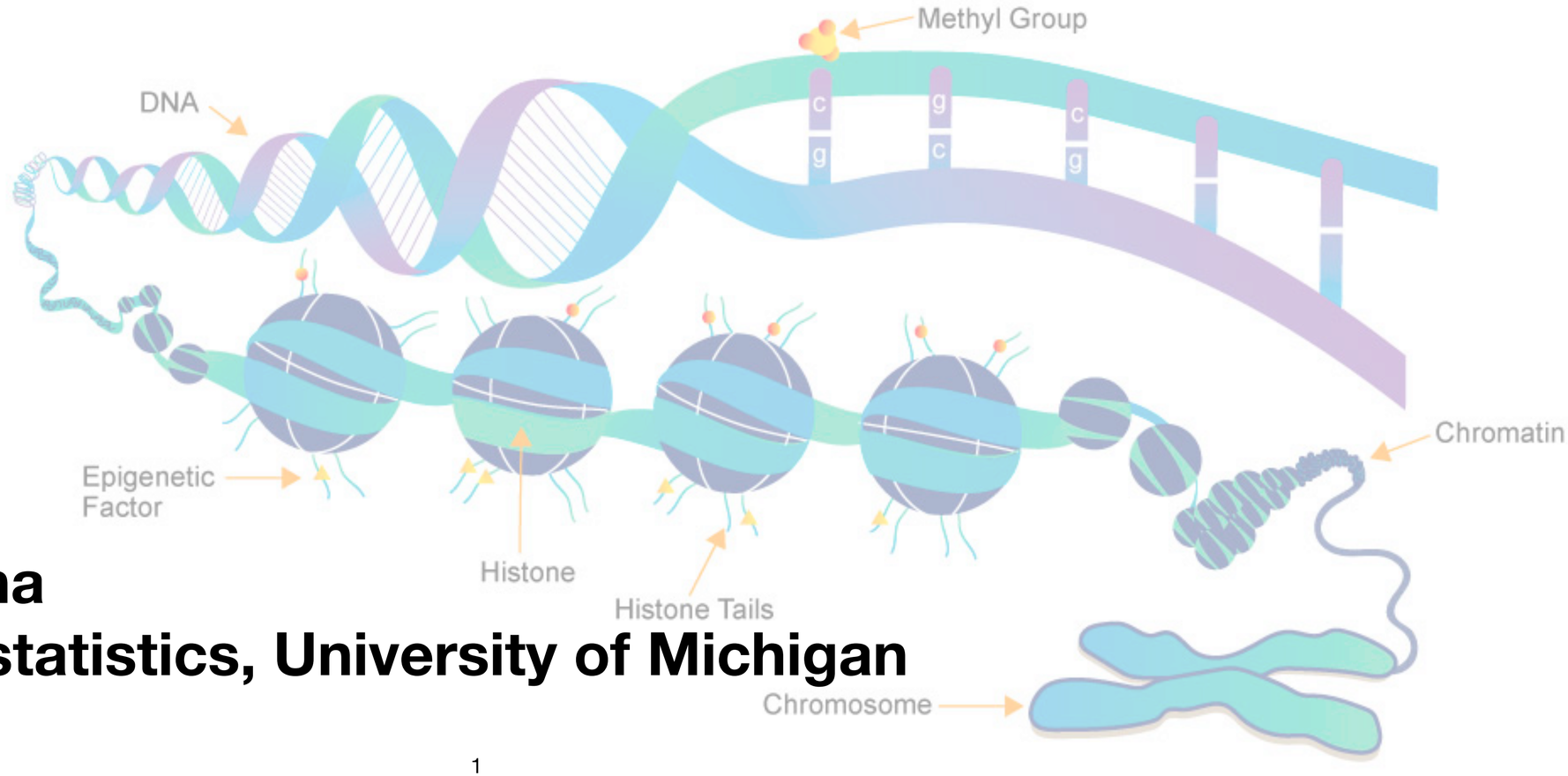
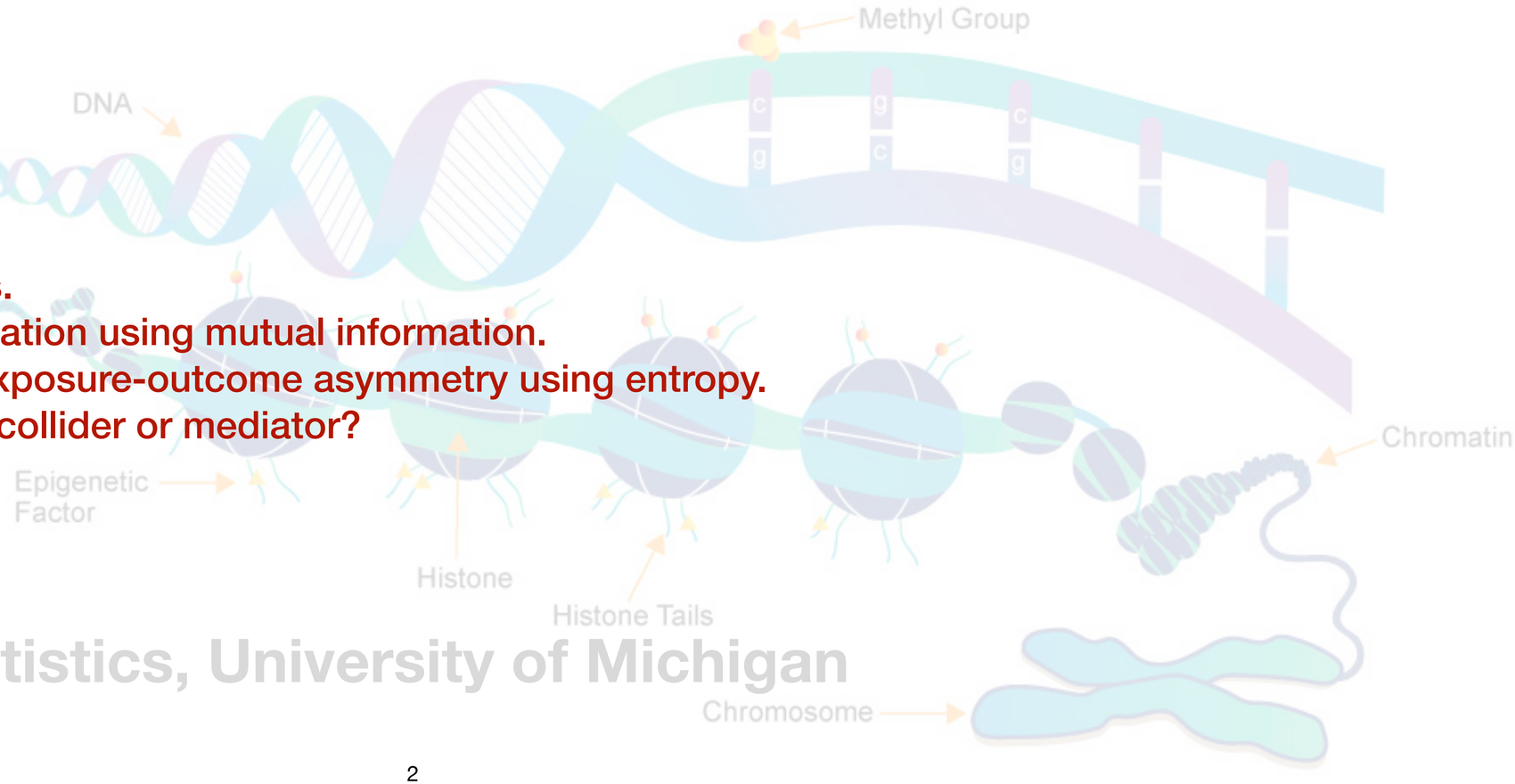
Statistical methods to investigate asymmetric association and directionality in biomedical studies



Soumik Purkayastha
Department of Biostatistics, University of Michigan

January 30, 2024

Statistical methods to investigate asymmetric association and directionality in biomedical studies



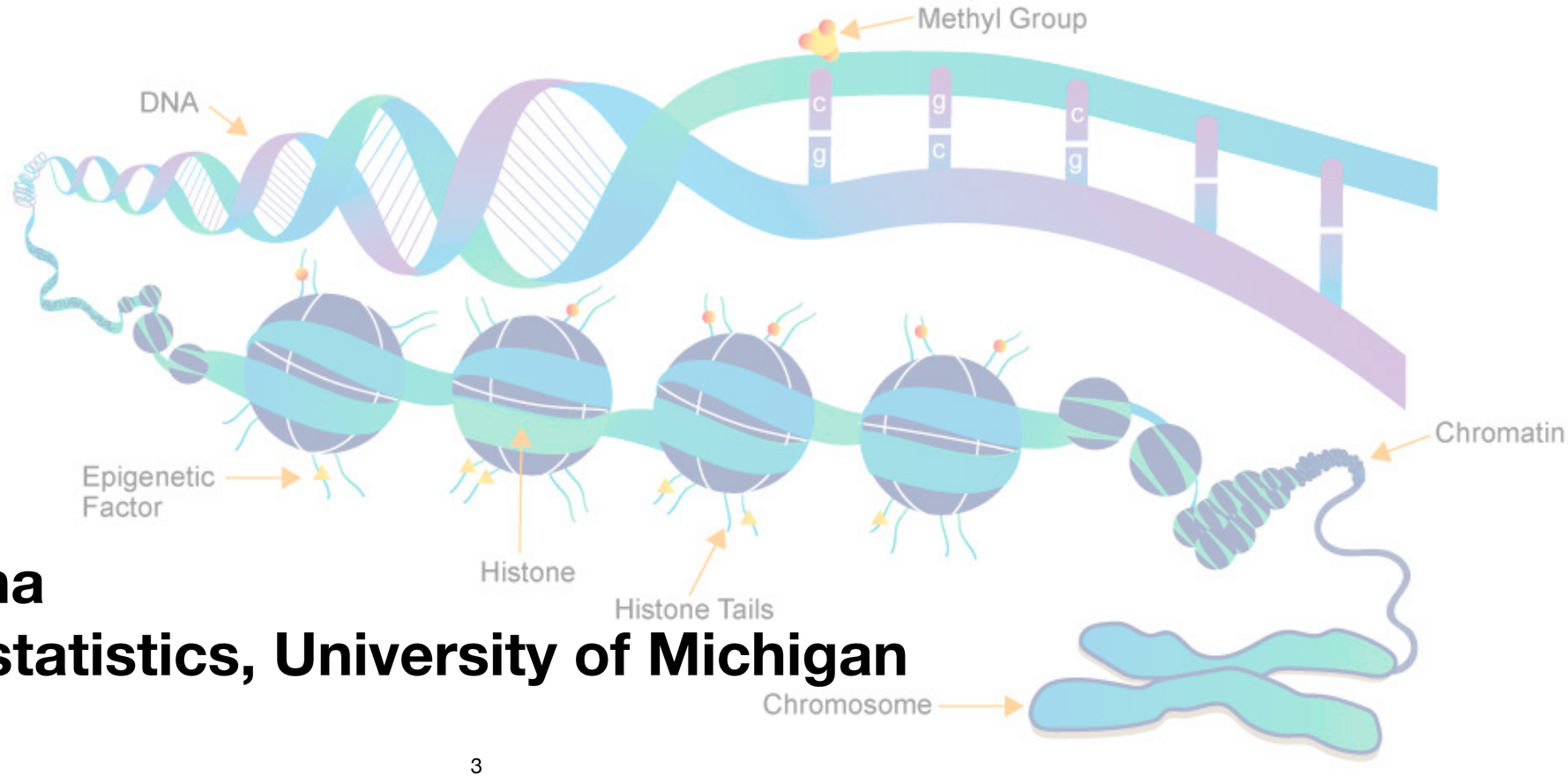
Overview:

1. Why care about directionality?
2. Introduction to epigenomics.
3. fastMI: studying association using mutual information.
4. GEMs: studying exposure-outcome asymmetry using entropy.
5. Future work: collider or mediator?

Soumik Purkayastha
Department of Biostatistics, University of Michigan

January 30, 2024

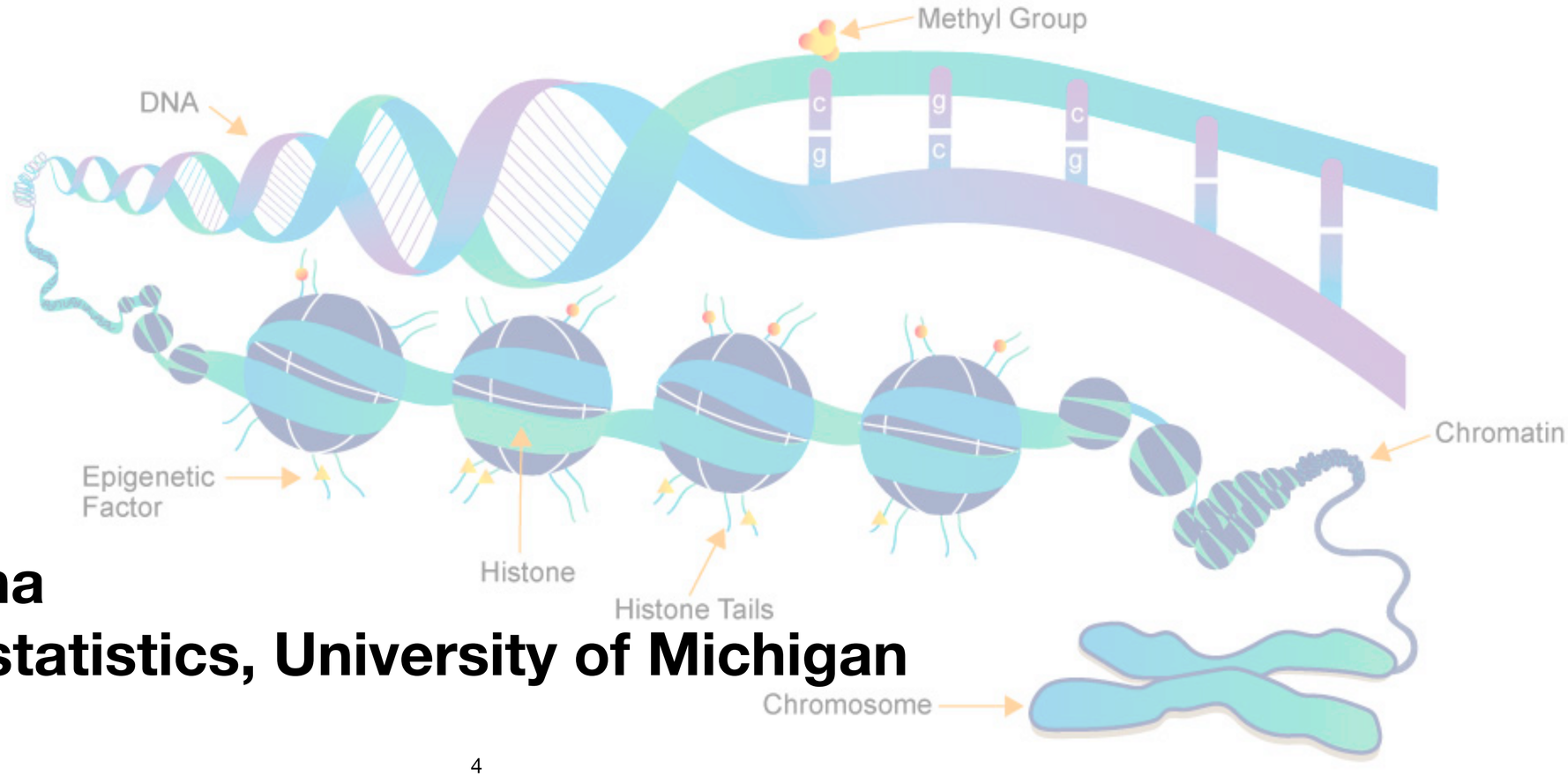
Statistical methods to investigate asymmetric *association* and *directionality* in biomedical studies



Soumik Purkayastha
Department of Biostatistics, University of Michigan

January 30, 2024

Statistical methods to investigate asymmetric *association* and *directionality* in *biomedical studies*



Soumik Purkayastha
Department of Biostatistics, University of Michigan

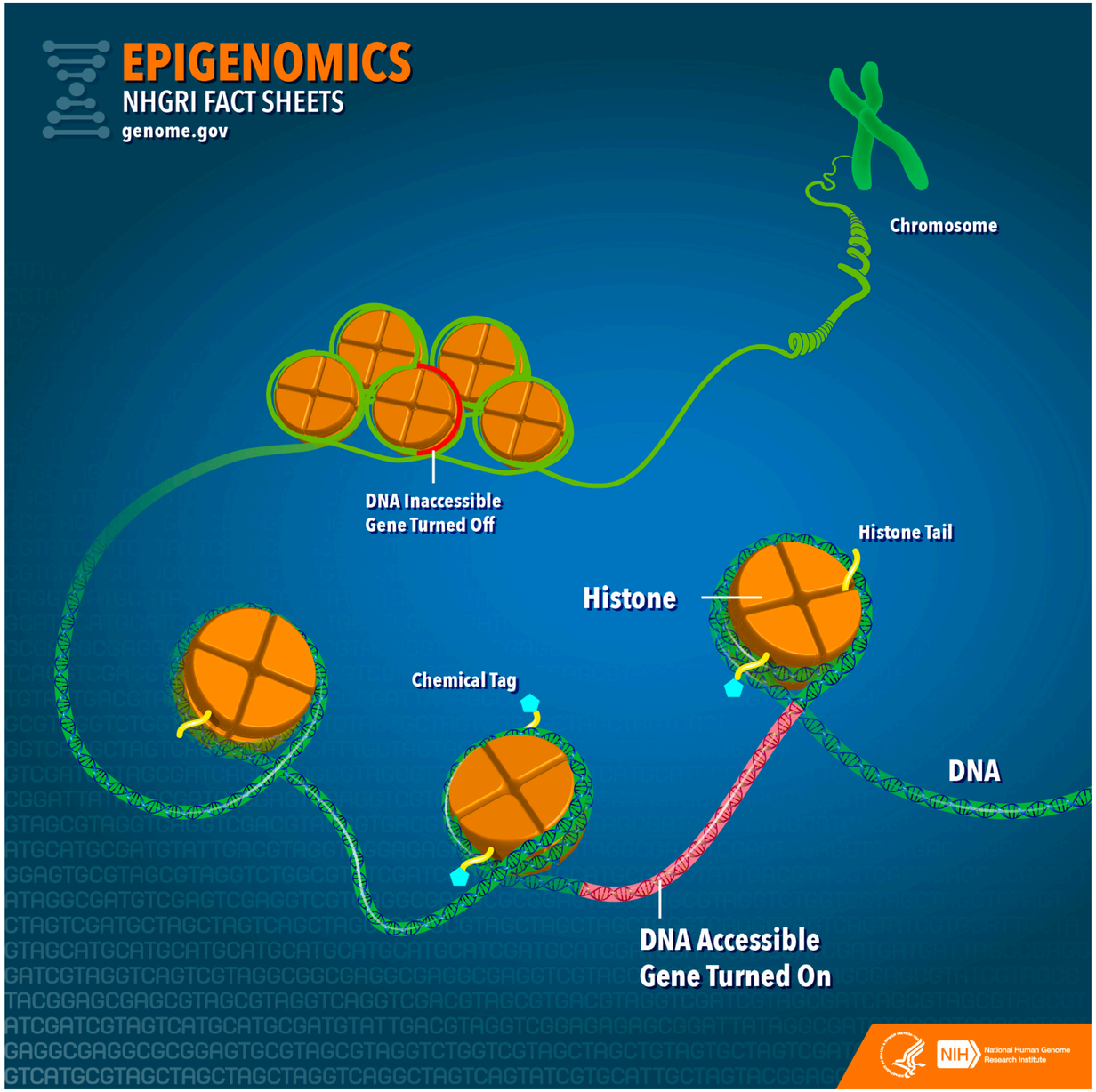
January 30, 2024

Epigenomics

Motivating problem

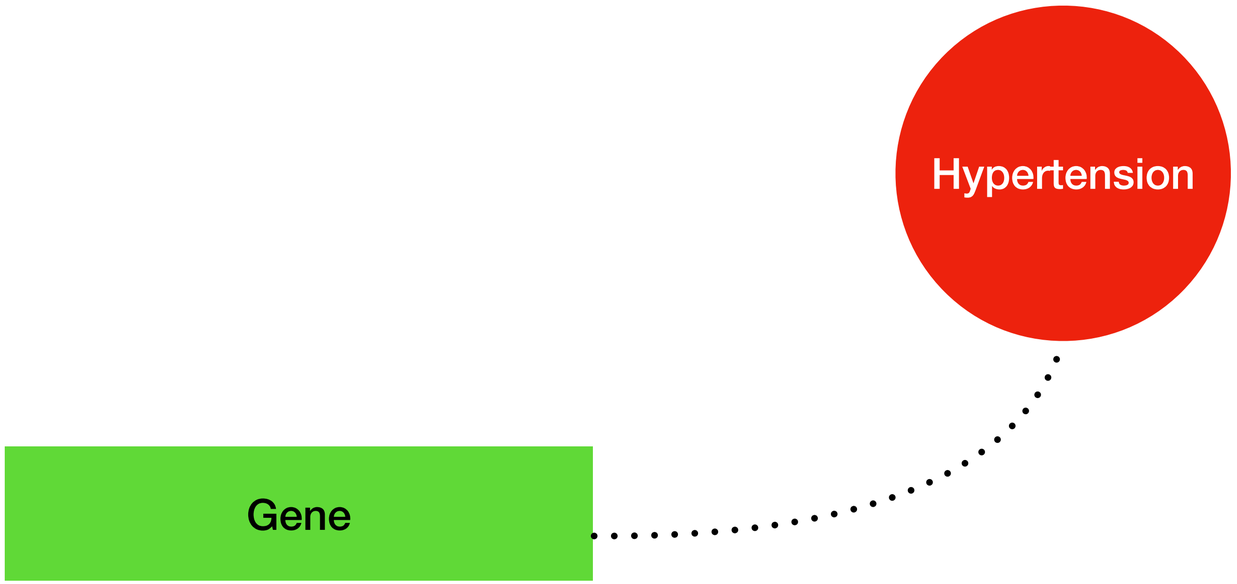
Epigenomics

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

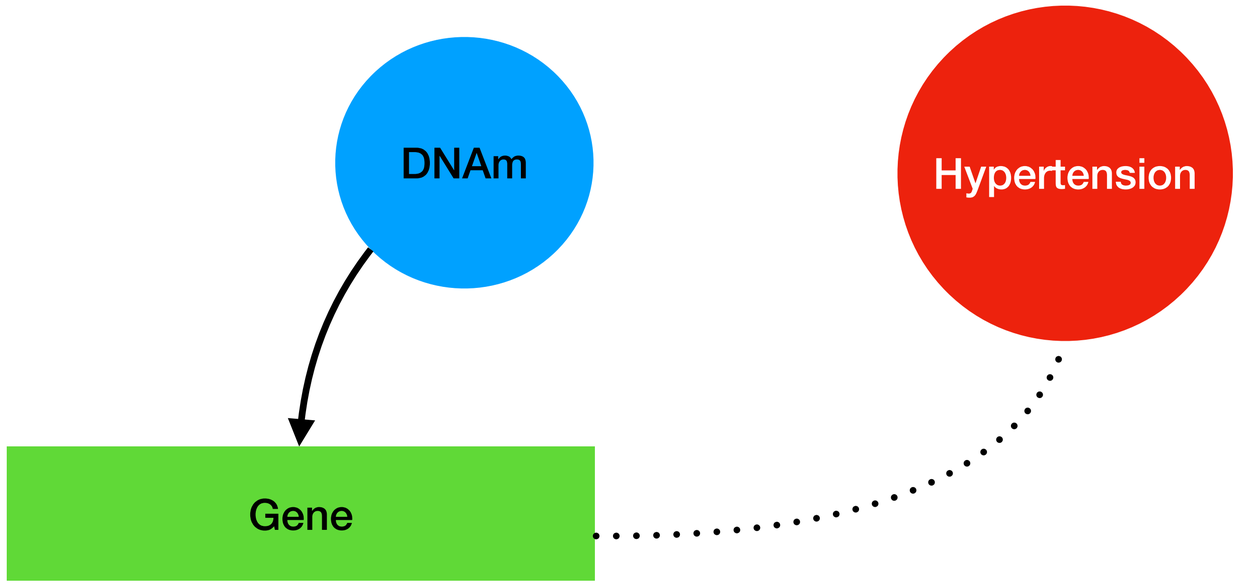


https://www.genome.gov/about-genomics/fact-sheets/Epigenomics-Fact-Sheet

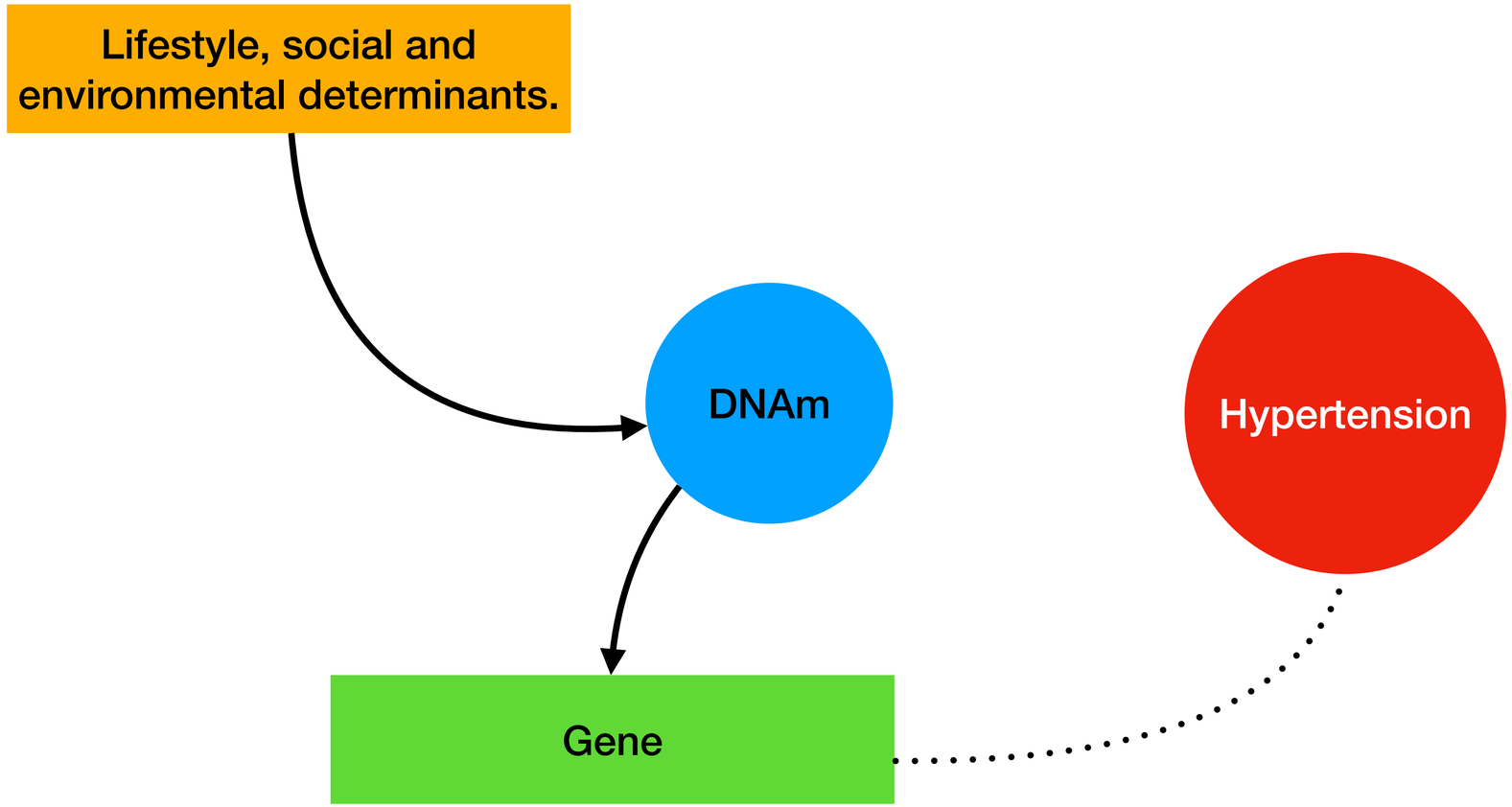
DNAm and cardiovascular diseases



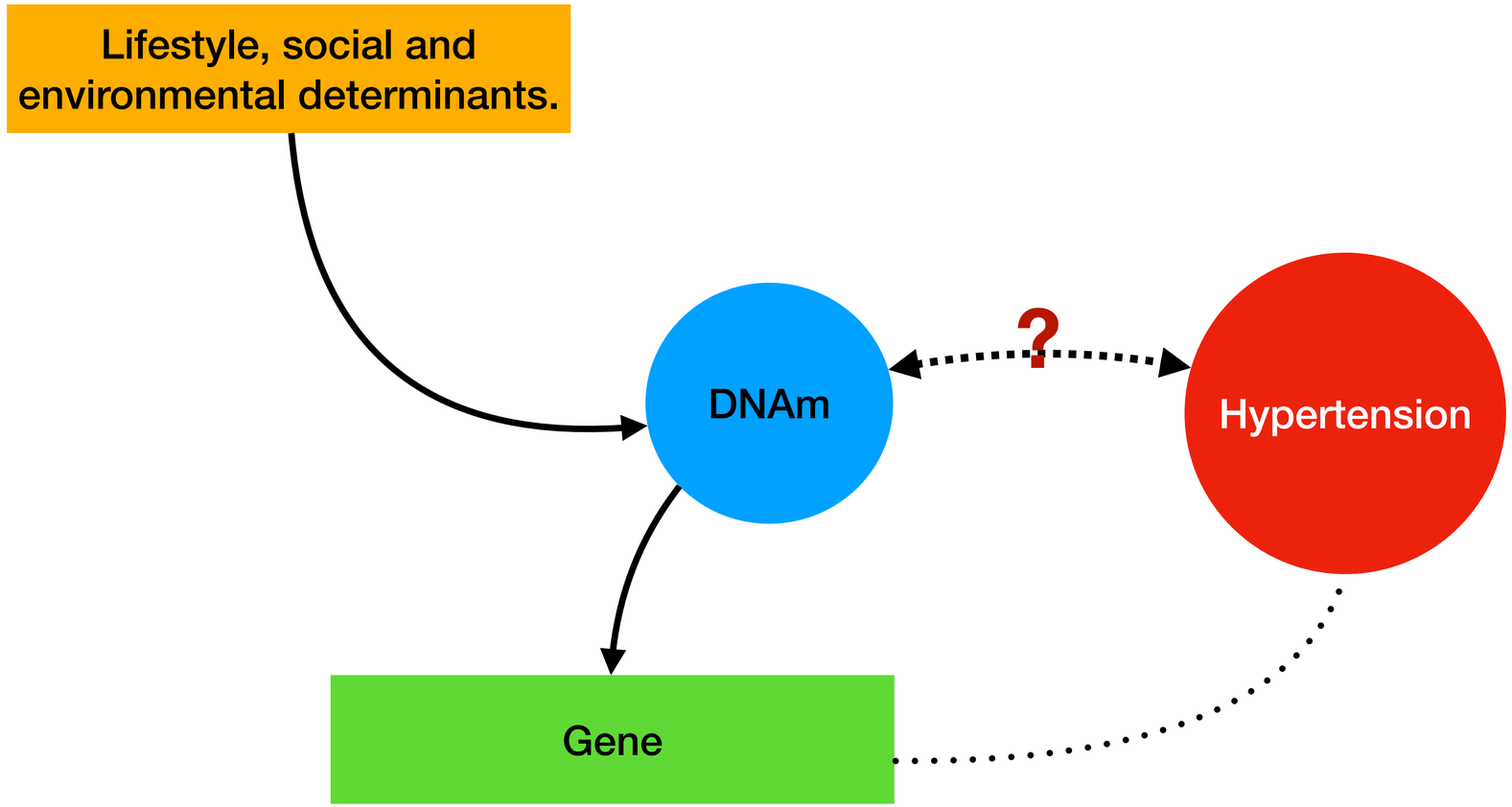
DNAm and cardiovascular diseases



DNAm and cardiovascular diseases



DNAm and cardiovascular diseases



Epigenetics and blood pressure

Leader or follower?

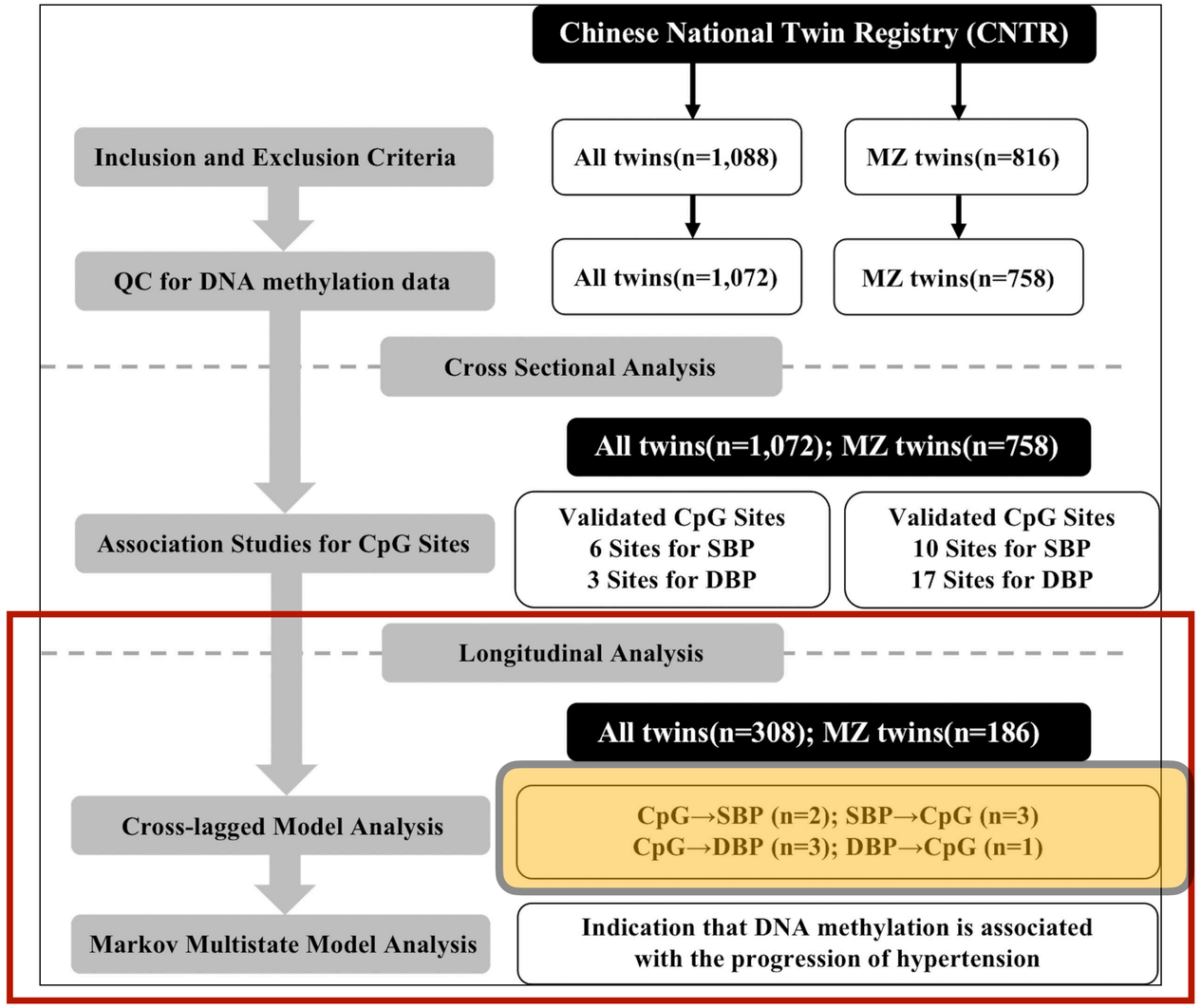
- Conventional knowledge: DNAm sites are associated with BP.
- Very few studies establish a direction or ordering between DNAm and BP.

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

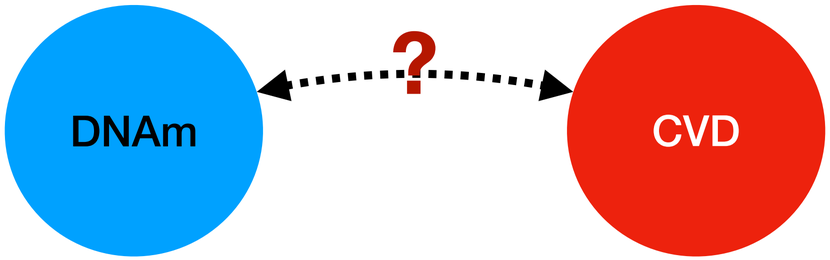
Epigenetics and blood pressure

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.



Does DNAm drive BP or vice-versa?



What data?

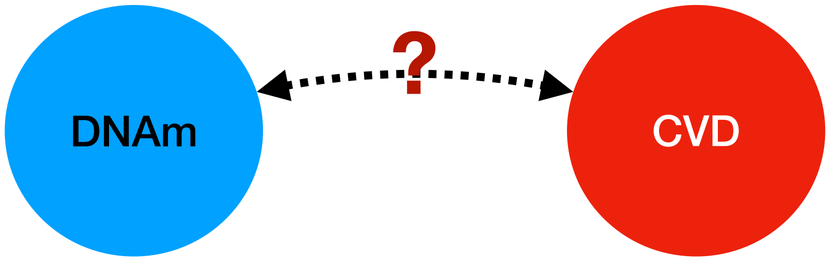
Which genes?

Biological considerations?

Does DNAm drive BP or vice-versa?

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

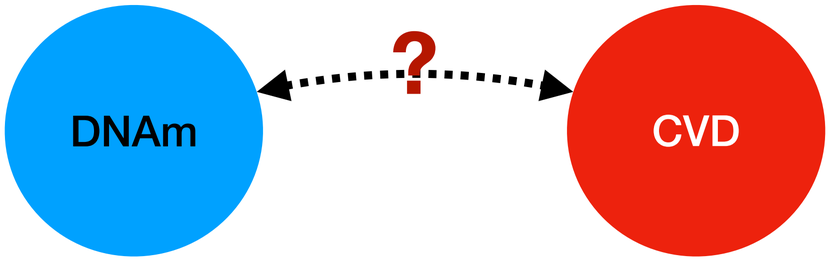
Perng, W., Tang, L., Song, P. X., Tellez-Rojo, M. M., Cantoral, A., & Peterson, K. E. (2019). Metabolomic profiles and development of metabolic risk during the pubertal transition: a prospective study in the ELEMENT Project. Pediatric research, 85(3), 262-268.



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.



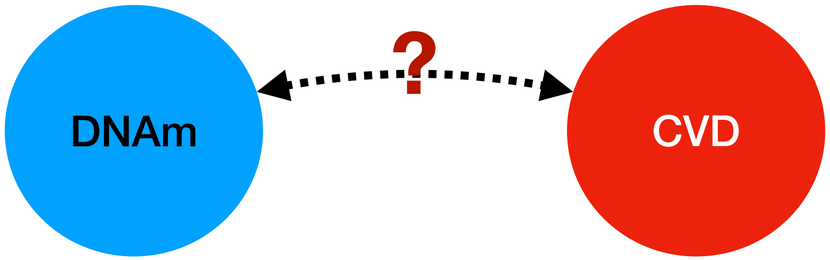
Candidate gene analysis: only examine DNAm for genes that are known to be associated with blood pressure.

ARHGAP42, ATP2B1, FGF5, HSD11B2, KCNK3, PRDM8

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

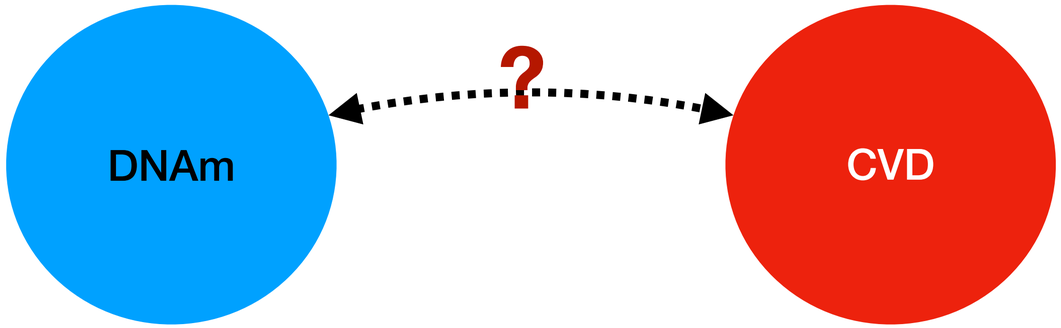
Biological considerations: n = 522 children aged 10 - 18 years, stratified by sex.

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.

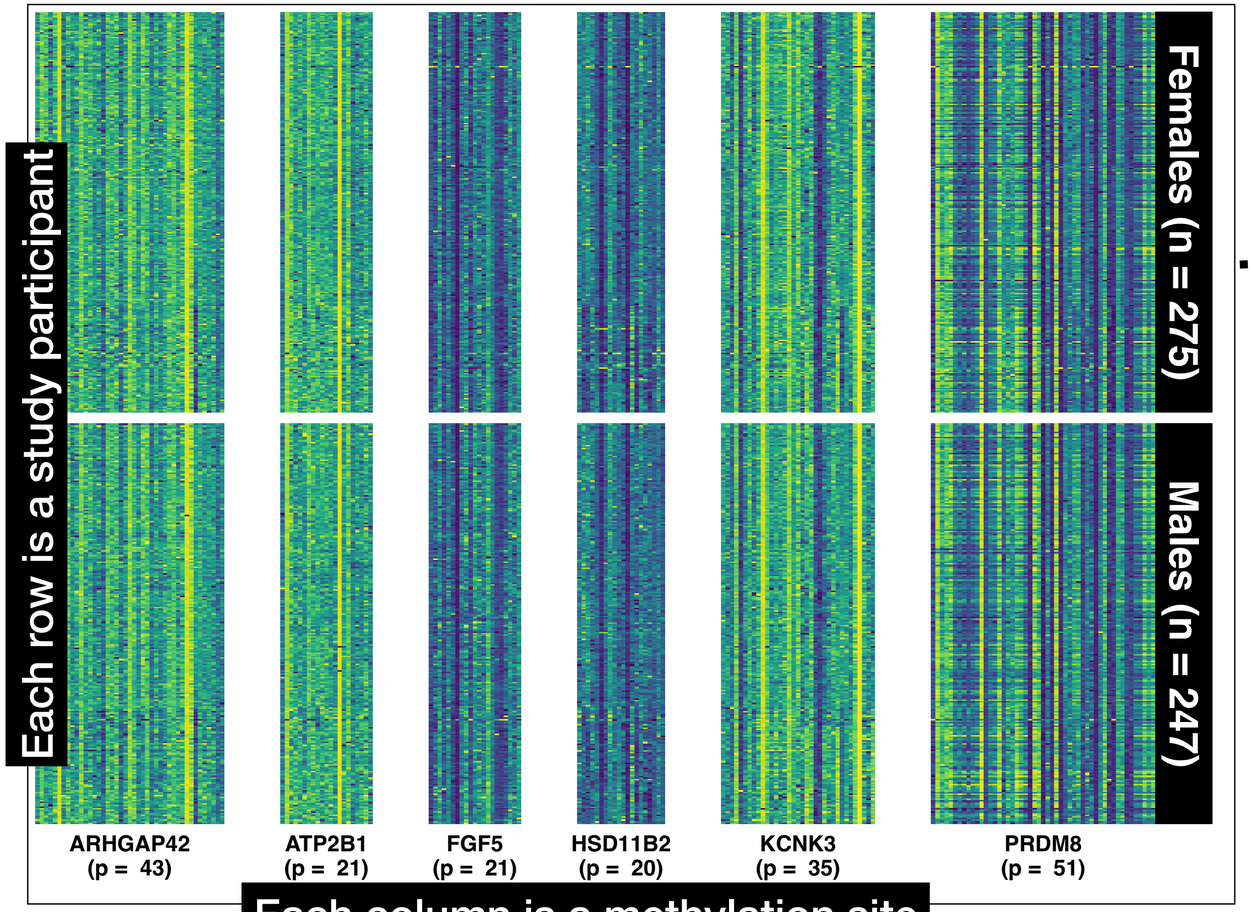


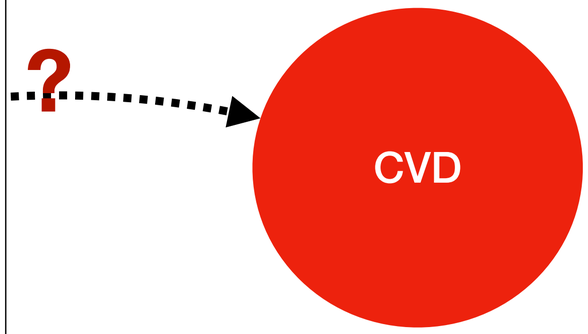
Does DNAm drive BP or vice-versa?

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

$0 \leq \beta \leq 1$

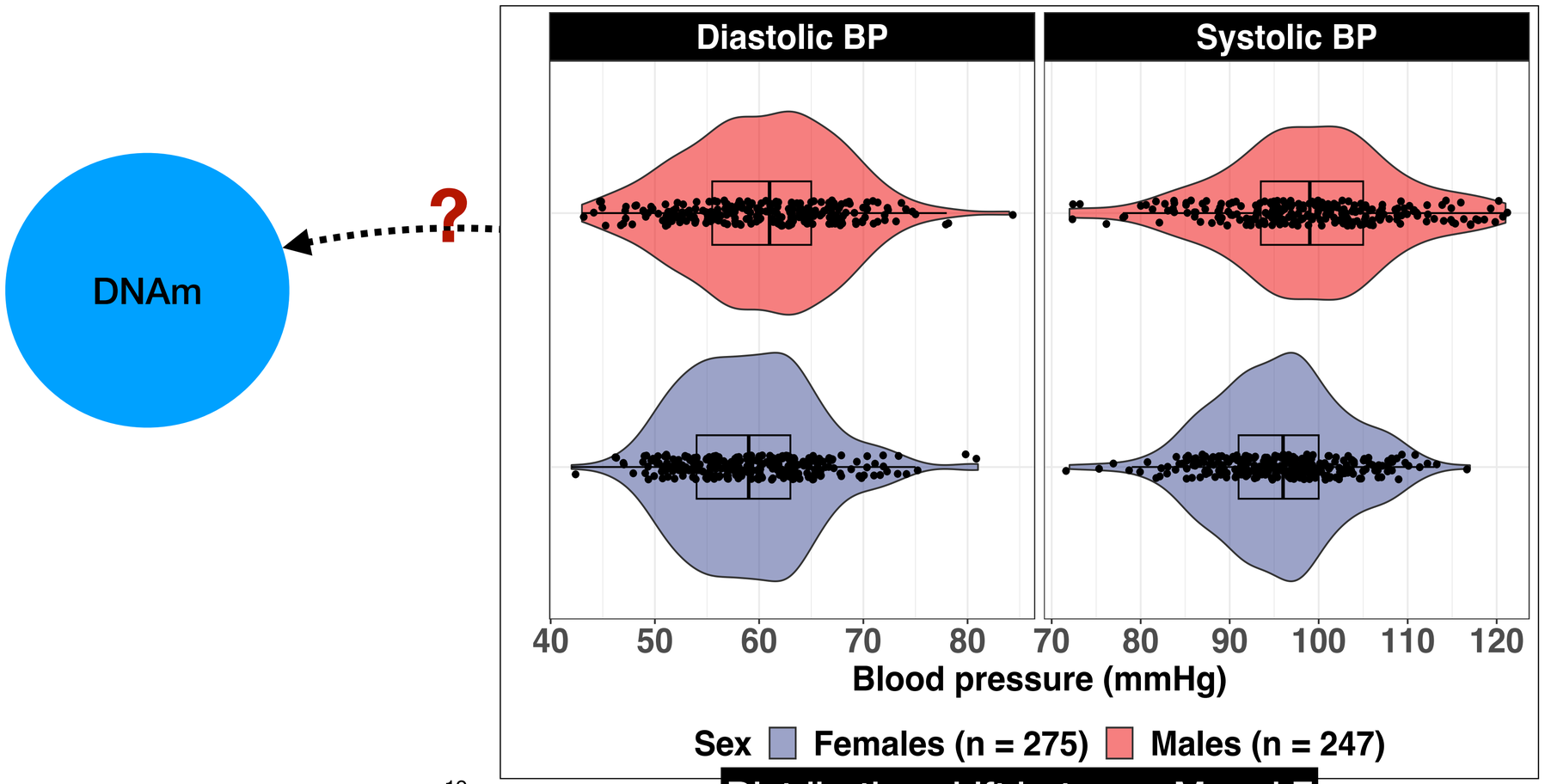
Higher β $\iff$ More methylation

Does DNAm drive BP or vice-versa?

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

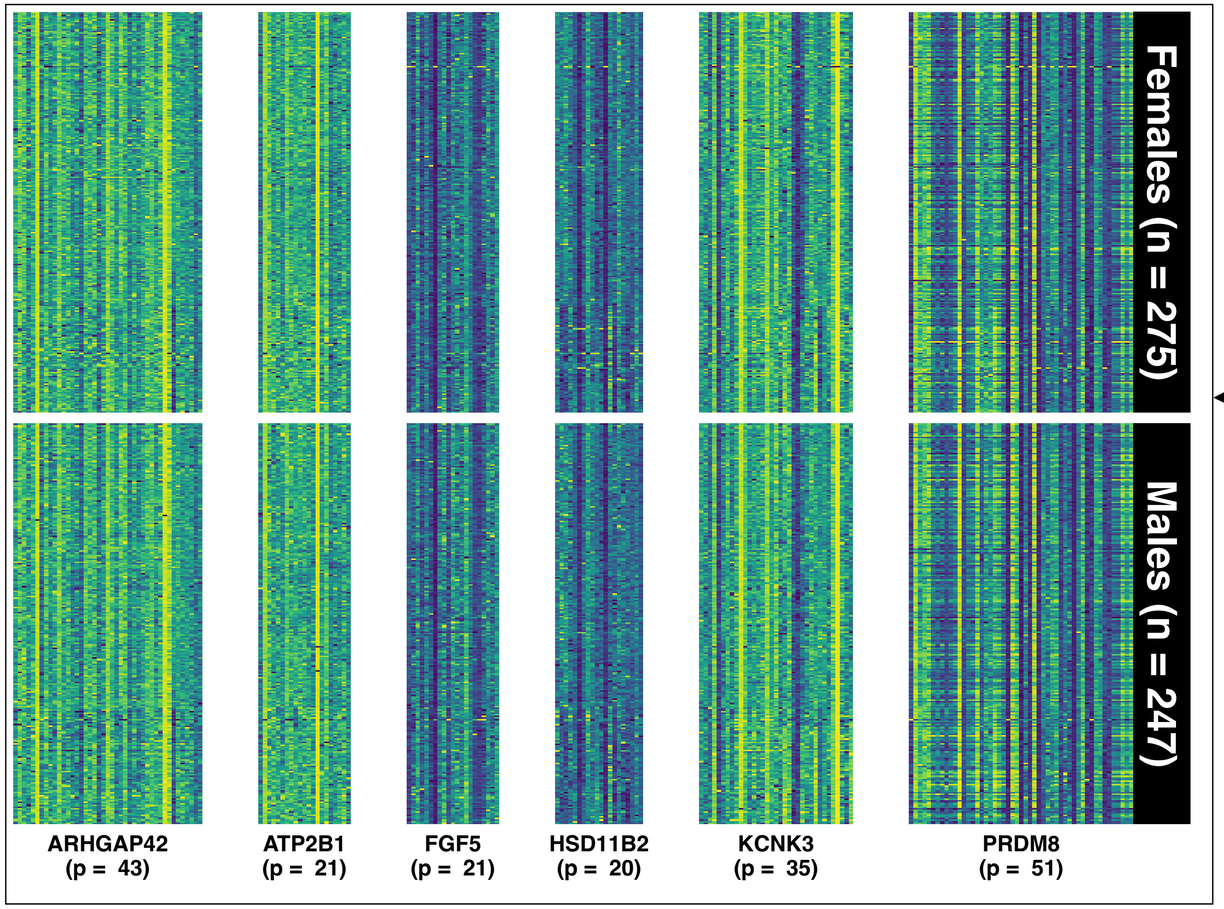
Candidate gene analysis: ARHGAP42, ATP2B1, FGF5, HSD11B2, KCNK3, PRDM8

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

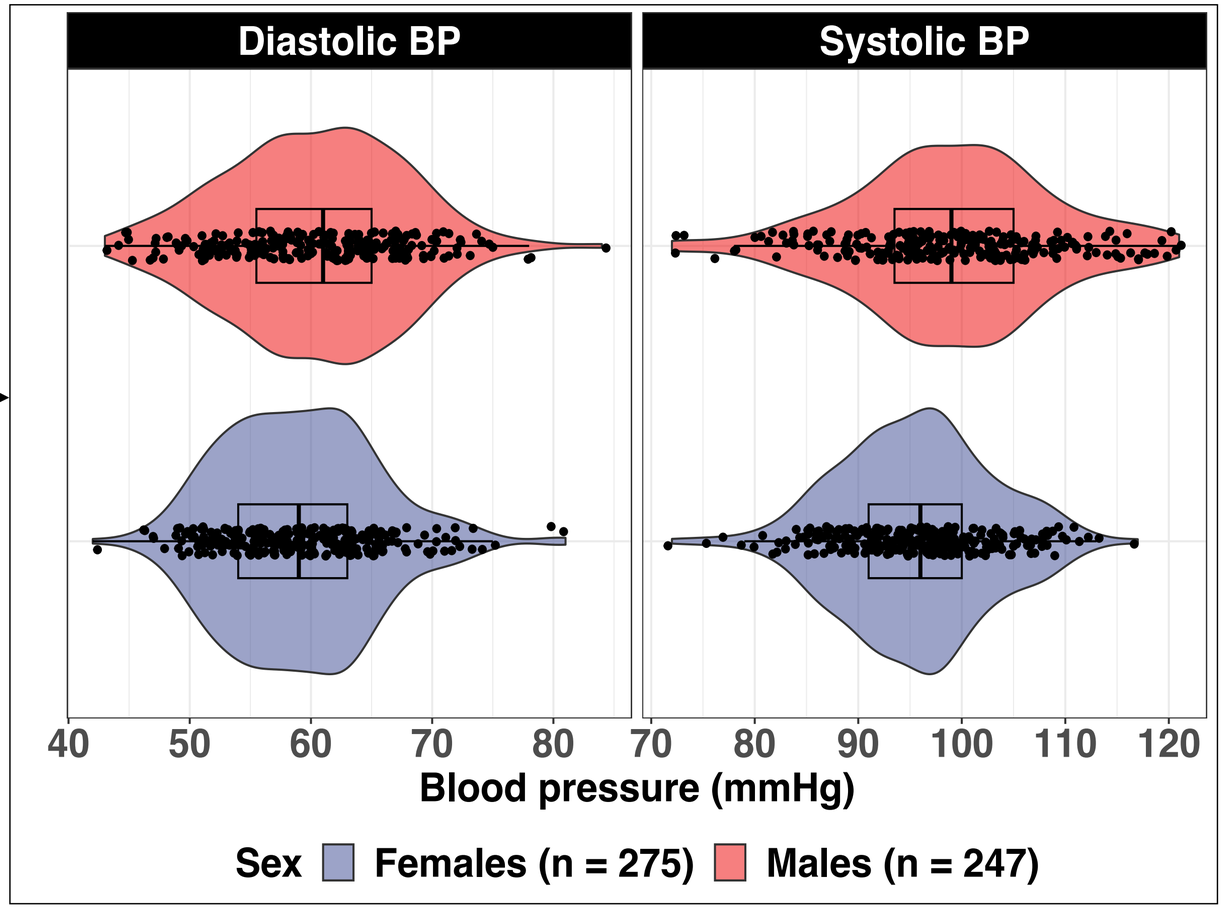


Distribution shift between M and F

Does DNAm drive BP or vice-versa?



?



Methods

Shannon’s information theory

Shannon's information theory

Shannon's information theory

Association and direction through distributions

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

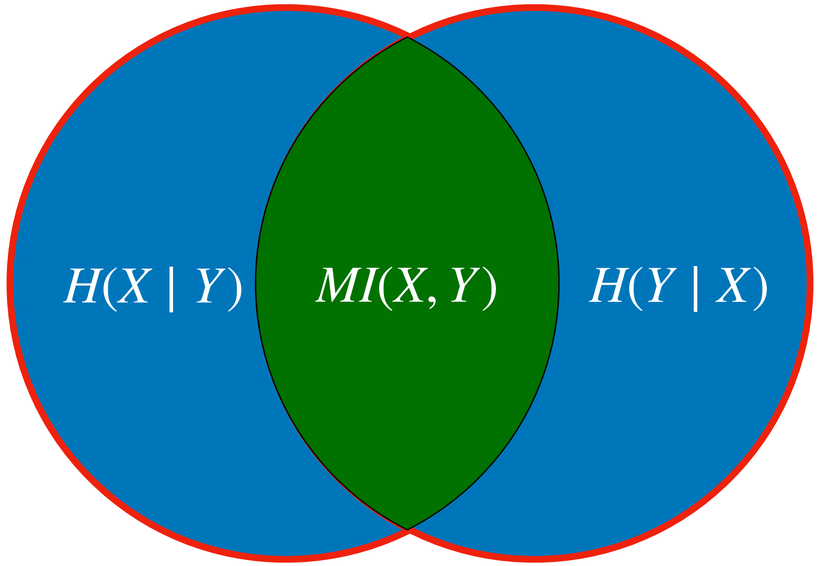
$$H(X \mid Y) + H(Y) = H(X, Y)$$

Entropy decomposition equation

Entropy decomposition equation

Attempt to study association and directionality

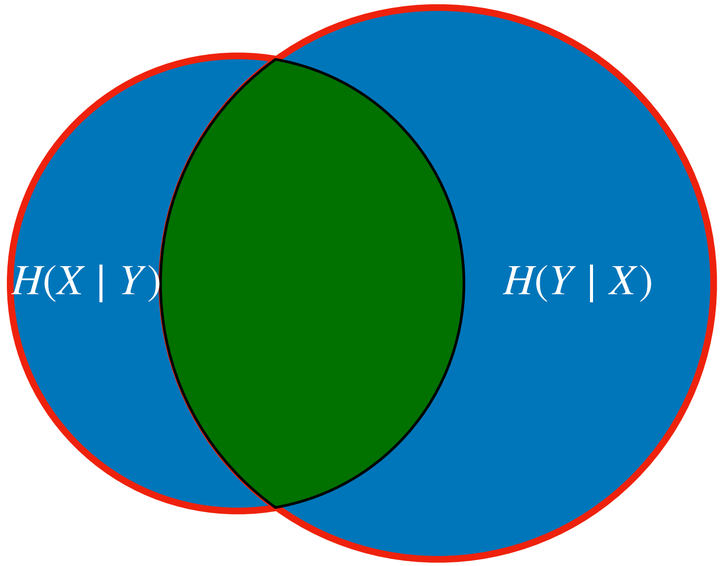
$$H(X, Y) = H(X \mid Y) + H(Y \mid X) + MI(X, Y)$$

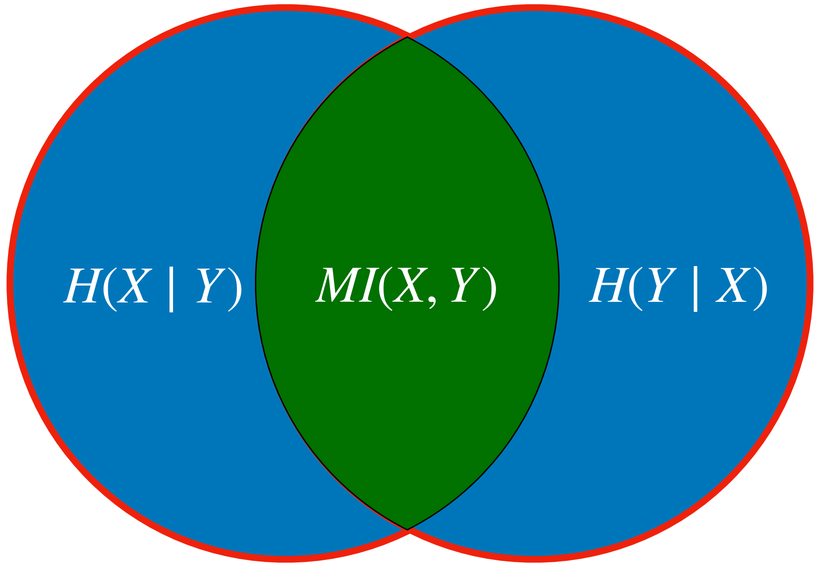


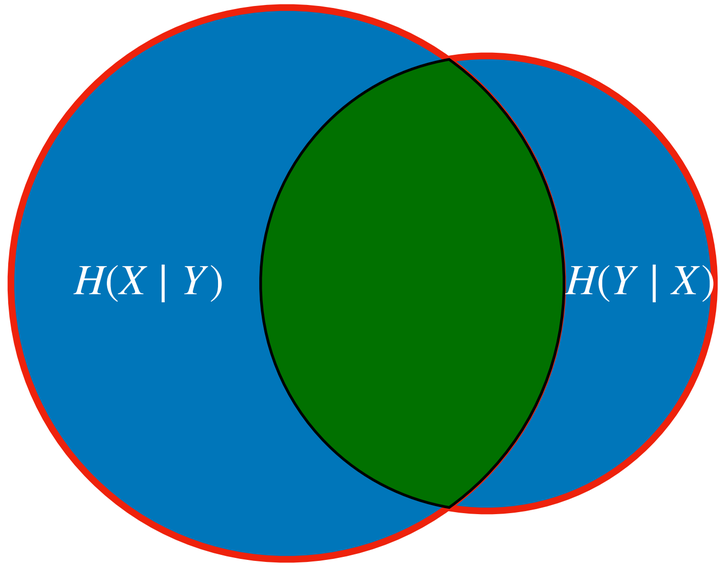
Entropy decomposition equation

Attempt to study association and directionality

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







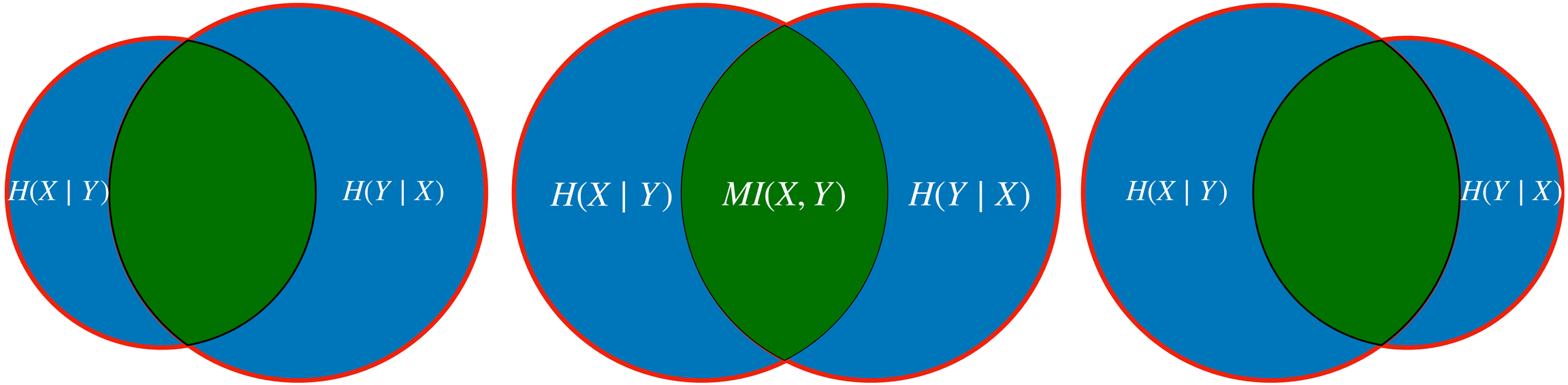
$$H(X \mid Y) < H(Y \mid X)$$

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

Entropy decomposition equation

Attempt to study association and directionality

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

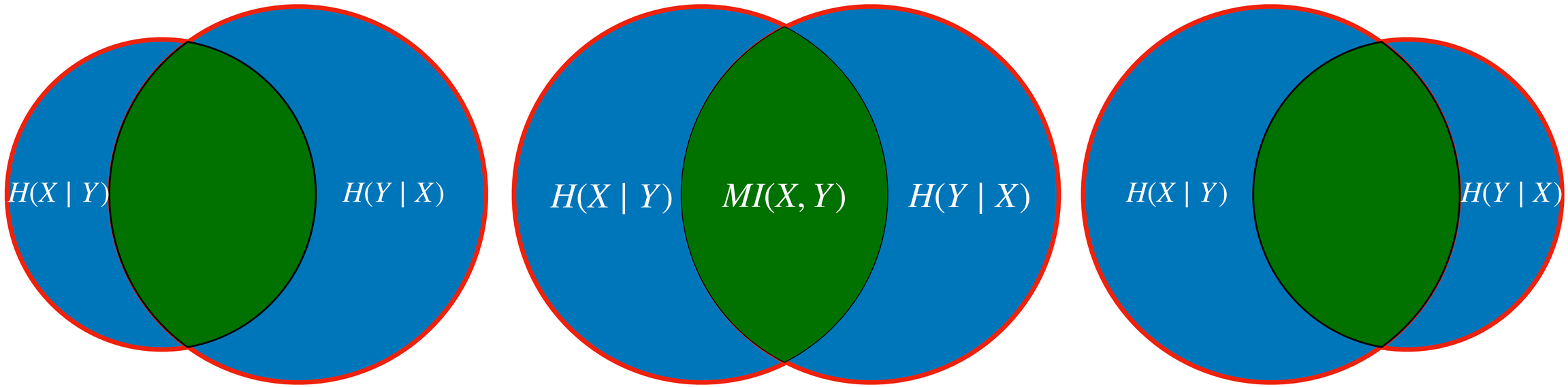


Plan 2: Use $H(X \mid Y)$ and $H(Y \mid X)$ to capture asymmetry/directionality.

Entropy decomposition equation

Attempt to study association and directionality

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



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 ⬆️⬇️**

Benefits and hurdles of MI

Need a fast, scalable and accurate estimator

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

Benefits and hurdles of MI

Need a fast, scalable and accurate estimator

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

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

Benefits and hurdles of MI

Need a fast, scalable and accurate estimator

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

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

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)

fastMI

Scalable and accurate estimation of MI

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

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 theorem:

$$f_{XY}(x,y) = c_{XY}\left(F_X(x), F_Y(y)\right) f_X(x) f_Y(y)$$

Rank transformation:

$$U_X = F_X(X), U_Y = F_Y(Y)$$

Mutual information is copula entropy:

$$MI(X,Y) = E_{U_X U_Y}\left[\log\left(c_{XY}(U_X, U_Y)\right)\right]$$

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 theorem:

$$f_{XY}(x, y) = c_{XY}\left(F_X(x), F_Y(y)\right) f_X(x) f_Y(y)$$

Rank transformation:

$$U_X = F_X(X), U_Y = F_Y(Y)$$

Mutual information is copula entropy:

$$MI(X, Y) = E_{U_X U_Y}\left[\log\left(c_{XY}(U_X, U_Y)\right)\right]$$

No longer have to estimate three densities!

Depends only on ranks of data!

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.
    - $MI = E[\log(c_{XY})]$

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.
    - $MI = E[\log(c_{XY})]$

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

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

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

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.

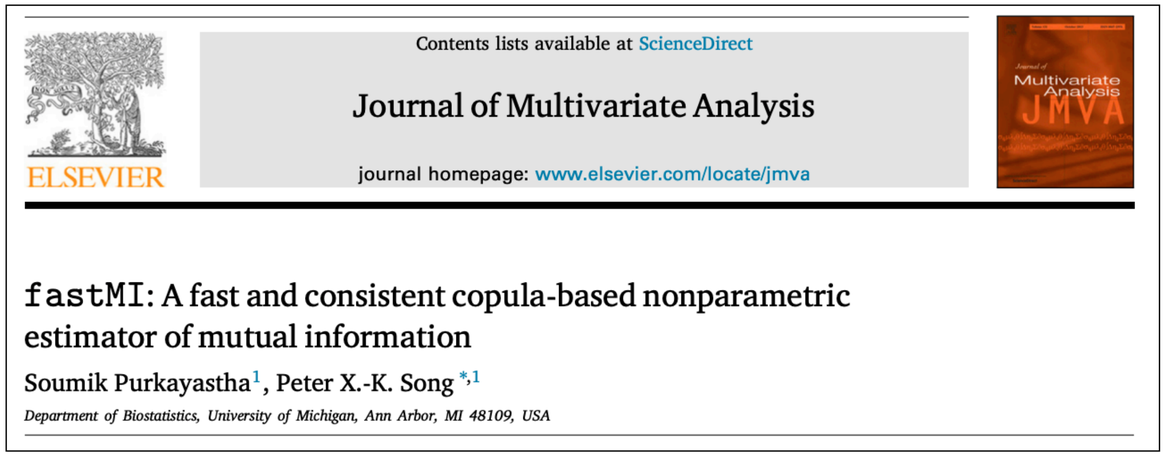
fastMI

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.

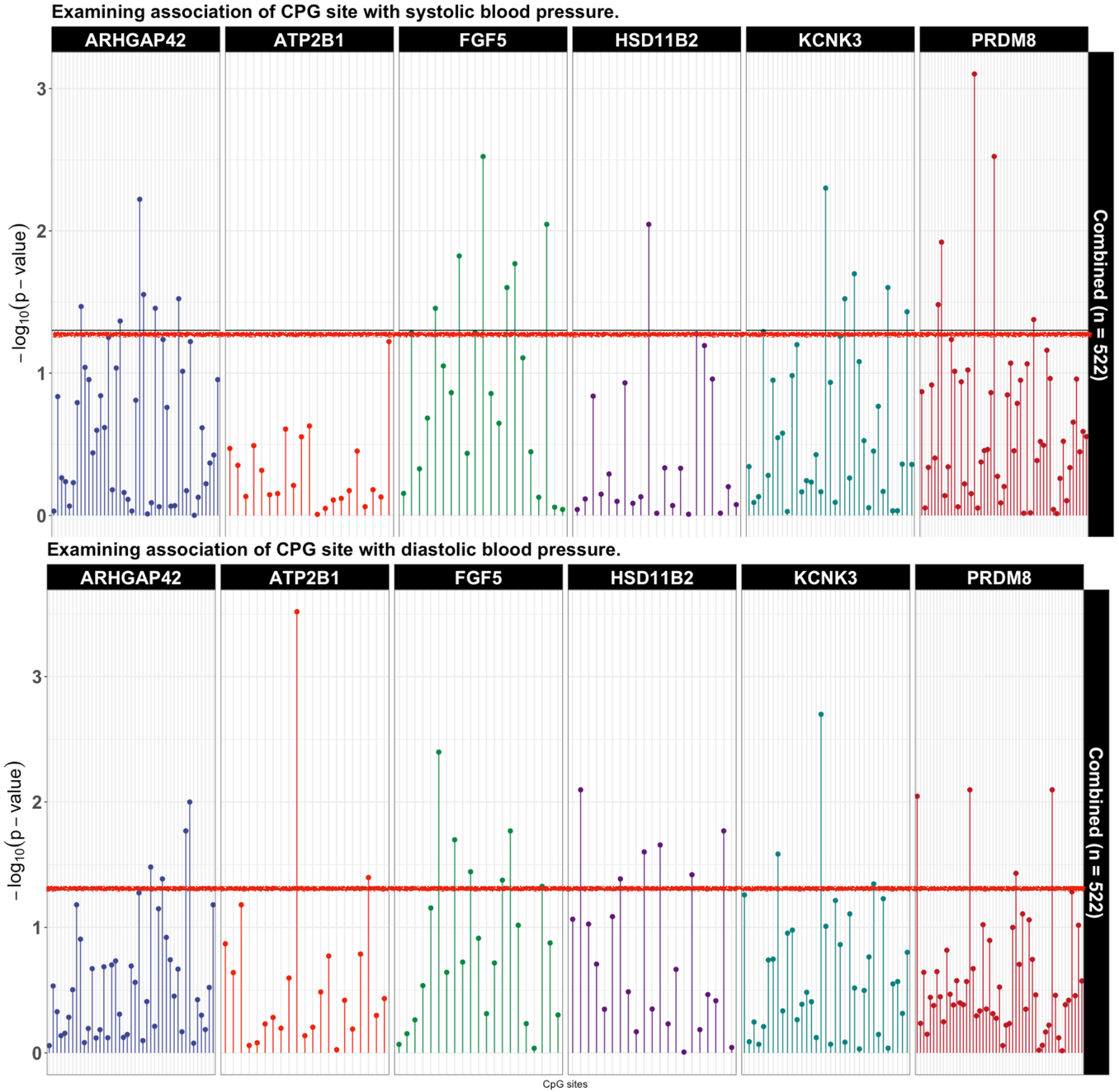
	Sample size (n)		
	1000	2500	5000
Empirical copula-based MI	4.360 (0.356)	5.368 (0.308)	64.040 (0.254)
Jackknifed MI	3.150 (0.107)	18.446 (0.116)	62.454 (4.601)
fastMI	**1.199 (0.135)**	**2.964 (0.181)**	**5.952 (0.125)**

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

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

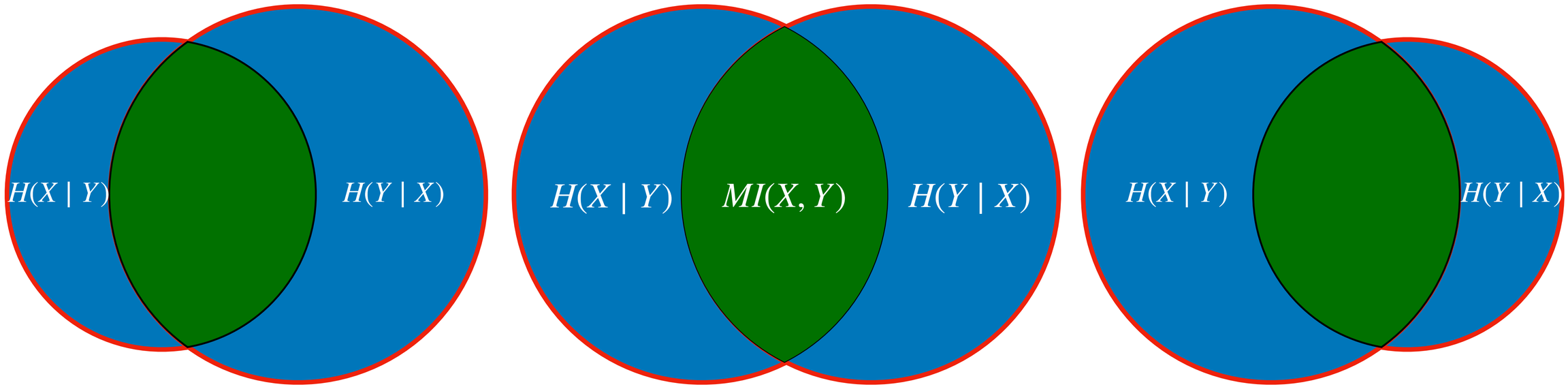
Still no sense of directionality

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.

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** $g(\cdot)$

$$Y = g(X)$$

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:

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

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"

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?

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"

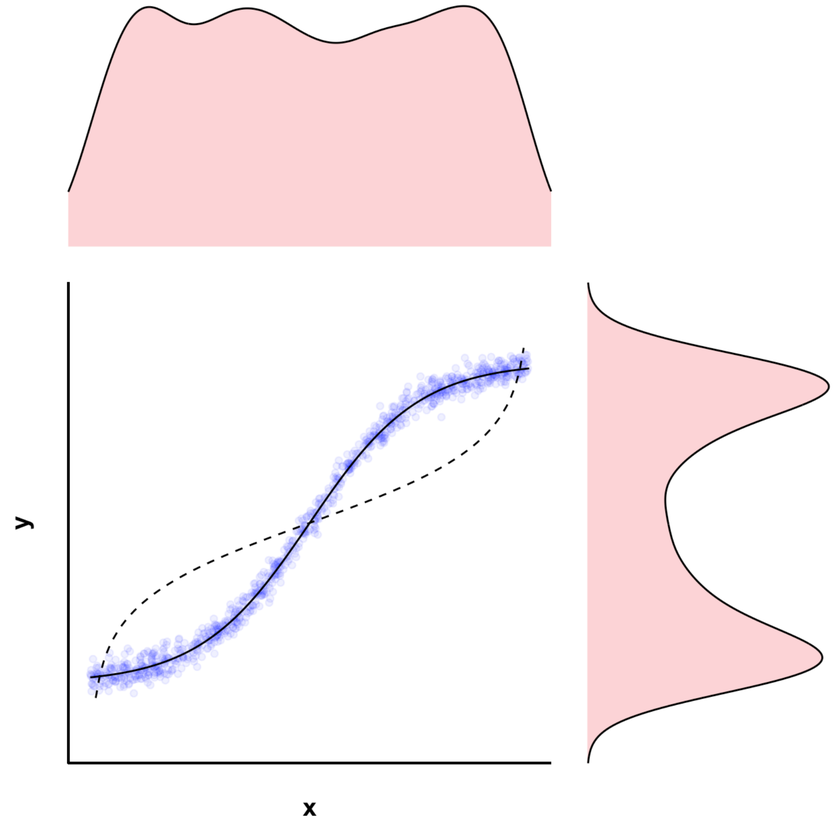
Impose orthogonality condition on $g(\cdot)$ and f_X

Identifiability condition for GEMs

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

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

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



Identifiability condition for GEMs

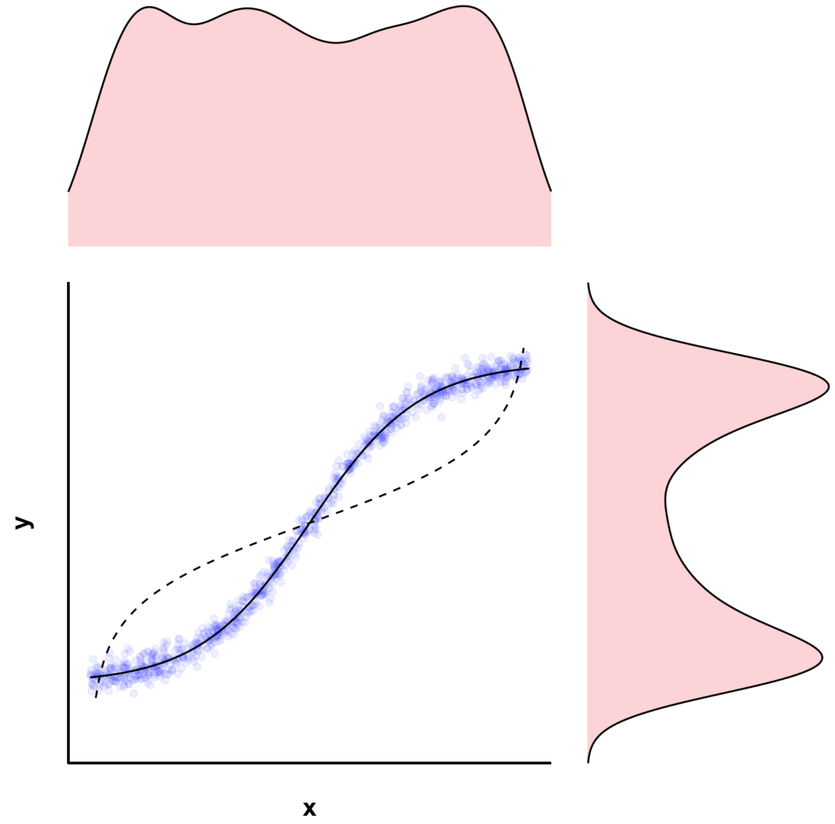
Orthogonality of f_X and g induces discrepancy in f_Y and g^{-1}

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

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

$U(0,1)$ is maximally random

Flat f_X and bijective g yields spikes in f_Y

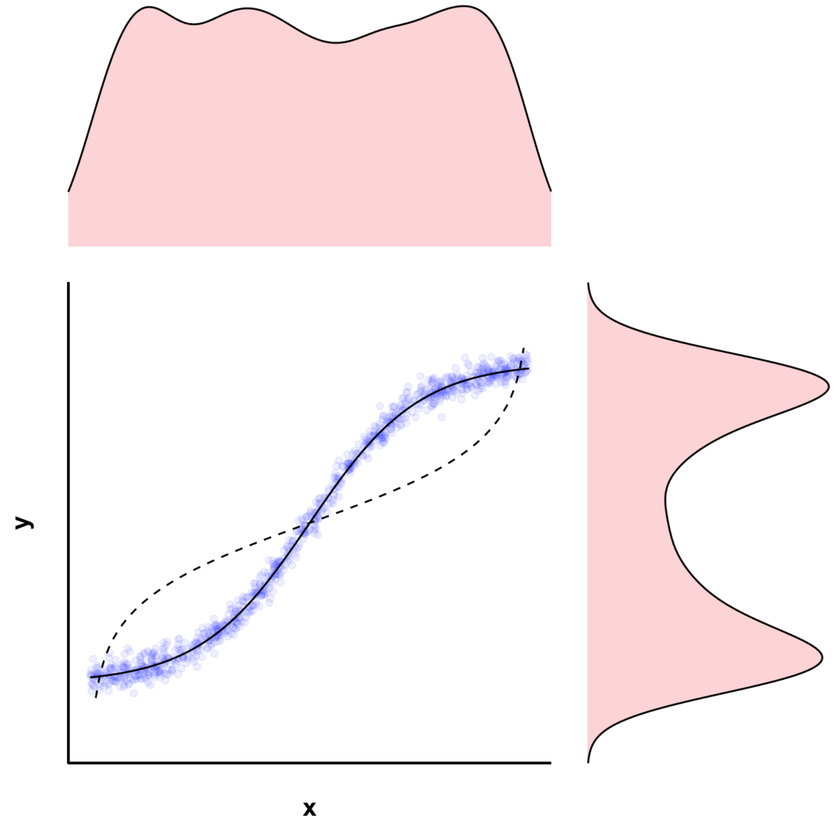


Identifiability condition for GEMs

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

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

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



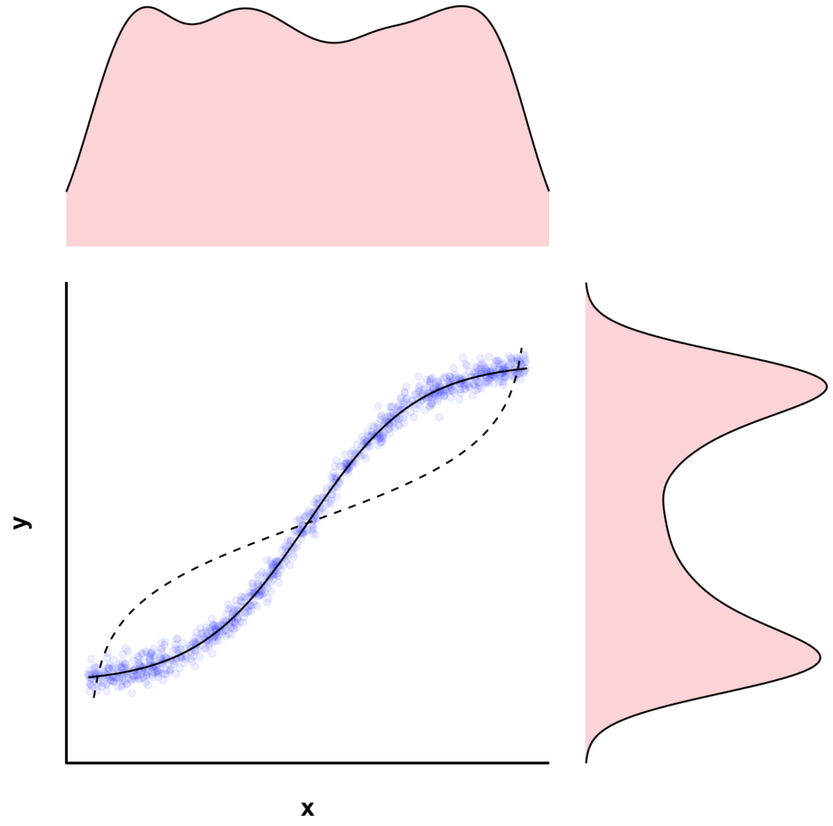
Identifiability condition for GEMs

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.

Asymmetry in GEMs using entropy

Approve or disprove $X \to Y$

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

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

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

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.

Asymmetry in GEMs using entropy

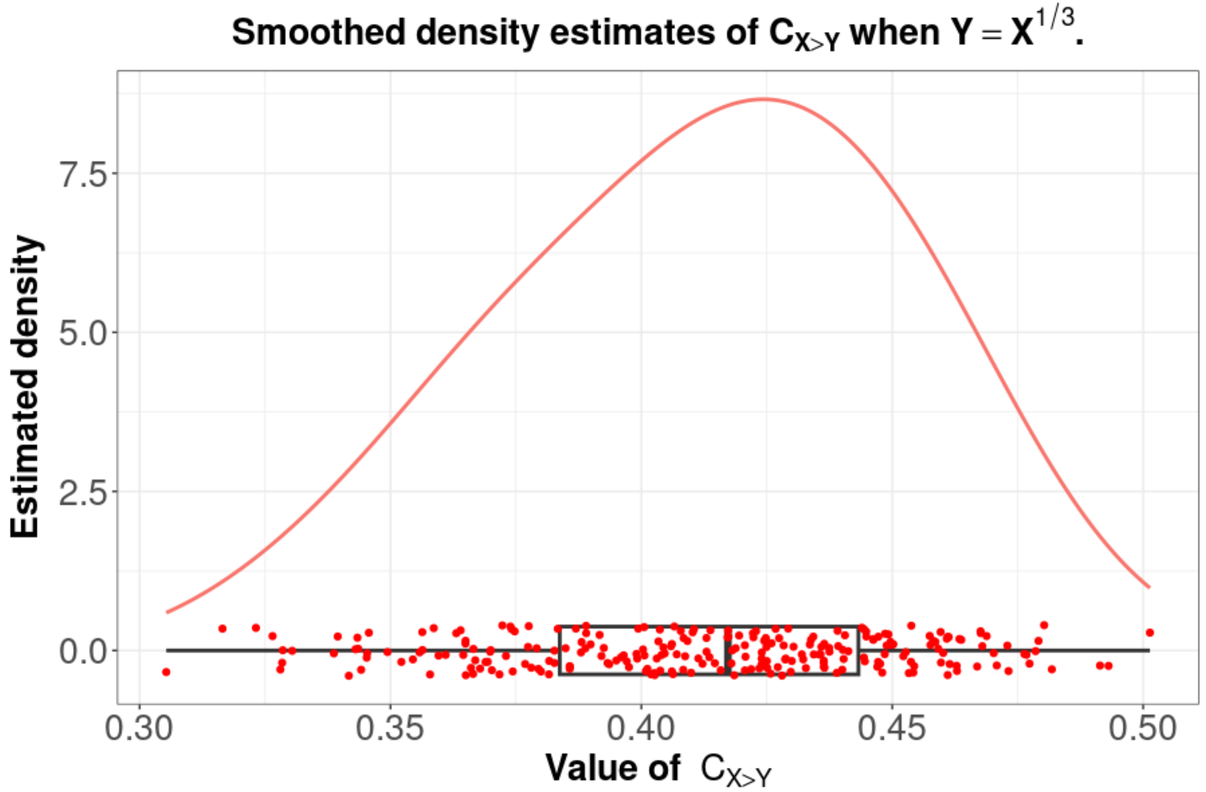
Approve or disprove $X \to Y$

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

Asymmetry in GEMs using entropy

Approve or disprove $X \to Y$

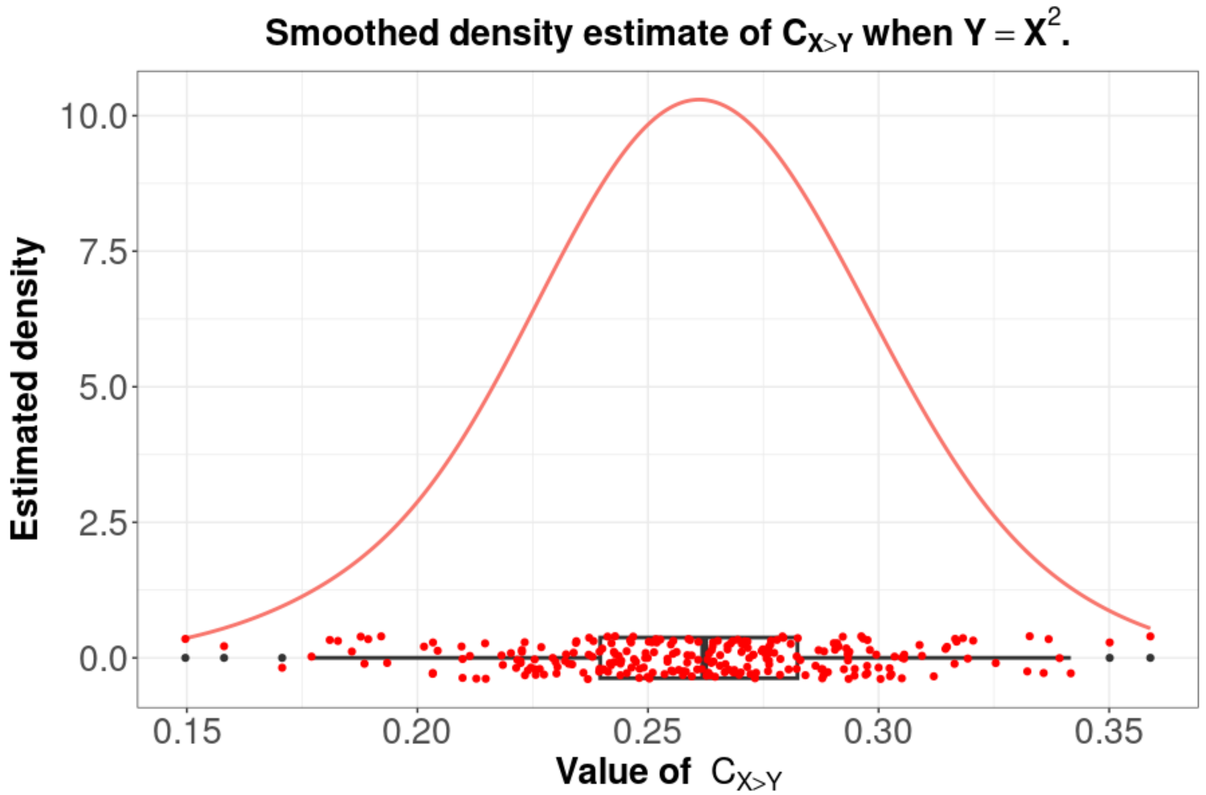
- Hypothesis: $X \to Y$.
- GEM-induced direction $Y = g(X)$.
- Experiment:
  - $X \sim U(0,1)$.
  - Generate $Y = g(X)$.
  - **Simulated behavior of $C_{X>Y}$?**



Asymmetry in GEMs using entropy

Approve or disprove $X \to Y$

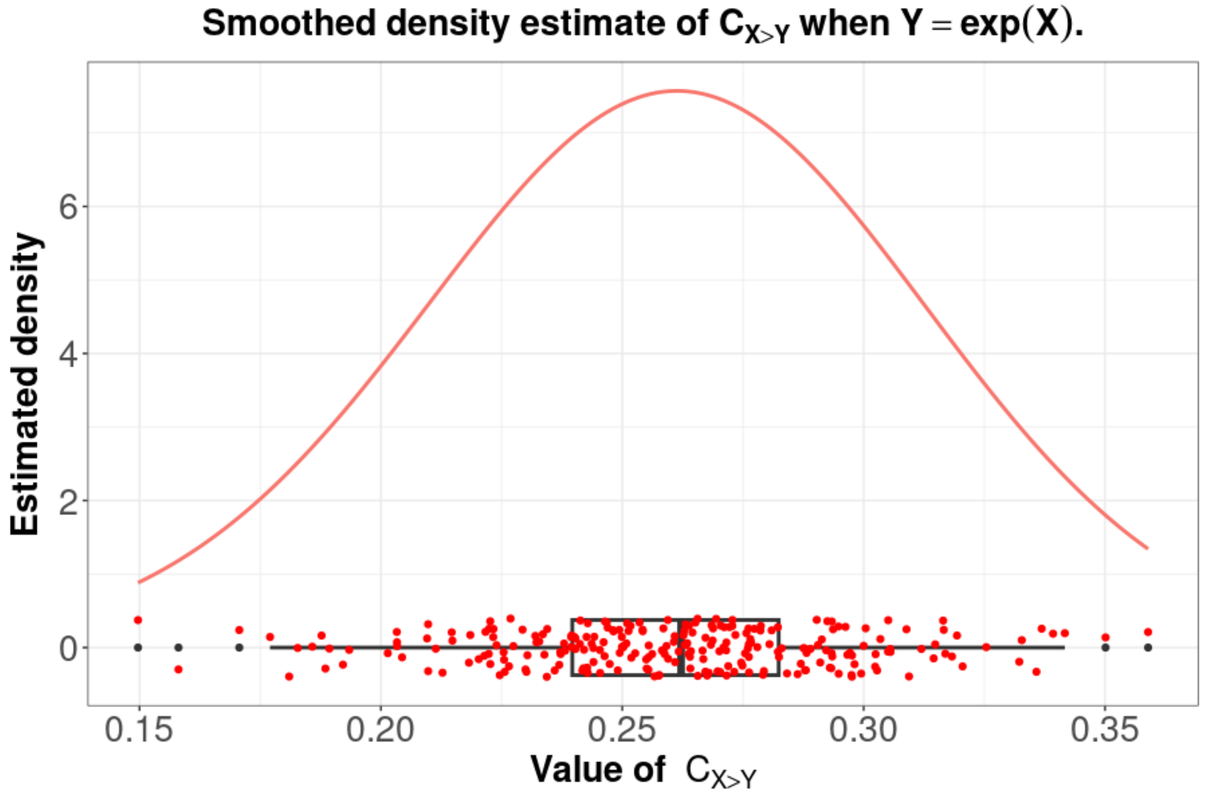
- Hypothesis: $X \to Y$.
- GEM-induced direction $Y = g(X)$.
- Experiment:
  - $X \sim U(0,1)$.
  - Generate $Y = g(X)$.
  - **Simulated behavior of $C_{X>Y}$?**



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 $C_{X>Y}$?**



Asymmetry coefficient $C_{X \succ Y}$

Advantages and challenges

Strong asymmetry:

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

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

Asymmetry coefficient $C_{X \succ Y}$

Advantages and challenges

Strong asymmetry:

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

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

- $C_{X \succ Y} = H(X) - H(Y) = H(X \mid Y) - H(Y \mid X)$
- Comparison of $H(X \mid Y) \lesseqgtr H(Y \mid X)$: which is the better predictor?

Data splitting and cross fitting

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

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!

Data splitting and cross fitting

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

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

Split data into two halves.

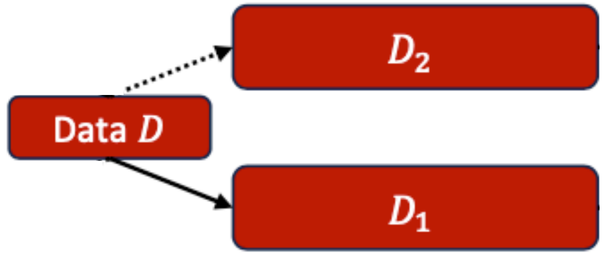
Use one half to estimate densities.

Use the other half to estimate entropies.

Data splitting and cross fitting

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

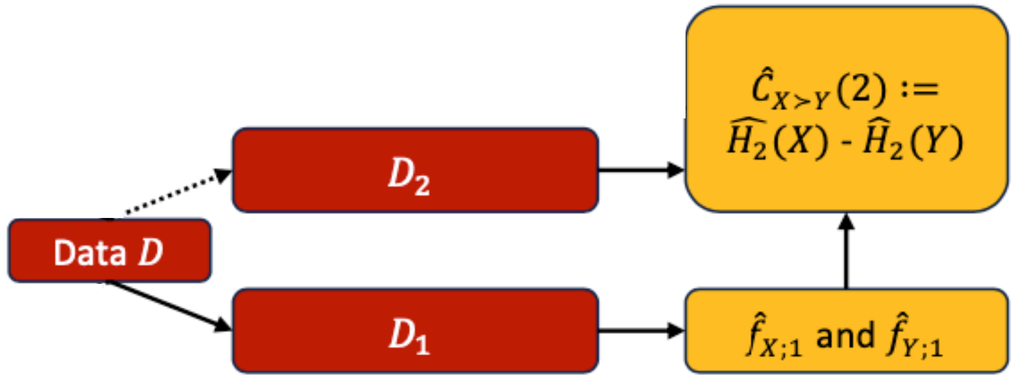
$$D = \{(X_i, Y_i)\}_{i=1}^{n} \cup \{(X_i, Y_i)\}_{i=n+1}^{2n} = D_1 \cup D_2$$



Data splitting and cross fitting

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

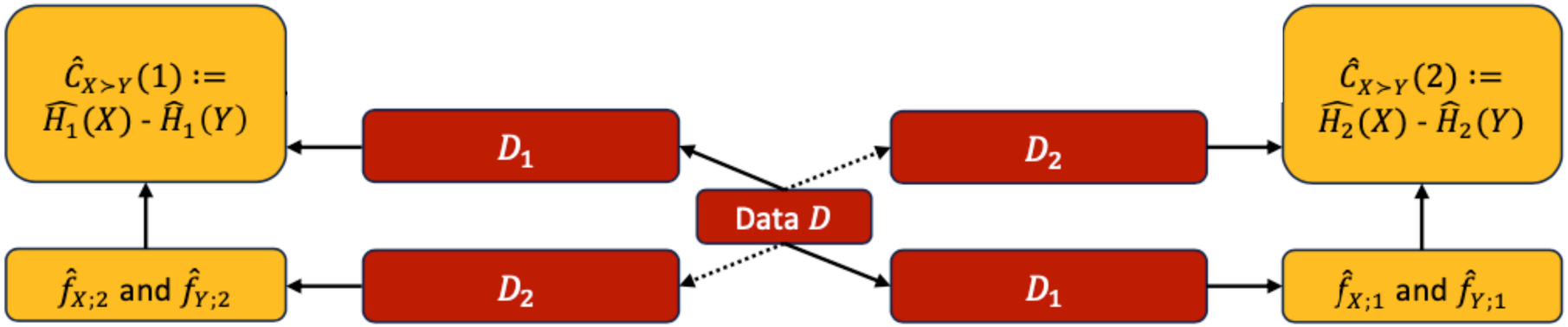
$$D = \{(X_i, Y_i)\}_{i=1}^{n} \cup \{(X_i, Y_i)\}_{i=n+1}^{2n} = D_1 \cup D_2$$



Data splitting and cross fitting

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

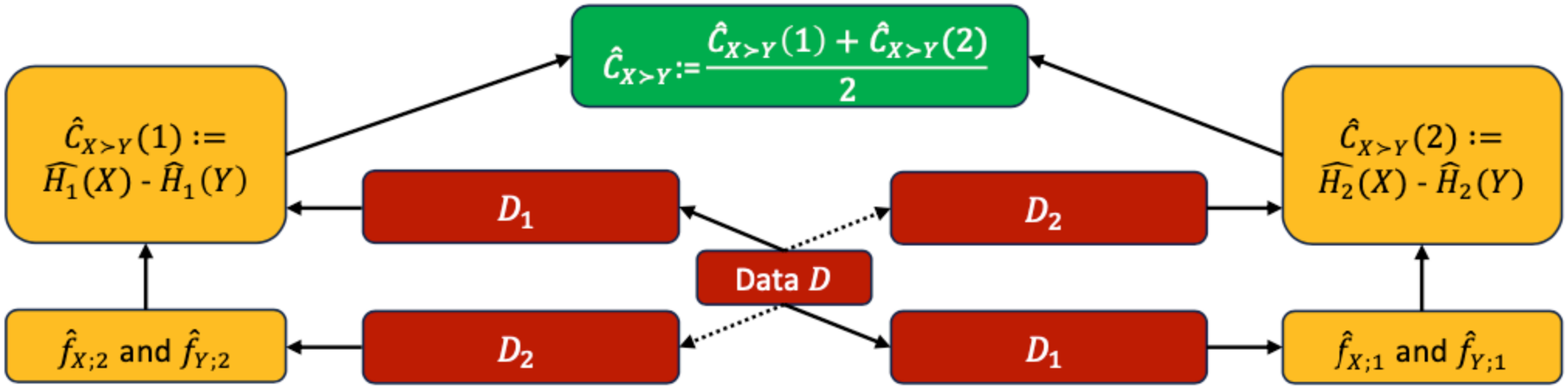
$$D = \{(X_i, Y_i)\}_{i=1}^{n} \cup \{(X_i, Y_i)\}_{i=n+1}^{2n} = D_1 \cup D_2$$



Data splitting and cross fitting

Estimation and inference for $C_{X>Y}$

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



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 $\hat{f}_Y$

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 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}

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

Coefficient can tolerate contaminated outcomes!

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

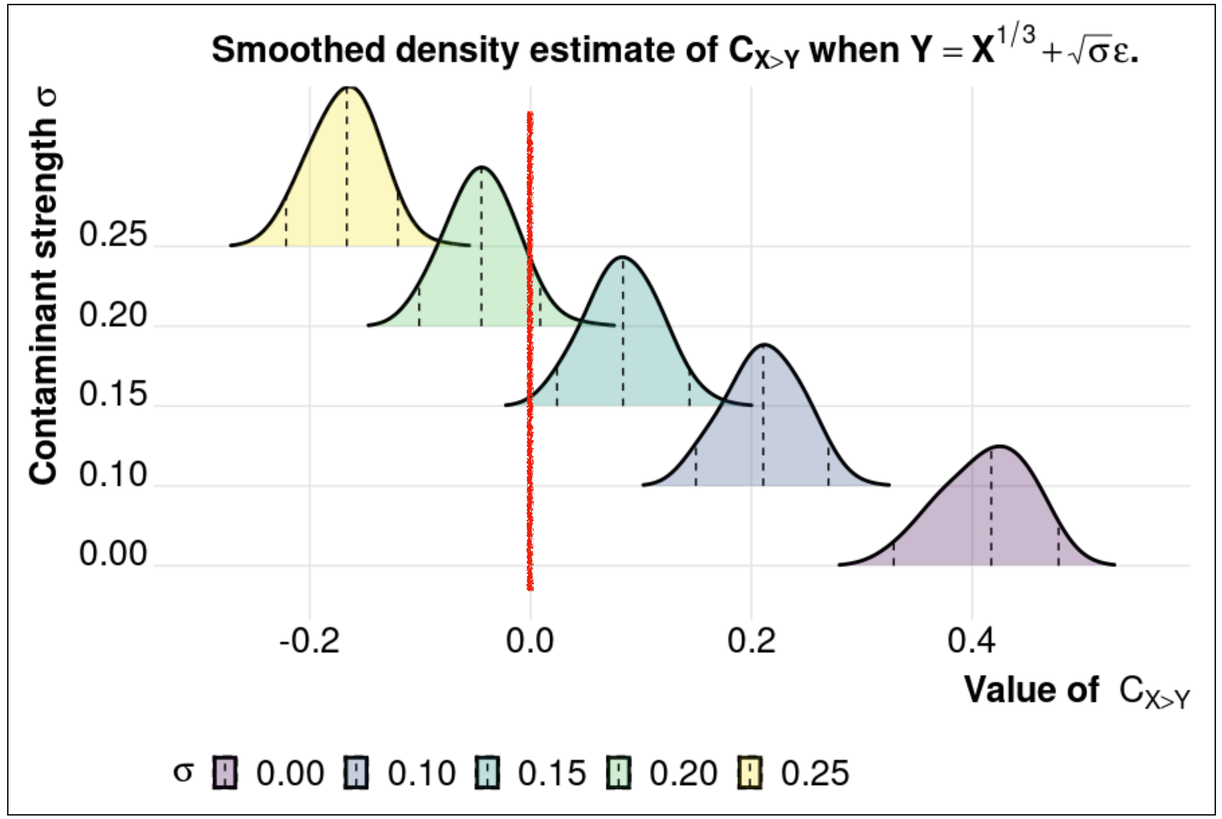
- Noise-perturbed GEM (NPGEM)

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

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

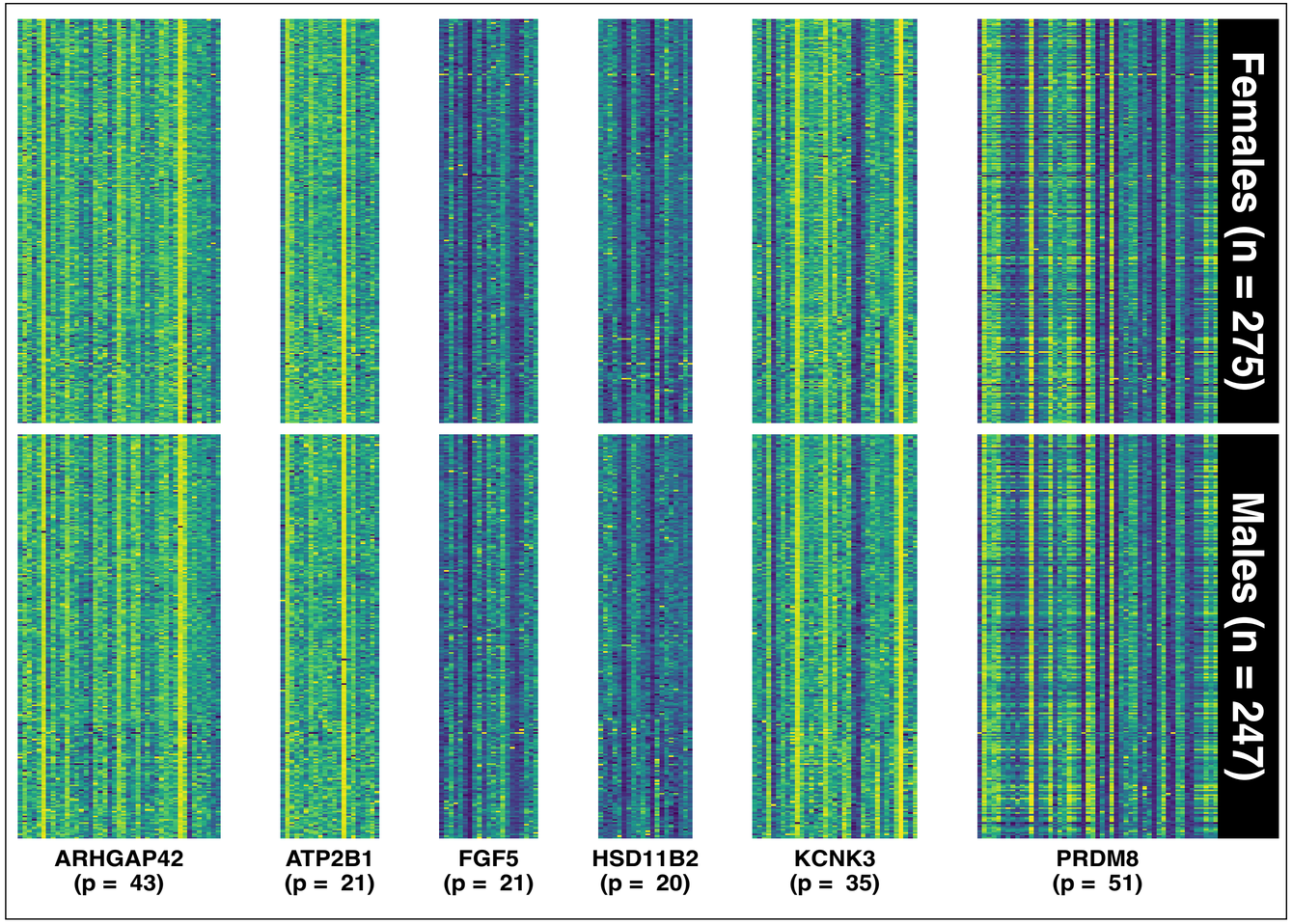
- Establish "critical value" σ_{CRIT}

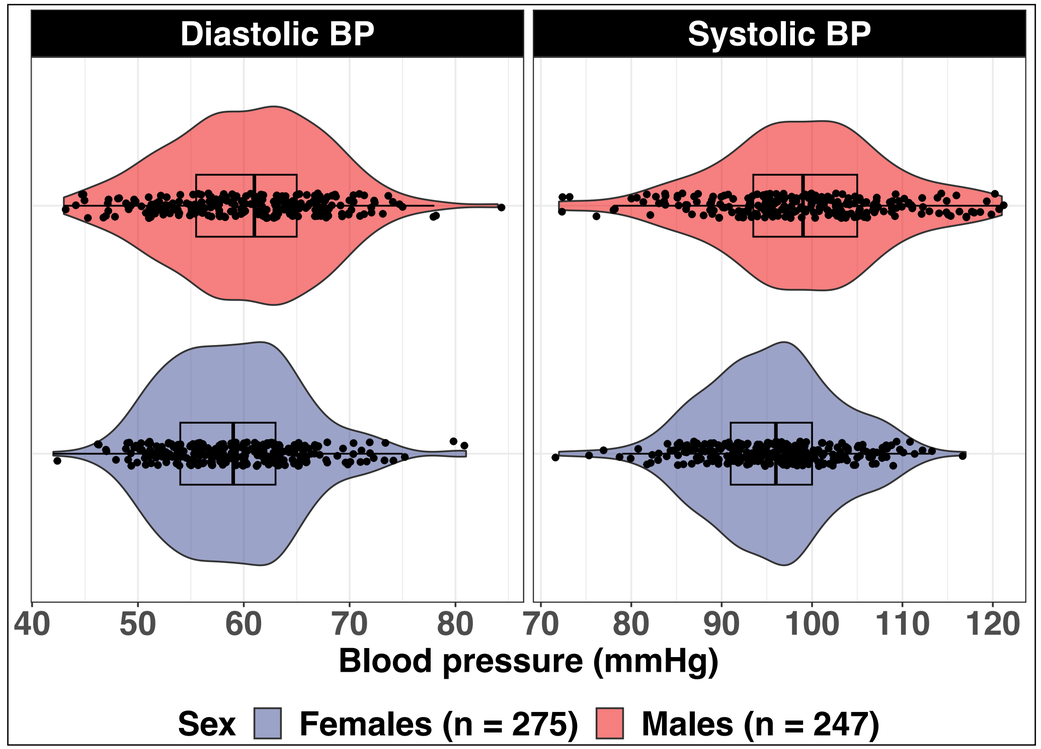
$$\sigma \leq \sigma_{CRIT} = \frac{\exp(2C_{X>Y}) - 1}{I(Y)}.$$



Adjusting for Z

Low-dimensional confounders





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

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

$BP \rightarrow DNAm?$

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

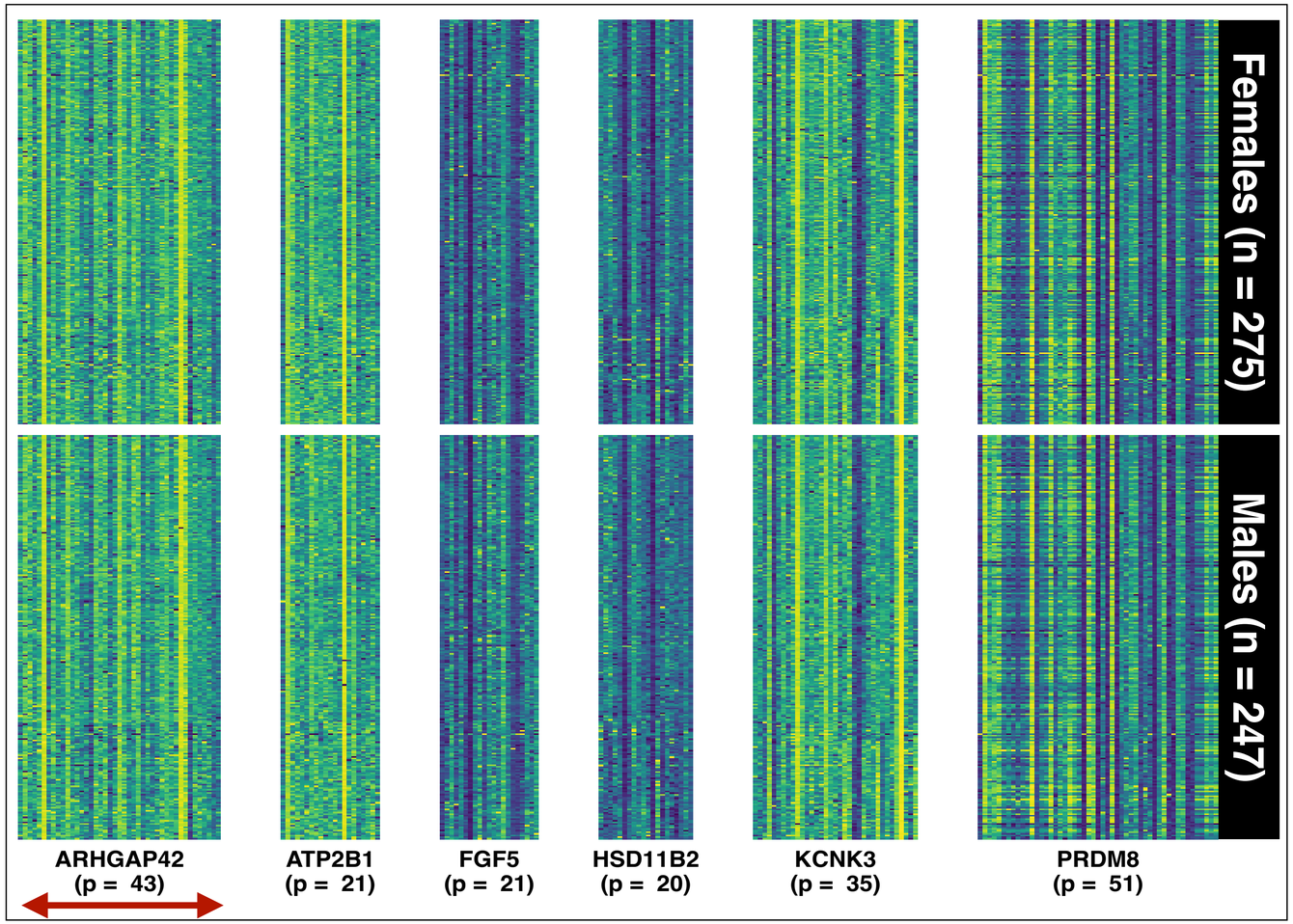
But what about directionality?

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

$BP \rightarrow DNAm?$

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

Stratify by Z: sex of study participant.

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

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

$BP \rightarrow DNAm?$

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

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

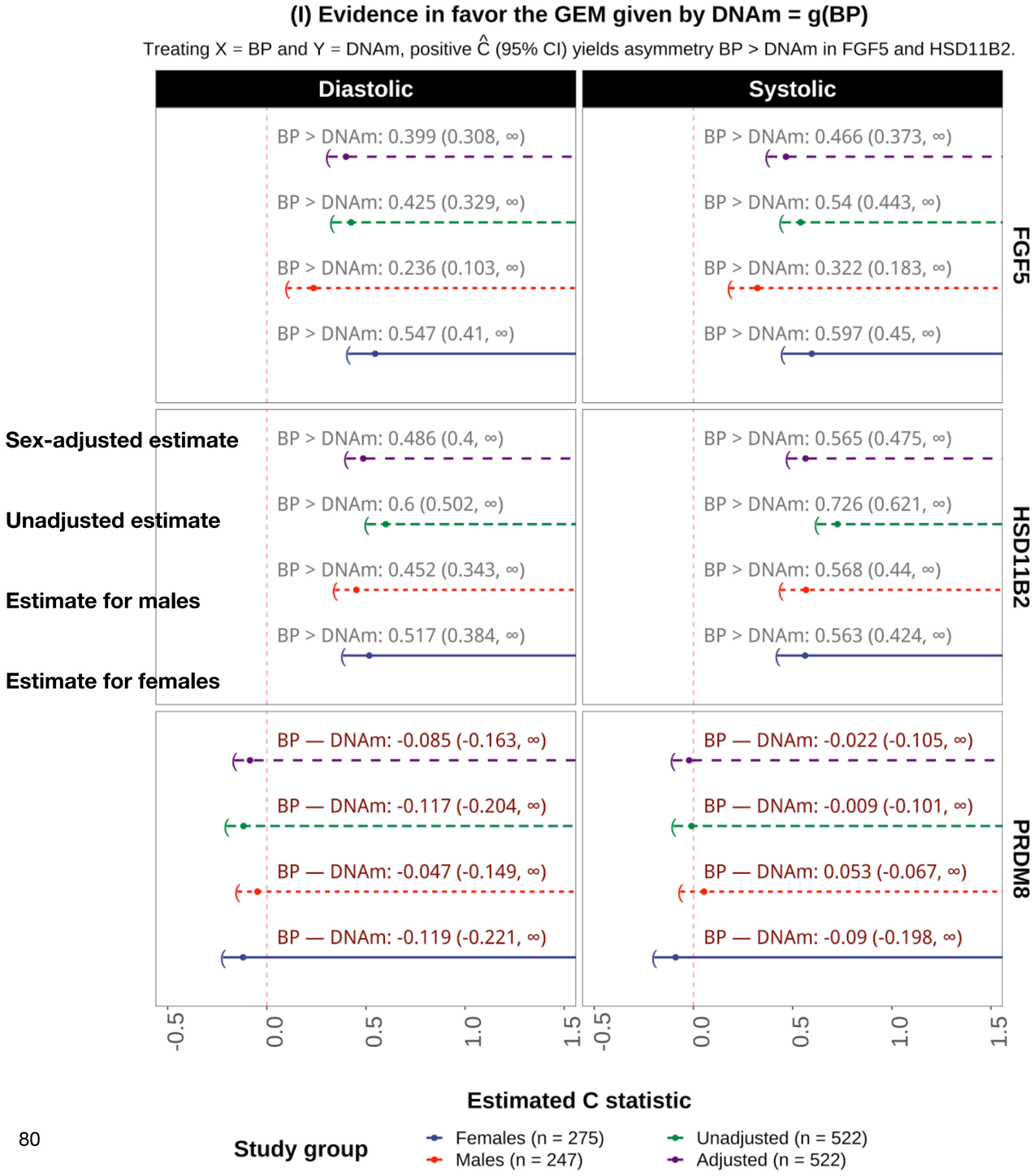
$X : BP$ (either diastolic or systolic)
$Y : DNAm$ for a given gene

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



$BP \rightarrow DNAm?$

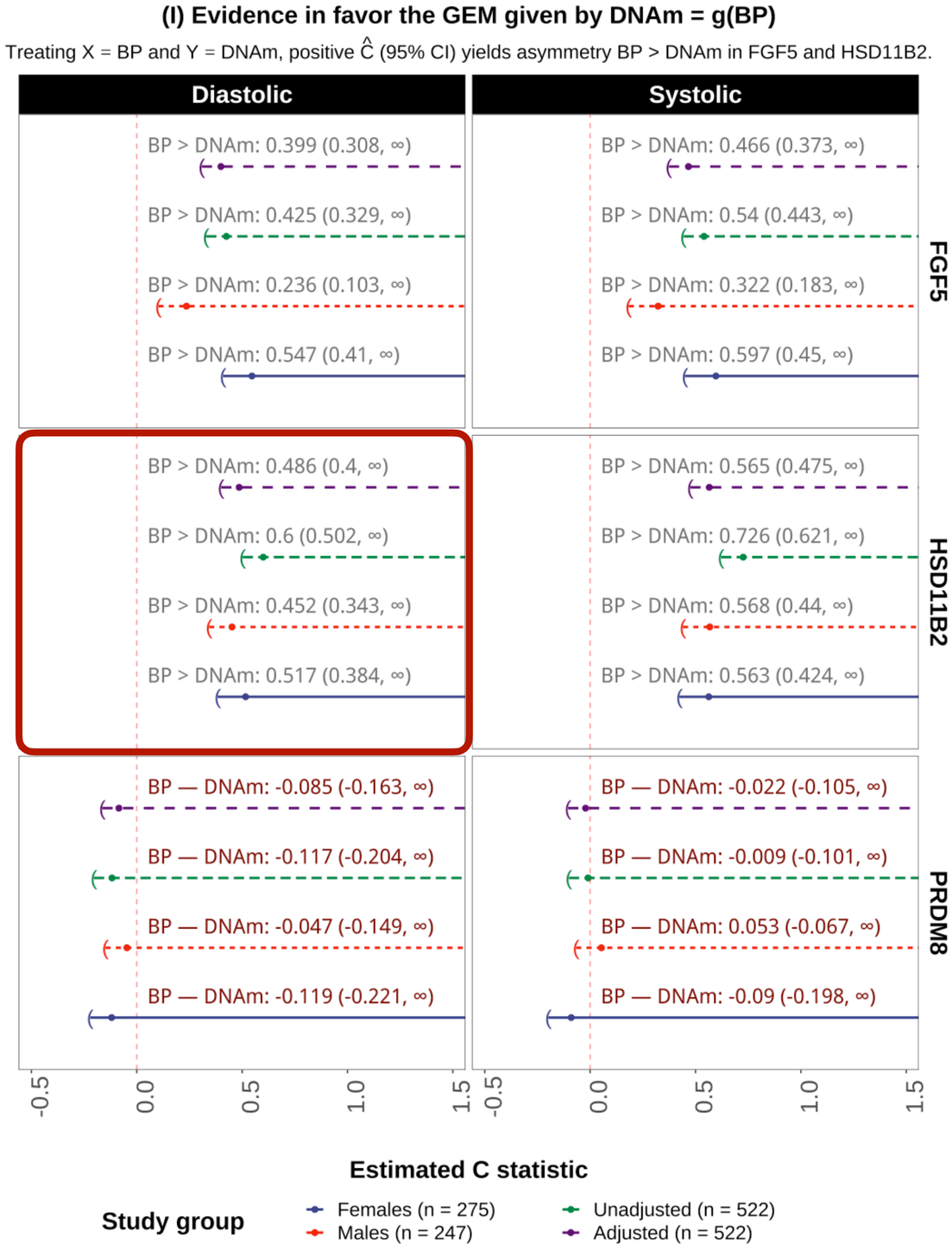
Use $\hat{C}_{X>Y}$ for clues!

$X = DBP$ $\quad$ $Y = DNAm$ 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 $CI)$

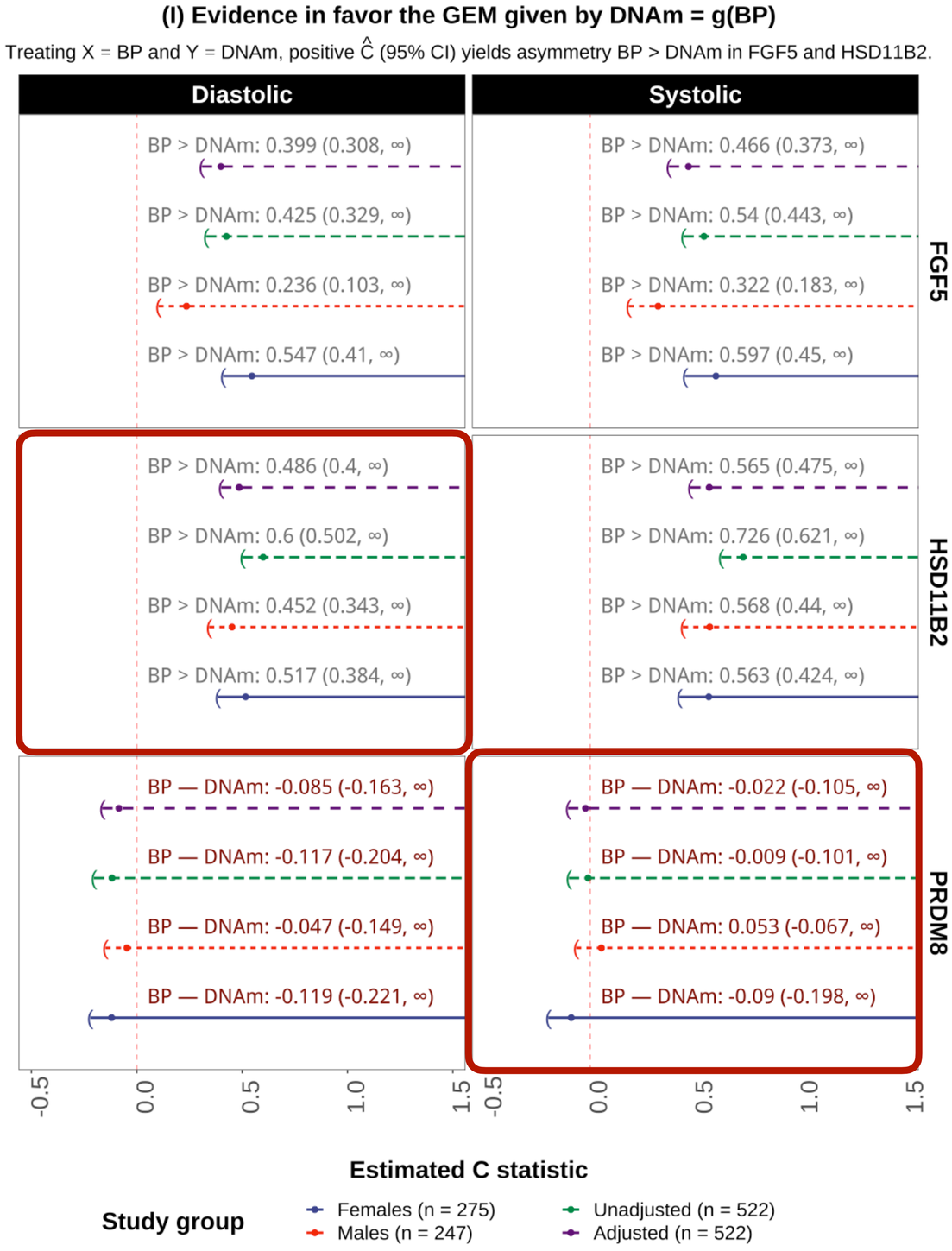
- For female subgroup: 0.517 $(0.384, \infty)$
- Evidence to protect the null.

$X = DBP$ $\quad$ $Y = DNAm$ 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!

$X = DBP$ $Y = DNAm$ 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)

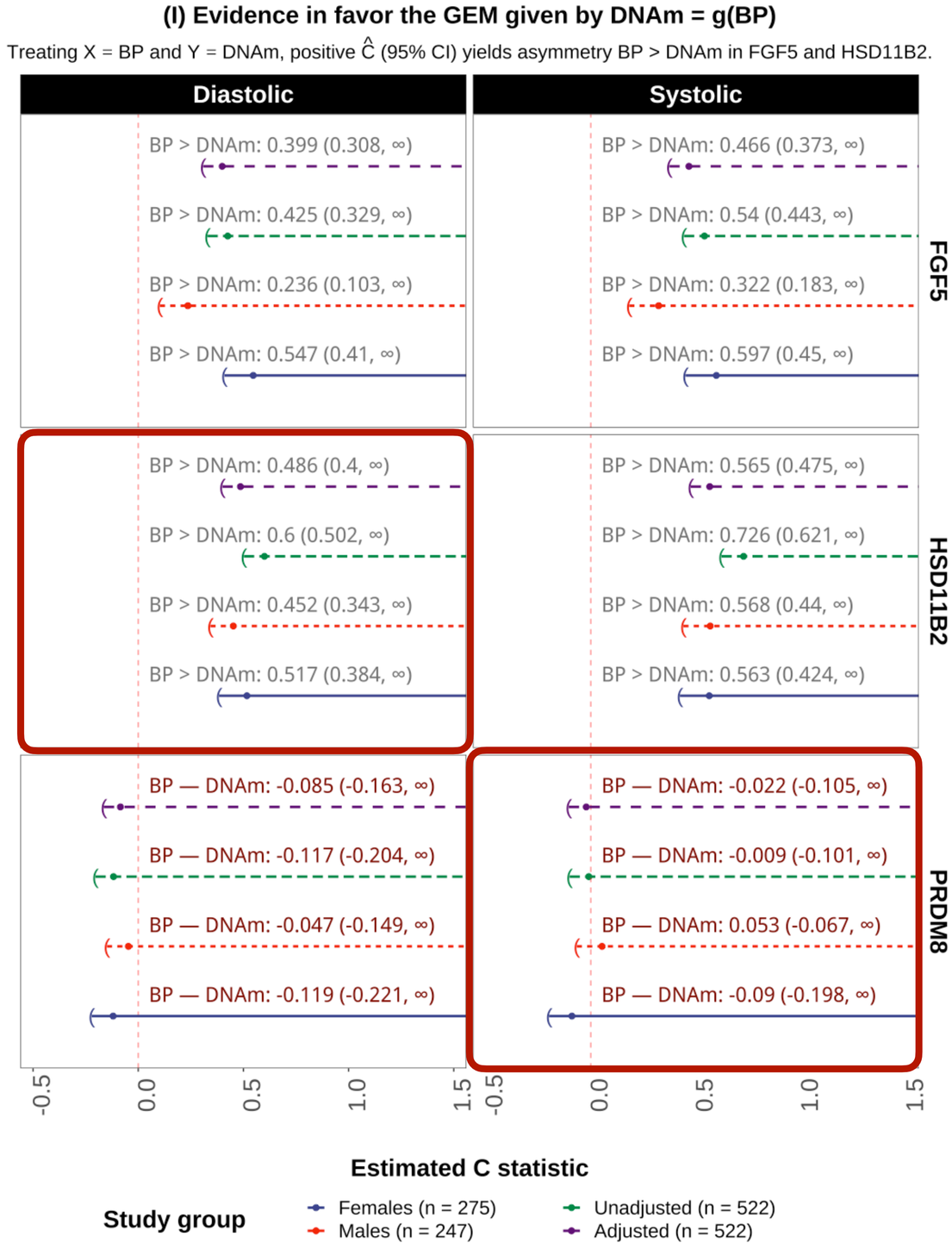
- For female subgroup: $0.517\ (0.384,\ \infty)$
- Evidence to protect the null.

$X = DBP$ $Y = DNAm$ 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??



$BP \rightarrow DNAm?$

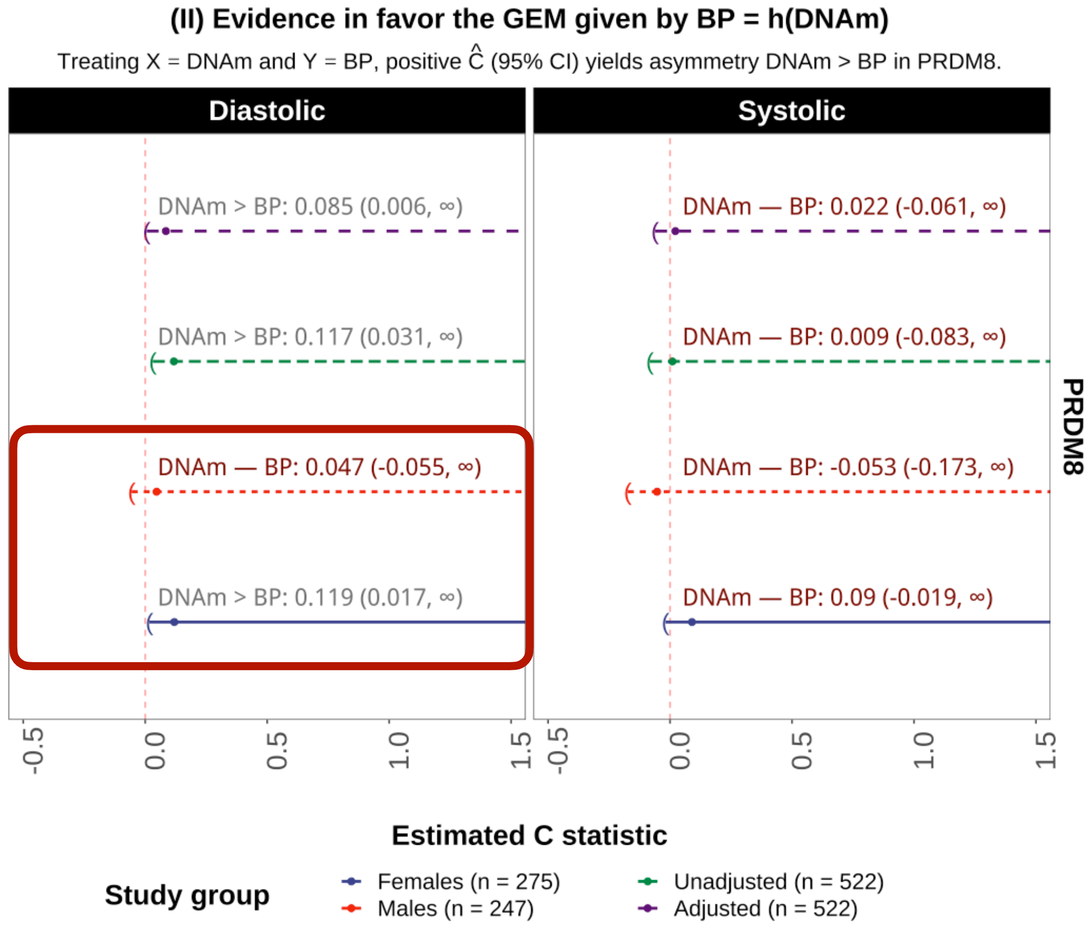
Use $\hat{C}_{X>Y}$ for clues!

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

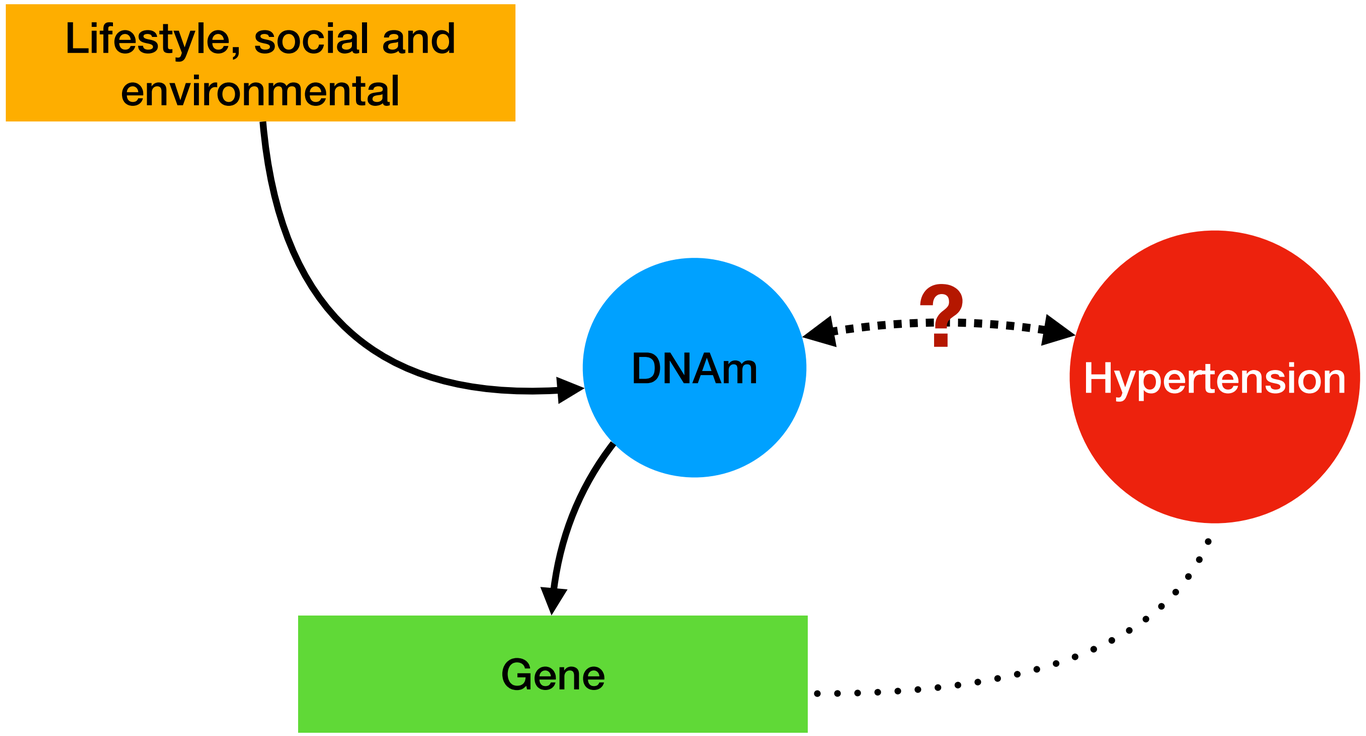
Hypothesized GEM under $H_0 : Y = g(X)$

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

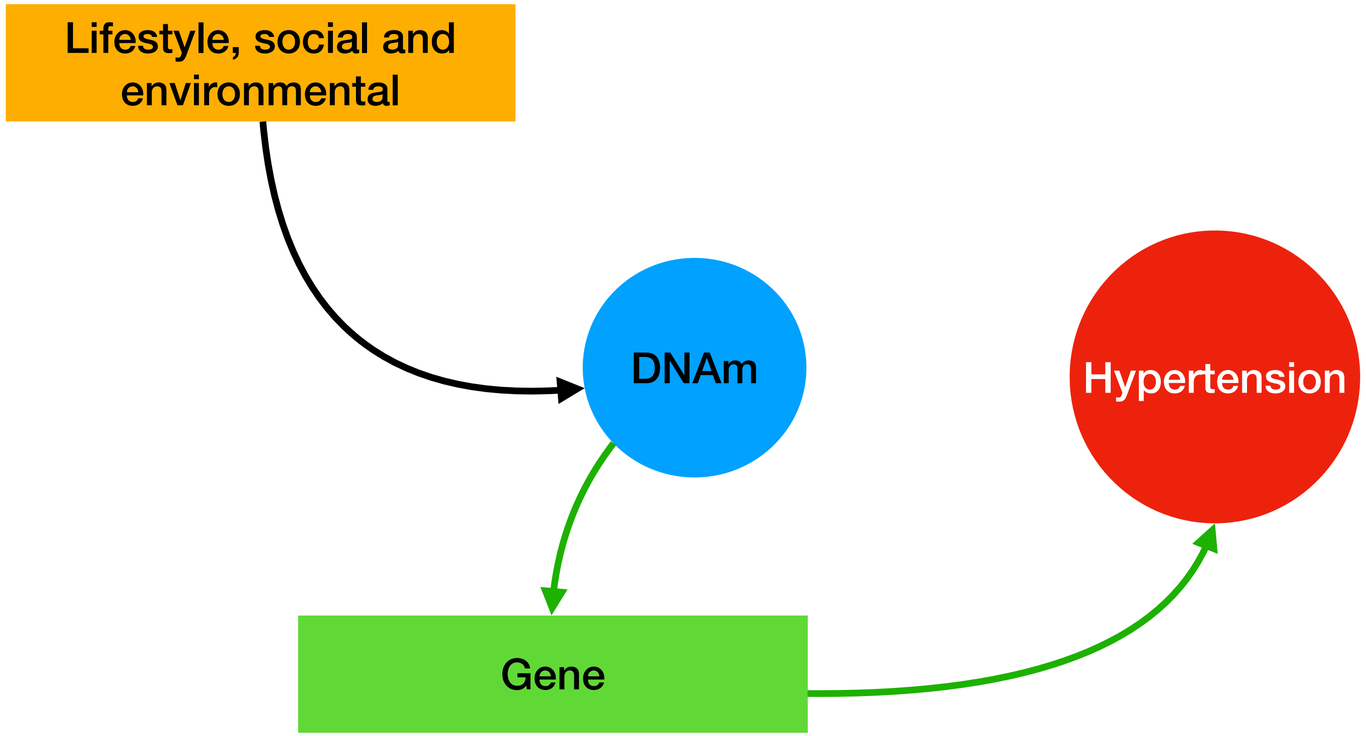
- Simpson's paradox.
- Stratifying is a good idea.



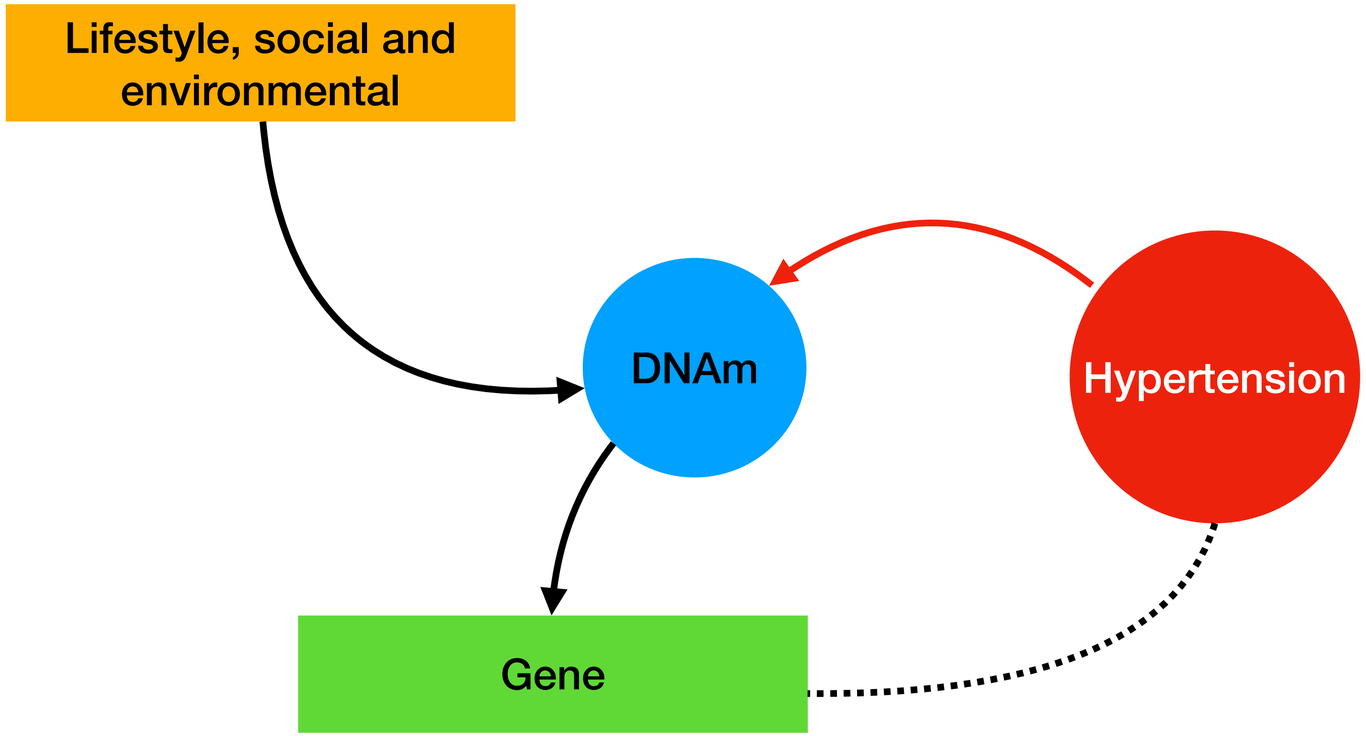
$BP \rightarrow DNAm?$



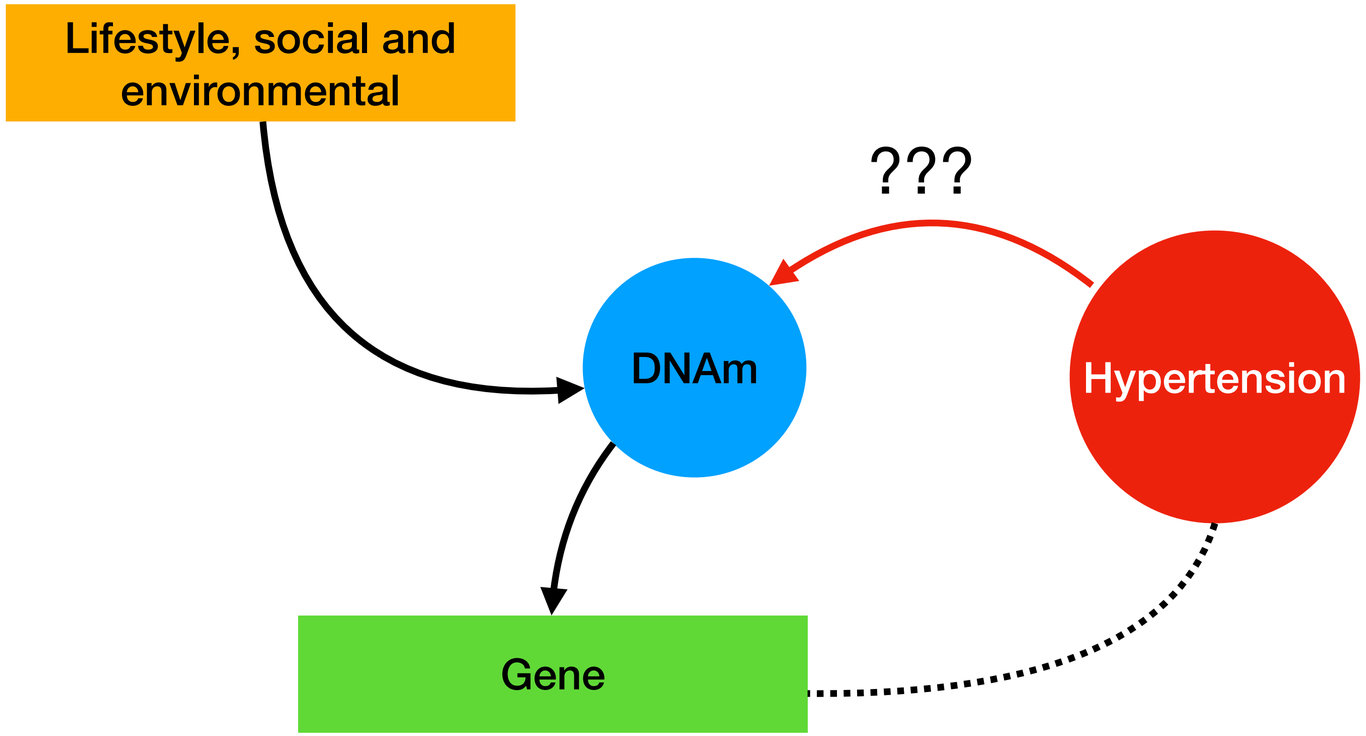
$BP \rightarrow DNAm?$



$BP \rightarrow DNAm?$

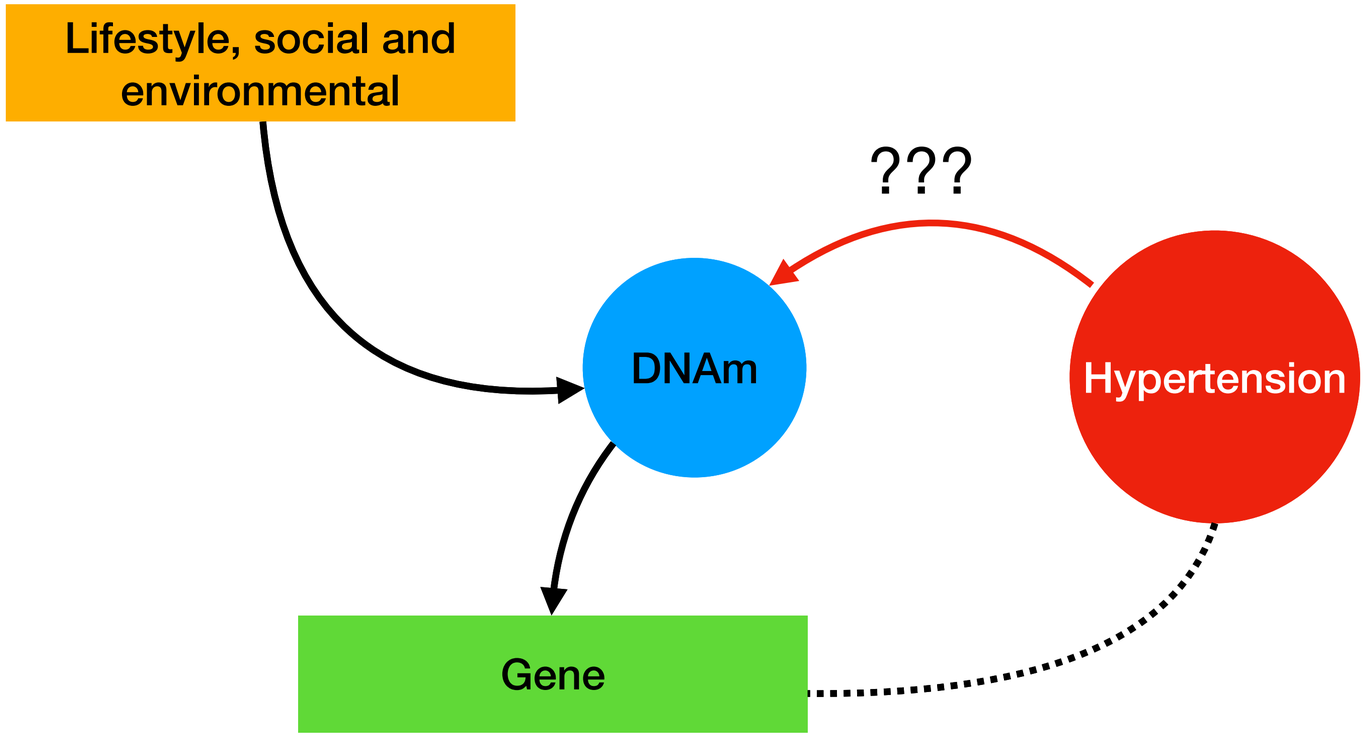


$BP \rightarrow DNAm?$



$BP \rightarrow DNAm?$

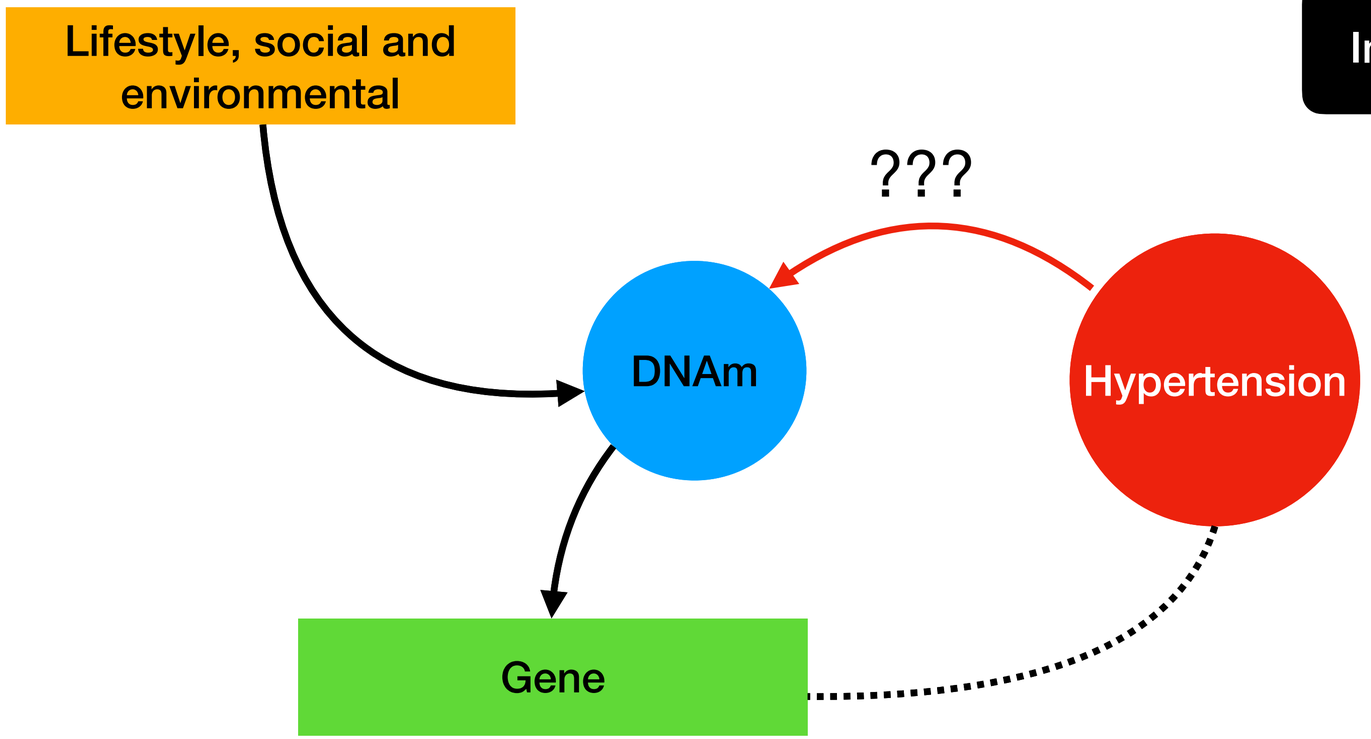
Epigenetic Changes in Response to Hypertension



$BP \rightarrow DNAm?$

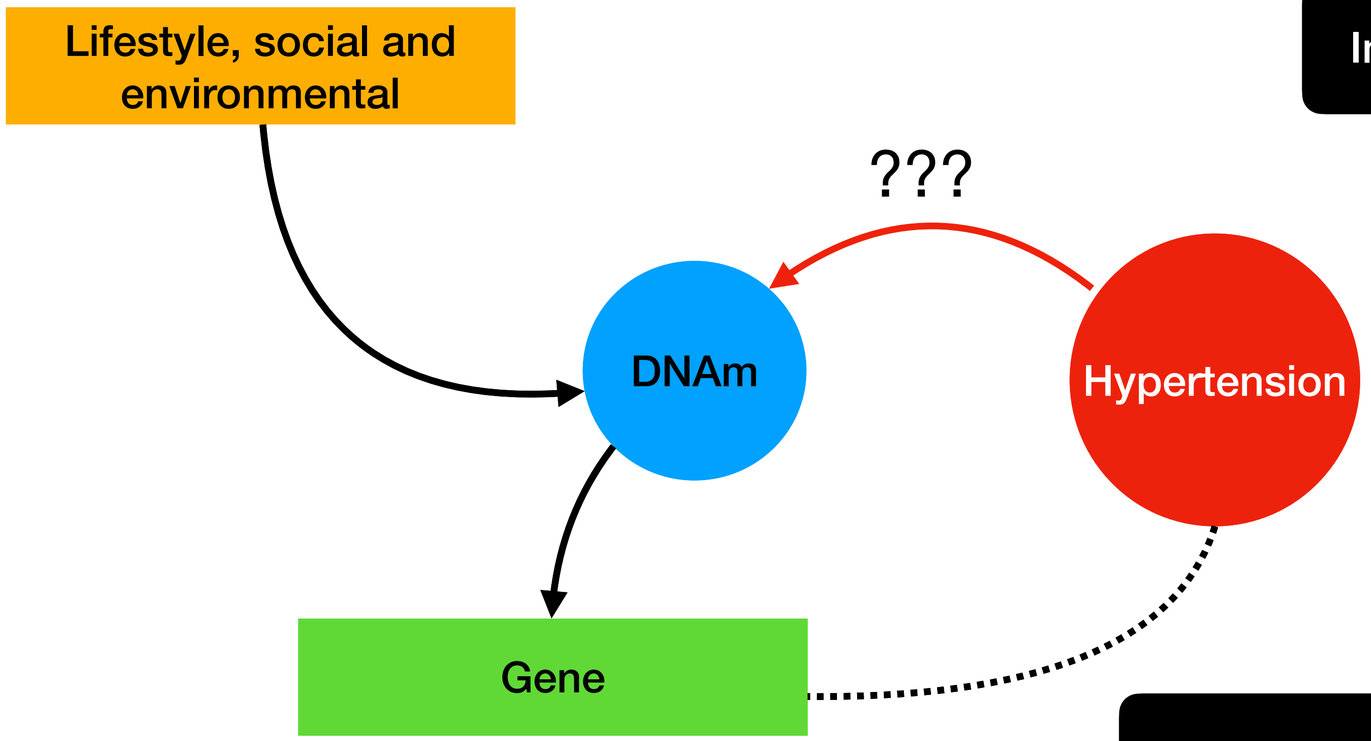
Epigenetic Changes in Response to Hypertension

Impact on Endothelial Cells



$BP \rightarrow DNAm?$

Epigenetic Changes in Response to Hypertension

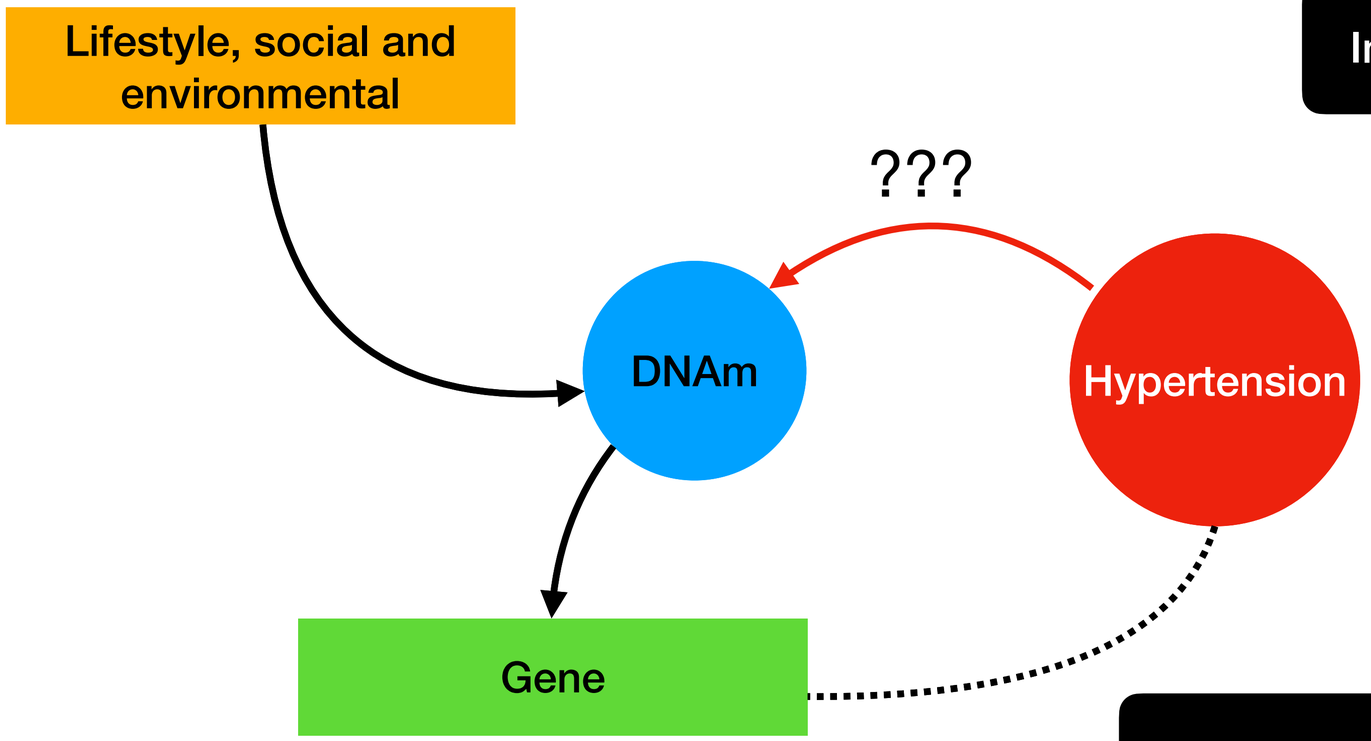


Impact on Endothelial Cells

Inflammation and Oxidative Stress

$BP \rightarrow DNAm?$

Epigenetic Changes in Response to Hypertension

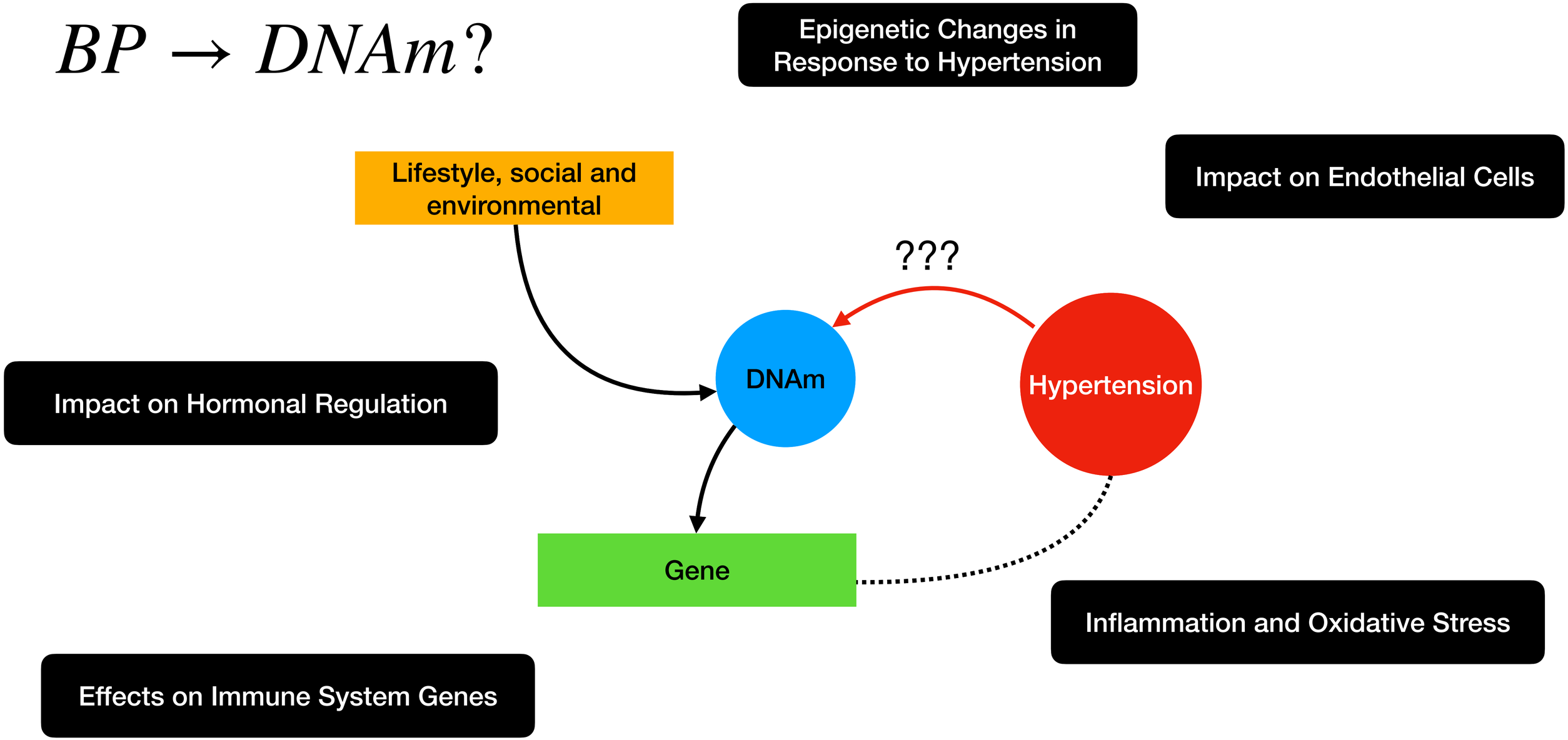


Impact on Endothelial Cells

Inflammation and Oxidative Stress

Effects on Immune System Genes

$BP \rightarrow DNAm?$



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.

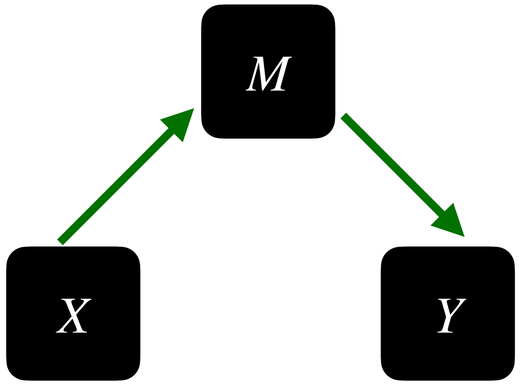
- Fast estimation for large n
- Reduced estimation error in simulations.
- Technical guarantees of data splitting.

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

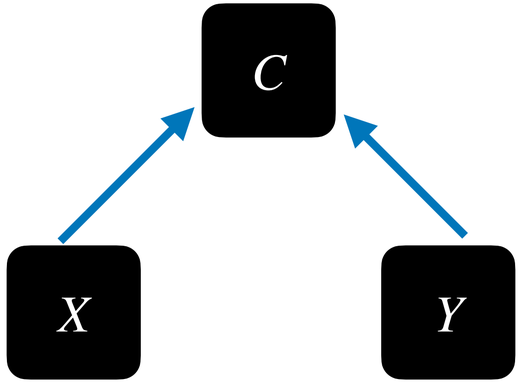
Future work

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

Extend GEM framework to detect third-variable effects

Third variable Z

Chain:

Z mediates pathway between X and Y

Fork:

Z is a common cause for X and Y

Collider:

Z is collider for X and Y

Extend GEM framework to detect third-variable effects

Third variable Z

Chain:

Z mediates pathway between X and Y

Fork:

Z is a common cause for X and Y

Collider:

Z is collider for X and Y

Extend GEM framework to detect third-variable effects

Third variable Z

$$(X \succ Z) \cap (Z \succ Y)$$

$$(X \succ Z) \cap (Y \succ Z)$$

$$(Y \succ Z) \cap (Z \succ X)$$

Extend GEM framework to detect third-variable effects

Third variable Z: mediator

$$\{H(X) - H(Y) > 0\} \cap \{H(Z) - H(Y) > 0\}$$

$$\{H(Y) - H(Z) > 0\} \cap \{H(Z) - H(X) > 0\}$$

Combined:

$$H_0 : \{H(X) - H(Z)\}\{H(Z) - H(Y)\} > 0$$

Test using $\hat{C}_{X>Z}$ and $\hat{C}_{Z>Y}$ [**possibly correlated**]

Extend GEM framework to detect third-variable effects

Third variable Z: collider

Combined:

$$H_0 : \{H(X) - H(Z) > 0\} \cap \{H(Y) - H(Z) > 0\}$$

Test using $\hat{C}_{X>Z}$ and $\hat{C}_{Y>Z}$ [**possibly correlated**]

$$(X \succ Z) \cap (Y \succ Z)$$

Extend GEM framework to detect third-variable effects

Third variable Z

$$(X \succ Z) \cap (Z \succ Y)$$

$$(X \succ Z) \cap (Y \succ Z)$$

$$(Y \succ Z) \cap (Z \succ X)$$

Planned work:

1. Separate out null hypotheses for mediator or collider.
2. Combined p-value?
3. Application to ELEMENT data

Thanks!*

Thanks!*

*I promise to add photos for my defense

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

   $MISE$ 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 $\mathscr{C}$

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

Appendix II

Asymmetry in GEMS reflects underlying directionality

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

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

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

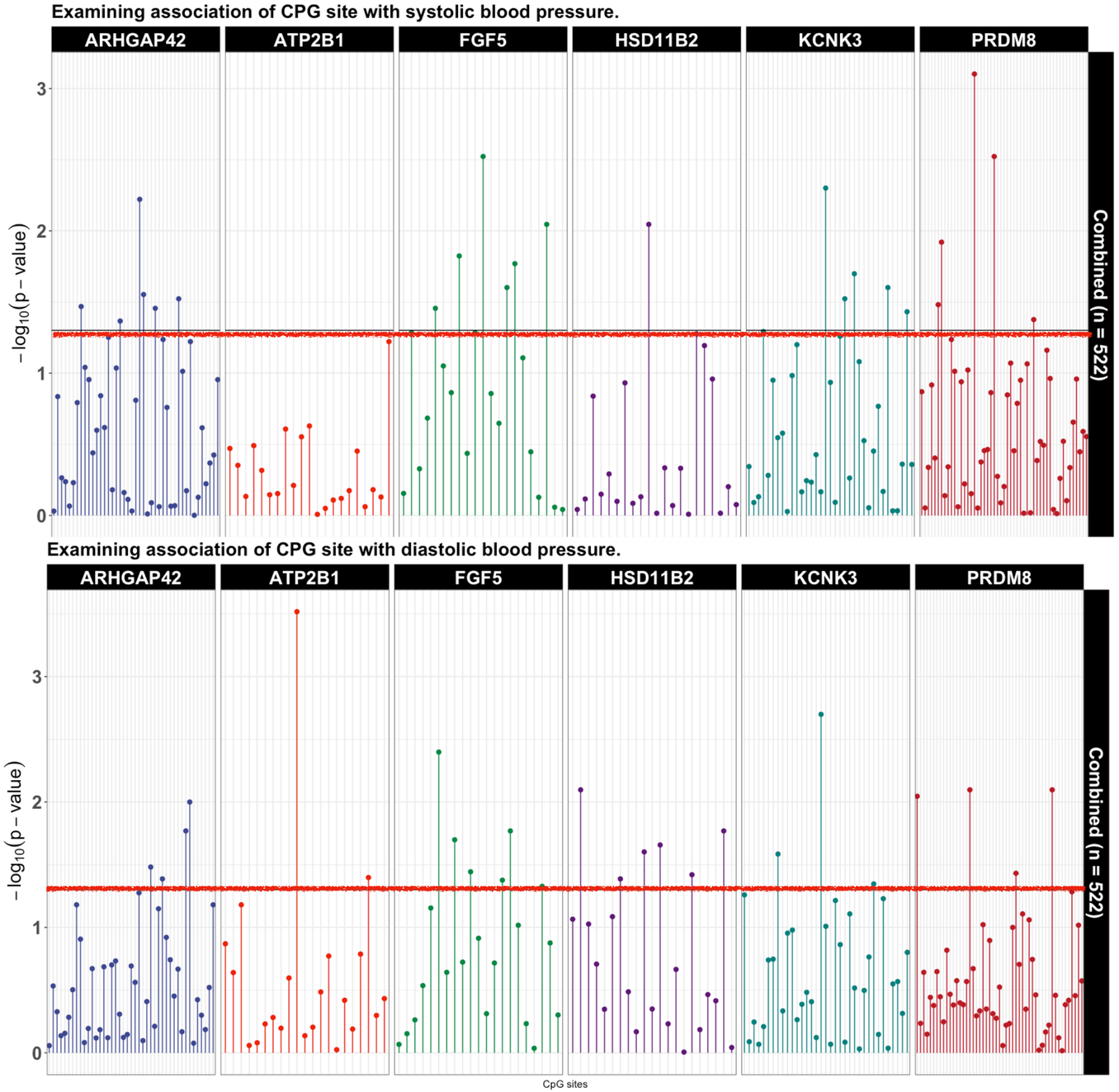
Appendix III

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

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)

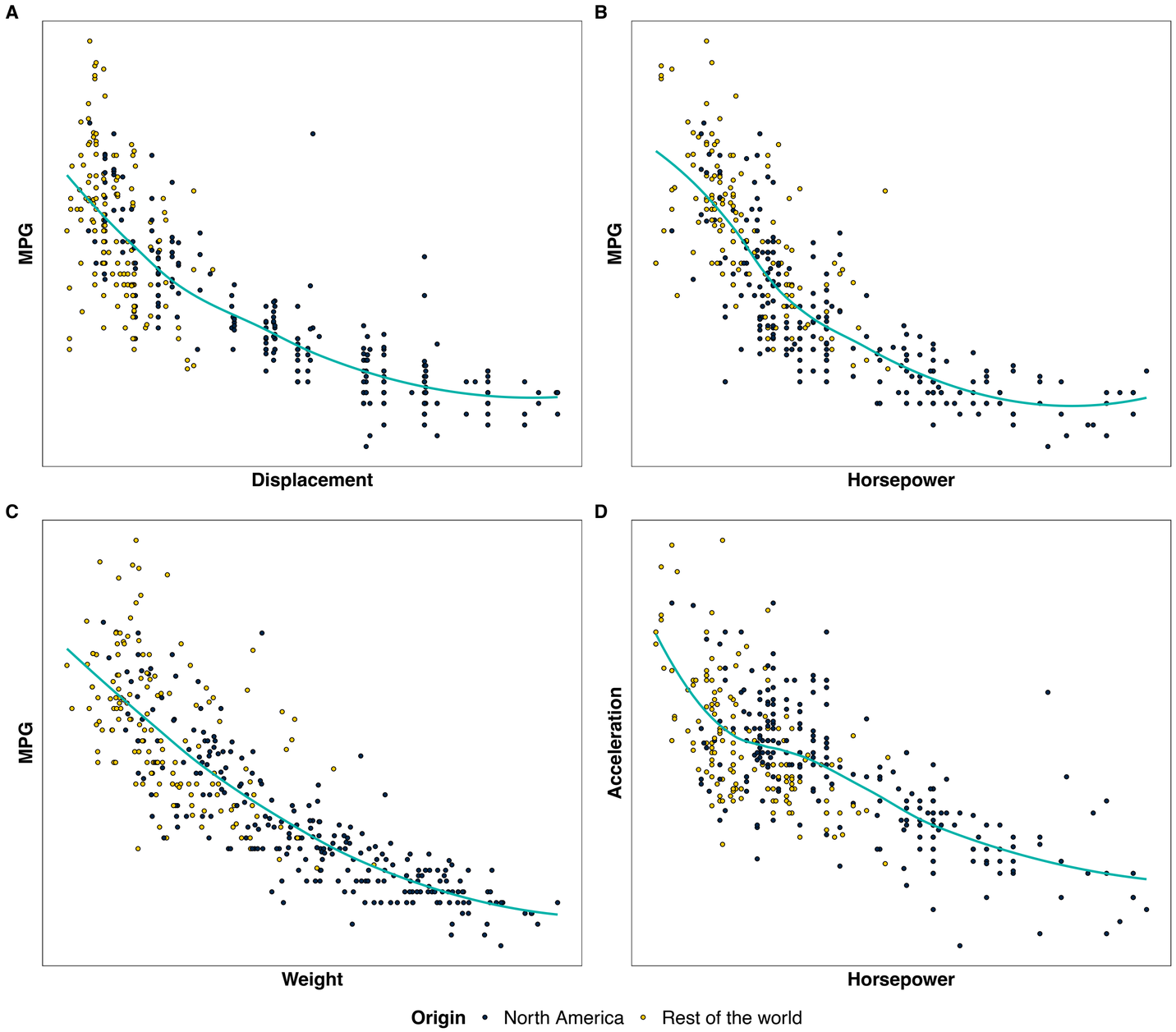
Plot C: Weight → MPG

- **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)

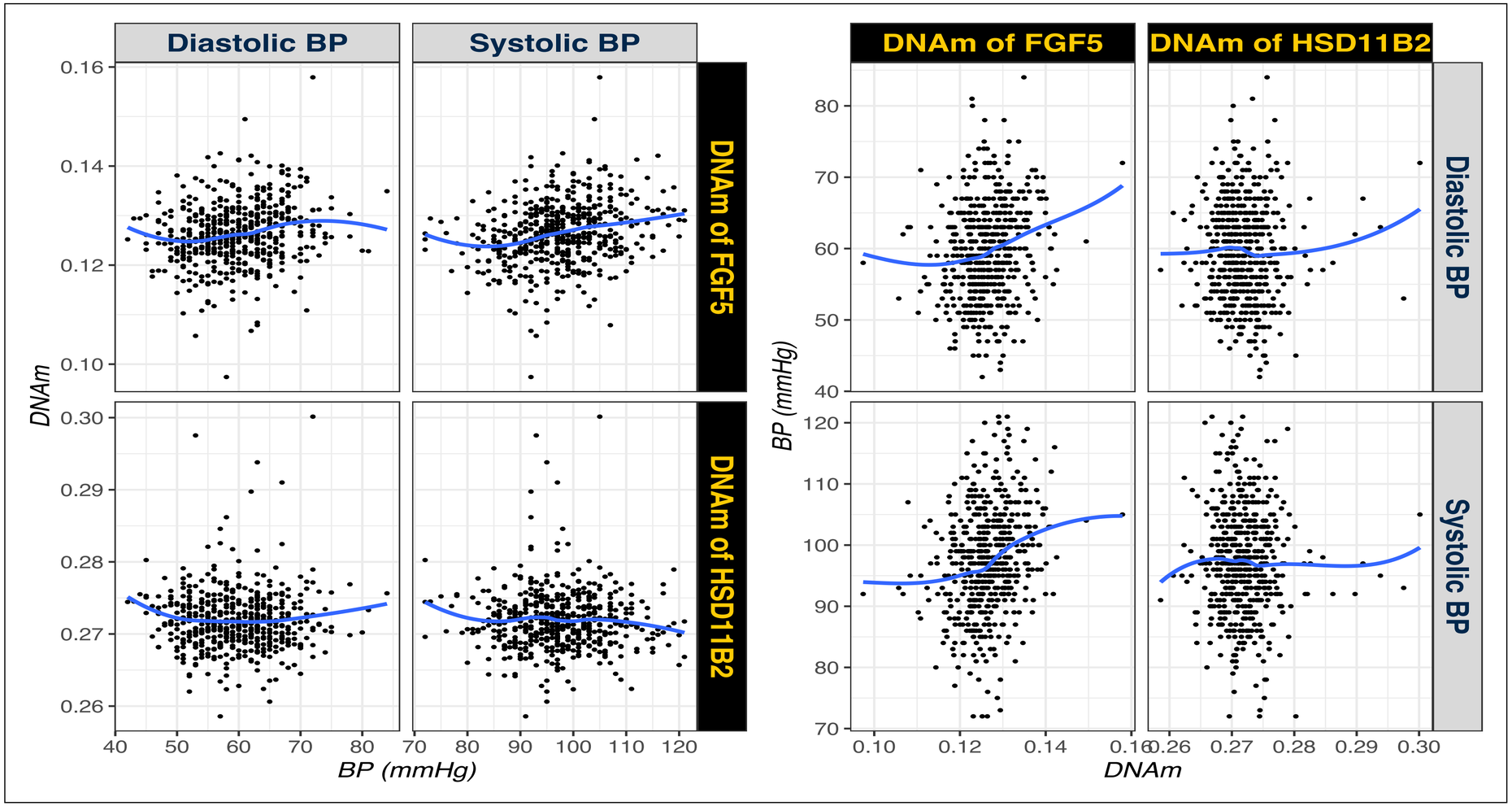
Pairwise scatterplots of features from the Auto MPG dataset of mechanical attributes of n = 390 cars.
Attributes: fuel consumption in miles per gallon (MPG), displacement, horsepower, weight, and acceleration, stratified by origin of car.



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

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

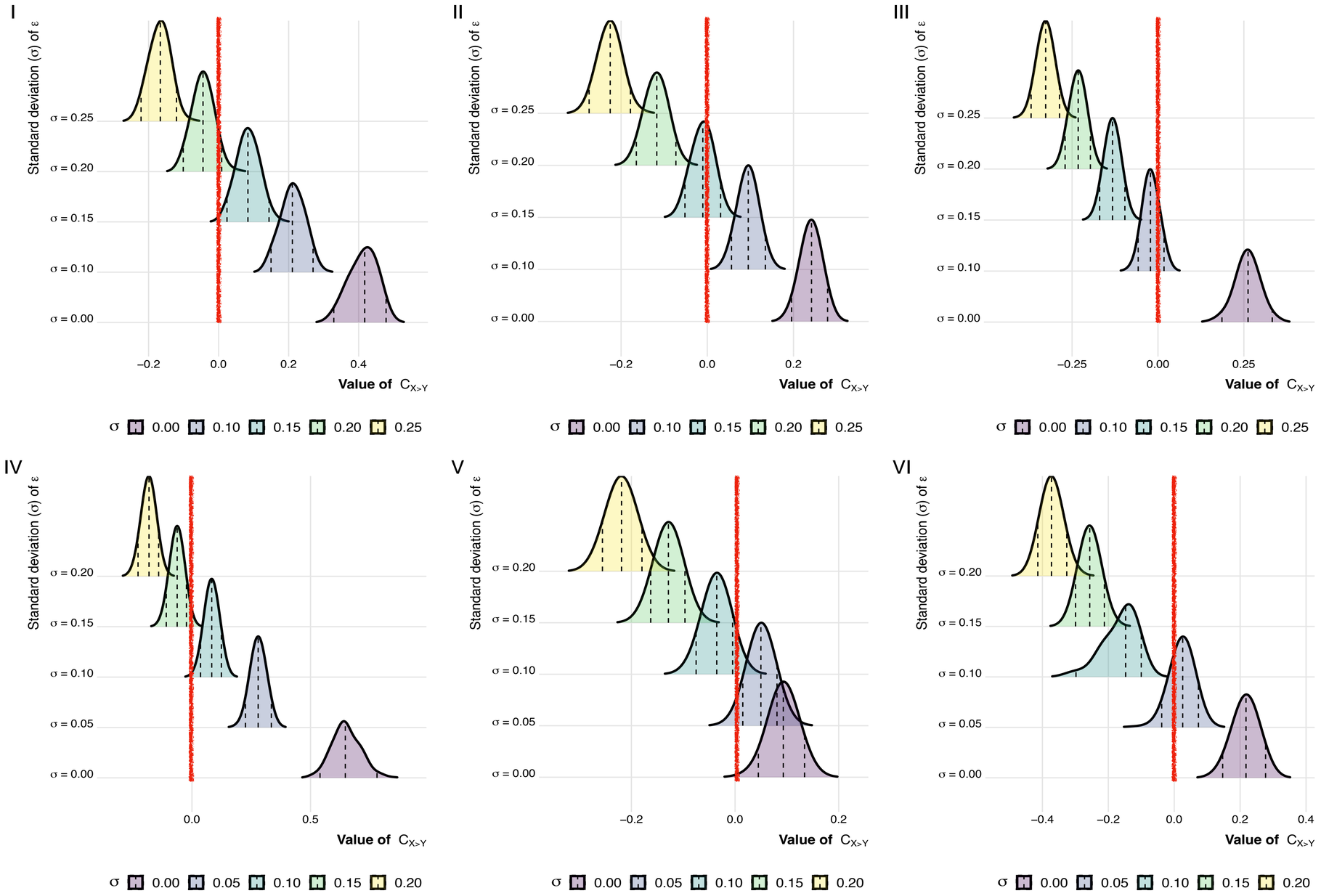
$$\hat{\mathscr{C}}(\mathbf{t}) := n^{-1} \sum_{j=1}^{n} \exp(i\mathbf{t}'\mathbf{Z}_j)$$

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

Appendix VII

Behaviour of $C_{X>Y}$ in the GEM model $Y = f(X) + \varepsilon$, where $X \sim U(0, 1)$ and $\varepsilon \sim N(0, \sigma^2)$. Smoothed density estimates of $C_{X>Y}$ are stratified by standard deviation (σ) of contaminating noise ε.

Case I: $f(x) = x^{1/3}$, II: $f(x) = x^{1/2}$, III: $f(x) = x^2$, IV: $f(x) = x^3$, V: $f(x) = \exp(x)$, VI: $f(x) = \sin(x)$. Dashed vertical lines in each density plot correspond to 2.5th, 50th and 97.5th percentiles of each distribution.



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 $\hat{\sigma}_C$.
3. Repeat steps 1 and 2 $r = 250$ times.

1. SD of simulated $\hat{C}$ gives empirical standard error (ESE).
2. Mean of estimated $\hat{\sigma}_C$ 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: $X \sim E(1); Y \sim Wb(1,3.2)$ $C_{X>Y} = H(X) - H(Y) = 0.213$		

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:

$$V_X = \Phi^{-1}\left(U_X\right); V_Y = \Phi^{-1}\left(U_Y\right)$$

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

Estimating c is replaced by estimating g

Fast and tuning free estimation of g

Appendix X

Simulation I: Estimation accuracy

Compute mean squared error of fastMI when estimating MI

Simulation II: Empirical power

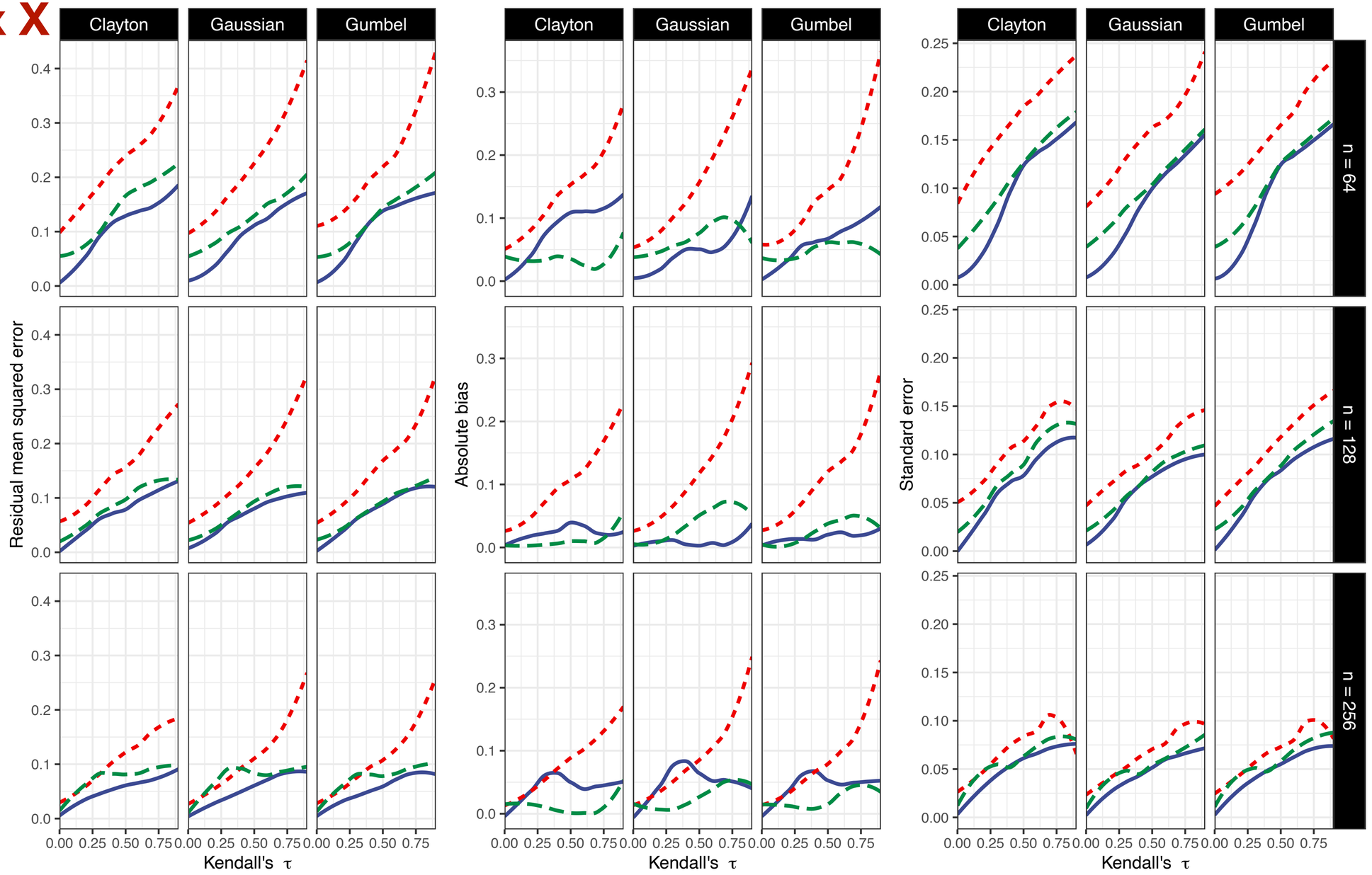
Compute empirical power of fastMI when testing for $H_0 : MI = 0$

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.

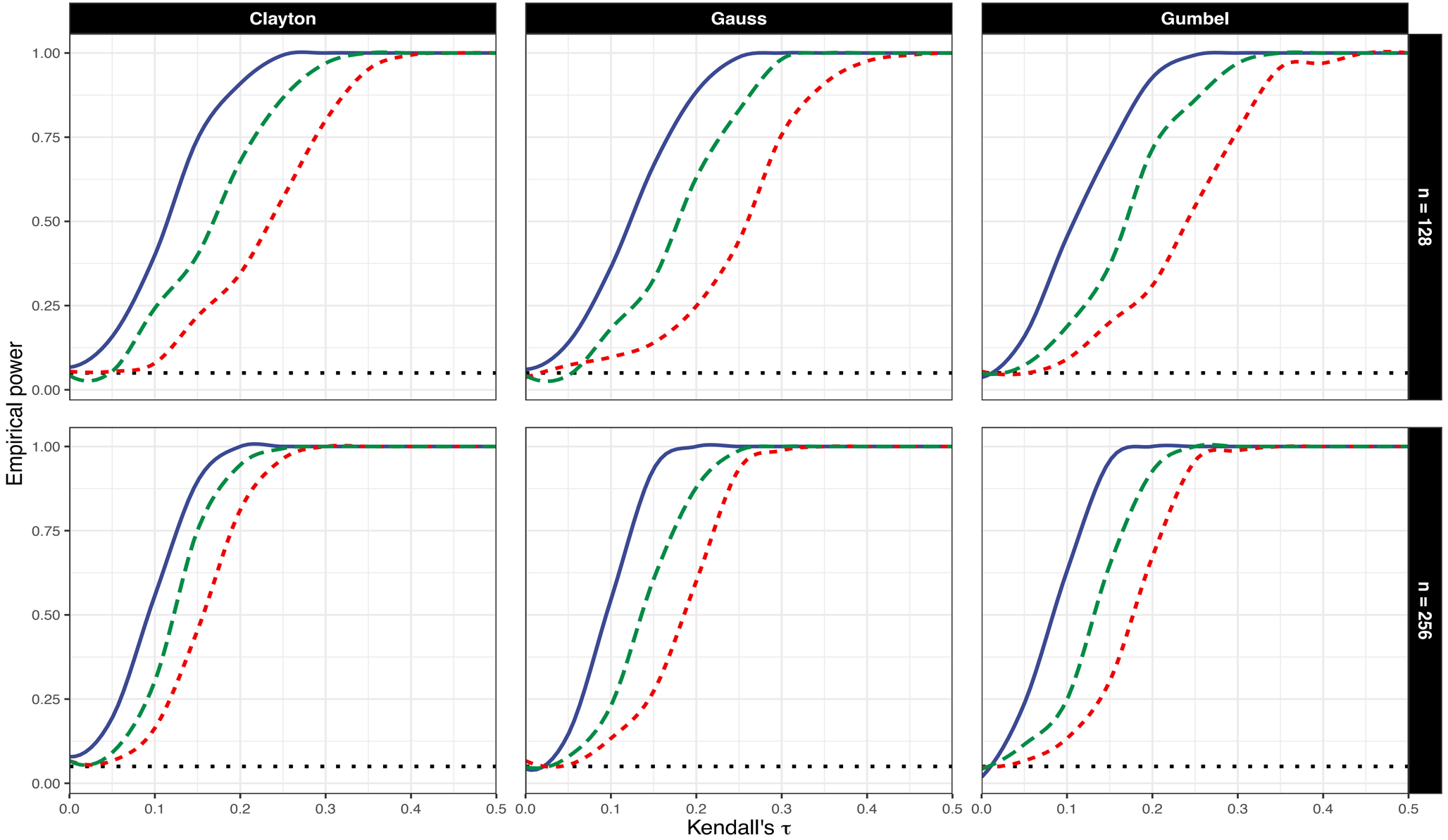
Increasing $\tau \leftrightarrow$ Increasing MI.

Appendix X


Clayton
Gaussian
Gumbel
n = 64
n = 128
n = 256
Residual mean squared error
Absolute bias
Standard error
Kendall's τ
0.00 0.25 0.50 0.75
0.0 0.1 0.2 0.3 0.4
0.00 0.05 0.10 0.15 0.20 0.25
Estimation method
fastMI
JMI
ECMI

Appendix X

The dotted black line parallel to the x–axis denotes specified level of significance = 0.05.

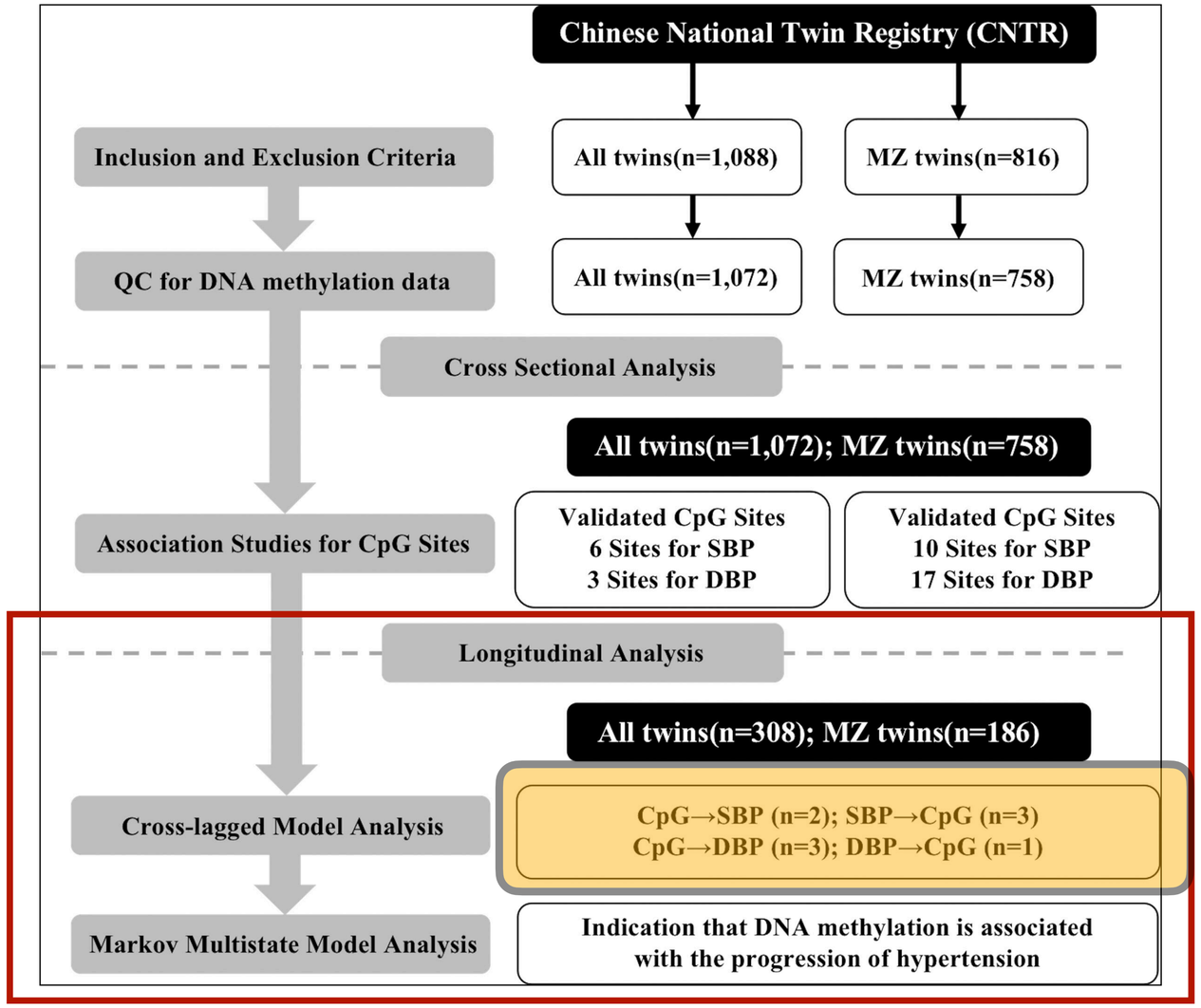


Appendix XI

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.



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