stronger' association

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

$H(X, Y)$ measures total randomness in (X, Y)

$X \sim f_X \qquad Y \sim f_Y$

Mutual i

$$\left[\log\left(\frac{f_{XY}}{f_X f_Y}\right)\right]$$

Entropy: measure of randomness

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

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

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

$H(X \mid Y)$ measures information needed to predict X if you know Y

Another key measure is entropy, which is a measure of uncertainty or randomness.
We can have joint entropy, marginal entropy and conditional 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

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

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

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

Another key measure is entropy, which is a measure of uncertainty or randomness.
We can have joint entropy, marginal entropy and conditional 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

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

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

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

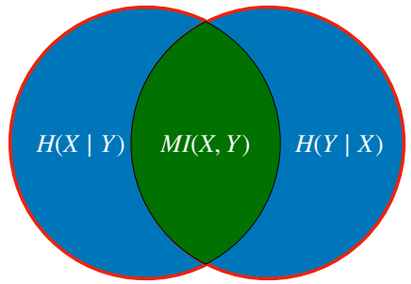
Entropy decomposition equation

And these quantities are all interlinked through the entropy decomposition equation.

Entropy decomposition equation

Attempt to study association and directionality

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



Which may be used to study association and directionality. I try to use Venn diagrams to visualise this.

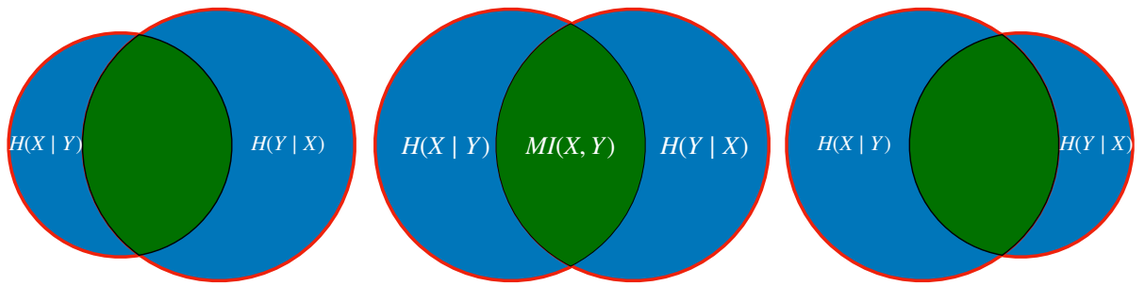
We see that the total joint entropy, measuring total information in the bivariate system can be decomposed into three parts.

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.

Entropy decomposition equation

Attempt to study association and directionality

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



$H(X \mid Y) < H(Y \mid X)$

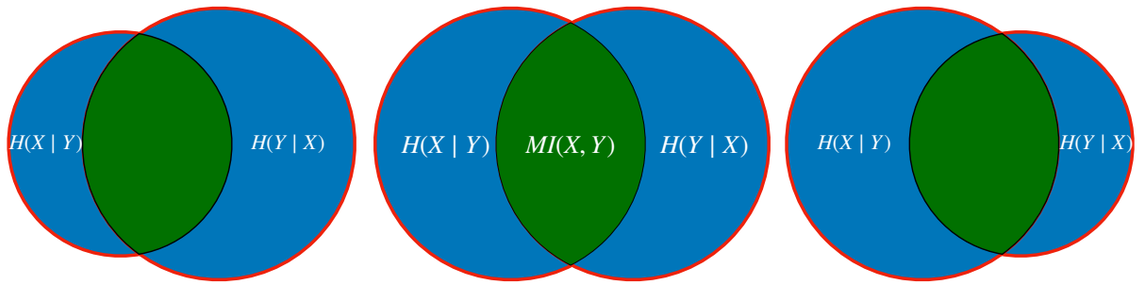
$H(X \mid Y) > H(Y \mid X)$

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.

Entropy decomposition equation

Attempt to study association and directionality

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



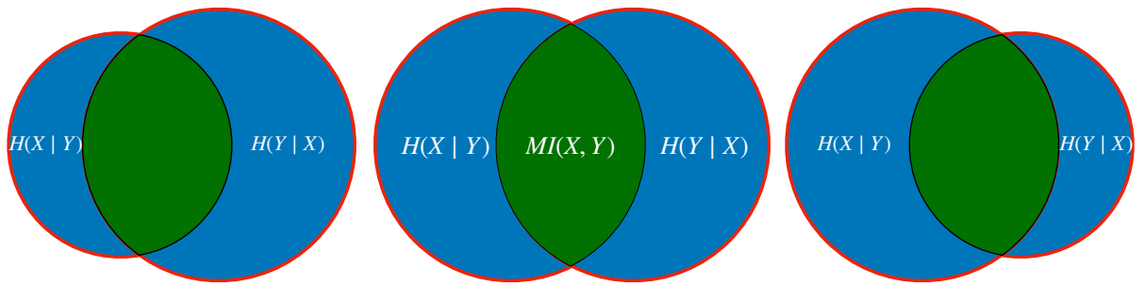
Plan 2: Use $H(X \mid Y)$ and $H(Y \mid X)$ to capture asymmetry/directionality.

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.



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

Part I

`fastMI`: fast and consistent nonparametric estimator of MI

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 ⬆️⬇️**

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.

Benefits and hurdles of MI

Need a fast, scalable and accurate estimator

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

- $MI = KL\left(f_{XY} \parallel f_X \otimes f_Y\right)$
- $MI = 0 \iff X \perp Y$

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

Benefits and hurdles of MI

Need a fast, scalable and accurate estimator

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

- $MI = KL\left(f_{XY} \parallel f_X \otimes f_Y\right)$
- $MI = 0 \iff X \perp Y$

Need $\hat{f}_{XY}, \hat{f}_X,$ and $\hat{f}_Y$: bandwidth tuning!

Table: Mean (SD) computation time (in seconds) of estimators of MI for bivariate data of varying sample size (n) for $s = 100$ iterations.

	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)

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.

fastMI

Scalable and accurate estimation of MI

- Want: $\hat{f}_{XY}, \hat{f}_X,$ and $\hat{f}_Y$, without tuning to get faster estimate $\hat{MI}$

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

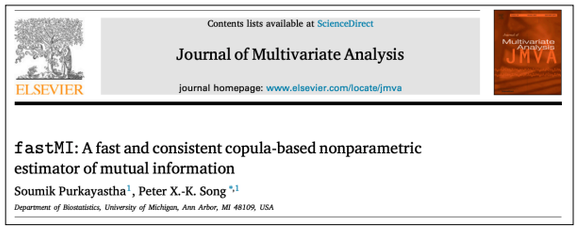
fastMI

Scalable and accurate estimation of MI

- Want: $\hat{f}_{XY}, \hat{f}_X,$ and $\hat{f}_Y$, 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.

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

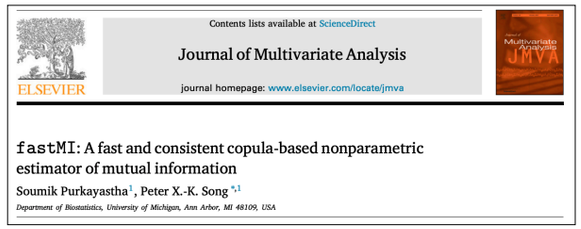
First, it turns out that instead of estimating the joint density along with the marginals, it is enough to simply estimate the underlying copula density function, so our technical complexity is greatly reduced.

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.

fastMI

Scalable and accurate estimation of MI

- Want: $\hat{f}_{XY}, \hat{f}_X,$ and $\hat{f}_Y$, 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.



Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

`fastMI`: A fast and consistent copula-based nonparametric estimator of mutual information

Soumik Purkayastha[1], Peter X.-K. Song [*,1]

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

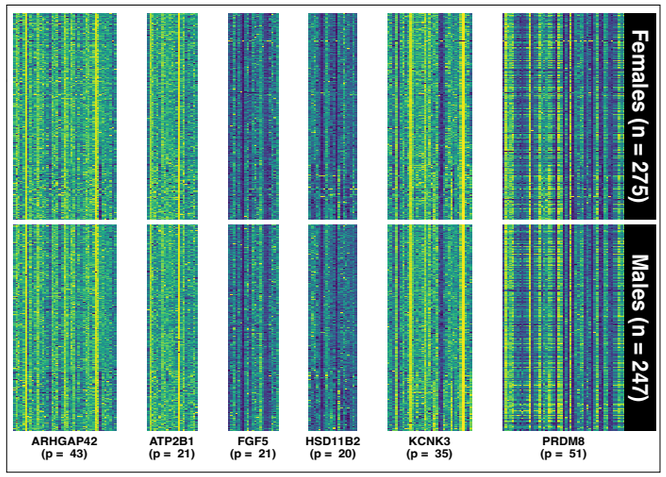
Purkayastha, S., & Song, P. X. K. (2023). fastMI: A fast and consistent copula-based nonparametric estimator of mutual information. Journal of Multivariate Analysis, 105270.

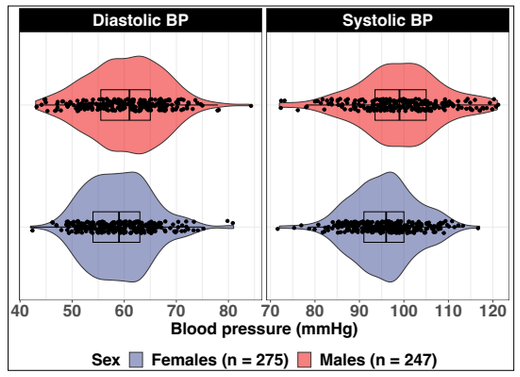
	Sample size (n)		
	1000	2500	5000
Empirical copula-based MI	4.360 (0.356)	5.368 (0.308)	64.040 (0.254)
Jackknifed MI	3.150 (0.107)	18.446 (0.116)	62.454 (4.601)
`fastMI`	**1.199 (0.135)**	**2.964 (0.181)**	**5.952 (0.125)**

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.

p-value plots of fastMI

$MI(SBP, DNAm) = ?$
$MI(DBP, DNAm) = ?$





- Genes: 6. Each gene has many CpG sites
- Compute pairwise MI
- Test if $H_0 : MI = 0$.

So here is my plan: for a given given gene and a specified sex I want to compute pairwise mutual information between methylation scores and blood pressure and test for association between the two.

p-value plots of fastMI

$MI(SBP, DNAm) \neq 0$
$MI(DBP, DNAm) \neq 0$

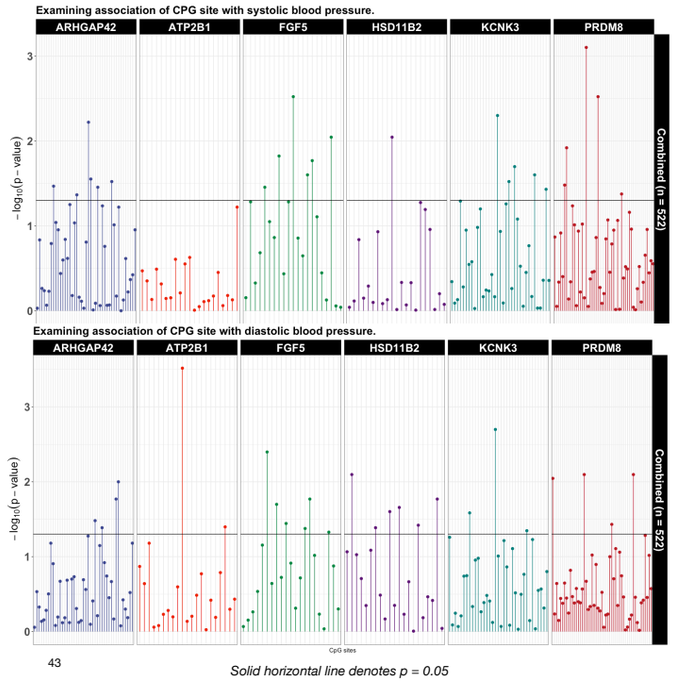
- Powerful test of independence.
- Faster computation leads to scalable estimator!
- fastMI unearths associations between DNAm and BP across all six candidate genes at various CpG sites.

Using the fastMI estimation and inference tool, I compute p-values for testing independence between methylation status and blood pressure.

p-value plots of fastMI

$MI(SBP, DNAm) \neq 0$
$MI(DBP, DNAm) \neq 0$

- Powerful test of independence.
- Faster computation leads to scalable estimator!
- fastMI unearths associations between DNAm and BP across all six candidate genes at various CpG sites.



Solid horizontal line denotes p = 0.05

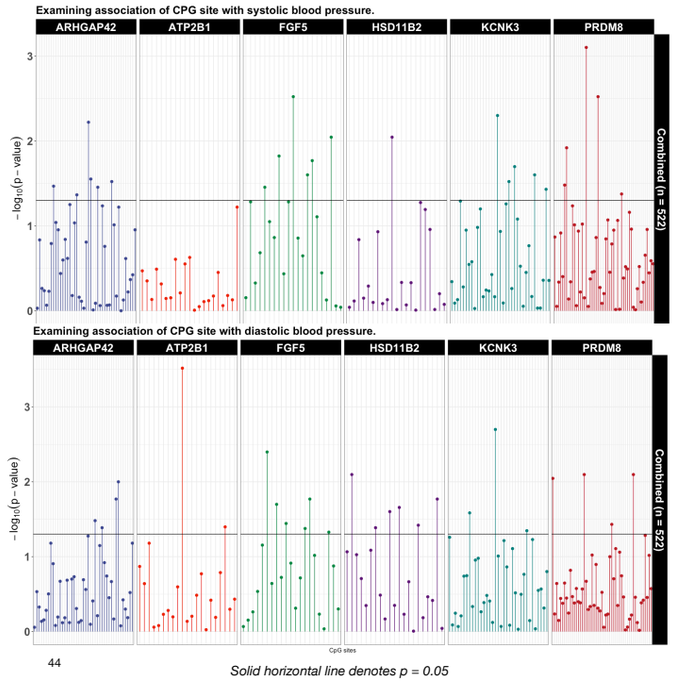
In each of the panels I report a log-transformed p-value. The higher the peak, the stronger the association. From a glance we see a lot of these correlated methylation sites report a p-value of less than 0.05.

p-value plots of fastMI

$MI(SBP, DNAm) \neq 0$
$MI(DBP, DNAm) \neq 0$

- Powerful test of independence.
- Faster computation leads to scalable estimator!
- fastMI unearths associations between DNAm and BP across all six candidate genes at various CpG sites.

But what about directionality?



Solid horizontal line denotes p = 0.05

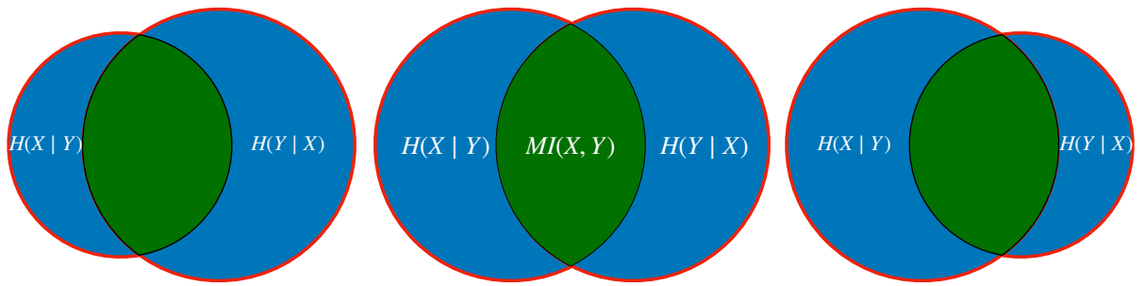
Of course there are other considerations such as multiple testing corrections and p-value aggregation to be considered here, but the key point I want to underline is even if there were remarkably strong signals here there would still be no way of telling directionality.

Part II

Entropy reflects underlying asymmetry

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.

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** $g(\cdot)$

$$Y = g(X)$$

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:

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

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***"**

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?

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

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

I am skipping the technical details on what the exact math notation is, and will focus on the intuition instead.

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.

Asymmetry in GEMs using entropy

Approve or disprove $X \to Y$

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

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

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

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

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

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

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 coefficient $C_{X>Y}$

Advantages and challenges

Strong asymmetry:

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

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 $C_{X>Y}$

Advantages and challenges

Strong asymmetry:

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

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

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

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 $C_{X>Y}$

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

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

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

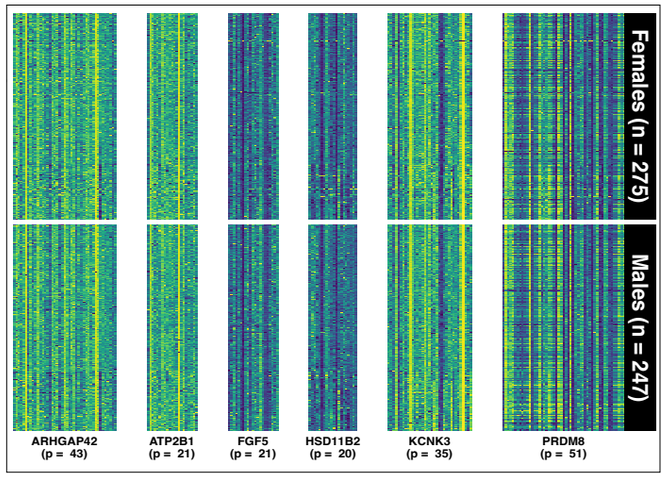
$$\sigma_C^2 = V[\log(f_X(X)) + \log(f_Y(Y))]$$

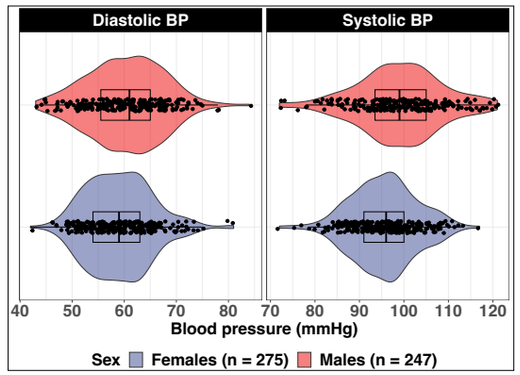
Estimated by Monte-Carlo methods with estimated $\hat{f}_X$ and $\hat{f}_Y$

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.

Adjusting for Z

Low-dimensional confounders

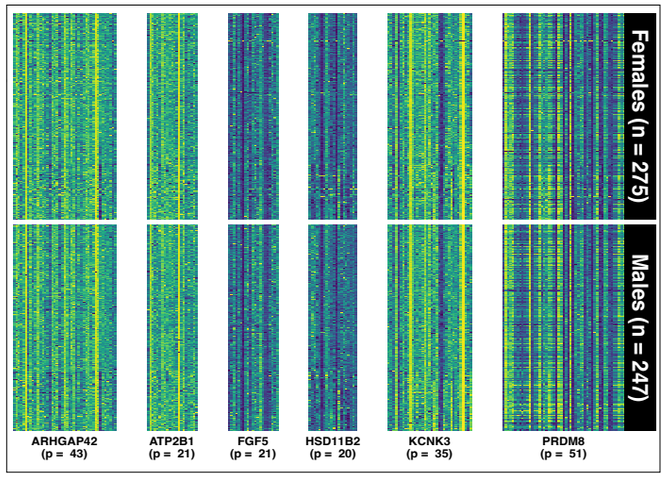


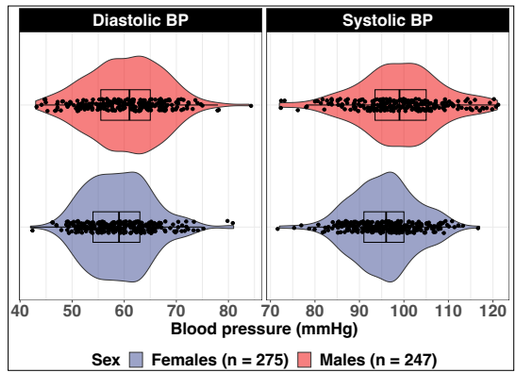


Also, given our biological considerations, a stratified analysis is must.

Adjusting for $\mathbf{Z}$

Low-dimensional confounders





Directionality in X and Y conditional on $\mathbf{Z}$:

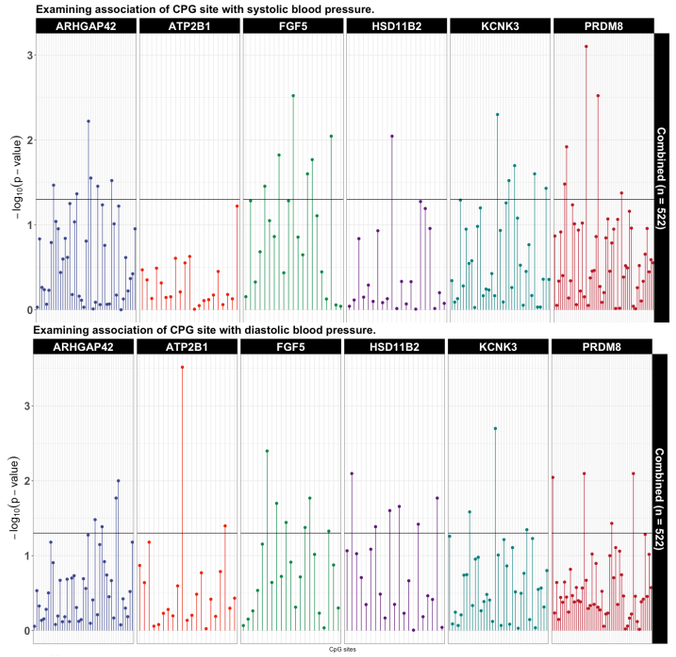
$$C_{X>Y|\mathbf{Z}} := H(X \mid \mathbf{Z} = \mathbf{z}) - H(Y \mid \mathbf{Z} = \mathbf{z})$$

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

$BP \rightarrow DNAm?$

Use $\hat{C}_{X>Y}$ for clues!

But what about directionality?



I want to go back to that plot of p-values and note how earlier we were unable to ascertain directionality.

$BP \rightarrow DNAm?$

Use $\hat{C}_{X>Y}$ for clues!

But what about directionality?

Correlated CpG sites: Aggregated mean DNAm for a given gene.

X : BP (either diastolic or systolic)
Y : $DNAm$ for a given gene

Obtain $\hat{C}_{X>Y}$ (95% CI)

Stratify by Z: sex of study participant.

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

$BP \rightarrow DNAm?$

Use $\hat{C}_{X>Y}$ for clues!

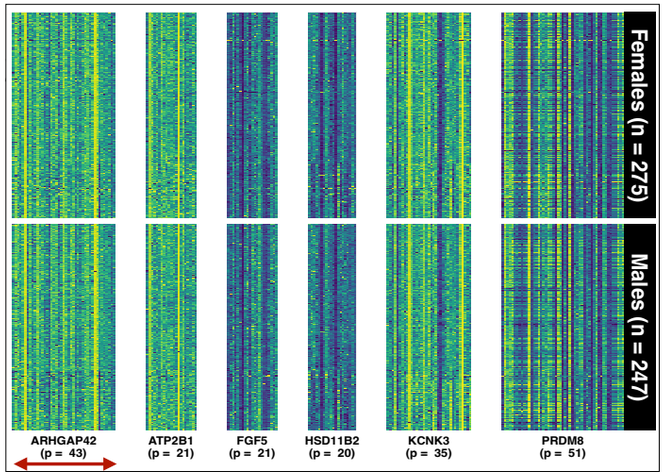
But what about directionality?

Correlated CpG sites: Aggregated mean DNAm for a given gene.

X : BP (either diastolic or systolic)
Y : $DNAm$ for a given gene

Test if $C_{X>Y} > 0$ using $\hat{C}_{X>Y}$ (95% CI)

Stratify by Z: sex of study participant.



Mean DNAm score across all CpG sites

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

$BP \rightarrow DNAm?$

Use $\hat{C}_{X>Y}$ for clues!

Correlated CpG sites: Aggregated mean DNAm for a given gene.

X : BP (either diastolic or systolic)
Y : $DNAm$ for a given gene

Test if $C_{X>Y} > 0$ using $\hat{C}_{X>Y}$ (95% CI)

Stratify by Z: sex of study participant.

Forest plot of $\hat{C}_{X>Y}$ (95% CI)

- Estimates for female group in blue
- Estimates for male group in red
- Unadjusted estimates in green
- Adjusted estimates in violet

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.

$BP \rightarrow DNAm?$

Use $\hat{C}_{X>Y}$ for clues!

Correlated CpG sites: Aggregated mean DNAm for a given gene.

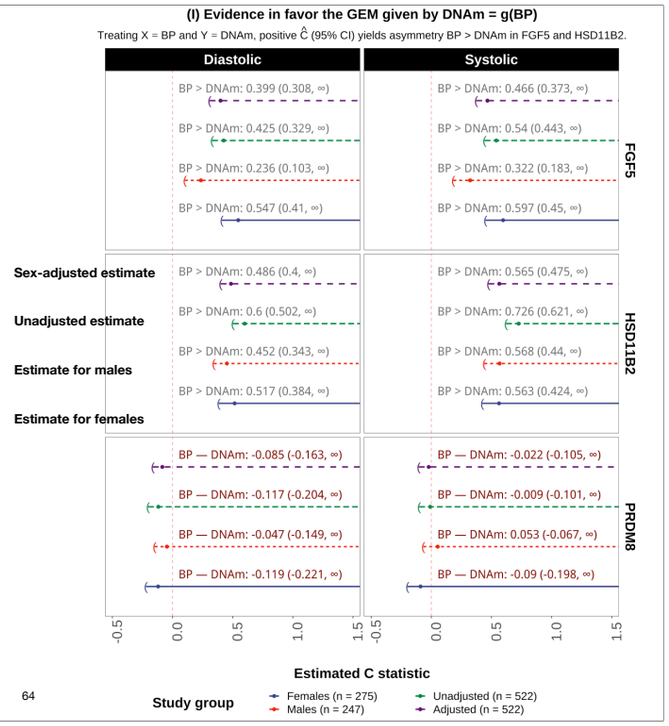
X : BP (either diastolic or systolic)
Y : $DNAm$ for a given gene

Test if $C_{X>Y} > 0$ using $\hat{C}_{X>Y}$ (95% CI)

Stratify by Z: sex of study participant.

Forest plot of $\hat{C}_{X>Y}$ (95% CI)

- Estimates for female group in blue
- Estimates for male group in red
- Unadjusted estimates in green
- Adjusted estimates in violet



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.

$BP \rightarrow DNAm$?

Use $\hat{C}_{X>Y}$ for clues!

$X = DBP$ $\quad$ $Y = DNAm$ of $HSD11B2$

Hypothesized GEM under $H_0 : Y = g(X)$

Want to test $H_0 : C_{X>Y} > 0$ using $\hat{C}_{X>Y}$ $(95\,\%$ one-sided $CI)$

- For female subgroup: $0.517\ (0.384,\ \infty)$
- Evidence to protect the null.

$BP \rightarrow DNAm$?

Use $\hat{C}_{X>Y}$ for clues!

$X = DBP$ $\quad$ $Y = DNAm$ of $HSD11B2$

Hypothesized GEM under $H_0 : Y = g(X)$

Want to test $H_0 : C_{X>Y} > 0$ using $\hat{C}_{X>Y}$ (95 % one-sided CI)

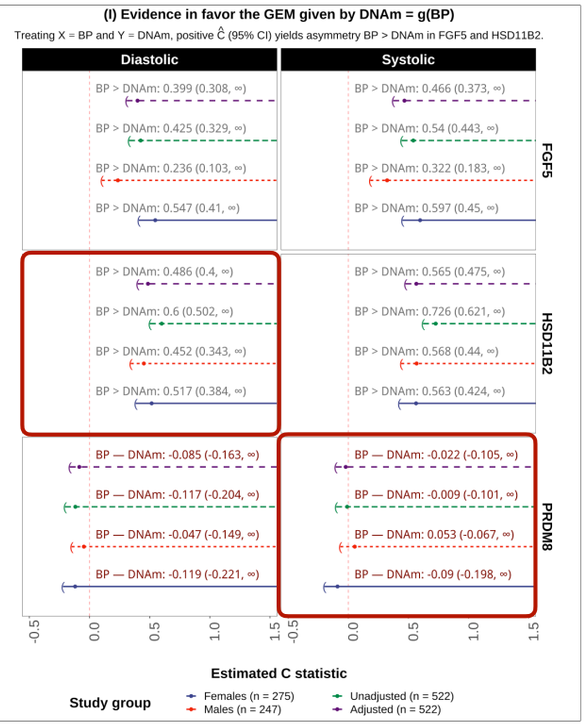
- For female subgroup: $0.517\ (0.384,\ \infty)$
- Evidence to protect the null.

$X = DBP$ $\quad$ $Y = DNAm$ of $PRDM8$

Hypothesized GEM under $H_0 : Y = g(X)$

Want to test $H_0 : C_{X>Y} > 0$ using $\hat{C}_{X>Y}$ (95 % one-sided CI)

- For female subgroup: $-0.090\ (-0.198,\ \infty)$
- Evidence to reject null.



$BP \rightarrow DNAm$?

Use $\hat{C}_{X>Y}$ for clues!

$X = DBP$ $\quad$ $Y = DNAm$ of $HSD11B2$

Hypothesized GEM under $H_0 : Y = g(X)$

Want to test $H_0 : C_{X>Y} > 0$ using $\hat{C}_{X>Y}$ (95 % one-sided CI)

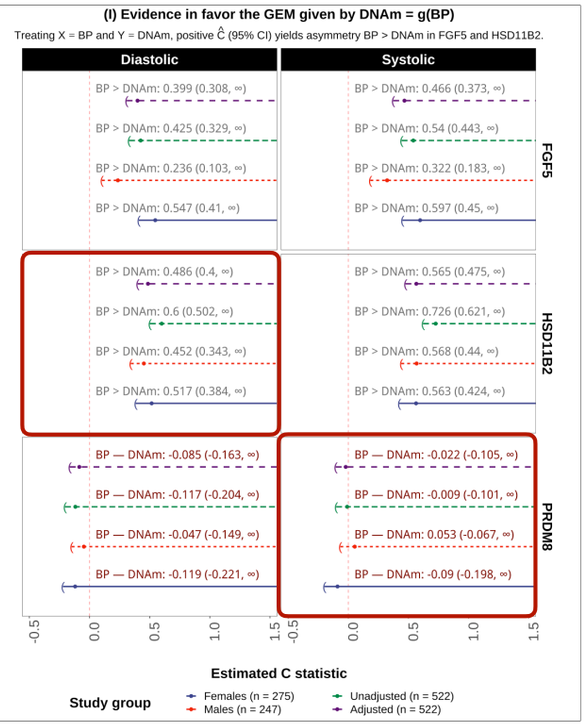
- For female subgroup: $0.517\ (0.384,\ \infty)$
- Evidence to protect the null.

$X = DBP$ $\quad$ $Y = DNAm$ of $PRDM8$

Hypothesized GEM under $H_0 : Y = g(X)$

Want to test $H_0 : C_{X>Y} > 0$ using $\hat{C}_{X>Y}$ (95 % one-sided CI)

- For female subgroup: $-0.090\ (-0.198,\ \infty)$
- Evidence to reject null. Alternate direction??



$BP \rightarrow DNAm?$

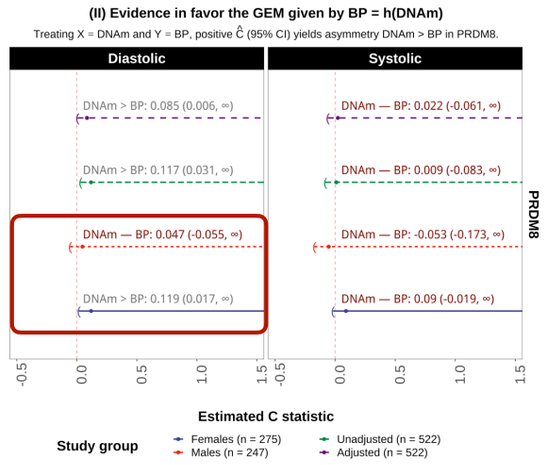
Use $\hat{C}_{X>Y}$ for clues!

$X = DNAm$ of $PRDM8$ $\quad\quad$ $Y = DBP$

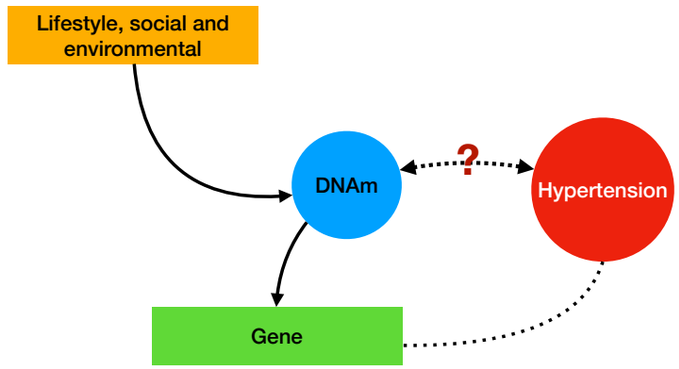
Hypothesized GEM under $H_0 : Y = g(X)$

Want to test $H_0 : C_{X>Y} > 0$ using $\hat{C}_{X>Y}$ $(95\,\%$ one-sided $CI)$

- Simpson's paradox.
- Stratifying is a good idea.

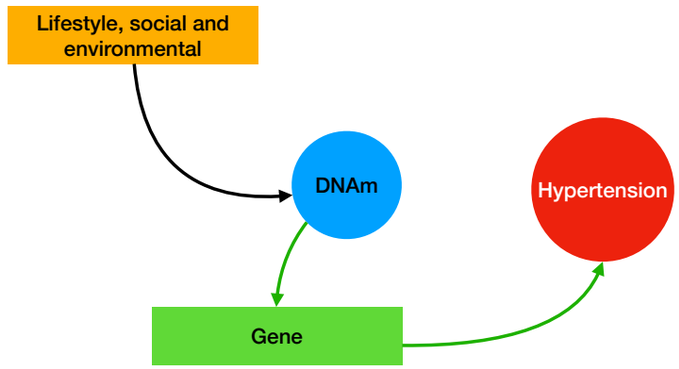


$BP \rightarrow DNAm?$



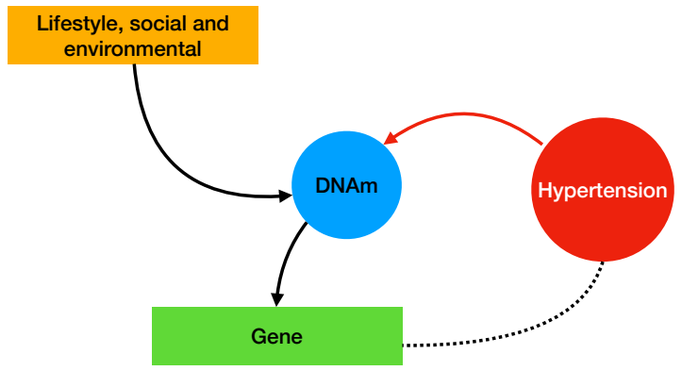
So to go back to our diagram,

$$BP \rightarrow DNAm?$$



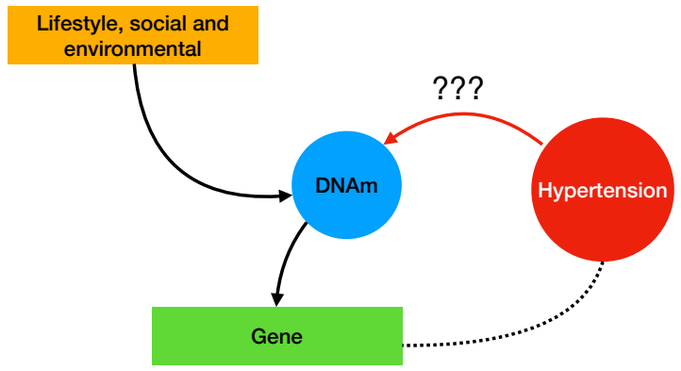
The pathway starting from dna methylation to hypertension is one that is probably intuitive. It is in line with our experience of writing snps as predictors of phenotypes, and the understanding that methylation is more like a controlling switch in that equation.

$BP \rightarrow DNAm?$



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

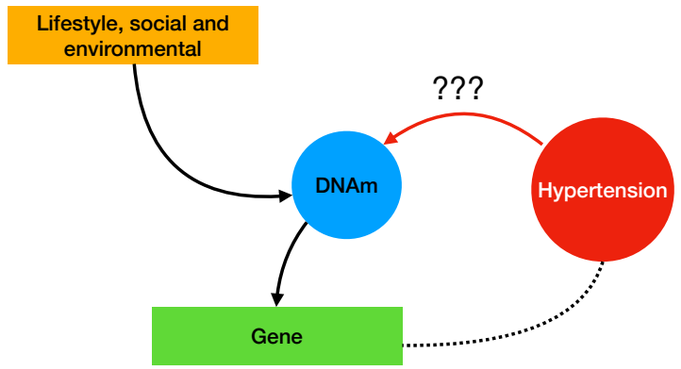
$BP \rightarrow DNAm?$



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

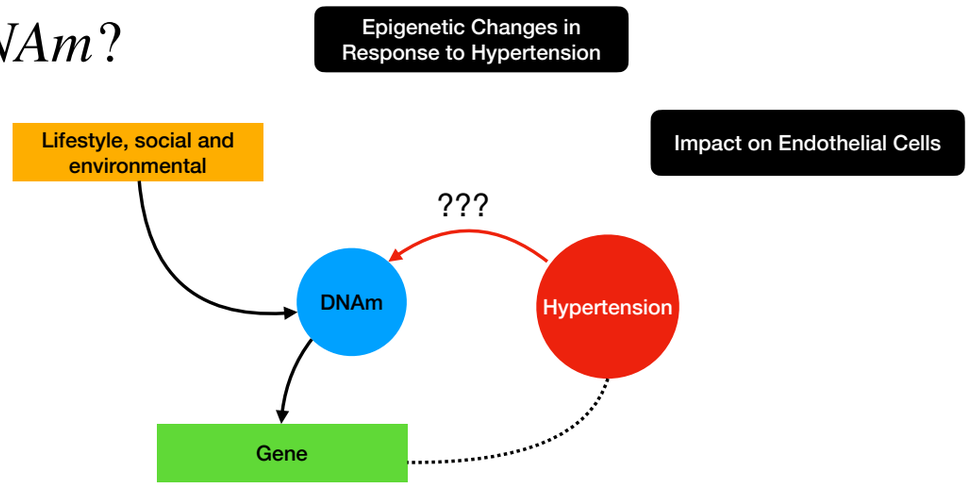
$BP \rightarrow DNAm?$

Epigenetic Changes in Response to Hypertension

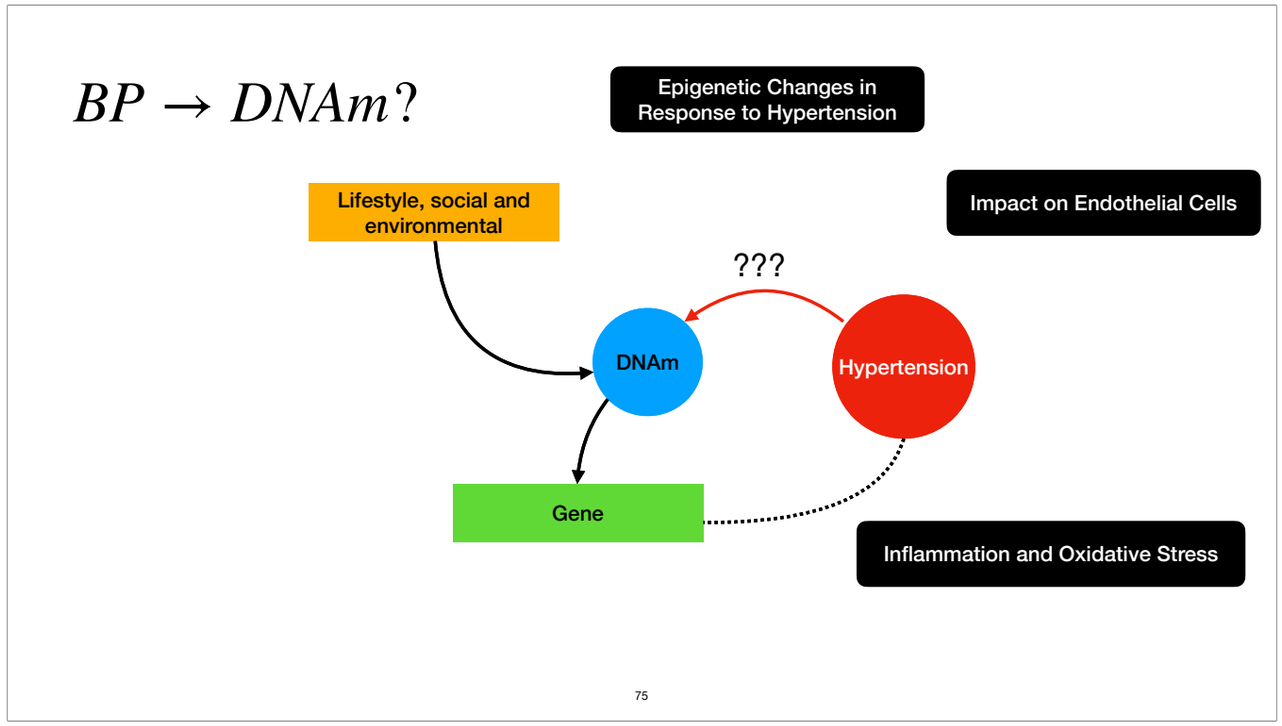


Stress responses and inflammation associated with hypertension can influence the epigenetic landscape.

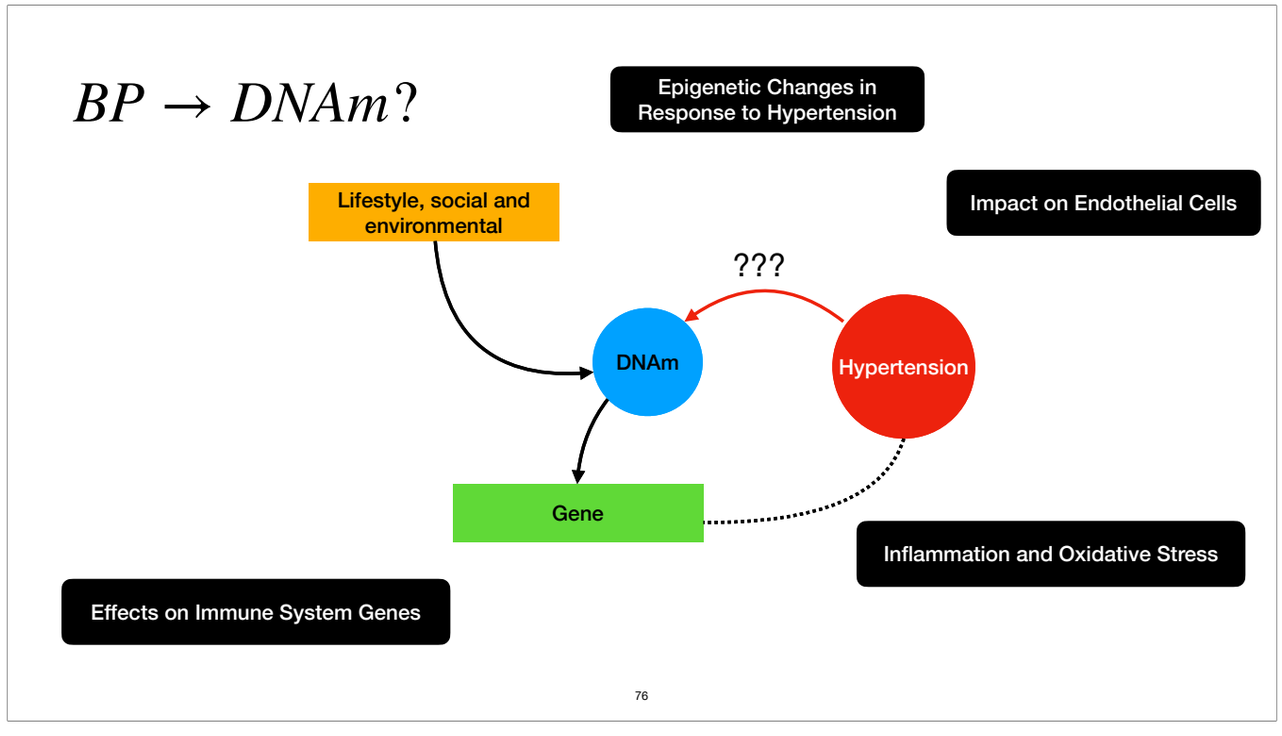
$BP \rightarrow DNAm?$



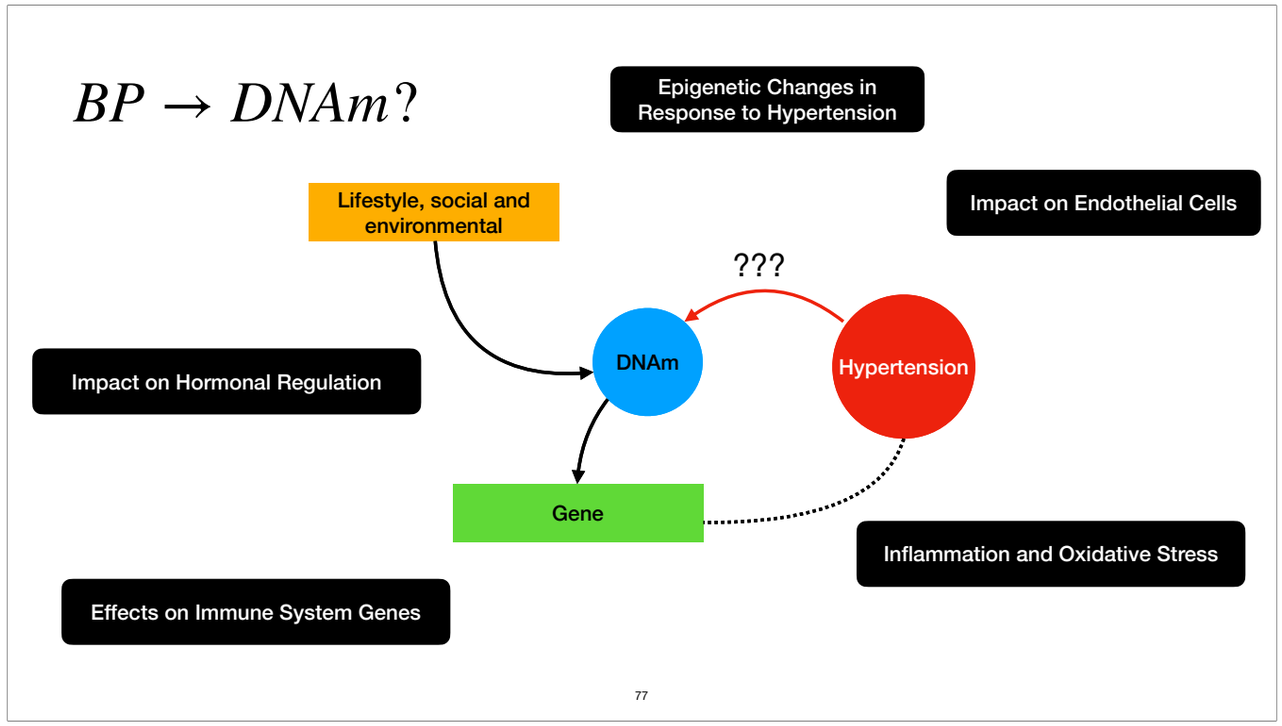
Further, endothelial cells are attached to blood pressure regulation. Changes in blood pressure might affect these cells, influencing DNA methylation patterns.



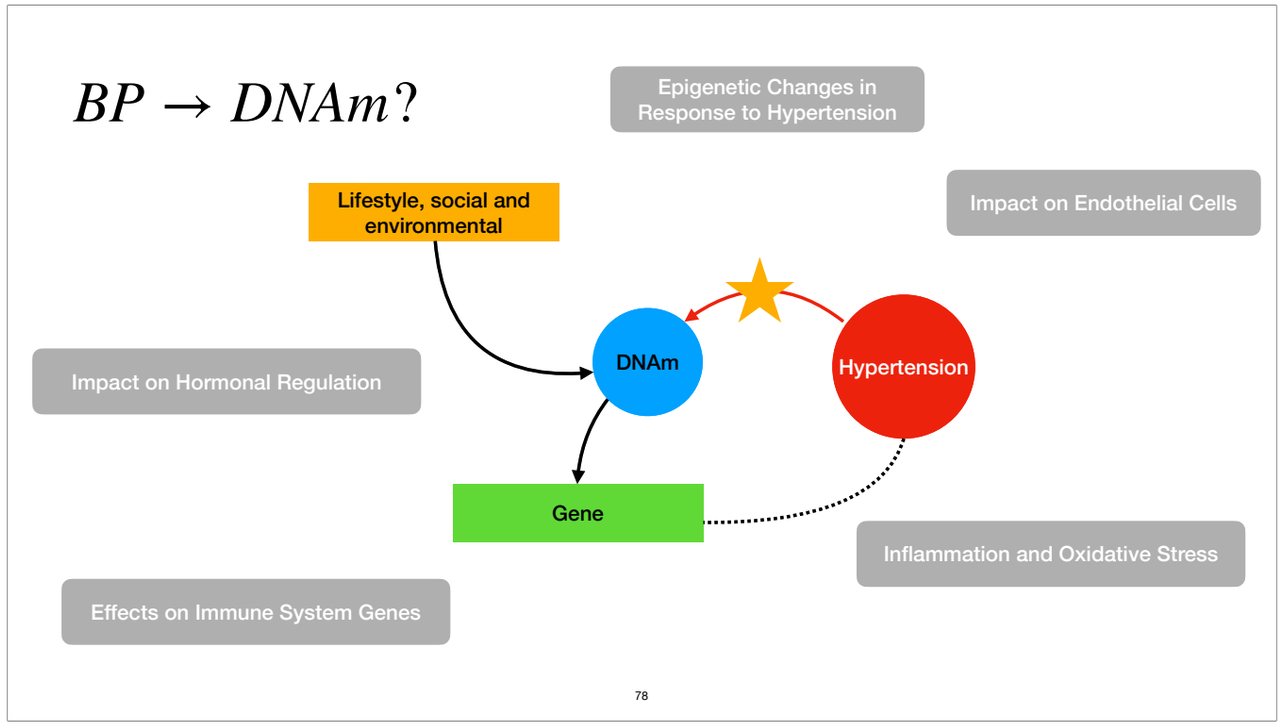
Hypertension is often associated with chronic inflammation and oxidative stress. Both inflammation and oxidative stress can modulate DNA methylation patterns.



Hypertension may influence the immune system, and alterations in immune cell DNA methylation patterns have been reported.



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.



So to summarise, the findings we report indicate support for potential pathways going from blood pressure to dna methylation and deserve more investigation.

Toolkit to study association and direction

Summary

Components of new toolkit

Technical strengths of toolkit

Scientific question examined

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

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

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.**

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.

Future work

Future work

Application: Genome-wide scalability?

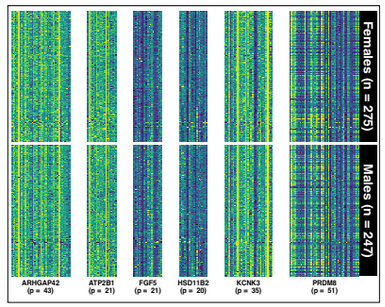
Application: tracking longitudinal BP measurement?

Methods: Mediator or collider?

Future work I

Application: Genome-wide scalability?

1. **Chose six candidate genes as a proof-of concept.**

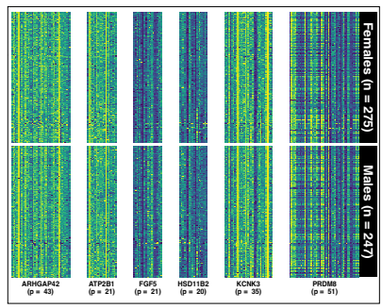


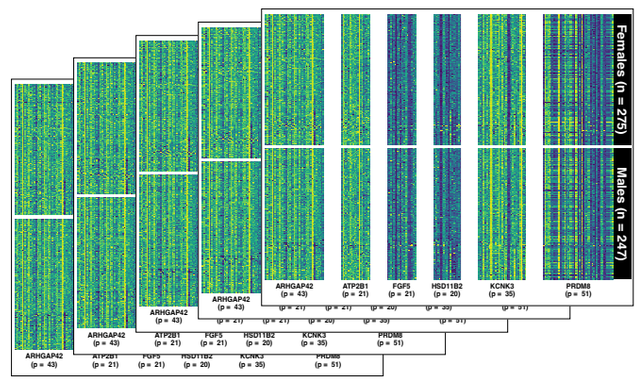
I want to explore third-variable structures using GEMs. What I mean by that are mediator and collider models that show various pathways between exposure and outcome.

Future work I

Application: Genome-wide scalability?

1. Chose six candidate genes as a proof-of concept.
2. Extend to entire genome?



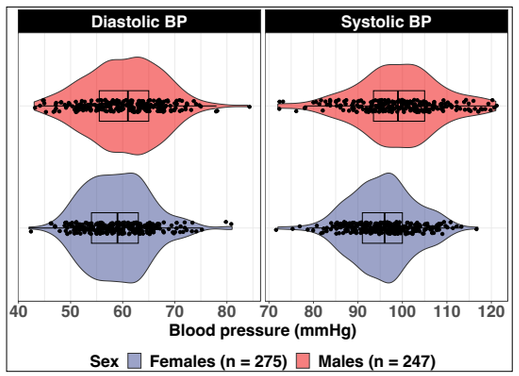


I want to explore third-variable structures using GEMs. What I mean by that are mediator and collider models that show various pathways between exposure and outcome.

Future work II

Application: tracking longitudinal BP measurement?

1. Evidence for $DNAm \rightarrow BP$ in *PRDM8* and *HSD11B2.*
2. SBP and DBP measured at a single time point.

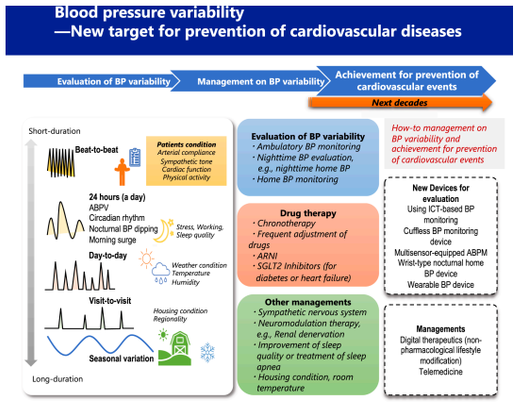


I want to explore third-variable structures using GEMs. What I mean by that are mediator and collider models that show various pathways between exposure and outcome.

Future work II

Application: tracking longitudinal BP measurement?

1. Evidence for $DNAm \rightarrow BP$ in *PRDM8* and *HSD11B2.*
2. SBP and DBP measured at a single time point.
3. Better measures of cardiovascular function?
    1. Variability of BP?



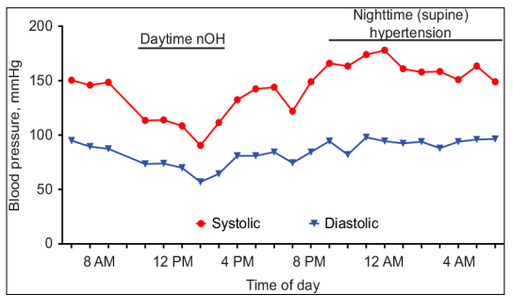
Narita, Keisuke, Satoshi Hoshide, and Kazuomi Kario. "Short-to long-term blood pressure variability: Current evidence and new evaluations." Hypertension Research 46.4 (2023): 950-958.

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.

Future work II

Application: tracking longitudinal BP measurement?

1. Evidence for $DNAm \rightarrow BP$ in *PRDM8* and *HSD11B2.*
2. SBP and DBP measured at a single time point.
3. Better measures of cardiovascular function?
    1. Variability of BP?
4. Longitudinal comparison of $DNAm \leftrightarrow BP$?



Biswas, Debashis, Beverly Karabin, and Debra Turner. "Role of nurses and nurse practitioners in the recognition, diagnosis, and management of neurogenic orthostatic hypotension: a narrative review." International Journal of General Medicine (2019): 173-184.

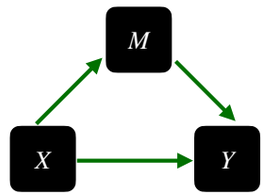
I want to explore third-variable structures using GEMs. What I mean by that are mediator and collider models that show various pathways between exposure and outcome.

Methods

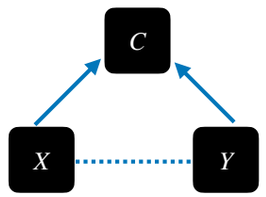
Future work III

Mediator or collider?

Extend GEM framework to detect third-variable DAGs



M mediates path from X to Y



C is a collider for X, Y

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.

SCAN ME

Thanks!

soumikp@umich.edu

soumikp.github.io

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 $\hat{c}_{\mathbf{Z}}$

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 $MISE = \mathbb{E}\left[\int \left\{\hat{c}_{\mathbf{z}}(\mathbf{z}) - c_{\mathbf{z}}(\mathbf{z})\right\}^2 d\mathbf{z}\right]$

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 $MISE = \mathbb{E}\left[\int \left\{\hat{c}_{\mathbf{z}}(\mathbf{z}) - c_{\mathbf{z}}(\mathbf{z})\right\}^2 d\mathbf{z}\right]$

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

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

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 $MISE = \mathbb{E}\left[\int \left\{\hat{c}_{\mathbf{z}}(\mathbf{z}) - c_{\mathbf{z}}(\mathbf{z})\right\}^2 d\mathbf{z}\right]$

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

   $MISE$ 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 $\hat{\mathscr{C}}$

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 $MISE = \mathbb{E}\left[\int \left\{\hat{c}_{\mathbf{z}}(\mathbf{z}) - c_{\mathbf{z}}(\mathbf{z})\right\}^2 d\mathbf{z}\right]$

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

   $MISE$ 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 $\mathscr{C}$

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

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

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 $MISE = \mathbb{E}\left[\int \left\{\hat{c}_{\mathbf{z}}(\mathbf{z}) - c_{\mathbf{z}}(\mathbf{z})\right\}^2 d\mathbf{z}\right]$

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

   $MISE$ 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 $\mathscr{C}$

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

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**.

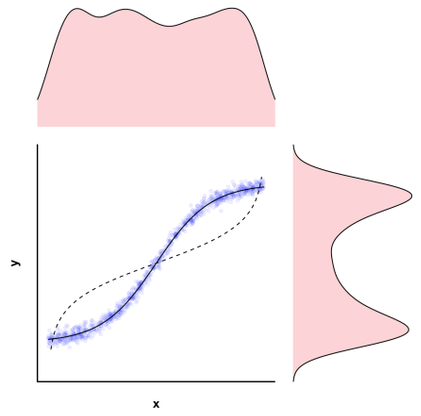
Appendix III

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.

Appendix IV

Another application of the GEM framework

Auto-MPG dataset: confirming known pathways

Plot A: Displacement → MPG

- **North America:** 2.590 (2.514, 2.666)
- **Rest:** 0.767 (0.636, 0.898)
- **Combined:** 1.908 (1.886, 1.930)

Plot B: Horsepower → MPG

- **North America:** 1.681 (1.559, 1.803)
- **Rest:** 0.749 (0.645, 0.853)
- **Combined:** 1.332 (1.308, 1.356)

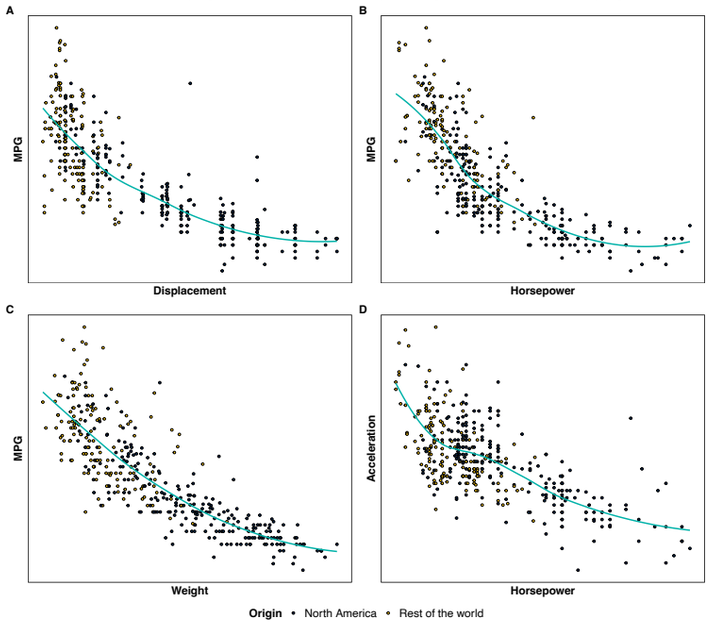
Plot C: Weight → MPG

- **North America:** 4.631 (4.558, 4.704)
- **Rest:** 3.664 (3.518, 3.810)
- **Combined:** 4.269 (4.247, 4.292)

Plot D: Horsepower → Acceleration

- **North America:** 2.798 (2.693, 2.904)
- **Rest:** 2.200 (2.066, 2.334)
- **Combined:** 2.574 (2.550, 2.598)

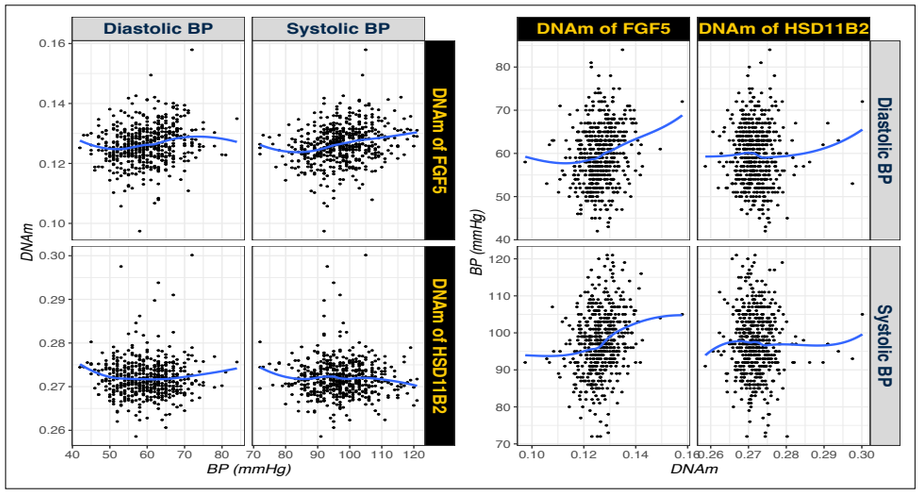
Pairwise scatterplots of features from the Auto MPG dataset of mechanical attributes of n = 390 cars.
Attributes: fuel consumption in miles per gallon (MPG), displacement, horsepower, weight, and acceleration, stratified by origin of car.



Appendix V

Epigenetics and blood pressure

Association and direction?



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 $\hat{c}_{\mathbf{Z}}$

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.

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

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.

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

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

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

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 $X \perp \epsilon$.

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.

Appendix VII

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}

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

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.

Appendix VII

Coefficient can tolerate contaminated outcomes!

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

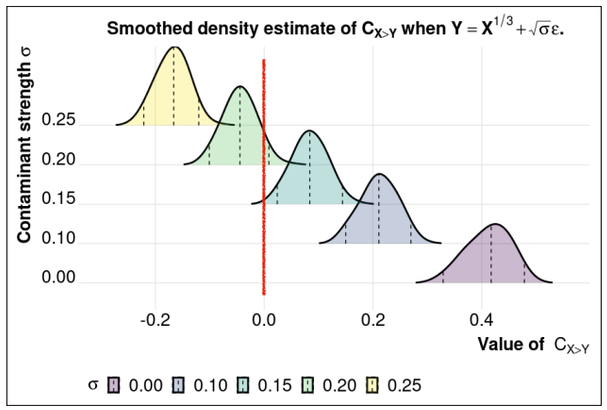
- Noise-perturbed GEM (NPGEM)

  $Y^* = g(X) + \sqrt{\sigma}\epsilon$, 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}

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



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.

Appendix VII

Coefficient can tolerate contaminated outcomes!

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

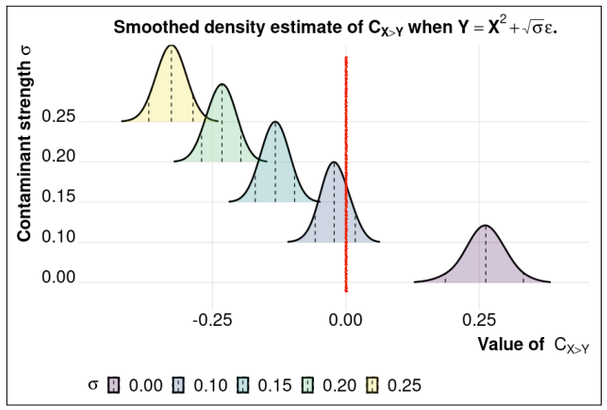
- Noise-perturbed GEM (NPGEM)

  $Y^* = g(X) + \sqrt{\sigma}\epsilon$, 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}

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



For some cases we see even small contamination is enough to throw our statistic off.