Orienting the causal relation between two variables

Wei Pan

Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455

Joint work with Haoran Xue, ...

Feb 26, 2019

at Dept of Biostatistics, U of Michigan
Outline

- Introduction: what and why?
- Existing methods
- New method
- Examples
- Future work
Question: causal direction between $X$ and $Y$?
$X \implies Y$, or $Y \implies X$?

Example: LDL $\implies$ CAD?
Statins; Mendelian randomization (MR) analyses.

Example: HDL $\implies$ CAD?
Failed drug trials; MR analyses: inconclusive.

Example: Education level $\implies$ AD?
A *Lancet* Commission (Livingston et al 2017): possible to prevent about 35% of dementia by controlling nine risk factors: education to a maximum of age 11-12 years, midlife hypertension, midlife obesity, hearing loss, late-life depression, diabetes, physical inactivity, smoking, and social isolation.
Introduction

- Imaging-wide association studies (IWAS) (Xu et al, 2017, *NeuroImage*).
  Atrophy in hippocampus $\implies$ AD?

  Main idea: If 1) GE sim SNPs ; 2) $D \sim \hat{GE}$ then
  GE $\implies$ D.
  Q: possible $D \implies$ GE?


- MR: assuming $X \implies Y$.
  $X$: exposure; $Y$: outcome.

- A tough and rough Q ...
Some existing methods

- Two **continuous** random variables (Peters et al 2017; Jiao et al 2018, *Front Genet*):
  \[ P_{X,Y} = P_{Y|X}P_X = P_{X|Y}P_Y \]

- If \( Y = f(X) + E_Y, X \perp E_Y, f \) is **nonlinear** (and ...), then possible!

- A simple example: \( Y = X^3 + e_Y \),

Approach: fit a nonlinear model; test the indep.
Genetics-based methods

- *Using genetic data to strengthen causal inference* ... (Pingault et al 2018, *Nat Rev Genet*).
- Use SNPs as anchors/instrumental variables (IVs) (Schadt et al 2005, *Nat Genet*; Chen et al 2007, *Genom Biol*; ...). SNPs $\implies ...$; not the reverse!
  SNPs: somewhat randomized.
  take advantage of many existing large-scale GWAS!

- Mediation analysis:
  Causal inference test (CIT) (Millstein et al 2009, *BMC Genet*)
  Limitations: 1) require data (SNP, $X$, $Y$);
  often have two samples: (SNP, $X$), (SNP, $Y$).
  2) less robust to measurement errors.

- MR: Steiger’s test (Hemani et al 2017, *PLOS Genet*)
  Theory: If SNP $\implies X \implies Y$, then $\rho_{gX} > \rho_{gY}$!
  Main idea: test their difference!
  Limitation: based on a single SNP, thus low statistical efficiency and low robustness! —–our task here!

- Others: Pickrell’s (2016, *Nat Genet*); bi-directional MR ...
Motivation: extending MR Steiger’s method from using a single SNP to multiple SNPs.
1) multiple correlated SNPs in a locus;
2) multiple independent loci.

Theory: If SNP $\implies X \implies Y$, then $\rho_{Yg} = \rho_{Xg} \rho_{YX}$.
\[
\left| \frac{\rho_{Yg}}{\rho_{Xg}} \right| = |\rho_{YX}| := K < 1,
\]
independent of $g$.
Similarly, if SNP $\implies Y \implies X$, then ...

Limitation: cannot distinguish $X \iff SNP \implies Y$

Main idea: combining multiple estimates $r_{Yg}/r_{Xg}$ across $g$’s...
1) one locus: GLSE;
2) multi-loci: IVW (meta-analysis).
Our method: set-up

- Given two indep samples of GWAS summary data:
  \((g, \hat{\beta}_X, se(\hat{\beta}_X)), n_X;\)
  \((g, \hat{\beta}_Y, se(\hat{\beta}_Y)), n_Y;\)

- Calculate sample correlations:

\[
rx_g = \frac{\hat{\beta}_X}{\sqrt{\hat{\beta}_X^2 + (n_X - 2) \cdot se(\hat{\beta}_X)}}
\]

- With SNPs \(g_1,\ldots,g_m\), denote \(rx_g = (rx_{g_1},\ldots,rx_{g_m})^T\), ... By Neudecker and Wesselman (1990),

\[
\sqrt{n_X} \cdot (rx_g - \rho_X) \rightarrow_d N(0, V_X),
\]

...
Our method: combining over multiple SNPs in a locus

- By the delta method,

\[ \frac{r_{Xg}}{r_{Yg}} \rightarrow_d N \left( \frac{\rho_{Xg}}{\rho_{Yg}}, V \right), \]

where \( V \) can be estimated...

- The GLSE is

\[ \hat{K}_{GLS} = 1^T \cdot V^{-1} \cdot \frac{r_{Xg}}{r_{Yg}} / 1^T \cdot V^{-1} \cdot 1. \]

where \( 1 = (1, \ldots, 1)^T \), and the variance of \( \hat{K}_{GLS} \) is

\[ \text{var}(\hat{K}_{GLS}) = 1/1^T \cdot V^{-1} \cdot 1. \]
Our method: combining over multiple loci

For each of \( k \) loci, we have \( \hat{K}_i^{GLS} \) and \( \text{var}(\hat{K}_i^{GLS}) \); by IVW

\[
\hat{K}_{IVW} = \frac{\sum_{i=1}^{k} \frac{\hat{K}_i^{GLS}}{\text{var}(\hat{K}_i^{GLS})}}{\sum_{i=1}^{k} \frac{1}{\text{var}(\hat{K}_i^{GLS})}},
\]

\[
\text{var}(\hat{K}_{IVW}) = \frac{1}{\sum_{i=1}^{k} \frac{1}{\text{var}(\hat{K}_i^{GLS})}}.
\]

So, we can construct a CI for each significant SNP, locus and all loci, respectively.

Remark: Fieller’s Theorem ...
Example: LDL/HDL vs CAD


- Partition the genome into 1703 (approximately) independent loci (Berisa and Pickrell 2016, *Bioinformatics*).

- Consider 8 (or 4) significant indep loci for both LDL (or HDL) and CAD.

- In each locus, pruned out highly correlated SNPs with $|r| > 0.8$. 
LDL vs CAD: Locus 1

LDL/CAD $5 \times 10^{-6}$

CAD/LDL $5 \times 10^{-6}$

Figure: LDL vs CAD locus 1.
LDL vs CAD: Locus 6
LDL vs CAD: all 8 loci

![Graph showing LDL vs CAD in all 8 loci](image-url)
LDL vs CAD: 7 loci

Figure: LDL vs CAD in 7 loci.
HDL vs CAD: all loci

**HDL/CAD, 5e−6**

**CAD/HDL, 5e−6**
Alternatives:

- **MR**: using SNPs as IVs. IV assumptions:

  - With a valid IV: \( \beta_{YX} = \beta_{Yg}/\beta_{Xg} \).
  - Wald ratio; IVW (meta-analysis for multiple indep SNPs).
  \[ \hat{\beta}_{Yg} = \beta_{YX}\hat{\beta}_{Xg} + \epsilon_g; \text{IVW, PS (Zhao et al 2018)} \]
  \[ \hat{\beta}_{Yg} = \beta_0 + \beta_{YX}\hat{\beta}_{Xg} + \epsilon_g; \text{Egger reg} \]
  \[ \hat{\beta}_{Yg} = \beta_{0g} + \beta_{YX}\hat{\beta}_{Xg} + \epsilon_g, \beta_{0g} \sim iid N(0, \tau^2); \text{APS/RAPS} \]

- **Bi-d MR** (Timpson et al 2011; Burgess et al 2015):
  1) use SNPs/IVs \( \Rightarrow X \Rightarrow Y \);
  2) use SNPs/IVs \( \Rightarrow Y \Rightarrow X \);
  Then the significant one gives the direction.

- **Pickrell’s method**: check \( \hat{\beta}_{Yg} \propto \hat{\beta}_{Xg} \). …

- Intuitive and reasonable!
Alternatives:

- But ...
- A special and ideal case: if i) SNPs \( \rightarrow X \rightarrow Y \); ii) no other pathway to \( Y \); iii) \( n \) large enough, then bi-d MR: significant in **both** directions! Pickrell’s: proportional in **both** directions!
- Might work if there exist other pathways to \( Y \)
- More Generally, need **two** sets of valid IVs; how to choose?
### Example: LDL vs CAD

**Table: Bi-directional MR vs our method.**

<table>
<thead>
<tr>
<th>SNPs as IVs</th>
<th>bi-d MR, p-val</th>
<th>Ours, CI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IVW</td>
<td>Egger reg</td>
</tr>
<tr>
<td>LDL 5E-8 (39)</td>
<td>9.0E-10</td>
<td>9.0E-5</td>
</tr>
<tr>
<td>LDL 5E-6 (65)</td>
<td>1.9E-12</td>
<td>1.3E-6</td>
</tr>
<tr>
<td>CAD 5E-8 (14)</td>
<td>0.306</td>
<td>0.898</td>
</tr>
<tr>
<td>CAD 5E-6 (40)</td>
<td>0.0660</td>
<td>0.562</td>
</tr>
<tr>
<td>Both LDL 5E-6 (7)</td>
<td>6.4E-6</td>
<td>0.421</td>
</tr>
<tr>
<td>Both CAD 5E-6 (7)</td>
<td>2.8E-6</td>
<td>0.421</td>
</tr>
</tbody>
</table>
Example: LDL vs CAD, by weighted median and weighted mode methods

<table>
<thead>
<tr>
<th>SNPs as IVs</th>
<th>median</th>
<th>mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDL 5E-8 (39)</td>
<td>0.452 (0.073), 6.55E-10</td>
<td>0.475 (0.079), 5.73E-7</td>
</tr>
<tr>
<td>LDL 5E-6 (65)</td>
<td>0.450 (0.070), 1.57E-10</td>
<td>0.480 (0.074), 1.43E-8</td>
</tr>
<tr>
<td>CAD 5E-8 (14)</td>
<td>-0.049 (0.021), 0.016</td>
<td>-0.052 (0.0196), 0.0195</td>
</tr>
<tr>
<td>CAD 5E-6 (40)</td>
<td>-0.034 (0.016), 0.030</td>
<td>-0.040 (0.017), 0.026</td>
</tr>
<tr>
<td>Both LDL 5E-6 (7)</td>
<td>0.637 (0.093), 9.03E-12</td>
<td>0.665 (0.085), 2.2E-4</td>
</tr>
<tr>
<td>Both CAD 5E-6 (7)</td>
<td>0.581 (0.059), 4.84E-23</td>
<td>0.630 (0.112), 2.0E-8</td>
</tr>
</tbody>
</table>

Table: Weighted median and mode methods to account for possible direct/pleiotropic effects. In each cell: $\hat{\beta}_{XY}$ or $\hat{\beta}_{YX}$ (SE), p-value.
Table: Methods (PS, APS, RAPS) of Zhao et al. (2018) with the squared error loss or Tukey’s loss, and possibly with random effects (RE) to model direct/pleiotropic effects.

<table>
<thead>
<tr>
<th>SNPs as IVs</th>
<th>L2</th>
<th>Tukey</th>
<th>L2+RE</th>
<th>Tukey+RE</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDL 5E-8 (39)</td>
<td>1.4E-34</td>
<td>2.7E-28</td>
<td>1.4E-34</td>
<td>5.9E-14</td>
</tr>
<tr>
<td>LDL 5E-6 (65)</td>
<td>2.5E-36</td>
<td>5.0E-31</td>
<td>2.5E-36</td>
<td>1.9E-19</td>
</tr>
<tr>
<td>CAD 5E-8 (14)</td>
<td>8.3E-13</td>
<td>8.0E-31</td>
<td>0.128</td>
<td>8.0E-31</td>
</tr>
<tr>
<td>CAD 5E-6 (40)</td>
<td>9.7E-22</td>
<td>1.9E-77</td>
<td>0.0124</td>
<td>0.0485</td>
</tr>
<tr>
<td>Both LDL 5E-6 (7)</td>
<td>1.8E-31</td>
<td>8.2E-35</td>
<td>1.8E-31</td>
<td>8.2E-35</td>
</tr>
<tr>
<td>Both CAD 5E-6 (7)</td>
<td>6.0E-36</td>
<td>3.2E-9</td>
<td>1.1E-7</td>
<td>2.0E-8</td>
</tr>
</tbody>
</table>
**LDL vs CAD: SNPs/IVs significant for LDL or CAD**

**Figure:** Left: significant SNPs for LDL; right: significant SNPs for CAD.
**Figure:** Left: most sig SNPs for LDL; right: most sig SNPs for CAD.
**Example: HDL vs CAD**

Table: Bi-directional MR vs our method.

<table>
<thead>
<tr>
<th>SNPs as IVs</th>
<th>bi-d MR, p-val</th>
<th>Ours, CI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IVW</td>
<td>Egger reg</td>
</tr>
<tr>
<td>HDL 5E-8 (38)</td>
<td>1.1E-3</td>
<td>0.882</td>
</tr>
<tr>
<td>HDL 5E-6 (63)</td>
<td>1.9E-4</td>
<td>0.377</td>
</tr>
<tr>
<td>CAD 5E-8 (14)</td>
<td>0.319</td>
<td>0.362</td>
</tr>
<tr>
<td>CAD 5E-6 (40)</td>
<td>0.0238</td>
<td>0.737</td>
</tr>
<tr>
<td>Both HDL 5E-6 (4)</td>
<td>0.0294</td>
<td>0.758</td>
</tr>
<tr>
<td>Both CAD 5E-6 (4)</td>
<td>1.9E-4</td>
<td>0.308</td>
</tr>
</tbody>
</table>
Future/on-going work

- Using larger lipid and CAD GWAS/seq data ...
- Model diagnostics; more robust methods:
  In the presence of invalid IVs.
- Fine mapping, co-localization testing, ...
- More applications:
  TWAS/integrating GWAS with omic data;
  risk factors for AD.
  (risk/trait prediction using large-scale GWAS summary data ...)
- Extend to high-dim settings:
  DAG (Yuan et al 2019, *Biometrika*);
  Inference (Zhu et al, in press, *JASA*) (for high-dim regression
  and Gaussian graphical models);
  Use of a non-convex penalty (TLP) (Shen et al 2012, *JASA*).
Acknowledgement:

- This research was supported by NIH, NSF and MSI.
- Many collaborators and (former and current) students!

Thank you!