Using sparsity to overcome unmeasured confounding: Two examples

Qingyuan Zhao

Statistical Laboratory, University of Cambridge

October 15, 2019 @ MRC-BSU Seminar

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

About me

- New University Lecturer in the Stats Lab (in West Cambridge).
- PhD (2011-2016) in Statistics from Stanford, advised by Trevor Hastie.
- Postdoc (2016-2019) at University of Pennsylvania, advised by Dylan Small and Sean Hennessy.
- Current research area: Causal Inference.
- Interested applications: public health, genetics, social sciences, computer science.

Growing interest in causal inference

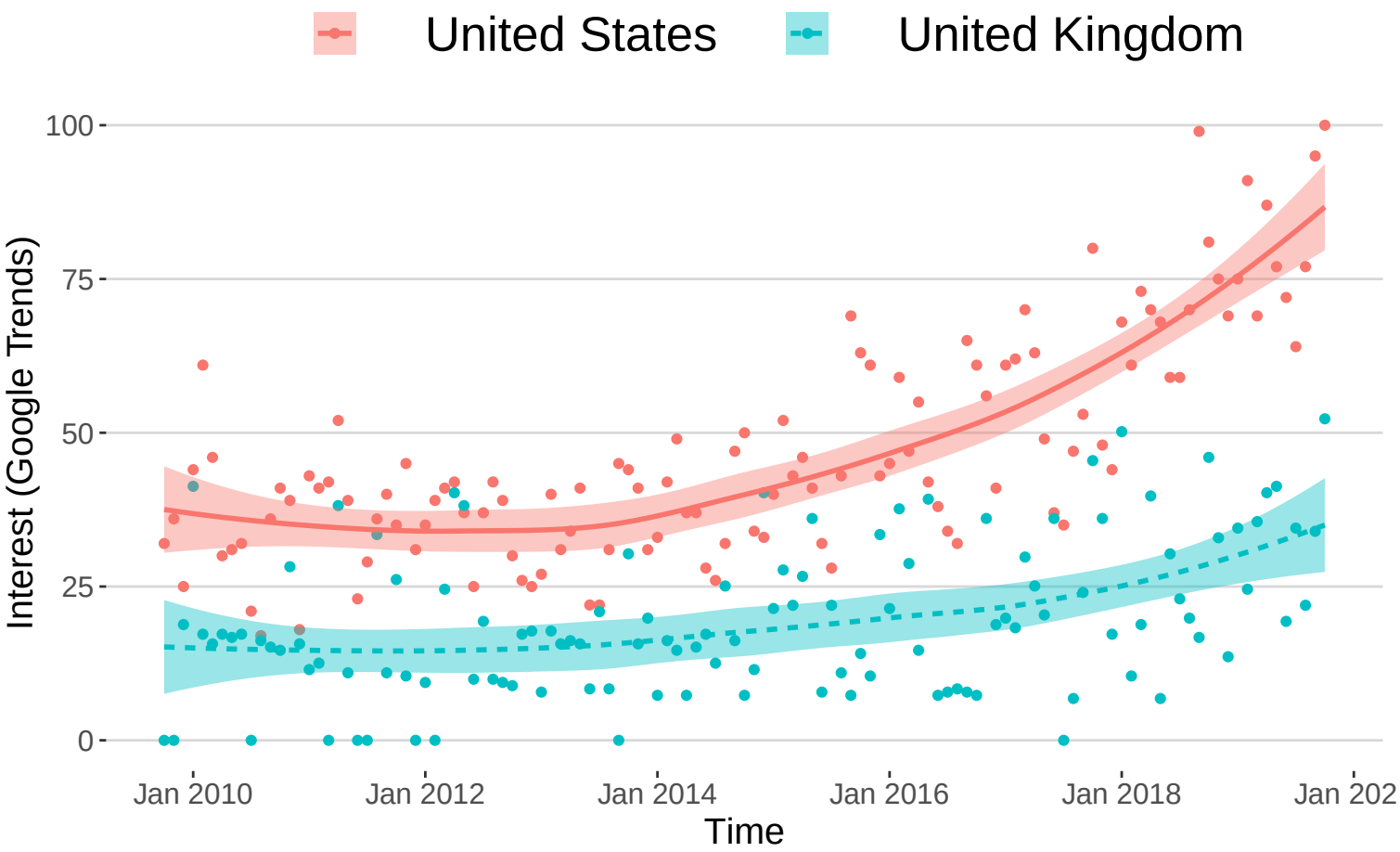


Figure: Data from Google Trends.

Old and new problems

- Epidemiology and public health: effectiveness of prevention/treatment, causal effect of risk factors, etc.
- Quantitative social sciences: evaluation of social programs, policy impact, etc.
- Precision medicine.
- Massive online experiments.
- Fairness of machine learning algorithms.
- Big Data $\neq$ better inference.

Causal inference in Cambridge

In Stats Lab

- A new 16-lecture Part III course in the Michaelmas term (Tuesday & Thursday 12-1).
- A new reading group (`http://talks.cam.ac.uk/show/index/105688`).

In BSU and the Clinical School

I would like to learn more!!

Cross schools?

Causal inference research requires inter-disciplinary collaboration.

Back to the main topic

Bradford Hill (1965) criteria

1. Strength (effect size);
2. Consistency (reproducibility);
3. Specificity;**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: Specificity

More precisely: How specificity/sparsity assumptions can help us overcome unmeasured confounding.

Growing awareness

- **Development in high-dimensional statistics:** multiple testing, lasso and sparsity, model selection, ...
- Growing interest in **using negative controls for causal inference**.
- **Biological mechanisms are often specific** (or more specific as we go more micro).

Two examples

Removing "batch effects" in multiple testing

A framework called Confounder Adjusted Testing and Estimation (CATE), proposed in

- Wang*, Zhao*, Hastie, Owen (2017) *Annals of Statistics*.

Invalid instrumental variables in Mendelian randomization

A class of methods called Robust Adjusted Profile Score (RAPS), proposed in

- Zhao, Wang, Hemani, Bowden, Small (2019+) *Annals of Statistics*.
- Zhao, Chen, Wang, Small (2019) *International Journal of Epidemiology*.

Connection

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

Batch effect: Motivating example

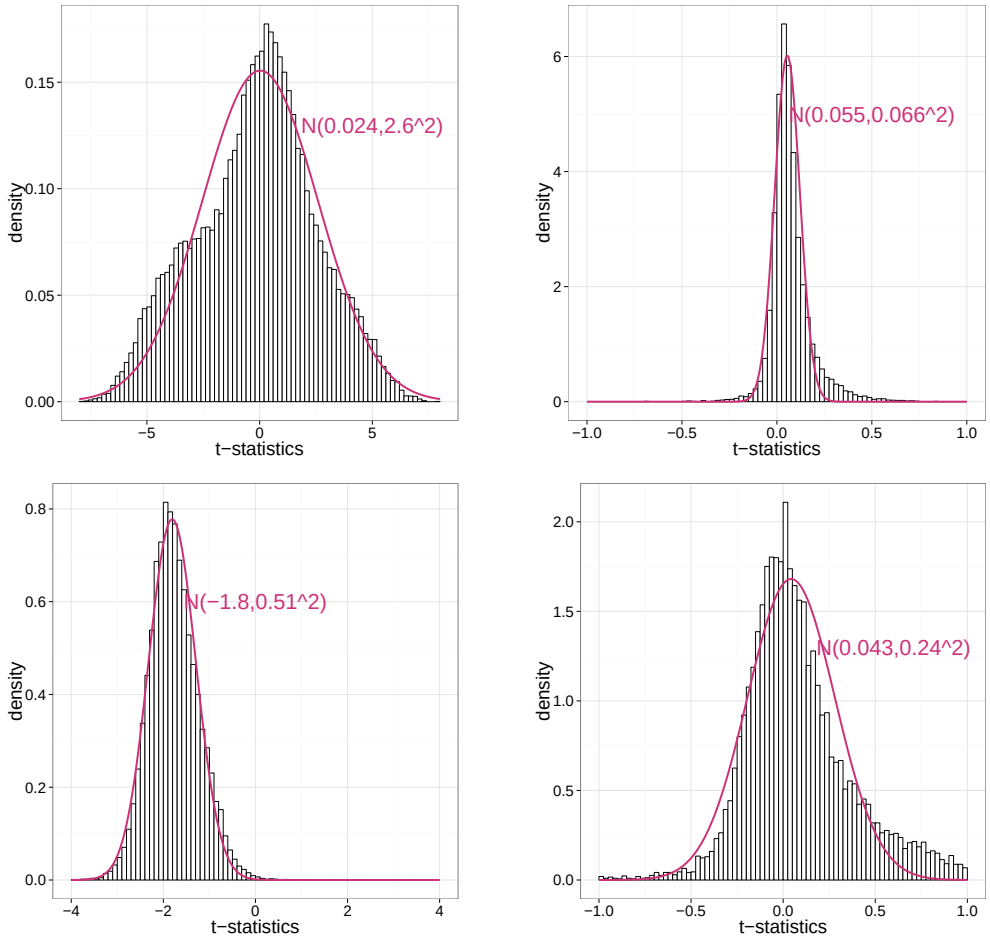


Figure: Empirical distribution of t-statistics for microarray datasets.

Motivating example

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.
- Modelling assumptions were unclear, basically no theory.
- Connections between the methods were unexplored.
- Probably most importantly (and surprisingly), nobody 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

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

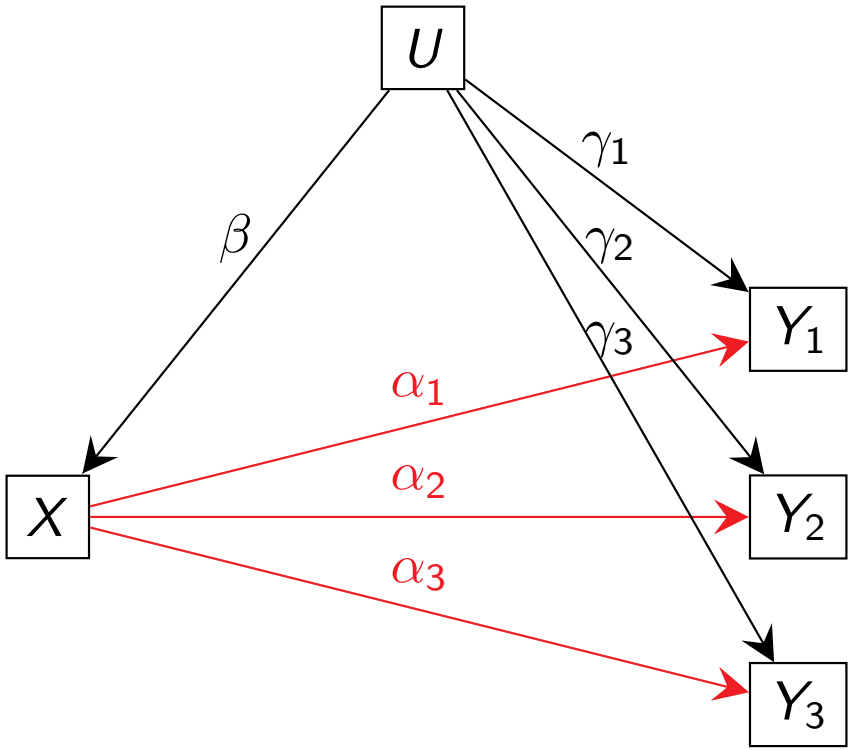
Can be identified without (much) assumption

- 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.

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):

Negative control

At least d known entries of α are zero.

Sparsity

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

The CATE procedure

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

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.)

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.

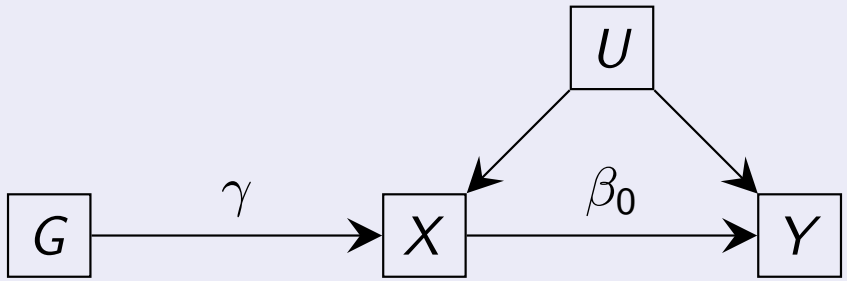
Highlight of the theory

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 even when these assumptions are not satisfied.

Second problem: Mendelian randomization with invalid IVs

Diagram for IV

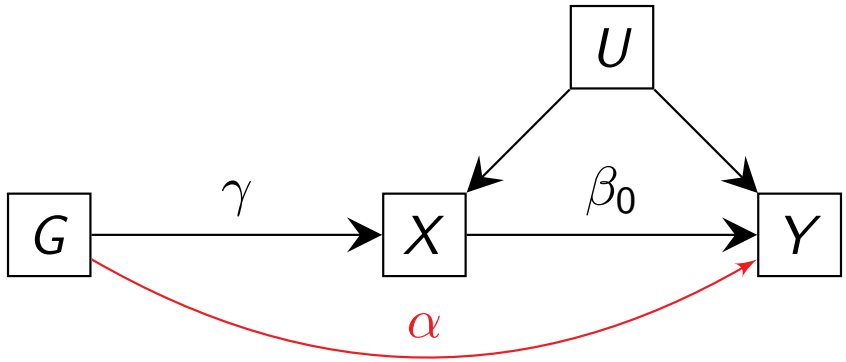


- G: Genetic variant as instrumental variable (IV);
- X: Epidemiological exposure (eg LDL-cholesterol);
- Y: Disease outcome (eg coronary heart disease);
- U: Unmeasured confounder.

Basic idea:

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

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

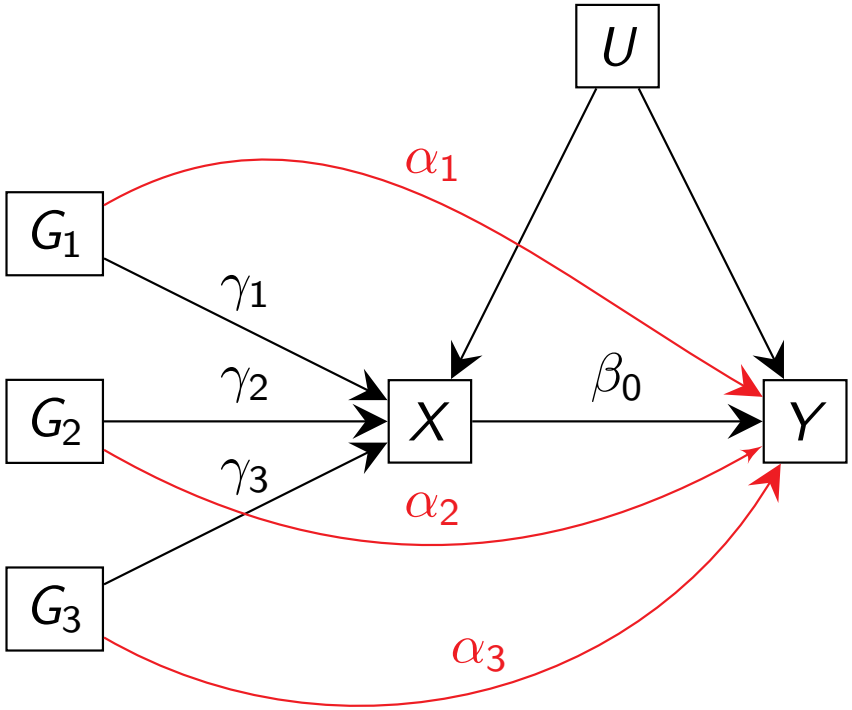
There are two main approaches (both requiring collecting many genetic IVs):

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)

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

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):

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

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

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

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

Related problem: Gene colocalization test

Stylized statistical problem

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

- MR is closely related to the problem of gene colocalization.
- In MR, the goal is to estimate β.
- In colocalization, the proportionality testing approach asks **if the above model fits the data for any** β (Wallace et al., 2012, *Hum Mol Genet*).
- A standard test uses (Plagnol et al, 2009, *Biostatistics*)

$$-2l(\hat{\beta}) \xrightarrow{d} \chi^2_{p-1} \text{ under the above model.}$$

- The factor $\dfrac{\|\gamma\|^2 + p/n}{\|\gamma\|^4}$ we obtained in weak IV asymptotics suggests that this approximation (based on Wilks' theorem) is only accurate if $\|\gamma\|^2 \gg p/n$.

Focal point 2: Robust adjusted profile score (RAPS)

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

It is illuminating to examine

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

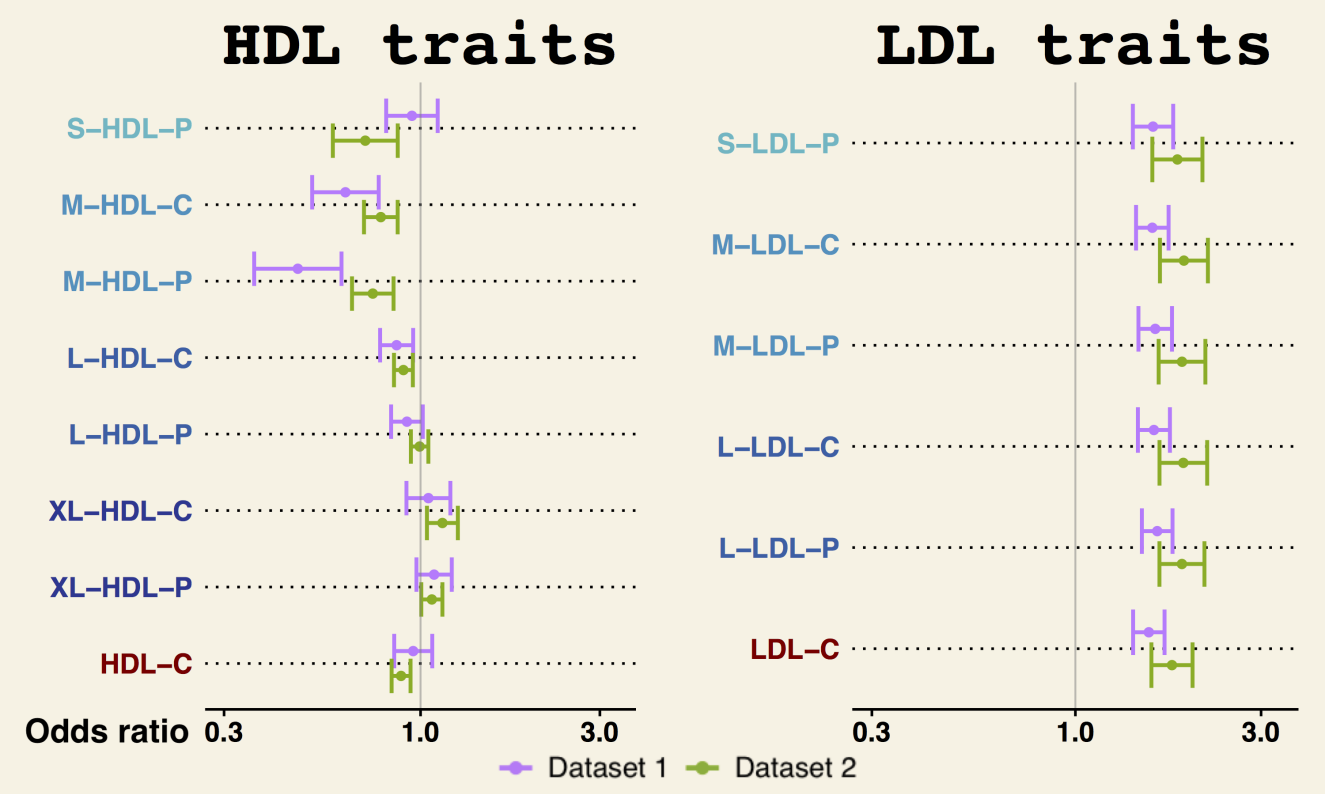
- $\hat{\gamma}_{j,\mathrm{MLE}}(\beta) = (\hat{\gamma}_j + \beta\hat{\Gamma}_j)/(1+\beta^2)$ estimates IV strength;
- $\hat{\alpha}_j(\beta) = (\hat{\Gamma}_j - \beta\hat{\gamma}_j)/\sqrt{(1+\beta^2)/n}$ estimates direct effect (standardized).

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.

Main messages

- Specificity/sparsity offers a way to overcome unmeasured confounding.
- High-dimensional data present challenges as well as opportunities:
  1. Learning the structure of unmeasured confounding;
  2. Selecting the invalid instrumental variables.

Future work

- Applying new statistical techniques learned in MR.RAPS to CATE.
- A more general statistical method for structural equation problems with specificity constraints?

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/MR.html`.

Acknowledgement

- Collaborators on CATE: Jingshu Wang, Trevor Hastie, Art B Owen; Yang Song (application in 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!!