Using sparsity to overcome unmeasured confounding: Two parametric tales

Qingyuan Zhao

Statistical Laboratory, University of Cambridge

25 August, 2020 @ ICSB 2020

Slides and more information are available at
http://www.statslab.cam.ac.uk/~qz280/.

Let's face the dragon

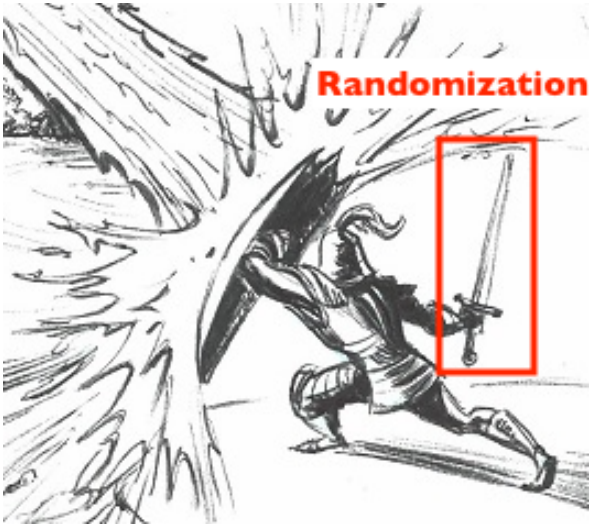
Image credit: Tony Bancroft.

Let's face the dragon

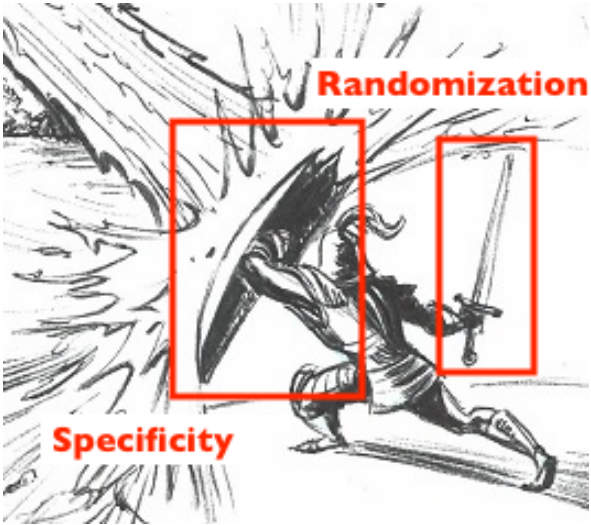
Image credit: Tony Bancroft.

Our two weapons

Our two weapons



Our two weapons



Credit of this idea

- Wang Miao (now at Peking University, China) told me about this idea during the Atlantic Causal Inference Conference (ACIC) in 2017.
- After being bombarded by machine learning talks for estimating heterogeneous treatment effect, he told me that he was going to talk about something different—specificity.

Bradford Hill's (1965) criteria for causality

1. Strength (effect size);
2. Consistency (reproducibility);
3. Specificity;
4. Temporality;
5. Biological gradient (dose-response relationship);
6. Plausibility (mechanism);
7. Coherence (between epidemiology and lab findings);
8. Experiment;
9. Analogy.

Bradford Hill's (1965) criteria for causality

1. Strength (effect size);
2. Consistency (reproducibility);
3. **Specificity;**
4. Temporality;
5. Biological gradient (dose-response relationship);
6. Plausibility (mechanism);
7. Coherence (between epidemiology and lab findings);
8. Experiment;
9. Analogy.

Hill's original specificity criterion

One reason, needless to say, is the specificity of the association. . . . If as here, **the association is limited to specific workers and to particular sites and types of disease** *and there is no association between the work and other modes of dying, then clearly that is a strong argument in favor of causation.*

- Now considered weak or irrelevant. Counter-example: smoking.
- In Hill's era, exposure = an occupational setting or a residential location (proxies for true exposures).
- Nowadays, exposure is much more precise.

This talk: Two parametric tales

Removing "batch effects" in multiple testing

- Wang, Zhao, Hastie, Owen (2017). Confounder adjustment in multiple hypothesis testing. *Annals of Statistics* 45(5).

This talk: Two parametric tales

Removing "batch effects" in multiple testing

- Wang, Zhao, Hastie, Owen (2017). Confounder adjustment in multiple hypothesis testing. *Annals of Statistics* 45(5).

Invalid instrumental variables in Mendelian randomization

- Zhao, Wang, Hemani, Bowden, Small (2020). Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *Annals of Statistics* 48(3).

This talk: Two parametric tales

Removing "batch effects" in multiple testing

- Wang, Zhao, Hastie, Owen (2017). Confounder adjustment in multiple hypothesis testing. *Annals of Statistics* 45(5).

Invalid instrumental variables in Mendelian randomization

- Zhao, Wang, Hemani, Bowden, Small (2020). Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *Annals of Statistics* 48(3).

Connection

The two share the same structure and are in some sense "dual" problems.

This talk: Two parametric tales

Removing "batch effects" in multiple testing

- Wang, Zhao, Hastie, Owen (2017). Confounder adjustment in multiple hypothesis testing. *Annals of Statistics* 45(5).

Invalid instrumental variables in Mendelian randomization

- Zhao, Wang, Hemani, Bowden, Small (2020). Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *Annals of Statistics* 48(3).

Connection

The two share the same structure and are in some sense "dual" problems.

Note: Wang Miao and Eric Tchetgen Tchetgen have done beautiful works on the nonparametric identification and semiparametric estimation using specificity.

First tale: Multiple testing in microarray data

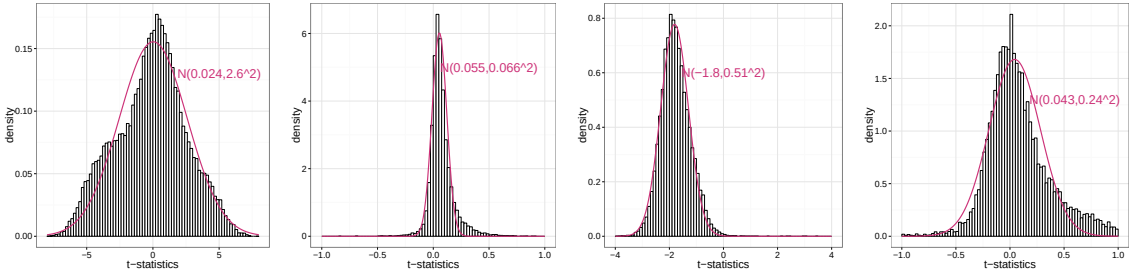


Figure: Empirical distribution of t-statistics for 4 microarray studies.

First tale: Batch effect

Table: Empirical distribution of the t-statistics

Dataset	Median	Median absolute deviation
1	0.024	2.6
2	0.055	0.066
3	-1.8	0.51
2 (adjusted for known batches)	0.043	0.24

- **Far from the "expected" null $\mathbf{N}(0,1)$ if true effect is sparse.**
- Most likely explanation: batch effect/unmeasured confounding.

Methods

Previous work

- Price et al. (2006) *Nat Gen*: Add principal components in GWAS.
- Leek and Storey (2008) *PNAS*: Surrogate variable analysis (SVA).
- Gagnon-Bartsch and Speed (2012) *Biostatistics*: Remove unwanted variation (RUV) using negative control genes.
- Sun, Zhang, Owen (2012) *AoAS*: Use sparsity to remove latent variable.

- A lot of great heuristics; Methods work well in some scenarios.
- However, modelling assumptions were unclear and the connections between the different methods were unexplored.
- Most surprisingly, nobody even called this problem "unmeasured confounding".

Statistical model

Notations

- X: treatment ($n \times 1$ vector).
- Y: outcome ($n \times p$ matrix). In this example, high-dimensional gene expressions.
- U: unobserved confounder ($n \times d$ matrix).
- Rows of X, Y, U are observations. Columns of Y are genes.

It turns out the everyone is (implicitly) using the following model:

$$Y = X\alpha^T + U\gamma^T + \text{noise},$$
$$U = X\beta^T + \text{noise}.$$

Therefore, ordinary least squares of Y vs. X estimate

$$\underset{p\times 1}{\Gamma} = \underset{p\times 1}{\alpha} + \underset{p\times d}{\gamma}\,\underset{d\times 1}{\beta}.$$

Identifiability problem

$$Y = X\alpha^T + U\gamma^T + \text{noise},$$
$$U = X\beta^T + \text{noise}.$$

Can be identified without (much) assumption

- OLS of $Y \sim X$:

$$\underset{p\times 1}{\Gamma} = \underset{p\times 1}{\alpha} + \underset{p\times d}{\gamma}\ \underset{d\times 1}{\beta}\ .$$

- Factor analysis on the residuals of $Y \sim X$ regression: γ.

Identifiability problem

$$Y = X\alpha^T + U\gamma^T + \text{noise},$$
$$U = X\beta^T + \text{noise}.$$

Can be identified without (much) assumption

- OLS of $Y \sim X$:

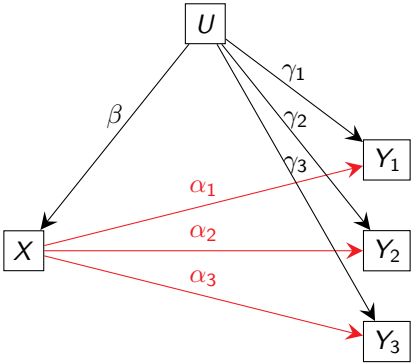
$$\underset{p\times 1}{\Gamma} = \underset{p\times 1}{\alpha} + \underset{p\times d}{\gamma}\ \underset{d\times 1}{\beta}\,.$$

- Factor analysis on the residuals of $Y \sim X$ regression: γ.

Specificity needed

- α and β cannot be immediately identified because there are more parameters $(p + d)$ than equations (p).
- Can be resolved by assuming α is "specific".

Diagram for CATE



Specificity

Some entries of α are zero (arrows are missing).

Specificity assumptions

$$\underset{p\times 1}{\Gamma} = \underset{p\times 1}{\alpha} + \underset{p\times d}{\gamma}\ \underset{d\times 1}{\beta}\ .$$

We can assume two kinds of specificity (either one is enough for identification):

Type 1: Negative control

At least d known entries of α are zero.

Type 2: Sparsity

Most entries of α are zero, though their positions are unknown.

The CATE procedure

Our procedure is called Confounder Adjusted Testing and Estimation (CATE).

$$\underset{p\times 1}{\Gamma} = \underset{p\times 1}{\alpha} + \underset{p\times d}{\gamma}\ \underset{d\times 1}{\beta}\,.$$

1. Obtain $\hat{\Gamma}$ by regressing Y on X;
2. Obtain $\hat{\gamma}$ by applying factor analysis on the residuals of $Y \sim X$ regression;

The CATE procedure

Our procedure is called Confounder Adjusted Testing and Estimation (CATE).

$$\underset{p\times 1}{\Gamma} = \underset{p\times 1}{\alpha} + \underset{p\times d}{\gamma}\ \underset{d\times 1}{\beta}\,.$$

1 Obtain $\hat{\Gamma}$ by regressing Y on X;

2 Obtain $\hat{\gamma}$ by applying factor analysis on the residuals of $Y \sim X$ regression;

3-1 With negative controls (say $\alpha_{1:k} = 0$), estimate β by regressing $\hat{\Gamma}_{1:k}$ on $\hat{\gamma}_{1:k}$.

The CATE procedure

Our procedure is called Confounder Adjusted Testing and Estimation (CATE).

$$\underset{p\times 1}{\Gamma} = \underset{p\times 1}{\alpha} + \underset{p\times d}{\gamma}\ \underset{d\times 1}{\beta}\ .$$

1 Obtain $\hat{\Gamma}$ by regressing Y on X;

2 Obtain $\hat{\gamma}$ by applying factor analysis on the residuals of $Y \sim X$ regression;

3-1 With negative controls (say $\alpha_{1:k} = 0$), estimate β by regressing $\hat{\Gamma}_{1:k}$ on $\hat{\gamma}_{1:k}$.

3-2 Or using sparsity, estimate β by regressing $\hat{\Gamma}$ on $\hat{\gamma}$ with robust loss function:

$$\hat{\beta} = \arg\min \sum_{j=1}^{p} \rho(\hat{\Gamma}_j - \hat{\gamma}_j^T \beta).$$

(Basically the same as putting lasso penalty on α).

The CATE procedure

Our procedure is called Confounder Adjusted Testing and Estimation (CATE).

$$\underset{p\times 1}{\Gamma} = \underset{p\times 1}{\alpha} + \underset{p\times d}{\gamma}\,\underset{d\times 1}{\beta}\,.$$

1. Obtain $\hat{\Gamma}$ by regressing Y on X;
2. Obtain $\hat{\gamma}$ by applying factor analysis on the residuals of $Y \sim X$ regression;

3-1. With negative controls (say $\alpha_{1:k} = 0$), estimate β by regressing $\hat{\Gamma}_{1:k}$ on $\hat{\gamma}_{1:k}$.

3-2. Or using sparsity, estimate β by regressing $\hat{\Gamma}$ on $\hat{\gamma}$ with robust loss function:

$$\hat{\beta} = \arg\min \sum_{j=1}^{p} \rho(\hat{\Gamma}_j - \hat{\gamma}_j^T \beta).$$

(Basically the same as putting lasso penalty on α).

4. Estimate α by $\hat{\alpha} = \hat{\Gamma} - \hat{\gamma}\hat{\beta}$.

Theory for CATE

Our paper derived an asymptotic theory for CATE (distribution of $\hat{\beta}$ and $\hat{\alpha}$, optimally, etc.)

Theory for CATE

Our paper derived an asymptotic theory for CATE (distribution of $\hat{\beta}$ and $\hat{\alpha}$, optimally, etc.)

Key assumptions

1. Factors are strong enough: $\|\gamma\|_F^2 = \Theta(p)$.
   - Recall γ is $p \times d$ matrix of the effect of confounders on gene expressions.
   - In real data: often a small number of strong factors + many weak factors.

Theory for CATE

Our paper derived an asymptotic theory for CATE (distribution of $\hat{\beta}$ and $\hat{\alpha}$, optimally, etc.)

Key assumptions

1. Factors are strong enough: $\|\gamma\|_F^2 = \Theta(p)$.
   - Recall γ is $p \times d$ matrix of the effect of confounders on gene expressions.
   - In real data: often a small number of strong factors + many weak factors.
2. In the sparsity scenario, α is quite sparse: $\|\alpha\|_1 \sqrt{n}/p \to 0$.
   - After working on the dual problem—MR, now I think this rate may be too stringent.

Theory for CATE

Our paper derived an asymptotic theory for CATE (distribution of $\hat{\beta}$ and $\hat{\alpha}$, optimally, etc.)

Key assumptions

1. Factors are strong enough: $\|\gamma\|_F^2 = \Theta(p)$.
   - Recall γ is $p \times d$ matrix of the effect of confounders on gene expressions.
   - In real data: often a small number of strong factors + many weak factors.
2. In the sparsity scenario, α is quite sparse: $\|\alpha\|_1 \sqrt{n}/p \to 0$.
   - After working on the dual problem—MR, now I think this rate may be too stringent.

Highlight of the theory

Under these two (perhaps unrealistic) assumptions, CATE may be **as efficient as the oracle** OLS estimator that observes Z!

Theory for CATE

Our paper derived an asymptotic theory for CATE (distribution of $\hat{\beta}$ and $\hat{\alpha}$, optimally, etc.)

Key assumptions

1. Factors are strong enough: $\|\gamma\|_F^2 = \Theta(p)$.
   - Recall γ is $p \times d$ matrix of the effect of confounders on gene expressions.
   - In real data: often a small number of strong factors + many weak factors.
2. In the sparsity scenario, α is quite sparse: $\|\alpha\|_1 \sqrt{n}/p \to 0$.
   - After working on the dual problem—MR, now I think this rate may be too stringent.

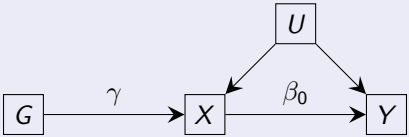
Highlight of the theory

Under these two (perhaps unrealistic) assumptions, CATE may be **as efficient as the oracle** OLS estimator that observes Z!

- Simulations show that CATE (with some tweaks) perform quite well in some scenarios when these assumptions are not satisfied.

Second tale: Mendelian randomization with invalid IVs

Diagram for Mendelian randomization (MR)

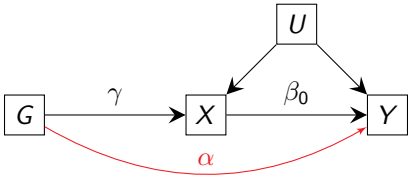


- G: Genetic variant as instrumental variable (IV);
- X: Epidemiological exposure (eg LDL-cholesterol);
- Y: Disease outcome (eg coronary heart disease);
- U: Unmeasured confounder.

Basic idea:

$$\underbrace{\textbf{Causal effect of } X \textbf{ on } Y \ (\beta_0)}_{\textbf{CONTROLLED experiment}} = \frac{\textbf{Effect of } Z \textbf{ on } Y \ (\Gamma = \gamma \cdot \beta_0)}{\underbrace{\textbf{Effect of } Z \textbf{ on } X \ (\gamma)}_{\textbf{NATURAL experiment}}}.$$

Invalid IV due to pleiotropy



- Pleiotropy: multiple functions of genes.
- Example: LDL-variant may also increase BMI.
- Invalid IV is the main challenge in designing an MR study.

Solutions to the invalid IV problem

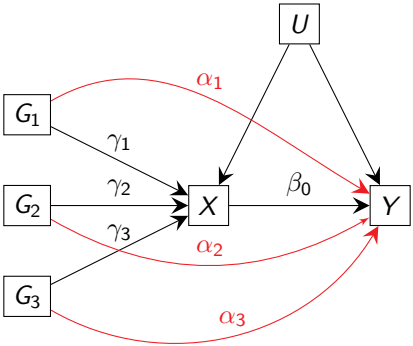
There are two main approaches (both requiring collecting many genetic IVs):

1. Assuming invalid IVs are **sparse**.
   - Kang et al., 2016, *JASA*.
2. InSIDE assumption: instrument strength (γ) independent of direct effect (α)
   - Bowden, Davey Smith, Burgess, 2015, *IJE*;
   - Kolesár et al., 2015, *JBES*.

MR.RAPS (Robust Adjusted Profile Score)

- A framework we developed that can accommodate both types of invalid instruments.
- I will focus on **sparse invalid IVs** today.

Diagram



Specificity

Some entries of α are zero (arrows are missing).

Correspondence between the two problems

Same problem structure

$$\underset{p\times 1}{\Gamma} = \underset{p\times 1}{\alpha} + \underset{p\times d}{\gamma}\ \underset{d\times 1}{\beta}\ .$$

Parameter	In batch-effect removal	In MR with invalid IV
α	**Effect of interest**	Direct effect of IV
β	Confounder effect on treatment	**Effect of interest**
γ	**Confounder effect on outcome**	**Effect of IV on exposure**
Γ	**Observed treatment effect**	**Effect of IV on outcome**

- In both problems, estimates of γ and Γ are immediately available.
- In both problems, specificity/sparsity of α is needed for identification.

MR.RAPS: A comprehensive framework

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.

Rest of the talk

Won't have time to discuss all of them...

Two focal points

1. Weak instrument asymptotics.
2. How MR.RAPS handles invalid IVs;

Focal point 1: Weak instrument asymptotics

Stylized statistical problem

We observe (p is the number of genetic instruments)

$$\begin{pmatrix} \hat{\gamma} \\ \hat{\Gamma} \end{pmatrix} \sim \mathrm{N}\Big(\begin{pmatrix} \gamma \\ \Gamma \end{pmatrix}, \frac{1}{n} I_{2p} \Big),$$

where most entries of the direct effect $\underset{p\times 1}{\alpha} = \underset{p\times 1}{\Gamma} - \beta \underset{p\times 1}{\gamma}$ are 0.

Focal point 1: Weak instrument asymptotics

Stylized statistical problem

We observe (p is the number of genetic instruments)

$$\begin{pmatrix} \hat{\gamma} \\ \hat{\Gamma} \end{pmatrix} \sim \mathrm{N}\Big(\begin{pmatrix} \gamma \\ \Gamma \end{pmatrix}, \frac{1}{n} I_{2p} \Big),$$

where most entries of the direct effect $\underset{p\times 1}{\alpha} = \underset{p\times 1}{\Gamma} - \beta \underset{p\times 1}{\gamma}$ are 0.

- Profile likelihood (different from a simple OLS):

$$l(\beta) = \max_{\gamma} l(\beta, \gamma) = -\frac{1}{2} \sum_{j=1}^{p} \frac{(\hat{\Gamma}_j - \beta \hat{\gamma}_j)^2}{1 + \beta^2}.$$

Focal point 1: Weak instrument asymptotics

Stylized statistical problem

We observe (p is the number of genetic instruments)

$$\begin{pmatrix} \hat{\gamma} \\ \hat{\Gamma} \end{pmatrix} \sim \mathrm{N}\Big(\begin{pmatrix} \gamma \\ \Gamma \end{pmatrix}, \frac{1}{n} I_{2p} \Big),$$

where most entries of the direct effect $\underset{p\times 1}{\alpha} = \underset{p\times 1}{\Gamma} - \beta \underset{p\times 1}{\gamma}$ are 0.

- Profile likelihood (different from a simple OLS):

$$l(\beta) = \max_{\gamma} l(\beta, \gamma) = -\frac{1}{2} \sum_{j=1}^{p} \frac{(\hat{\Gamma}_j - \beta \hat{\gamma}_j)^2}{1 + \beta^2}.$$

- Assuming $\alpha = 0$, the maximum likelihood estimator $\hat{\beta}$ converges to

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathrm{N}\Big(0, (1+\beta^2) \frac{\|\gamma\|^2 + p/n}{\|\gamma\|^4}\Big).$$

- Classical asymptotics: $\|\gamma\|^2$ fixed, **p fixed**, $n \to \infty$.

Focal point 2: Robust adjusted profile score (RAPS)

Profile score ($= \partial/\partial\beta$ profile likelihood) equation

It is illuminating to examine

$$\sum_{j=1}^{p} \hat{\gamma}_{j,\mathrm{MLE}}(\boldsymbol{\beta}) \cdot \hat{\boldsymbol{\alpha}}_{\boldsymbol{j}}(\boldsymbol{\beta}) = 0, \text{ where}$$

- $\hat{\gamma}_{j,\mathrm{MLE}}(\boldsymbol{\beta}) = (\hat{\gamma}_j + \beta\hat{\Gamma}_j)/(1+\beta^2)$ estimates IV strength;
- $\hat{\boldsymbol{\alpha}}_{\boldsymbol{j}}(\boldsymbol{\beta}) = (\hat{\Gamma}_j - \beta\hat{\gamma}_j)/\sqrt{(1+\beta^2)/n}$ estimates direct effect (standardized).

Focal point 2: Robust adjusted profile score (RAPS)

Profile score ($= \partial/\partial\beta$ profile likelihood) equation

It is illuminating to examine

$$\sum_{j=1}^{p} \hat{\gamma}_{j,\mathrm{MLE}}(\beta) \cdot \hat{\alpha}_j(\beta) = 0, \text{ where}$$

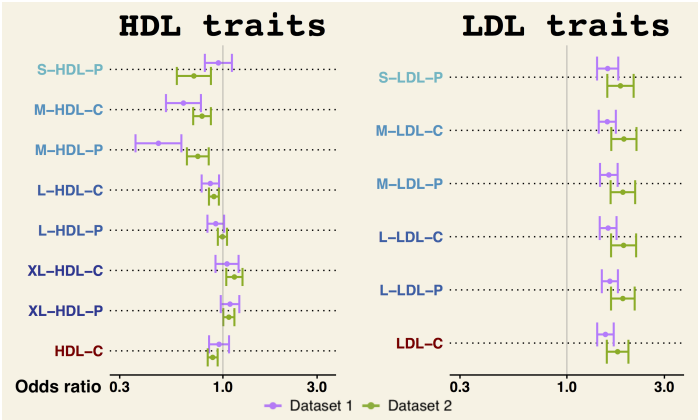
- $\hat{\gamma}_{j,\mathrm{MLE}}(\beta) = (\hat{\gamma}_j + \beta\hat{\Gamma}_j)/(1+\beta^2)$ estimates IV strength;
- $\hat{\alpha}_j(\beta) = (\hat{\Gamma}_j - \beta\hat{\gamma}_j)/\sqrt{(1+\beta^2)/n}$ estimates direct effect (standardized).

Two innovations in MR.RAPS

$$\sum_{j=1}^{p} f(\hat{\gamma}_{j,\mathrm{MLE}}(\beta)) \cdot \psi(\hat{\alpha}_j(\beta)) = 0.$$

- f function: Selectively shrink IV strength estimates (increases efficiency).
- ψ function: Bounded function (robust to large direct effect α).

New MR results



- Exposures: **Lipoprotein subfractions**; Outcome: **Coronary heart disease**.
- Main finding: **Heterogeneous effect of HDL subfractions** across different partial size.
- Estimates much more precise than IVW, MR-Egger, weighted median, ...
- More detail: bioRxiv:691089.

Wrap up

Two problems, same structure

1. CATE: Remove batch effects in multiple testing;
2. MR.RAPS: Tackling invalid IVs in Mendelian randomization.

Wrap up

Two problems, same structure

1. CATE: Remove batch effects in multiple testing;
2. MR.RAPS: Tackling invalid IVs in Mendelian randomization.

Main messages

- Randomization and Specificity are our two (only?) weapons against the dragon (unmeasured confounding).
- High-dimensional data present challenges as well as opportunities:
  1. Possibility to learn the structure of unmeasured confounding;
  2. Sparsity as "unspecified specificity" for causal inference.

Wrap up

Software

- R package *cate* available on CRAN.
- R package *mr.raps* on `github.com/qingyuanzhao`.
- More information about MR.RAPS can be found at `http://www.statslab.cam.ac.uk/~qz280/project/iv-mr/`.

Wrap up

Software

- R package *cate* available on CRAN.
- R package *mr.raps* on `github.com/qingyuanzhao`.
- More information about MR.RAPS can be found at `http://www.statslab.cam.ac.uk/~qz280/project/iv-mr/`.

Acknowledgement

- Collaborators on CATE: Jingshu Wang, Trevor Hastie, Art B Owen; Yang Song (applications to financial data).
- Collaborators on MR.RAPS: Jingshu Wang, Dylan S Small, Jack Bowden, Yang Chen, Gibran Hemani, George Davey Smith, Nancy R Zhang, Daniel J Rader, Sean Hennessy.

Wrap up

Software

- R package *cate* available on CRAN.
- R package *mr.raps* on `github.com/qingyuanzhao`.
- More information about MR.RAPS can be found at `http://www.statslab.cam.ac.uk/~qz280/project/iv-mr/`.

Acknowledgement

- Collaborators on CATE: Jingshu Wang, Trevor Hastie, Art B Owen; Yang Song (applications to financial data).
- Collaborators on MR.RAPS: Jingshu Wang, Dylan S Small, Jack Bowden, Yang Chen, Gibran Hemani, George Davey Smith, Nancy R Zhang, Daniel J Rader, Sean Hennessy.

Thank you!!