When the Causal Graph is Unknown Embrace Model Pluralism

Rohit Bhattacharya

Williams College

Collaborators

Junhui Yang[†]

Ted Westling[†]

Youjin Lee[‡]

He Bai[†]

Ina Ocelli[⋆]

[†] Dept. of Mathematics and Statistics, UMass Amherst

[‡] Dept. of Biostatistics, Brown University

[⋆] Dept. of Computer Science, Williams College

RB acknowledges funding from NSF CRII grant 2348287

Presentation overview

Papers:

- Yang, Bhattacharya, Lee, and Westling (2024). Statistical and Causal Robustness for Causal Null Hypothesis Tests. *40th Conference on Uncertainty in Artificial Intelligence.*
- Bhattacharya, Ocelli, and Westling (2026). Robust Weighted Triangulation of Causal Effects Under Model Uncertainty. *pre-print on arXiv.*

Presentation overview

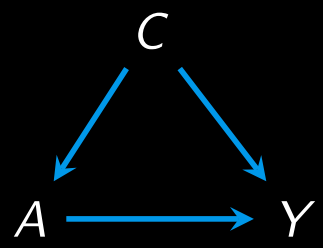
- Propose causal null hypothesis tests that are "causally robust"—the test is valid if at least one causal model is correct among $K \geq 2$ models
- Propose a triangulation functional that gives robust effect estimates in the sense that we can bound its distance from the true causal effect, and the distance is small if at least one model is correct and testable from observed data

Motivation: Statistical robustness

- Suppose our target is the Average Causal Effect (ACE)

$$\theta = \mathbb{E}[Y^{\text{do}(A=1)}] - \mathbb{E}[Y^{\text{do}(A=0))}]$$

Motivation: Statistical robustness

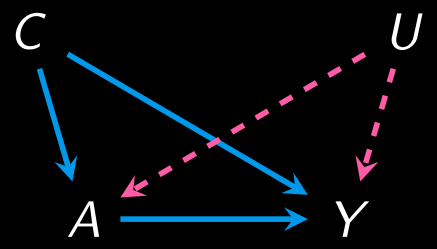


$\mathcal{M}_1$

- Let $\mu(a, c) \coloneqq \mathbb{E}[Y \mid A, C]$ and $\pi(c) \coloneqq p(A = 1 \mid C)$
- The AIPW estimator (Bang and Robins 2005) of θ with influence function ϕ_{ψ_1} is doubly robust with respect to estimators μ_n and π_n

$$\phi_{\psi_1} = \{y - \mu(a, c)\}\left\{\frac{a - \pi(c)}{\pi(c)(1 - \pi(c))}\right\} + \{\mu(1, c) - \mu(0, c)\} - \psi_1$$

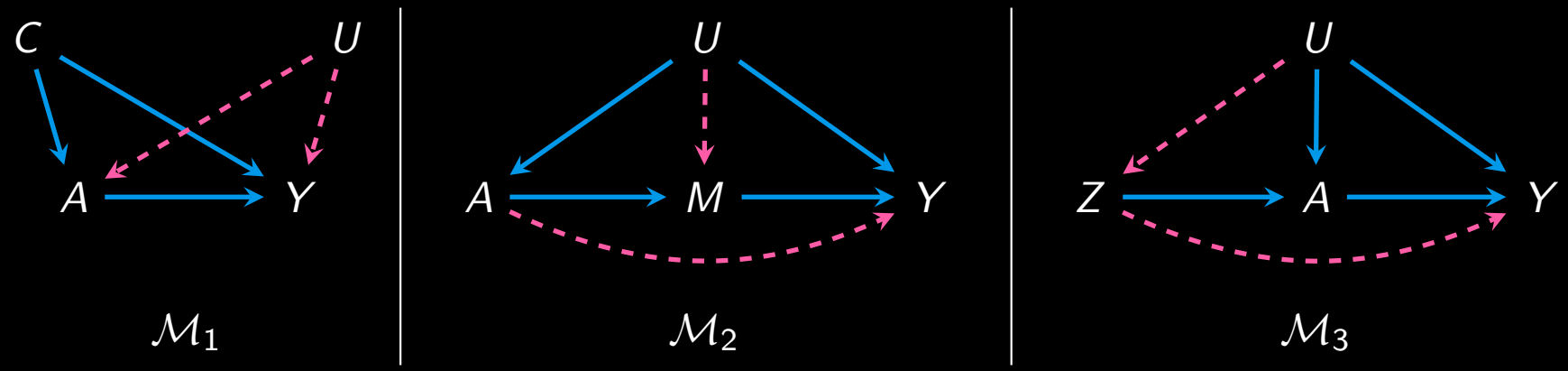
Causal robustness under model pluralism



$\mathcal{M}_1$

- AIPW (and other estimators like it) provides robustness to statistical model misspecification
- However, this robustness property cannot protect against misspecification of $\mathcal{M}_1$—we get valid inference, but for an observed data parameter $\psi_1 \neq \theta$
- A natural response to uncertainty about $A \leftarrow U \rightarrow Y$ is to try methods that use different identifying assumptions

Causal robustness under model pluralism



- $\mathcal{M}_2$ is the frontdoor model (Pearl 1995)

$$\psi_2 = \sum_{m,a} \{p(m \mid A = 1) - p(m \mid A = 0)\} \times p(a) \times \mathbb{E}[Y \mid a, m]$$

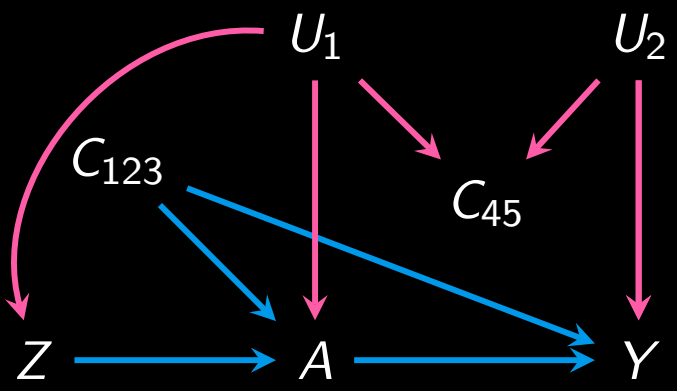
- $\mathcal{M}_3$ is the IV model (Angrist et al. 1996)

$$\psi_3 = \frac{\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0]}{\mathbb{E}[A \mid Z = 1] - \mathbb{E}[A \mid Z = 0]}$$

- Both trade $A \leftarrow U \rightarrow Y$ for different assumptions

And a whole universe of other models like the napkin model, proximal causal inference model, three parallel outcomes model, . . .

Another common example of causal model uncertainty



- Could be unsure which of the following adjustment sets to use when $\mathcal{G}$ is unknown. Related to the classic M-bias controversy (Sjölander 2009; Pearl 2009; Ding and Miratrix 2015)

$$\mathcal{M}_1 : \{C_1, C_2, C_3\}$$
$$\mathcal{M}_2 : \{C_1, C_2, C_3, C_4, C_5\}$$
$$\mathcal{M}_3 : \{C_1, C_2, C_3, C_4\}$$

- In this case, only $\psi_1 = \theta$

Causal robustness under model pluralism

- Intuitively, methodological pluralism should help if done in a principled manner—relying on a single model is risky when there is model uncertainty, however, we need to be careful about inference when dipping into the data several times
- Goal: Let analyst propose $K \geq 2$ causal models $\mathcal{M}_1, \ldots, \mathcal{M}_K$. Provide valid inference for the causal parameter θ using the functionals $\psi_1, \ldots, \psi_K$ if at least one model $\mathcal{M}_k$ is correct
- These models could be qualitatively distinct models like backdoor, frontdoor, and IV. Or qualitatively similar ones like multiple candidate adjustment sets

Causal robustness under model pluralism

- Triangulation in the social sciences dates back to at least 1899—writings of WEB Du Bois (Heesen et al. 2019)
- More recent work in stats/epi: Rosenbaum 2010; Rosenbaum 2011; Kang et al. 2016; Sun et al. 2021; Yao et al. 2024; Rakshit et al. 2025; Lawlor et al. 2016; Shapland et al. 2024
- These methods typically rely on one of the following: the majority of models being correct (so that a plurality strategy can be used), non-overlapping sources of bias, or are tailored to specific classes of causal models. These assumptions can be difficult to justify in observational settings

Benefits of model pluralism need not rely on plurality correctness

Want our method to be general, so it can be applied to many settings

Main idea for robust hypothesis testing

- We suppose $O_1, \ldots, O_n \sim P$ are IID observed data
- Let $\mathcal{M}_1, \ldots, \mathcal{M}_K$ be candidate causal models, and let ψ_k be an identifying functional for θ under $\mathcal{M}_k$
- For each k, suppose we can construct an asymptotically linear (AL) estimator of ψ_k with influence function ϕ_k, meaning that

$$\psi_{k,n} - \psi_k = \frac{1}{n} \sum_{i=1}^{n} \phi_k(O_i) + o_P(n^{-1/2})$$

  where ϕ_k are assumed to satisfy $\mathbb{E}[\phi_k] = 0$ and $\mathbb{E}[\phi_k^2] < \infty$
- Influence function-based estimators are typically AL under sufficient rates of convergence of nuisance estimators

Main idea for robust hypothesis testing

- Now consider the product $\prod_{k=1}^{K} \psi_k$. By the delta method

$$n^{1/2}\left(\prod_{k=1}^{K} \psi_{k,n} - \prod_{k=1}^{K} \psi_k\right) \to_d N\left(0, \gamma'\Sigma\gamma\right)$$

  where $\gamma := (\gamma_1, \ldots, \gamma_K)'$ for $\gamma_k := \prod_{j \neq k} \psi_j$

- Now, if at least one causal model, say $\mathcal{M}_k$ is correct, then the causal null H_0 implies that $\theta = \psi_k = 0$ and $\prod_{k=1}^{K} \psi_k = 0$. So,

$$n^{1/2} \prod_{k=1}^{K} \psi_{k,n} \to_d N\left(0, \gamma'\Sigma\gamma\right)$$

- Hence, for PD Σ, under H_0 and at least one $\mathcal{M}_k$ being correct

$$T_n := n^{1/2}\left(\gamma_n'\Sigma_n\gamma_n\right)^{-1/2} \prod_{k=1}^{K} \psi_{k,n} \to_d N(0,1)$$

- We reject H_0 at level α if $|T_n| > q_{1-\alpha/2}$, where q_p denotes the pth quantile of a standard normal distribution

Asymptotic size

Theorem 1: Asymptotic Type I Error Rate

If $\psi_{k,n}$ is an AL estimator of ψ_k for each k, $\prod_{k=1}^{K} \psi_k = 0$, $\Sigma_n \to_P \Sigma$, and $\gamma'\Sigma\gamma > 0$, then $P\left(|T_n| > q_{1-\alpha/2}\right) \longrightarrow \alpha$.

If $\gamma'\Sigma\gamma = 0$, then our test is conservative. This happens when:

- Two or more causal models are correct
- Identified functional in wrong model is zero
- An influence function is 0 under H_0
- Two or more of the influence functions are linearly dependent under H_0

Simulation study

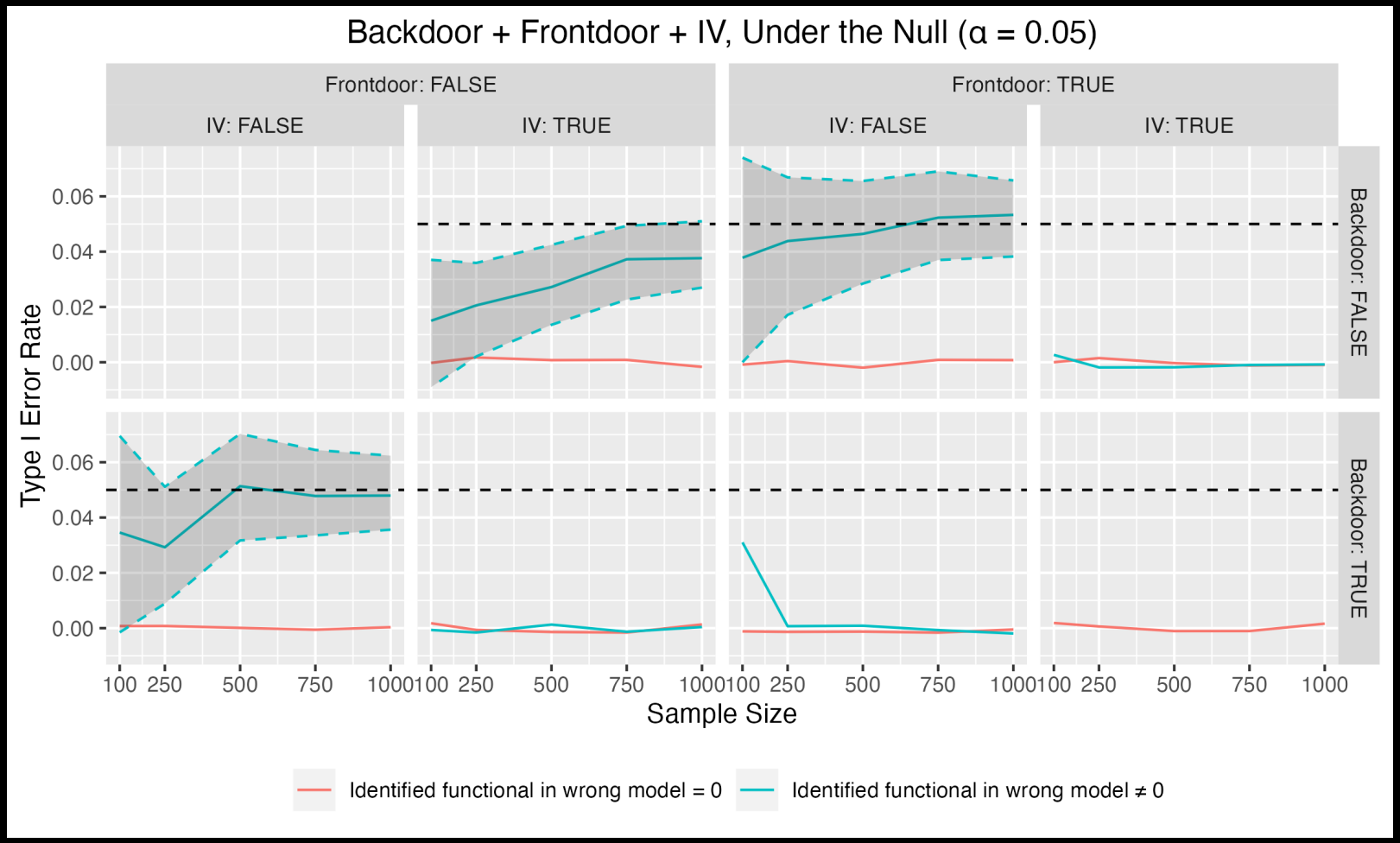


Figure: Size of the test as a function of sample size. Panel labels indicate which model(s) are correct (TRUE) and incorrect (FALSE).

Asymptotic power

Theorem 2: Asymptotic Power

If $\psi_{k,n} \to_P \psi_k$ for each k, $\prod_{k=1}^{K} \psi_k \neq 0$, and $\Sigma_n = O_P(1)$, then $P\left(|T_n| > q_{1-\alpha/2}\right) \longrightarrow 1$.

- If $\prod_{k=1}^{K} \psi_k = 0$, then we do not have power
- This happens under the alternative when $\psi_k = 0$ for an incorrect causal model $\mathcal{M}_k$—can happen sometimes as a violation of faithfulness, for example. Or if you have some models in which ψ_k is always 0
- The price of additional robustness is lower power in some cases

Simulation study

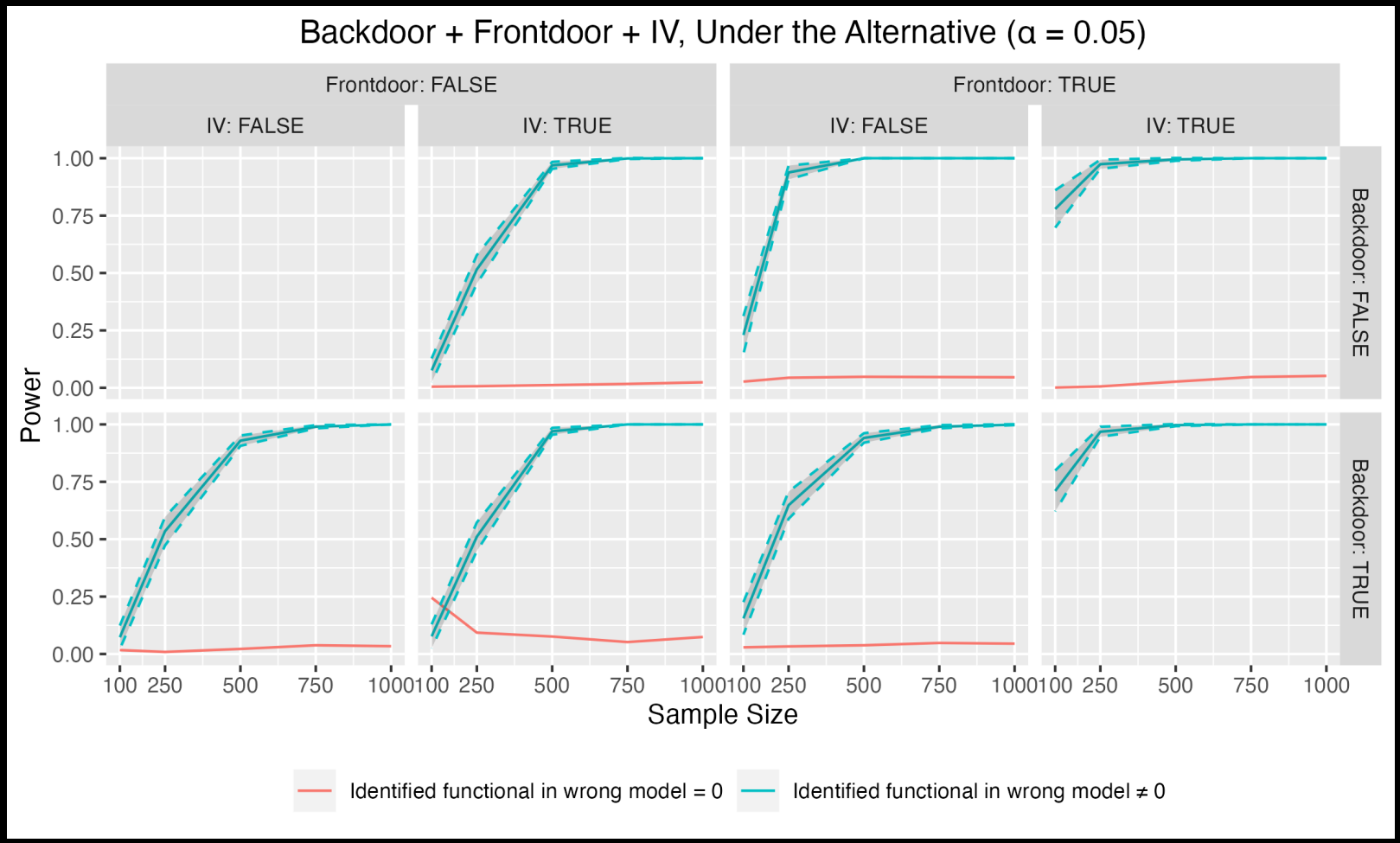


Figure: Power of the test as a function of sample size. Panel labels indicate which model(s) are correct (TRUE) and incorrect (FALSE).

Adaptive test

- As mentioned earlier, we are leaving some power on the table, especially when multiple causal models are correct
- We propose an adaptive test in Yang, Bai, Bhattacharya, and Westling (2026+). Robust and Adaptive Causal Inference Under Model Uncertainty. *pre-print coming soon*
- Skip describing this and move on to effect estimation

See also Drton and Xiao 2016; Miles and Chambaz 2021

Inverting the test

- We can invert the test to obtain confidence sets
- However, in general, they may be quite difficult to interpret as they are multisets, so we also develop a different approach for effect estimation

Main idea for effect estimation via weighted triangulation

- The test from the previous part operates under fairly weak assumptions and uses only the effects computed from each model when combining them
- We can do more if we allow ourselves to use some extra information on how reliable each model is based on testable implications of the identifying assumptions

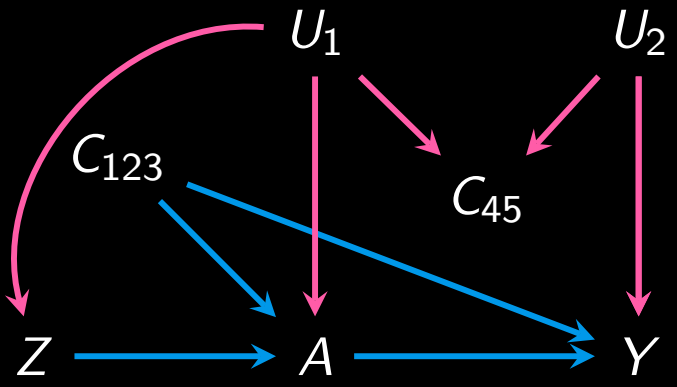
Main idea for effect estimation via weighted triangulation

- Suppose, under some set of assumptions $\mathcal{A}$, there exist observed data parameters β_k for each $\mathcal{M}_k$, such that

$$\beta_k = 0 \implies \mathcal{M}_k \text{ is correct}$$

When $\beta_k \neq 0$, the model $\mathcal{M}_k$ is either untestable from observed data, or simply incorrect

A concrete example of assumptions $\mathcal{A}$ and β_k



- Say we want to test correctness of backdoor adjustment sets:

$$\mathcal{M}_1 : \{C_1, C_2, C_3\}$$
$$\mathcal{M}_2 : \{C_1, C_2, C_3, C_4, C_5\}$$
$$\mathcal{M}_3 : \{C_1, C_2, C_3, C_4\}$$

A concrete example of assumptions $\mathcal{A}$ and β_k

$\mathcal{A}_1$: P is faithful wrt a causal DAG $\mathcal{G}(V \cup U)$ i.e. $X \perp\!\!\!\perp_{\text{d-sep}} Y \mid Z \iff X \perp\!\!\!\perp Y \mid Z$

$\mathcal{A}_2$: $\mathcal{G}$ satisfies the causal ordering $\{Z, C\} < A < Y$

$\mathcal{A}_3$: $Z \to A \to Y$ exists in $\mathcal{G}$

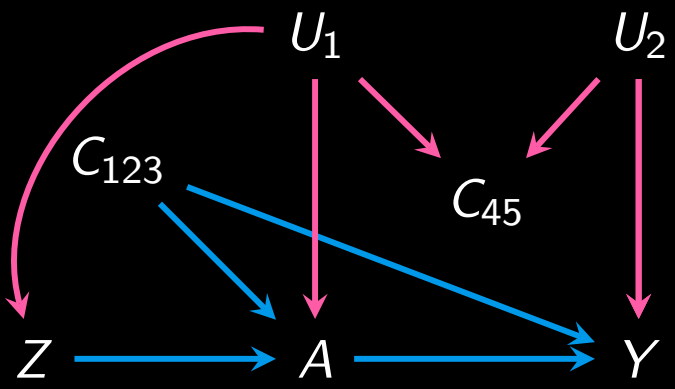
- Entner et al. 2013 show under $\mathcal{A}_1 - \mathcal{A}_3$ that if $Z \perp\!\!\!\perp Y \mid A, W$ for any $W \subseteq C$, then W is a valid backdoor set
- Let β_k be any non-parametric/semi-parametric measure of association such that $\beta_k = 0 \iff Z \perp\!\!\!\perp Y \mid A, W$. One choice is the log-odds ratio. Given reference values z_0, y_0, the odds ratio function $\text{OR}(Z, Y \mid A, W)$ is (Chen 2007),

$$\text{OR}(z, y \mid a, w) = \frac{p(z \mid a, y, w)}{p(z_0 \mid a, y, w)} \times \frac{p(z_0 \mid y_0, a, w)}{p(z \mid y_0, a, w)},$$

and $\log(\text{OR}(Z, Y \mid A, W)) = 0 \iff Z \perp\!\!\!\perp Y \mid A, W$

Could also use, e.g., generalized covariance measure (Shah and Peters 2020)

A concrete example of assumptions $\mathcal{A}$ and β_k

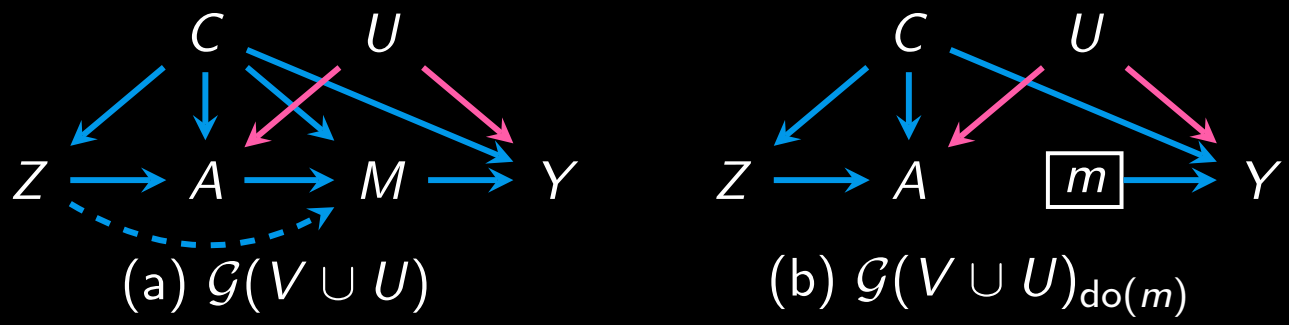


- Say we want to test correctness of backdoor adjustment sets:

$$\mathcal{M}_1 : \{C_1, C_2, C_3\}$$
$$\mathcal{M}_2 : \{C_1, C_2, C_3, C_4, C_5\}$$
$$\mathcal{M}_3 : \{C_1, C_2, C_3, C_4\}$$

- Per previous slide, only
  $\beta_1 = \log(\text{OR}(Z, Y \mid A, C_1, C_2, C_3)) = 0$ and this implies
  $\{C_1, C_2, C_3\}$ is a valid adjustment set

Aside: Tests for other popular models



(a) $\mathcal{G}(V \cup U)$ (b) $\mathcal{G}(V \cup U)_{\text{do}(m)}$

- From Bhattacharya and Nabi 2022, under Verma faithfulness and similar causal ordering + relevance assumptions as 2 slides ago, the Verma constraint $Z \perp\!\!\!\perp Y \mid C$ in $\frac{p(Z,C,A,M,Y)}{p(M|A,Z,C)}$ implies the frontdoor model is valid
- From the same paper, if the Verma constraint holds and $M \perp\!\!\!\perp Z \mid A, C$, then Z is a valid IV—different over-identifying restriction that allows testing graphical assumptions of IV
- Xie et al. 2024 show tetrad constraints can be used to test assumptions of proximal causal inference model

Back to main idea for weighted triangulation

- Suppose, under some set of assumptions $\mathcal{A}$, there exist observed data parameters β_k for each $\mathcal{M}_k$, such that

$$\beta_k = 0 \implies \mathcal{M}_k \text{ is correct}$$

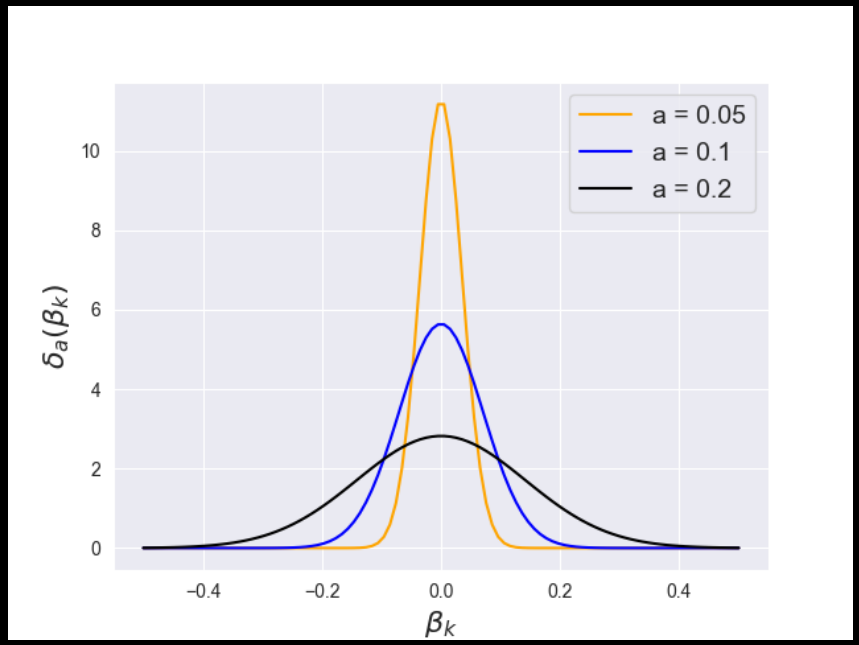
- Then consider the following naive triangulation functional,

$$\psi_{\text{naive}} = \frac{\sum_{k=1}^{K} \mathbb{I}(\beta_k = 0) \cdot \psi_k}{\sum_{j=1}^{K} \mathbb{I}(\beta_j = 0)}$$

  This is causally robust, because $\psi_{\text{naive}} = \theta$ if at least one model $\mathcal{M}_k$ is correct and testable from observed data

- However, this won't work in practice because our estimates $\beta_{k,n}$ are unlikely to be exactly 0. We also want to avoid explicit model selection based on $\beta_{k,n}$ (post-selection issues)

Main idea for weighted triangulation



- We can smooth $\mathbb{I}(\beta_k = 0)$ with Gaussian kernels of the form

$$\delta_a(\beta_k) = \frac{1}{|a|\sqrt{\pi}} e^{-(\beta_k/a)^2}$$

  So higher weight is assigned as $\beta_k \to 0$

- The parameter a controls sharpness of the approximation

Triangulation functional

- Based on these ideas, our triangulation functional is

$$\psi = \sum_{k=1}^{K} w_k \psi_k, \quad \text{where} \quad w_k = \frac{\delta_a(\beta_k)}{\sum_{j=1}^{K} \delta_a(\beta_j)}$$

- The robustness property of this functional is more complicated, but can be formalized in an intuitive manner

Properties of the triangulation functional

Theorem

Suppose assumptions $\mathcal{A}$ hold. Let $\mathcal{C} \subseteq \{1, \ldots, K\}$ and $\mathcal{I} = \{1, \ldots, K\} \setminus \mathcal{C}$ be the subsets of indices k such that model $\mathcal{M}_k$ is correct and incorrect, respectively. Then

$$|\psi - \theta| \leq \frac{\max_k |\psi_k - \theta|}{1 + D_a} \tag{1}$$

where $D_a = \left[\sum_{k \in \mathcal{C}} \delta_a(\beta_k)\right] / \left[\sum_{k \in \mathcal{I}} \delta_a(\beta_k)\right]$. Further, if at least one model $\mathcal{M}_k$ is correct and testable using the observed data, then $D_a \geq e^{\varepsilon^2/a^2}/|\mathcal{I}|$, where $\varepsilon = \min_{k \in \mathcal{I}} |\beta_k|$.

The first part of the theorem describes how putting more weight on correct models can make the bias between ψ and θ small.
The second part describes a kind of causal robustness

Inference procedure summary

- Similar to our robust hypothesis test, we rely on joint convergence in distribution of AL estimators
- Influence function-based estimators for odds ratios $\mathrm{OR}(Z, Y \mid A, W)$ are provided in Tchetgen Tchetgen et al. 2010. These are AL under doubly robust conditions on $\eta := p(Z \mid Y = y_0, A, W)$ and $\zeta := p(Y \mid Z = z_0, A, W)$
- In case influence functions have not been derived for some complicated functionals ψ_k and β_k, we also provide some contingency plans based on bootstrap (Efron and Tibshirani 1994) and subsampling (Politis et al. 2001)

I'm grateful to Daniel Malinsky for helpful discussions on DR estimation of odds ratios, and Hyunseung Kang and Oliver Dukes for helpful discussions on subsampling

Inference procedure with AL estimators

- Recall triangulation estimator is,

$$\psi_n = \sum_{k=1}^{K} w_{k,n}\psi_{k,n}, \text{ for } w_{k,n} = \frac{\delta_a(\beta_{k,n})}{\lambda_n + \sum_j \delta_a(\beta_{j,n})}$$

- By the delta method

$$n^{1/2}(\psi_n - \psi) \xrightarrow{d} N(0, \gamma^T \Sigma \gamma),$$

- The partial derivatives of ψ in γ are $\frac{\partial \psi}{\partial \psi_k} = w_k$ and

$$\frac{\partial \psi}{\partial \beta_k} = \frac{2\psi_k \beta_k w_k}{a^2} \left(\sum_{j \neq k} w_j - \frac{\lambda_n + \sum_{j \neq k} \delta(\beta_j)}{\lambda_n + \sum_{j=1}^{K} \delta(\beta_j)} \right)$$

- Can construct a $(1-\alpha)$ CI as

$$\psi_n \pm z_{1-\alpha/2}\, \widehat{\mathrm{SE}}(\psi_n) \text{ for } \widehat{\mathrm{SE}}(\psi_n) = \sqrt{\gamma_n^\top \Sigma_n \gamma_n / n}$$

Simulation study

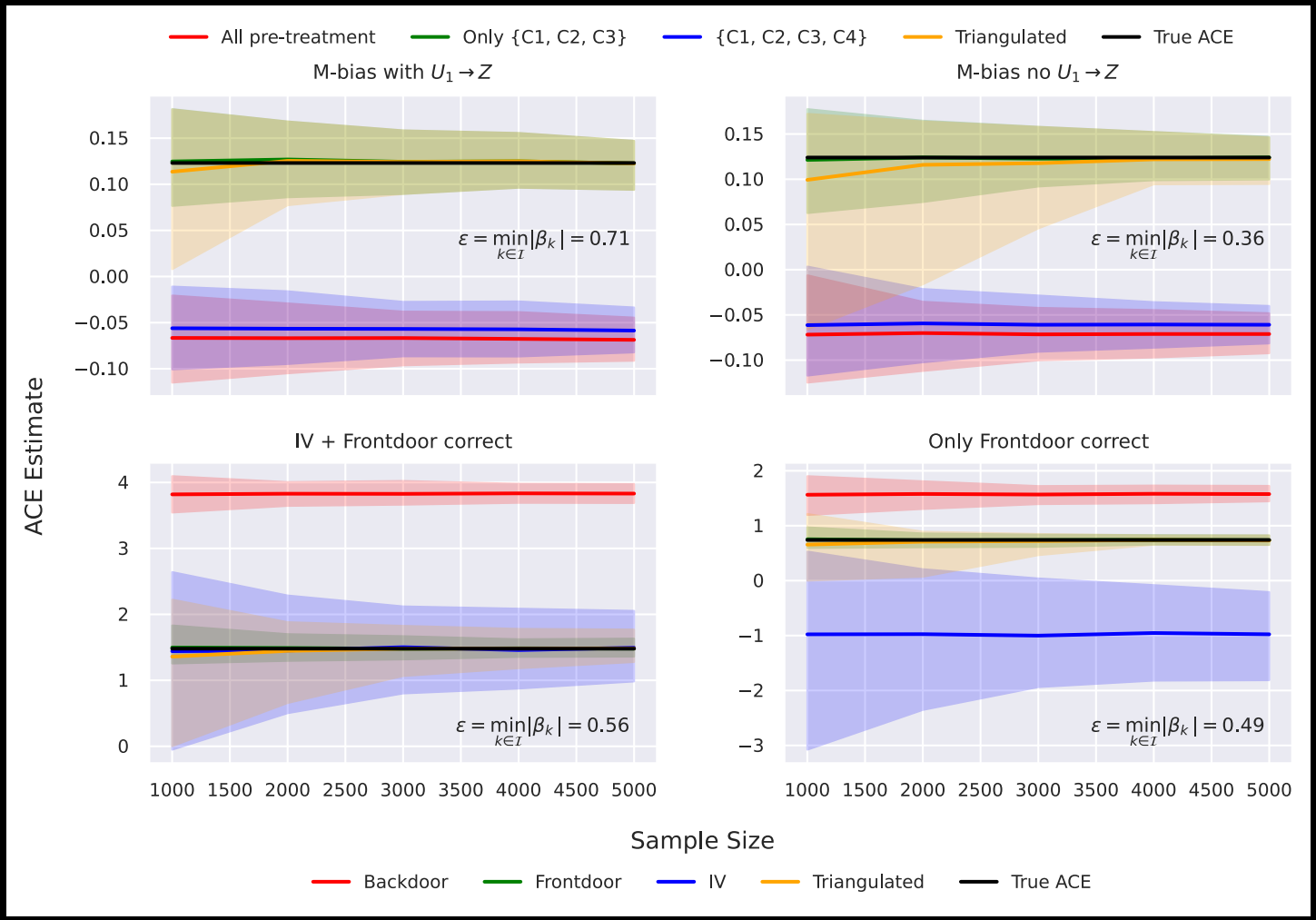


Figure: Point estimates averaged over 200 trials; shaded bands correspond to 2.5 and 97.5 percentiles. The estimator has 95% coverage of *both* ψ and θ at $n = 5000$, even though ψ has some small bias.

Simulation study takeaway

- Plurality of models can agree on the wrong effect. Since our procedure does not rely on plurality correctness, we are fine as long as at least one model is correct and testable
- Besides robustness, it is sometimes beneficial to triangulate if the single model you were considering is correct but inefficient

Application to Framingham Heart Study

Method	$\widehat{\textbf{ACE}}$	$\beta_{k,n}, w_{k,n}$
Backdoor	$0.086\ (0.048, 0.115)$	$-0.04, 0.48$
Frontdoor	$0.011\ (0.006, 0.016)$	$-0.02, 0.52$
IV	$-3.12\ (-14.7, 11.09)$	$-0.41, \approx 0$
Triangulation	$0.047(0.026, 0.062)$	–

Table: Z = educational attainment, C = sex, and M = hypertension.

- We estimate the effect of blood glucose on coronary heart disease. We treat glucose levels as a binary treatment, with $A = 1$ corresponding to high bood glucose. Y is also binary
- We intentionally adjust for only one confounder C, so that the assumptions of backdoor in particular are difficult to justify

Summary

- Proposed a test that is asymptotically valid as long as at least one causal model is correct
- We proposed a triangulation functional for effect estimation that exhibits a certain robustness property based on testability of causal models from observed data
- The second proposal relies on a tuning parameter a. This could be selected in a data-adaptive manner, or used for sensitivity analysis
- Many open interesting problems—what if no model is exactly correct à la "all models are wrong, but some are useful"

References I

- Angrist, Joshua D et al. (1996). "Identification of causal effects using instrumental variables". In: *Journal of the American Statistical Association* 91.434, pp. 444–455.
- Bang, Heejung and James M Robins (2005). "Doubly robust estimation in missing data and causal inference models". In: *Biometrics* 61.4, pp. 962–973.
- Bhattacharya, Rohit and Razieh Nabi (2022). "On testability of the front-door model via Verma constraints". In: *Uncertainty in Artificial Intelligence*. PMLR, pp. 202–212.
- Chen, Hua Yun (2007). "A semiparametric odds ratio model for measuring association". In: *Biometrics* 63.2, pp. 413–421.
- Ding, Peng and Luke W Miratrix (2015). "To adjust or not to adjust? Sensitivity analysis of M-bias and butterfly-bias". In: *Journal of Causal Inference* 3.1, pp. 41–57.
- Drton, Mathias and Han Xiao (2016). "Wald tests of singular hypotheses". In: *Bernoulli* 22.1, pp. 38–59.

References II

- Efron, Bradley and Robert J Tibshirani (1994). *An Introduction To The Bootstrap*. Chapman and Hall/CRC.
- Entner, Doris et al. (2013). "Data-driven covariate selection for nonparametric estimation of causal effects". In: *Artificial Intelligence and Statistics*. PMLR, pp. 256–264.
- Heesen, Remco et al. (2019). "Vindicating methodological triangulation". In: *Synthese* 196, pp. 3067–3081.
- Kang, Hyunseung et al. (2016). "Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization". In: *Journal of the American statistical Association* 111.513, pp. 132–144.
- Lawlor, Debbie A et al. (2016). "Triangulation in aetiological epidemiology". In: *International Journal of Epidemiology* 45.6, pp. 1866–1886.
- Miles, Caleb H and Antoine Chambaz (2021). "Optimal tests of the composite null hypothesis arising in mediation analysis". In: *arXiv preprint arXiv:2107.07575*.

References III

- Pearl, Judea (1995). "Causal diagrams for empirical research.". In: *Biometrika* 82.4, pp. 669–710.
- — (2009). "Remarks on the method of propensity score". In: *Statistics in Medicine* 28, pp. 1415–1416.
- Politis, Dimitris N et al. (2001). "On the asymptotic theory of subsampling". In: *Statistica Sinica*, pp. 1105–1124.
- Rakshit, Prabrisha et al. (2025). "Adaptive proximal causal inference with some invalid proxies". In: *arXiv preprint arXiv:2507.19623*.
- Rosenbaum, Paul R (2010). "Evidence factors in observational studies". In: *Biometrika* 97.2, pp. 333–345.
- — (2011). "Some approximate evidence factors in observational studies". In: *Journal of the American Statistical Association* 106.493, pp. 285–295.
- Shah, Rajen D and Jonas Peters (2020). "The hardness of conditional independence testing and the generalized covariance measure". In: *The Annals of Statistics* 48.3, pp. 1514–1538.

References IV

- Shapland, Chin Yang et al. (2024). "A quantitative approach to evidence triangulation: development of a framework to address rigour and relevance". In: *medRxiv*, pp. 2024–09.
- Sjölander, Arvid (2009). "Propensity scores and M-structures". In: *Statistics in medicine* 28.9, pp. 1416–1420.
- Sun, Baoluo et al. (2021). "On multiply robust Mendelian randomization (MR2) with many invalid genetic instruments". In: *medRxiv*, pp. 2021–10.
- Tchetgen Tchetgen, Eric J et al. (2010). "On doubly robust estimation in a semiparametric odds ratio model". In: *Biometrika* 97.1, pp. 171–180.
- Xie, Feng et al. (2024). "Automating the Selection of Proxy Variables of Unmeasured Confounders". In: *Forty-first International Conference on Machine Learning*.
- Yao, Minhao et al. (2024). "Deciphering proteins in Alzheimer's disease: A new Mendelian randomization method integrated with AlphaFold3 for 3D structure prediction". In: *Cell Genomics* 4.12.