Sparse Causal Effect Estimation using Two-Sample Summary Statistics in the Presence of Unmeasured Confounding

Shimeng Huang, Niklas Pfister, and Jack Bowden

ICSDS 2024

Nice, France

December 17, 2024

Motivation

In genetic epidemiology, Mendelian randomization (MR)—instrumental variable estimation (IV) with genetic variants being the instruments—is often used to estimate causal effects, which accounts for unmeasured confounding.

- Only summary-level data from two samples are available: estimated marginal associations from large-scale genome-wide association study (GWAS).
- Not enough strong & valid instruments for all covariates, but also not all covariates may have a direct causal effect on the response.

Setting

Assume the underlying SCM is linear and we observe two-sample summary statistics

- Estimated association between Z and Y from sample 1 $(\widehat{\pi}, \widehat{\Sigma}_{\pi})$
- Estimated association between Z and X from sample 2 $(\widehat{\Pi}, \widehat{\Sigma}_{\Pi})$

The goal is to identify the average causal effect of X on Y.

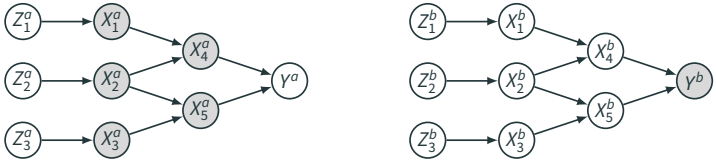


Figure 1: Example: A two-sample IV scenario that is considered as underidentified in the usual sense. *Hidden confounders between X and Y are omitted.* Unobserved variables are in gray.

Methods

Pfister and Peters (2022) discuss identifiability conditions of sparse causal effects, and propose an estimator (spaceIV) based on one-sample individual-level data using subset selection.

In this work, we consider the two-sample summary-statistics counterpart using

- **Subset selection** (L0) for which we show consistency
- **Lasso-type estimation** (L1) as a computational speed-up

Methods

Q statistic (Theorem 2.2, HPB, 2024)

Assume certain regularity conditions hold. For all $\beta \in \mathbb{R}^d$, define the Q statistic as

$$Q(\beta) := (\widehat{\pi} - \widehat{\Pi}\beta)^\top (\tfrac{1}{n_a}\widehat{\Sigma}_\pi + \tfrac{1}{n_b}\widehat{\Sigma}_\Pi(\beta))^{-1}(\widehat{\pi} - \widehat{\Pi}\beta),$$

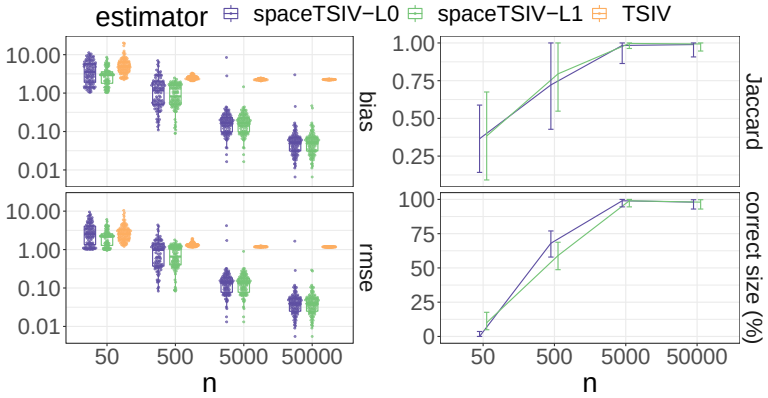
where $\widehat{\Sigma}_\Pi(\beta) := \xi(\beta)\widehat{\Sigma}_\Pi\xi^\top(\beta)$ with $\xi(\beta) := \beta^\top \otimes I_m$. Then it holds for all $\beta \in \mathbb{R}^d$ and all $r \in (0, \infty)$ that

$$\lim_{\substack{n_a, n_b \to \infty \\ n_a/n_b \to r}} \sup_{t \in \mathbb{R}} \sup_{\substack{P \in \mathcal{P}: \\ \beta \in \mathcal{B}^{\text{sum}}(P)}} \left| \mathbb{P}_P \left(Q(\beta) \leq t \right) - \kappa_m(t) \right| = 0,$$

where κ_m is the CDF of the χ^2 distribution with m-degrees of freedom.

Simulation

Data-generating processes: A linear SCM corresponding to Figure 1.



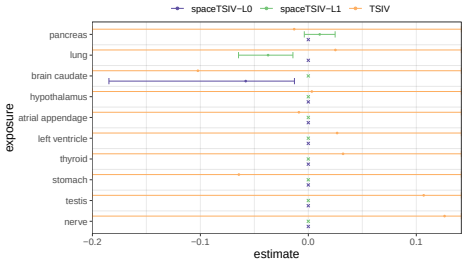
Application

We apply sparse MVMR to the tissue-specific GLP1R expression on the risk of coronary artery disease (Patel et al., 2024).

Table 1: Top 5 ranked models and corresponding posterior probabilities from MR-BMA (Patel et al., 2024, Table 2).

Rank	Tissues	Posterior probability (%)
1	Brain-caudate	45.9
2	Stomach	14.8
3	Nerve	8.0
4	Testis	6.6
5	Brain-hypothalamus	5.4

Figure 2: spaceTSIV using 17 genetic variants as instruments. Error bars represent 90% CIs constructed by inverting Q test.



Discussions

Future directions:

- Variable selection on the instruments as well
- Overlapping samples
- Heterogeneous samples
- Inference of the spaceTSIV estimator (post-selection inference)

References

F. Batool, A. Patel, D. Gill, and S. Burgess. Disentangling the effects of traits with shared clustered genetic predictors using multivariable mendelian randomization. *Genetic Epidemiology*, 46(7):415–429, 2022.

L. P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the econometric society*, pages 1029–1054, 1982. doi:10.2307/1912775.

A. Patel, D. Gill, D. Shungin, C. S. Mantzoros, L. B. Knudsen, J. Bowden, and S. Burgess. Robust use of phenotypic heterogeneity at drug target genes for mechanistic insights: Application of cis-multivariable mendelian randomization to glp1r gene region. *Genetic Epidemiology*, 2024. doi:10.1002/gepi.22551.

N. Pfister and J. Peters. Identifiability of sparse causal effects using instrumental variables. In *Uncertainty in Artificial Intelligence*, pages 1613–1622. PMLR, 2022.

M. Schmidt, G. Fung, and R. Rosales. Fast optimization methods for l1 regularization: A comparative study and two new approaches. In *Machine Learning: ECML 2007: 18th European Conference on Machine Learning, Warsaw, Poland, September 17-21, 2007. Proceedings 18*, pages 286–297. Springer, 2007.