The Statistics of Summary-Data MR

Qingyuan Zhao

Department of Statistics, Wharton School, University of Pennsylvania
(**From August 1st**: Statistical Laboratory, University of Cambridge)

July 17, 2019 @ MRC-IEU Mendelian randomization conference, Bristol

Slides and more information are available at
`http://www-stat.wharton.upenn.edu/~qyzhao/MR.html`.

Outline of this talk

Design

I **Three-sample** MR: ~~**winner's curse**~~.

II **Genome-wide** MR: exploit weak instruments.

Model

I **Measurement error** in GWAS summary data: ~~**NOME assumption**~~.

II Both **systematic** and **idiosyncratic** pleiotropy.

Analysis

I **Robust adjusted profile score (RAPS)**: robust and efficient inference.

II Extension to **multivariate MR** and **sample overlap**.

Diagnostics

I **Q-Q plot** and **InSIDE plot**: falsify modeling assumptions.

II **Modal plot**: discover mechanistic heterogeneity.

Design I: Three-sample MR

Example: LDL-CAD

- Genetic instruments $Z_1, Z_2, \ldots, Z_n$;
- Exposure X: LDL-cholesterol;
- Outcome Y: coronary artery disease (CAD).

Data pre-processing

Name	**Selection GWAS**	Exposure GWAS	Outcome GWAS
Dataset	**GLGC (2010)**	GLGC (2013)	CARDIoGRAM + C4D + UKBB
GWAS	**Linear regression** $X \sim Z_j$	Linear regression $X \sim Z_j$	Logistic regression $Y \sim Z_j$
Coefficient	**Used for selection**	$\hat{\gamma}_j$	$\hat{\Gamma}_j$
Std. Err.		σ_{Xj}	σ_{Yj}

- Use **selection GWAS** to select **independent instruments** that are associated with the exposure (p-value $\leq p_{\text{sel}}$).

Selection GWAS must be independent

Common misconception

We do not need the third selection GWAS if only "genome-wide significant" SNPs are used (e.g. p-value $\leq 5 \times 10^{-8}$).

This is wrong because, although the SNPs are most likely "true hits", the associations are **still overestimated due to selection**.

A simple example

```
> z <- rnorm(10^6); z[1:100] <- z[1:100] + 5
> pval <- 2*pnorm(-abs(z))
> sum(pval < 5e-8)
[1] 33
> mean(z[pval < 5e-8])
[1] 6.112361
```

Selection GWAS must be independent (cont.)

A real data example: BMI-BMI

- Exposure X = Outcome Y = BMI, so **true "causal effect" = 1**.
- **Selection GWAS = Exposure GWAS** using 50% UKBB; Outcome GWAS computed using the other 50%.

p_{sel}	# SNPs	Mean F	IVW	W. Median	W. Mode
1e-8	168	57.00	0.823 (0.017)	0.8 (0.022)	0.885 (0.053)
1e-6	305	43.92	0.761 (0.015)	0.736 (0.019)	0.865 (0.079)
1e-4	652	30.68	0.678 (0.012)	0.616 (0.015)	0.593 (0.122)
1e-2	1289	20.70	0.592 (0.01)	0.528 (0.013)	0.554 (0.093)

p_{sel}	# SNPs	Median F	Egger	PS	RAPS
1e-8	168	41.12	1.018 (0.046)	0.848 (0.014)	0.831 (0.018)
1e-6	305	33.68	1.006 (0.041)	0.793 (0.011)	0.763 (0.016)
1e-4	652	23.23	0.89 (0.033)	0.724 (0.009)	0.66 (0.014)
1e-2	1289	15.26	0.749 (0.025)	0.657 (0.008)	0.541 (0.012)

Design II: Genome-wide MR

Instrument selection

- **No p-value threshold is used when selecting IVs**.
- The only requirement is that the SNPs are independent.

Weak IV bias?

Wait... Didn't you just show that **weaker IVs bring more bias**?

Three sources of bias

1. Winner's curse.
   Solution: Three-sample design.
2. Weak IV bias (dividing by a small number).
   Solution: Use appropriate model and statistical methods.
3. Weak IVs have more pleiotropic effect.
   "Solution": InSIDE assumption..

Validation of genome-wide MR

The BMI-BMI example

- Exposure X = Outcome Y = BMI, so true "causal effect" = 1.
- **Selection GWAS = GIANT consortium;**
- Exposure GWAS using 50% UKBB;
- Outcome GWAS computed using the other 50%.

p_{sel}	# SNPs	Mean F	IVW	W. Median	W. Mode
1e-8	58	69.2	0.983 (0.024)	0.945 (0.039)	0.939 (0.044)
1e-6	126	44.1	0.986 (0.022)	0.944 (0.034)	0.931 (0.038)
1e-4	287	26.1	0.981 (0.017)	0.941 (0.031)	0.929 (0.035)
1e-2	812	12.7	0.928 (0.014)	0.879 (0.023)	0.739 (7.130)

p_{sel}	# SNPs	Median F	Egger	**PS**	**RAPS**
1e-8	58	42.0	0.928 (0.050)	**0.999 (0.023)**	**0.998 (0.025)**
1e-6	126	27.4	0.881 (0.043)	**1.017 (0.019)**	**1.009 (0.023)**
1e-4	287	15.8	0.921 (0.031)	**1.023 (0.017)**	**1.018 (0.018)**
1e-2	812	5.6	0.909 (0.022)	**1.010 (0.015)**	**1.005 (0.015)**

Validation of genome-wide MR (cont.)

In many (but not all) real examples, the MR results are stable across different instrument strength.

Example: LDL-CAD

Selection threshold	**RAPS Results** Only	Cumulative
$0 \leq p \leq 10^{-8}$	0.48 (0.04)	0.48 (0.04)
$10^{-8} \leq p \leq 10^{-4}$	0.36 (0.11)	0.46 (0.04)
$10^{-4} \leq p \leq 1$	0.34 (0.26)	0.48 (0.03)

Example: BMI-CAD

Selection threshold	**RAPS Results** Only	Cumulative
$0 \leq p \leq 10^{-8}$	0.34 (0.13)	0.34 (0.13)
$10^{-8} \leq p \leq 10^{-4}$	0.34 (0.15)	0.34 (0.09)
$10^{-4} \leq p \leq 1$	0.45 (0.11)	0.39 (0.07)

Model I: Measurement error in GWAS summary data

Simplifying requirement

Exposure GWAS and outcome GWAS have no sample overlap.

Assumption 1

Let $\hat{\gamma} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_n)$ be the vector of exposure coefficients (similarly $\hat{\Gamma}$):

$$\begin{pmatrix} \hat{\gamma} \\ \hat{\Gamma} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \gamma \\ \Gamma \end{pmatrix}, \ \mathrm{diag}(\sigma_{X1}^2, \ldots, \sigma_{Xn}^2, \sigma_{Y1}^2, \ldots, \sigma_{Yn}^2) \right).$$

Three-sample design warrants Assumption 1

Name	Selection GWAS	Exposure GWAS	Outcome GWAS
GWAS	$\mathrm{lm}(X \sim Z_j)$	$\mathrm{lm}(X \sim Z_j)$	$\mathrm{lm}(Y \sim Z_j)$
Coefficient	Used for selection	$\hat{\gamma}_j$	$\hat{\Gamma}_j$
Std. Err.		σ_{Xj}	σ_{Yj}

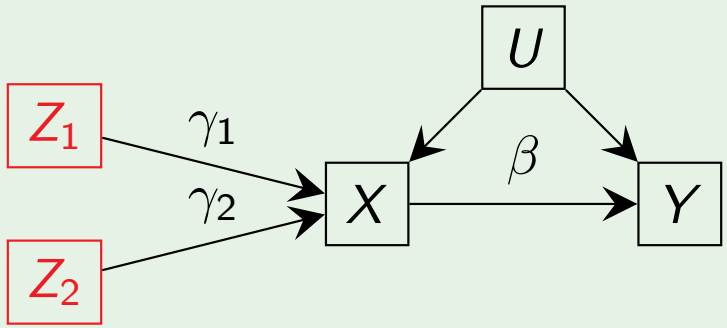
- Large sample size $\Rightarrow$ **normal distribution** (central limit theorem).
- Independence (**diagonal covariance matrix**) due to
  1. Non-overlapping samples (between all three GWAS).
  2. Independent SNPs.

Ideal setting

The causal effect β satisfy $\Gamma_j = \beta\gamma_j$ **for all** j if

- All the genetic IVs are valid and mutually independent;
- The variables follow a linear structural model;

Heuristic



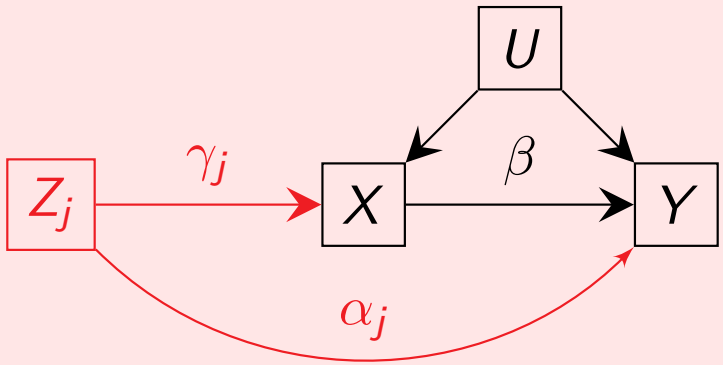
$$X = \sum_{j=1}^{p} \gamma_j Z_j + \eta_X U + E_X,$$

$$Y = \beta X + \sum_{j=1}^{p} \alpha_j Z_j + \eta_Y U + E_Y$$

$$= \sum_{j=1}^{p} \underbrace{(\beta\gamma_j) Z_j}_{\Gamma_j} + \underbrace{\sum_{j=1}^{p} \alpha_j Z_j}_{\text{0 by exclusion restriction}} + \underbrace{f(U, E_X, E_Y)}_{\text{independent of Z}}$$

Model II: Invalid IV

Pleiotropy $\Longrightarrow$ Violation of exclusion restriction



Assumption 2

Let $\alpha_j = \Gamma_j - \beta\gamma_j$ be the "direct effect". We allow for two kinds of deviation:

Systematic pleiotropy For most j, $\alpha_j \perp\!\!\!\perp \gamma_j$ (InSIDE) and $\alpha_j \sim \mathrm{N}(0, \tau^2)$.

Idiosyncratic pleiotropy For a few j, $|\alpha_j|$ might be much larger.

Both kinds of pleiotropy exist in exploratory data analysis.

Invariance to allele coding

Assumption 2

Let $\alpha_j = \Gamma_j - \beta\gamma_j$ be the "direct effect". We assume

Systematic pleiotropy For most j, $\alpha_j \perp\!\!\!\perp \gamma_j$ (InSIDE) and $\boldsymbol{\alpha_j \sim \mathbf{N}(0, \tau^2)}$.

Idiosyncratic pleiotropy For a few j, $|\alpha_j|$ might be much larger.

No "directional" pleiotropy?

Why do you assume the mean of α_j is 0?

Allele recoding

In GWAS, switching **effective allele** $\leftrightarrow$ **reference allele** of SNP j amounts to:

$$\hat{\gamma}_j \leftarrow -\hat{\gamma}_j,\ \hat{\Gamma}_j \leftarrow -\hat{\Gamma}_j,\ \textbf{thus } \alpha_j \leftarrow -\alpha_j.$$

- "Directional" pleiotropy is always **relative** to the allele coding we use.
- Instead, RAPS is **invariant** to allele coding.

Analysis I: RAPS

Heuristics

In the ideal setting where $\alpha_j \equiv 0$, we would like to solve the equation:

$$\sum_{j=1}^{n} \textbf{Estimated IV strength}_j(\beta) \cdot \textbf{Estimated direct effect}_j(\beta) = 0.$$

Statistical equivalence:

$$\hat{\gamma}_{j,\mathrm{MLE}}(\beta, \tau^2) = \frac{\hat{\gamma}_j/\sigma_{Xj}^2 + \beta\hat{\Gamma}_j/(\sigma_{Yj}^2+\tau^2)}{1/\sigma_{Xj}^2 + \beta^2/(\sigma_{Yj}^2+\tau^2)} \perp\!\!\!\perp \hat{\alpha}_j(\beta, \tau^2) = \frac{\hat{\Gamma}_j - \beta\hat{\gamma}_j}{\sqrt{\sigma_{Yj}^2 + \beta^2\sigma_{Xj}^2 + \tau^2}}.$$

Robust adjusted profile score (invariant to allele coding!)

$$\frac{1}{n}\sum_{j=1}^{n} f\Big(\hat{\gamma}_{j,\mathrm{MLE}}(\beta, \tau^2)\Big) \cdot \psi\Big(\hat{\alpha}_j(\beta, \tau^2)\Big) = 0,$$

$$\frac{1}{n}\sum_{j=1}^{n} \hat{\alpha}_j(\beta, \tau^2) \cdot \psi\Big(\hat{\alpha}_j(\beta, \tau^2)\Big) = \mathbb{E}\big[T \cdot \psi(T)\big], \text{ for } T \sim \mathrm{N}(0,1).$$

ψ is the derivative of a robust loss function and f is (empirical Bayes) shrinkage.

Analysis II: Extensions

Multivariate MR

Modify the RAPS equations straightforwardly.

Sample overlap

- The modified RAPS equations depend on $\mathbf{cor}(\hat{\Gamma}_j, \hat{\gamma}_j)$.
- If no missing data, one can show quite generally

$$\mathbf{cor}(\hat{\Gamma}_j, \hat{\gamma}_j) \approx \sqrt{n^2/(n_X n_Y)} \cdot \mathrm{cor}(X, Y)$$

  does not depend on j (n is the #overlap, n_X and n_Y are the total #sample).
- Can thus estimate $\mathbf{cor}(\hat{\Gamma}_j, \hat{\gamma}_j)$ by sample correlation of the "null" SNPs (or the intercept in LD-score regression).

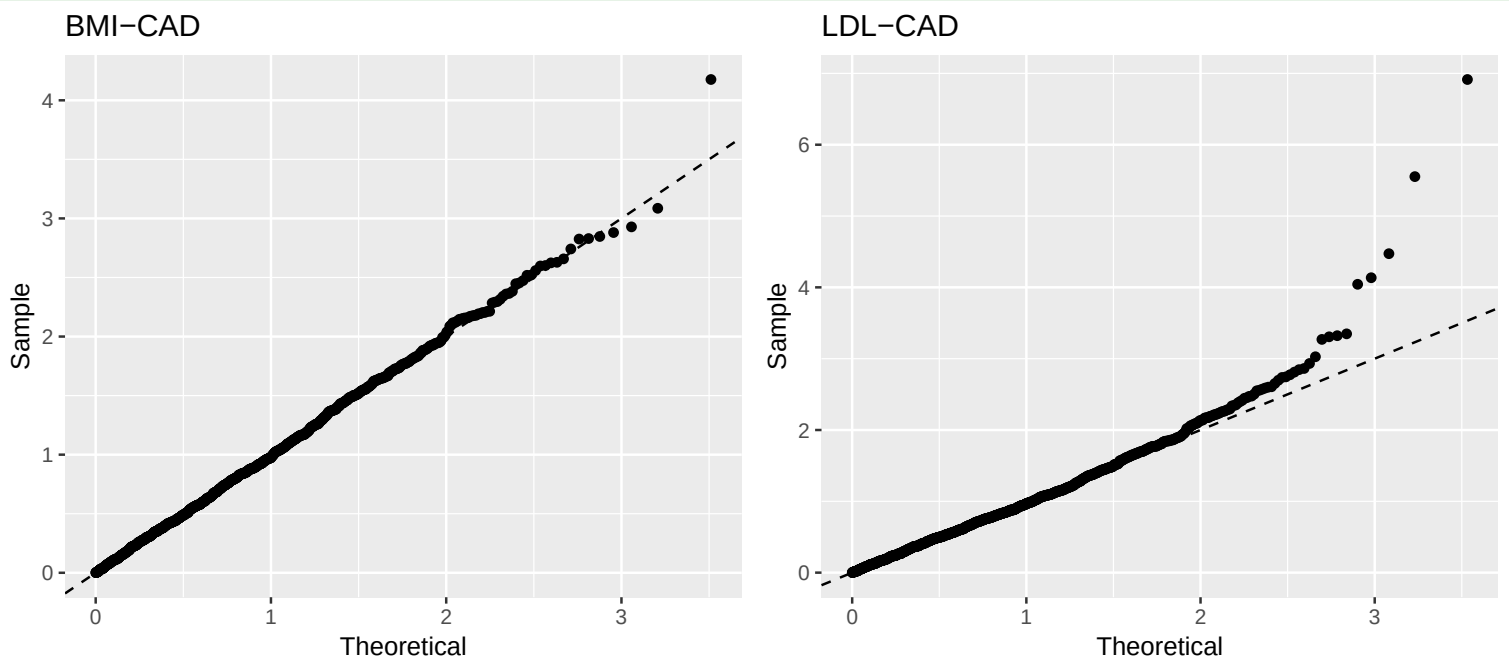
Diagnostics I: Falsifications

Key implication of **Assumption 1**

$$\hat{\alpha}_j(\beta, \tau^2) = \frac{\hat{\Gamma}_j - \beta\hat{\gamma}_j}{\sqrt{\sigma^2_{Yj} + \beta^2\sigma^2_{Xj} + \tau^2}}.$$

Under the measurement error model, $\hat{\alpha}_j(\beta, \tau^2)$ **at the truth** $\sim \mathbf{N(0,1)}$.

Quantile-Quantile plot: $\left|\hat{\alpha}_j(\hat{\beta}, \hat{\tau}^2)\right|$ against $\left|\mathrm{N}(0,1)\right|$



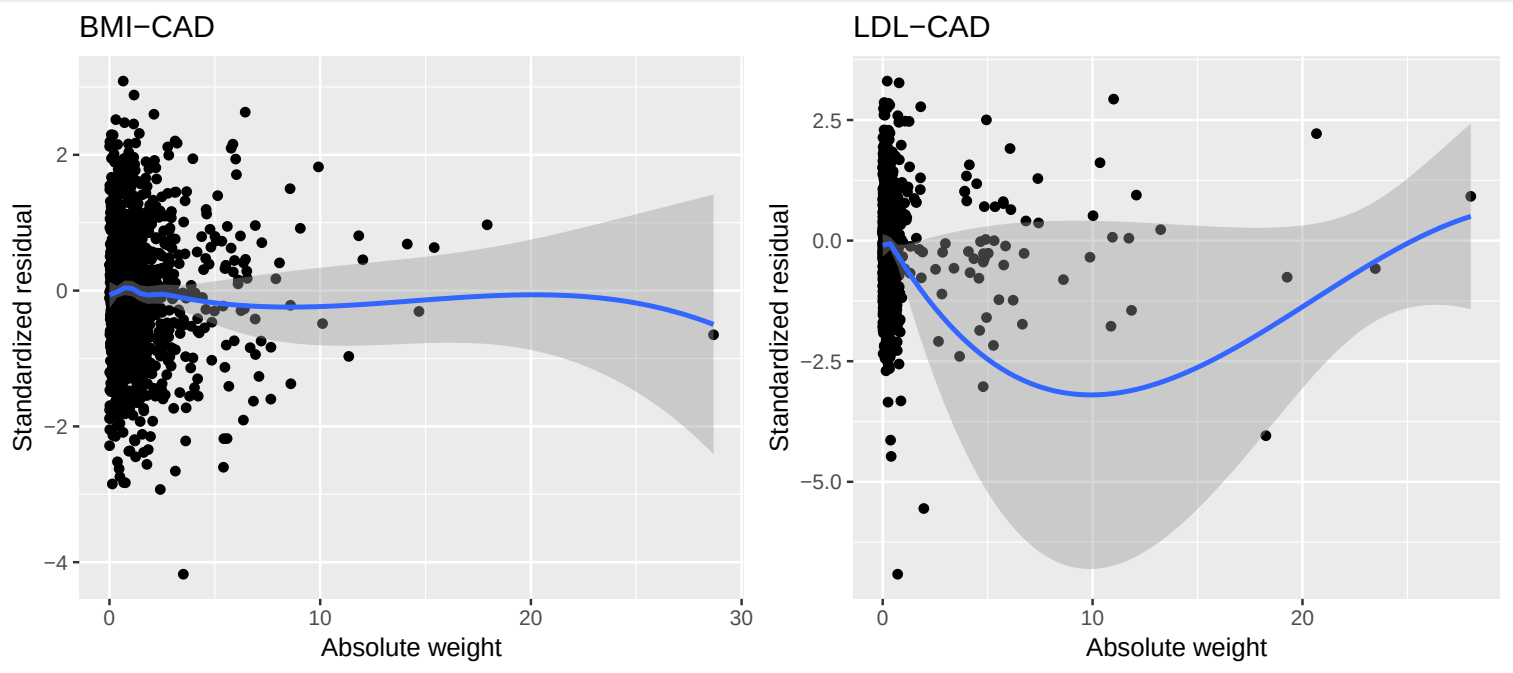
Diagnostics I: Falsifications

Key implication of **Assumption 2**

Under the InSIDE assumption,

$$\hat{\gamma}_{j,\mathsf{MLE}}(\beta,\tau^2)=\frac{\hat{\gamma}_j/\sigma_{Xj}^2+\beta\hat{\Gamma}_j/(\sigma_{Yj}^2+\tau^2)}{1/\sigma_{Xj}^2+\beta^2/(\sigma_{Yj}^2+\tau^2)} \perp\!\!\!\perp \hat{\alpha}_j(\beta,\tau^2)=\frac{\hat{\Gamma}_j-\beta\hat{\gamma}_j}{\sqrt{\sigma_{Yj}^2+\beta^2\sigma_{Xj}^2+\tau^2}}.$$

InSIDE plot: $\hat{\alpha}_j(\hat{\beta},\hat{\tau}^2)$ against $\hat{\gamma}_j(\hat{\beta},\hat{\tau}^2)$



Falsification $\neq$ Validation stamp

Diagnostics CAN tell us

Our assumptions reasonably model GWAS summary data for the selected SNPs:

1. $\begin{pmatrix}\hat{\gamma}\\ \hat{\Gamma}\end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix}\gamma\\ \Gamma\end{pmatrix}, \mathrm{diag}(\sigma^2_{X1},\ldots,\sigma^2_{Xn},\sigma^2_{Y1},\ldots,\sigma^2_{Yn})\right)$;
2. For most j and **some** $\tilde{\beta}$, $\alpha_j = \Gamma_j - \tilde{\beta}\gamma_j$ (InSIDE) and $\alpha_j \sim \mathrm{N}(0,\tau^2)$;

Diagnostics CANNOT tell us

InSIDE assumption is satisfied (aka $\tilde{\beta} = \beta$), because

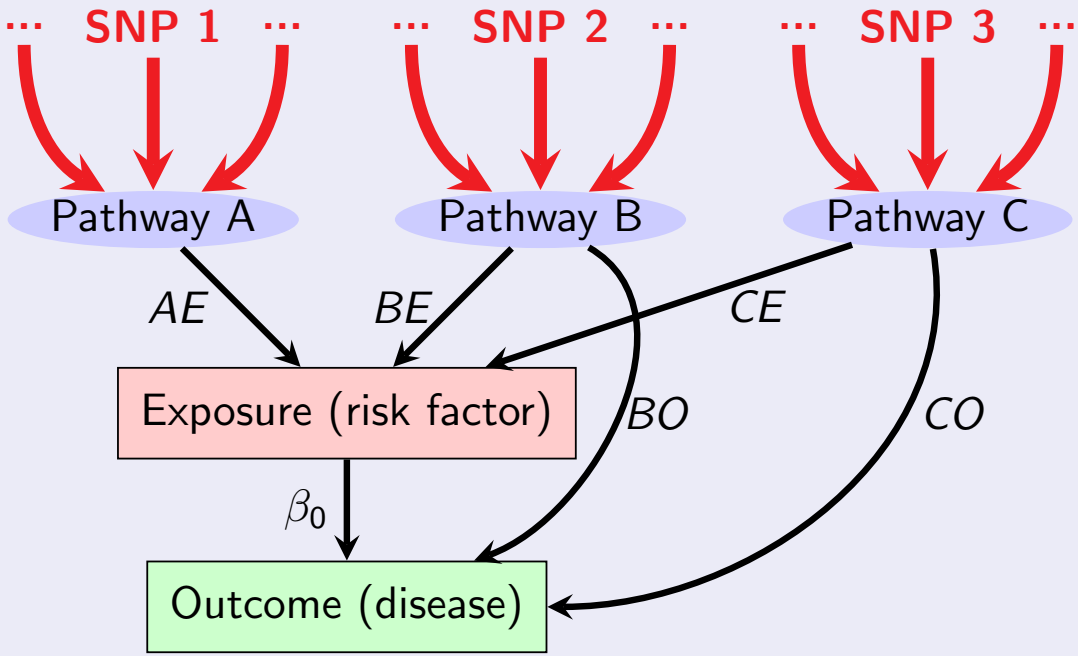
$$\tilde{\beta} = \underbrace{\beta}_{\text{causal effect}} + \underbrace{\mathbf{slope}(\alpha_j \sim \gamma_j)}_{\text{InSIDE assumes = 0}}.$$

It is impossible to distinguish between

- **True causal effect** β;
- **Correlation between** γ_j **and** α_j.

Motivations for mechanistic heterogeneity

Multiple genetic pathways ⇒ Multiple modes of β



	Exposure effect γ	Outcome effect Γ	**Ratio**
SNP 1	$1A \cdot AE$	$1A \cdot AE \cdot \beta_0$	$\boldsymbol{\beta_0}$
SNP 2	$2B \cdot BE$	$2B \cdot BE \cdot \beta_0 + 2B \cdot BO$	$\boldsymbol{\beta_0 + (BO/BE)}$
SNP 3	$3C \cdot CE$	$3C \cdot CE \cdot \beta_0 + 3C \cdot CO$	$\boldsymbol{\beta_0 + (CO/CE)}$

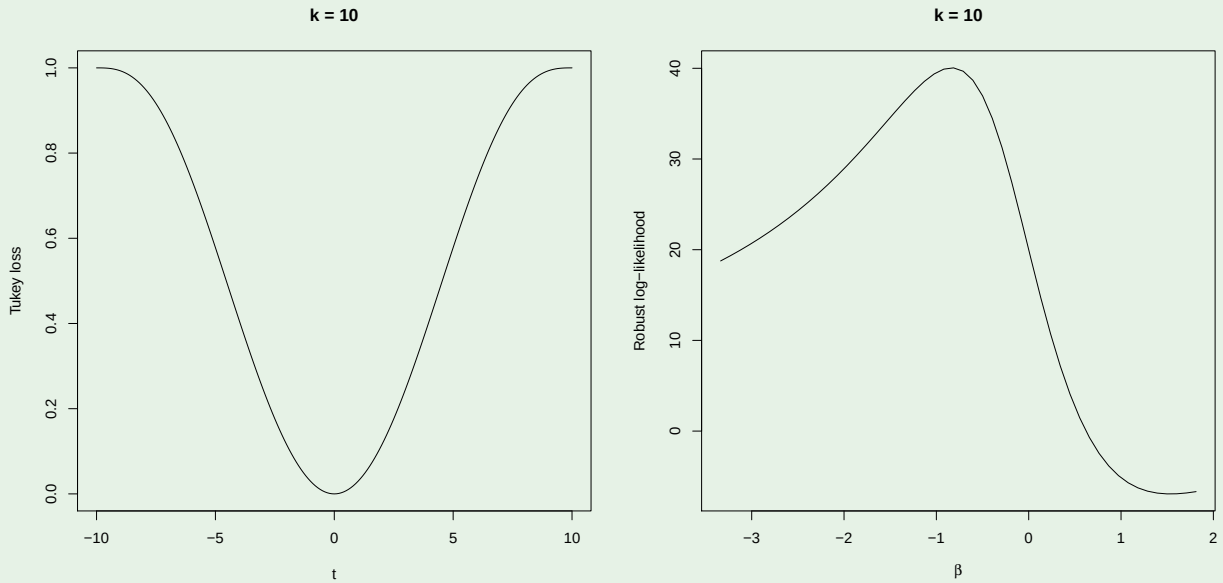
Diagnostics II: Modal plot

Plot robust profile likelihood

$$l_\rho(\beta) = -\sum_{j=1}^{p} \rho\left(\frac{\hat{\Gamma}_j - \beta\hat{\gamma}_j}{\sqrt{1+\beta^2}}\right) \text{ for robust loss function } \rho.$$

Simulation example

$\gamma_j \sim \mathrm{N}(0,4)$, $\Gamma_j = \gamma_j$ for $1 \le j \le 15$, $\Gamma_j = -\gamma_j$ for $16 \le j \le 50$, $\sigma_{Xj} = \sigma_{Yj} = 1$.



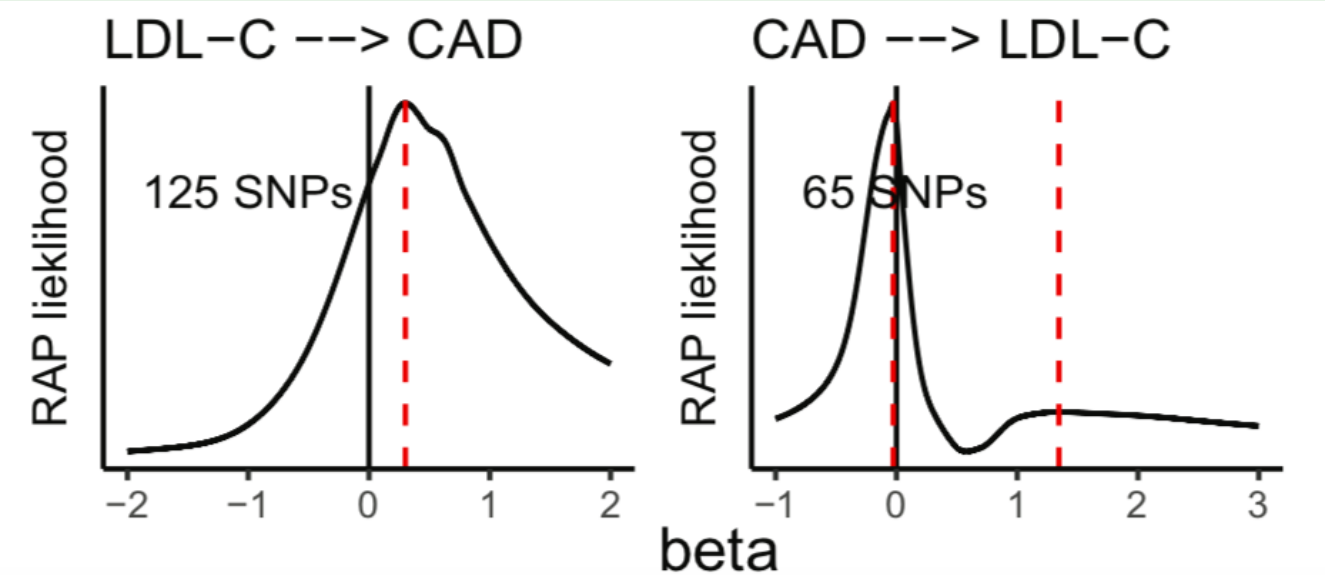
Identify causal direction

Heuristic

When reversing the role of exposure and outcome, the modal plot should show two modes:

- **A smaller one at** $1/\beta$ (SNPs associated with the true exposure);
- **A larger one at** 0 (all other genetic determinants of the true outcome).

Example: LDL-CAD



Summary

Design

I **Three-sample** MR: ~~**winner's curse**~~.

II **Genome-wide** MR: exploit weak instruments.

Model

I **Measurement error** in GWAS summary data: ~~**NOME assumption**~~.

II Both **systematic** and **idiosyncratic** pleiotropy.

Analysis

I **Robust adjusted profile score (RAPS)**: robust and efficient inference.

II Extension to **multivariate MR** and **sample overlap**.

Diagnostics

I **Q-Q plot** and **InSIDE plot**: falsify modeling assumptions.

II **Modal plot**: discover mechanistic heterogeneity.

Summary

Acknowledgement

- Jingshu Wang, Dylan Small, Nancy Zhang (University of Pennsylvania);
- Jack Bowden, Gib Hemani (MRC-IEU);
- Yang Chen (University of Michigan).

References

- Zhao et al. (2019+) Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *Annals of Statistics* (to appear).
- Zhao et al. (2019+) Powerful three-sample genome-wide design and robust statistical inference in summary-data Mendelian randomization. *International Journal of Epidemiology* (to appear).
- Wang et al. (2019) Estimating Causal Relationship for Complex Traits with Weak and Heterogeneous Genetic Effects. *Manuscript available upon request*.

- R package `mr.raps`: `https://github.com/qingyuanzhao/mr.raps`.
- `http://www-stat.wharton.upenn.edu/~qyzhao/MR.html`.

Case study: Friday 19th Session 20 (Data Challenge).