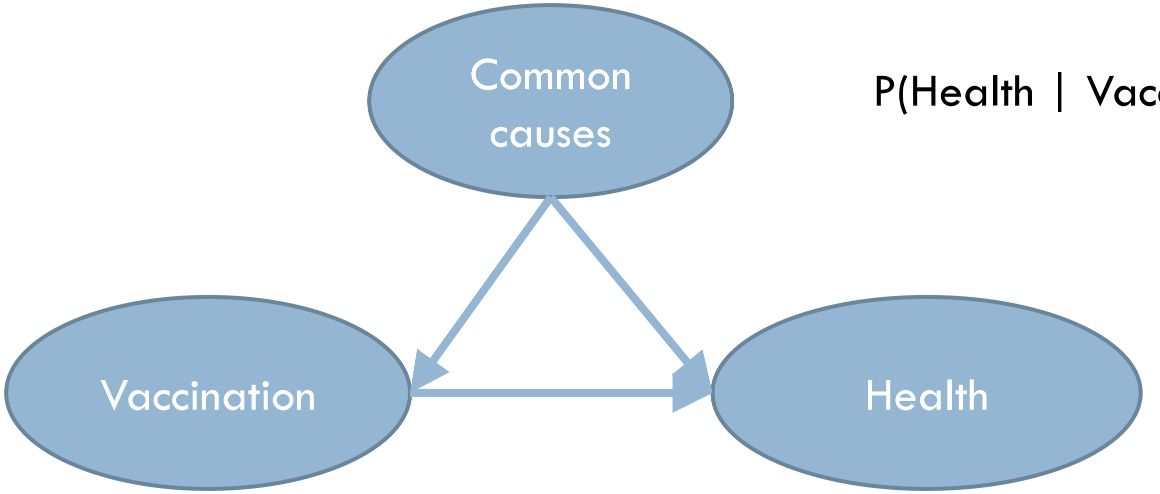
CAUSAL INFERENCE THROUGH A WITNESS PROTECTION PROGRAM

Ricardo Silva
Department of Statistical Science
Centre for Computational Statistics and Machine Learning
ricardo@stats.ucl.ac.uk

Joint work with Robin Evans (Statistics, Oxford)

Background: Causal Inference

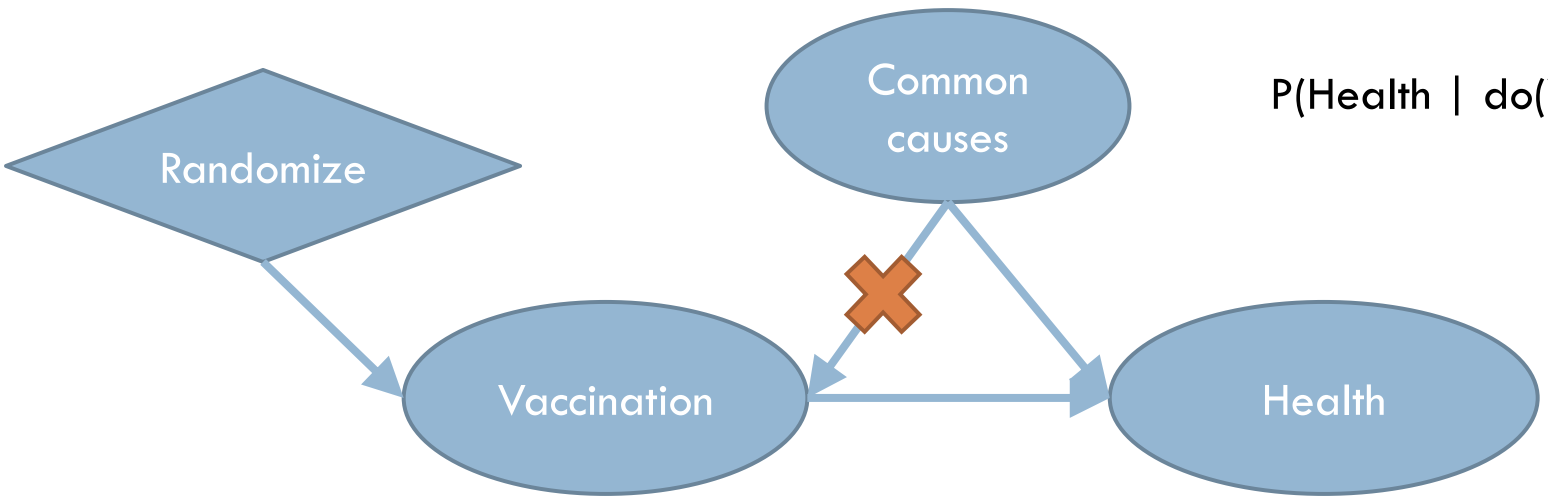
- The task: estimate the effect of an intervention
  - Medical treatments, public policy, gene knock-outs and so on
- Gold standard: randomized controlled trial



P(Health | Vaccination)

Background: Causal Inference

- The task: estimate the effect of an intervention
  - Medical treatments, public policy, gene knock-outs and so on
- Gold standard: randomized controlled trial



P(Health | do(Vaccination))

Goals of this talk

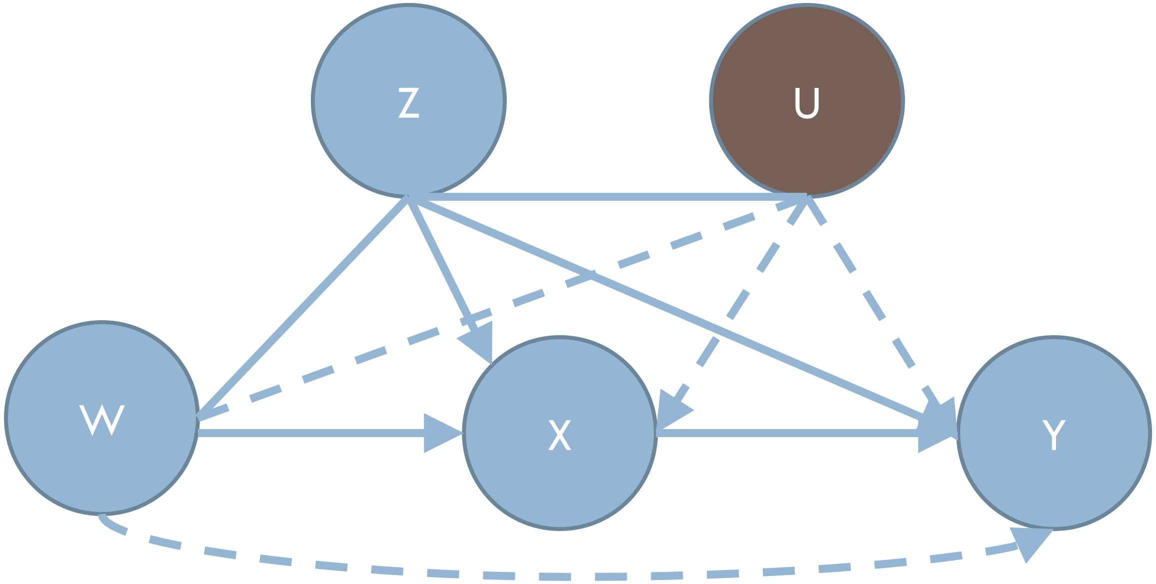
- Given binary X precedes binary Y causally, estimate average causal effect (ACE) **using observational data**

$$ACE \equiv E[Y \mid do(X = 1)] - E[Y \mid do(X = 0)] =$$

$$P(Y = 1 \mid do(X = 1)) - P(Y = 1 \mid do(X = 0))$$

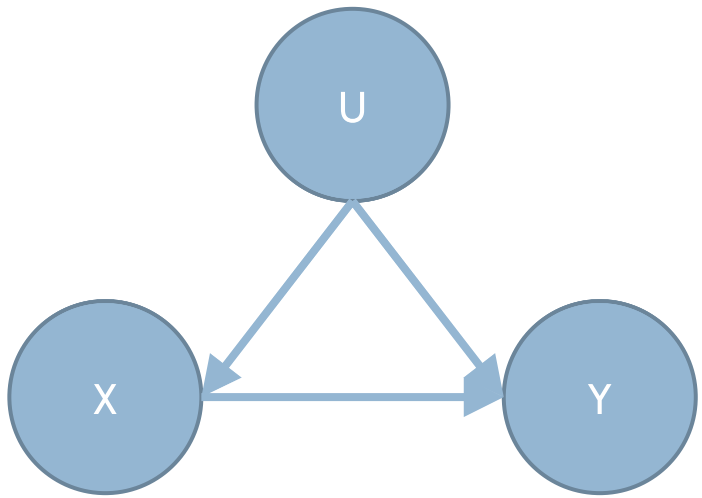
Goal

- To get an estimate of **bounds** of the ACE
- Rely on the identification of an auxiliary variable W (**witness**), an auxiliary set Z (**background set**), and **assumptions about strength of dependencies on latent variables**



Observational Studies: Tricks of the Trade

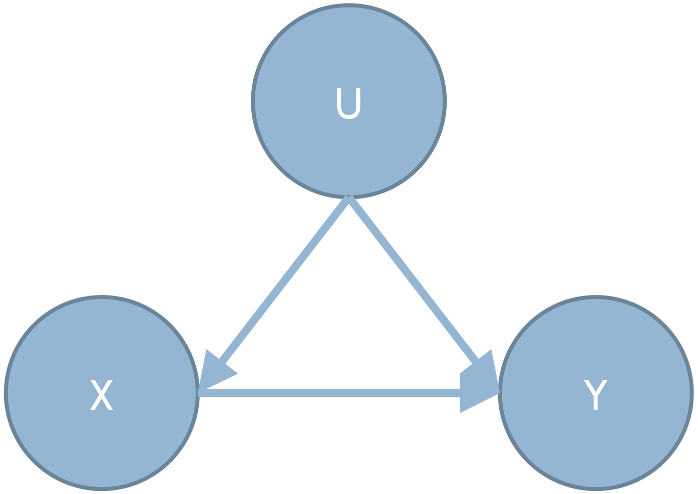
- Backdoor adjustment (Pearl and others):



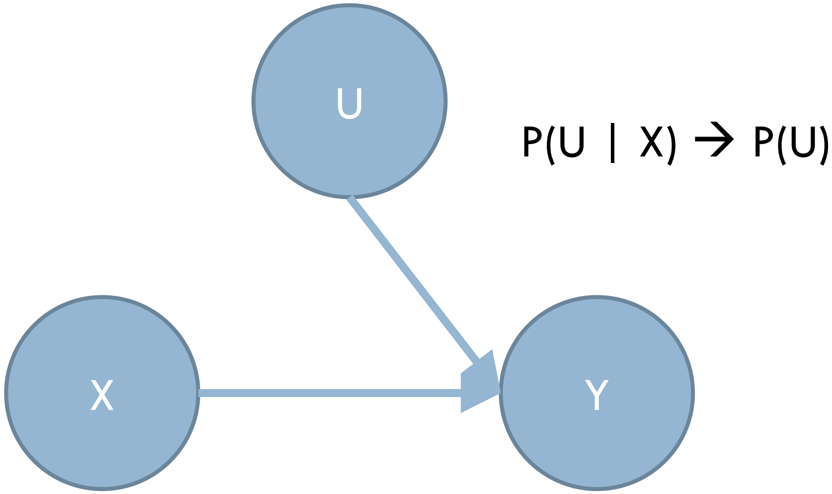
- $P(Y = 1 \mid do(X = x)) = \sum_{u} P(Y = 1 \mid x, u)P(u)$

Observational Studies: Tricks of the Trade

- Backdoor adjustment (justification):



Natural state

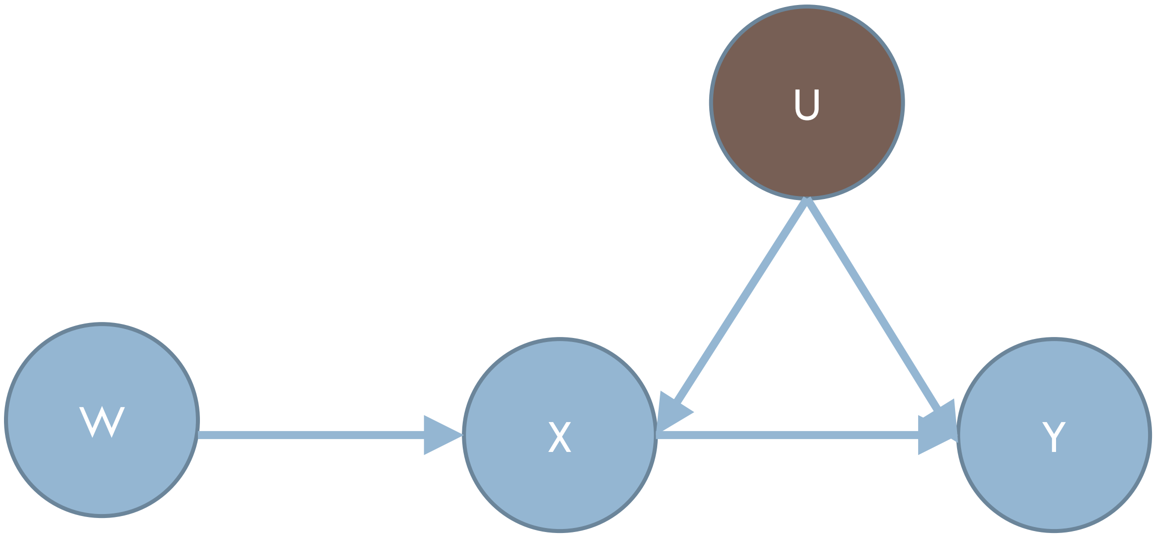


Controlled state

- Problem: where causal knowledge about U comes from?
- Problem: what if we do not have P(Y | X, U), P(U)?

Observational Studies: Tricks of the Trade

- Instrumental variables



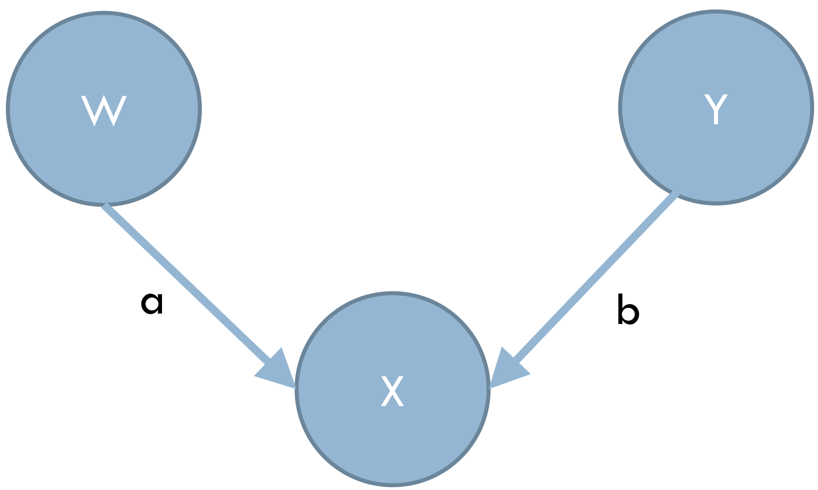
$$L_{P(Y,\, X \mid W)} \leq \text{ACE} \leq U_{P(Y,\, X \mid W)}$$

Exploiting Independence Constraints

- **Faithfulness** provides a way of sometimes finding a point estimator
  - Faithfulness means independence in probability iif "structural" independence (Spirtes et al., 1993)

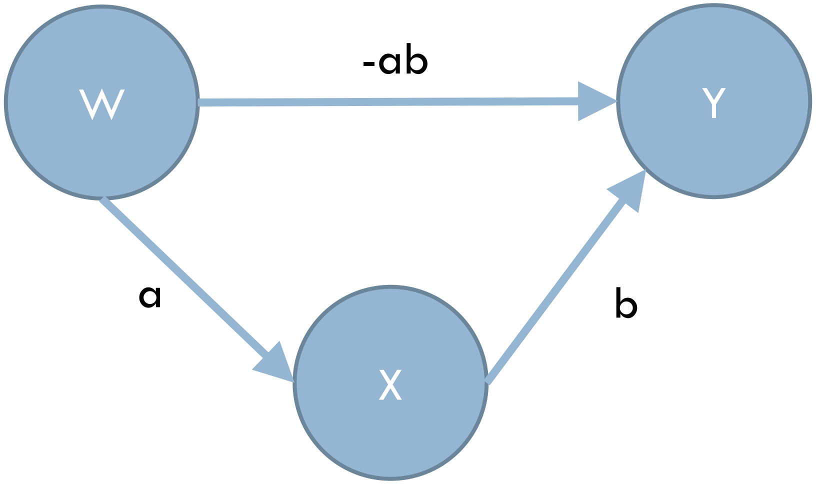
Faithfulness

- W independent of Y, but not when given X: conclude the following (absentia hidden common causes)



(Lack of) Faithfulness

- W independent of Y, but not when given X: different structure



Exploiting Independence Constraints

- In what follows, we will assume that we have access to a set of variables which we know are not effects of neither X nor Y

The Problem with Naïve Back-Door Adjustment

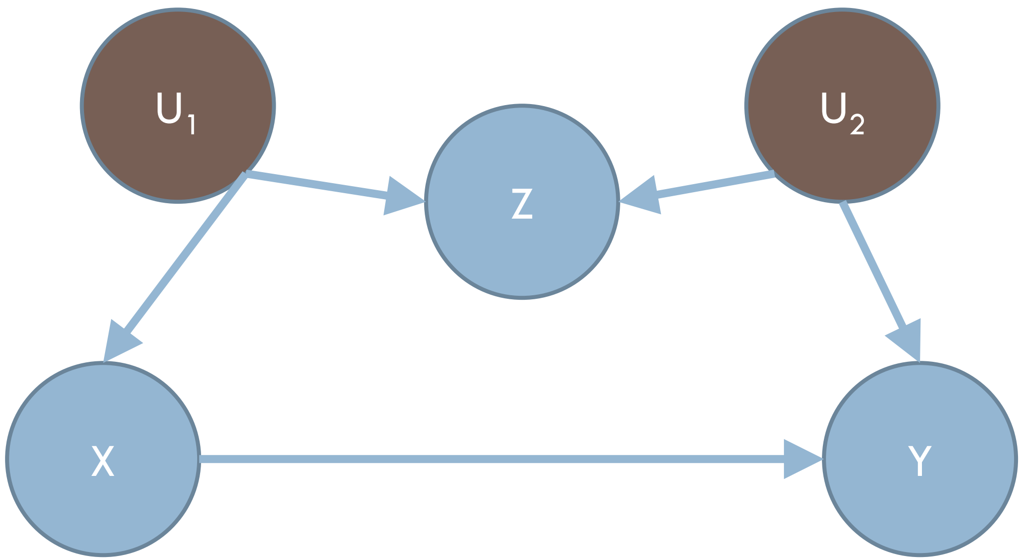
- It is not uncommon in applied sciences to posit that, given a large number of covariates Z that are plausible common causes of X and Y, we should adjust for all

$$P_{est}(Y = 1 \mid do(X = x)) = \sum_{z} P(Y = 1 \mid x, z)P(z)$$

- Even if there are remaining unmeasured confounders, a common assumption is that adding elements of Z will in general decrease bias $ACE_{true} - ACE_{hat}$

The Problem with Naïve Back-Door Adjustment

- Example of failure:

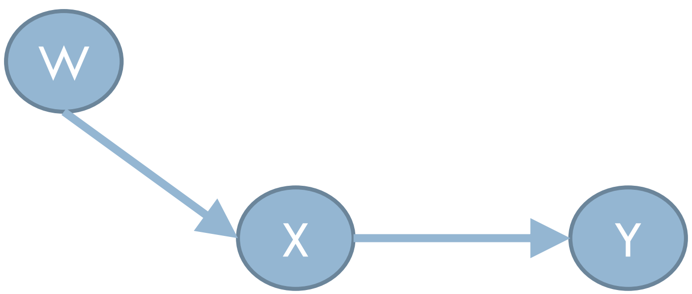


$$P(Y = 1 \mid do(X = x)) = P(Y = 1 \mid X = x) \neq \sum P(Y = 1 \mid x, z)P(z)$$

Pearl (2009). Technical Report R-348

Exploiting Faithfulness: A Very Simple Example

- W not caused by X nor Y, X → Y
- W $\not\perp\!\!\!\perp$ X, W $\perp\!\!\!\perp$ Y | X + Faithfulness. Conclusion?



No unmeasured confounding

- Naïve estimator vindicated:

  $$ACE = P(Y = 1 \mid X = 1) - P(Y = 1 \mid X = 0)$$

- This super-simple nugget of causal information has found some practical uses on large-scale problems

A Very Simple Example

- Consider "the genotype at a fixed locus *L* is a random variable, whose random outcome occurs before and independently from the subsequently measured expression values"
- Find genes Ti, Tj such that L → Ti → Tj

Chen, Emmert-Streib12 and Storey (2007) Genome Biology, 8:R219

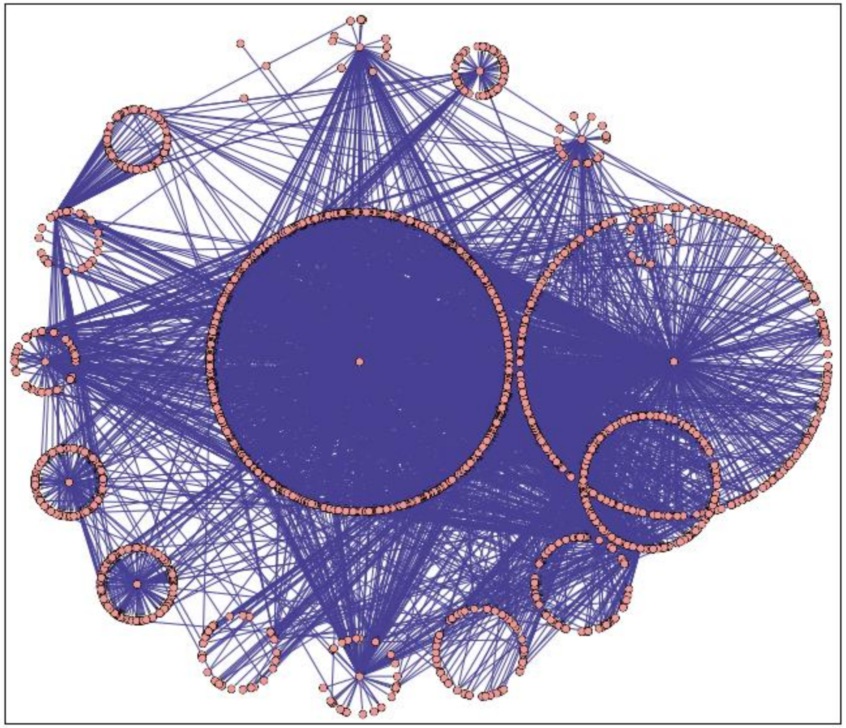


Figure 2
A transcriptional regulatory network drawn from a Trigger probability threshold of 90%. The network consists of 4,394 genes, 2,145 causal relationships, and 127 causal genes. Genes are represented by orange circles and causal relationships are represented by directed edges with black arrows.

Entner et al.'s Background Finder

- Entner, Hoyer and Spirtes (2013) AISTATS: two simple rules based on finding a **witness** W for a correct **admissible background set** Z
  - Generalizes "chain models" W → X → Z

R1: If there exists a variable $w \in \mathcal{W}$ and a set $\mathcal{Z} \subseteq \mathcal{W} \setminus \{w\}$ such that

(i) $w \not\perp\!\!\!\perp y \mid \mathcal{Z}$, and

(ii) $w \perp\!\!\!\perp y \mid \mathcal{Z} \cup \{x\}$

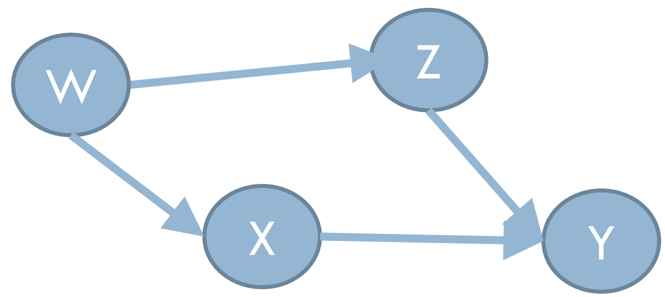
then infer '±' and give $\mathcal{Z}$ as an admissible set.

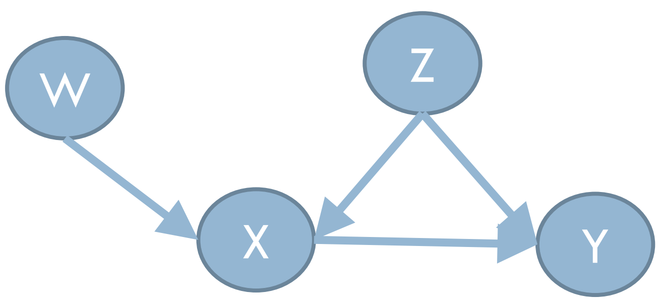
Rule 1: Illustration

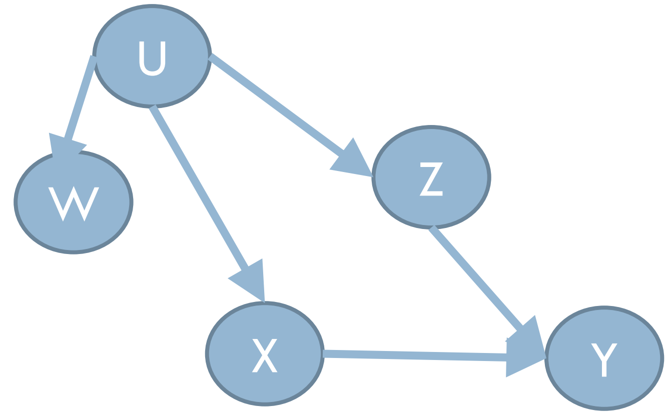
R1: If there exists a variable $w \in \mathcal{W}$ and a set $\mathcal{Z} \subseteq \mathcal{W} \setminus \{w\}$ such that

(i) $w \not\perp\!\!\!\perp y \mid \mathcal{Z}$, and
(ii) $w \perp\!\!\!\perp y \mid \mathcal{Z} \cup \{x\}$

then infer '$\pm$' and give $\mathcal{Z}$ as an admissible set.







- Note again the necessity of the dependence of W and Y

Reverting the Question

- What if instead of using W to find Z to make an adjustment by the back-door criterion, we find a Z to allow W to be an instrumental variable to find bounds on the ACE?

Why do We Care?

- A way to weaken the faithfulness assumption
  - Suppose also by "independence", we might mean "weak dependence" (and by "dependence", we might mean "strong dependence")
- How would interpret the properties of W in this case, given Rule 1?

R1: If there exists a variable $w \in \mathcal{W}$ and a set $\mathcal{Z} \subseteq \mathcal{W} \setminus \{w\}$ such that

(i) $w \not\perp\!\!\!\perp y \mid \mathcal{Z}$, and

(ii) $w \perp\!\!\!\perp y \mid \mathcal{Z} \cup \{x\}$

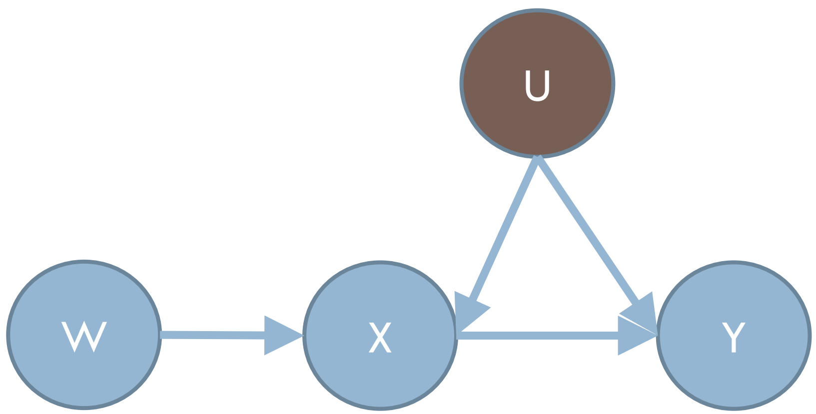
then infer '±' and give $\mathcal{Z}$ as an admissible set.

Modified Setup: Main Assumption Statement

- Given Rule 1, assume W is a “conditional IV for X → Y” in the sense that given Z
  - All active paths between W and X are into X
  - There is no “strong direct effect” of W on Y
  - There are no “strong active paths” between W and X, nor W and Y, through common ancestors of X and Y
- The definition of “strong effect/path” creates free parameters we will have to deal with. More on that later

Motivation

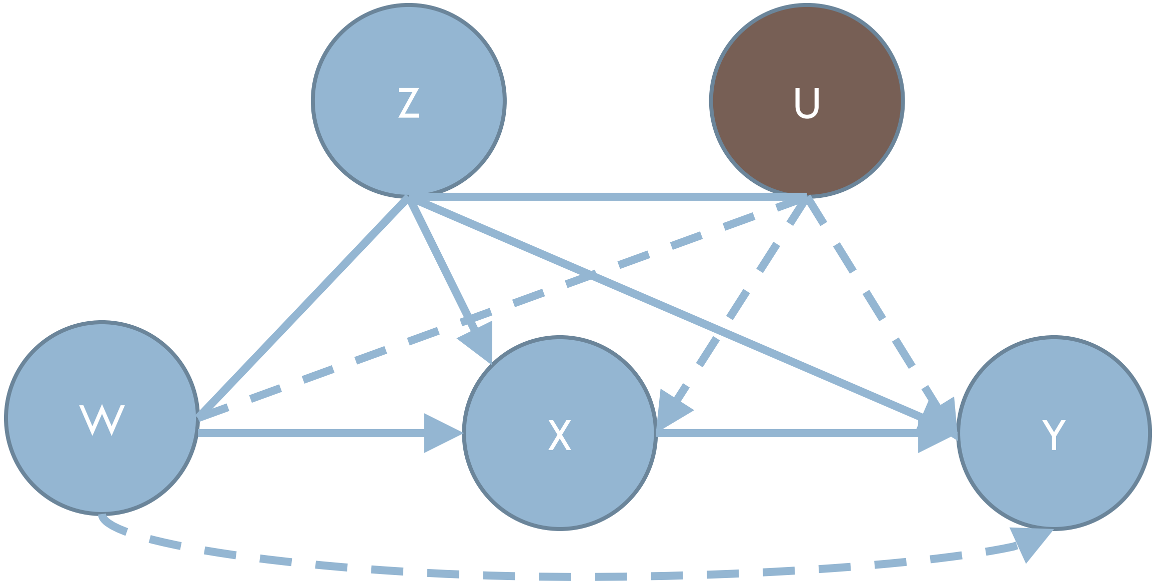
- Bounds on the ACE in the "standard IV model" can be *quite wide even when* $W \perp\!\!\!\perp Y \mid X$



- This means faithfulness can be quite a strong assumption, and/or "worst-case" analysis can be quite conservative

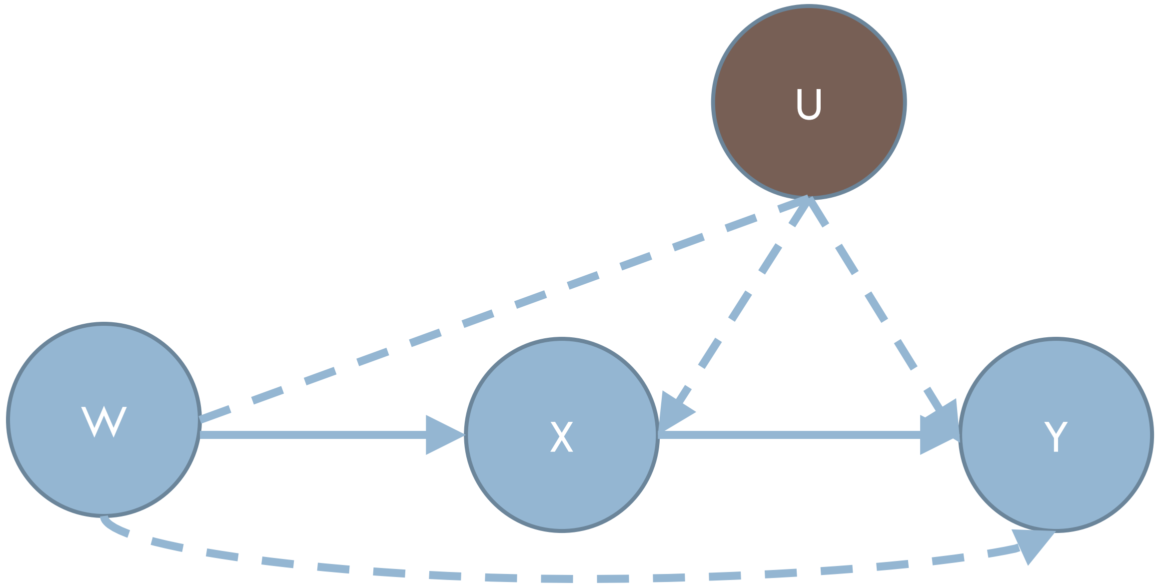
Motivation

- Our analysis can be seen as a way of bridging the two extremes of point estimators of faithfulness analysis and IV bounds without effect constraints
- Notice: this does ***not*** mean making stronger assumptions than the standard IV model



Stating Assumptions

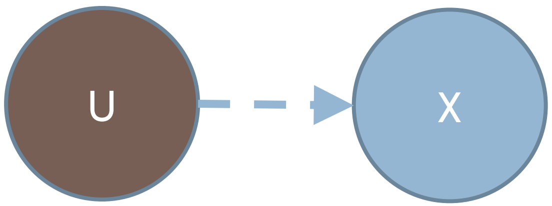
- Some notation first, ignoring Z for now:



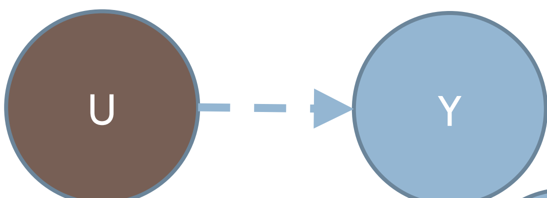
$$
\begin{aligned}
\zeta^{\star}_{yx.w} &\equiv P(Y = y, X = x \mid W = w, U) \\
\eta^{\star}_{xw} &\equiv P(Y = 1 \mid X = x, W = w, U) \\
\delta^{\star}_{w} &\equiv P(X = 1 \mid W = w, U)
\end{aligned}
$$

Stating Assumptions

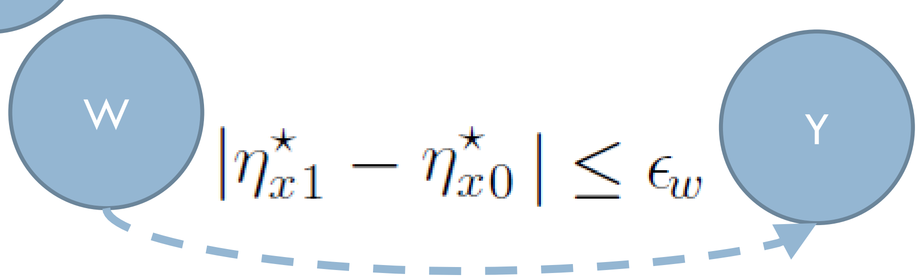
$$
\begin{aligned}
\zeta^\star_{yx.w} &\equiv P(Y = y, X = x \mid W = w, U) \\
\eta^\star_{xw} &\equiv P(Y = 1 \mid X = x, W = w, U) \\
\delta^\star_{w} &\equiv P(X = 1 \mid W = w, U)
\end{aligned}
$$



$$|\delta^\star_w - P(X = 1 \mid W = w)| \leq \epsilon_x$$

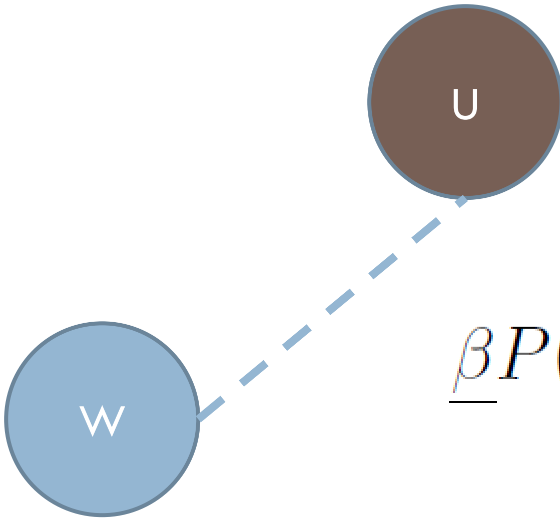


$$|\eta^\star_{xw} - P(Y = 1 \mid X = x, W = w)| \leq \epsilon_y$$



$$|\eta^\star_{x1} - \eta^\star_{x0}| \leq \epsilon_w$$

Stating Assumptions



$$\underline{\beta}P(U) \leq P(U \mid W = w) \leq \bar{\beta}P(U)$$

Relation to Observations

$$\begin{aligned} \zeta^\star_{yx.w} &\equiv P(Y = y, X = x \mid W = w, U) \\ \eta^\star_{xw} &\equiv P(Y = 1 \mid X = x, W = w, U) \\ \delta^\star_{w} &\equiv P(X = 1 \mid W = w, U) \end{aligned}$$

- Let $\zeta_{yx.w}$ be the expectation of the first entry by P(U | W): this is P(Y = y, X = x | W = w)
- Similarly, let η_{xw} be the expectation of the second entry: this is P(Y = 1 | do(X = x), W = w)

Context

- The parameterization given was originally exploited by Dawid (2000) and Ramsahai (2012)
- It provides an alternative to the structural equation model parameterization of Balke and Pearl (1997)
- Both approaches work by mapping the problem of testing the model and bounding the ACE by a linear program
- We build on this strategy, with some generalizations

Estimation

- Simpler mapping on $(\delta^*, \eta^*) \rightarrow P(W, X, Y \mid U)$, marginalized, gives constraints on $\zeta \equiv P(W, X, Y)$
- Test whether constraints hold, if not provide no bounds
- Plug-in estimates for ζ to get (ζ, η) polytope. Find upper bounds and lower bounds on the ACE by solving linear program and maximizing/minimizing objective function

$$f(\eta) = (\eta_{11} - \eta_{01})P(W = 1) + (\eta_{10} - \eta_{00})P(W = 0)$$

Coping with Non-linearity

- Notice that because of constraints such as

$$|\delta_w^\star - P(X = 1 \mid W = w)| \leq \epsilon_x$$

there will be non-linear constraints in $\zeta \equiv P(W, X, Y)$

- The implied constraints are still linear in $\eta \equiv P(Y \mid do(X), W)$. So linear programming formulation still holds, treating ζ as a constant.
  - Non-linearity on ζ can be a problem for estimation of ζ and derivation of confidence intervals. We will describe later a Bayesian approach that does that simply by rejection sampling

Algorithm

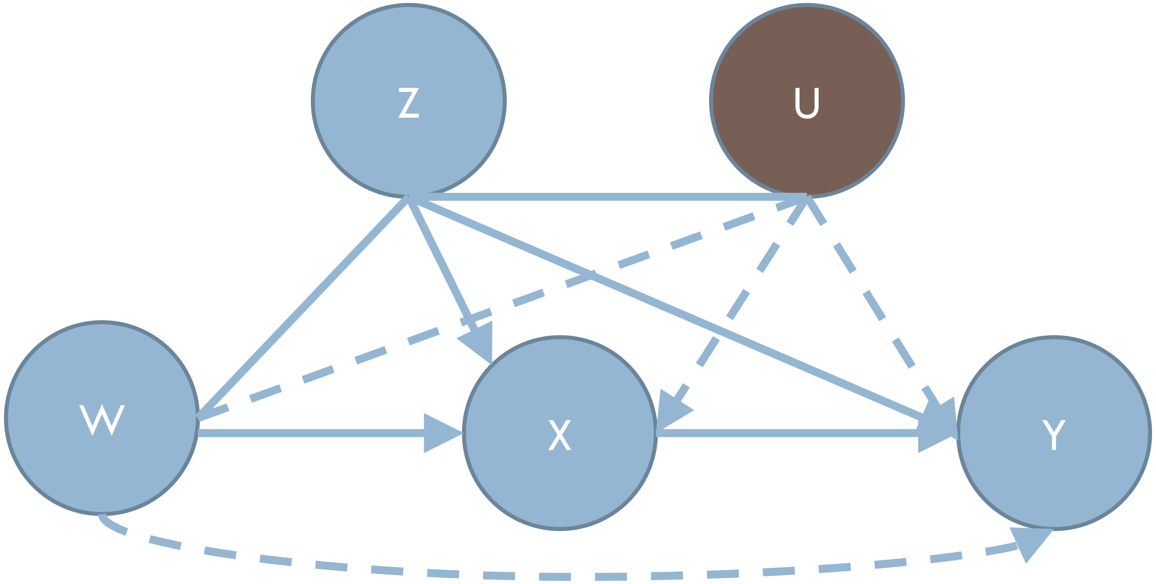
In what follows, we assume dimensionality of Z is small, |Z| < 10

input : Binary data matrix $\mathcal{D}$; set of relaxation parameters θ; covariate index set $\mathcal{W}$; cause-effect indices X and Y
output: A list of pairs (witness, admissible set) contained in $\mathcal{W}$

$\mathcal{L} \leftarrow \emptyset$;
for *each* $W \in \mathcal{W}$ **do**
  for *every admissible set* $\mathbf{Z} \subseteq \mathcal{W} \backslash \{W\}$ *identified by* W *and* θ *given* $\mathcal{D}$ **do**
    $\mathcal{B} \leftarrow$ posterior over upper/lowed bounds on the ACE as given by $(W, \mathbf{Z}, X, Y, \mathcal{D}, \theta)$;
    if *there is no evidence in* $\mathcal{B}$ *to falsify the* $(W, \mathbf{Z}, \theta)$ *model* **then**
      $\mathcal{L} \leftarrow \mathcal{L} \cup \{\mathcal{B}\}$;
    end
  end
end
return $\mathcal{L}$

Recap: So far, everything in the population

- “Rely on the identification of an auxiliary variable W (**witness**), an auxiliary set Z (**background set**), and **assumptions about strength of dependencies on latent variables**”

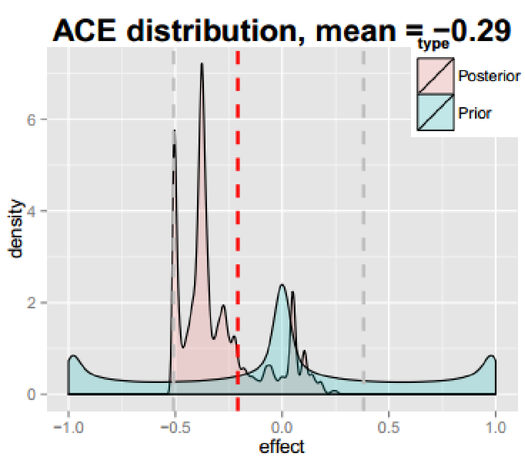


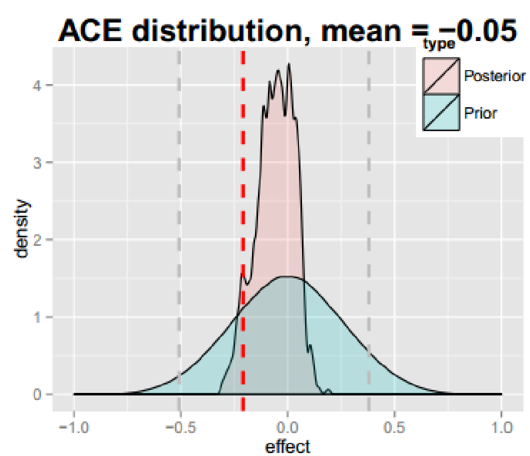
Bayesian Learning

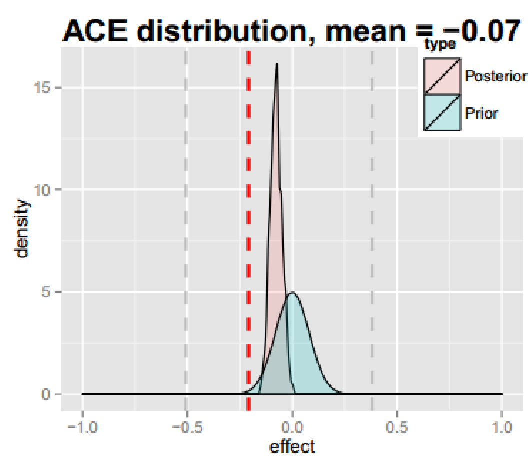
- To decide on independence, we do Bayesian model selection with a contingency table model with Dirichlet priors
- For each pair (W, Z), find posterior bounds for each configuration of Z
  - Use Dirichlet prior for ζ (for each $Z = z$), conditioned on the constraints of the model, using rejection sampling
    - Propose from unconstrained Dirichlet
  - Reject model if 95% or more of proposed parameters are rejected in the initial round of rejection sampling
  - Feed sample from the posterior of ζ into linear program to get a sample for the upper bound and lower bound

Difference wrt ACE Bayesian Learning

- How not put a prior directly on the latent variable model?
  - However, model is unidentifiable → results extremely sensitive to priors
  - Putting priors directly into ζ produces no point estimates, but avoids prior sensibility







Wrapping Up

- Finally, one is left with different posterior distributions over different bounds on the ACE
- Final step is how to summarize possibly conflicting information. Possibilities are:
  - Report tightest bound
  - Report widest bound
  - Report combined smallest lower bound with largest upper bound
  - Use "posterior of Rule 1" to pick a handful of bounds and discard others

Recap

- Invert usage of Entner's Rules towards the instrumental variable point of view
- Obtain bounds, not point estimates
- Use Bayesian inference, set up a rule to combine possibly conflicting information
- Because the framework relies on using a linear program to protect a witness variable against violations of faithfulness, we call this the *Witness Protection Program* (WPP) framework

Scaling Up

- There are four main bottlenecks:
  - The witness search procedure
  - Posterior sampling of parameters
    - Rejection criterion
    - Averaging over P(Z)
  - Running linear programs to obtain bounds (potentially expensive if done separately for each posterior sample)
- We address here problems of sampling and bound optimization, which can be solved by the same idea

Direct Polytope Manipulation

$$
\begin{aligned}
\eta_1^\star &\leq 1 \\
\eta_1^\star(1 - \delta_1^\star) &\leq 1 - \delta_1^\star \\
\eta_1 - \zeta_{11.1} &\leq 1 - (\zeta_{11.1} + \zeta_{01.1}) \quad \text{(marginalization)} \\
\zeta_{11.0} - \zeta_{11.1} &\leq 1 - (\zeta_{11.1} + \zeta_{01.1}) \quad (\text{since } \eta_1 = \eta_{10} \geq \zeta_{11.0}) \\
\zeta_{11.0} + \zeta_{01.1} &\leq 1
\end{aligned}
$$

- This is one of the "instrumental inequalities" of the standard IV model, derived directly
  - Bounding η^* by one of its extreme points
  - Modify factor in a way to map it to ζ and η, perform further manipulations
- Useful as a way of deriving symbolic bounds as a function of the extreme points of the original parameter space

Direct Polytope Manipulation

- In the accompanying paper, we describe several analytical bounds on P(Y | do(X), W) as a function of P(W, X, Y) and constraints

$$
\begin{aligned}
\omega_{xw} &\geq \kappa_{1x.w} + L^{YU}_{xw}(\kappa_{0x'.w} + \kappa_{1x'.w}) \\
\omega_{xw} &\leq 1 - (\kappa_{0x.w'} - \epsilon_w(\kappa_{0x.w'} + \kappa_{1x.w'}))/U^{XU}_{xw'} \\
\omega_{xw} - \omega_{xw'}U^{XU}_{x'w} &\leq \kappa_{1x.w} + \epsilon_w(\kappa_{0x'.w} + \kappa_{1x'.w}) \\
\omega_{xw} + \omega_{x'w} - \omega_{x'w'} &\geq \kappa_{1x'.w} + \kappa_{1x.w} - \kappa_{1x'.w'} + \kappa_{1x.w'} - \chi_{xw'}(\bar{U} + \underline{L} + 2\epsilon_w) + \underline{L}
\end{aligned}
$$

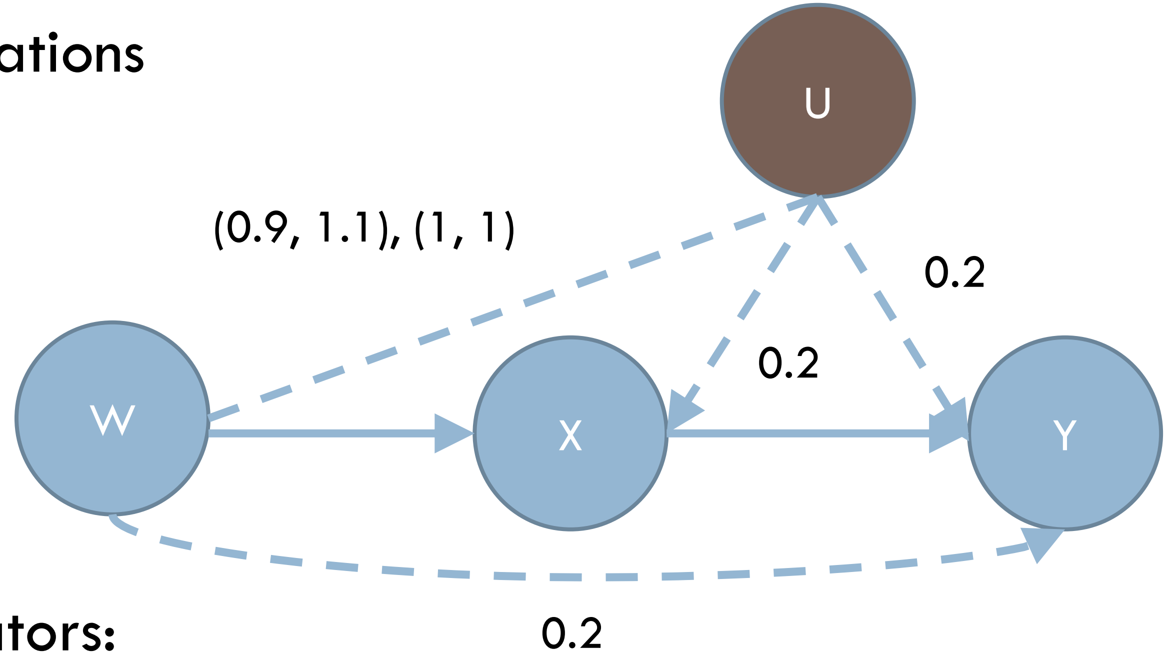
- This are used to generate relaxed (i.e., underconstrained) linear programming problems which are much more efficient to solve

Illustration: Synthetic Studies

- 4 observable nodes, “basic set”, form a pool that can generate a possible (witness, background set) pair
- 4 observable nodes form a “decoy set”: none of them should be included in the background set
- Graph structures over “basic set” + {X, Y} are chosen randomly
- Observable parents of “decoy set” are sampled from “basic set”
- Each decoy has another four latent parents, $\{L_1, L_2, L_3, L_4\}$
- Latents are mutually independent
- Each latent variable L_i uniformly chooses either X or Y as a child
- Conditional distributions are logistic regression models with pairwise interactions

Illustration: Synthetic Studies

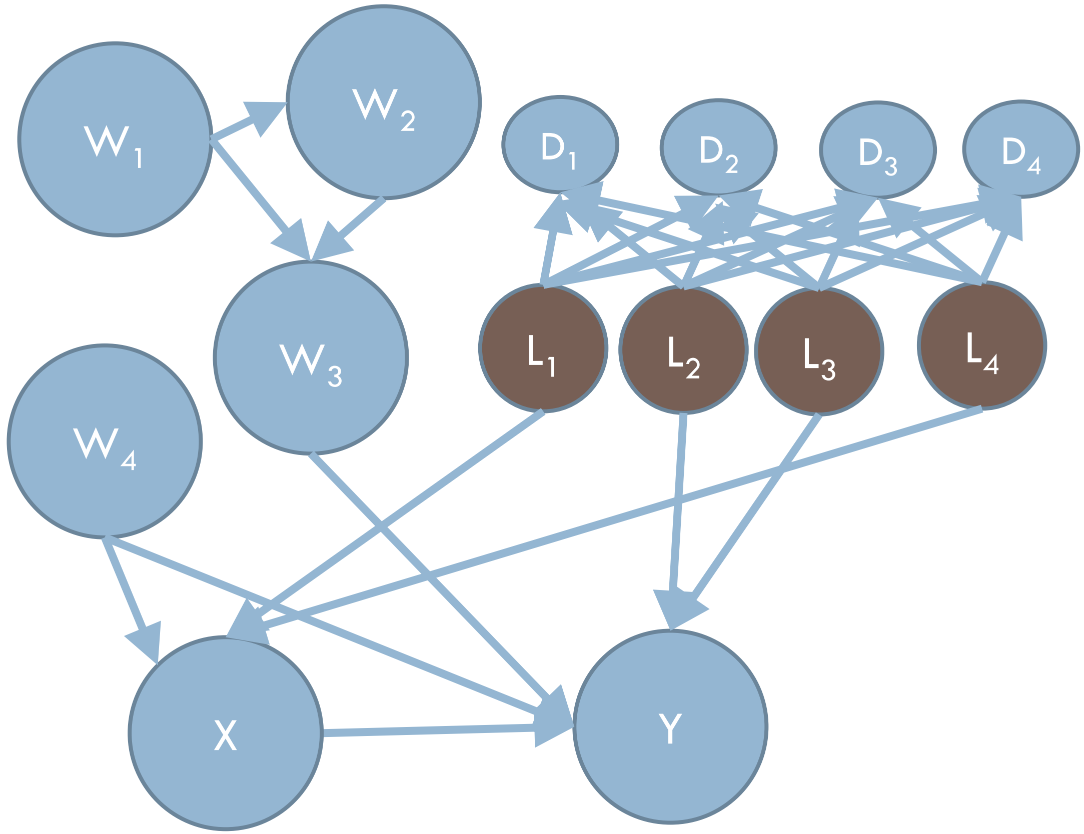
- Relaxations



- Estimators:
  - Posterior expected bounds
  - Naïve 1: back-door adjustment conditioning on everybody
  - Naïve 2: plain $P(Y = 1 \mid X = 1) - P(Y = 1 \mid X = 0)$
  - Backdoor by faithfulness

Example

- Note: no theoretical witness solution



Evaluation

- Bias definition:
  - For point estimators, just absolute value of difference between true ACE and estimate
  - For bounds, Euclidean distance between true ACE and nearest point in the bound
- Summaries (over 100 simulations):
  - Bias average
  - Bias tail mass at 0.1
    - proportion of cases where bias exceeds 0.1
- Notice difficulty of direct comparisons

Summary

Hard, Solvable: NE1 $= (0.12, 1.00)$, NE2 $= (0.02, 0.03)$

k_ϵ	Found	Faith.1		WPP1		Width1	WPP2		Width2
0.05	0.74	0.03	0.05	0.02	0.05	0.05	0.00	0.00	0.34
0.10	0.94	0.04	0.05	0.01	0.01	0.11	0.00	0.00	0.41
0.15	0.99	0.04	0.05	0.01	0.02	0.16	0.00	0.00	0.46
0.20	1.00	0.05	0.05	0.01	0.01	0.24	0.00	0.00	0.53
0.25	1.00	0.05	0.07	0.00	0.00	0.32	0.00	0.00	0.60
0.30	1.00	0.05	0.10	0.00	0.00	0.41	0.00	0.00	0.69

Easy, Solvable: NE1 $= (0.01, 0.01)$, NE2 $= (0.07, 0.24)$

k_ϵ	Found	Faith.1		WPP1		Width1	WPP2		Width2
0.05	0.81	0.03	0.02	0.02	0.04	0.04	0.00	0.01	0.34
0.10	0.99	0.02	0.02	0.01	0.02	0.09	0.00	0.00	0.40
0.15	1.00	0.02	0.01	0.00	0.00	0.17	0.00	0.00	0.46
0.20	1.00	0.02	0.01	0.00	0.00	0.24	0.00	0.00	0.54
0.25	1.00	0.02	0.01	0.00	0.00	0.32	0.00	0.00	0.61
0.30	1.00	0.02	0.01	0.00	0.00	0.41	0.00	0.00	0.67

Bias average

Bias tail mass at 0.1

Summary

Hard, Not Solvable: NE1 = (0.16, 1.00), NE2 = (0.20, 0.88)									
k_ϵ	Found	Faith.1		WPP1		Width1	WPP2		Width2
0.05	0.67	0.20	0.90	0.17	0.76	0.06	0.04	0.14	0.32
0.10	0.91	0.19	0.91	0.13	0.63	0.10	0.02	0.07	0.39
0.15	0.97	0.19	0.92	0.10	0.41	0.18	0.01	0.03	0.45
0.20	0.99	0.19	0.95	0.07	0.25	0.24	0.01	0.01	0.51
0.25	1.00	0.19	0.96	0.03	0.13	0.31	0.00	0.00	0.58
0.30	1.00	0.19	0.96	0.02	0.06	0.39	0.00	0.00	0.66
Easy, Not Solvable: NE1 = (0.09, 0.32), NE2 = (0.14, 0.56)									
k_ϵ	Found	Faith.1		WPP1		Width1	WPP2		Width2
0.05	0.68	0.13	0.51	0.10	0.37	0.05	0.02	0.07	0.33
0.10	0.97	0.12	0.53	0.08	0.28	0.10	0.01	0.05	0.39
0.15	1.00	0.12	0.52	0.05	0.17	0.16	0.01	0.03	0.46
0.20	1.00	0.12	0.53	0.03	0.08	0.23	0.01	0.03	0.52
0.25	1.00	0.12	0.48	0.02	0.05	0.31	0.00	0.02	0.59
0.30	1.00	0.12	0.48	0.01	0.04	0.39	0.00	0.01	0.65

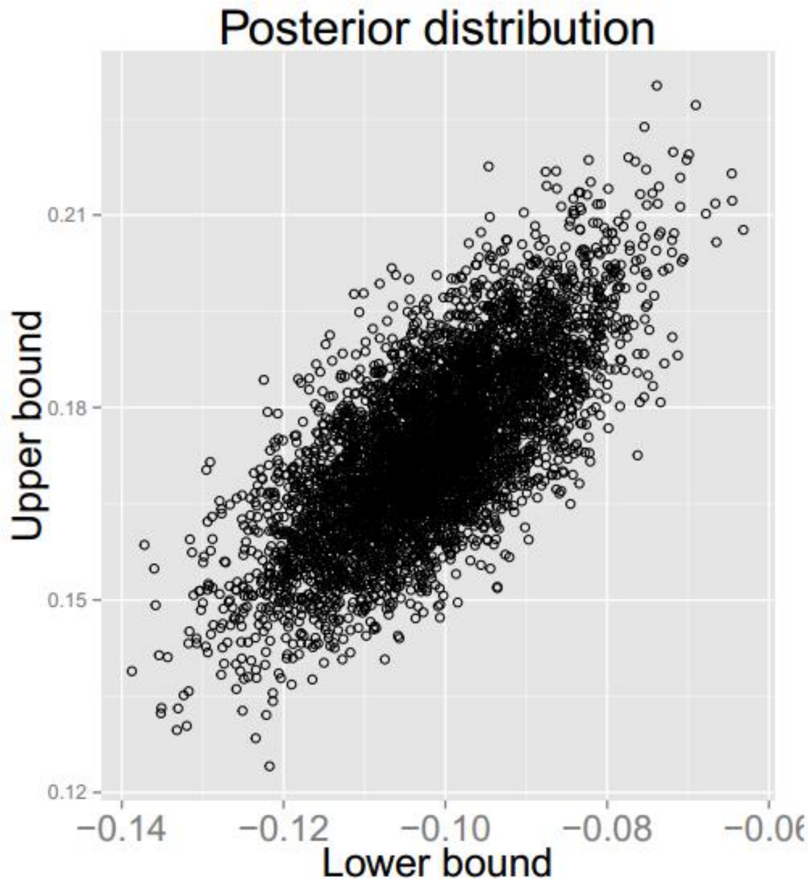
Influenza Data

- Effect of influenza vaccination (X) on hospitalization (Y = 1 means hospitalized)
- Covariate GRP: randomized, doctor of that patient received letter to encourage vaccination
  - (GRP, X, Y) ACE bound using standard IV: [-0.23, 0.64]
- WPP could *not* validate GRP. Instead it picked DM (diabetes history) as a witness, and AGE (dichotomized at 60 years) and SEX as admissible set

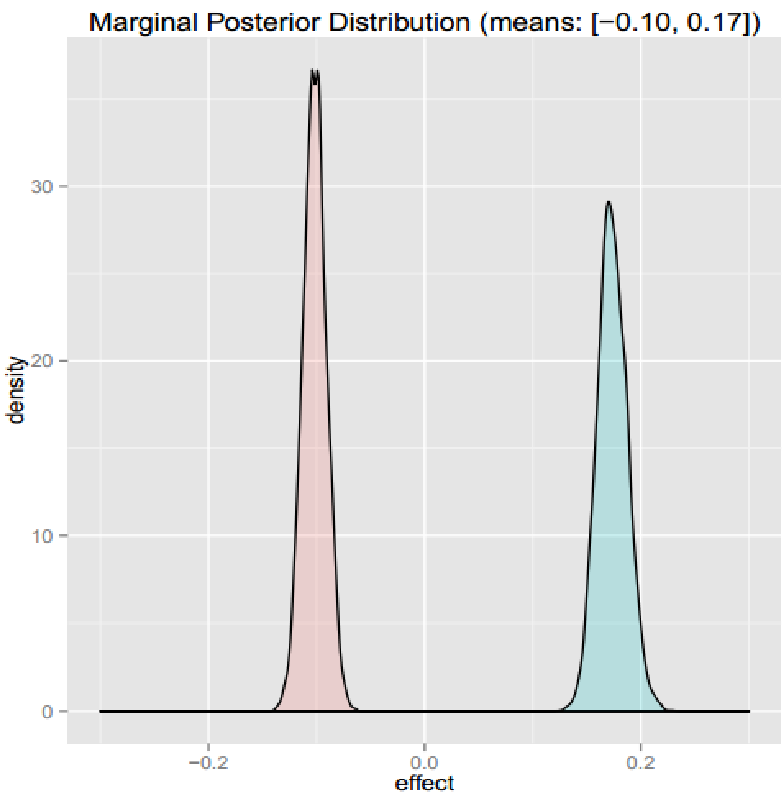
Influenza Data

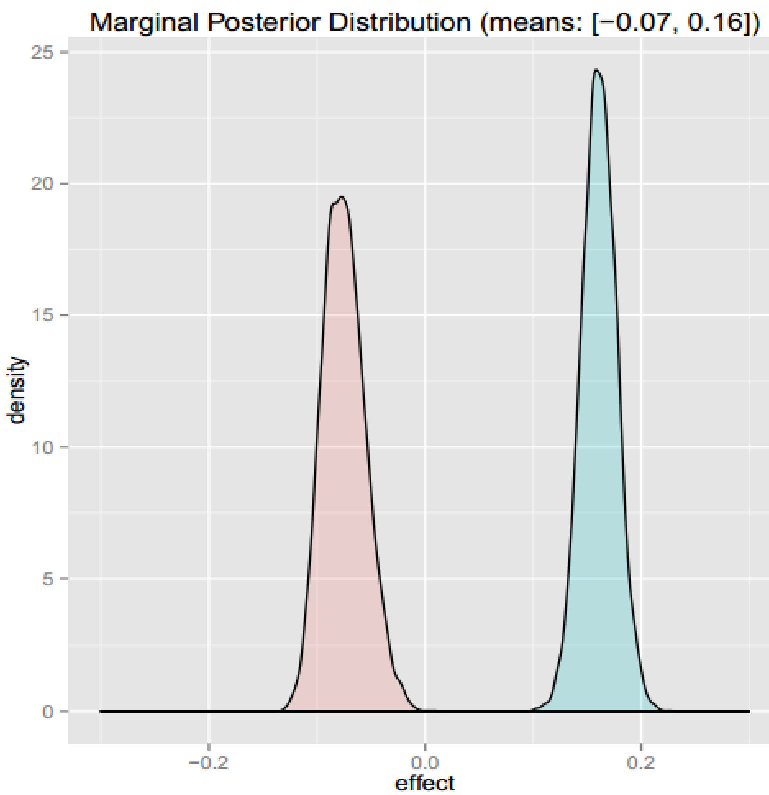
- Using same parameters as synthetic case study (0.9-1.1 for β), WPP estimated interval as [-0.10, 0.17]

Influenza Data: Full Posterior Plots



Influenza Data: Full Posterior Plots





On-going Work

- Finding a more primitive default set of assumptions where assumptions about the relaxations can be derived from
- Doing without a given causal ordering
- Large scale experiments
- Scaling up for a large number of covariates
- Continuous data
- More real data experiments
- R package to follow

http://arxiv.org/abs/1406.0531

Thank You

Extra

Mapping IV Model to Observations

- For now, assume model where W⊥⊥U
- Let

$$\zeta_{yx.w} \equiv \sum_u P(y, x \mid w, u) P(u)$$

and recall

$$\begin{aligned}
\zeta^\star_{yx.w} &\equiv P(Y = y, X = x \mid W = w, U) \\
\eta^\star_{xw} &\equiv P(Y = 1 \mid X = x, W = w, U) \\
\delta^\star_w &\equiv P(X = 1 \mid W = w, U)
\end{aligned}$$

- Idea: define a mapping from (η^*, δ^*) to ζ^*, then take convex combinations

Mapping

$$\eta^\star_{00} \quad \eta^\star_{01} \quad \eta^\star_{10} \quad \eta^\star_{11} \quad \delta^\star_0 \quad \delta^\star_1$$

$$\downarrow$$

$$\zeta^\star_{00.0} \quad \zeta^\star_{01.0} \quad \zeta^\star_{10.0} \quad \zeta^\star_{11.0} \quad \zeta^\star_{00.1} \quad \zeta^\star_{01.1} \quad \zeta^\star_{10.1} \quad \zeta^\star_{11.1} \quad \eta^\star_{00} \quad \eta^\star_{01} \quad \eta^\star_{10} \quad \eta^\star_{11}$$

$$\begin{aligned}
\zeta^\star_{00.0} &= (1-\eta^\star_{00})(1-\delta^\star_0) \\
\zeta^\star_{01.0} &= (1-\eta^\star_{10})\delta^\star_0 \\
\zeta^\star_{10.0} &= \eta^\star_{00}(1-\delta^\star_0) \\
\zeta^\star_{11.0} &= \eta^\star_{10}\delta^\star_0 \\
\zeta^\star_{00.1} &= (1-\eta^\star_{01})(1-\delta^\star_1) \\
\zeta^\star_{01.1} &= (1-\eta^\star_{11})\delta^\star_1 \\
\zeta^\star_{10.1} &= \eta^\star_{01}(1-\delta^\star_1) \\
\zeta^\star_{11.1} &= \eta^\star_{11}\delta^\star_1
\end{aligned}$$

Recipe

- Map the extreme points of (η^*, δ^*) to the extreme points of (ζ^*, η^*)
- Find convex hull of (ζ^*, η^*) → Show to be equivalent to the set of (ζ, η) allowable by the IV model. And

$$
\begin{aligned}
\eta_{xw} &\equiv \sum_U P(Y = 1 \mid X = x, W = w, U) P(U) \\
&= P(Y = 1 \mid do(X = x), W = w)
\end{aligned}
$$

- Re-express convex hull as linear inequalities (and equalities)
- ζ is observable/possible to estimate. Fixing ζ gives bounds on η

Estimation

- Simpler mapping on $(\delta^*, \eta^*) \rightarrow P(W, X, Y \mid U)$, marginalized, gives constraints on $\zeta \equiv P(W, X, Y)$
- Test whether constraints hold, if not provide no bounds
- Plug-in estimates for ζ to get (ζ, η) polytope. Find upper bounds and lower bounds on the ACE by solving linear program and maximizing/minimizing objective function

$$f(\eta) = (\eta_{11} - \eta_{01})P(W = 1) + (\eta_{10} - \eta_{00})P(W = 0)$$

All is Well?

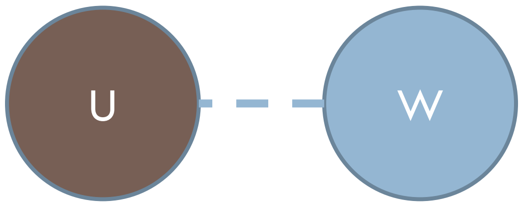
- It follows then min f(η) ≤ ACE ≤ max f(η)
- However, recall we mentioned this always has width 1… and actually there are no constraints on ζ!
- Further assumptions required. For instance:
  - Assume no direct effect of W on Y (change parameterization and mapping)
  - Assume monotonicity
    P(Y = 1 | do(X = 0)) ≤ P(Y = 1 | do(X =1))
  - Allow for bounded effect of W on Y, $|\eta^\star_{x1} - \eta^\star_{x0}| \leq \epsilon_w$
  - See Ramsahai (2012) for details

Adding More Assumptions

- In the linear programming formulation, an assumption such as $|\eta^\star_{x1} - \eta^\star_{x0}| \leq \epsilon_w$ is translated into a set of extreme points different from {(0, 0), (0, 1), (1, 0), (1, 1)}
  - Ramsahai (2012) provides analytical bounds for a given, numerical, value of ε_w
- Constraints such as $|\delta^\star_w - P(X = 1 \mid W = w)| \leq \epsilon_x$ are included by fixing P(X = 1 | W = w) first, the redefining the extreme points of parameter
  - Notice this implies non-linear constraints on ζ

Linking U and W

- What about



$$\underline{\beta} P(U) \leq P(U \mid W = w) \leq \bar{\beta} P(U)$$

- This redefines our expectations

$$\begin{aligned} \eta_{xw} &\equiv \sum_U P(Y = 1 \mid X = x, W = w, U) P(U \mid W) \\ &= P(Y = 1 \mid do(X = x), W = w) \end{aligned}$$

- Without further assumptions on P(U | W), linear program can be done as before, obtaining bounds for each value of W (Ramsahai, 2012)
  - Bounds always span zero

Linking U and W

- An additive relaxation
  $P(U) - \varepsilon \leq P(U \mid W) \leq P(U) + \varepsilon$ would however be problematic. Hence, the multiplicative relaxation
- Introduce intermediate parameterization

$$
\begin{aligned}
\zeta_{yx.w} &\equiv \textstyle\sum_U P(Y = y, X = x \mid W = w, U) P(U \mid W = w) \\
\kappa_{yx.w} &\equiv \textstyle\sum_U P(Y = y, X = x \mid W = w, U) P(U)
\end{aligned}
$$

$$
\begin{aligned}
\eta_{xw} &\equiv \textstyle\sum_U P(Y = 1 \mid X = x, W = w, U) P(U \mid W) \\
\omega_{xw} &\equiv \textstyle\sum_U P(Y = 1 \mid X = x, W = w, U) P(U)
\end{aligned}
$$

$$
\begin{aligned}
\delta_w &\equiv P(X = 1 \mid W = w) \\
\chi_{xw} &\equiv \textstyle\sum_U P(X = x \mid W = w, U) P(U)
\end{aligned}
$$

Linking U and W

- Follow recipe as before, but applying to the new – unobservable – variables
- Link them to observable ζ and target η using

$$\underline{\beta} P(U) \leq P(U \mid W = w) \leq \bar{\beta} P(U)$$

- For instance

$$\begin{array}{rcl}
\kappa_{yx.w} & \geq & P(Y = y, X = x \mid W = w)/\bar{\beta} \\
\kappa_{yx.w} & \leq & P(Y = y, X = x \mid W = w)/\underline{\beta} \\
\chi_{xw} & \geq & P(X = x \mid W = w)/\bar{\beta} \\
\chi_{xw} & \leq & P(X = x \mid W = w)/\underline{\beta}
\end{array}$$

Rejection Sampling

- If we have the polytope, then this is a very cheap check of whether linear inequalities are satisfied
- However, we need to obtain the polytope as a function of ζ. Better do that in an analytic way, or otherwise a numerical polytope calculation procedure for each sample will not be feasible
- Difficulty: extreme points of (δ^*, ζ^*) are not the extremes of the unit hypercube anymore

Main Idea

- Let's go back to the original mapping:

$$
\begin{aligned}
\zeta^\star_{00.0} &= (1-\eta^\star_{00})(1-\delta^\star_0)\\
\zeta^\star_{01.0} &= (1-\eta^\star_{10})\delta^\star_0\\
\zeta^\star_{10.0} &= \eta^\star_{00}(1-\delta^\star_0)\\
\zeta^\star_{11.0} &= \eta^\star_{10}\delta^\star_0\\
\zeta^\star_{00.1} &= (1-\eta^\star_{01})(1-\delta^\star_1)\\
\zeta^\star_{01.1} &= (1-\eta^\star_{11})\delta^\star_1\\
\zeta^\star_{10.1} &= \eta^\star_{01}(1-\delta^\star_1)\\
\zeta^\star_{11.1} &= \eta^\star_{11}\delta^\star_1
\end{aligned}
$$

- Without further assumptions, what can we say?

Main Idea

- Implied bounds follow from the probability simplex constraints
  - Notice the need for X to be discrete
- As pointed out by Balke and Pearl, ζ is feasible if no upper bound on η is smaller than any lower bound
- What happens when we introduce the assumption "no direct effect of W on Y"?

$$
\begin{aligned}
\eta_{11} &\geq \zeta_{11.1} \\
\eta_{10} &\geq \zeta_{11.0} \\
\eta_{11} &\leq 1 - \zeta_{01.1} \\
\eta_{10} &\leq 1 - \zeta_{01.0} \\
\eta_{01} &\geq \zeta_{10.1} \\
\eta_{01} &\leq 1 - \zeta_{00.1} \\
\eta_{00} &\geq \zeta_{10.0} \\
\eta_{00} &\leq 1 - \zeta_{00.0}
\end{aligned}
$$

Direct Polytope Manipulation

$$
\begin{aligned}
\eta_1^\star &\leq 1 \\
\eta_1^\star(1-\delta_1^\star) &\leq 1-\delta_1^\star \\
\eta_1 - \zeta_{11.1} &\leq 1-(\zeta_{11.1}+\zeta_{01.1}) \quad \text{(marginalization)} \\
\zeta_{11.0} - \zeta_{11.1} &\leq 1-(\zeta_{11.1}+\zeta_{01.1}) \quad (\text{since } \eta_1 = \eta_{10} \geq \zeta_{11.0}) \\
\zeta_{11.0} + \zeta_{01.1} &\leq 1
\end{aligned}
$$

- This is one of the "instrumental inequalities" of the standard IV model, derived directly
  - Bounding η^* by one of its extreme points
  - Modify factor in a way to map it to ζ and η, perform further manipulations
- Useful as a way of deriving symbolic bounds as a function of the extreme points of the original parameter space

Our More General Case

- Start from

$$\begin{aligned}
\max(P(Y = 1|X = x, W = w) - \epsilon_y, 0) &\equiv L_{xw}^{YU} \\
\min(P(Y = 1|X = x, W = w) + \epsilon_y, 1) &\equiv U_{xw}^{YU} \\
\max(P(X = 1|W = w) - \epsilon_x, 0) &\equiv L_{w}^{XU} \\
\min(P(X = 1|W = w) + \epsilon_x, 1) &\equiv U_{w}^{XU}
\end{aligned}$$

$$\begin{aligned}
L_{xw}^{YU} \leq \; &\eta_{xw}^{\star} \leq U_{xw}^{YU} \\
L_{w}^{XU} \leq \; &\delta_{w}^{\star} \leq U_{w}^{XU}
\end{aligned}$$

- Like in the previous slide, we create new bounds by multiplying and marginalizing pieces of the latent variable model

Examples

- Case 1 (Fails to obtain new bound)

$$
\begin{aligned}
\eta^{\star}_{1w} &\leq U^{YU}_{1w} \\
\eta^{\star}_{1w}\delta^{\star}_{w} &\leq U^{YU}_{1w}\delta^{\star}_{w} \quad \text{(Marginalize over } P(U)\text{)} \\
\kappa_{11.w} &\leq U^{YU}_{1w}\chi_{w} \quad \text{(Always true)}
\end{aligned}
$$

- Case 2 (Generalizes $\omega_{0w} \leq 1 - \kappa_{00.w}$)

$$
\begin{aligned}
\eta^{\star}_{0w} &\leq U^{YU}_{0w} \\
\eta^{\star}_{0w}(1-(1-\delta^{\star}_{w})) &\leq U^{YU}_{0w}\delta^{\star}_{w} \\
\omega_{0w} - \kappa_{10.w} &\leq U^{YU}_{0w}\chi_{w} \\
\omega_{0w} &\leq \kappa_{10.w} + U^{YU}_{0w}(\kappa_{01.w} + \kappa_{11.w})
\end{aligned}
$$

Solving the Linear Program

- The very same (symbolic) bounds used for verifying the feasibility of ζ can be used in a straightforward way to bound the ACE