北京工商大學
BEIJING TECHNOLOGY AND BUSINESS UNIVERSITY

NEURAL INFORMATION PROCESSING SYSTEMS

Identification and Estimation of the Bi-Directional MR with Some Invalid Instruments

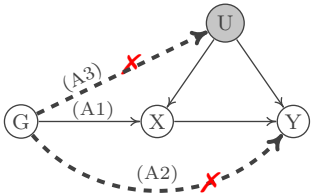
Feng Xie, Zhen Yao, Lin Xie, Yan Zeng, Zhi Geng

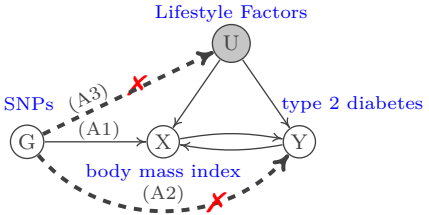
School of Mathematics and Statistics

December 14, 2024

Background

- Mendelian Randomization (MR)
  - A powerful method to estimate causal effects.
  - Widely applied in fields of clinics, socioeconomics, drug targets, etc.
  - Implemented as a form of instrumental variables analysis.
- Instrumental Variables (IVs)
  - (A1) Relevance
  - (A2) Exclusion Restriction
  - (A3) Randomness

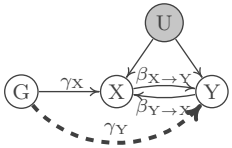




- Most existing methods focus on the one-directional MR. However, bi-directional relationships are ubiquitous in real-life scenarios.

How to identify valid IV sets for bi-directional MR models from observational data?

Model Definition



Two Stage Least Square(TSLS) IV Estimator:

$$\hat{\beta}_{X\to Y} = [X^\intercal PX]^{-1} X^\intercal PY = \beta_{X\to Y},$$

where $P = (G_{\mathcal{V}}^{X\to Y})^\intercal \left[G_{\mathcal{V}}^{X\to Y}(G_{\mathcal{V}}^{X\to Y})^\intercal\right]^{-1} G_{\mathcal{V}}^{X\to Y}$.

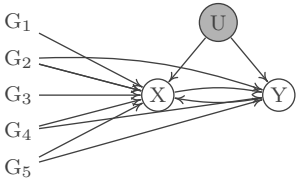
Definition (Bi-directional MR Causal Models)

$$\begin{aligned} X &= Y\beta_{Y\to X} + G^\intercal\gamma_X + \varepsilon_X, \\ Y &= X\beta_{X\to Y} + G^\intercal\gamma_Y + \varepsilon_Y. \end{aligned} \qquad \begin{aligned} X &= (G^\intercal\gamma_X + G^\intercal\gamma_Y\beta_{Y\to X} + \varepsilon_X + \varepsilon_Y\beta_{Y\to X})\,\Delta, \\ Y &= (G^\intercal\gamma_X\beta_{X\to Y} + G^\intercal\gamma_Y + \varepsilon_X\beta_{X\to Y} + \varepsilon_Y)\,\Delta. \end{aligned}$$

Goals

1. Identify valid IVs and invalid IVs.
2. Estimate the causal effects $\beta_{X\to Y}$ and $\beta_{Y\to X}$.

Motivating Example



$$
\begin{aligned}
\mathrm{X} &= \mathrm{Y}\beta_{\mathrm{Y}\to\mathrm{X}} + \mathrm{G}_1\gamma_{\mathrm{X},1} + \mathrm{G}_2\gamma_{\mathrm{X},2} + \mathrm{G}_3\gamma_{\mathrm{X},3} \\
&\quad + \mathrm{G}_4\gamma_{\mathrm{X},4} + \mathrm{G}_5\gamma_{\mathrm{X},5} + \varepsilon_{\mathrm{X}}, \\
\mathrm{Y} &= \mathrm{X}\beta_{\mathrm{X}\to\mathrm{Y}} + \mathrm{G}_2\gamma_{\mathrm{Y},2} + \mathrm{G}_4\gamma_{\mathrm{Y},4} + \mathrm{G}_5\gamma_{\mathrm{Y},5} \\
&\quad + \varepsilon_{\mathrm{Y}}.
\end{aligned}
$$

$$
\begin{cases}
\mathrm{corr}(\mathrm{Y} - \mathrm{X}\omega_{\{\mathrm{G}_3\}}, \mathrm{G}_1) = 0, \\
\mathrm{corr}(\mathrm{Y} - \mathrm{X}\omega_{\{\mathrm{G}_1\}}, \mathrm{G}_3) = 0.
\end{cases}
$$

$$
\begin{cases}
\mathrm{corr}(\mathrm{Y} - \mathrm{X}\omega_{\{\mathrm{G}_4,\mathrm{G}_5\}}, \mathrm{G}_2) \neq 0, \\
\mathrm{corr}(\mathrm{Y} - \mathrm{X}\omega_{\{\mathrm{G}_2,\mathrm{G}_5\}}, \mathrm{G}_4) \neq 0, \\
\mathrm{corr}(\mathrm{Y} - \mathrm{X}\omega_{\{\mathrm{G}_2,\mathrm{G}_4\}}, \mathrm{G}_5) \neq 0.
\end{cases}
$$

where,
$\mathrm{corr}(\cdot)$ denotes the Pearson's correlation coefficient between two random variables,
$\omega_{\{\mathrm{G_i}\}} = \mathrm{TSLS}(\mathrm{X}, \mathrm{Y}, \{\mathrm{G_i}\})$ with $\mathrm{i} \in \{1, 3\}$, and
$\omega_{\{\mathrm{G_i},\mathrm{G_j}\}} = \mathrm{TSLS}(\mathrm{X}, \mathrm{Y}, \{\mathrm{G_i}, \mathrm{G_j}\})$ with $\mathrm{i} \neq \mathrm{j}$ and $\mathrm{i}, \mathrm{j} \in \{2, 4, 5\}$.

Identifying Invalid IV Sets

Definition (Pseudo-Residual)

Let $\mathbb{G}$ be a subset of candidate genetic variants. A pseudo-residual of $\{\mathrm{X}, \mathrm{Y}\}$ relative to $\mathbb{G}$ is defined as $\mathcal{PR}_{(\mathrm{X},\mathrm{Y}\,|\,\mathbb{G})} := \mathrm{Y} - \mathrm{X}\omega_{\mathbb{G}},$, where $\omega_{\mathbb{G}} = \mathrm{TSLS}(\mathrm{X}, \mathrm{Y}, \mathbb{G})$.

Assumption 1. (Valid IV Set)

For a given causal relationship (if the relationship exists), there exists a valid IV set that consists of at least two valid IVs. For example, for the causal relationship $\mathrm{X} \to \mathrm{Y}$, $|\mathrm{G}_{\mathcal{V}}^{\mathrm{X}\to\mathrm{Y}}| \geq 2$.

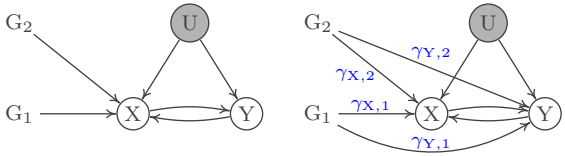
Proposition (Identifying Invalid IV Sets)

Let $\mathbb{G} = \{\mathrm{G}_1, \ldots, \mathrm{G}_\mathrm{p}\}, \mathrm{p} \geq 2$ be a subset of candidate genetic variants G. Suppose that Assumption 1 and the size of participants $\mathrm{n} \to \infty$ hold. If there exists a $\mathrm{G}_\mathrm{j} \in \mathbb{G}$ such that, $\mathrm{corr}(\mathcal{PR}_{(\mathrm{X},\mathrm{Y}\,|\,\mathbb{G}\setminus\mathrm{G}_\mathrm{j})}, \mathrm{G}_\mathrm{j}) \neq 0$, then $\mathbb{G}$ is an invalid IV set for any one of the causal relationships.

Counter Example

Question: Does the above proposition identify all invalid IV sets?

$\Rightarrow$ Assumption 1 is not sufficient. See the following example.



$$\begin{cases} \mathrm{corr}(\mathcal{PR}_{(\mathrm{X},\mathrm{Y}\,|\,\{\mathrm{G}_2\})}, \mathrm{G}_1) = 0 \\ \mathrm{corr}(\mathcal{PR}_{(\mathrm{X},\mathrm{Y}\,|\,\{\mathrm{G}_1\})}, \mathrm{G}_2) = 0 \end{cases}$$

$$\begin{cases} \mathrm{corr}(\mathcal{PR}_{(\mathrm{X},\mathrm{Y}\,|\,\{\mathrm{G}_2\})}, \mathrm{G}_1) = \dfrac{\gamma_{\mathrm{Y},1}\gamma_{\mathrm{X},2} - \gamma_{\mathrm{Y},2}\gamma_{\mathrm{X},1}}{\beta_{\mathrm{Y}\to\mathrm{X}}\gamma_{\mathrm{Y},2} + \gamma_{\mathrm{X},2}}. \\ \mathrm{corr}(\mathcal{PR}_{(\mathrm{X},\mathrm{Y}\,|\,\{\mathrm{G}_1\})}, \mathrm{G}_2) = \dfrac{\gamma_{\mathrm{Y},2}\gamma_{\mathrm{X},1} - \gamma_{\mathrm{Y},1}\gamma_{\mathrm{X},2}}{\beta_{\mathrm{Y}\to\mathrm{X}}\gamma_{\mathrm{Y},1} + \gamma_{\mathrm{X},1}}. \end{cases}$$

Identifying Valid IV Sets

Assumption 2. (Generic Identifiability)

For a given MR causal model, parameters γ and β live in a set of Lebesgue measure non-zero.

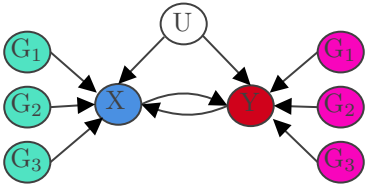
Proposition (Identifying Valid IV Sets)

Let $\mathbb{G} = \{G_1, \ldots, G_p\}, p \geq 2$ be a subset of candidate genetic variants G. Suppose that Assumption 1 and Assumption 2, and the size of participants $n \to \infty$ hold. If each $G_j \in \mathbb{G}$ satisfies

$$\operatorname{corr}(\mathcal{PR}_{(X,Y \mid \mathbb{G} \setminus G_j)}, G_j) = 0, \tag{1}$$

then $\mathbb{G}$ is a proper valid IV set for one of the causal relationships, i.e., $\mathbb{G} \subseteq G_{\mathcal{V}}^{X \to Y}$ or $\mathbb{G} \subseteq G_{\mathcal{V}}^{Y \to X}$.

Identifying the Causal Direction



Assumption 3.

In a bi-directional MR model, for a given causal relationship $X \to Y$, $\beta^2_{X\to Y} \cdot \mathrm{Var}(X) < \mathrm{Var}(Y)$. Similarly, for a given causal relationship $Y \to X$, $\beta^2_{Y\to X} \cdot \mathrm{Var}(Y) \leq \mathrm{Var}(X)$ [Xue and Pan, 2020].

Proposition (Identifying the Direction for the Identified IV set)

Suppose Assumption 3 holds. Given a valid IV set denoted as $\mathbb{G}$, if each $G_j \in \mathbb{G}$ satisfies the condition $|\mathrm{corr}(G_j, Y)/\mathrm{corr}(G_j, X)| < 1$, then $\mathbb{G}$ is a valid IV set for the causal relationship $X \to Y$, i.e., $\mathbb{G} \subseteq G_{\mathcal{V}}^{X\to Y}$. Conversely, if this condition is not satisfied, $\mathbb{G}$ is a valid IV set for the causal relationship $Y \to X$, i.e., $\mathbb{G} \subseteq G_{\mathcal{V}}^{Y\to X}$.

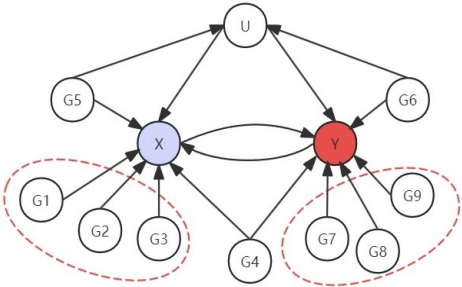
Identifiability Result

Identifiability of Bi-directional MR model

Suppose that Assumptions 1 - 3, and the size of participants $n \to \infty$ hold, the Bi-directional MR model is fully identifiable, which includes the identification of valid IV sets, the causal relationships the identified sets relate to, the causal directions between X and Y, and the causal effects of X on Y as well as of Y on X.

Our Framework

- Find valid IV sets.
- Infer directions and estimate effects.



Algorithm 1 PReBiM

Input: A dataset of measured genetic variants $\mathbf{G} = (G_1, ..., G_g)^\intercal$, two phenotypes X and Y, significance level α, and parameter W, maximum number of IVs to consider.

1: Valid IV sets $\mathcal{V} \leftarrow$ FindValidIVSets($\mathbf{G}$, X, Y, α, W); ▷ *Step I*

2: $(\hat{\beta}_{X\to Y}, \hat{\beta}_{Y\to X}) \leftarrow$ InferCausalDirectionEffects(X, Y, $\mathcal{V}$); ▷ *Step II*

Output: $\hat{\beta}_{X\to Y}$ and $\hat{\beta}_{Y\to X}$, the causal effects of X on Y and Y on X, respectively.

Experiments on Synthetic Data

Table 1: Performance comparison of NAIVE, MR-Egger, sisVIVE, IV-TETRAD, TSHT, and PRe- BiM in estimating bi-directional MR models across various sample sizes and three scenarios.

		$\mathcal{S}(2,2,6)$				$\mathcal{S}(3,3,8)$				$\mathcal{S}(4,4,10)$			
		$X \to Y$		$Y \to X$		$X \to Y$		$Y \to X$		$X \to Y$		$Y \to X$	
Size	Algorithm	CSR↑	MSE↓	CSR↑	MSE↓	CSR↑	MSE↓	CSR↑	MSE↓	CSR↑	MSE↓	CSR↑	MSE↓
2k	NAIVE	-	0.354	-	0.349	-	0.351	-	0.320	-	0.335	-	0.304
	MR-Egger	-	0.800	-	0.734	-	0.571	-	0.569	-	0.457	-	0.494
	sisVIVE	0.30	0.398	0.25	0.447	0.34	0.379	0.32	0.379	0.37	0.346	0.38	0.330
	IV-TETRAD	0.60	0.859	0.30	0.867	0.60	0.773	0.28	0.782	0.67	0.688	0.23	0.622
	TSHT	0.07	0.606	0.06	0.660	0.07	0.549	0.07	0.613	0.12	0.653	0.11	0.689
	PReBiM	0.85	0.046	0.85	0.070	0.89	0.083	0.87	0.075	0.88	0.039	0.89	0.057
5k	NAIVE	-	0.350	-	0.339	-	0.347	-	0.319	-	0.357	-	0.306
	MR-Egger	-	0.737	-	0.836	-	0.604	-	0.508	-	0.515	-	0.520
	sisVIVE	0.32	0.421	0.30	0.407	0.33	0.378	0.35	0.384	0.39	0.439	0.38	0.365
	IV-TETRAD	0.62	0.804	0.31	0.844	0.63	0.836	0.29	0.659	0.67	0.686	0.25	0.640
	TSHT	0.04	0.544	0.06	0.565	0.08	0.528	0.06	0.570	0.10	0.549	0.07	0.541
	PReBiM	0.93	0.020	0.91	0.027	0.90	0.019	0.92	0.020	0.90	0.009	0.91	0.010
10k	NAIVE	-	0.366	-	0.348	-	0.319	-	0.342	-	0.349	-	0.300
	MR-Egger	-	0.800	-	0.763	-	0.524	-	0.620	-	0.487	-	0.424
	sisVIVE	0.31	0.457	0.29	0.503	0.37	0.383	0.31	0.394	0.41	0.399	0.41	0.328
	IV-TETRAD	0.64	0.811	0.31	0.804	0.64	0.748	0.29	0.696	0.71	0.635	0.24	0.577
	TSHT	0.04	0.449	0.02	0.475	0.05	0.525	0.05	0.466	0.05	0.479	0.07	0.498
	PReBiM	0.95	0.044	0.93	0.018	0.93	0.026	0.93	0.039	0.93	0.027	0.94	0.011

Experiments on Real-World Data

- Major IV information for the bi-directional dataset. (Here X is obesity and Y denotes Vitamin D Status)
  - Valid IVs ($X \rightarrow Y$):
    - rs9939609 (FTO)
    - rs2867125 (TMEM18)
    - rs4074134 (BDNF)
    - rs7647305 (ETV5)
    - rs7138803 (FAIM2)
    - rs3101336 (NEGR1)
    - rs10938397 (GNPDA2)
  - Valid IVs ($Y \rightarrow X$):
    - rs12785878 (DHCR7)
    - rs10741657 (CYP2R1)
    - rs2282679 (GC)
    - rs6013897 (CYP24A1)
  - Our Results: FTO and FAIM2 ($X \rightarrow Y$); DHCR7 and CYP24A1 ($Y \rightarrow X$)

Conclusions

- We provided the identifiability conditions for the bi-directional MR model, enabling both valid IV sets for each direction and the causal effects of interests to be correctly identified.
- We proposed a practical and effective cluster fusion-like algorithm for unbiased estimation of valid IV sets.
- Experimental results on synthetic and real-world data verified the effectiveness of our algorithm.

Thank you!