Causal Inference with Hidden Confounders

Instrumental Variables and the Generalized Causal Dantzig

James Long

University of Texas MD Anderson Cancer Center

ENAR – March 28, 2022

Contact: jplong@mdanderson.org

Collaboration

James Long
MDA

Kim-Anh Do
MDA

Min Jin Ha
Yonsei

Hongxu Zhu
UT Health

THE UNIVERSITY OF TEXAS
MD Anderson Cancer Center
Making Cancer History®

YONSEI UNIVERSITY HEALTH SYSTEM

UTHealth
The University of Texas
Health Science Center at Houston

Outline

Hidden Confounding, Reverse Causality and Instrumental Variables

Generalized Causal Dantzig

Simulation

Outline

Hidden Confounding, Reverse Causality and Instrumental Variables

Generalized Causal Dantzig

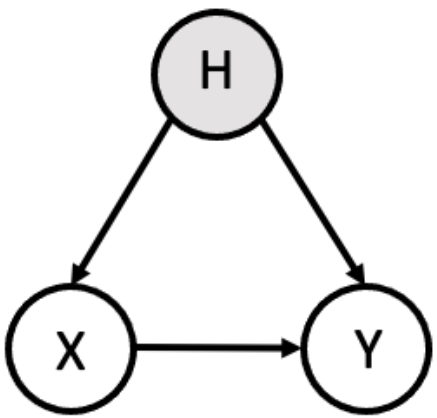
Simulation

Hidden Confounders and Reverse Causality

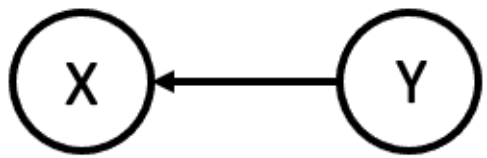
Goal: Estimate causal effect of X on Y.

Problem: Regression may produce inconsistent estimates.

Reason 1: Hidden confounding



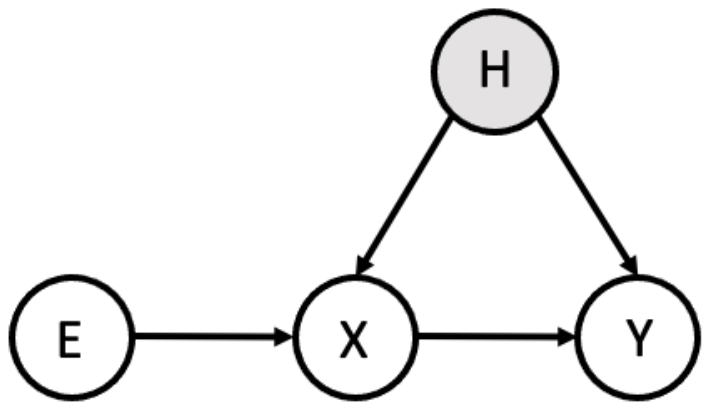
Reason 2: Reverse causality



Instrumental Variables (IV)

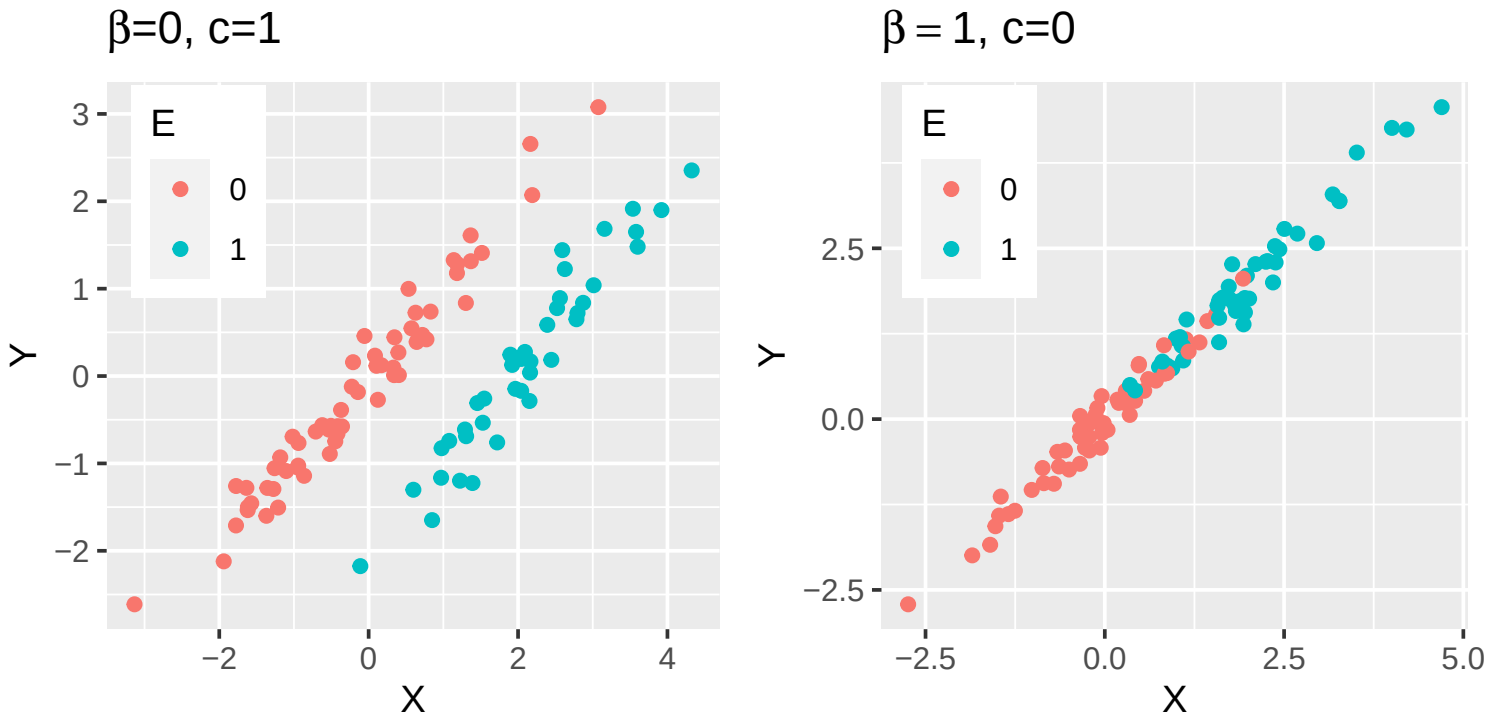
E is an instrument if:

1. correlated with exposure X
2. does not have a direct effect on Y



How Do IVs Identify Causal Effects?

$$E \sim Bernoulli(1/2)$$
$$X \leftarrow 2E + H + \epsilon_X$$
$$Y \leftarrow \beta X + cH + \epsilon_Y$$



IV Estimator via Generalized Method of Moments

- **Main Idea:**

$$E \perp \underbrace{Y - X^T \overbrace{\beta_0}^{\text{true } \beta}}_{=h+\epsilon_Y}$$

- Moment conditions:

$$m_{IV}(\beta) \equiv \mathbb{E}[E(Y - X^T\beta)] \underbrace{=}_{\text{iff } \beta=\beta_0} 0$$

$$\widehat{m}_{IV}(\beta) \equiv \frac{1}{n}\sum_{i=1}^{n} E_i(Y_i - X_i^T\beta) \underbrace{\approx}_{\text{when } \beta=\beta_0} 0$$

- Estimate β with:

$$\widehat{\beta}_{IV}(\widehat{W}) = \underset{\beta}{\text{argmin}} \; ||\widehat{m}(\beta)||_{\widehat{W}} = \underset{\beta}{\text{argmin}} \; \widehat{m}(\beta)^T \widehat{W} \widehat{m}(\beta)$$

- Two Stage Least Squares (TSLS) uses $\widehat{W}$ which is optimal under some conditions.

IV References: IVs date back to Wright [1928]. See Mátyás [1999] for GMM estimation in IV models.

Overview of Remainder of Talk

Background on Causal Dantzig (CD):

- Proposed in Rothenhäusler et al. [2019] AOS
- Consistent under hidden confounding

Contributions:

- New Estimator: Generalized Causal Dantzig (GCD)
- GCD and IV Comparisons

Outline

Hidden Confounding, Reverse Causality and Instrumental Variables

Generalized Causal Dantzig

Simulation

Environments and CD Response Model

- (X, Y) data may be collected in different **environments**
  - Example: Perturbation Biology
    - Environment 0: No stimulation
    - Environment 1: Stimulation with reagents
    - Data: Protein expression (X) and phenotype (Y) for many cells in each environment
- Goal: Use environment to infer $X \to Y$ causal effects
- Several environment based estimators have been proposed:
  - Invariant Causal Prediction (ICP) Peters et al. [2016], Sequential ICP Pfister et al. [2019], Nonlinear ICP Heinze-Deml et al. [2018]
  - **Causal Dantzig**
- Main Assumptions of Causal Dantzig:

$$\underbrace{Y^e}_{\text{env. e resp.}} = \overbrace{X^{e^T}}^{\text{env. e exp.}} \beta + \delta_Y$$

Causal Dantzig (CD) Estimator

Let:

- $\mathbf{X^j} \in \mathbb{R}^{\mathbf{n_j} \times \mathbf{p}}$ be the environment j design matrix
- $\mathbf{Y^j} \in \mathbb{R}^{\mathbf{n_j}}$ be the environment j response

With 2 environments the CD is:

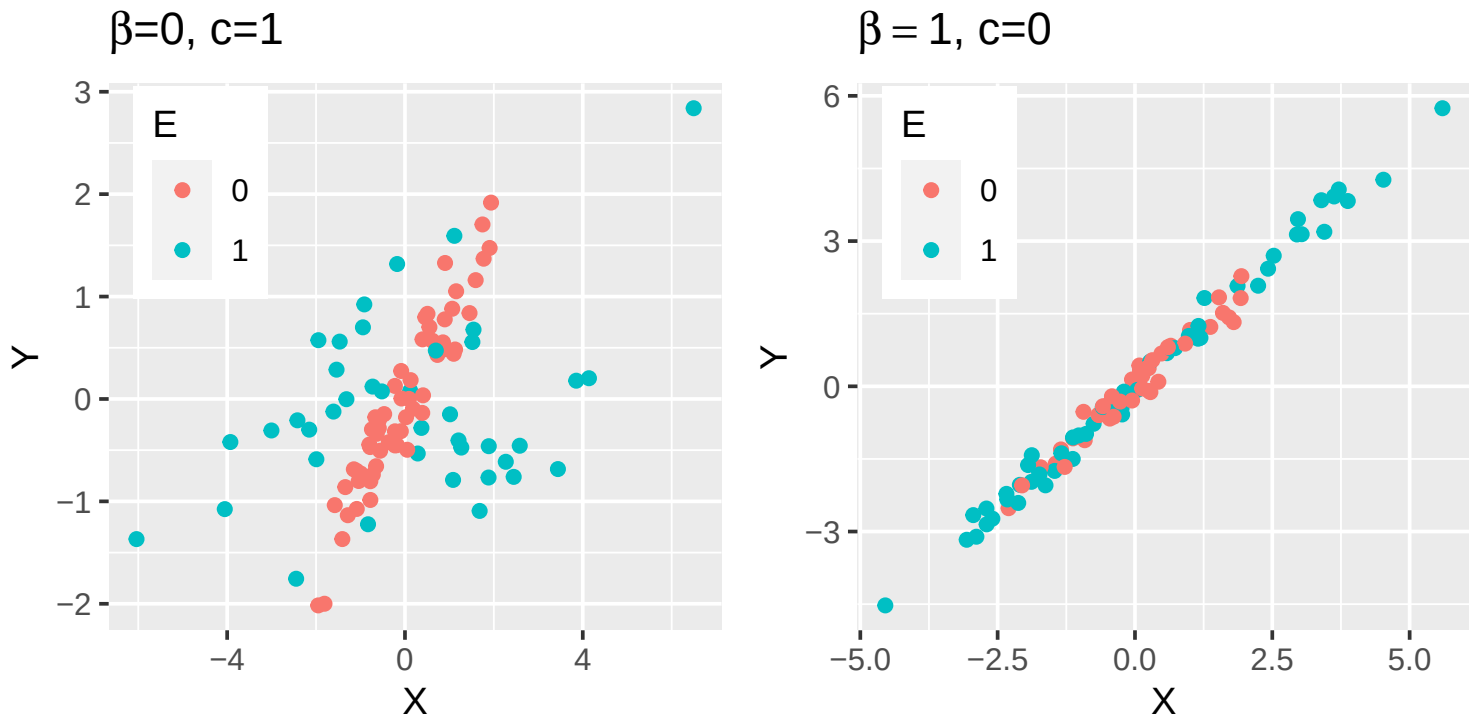
$$\widehat{\beta}_{CD} = \left(\overbrace{\frac{1}{n_1}\mathbf{X}^{1^T}\mathbf{X}^1 - \frac{1}{n_0}\mathbf{X}^{0^T}\mathbf{X}^0}^{\Delta \text{ 2nd moment of exposures}} \right)^{-1} \left(\underbrace{\frac{1}{n_1}\mathbf{X}^{1^T}\mathbf{Y}^1 - \frac{1}{n_0}\mathbf{X}^{0^T}\mathbf{Y}^0}_{\Delta \text{ exposure} \times \text{ response}} \right)$$

Notes:

- Proposed in Rothenhäusler et al. [2019]
- Motivated by notion of inner product invariance

When Can CD Identify Causal Effects?

$$E \sim Bernoulli(1/2)$$
$$X \leftarrow H + (2E + .2)\epsilon_X$$
$$Y \leftarrow \beta X + cH + \epsilon_Y$$



Note: E does not shift mean of X, so IV/TSLS inconsistent.

Notes on CD

Strengths:

- CD is consistent for some models where TSLS fails (e.g. variance shifts)
- Only need two environments to identify causal effects (under assumptions on how environment effects $X \in \mathbb{R}^p$)

Weaknesses:

- Environments must be discrete.
- Rothenhäusler et al. [2019] has no theory or optimality guarantees for estimators with more than 2 environments.

Generalized Causal Dantzig (GCD) Estimator

New Estimator: The GCD is a GMM with moment conditions:

$$m_{GCD}(\beta) \equiv \mathbb{E}[vec(EX^T)(Y - X^T\beta)].$$

- $E \in \mathbb{R}^q$ is set of instruments.
- $vec(EX^T) \in \mathbb{R}^{qp}$ column stacks matrix EX^T
- $\widehat{\beta}_{GCD}(\widehat{W}) = \underset{\beta}{\text{argmin}} \; ||\widehat{m}_{GCD}(\beta)||_{\widehat{W}}$

GCD Estimator

- The GCD exactly matches CD in 2 environment case.
  - Idea: Encode (categorical) environment $\mathcal{E} \in \{0, 1, 2, \ldots\}$ as instrument $E \in \mathbb{R}^{\#\mathcal{E}-1}$.
- GCD can be used with continuous environments, CD cannot.
- GMM theory can be used to optimally weight environments (2 step estimators).
- Conceptual Result: Environments and instruments are very closely related.

$$m_{IV}(\beta) = \mathbb{E}[E(Y - X^T\beta)]$$
$$m_{GCD}(\beta) = \mathbb{E}[vec(EX^T)(Y - X^T\beta)]$$

Outline

Hidden Confounding, Reverse Causality and Instrumental Variables

Generalized Causal Dantzig

Simulation

Simulation Parameters

Exposure and response models:

$$X \leftarrow 9h + (10E + 1)\epsilon_X$$
$$Y \leftarrow 3h + X + \epsilon_Y$$

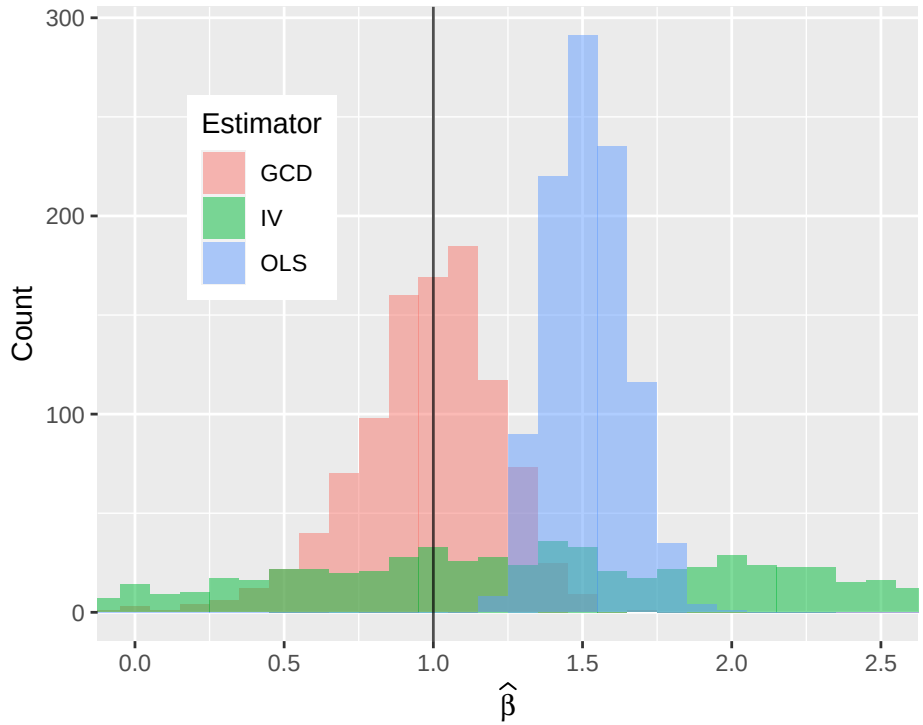
Exogenous variables (all independent) are:

$$E \sim Unif[0, 1]$$
$$h, \epsilon_X, \epsilon_Y \sim N(0, 1)$$

Sample size $n = 100$.

Fit 3 estimators $\widehat{\beta}_{GCD}$, $\widehat{\beta}_{IV}$, $\widehat{\beta}_{OLS}$.

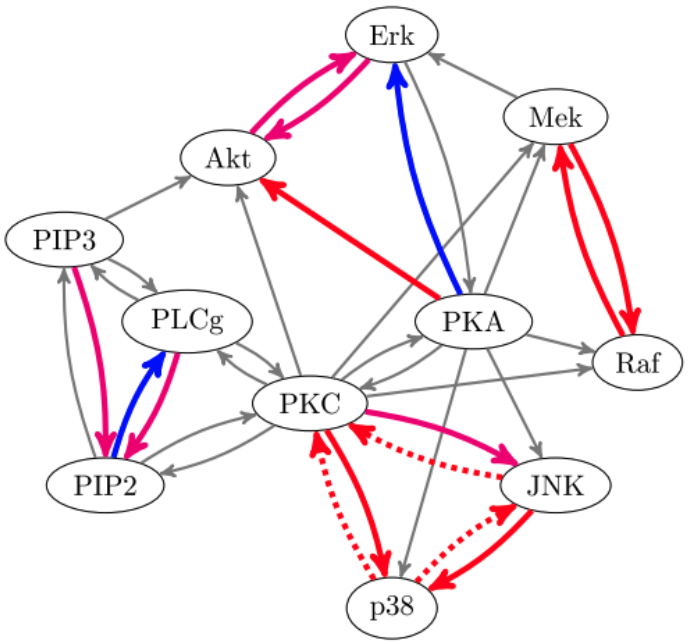
Empirical Sampling Distributions



- OLS is inconsistent due to hidden confounding.
- IV inconsistent because E does not shift mean of X.
- GCD performs well.
- CD not applicable because instrument is continuous.

Ongoing Work

- Currently working on applications in causal discovery with Flow Cytometry Data



- Manuscript available on arxiv soon

Source: Meinshausen et al. [2016]

Thank you. Questions?

Bibliography I

C. Heinze-Deml, J. Peters, and N. Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6 (2), 2018.

L. Mátyás. Generalized method of moments estimation. 1999.

N. Meinshausen, A. Hauser, J. M. Mooij, J. Peters, P. Versteeg, and P. Bühlmann. Methods for causal inference from gene perturbation experiments and validation. *Proceedings of the National Academy of Sciences*, 113(27):7361–7368, 2016.

J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

N. Pfister, P. Bühlmann, and J. Peters. Invariant causal prediction for sequential data. *Journal of the American Statistical Association*, 114(527):1264–1276, 2019.

D. Rothenhäusler, P. Bühlmann, and N. Meinshausen. Causal dantzig: fast inference in linear structural equation models with hidden variables under additive interventions. *The Annals of Statistics*, 47(3):1688–1722, 2019.

P. G. Wright. *Tariff on animal and vegetable oils*. Macmillan Company, New York, 1928.