Slide 12 (Introduction)

- n experimental units (for measurement)
- Unit i : covariates $X_i \in \mathcal{X}$ measured before treatment.
  outcome $Y_i \in \mathcal{Y} \subseteq \mathbb{R}$ measured after treatment.
- Treatment $Z \in \mathcal{Z}$. Often $Z = (Z_1, \cdots, Z_n)$, but not required.
- Exposure : $A_i = A_i(Z) \in \mathcal{A}$. Usually $A = Z$ (identity mapping)
  $\mathcal{A} = \{0, 1\}$ (binary exposure)
- $X = (X_1, \cdots, X_n)$, same for A, Y.

Slide 13 (Design of Experiments)

- Bernoulli trial : $\pi(z \mid x) = \prod_{i=1}^{n} (\pi(x_i))^{z_i} (1 - \pi(x_i))^{1-z_i}$.

  [Simple/sampling w. replacement : $\pi(x_i) = \pi$]

- Sampling w/o replacement : $\pi(z \mid x) = \begin{cases} \binom{n}{n_1}^{-1}, & \text{if } \sum_{i=1}^{n} z_i = n_1 \\ 0, & \text{otherwise.} \end{cases}$

- Randomized complete block design :

  k treatment levels : $0, 1, \cdots, k-1$. m blocks.

  $Z = (Z_{ij})_{i \in [m], j \in [k]}$.

  $\pi(z) = \begin{cases} (k!)^{-m}, & \text{if } (z_{ij})_{j \in [k]} \text{ is permutation of } (0, 1, \cdots, k-1) \text{ for all } i. \\ 0, & \text{otherwise.} \end{cases}$

Slide 14 (Towards a formal theory)

- Potential outcomes : $Y(z) = (Y_1(z), \cdots, Y_n(z))$, $z \in \mathcal{Z}$.
- Causal effect is understood as the contrast between $Y(z)$ for different z.

<u>Assumption</u> (Consistency of p.o.) $Y = Y(Z)$

realized treatment.

- Fundamental problem : only 1 p.o. is observed for every unit.

<u>Assumption</u>

$$A: \mathcal{Z} \to \mathcal{A}^n$$
$$z \mapsto (A_1(z), \cdots, A_n(z))$$

is a valid exposure mapping

in the sense that $Y_i(z) = Y_i(z')$ for all i, z, z' s.t. $A_i(z) = A_i(z')$.

[In this case, we often write p.o. as $Y_i(a)$ for $a \in \mathcal{A}$.]

<u>Definition</u> The Neyman - Rubin causal model assumes

1. Consistency of p.o.
2. Identity exposure mapping is valid.

Usually $\mathcal{A} = \{0, 1\}$, and $Y_i(1) - Y_i(0)$ is called individual treatment effect.

Slide 15 (Example : Interference)

- Network : $\mathcal{G} = (\mathcal{V} = [n] , \mathcal{E} \subseteq [n] \times [n])$

vertex set   edge set

- $A_i(z) = (z_i , \sum_{(i',i) \in \mathcal{E}} z_{i'}) \in \{0,1\} \times \{0, \cdots, n-1\}$.

Slide 16 (Example: Lady tasting tea)

- Data and potential outcome schedule.

[Coding : 0 means milk first, 1 means tea first.]

i	Z_i [$=A_i$]	Y_i	$Y_i(0)$	$Y_i(1)$	$Y_i(1) - Y_i(0)$
1	1	1	?	1	?
2	1	1	?	1	?
3	0	0	0	?	?
4	0	1	1	?	?
5	1	1	?	1	?
6	0	0	0	?	?
7	1	0	?	0	?
8	0	0	0	?	?

- Data can be summarized by the following 2×2 contingency table.

	$Y=0$	$Y=1$
$A=0$	3	1
$A=1$	1	3 ← test stat.

- Fisher's exact test: $\binom{8}{4} = 70$ possibilities.

Test statistic: how many correct tea-first guesses.

$$\mathbb{P}(T=4) = \binom{4}{4}\binom{4}{4}/70 = 1/70$$

$$\mathbb{P}(T=3) = \binom{4}{3}\binom{4}{1}/70 = 16/70.$$

p-value is $17/70 \approx 24.3\%$.

Slide 17 (Randomization inference: General tests)

<u>Assumption</u> (Exogeneity of randomization)

① $Z \perp\!\!\!\perp (Y(z))_{z \in \mathcal{Z}} \mid X$
$[= W]$

② $\mathbb{P}(Z=z \mid X=x) = \pi(z \mid x)$ is known

- Fisher's sharp null: H_0: $Y(z)$ doesn't depend on z.
- In N-R model: H_0: $Y_i(0) = Y_i(1)$, $\forall\, i \in [n]$.

This allows us to impute the p.o. schedule

i	$Z_i^{[=A_i]}$	Y_i	$Y_i(0)$	$Y_i(1)$	$Y_i(1) - Y_i(0)$
1	1	1	1	1	0
2	1	1	1	1	0
3	0	0	0	0	0
4	0	1	1	1	0
5	1	1	1	1	0
6	0	0	0	0	0
7	1	0	0	0	0
8	0	0	0	0	0

- Test statistic: $T: \mathcal{Z} \times \mathcal{X} \times \mathcal{W} \to \mathbb{R}$

- Randomization p-value:

$$P(Z, X, W) = \mathbb{P}\left(T(Z', X, W) \le T(Z, X, W) \mid Z, X, W\right)$$

where $Z' \overset{d}{=} Z \mid X, W$, $\quad Z' \perp\!\!\!\perp Z \mid X, W$.

<u>Thm</u> $\quad \mathbb{P}\left(P(Z, X, W) \le \alpha \mid X, W\right) \le \alpha, \quad \forall\ 0 < \alpha < 1.$

Pf Probability integral transform: Let F be the C.D.F. of r.v. T. Then $F(T)$ is (almost) uniform and

$$P(F(T) \le \alpha) \le \alpha, \quad \forall \; 0<\alpha<1. \qquad \square$$

Rem Computing $P(Z, X, W)$ requires

1. Randomization: $Z \perp\!\!\!\perp W \mid X$, $\pi(Z \mid X)$ is known.

2. W is known under H_0.

   What if H_0 is not exhaustive? Example: No spillover effect for vaccinated students.

Tool: Further condition on $g(Z') = g(Z)$ for some function g.

Ex N-R model. $\mathcal{A} = \mathcal{Y} = \{0, 1\}$. No covariates.

Sampling w.o. replacement.

Data can be summarized by

		Y: 0	1	
A	0	N_{00}	N_{01}	$N_{0\cdot}$
	1	N_{10}	N_{11}	$N_{1\cdot} = n_1$
		$N_{\cdot 0}$	$N_{\cdot 1}$	$N_{\cdot\cdot} = n$

Under Fisher's sharp null, the probability of observing $(N_{00}, N_{01}, N_{10}, N_{11})$ is given by the hypergeometric prob.

$$\frac{\binom{N_{0\cdot}}{N_{00}}\binom{N_{1\cdot}}{N_{10}}}{\binom{N}{N_{\cdot 0}}}. \Longrightarrow \text{Fisher's exact test.}$$

Slide 18 (Randomization inference : Estimation)

- Randomization dist. of $\hat{\beta}_{DIM} = \bar{Y}_1 - \bar{Y}_0$,

$$\left[\bar{Y}_1 = \frac{\sum A_i Y_i}{\sum A_i}, \quad \bar{Y}_0 = \frac{\sum (1-A_i) Y_i}{\sum (1-A_i)} \right]$$

consistency of p.o. $\rightarrow$

$$= \frac{1}{n_1}\sum_{i=1}^{n} A_i Y_i(1) - \frac{1}{n_0}\sum_{i=1}^{n}(1-A_i) Y_i(0)$$

sampling w.o. replacement

So $$\mathbb{E}[\hat{\beta} \mid W] = \frac{1}{n_1}\sum_{i=1}^{n} \mathbb{E}(A_i)\cdot Y_i(1) - \frac{1}{n_0}\sum_{i=1}^{n} \mathbb{E}(1-A_i) Y_i(0)$$

exogeneity

$$= \frac{1}{n_1}\sum_{i=1}^{n}\frac{n_1}{n}\cdot Y_i(1) - \frac{1}{n_0}\sum_{i=1}^{n}\frac{n_0}{n}\cdot Y_i(0)$$

$$= \frac{1}{n}\sum_{i=1}^{n} Y_i(1) - Y_i(0)$$

$$= \beta_n.$$

Exercise: Show that $\mathrm{Var}(\hat{\beta}_{DIM} \mid W) = \frac{1}{n_0}S^2(0) + \frac{1}{n_1}S^2(1) - \frac{1}{n}S^2(0,1)$.

Hint: $\hat{\beta}_{DIM}$ is linear in W, so $\mathrm{Var}(\hat{\beta}_{DIM} \mid W)$ is quadratic.

Use symmetry to argue

$$\mathrm{Var}(\hat{\beta}_{DIM} \mid W) = c_0 S^2(0) + c_1 S^2(1) + c_{01} S^2(0,1).$$

Now consider the model

$$\begin{pmatrix} Y_i(0) \\ Y_i(1) \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \rho\sigma_0\sigma_1 \\ \rho\sigma_0\sigma_1 & \sigma_1^2 \end{pmatrix} \right).$$

- Variance estimators:

$$\hat{S}_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^n A_i (Y_i - \bar{Y}_1)^2$$

$$\hat{S}_0^2 = \frac{1}{n_0 - 1} \sum_{i=1}^n (1 - A_i)(Y_i - \bar{Y}_0)^2.$$

<u>Exercise</u> $\mathbb{E}(\hat{S}_a^2 \mid W) = S^2(a), \quad a = 0, 1.$

However, $S^2(0,1)$ cannot be directly estimated.

- Finite-sample CLT: Under additional assumptions on p.o., one can show

$$\frac{\hat{\beta} - \beta_n}{\sqrt{\mathrm{Var}(\hat{\beta} \mid W)}} \xrightarrow{d} N(0, 1) \quad \text{as } n \to \infty.$$

Slide 19 (Randomization inference: F-test)

- Randomization dist. of S_A and S_E.

<u>Assume</u>: $\mathcal{A} = \{0, 1\}. \Rightarrow S_A = \frac{n_1 n_0}{n} (\bar{Y}_1 - \bar{Y}_0)^2.$

sampling w.o. replacement

Fisher's sharp null: $H_0: Y_i(0) = Y_i(1), \ \forall i. \Rightarrow S^2(0) = S^2(1)$

$S^2(0,1) = 0.$

So $\mathbb{E}(S_A \mid W) = \frac{n_1 n_0}{n} \mathrm{Var}(\hat{\beta}_{DIM} \mid W) = \frac{n_1 n_0}{n} S^2(0) \left(\frac{1}{n_0} + \frac{1}{n_1}\right) = S^2(0).$

On the other hand,

$$S_A + S_E = S_T = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

$$\mathbb{E}(S_T \mid W) = \sum_{i=1}^{n} (Y_i(0) - \bar{Y}(0))^2 = (n-1) S^2(0).$$

$$\Rightarrow \quad \frac{\mathbb{E}(S_A \mid W)}{\mathbb{E}(S_E \mid W)} = \frac{1}{n-2}.$$

It can be further shown that the F-test is approximately valid in the randomization model, without assuming the normal linear model.

Slide 20 (Repeated sampling)

- Positivity / overlap assumption: $\mathbb{P}(A=a \mid X=x) > 0, \quad \forall a \in \mathcal{A}, x \in \mathcal{X}.$
- Causal identification:

<u>Thm</u> Assume exogeneity ($\Rightarrow$ stratified Bernoulli trials) and positivity.

Then $\mathbb{P}(Y(a) \le y \mid X=x) = \mathbb{P}(Y \le y \mid A=a, X=x), \quad \forall a, x, y.$

<u>Pf</u>

$$\mathbb{P}(Y(a) \le y \mid X=x) = \mathbb{P}(Y(a) \le y \mid X=x, A=a)$$

$$\underset{\substack{\uparrow \\ \text{consistency}}}{=} \mathbb{P}(Y \le y \mid X=x, A=a). \qquad \square$$

- Average treatment effect: $\beta_{ATE} = \mathbb{E}(Y(1) - Y(0)).$
- Estimators: Outcome regression $\hat{\beta}_{OR} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i).$

Inverse-probability weighting $\hat{\beta}_{IPW} = \frac{1}{n}\sum_{i=1}^{n} \frac{A_i Y_i}{\pi(X_i)} - \frac{(1-A_i)Y_i}{1-\pi(X_i)}.$

<u>Exercise</u> $\sqrt{n}(\hat{\beta}_{IPW} - \beta_{ATE}) \xrightarrow{d} N(0, \sigma^2_{IPW}).$

Slide 21 (M-estimation)

- Taylor expansion

$$
\begin{aligned}
0 &= \frac{1}{n}\sum_{i=1}^{n} \psi(\hat{\theta}; D_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \psi(\theta_0; D_i) + (\hat{\theta}-\theta_0)^T \left\{\frac{\partial}{\partial\theta}\psi(\theta_0; D_i)\right\} \\
&\quad + \underbrace{O_p(\|\hat{\theta}-\theta_0\|^2)}.
\end{aligned}
$$

Negligible if: 1. $\frac{\partial^2}{\partial\theta^2}\psi$ bounded

2. $\hat{\theta}-\theta_0 \xrightarrow{p} 0$

$$
\Rightarrow \sqrt{n}(\hat{\theta}-\theta_0) \approx \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \underbrace{\left\{\mathbb{E}\left[\frac{\partial}{\partial\theta}\psi(\theta_0; D)\right]\right\}^{-1} \psi(\theta_0; D_i)}_{\text{Influence function}}
$$

$$
\xrightarrow{d} N\left(0, \left\{\mathbb{E}\left[\frac{\partial}{\partial\theta}\psi(\theta_0)\right]\right\}^{-1} \mathbb{E}[\psi(\theta_0)\psi(\theta_0)^T] \left\{\mathbb{E}\left[\frac{\partial}{\partial\theta}\psi(\theta_0)\right]\right\}^{-1}\right)
$$

"sandwich" variance.

Apply this to $\ell(\theta; D) = (Y - \theta^T V)^2$:

Let $\varepsilon = Y - \theta_0^T V$

$$
\sqrt{n}(\hat{\theta}-\theta) \xrightarrow{d} N\left(0, \{\mathbb{E}(VV^T)\}^{-1} \mathbb{E}(VV^T\varepsilon^2)\{\mathbb{E}(VV^T)\}^{-1}\right).
$$

This is robust to heteroscedasticity (i.e. $\mathrm{Var}(\varepsilon|V)$ depends on V).

Slide 22 (Regression adjustment)

Let $$(\alpha_k, \beta_k, \gamma_k, \delta_k) = \underset{\substack{\gamma \in \Gamma_k \\ \delta \in \Delta_k}}{\arg\min}\ \mathbb{E}\left[\left\{ Y - \alpha - \beta A - \gamma^T X - A\cdot(\delta^T X)\right\}^2\right].$$

Suppose $X \in \mathbb{R}^p$. Then $\Gamma_1, \Delta_1 = \{0\}$. $\Gamma_2 = \mathbb{R}^p$, $\Delta_2 = \{0\}$. $\Gamma_3 = \Delta_3 = \mathbb{R}^p$.

Lem $\alpha_1 = \alpha_2 = \alpha_3$, $\beta_1 = \beta_2 = \beta_3 = \beta_{ATE}$.

Proof Assume ∂ and $\mathbb{E}$ can always be interchanged.

$$\frac{\partial}{\partial \alpha} = 0 \Rightarrow \mathbb{E}\left[Y - \alpha_k - \beta_k A - \gamma_k^T X - A_i(\delta_k^T X)\right] = 0.$$

$$\frac{\partial}{\partial \beta} = 0 \Rightarrow \mathbb{E}\left[A\left(- \ - \ - \ - \ - \ - \right)\right] = 0.$$

By using $\mathbb{E}(X) = 0$ & $A \perp\!\!\!\perp X$

$$\left.\begin{aligned} &\mathbb{E}\left[Y - \alpha_k - \beta_k A\right] = 0 \quad (1) \\ &\mathbb{E}\left[A(Y - \alpha_k - \beta_k A)\right] = 0. \quad (2) \end{aligned}\right\} \text{ for } k = 1,2,3.$$

$$(2) - (1) \times \pi \Rightarrow \beta_k = \frac{\mathbb{E}(AY) - \pi\mathbb{E}(Y)}{\pi - \pi^2}$$

$$= \frac{\pi\cdot\mathbb{E}(Y|A=1) - \pi\{\pi\mathbb{E}(Y|A=1) + (1-\pi)\mathbb{E}(Y|A=0)\}}{\pi - \pi^2}$$

$$= \mathbb{E}(Y|A=1) - \mathbb{E}(Y|A=0). \quad \square$$

- Applying M-estimation to

$$V_1 = \begin{pmatrix} 1 \\ A \end{pmatrix}, \quad V_2 = \begin{pmatrix} 1 \\ A \\ X \end{pmatrix}, \quad V_3 = \begin{pmatrix} 1 \\ A \\ X \\ AX \end{pmatrix}.$$

Let $\varepsilon_k = Y - \alpha_k - \beta_k A - \gamma_k^T X - A(\delta_k^T X)$.

<u>Thm</u> $\sqrt{n}(\hat{\beta}_k - \beta) \xrightarrow{d} N(0, V_k)$, $\quad \underbrace{V_k = \dfrac{\mathbb{E}[(A-\pi)^2 \varepsilon_k^2]}{\pi^2(1-\pi)^2},}_{\text{Verify in E.S.}}$ $k=1,2,3$

And $V_3 \le \min\{V_1, V_2\}$.

<u>Proof</u> $\mathbb{E}(\varepsilon V) = 0 \Rightarrow \mathbb{E}(\varepsilon_3) = \mathbb{E}(A\varepsilon_3) = 0$

$$\mathbb{E}(\varepsilon_3 X) = \mathbb{E}(\varepsilon_3 AX) = 0.$$

$$\varepsilon_1 = \varepsilon_3 + \gamma_3^T X + A(\delta_3^T X)$$

$$\varepsilon_2 = \varepsilon_3 + (\gamma_3 - \gamma_2)^T X + A(\delta_3^T X).$$

So for $k = 1, 2$

$$\mathbb{E}\{(A-\pi)^2 \varepsilon_k^2\} - \mathbb{E}\{(A-\pi)^2 \varepsilon_3^2\}$$
$$= \mathbb{E}\{(A-\pi)^2 [(\gamma_3 - \gamma_k)^T X + A(\delta_3^T X)]^2\} \ge 0.$$

- Remark: We do not assume linear model is correct.

- When $\mathbb{E}(X) \neq 0$, need to center X first.

$$X_i \longrightarrow X_i - \bar{X}, \qquad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i .$$

$\hat{\beta}_1$, $\hat{\beta}_2$ unchanged.

$\hat{\beta}_3$ becomes $\tilde{\beta}_3 = \hat{\beta}_3 + \hat{\delta}_3 \bar{X}$.

Can show $\tilde{\beta}_3$ is still more efficient.

Slide 28

- Example:

$$\begin{tikzcd} V_1 \arrow[r, "\beta_{12}"] \arrow[rr, bend left, "\beta_{13}"] & V_2 \arrow[r, "\beta_{23}"] & V_3 \end{tikzcd}$$

Bidirected self-loop often omitted.

$$\begin{aligned} V_1 &= E_1 \\ V_2 &= \beta_{12} V_1 + E_2 \\ V_3 &= \beta_{13} V_1 + \beta_{23} V_2 + E_3 \end{aligned} \quad \xRightarrow[V_1 = v_1]{\text{Intervention}} \quad \begin{aligned} V_2(v_1) &= \beta_{12} v_1 + E_2 \\ V_3(v_1) &= \beta_{13} v_1 + \beta_{23} V_2(v_1) + E_3 \\ &= (\beta_{13} + \beta_{12}\beta_{23}) v_1 + \cdots \end{aligned}$$

- This suggests us to define the total causal effect of V_j on V_k as $\sigma(\underbrace{W(j \rightsquigarrow k)}_{\text{the set of all directed walks } j \to \cdots \to k.})$.

Slide 29 (Path analysis)

- Trek rule:

$$V = \alpha + B^T V + E \Rightarrow V = (I - B)^{-T}(\alpha + E)$$

$$\Rightarrow \operatorname{Cov}(V) = (I-B)^{-T} \Lambda (I-B)^{-1}.$$

$$(I-B)^{-1} = I + B + B^2 + \cdots = \sigma(I + W(V \rightsquigarrow V))$$

$$\Lambda = \sigma(W(V \leftrightarrow V))$$

$$\Rightarrow \operatorname{Cov}(V) = \sigma(W(V \overset{t}{\leftrightsquigarrow} V))$$

where
$$W(V \overset{t}{\leftrightsquigarrow} V) = \{I + W(V \leftsquigarrow V)\} \cdot W(V \leftrightarrow V) \cdot \{I + W(V \rightsquigarrow V)\}$$

$$= W(V \leftrightarrow V) + W(V \leftsquigarrow V \leftrightarrow V) + W(V \leftrightarrow V \rightsquigarrow V) + W(V \leftsquigarrow V \leftrightarrow V \rightsquigarrow V)$$

$\overset{t}{\leftrightsquigarrow}$ is called a trek: a walk with 0 collider and 1 $\leftrightarrow$

- m-connectedness: a walk is an arc or (uncond.) m-connected if it has no collider:

$$W(V \sim\!\!\sim V) = W(V \rightsquigarrow V) + W(V \leftsquigarrow V) + W(V \leftrightsquigarrow V)$$

<u>Exercise</u> How many $\leftrightarrow$ can an arc have?

<u>Thm</u> Suppose ADMG $\mathcal{G}$ has all bidirected self-loops: $(j, j) \in \mathcal{B}$ for all $j \in V$. Then

$$W(j \overset{t}{\leftrightsquigarrow} k) \neq \emptyset \iff P(j \sim\!\!\sim k) \neq \emptyset.$$

<u>Pf</u> $\Leftarrow$ Insert a bidirected self-loop if needed.

$\Rightarrow$ Suppose $\pi \in W(j \overset{t}{\leftrightsquigarrow} k)$

So $\pi = j \overset{\circ}{\leftsquigarrow} l \leftrightarrow l' \overset{\circ}{\rightsquigarrow} k$ (possibly $l = j$, $l' = k$).

If π is not already a path, suppose r is the repeated vertex closest to j.

r must appear once on $j \overset{\circ}{\leftsquigarrow} l$ and once on $l' \overset{\circ}{\rightsquigarrow} k$.

So $\pi = j \overset{\circ}{\leftsquigarrow} \underbrace{r \overset{\circ}{\leftsquigarrow} l \leftrightarrow l' \overset{\circ}{\rightsquigarrow} r}_{\in W(r \overset{t}{\leftrightsquigarrow} r)} \overset{\circ}{\rightsquigarrow} k$

We say r is the root of π, and denote all treks from j to k with root r as $W(j \overset{t}{\leftrightsquigarrow} k, \text{root } r)$

By assumption, $W(r \overset{t}{\leftrightsquigarrow} r) \neq \phi$.

So $W(j \overset{t}{\leftrightsquigarrow} k, \text{root } r) \neq \phi \iff P(j \overset{\circ}{\leftsquigarrow} r \overset{\circ}{\rightsquigarrow} k) \neq \phi$.

The desired conclusion follows from

$$W(j \overset{t}{\leftrightsquigarrow} k) = P(j \overset{t}{\leftrightsquigarrow} k) + \sum_{r \in V} W(j \overset{t}{\leftrightsquigarrow} k, \text{root } r)$$

$$P(j \leftrightsquigarrow k) = P(j \overset{t}{\leftrightsquigarrow} k) + \sum_{r \in V} P(j \overset{\circ}{\leftsquigarrow} r \overset{\circ}{\rightsquigarrow} k) \qquad \square$$

- Wright's path analysis: As a corollary,

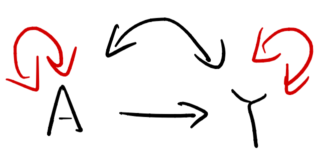
$$\mathrm{Cov}(V_j, V_k) = \sigma(P(j \overset{t}{\leftrightsquigarrow} k)) + \sum_{r \in V} \sigma\left(P(j \overset{\circ}{\leftsquigarrow} r \overset{\circ}{\rightsquigarrow} k)\right) \cdot \mathrm{Var}(V_r).$$

If V is standardized so that $\mathrm{Var}(V_j) = 1$, $\forall j \in V$, then

$$\mathrm{Cov}(V_j, V_k) = \sigma(P(j \leftrightsquigarrow k))$$

Compare this to the total causal effect $\sigma(P(j \rightsquigarrow k))$.

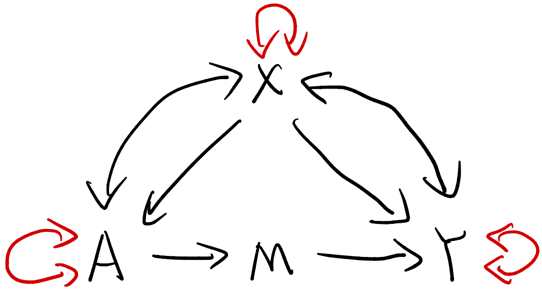
Slide 30 (Two examples)



$\mathrm{Var}(A) = \lambda_{AA}$ Q: Why?

$\mathrm{Var}(Y) = \lambda_{YY} + \lambda_{AA} \beta_{AY}^2 + 2 \lambda_{AY} \beta_{AY}$

$\mathrm{Cov}(A, Y) = \lambda_{AY} + \lambda_{AA} \beta_{AY}$.



$$\mathrm{Cov}(A, Y) = \beta_{AM}\beta_{AY} \cdot \mathrm{Var}(A) + \beta_{XA}\beta_{XY} \cdot \mathrm{Var}(X) + \beta_{XA}\lambda_{XY} + \beta_{XY}\lambda_{AX}$$

Slide 31 (Correlation vs. causation)

Examples		Causal effect	Correlation	Partial cor.
1. Confounder:	$A \leftarrow X \rightarrow Y$	$=0$	$\neq 0$	$\neq 0$
2. Mediator:	$A \rightarrow M \rightarrow Y$	$\neq 0$	$\neq 0$	$=0$
3. Collider:	$A \rightarrow C \leftarrow Y$	$=0$	$=0$	$\neq 0$

Slide 34 (Marginalization and latent projection)

$$(I-B)^{-1} = \begin{pmatrix} ((I-B)^{-1})_{\tilde{V},\tilde{V}} & * \\ ((I-B)^{-1})_{U,\tilde{V}} & * \end{pmatrix}$$

$$((I-B)^{-1})_{\tilde{V},\tilde{V}} = (I-\tilde{B})^{-1}, \text{ where}$$

$$\tilde{B} = B_{\tilde{V},\tilde{V}} + B_{\tilde{V},U}(I-B_{U,U})^{-1}B_{U,\tilde{V}}$$

$$= \sigma(W(\tilde{V} \overset{\text{via } U}{\rightsquigarrow} \tilde{V})).$$

$$((I-B)^{-1})_{U,\tilde{V}} = (I-B_{U,U})^{-1} B_{U,\tilde{V}} (I-\tilde{B})^{-1}$$

- $$\tilde{\Sigma} = ((I-B)^{-T}\Lambda(I-B)^{-1})_{\tilde{V},\tilde{V}}$$

$$= (I-\tilde{B})^{-T}\Lambda_{\tilde{V},\tilde{V}}(I-\tilde{B})^{-1}$$

$$+ (I-\tilde{B})^{-T}\Lambda_{\tilde{V},U}((I-B)^{-1})_{U,\tilde{V}}$$

$$+ ((I-B)^{-T})_{\tilde{V},U} \cdot \Lambda_{U,\tilde{V}} (I-\tilde{B})^{-1}$$

$$+ ((I-B)^{-T})_{\tilde{V},U} \cdot \Lambda_{U,U}((I-B)^{-1})_{U,\tilde{V}}$$

$$= (I-\tilde{B})^{-T}\tilde{\Lambda}(I-\tilde{B})^{-1}.$$

where $\tilde{\Lambda} = \Lambda_{\tilde{V},\tilde{V}} + \Lambda_{\tilde{V},U}(I-B_{U,U})^{-1}B_{U,\tilde{V}}$

$$+ B_{U,\tilde{V}}^{T}(I-B_{U,U})^{-T}\Lambda_{U,\tilde{V}}$$

$$+ \{(I-B_{U,U})^{-1}B_{U,\tilde{V}}\}^{T}\Lambda_{U,U}\{(I-B_{U,U})^{-1}B_{U,\tilde{V}}\}$$

$$= \sigma(W(\tilde{V} \leftrightarrow \tilde{V})) + \sigma(W(\tilde{V} \leftrightarrow U \overset{\circ}{\rightsquigarrow} U \rightarrow \tilde{V}))$$

$$+ \sigma(W(\tilde{V} \leftarrow U \overset{\circ}{\leftsquigarrow} U \leftrightarrow \tilde{V}))$$

$$+ \sigma(W(\tilde{V} \leftarrow U \overset{\circ}{\leftsquigarrow} U \leftrightarrow U \overset{\circ}{\rightsquigarrow} U \rightarrow \tilde{V}))$$

$$= \sigma(W(\tilde{V} \overset{t,\text{via } U}{\leftrightsquigarrow} \tilde{V})).$$

Exercise: Show that $W(\tilde{V} \overset{t, \text{via } \mathcal{U}}{\leftrightsquigarrow} \tilde{V}) = \emptyset$

iff $W(\tilde{V} \overset{\text{via } \mathcal{U}}{\leftrightsquigarrow} \tilde{V}) = \emptyset$.

- Definition: $j \rightsquigarrow k \mid L \iff W(j \overset{\text{via } \mathcal{U}}{\rightsquigarrow} k \mid L) \neq \emptyset$.

  "j has a direct causal effect on k given L"

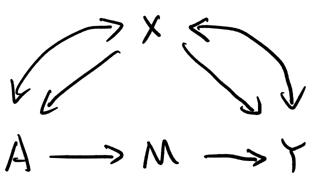
  $j \leftrightsquigarrow k \mid L \iff W(j \overset{\text{via } \mathcal{U}}{\leftrightsquigarrow} k \mid L) \neq \emptyset$

  "There is endogeneity between j and k given L".

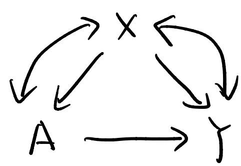
- Corollary: $\forall \tilde{V} \subset V$, $a, b \in \tilde{V}$, $C \subseteq \tilde{V}$, $a \neq b$, $a, b \notin C$.

$$a \left\{ \begin{matrix} \rightsquigarrow \\ \leftsquigarrow \\ \leftrightsquigarrow \end{matrix} \right\} b \mid C \; [\mathcal{G}] \iff a \left\{ \begin{matrix} \rightsquigarrow \\ \leftsquigarrow \\ \leftrightsquigarrow \end{matrix} \right\} b \mid C \; [\tilde{\mathcal{G}}]$$

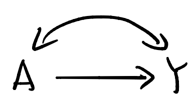
- Example:



$\xrightarrow[\{A, X, Y\}]{\text{projection}}$



$\xrightarrow[\{A, Y\}]{\text{projection}}$



$\xrightarrow[\{X, M\}]{\text{projection}}$

Slide 35 (A graphical criterion for conditional independence)

Definitions

- A walk is <u>m^*-connected</u> given L if all its colliders are in L and none of its non-colliders is in L.

- We say j and k are <u>m-connected</u> given L if there exists a m^*-connected walk like $j \leftrightsquigarrow * \leftsquigarrow k \mid L$.

- We say j and k are <u>confounded</u> given L if there exists a m^*-connected walk like $j \leftrightsquigarrow * \leftrightsquigarrow k \mid L$

- Unconfoundedness:

$$j \leftrightsquigarrow \cancel{*} \leftrightsquigarrow k \mid L \,[\mathcal{G}] \iff j \leftrightsquigarrow \cancel{*} \leftrightsquigarrow k \mid L \,[\tilde{\mathcal{G}}]$$

$$\iff j \leftrightarrow \cancel{*} \leftrightarrow k \,[\tilde{\mathcal{G}}] \Rightarrow \tilde{\Lambda} = \begin{array}{c} \\ \to j \\ \\ \to k \end{array} \begin{pmatrix} * & 0 \\ 0 & * \end{pmatrix}$$

Q: Why?

$$\Rightarrow (\tilde{\Lambda}^{-1})_{jk} = 0.$$

- m-separation: Recall $\tilde{\Omega} = \tilde{\Sigma}^{-1} = (I - \tilde{B}) \tilde{\Lambda}^{-1} (I - \tilde{B})^T$.

$$j \leftrightsquigarrow \cancel{*} \leftsquigarrow k \mid L \,[\mathcal{G}] \iff j \hookrightarrow \cancel{*} \leftarrow k \,[\mathcal{G}^*]$$

$$\Rightarrow (\tilde{\Omega})_{jk} = 0$$

- Completeness: For almost all linear SEM w.r.t. $\mathcal{G}$,

$$j \leftrightsquigarrow \cancel{*} \leftsquigarrow k \mid L \,[\mathcal{G}] \iff (\tilde{\Omega})_{jk} = 0$$

- m-connected path: we say a path is m-connected given L if

  1. none of its non-colliders is in L

  2. for any collider on the path, it is either in L or has a descendant in L.

It can be shown that there is no m^*-connected walk from j to k given L iff there is no m-connected path from j to k given L.

Slide 37 (Identifiability problems in linear SEM)

- IV graph

$$\begin{tikzcd} Z \arrow[r, "\beta_{ZA}"] & A \arrow[r, "\beta_{AY}"] \arrow[r, leftrightarrow, bend left, "\lambda_{AY}"] & Y \end{tikzcd}$$

Generic identifiability: $\beta_{AY} = \dfrac{\mathrm{Cov}(Z, Y)}{\mathrm{Cov}(Z, A)}$ if $\beta_{ZA} \neq 0$.

- Factor analysis:

$$\begin{tikzcd} & F \arrow[dl, "\beta_1"'] \arrow[d, "\beta_2"] \arrow[drr, "\beta_p"] & & \\ X_1 & X_2 & \cdots & X_p \end{tikzcd}$$

F is unobserved.

<u>Thm</u> Assume $\mathrm{Var}(F) = 1$. Then $(\beta_1, \cdots, \beta_p)$ is generically identifiable up to a sign change iff $p \geq 3$.

Pf Denote $Var(\varepsilon_i) = \sigma_i^2$.

$$\Sigma = Cov(X) = \beta\beta^T + diag(\sigma_1^2, \cdots, \sigma_p^2).$$

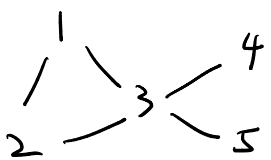
$p=3$:

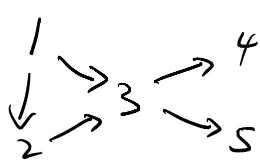
$$\Sigma = \begin{pmatrix} \Sigma_{11} & & \\ \Sigma_{12} & \Sigma_{22} & \\ \Sigma_{13} & \Sigma_{23} & \Sigma_{33} \end{pmatrix} = \begin{pmatrix} \beta_1^2+\sigma_1^2 & & \\ \beta_1\beta_2 & \beta_2^2+\sigma_2^2 & \\ \beta_1\beta_3 & \beta_2\beta_3 & \beta_3^2+\sigma_3^2 \end{pmatrix}$$

$$\Rightarrow \beta_1^2 = \frac{\Sigma_{12}\Sigma_{13}}{\Sigma_{23}}.$$ Similary for β_2^2, β_3^2. □

- Measurement model: In psychometrics, it is often of interest to infer causal relationships between abstract concepts that are measured by questionnaires. See example paper.

Slide 40 (Factorization)





$f(v) = \psi(v_1, v_2, v_3)\,\psi(v_3, v_4)\,\psi(v_3, v_5)$

$f(v) = f(v_1)\,f(v_2|v_1)\,f(v_3|v_1,v_2) \cdot f(v_4|v_3)\,f(v_5|v_3)$

Slide 44 (Global Markov $\Longleftrightarrow$ Factorization in DAG models)

WLOG suppose $(1, \cdots, p)$ is a topological ordering.

It's obvious that factorization is equivalent to

$$V_j \perp\!\!\!\perp V_{[j-1] \setminus pa(j)} \mid V_{pa(j)} \ , \forall j \qquad \text{(Ordered Markov)}$$

It suffices to show Order Markov $\Longrightarrow$ Global Markov.

Below: Proof by induction.

<u>Induction hypothesis</u> If $J \cup K \cup L \subseteq [m]$,

$$J \not\leftrightsquigarrow K \mid L \Rightarrow V_J \perp\!\!\!\perp V_K \mid V_L.$$

$m = 2$ ✓

Now suppose it is true up to $m-1$.

Also given: $V_m \perp\!\!\!\perp V_{[m-1] \setminus pa(m)} \mid V_{pa(m)}$.

$$J \not\leftrightsquigarrow K \mid L$$

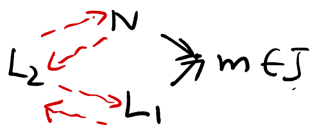
Want to prove: $V_J \perp\!\!\!\perp V_K \mid X_L$.

Idea: Apply the chain rule.

Let $N = pa(m) \setminus L$. $L_1 = L \cap pa(m)$, $L_2 = L \setminus pa(m) \setminus \{m\}$

Three cases: ① $m \in J$ ② $m \in K$ ③ $m \in L$.

① $m \in J$ $\quad m \not\leftrightsquigarrow K \mid L$.



So $K \not\to m$, $K \cap N = \phi$, $K \not\leftrightsquigarrow N \mid L$.

$$\Rightarrow J \cup N \not\leftrightsquigarrow K \mid L.$$

$$\overset{(1)}{\Rightarrow} V_{J \cup N \setminus \{m\}} \perp\!\!\!\perp V_K \mid V_L.$$ By (Induction hypothesis)

Further,

$$m \not\leftrightsquigarrow K \mid J \cup N \cup L \setminus \{m\}$$

$$\overset{(2)}{\Rightarrow} V_m \perp\!\!\!\perp V_K \mid V_L, V_{J \cup N \setminus \{m\}}$$ By (Ordered Markov)

$$\overset{(1)}{\underset{(2)}{\Rightarrow}} V_{J \cup N} \perp\!\!\!\perp V_K \mid V_L$$ (chain rule)

$$\Rightarrow V_J \perp\!\!\!\perp V_K \mid V_L.$$

② $m \in K$. $\quad$ Symmetric.

③ $m \in L$

$$\begin{tikzcd} & N \arrow[dl, dashed, bend left] \arrow[dr] & \\ L_2 \arrow[ur, dashed] \arrow[dr, dashed] & & m \in L \\ & L_1 \arrow[ul, dashed, bend left] \arrow[ur] & \end{tikzcd}$$

<u>Claim</u> $J \cap N = \emptyset$ and $N \rightsquigarrow \not{*} \leftsquigarrow J \mid L$ or $K \cap N = \emptyset$ and $N \rightsquigarrow \not{*} \leftsquigarrow K \mid L$

Otherwise $J \rightsquigarrow * \leftsquigarrow N \to m \leftarrow N \rightsquigarrow * \leftsquigarrow K \mid L$.

WLOG, suppose $N \rightsquigarrow \not{*} \leftsquigarrow K \mid L$.

$\Rightarrow$ $K \not\to m$, $J \cup N \rightsquigarrow \not{*} \leftsquigarrow K \mid L$.

Because m is the last vertex and only has edges like $\to m$, this shows

$$J \cup N \rightsquigarrow \not{*} \leftsquigarrow K \mid L \setminus \{m\}$$

$$\overset{(3)}{\Rightarrow} V_{J \cup N} \perp\!\!\!\perp V_K \mid V_{L \setminus \{m\}}.$$

Ordered local Markov ($pa(m) = N \cup L_1$)

$$\Rightarrow V_m \perp\!\!\!\perp V_{[m-1] \setminus N \setminus L_1} \mid V_{N \cup L_1}.$$

$$\overset{(4)}{\Rightarrow} V_m \perp\!\!\!\perp V_K \mid V_{J \cup N \cup L \setminus \{m\}} \qquad \text{(Chain rule)}$$

$$(3), (4) \Rightarrow V_K \perp\!\!\!\perp V_{J \cup N \cup \{m\}} \mid V_{L \setminus \{m\}}$$

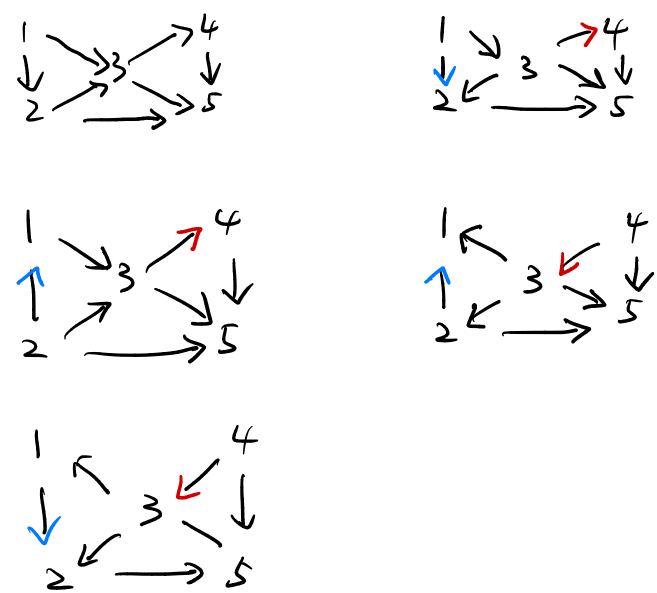
$$\Rightarrow V_K \perp\!\!\!\perp V_J \mid V_L. \qquad \text{(Chain rule)}$$

$\square$

Slide 47 (Markov equivalence of DAG models)

Markov equivalence class:

Ex 1



Slide 50 (NPSEM)

- Recursive substitution
  - Basic potential outcomes: $V_j(v_{[p]}) = f_j(v_{pa(j)}, \varepsilon_j)$, $j \in \mathcal{V}$.
  - Derived potential outcomes: $V_j(v_I) = V_j(v_{pa(j) \cap I}, V_{pa(j) \setminus I}(v_I))$.

  This defines a "natural counterfactual" $V_i(v_I)$ for $i \in I$.

- Example

$$\begin{tikzcd} V_1 \arrow[r] \arrow[rr, bend left] & V_2 \arrow[r] & V_3 \end{tikzcd}$$

Basic: $\quad V_1(v_1,v_2,v_3) = V_1, \quad V_2(v_1,v_2,v_3) = V_2(v_1)$

$V_3(v_1,v_2,v_3) = V_3(v_1,v_2).$

$I = \{2\}$: $\quad V_1(v_2) = V_1, \quad V_2(v_2) = V_2, \quad V_3(v_2) = V_3(V_1, v_2).$

- Simplification of potential outcomes.

Prop Disjoint $I, I' \subseteq V$. For any $j \in V$,

$$I' \not\rightsquigarrow j \mid I \implies V_j(v_I, v_{I'}) = V_j(v_I).$$

Corollary $\quad V_j(v_I) = V_j(v_{I \cap an(j)}).$

- Consistency of potential outcomes.

It follows from the definition that $V_j = V_j(v_\phi) = V_j(V_{pa(j)}).$

Prop Disjoint $I, I' \subseteq V$.

$$V_{I'}(v_I) = v_{I'} \implies V(v_I, v_{I'}) = V(v_I).$$

Slide 51 (Markov properties of basic p.o.)

Ex

$$\begin{tikzcd} V_1 \arrow[r] \arrow[rr, bend left] & V_2 \arrow[r] & V_3 \end{tikzcd}$$

- Single-world: $\quad V_1 \perp\!\!\!\perp V_2(v_1) \perp\!\!\!\perp V_3(v_1,v_2), \quad \forall v_1, v_2.$
- Multiple-world: $\quad V_1 \perp\!\!\!\perp (V_2(v_1) : v_1) \perp\!\!\!\perp (V_3(v_1,v_2) : v_1, v_2)$

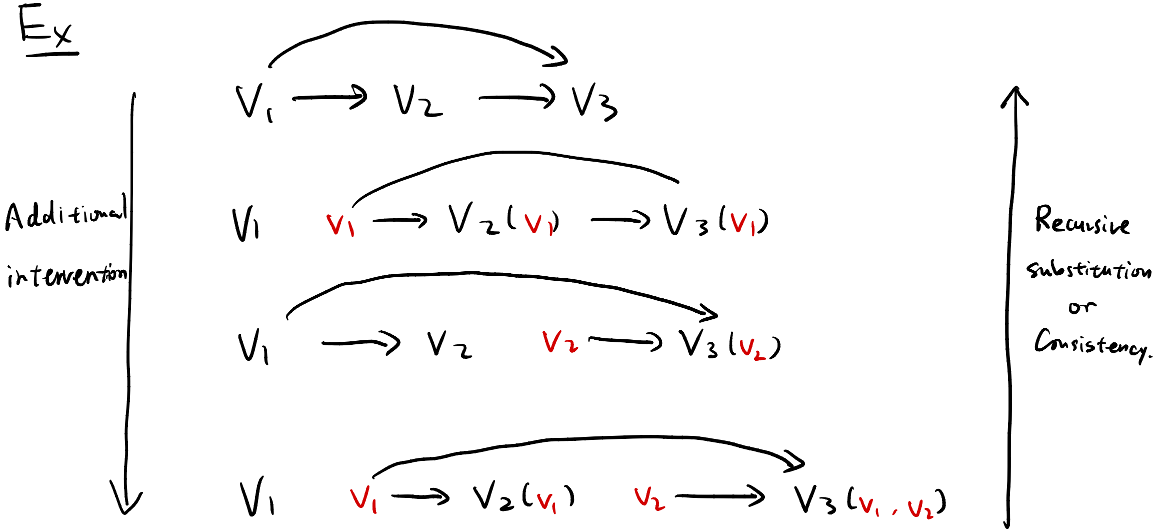
which includes all single-world independence but also cross-world independence like $V_2(v_1) \perp\!\!\!\perp V_3(v_1', v_2)$.

- NPSEM via basic p.o. : $E_j = (V_j(v_{pa(j)}) : v_{pa(j)})$.

  f_j selects the corresponding p.o.

Slide 52 (Representing recursive substitution as a graph)

Ex



Slide 53 (Recursive substitution preserves global Markov)

Proof by induction : Suppose $de(j) \subseteq I$ and $I' = I \cup \{j\}$. If $V(v_{I'})$ is Markov w.r.t. $\mathcal{G}[I']$, then $V(v_I)$ is Markov w.r.t. $\mathcal{G}[I]$.

Claim 1 Disjoint $K, L, M \subseteq V$, $j \in L$. Then

$$V_K(v_I) \not\leftrightsquigarrow V_L(v_I) \mid V_M(v_I) \Rightarrow V_K(v_I) \perp\!\!\!\perp V_L(v_I) \mid V_M(v_I).$$

Pf By assumption, $j \not\to K$. Then

$$V_K(v_I) \not\leftrightsquigarrow V_L(v_I) \mid V_M(v_I)$$

$$\Rightarrow V_K(v_I) \rightsquigarrow \not* \leftsquigarrow \left(V_L(v_I), V_{M\cap ch(j)}(v_I)\right) \mid V_{M\setminus ch(j)}(v_I). \quad (1)$$

[Because $V_{ch(j)}(v_I)$ has no outgoing edges like $V_{ch(j)}(v_I) \to *$
See ES2 Q9.]

$$\Rightarrow V_K(v_{I'}) \rightsquigarrow \not* \leftsquigarrow \left(V_L(v_{I'}), V_{M\cap ch(j)}(v_{I'})\right) \mid V_{M\setminus ch(j)}(v_{I'}).$$

$$\Rightarrow V_K(v_I, v_j) \perp\!\!\!\perp \left(V_L(v_I, v_j), V_{M\cap ch(j)}(v_I, v_j)\right) \mid V_{M\setminus ch(j)}(v_I, v_j).$$

$$\Rightarrow V_K(v_I) \perp\!\!\!\perp \left(V_L(v_I), V_{M\cap ch(j)}(v_I)\right) \mid V_{M\setminus ch(j)}(v_I).$$

[$j \in L$. consistency. Lemma 2]

<u>Claim 2</u> Disjoint $K, L, M \subseteq V$, $j \notin K \cup L$. Then

$$V_K(v_I) \rightsquigarrow \not* \leftsquigarrow V_L(v_I) \mid V_M(v_I) \Rightarrow V_K(v_I) \perp\!\!\!\perp V_L(v_I) \mid V_M(v_I).$$

<u>Pf</u> If $(K \cup L \cup M) \cap ch(j) = \emptyset$. Trivial.

Now suppose $j \to K \cup L \cup M$.

<u>Observation 1</u> $V_K(v_I) \rightsquigarrow \not* \leftsquigarrow V_L(v_I) \mid V_M(v_I), V_j(v_I)$.

Otherwise $V_K(v_I) \rightsquigarrow * \leftrightsquigarrow V_j(v_I) \leftrightsquigarrow * \leftsquigarrow V_L(v_I) \mid V_M(v_I), V_j(v_I)$

This leads to contradictions when $j \to K \cup L \cup M$.

<u>Observation 2</u> Either $V_K(v_I) \rightsquigarrow \not* \leftsquigarrow V_j(v_I) \mid V_M(v_I)$

or $V_L(v_I) \rightsquigarrow \not* \leftsquigarrow V_j(v_I) \mid V_M(v_I)$

Otherwise $V_K(v_I) \rightsquigarrow * \leftsquigarrow V_j(v_I) \rightsquigarrow * \leftsquigarrow V_L(v_I) \mid V_M(v_I)$

This contradicts the assumption or Observation 1.

WLOG Suppose $\quad V_K(v_I) \not\leftrightsquigarrow V_j(v_I) \mid V_M(v_I)$

$$\Rightarrow \quad V_K(v_I) \not\leftrightsquigarrow V_{L\cup\{j\}}(v_I) \mid V_M(v_I).$$

$$\overset{\text{Claim 1}}{\Rightarrow} \quad V_K(v_I) \perp\!\!\!\perp V_{L\cup\{j\}}(v_I) \mid V_M(v_I). \qquad \square$$

Slide 54 (DAG causal models)

- Example:

So $\quad A \perp\!\!\!\perp Y(a) \mid X.$

$$\mathbb{P}(A=\tilde{a}, X=\tilde{x}, Y(a)=\tilde{y}) = \mathbb{P}(A=\tilde{a} \mid X=\tilde{x})\,\mathbb{P}(X=\tilde{x})\cdot\mathbb{P}(Y=\tilde{y} \mid A=a, X=\tilde{x})$$

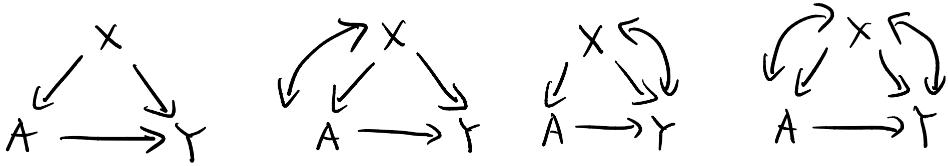
$$\Rightarrow \quad \mathbb{P}(Y(a)=\tilde{y}) = \sum_{\tilde{a},\tilde{x}} \mathbb{P}(A=\tilde{a} \mid X=\tilde{x})\,\mathbb{P}(X=\tilde{x})\cdot\mathbb{P}(Y=\tilde{y} \mid A=a, X=\tilde{x})$$

$$= \sum_{\tilde{x}} \mathbb{P}(X=\tilde{x})\,\mathbb{P}(Y=\tilde{y} \mid A=a, X=\tilde{x})$$

$$= \mathbb{E}\{\mathbb{P}(Y=\tilde{y} \mid A=a, X)\}.$$

Slide 55 (Back-door criterion)

Ex



Slide 56 (Front-door criterion)

$$\mathbb{P}(Y(a)=y) = \mathbb{P}(Y(a, M(a))=y)$$

$$= \mathbb{P}(Y(M(a))=y) \qquad [Y(a,m)=Y(m)]$$

$$= \sum_m \mathbb{P}(Y(m)=y \mid M(a)=m)\, \mathbb{P}(M(a)=m)$$

$$= \sum_m \mathbb{P}(Y(m)=y)\, \mathbb{P}(M(a)=m)$$

$$= \sum_m \Big\{ \sum_{a'} \mathbb{P}(Y=y \mid M=m, A=a')\, \mathbb{P}(A=a') \Big\}\, \mathbb{P}(M=m \mid A=a)$$

"Nonparametric path analysis".

Slide 57 (The fixing operator)

Pf $\quad \mathcal{V} = \{i\} \cup de(i) \cup nd(i)$. $\qquad nd =$ non-descendant.

Claim $\quad V_i(v_i)$ has $\not\ast\!\leftarrow$ any $V_{de(i)}(v_i) \mid V_{nd(i)}(v_i)$

$\Longleftrightarrow \quad V_i(v_i) \leftrightarrow \not\ast \leftarrow V_{de(i)}(v_i) \mid V_{nd(i)}(v_i)$

$\Longleftrightarrow \quad V_i(v_i) \leftrightarrow \not\ast \leftrightarrow V_{de(i)}(v_i) \mid V_{nd(i)}(v_i)$

because V_i is split $\quad$ acyclicity (so $V_{de(i)} \not\to V_{nd(i)}$)

$\Longleftrightarrow \quad i$ fixable.

Thus $\quad \mathbb{P}(V(v_i) = \tilde{v})$

$$= \mathbb{P}(V_{nd(i)}(v_i) = \tilde{v}_{nd(i)})\, \mathbb{P}(V_i(v_i) = \tilde{v}_i \mid V_{nd(i)}(v_i) = \tilde{v}_{nd(i)})$$

$$\cdot\, \mathbb{P}(V_{de(i)}(v_i) = \tilde{v}_{de(i)} \mid V_{nd(i)}(v_i) = \tilde{v}_{nd(i)},\ V_i(v_i) = \tilde{v}_i)$$

$$= \mathbb{P}(V_{nd(i)} = \tilde{v}_{nd(i)}) \cdot \mathbb{P}(V_i = \tilde{v}_i \mid V_{nd(i)} = \tilde{v}_{nd(i)})$$

$$\cdot\ \mathbb{P}(V_{de(i)}(v_i) = \tilde{v}_{de(i)} \mid V_{nd(i)}(v_i) = \tilde{v}_{nd(i)},\ V_i(v_i) = \underset{\text{Claimed } \perp\!\!\!\perp}{v_i})$$

$$= \mathbb{P}(V_{nd(i)} = \tilde{v}_{nd(i)}) \cdot \mathbb{P}(V_i = \tilde{v}_i \mid V_{nd(i)} = \tilde{v}_{nd(i)})$$

$$\cdot\ \mathbb{P}(V_{de(i)} = \tilde{v}_{de(i)} \mid V_{nd(i)} = \tilde{v}_{nd(i)},\ V_i = v_i)$$

So
$$\frac{\mathbb{P}(V(v_i) = \tilde{v})}{\mathbb{P}(V_i = v_i,\ V_{V\setminus\{i\}} = \tilde{v}_{V\setminus\{i\}})} = \frac{\mathbb{P}(V_i = \tilde{v}_i \mid V_{nd(i)} = \tilde{v}_{nd(i)})}{\mathbb{P}(V_i = v_i \mid V_{nd(i)} = \tilde{v}_{nd(i)})}$$

$$= \frac{\mathbb{P}(V_i = \tilde{v}_i \mid V_{mb(i)} = \tilde{v}_{mb(i)})}{\mathbb{P}(V_i = v_i \mid V_{mb(i)} = \tilde{v}_{mb(i)})}$$

__Exercise__ Show that if i is fixable, then $i \rightsquigarrow \ast \leftsquigarrow nd(i)\setminus mb(i) \mid mb(i)$.

$\square$

- Examples:

1.

$$\begin{tikzcd} A \arrow[r] \arrow[rr, leftrightarrow, bend left] & M \arrow[r] & Y \end{tikzcd}$$

M is fixable

$$\begin{tikzcd} A \arrow[r] \arrow[rr, leftrightarrow, bend left] & M \quad m \arrow[r] & Y(m) \end{tikzcd}$$

A is fixable

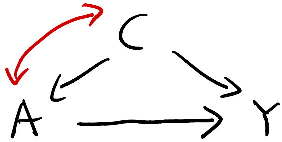
$$\begin{tikzcd} A \quad a \arrow[r] \arrow[rr, leftrightarrow, bend left] & M(a) \quad m \arrow[r] & Y(m) \end{tikzcd}$$

$\Rightarrow$ Prob. dist. of $(A, M(a), Y(m))$ identifiable.

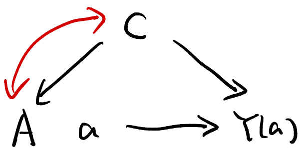
$$\Rightarrow \mathbb{P}(Y(a) = y) = \mathbb{E}\{\mathbb{P}(Y(a) = y \mid M(a))\}$$

$$= \mathbb{E}\{\mathbb{P}(Y(M(a)) = y \mid M(a))\} \quad \text{identifiable.}$$

2.

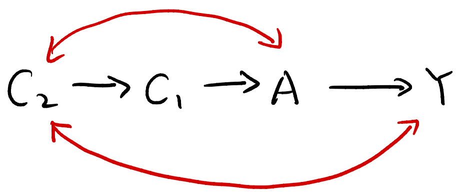


A is fixable



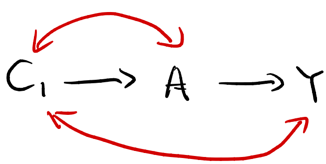
$\Longrightarrow$ Prob. dist. of $(A, C, Y(a))$ identifiable.

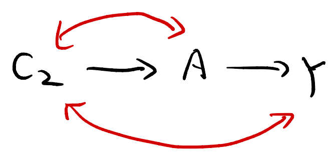
3.



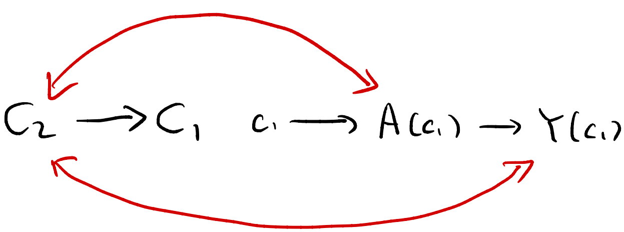
This is a challenging problem because the following are not identifiable

Projection on $\{C_1, A, Y\}$ $\{C_2, A, Y\}$





But we can first fix C_1 and then marginalize C_1, C_2



$c_1 \to A(c_1) \to Y(c_1)$

Now $A(c_1)$ is fixable: $\quad C_1 \longrightarrow A(c_1) \quad a \longrightarrow Y(a)$

Identification formula:

$$\mathbb{P}(A(c_1) = a, Y(c_1) = y)$$

$$= \sum_{\tilde{c}_1, \tilde{c}_2} \mathbb{P}(C_2 = \tilde{c}_2, C_1 = c_1, A = a, Y = y) \cdot \frac{\mathbb{P}(C_1 = \tilde{c}_1 \mid C_2 = \tilde{c}_2)}{\mathbb{P}(C_1 = c_1 \mid C_2 = \tilde{c}_2)}$$

$$= \sum_{c_2} \mathbb{P}(C_2 = c_2) \cdot \mathbb{P}(A = a, Y = y \mid C_1 = c_1, C_2 = c_2)$$

So $\mathbb{P}(Y(a) = y)$

$$= \mathbb{P}(Y(c_1) = y \mid A(c_1) = a)$$

$$= \frac{\mathbb{E}\{\mathbb{P}(A = a, Y = y \mid C_1 = c_1, C_2)\}}{\mathbb{E}\{\mathbb{P}(A = a \mid C_1 = c_1, C_2)\}}$$

Slide 67 (von Mises expansion)

- Empirical distribution: $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{V_i} \rightarrow \mathbb{P}$ (Glivenko-Cantelli)

  at rate $\frac{1}{\sqrt{n}}$ (Donsker)

  δ_V means the point mass at V.

  Thus we may expect $\hat{\beta}_{\text{plug-in}} = \beta(\mathbb{P}_n)$ to converge to $\beta(\mathbb{P})$ at the same rate if β is "smooth".

- von-Mises expansion.

  The definition of Gateaux differentiability implies

$$\beta(\mathbb{P} + \varepsilon(\mathbb{Q} - \mathbb{P})) - \beta(\mathbb{P}) = \varepsilon\, \beta'_{\mathbb{P}}(\mathbb{Q} - \mathbb{P}) + o(\varepsilon).$$

  Substituting $\varepsilon = \frac{1}{\sqrt{n}}$, $\mathbb{Q} - \mathbb{P} = \sqrt{n}(\mathbb{P}_n - \mathbb{P})$, we obtain

$$\beta(\mathbb{P}_n) - \beta(\mathbb{P}) = \frac{1}{\sqrt{n}} \beta'_{\mathbb{P}}(\sqrt{n}(\mathbb{P}_n - \mathbb{P})) + \underbrace{R(\mathbb{P}_n, \mathbb{P})}_{\text{should be } o_{\mathbb{P}}(\frac{1}{\sqrt{n}})}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \underbrace{\beta'_{\mathbb{P}}(\delta_{V_i} - \mathbb{P})}_{\text{"Influence" of } V_i.} + R(\mathbb{P}_n, \mathbb{P}).$$

Slide 68 (Influence function)

Define $\phi_{\mathbb{P}}(v) = \beta'_{\mathbb{P}}(\delta_v - \mathbb{P})$.

- Properties : ① $\mathbb{E}_{\mathbb{P}}(\phi_{\mathbb{P}}(V)) = \mathbb{E}_{\mathbb{P}}\{\beta'_{\mathbb{P}}(\delta_V - \mathbb{P})\}$

$$= \beta'_{\mathbb{P}}\{\mathbb{E}_{\mathbb{P}}(\delta_V - \mathbb{P})\}$$

$$= \beta'_{\mathbb{P}}(0) = 0.$$

② $\mathbb{E}_{\mathbb{Q}}(\phi_{\mathbb{P}}(V)) = \mathbb{E}_{\mathbb{Q}}\{\beta'_{\mathbb{P}}(\delta_V - \mathbb{P})\}$

$$= \beta'_{\mathbb{P}}\{\mathbb{E}_{\mathbb{Q}}(\delta_V - \mathbb{P})\}$$

$$= \beta'_{\mathbb{P}}(\mathbb{Q} - \mathbb{P})$$

$$= \frac{\partial}{\partial \varepsilon}\beta(\mathbb{P}_\varepsilon)\Big|_{\varepsilon=0}.$$

In other words, $\phi_{\mathbb{P}}$ is the Riesz representation of $\beta'_{\mathbb{P}}$.

- Examples :

① Mean : $\beta(\mathbb{P}) = \mathbb{E}_{\mathbb{P}}(V)$.

$$\hat{\beta}_{\text{plug-in}} = \beta(\mathbb{P}_n) = \mathbb{E}_{\mathbb{P}_n}(V) = \frac{1}{n}\sum_{i=1}^{n} V_i.$$

Gateaux : $\beta'_{\mathbb{P}}(\mathbb{Q} - \mathbb{P}) = \mathbb{E}_{\mathbb{Q}}(V) - \mathbb{E}_{\mathbb{P}}(V)$.

IF : $\phi_{\mathbb{P}}(V) = V - \mathbb{E}_{\mathbb{P}}(V)$.

$R(\mathbb{P}_n, \mathbb{P}) = 0$. Not surprising because β is linear

② Z-estimation. $\beta = \beta(\mathbb{P})$ defined by

$$\mathbb{E}_{\mathbb{P}}(m(\beta; V)) = 0.$$

Plug-in estimator solves

$$\frac{1}{n}\sum_{i=1}^{n} m(\beta; V_i) = 0$$

IF: is $\phi_{\mathbb{P}}(V) = \mathbb{E}\left(\left\{\frac{\partial}{\partial\beta} m(\beta; V)\right\}^{-1}\right) m(\beta; V)$.

Slide 69 (Bias correction using IF)

- Heuristics for $\hat{\beta}_{1\text{-step}}$.

Let's restate the von Mises expansion:

$$\beta(\mathbb{Q}) - \beta(\mathbb{P}) = \mathbb{E}_{\mathbb{Q}}(\phi_{\mathbb{P}}(V)) + R(\mathbb{Q}, \mathbb{P}) \quad (*)$$

If $\mathbb{Q}$ is "close" to $\mathbb{P}$, consider

$$\tilde{\mathbb{P}}_\varepsilon = \mathbb{P} + \varepsilon \frac{\mathbb{Q} - \mathbb{P}}{\|\mathbb{Q} - \mathbb{P}\|}$$

Then we may expect

$$|R(\mathbb{Q}, \mathbb{P})| \le O(\varepsilon^2) = O(\|\mathbb{Q} - \mathbb{P}\|^2)$$

Now setting $\mathbb{Q} = \mathbb{P}$ and $\mathbb{P} = \widehat{\mathbb{P}}$ in $(*)$,

$$\beta(\mathbb{P}) - \beta(\widehat{\mathbb{P}}) = \underbrace{\mathbb{E}_{\mathbb{P}}(\phi_{\widehat{\mathbb{P}}}(V))}_{-\text{ bias}} + \underbrace{R(\mathbb{P}, \widehat{\mathbb{P}})}_{\text{remaider}}.$$

We may thus improve $\beta(\widehat{\mathbb{P}})$ by

$$\widehat{\beta}_{1\text{-step}} = \beta(\widehat{\mathbb{P}}) + \mathbb{E}_{\mathbb{P}_n}(\phi_{\widehat{\mathbb{P}}}(V)).$$

This is similar to take one Newton-Raphson step at $\widehat{\mathbb{P}}$.

- Expansion

$$\widehat{\beta}_{1\text{-step}} - \beta(\mathbb{P}) = \mathbb{E}_{\mathbb{P}_n}(\phi_{\widehat{\mathbb{P}}}(V)) - \mathbb{E}_{\mathbb{P}}(\phi_{\widehat{\mathbb{P}}}(V)) - R(\mathbb{P}, \widehat{\mathbb{P}})$$

$$= \underbrace{\mathbb{E}_{\mathbb{P}_n - \mathbb{P}}(\phi_{\mathbb{P}}(V))}_{\text{CLT}} + \underbrace{\mathbb{E}_{\mathbb{P}_n - \mathbb{P}}(\phi_{\widehat{\mathbb{P}}}(V) - \phi_{\mathbb{P}}(V))}_{\text{Empirical process}} - \underbrace{R(\mathbb{P}, \widehat{\mathbb{P}})}_{\text{Second-order}}.$$

where $\mathbb{E}_{\mathbb{Q}-\mathbb{P}}(\cdot) = \mathbb{E}_{\mathbb{Q}}(\cdot) - \mathbb{E}_{\mathbb{P}}(\cdot)$.

- Cross-fitting: the empirical process term shows up because $\widehat{\mathbb{P}}$ depends on $\mathbb{P}_n$. This prevents us from claiming

$$\mathbb{E}_{\mathbb{P}_n - \mathbb{P}}(\phi_{\widehat{\mathbb{P}}}(V) - \phi_{\mathbb{P}}(V)) = O\left(\frac{1}{\sqrt{n}} \| \phi_{\widehat{\mathbb{P}}} - \phi_{\mathbb{P}} \|\right).$$

Let $\mathbb{P}_n^{(1)} = \frac{1}{n/2}\sum_{i=1}^{n/2} \delta_{V_i}$, $\mathbb{P}_n^{(2)} = \frac{1}{n/2}\sum_{i=n/2+1}^{n} \delta_{V_i}$.

So $\mathbb{P}_n = \frac{1}{2}(\mathbb{P}_n^{(1)} + \mathbb{P}_n^{(2)})$.

Let $\widehat{\mathbb{P}}^{(k)}$ be a smooth estimator of $\mathbb{P}$ based on $\mathbb{P}_n^{(k)}$, $k=1,2$. Denote $\mathbb{P}_n^{(-1)} = \mathbb{P}_n^{(2)}$ and $\mathbb{P}_n^{(-2)} = \mathbb{P}_n^{(1)}$.

Define $\widehat{\beta}^{(k)} = \beta(\widehat{\mathbb{P}}^{(-k)}) + \mathbb{E}_{\mathbb{P}_n^{(k)}}\left(\phi_{\widehat{\mathbb{P}}^{(-k)}}(V)\right)$, $k=1,2$

and $\widehat{\beta} = \frac{1}{2}(\widehat{\beta}^{(1)} + \widehat{\beta}^{(2)})$.

Then
$$\begin{aligned}\widehat{\beta} - \beta &= \frac{1}{2}\sum_{k=1}^{2}\Big\{\mathbb{E}_{\mathbb{P}_n^{(k)} - \mathbb{P}}\left(\phi_{\mathbb{P}}(V)\right)\\
&\quad + \mathbb{E}_{\mathbb{P}_n^{(k)} - \mathbb{P}}\left(\phi_{\widehat{\mathbb{P}}^{(-k)}}(V) - \phi_{\mathbb{P}}(V)\right)\\
&\quad + R(\mathbb{P}, \widehat{\mathbb{P}}^{(-k)})\Big\}\\
&= \mathbb{E}_{\mathbb{P}_n - \mathbb{P}}\left(\phi_{\mathbb{P}}(V)\right) + \frac{1}{2}\sum_{k=1}^{2} R(\mathbb{P}, \widehat{\mathbb{P}}^{(-k)})\\
&\quad + \underbrace{o_P\left(\tfrac{1}{\sqrt{n}}\right)}\end{aligned}$$

If $\phi_{\widehat{\mathbb{P}}^{(-k)}} \xrightarrow{P} \phi_{\mathbb{P}}$, $k=1,2$.

Slide 70 (Calculus of IF)

Let $V = (X, Y)$. Consider the submodel induced by ε:

$$\mathbb{P}_\varepsilon(V=v) = (1-\varepsilon)\,\mathbb{P}_\varepsilon(V=v) + \varepsilon\,\delta_{\tilde{v}}$$

Then
$$\frac{\partial}{\partial\varepsilon}\mathbb{P}_\varepsilon(X=x)\Big|_{\varepsilon=0} = \left(\frac{\partial}{\partial\varepsilon}(1-\varepsilon)\cdot\mathbb{P}(X=x) + \varepsilon\mathbb{1}\{x=\tilde{x}\}\right)\Big|_{\varepsilon=0}$$
$$= \mathbb{1}\{x=\tilde{x}\} - \mathbb{P}(X=x)$$

So the IF is $\delta_x - \mathbb{P}(X=x)$

$$\frac{\partial}{\partial\varepsilon}\mathbb{E}_\varepsilon(Y|X=x)\Big|_{\varepsilon=0}$$
$$= \frac{\partial}{\partial\varepsilon}\sum_y y\cdot\frac{(1-\varepsilon)\cdot\mathbb{P}(V=v) + \varepsilon\cdot\mathbb{1}\{v=\tilde{v}\}}{(1-\varepsilon)\cdot\mathbb{P}(X=x) + \varepsilon\cdot\mathbb{1}\{x=\tilde{x}\}}\Bigg|_{\varepsilon=0}$$
$$= \sum_y y\cdot\frac{\left(\mathbb{1}\{v=\tilde{v}\} - \cancel{\mathbb{P}(V=v)}\right)\mathbb{P}(X=x) - \mathbb{P}(V=v)\cdot\left(\mathbb{1}\{x=\tilde{x}\} - \cancel{\mathbb{P}(X=x)}\right)}{\mathbb{P}(X=x)^2}$$
$$= \sum_y y\cdot\frac{\mathbb{1}\{v=\tilde{v}\} - \mathbb{1}\{x=\tilde{x}\}\cdot\mathbb{P}(Y=y|X=x)}{\mathbb{P}(X=x)}$$
$$= \frac{(\tilde{y} - \mathbb{E}(Y|X=x))\,\mathbb{1}\{x=\tilde{x}\}}{\mathbb{P}(X=x)}$$

So the IF is $\{Y - \mathbb{E}(Y|X=x)\}\cdot\dfrac{\delta_x}{\mathbb{P}(X=x)}$

Now consider $\beta = \mathbb{E}\left(\mathbb{E}(Y \mid A=1, X) \right)$.

$$= \sum_x \mathbb{E}(Y \mid A=1, X=x) \cdot \mathbb{P}(X=x).$$

$$IC(\beta) = \sum_x IC\left(\mathbb{E}(Y \mid A=1, X=x) \right) \cdot \mathbb{P}(X=x) + IC\left(\mathbb{P}(X=x) \right) \cdot \mathbb{E}(Y \mid A=1, X=x)$$

$$= \sum_x \frac{\mathbb{1}\{X=x, A=1\}}{\mathbb{P}(X=x, A=1)} \cdot \left(Y - \mathbb{E}(Y \mid X=x, A=1) \right) \cdot \mathbb{P}(X=x)$$

$$+ \left(\mathbb{1}\{X=x\} - \mathbb{P}(X=x) \right) \cdot \mathbb{E}(Y \mid A=1, X=x).$$

$$= \frac{A}{\mathbb{P}(A=1 \mid X)} \cdot \left(Y - \mathbb{E}(Y \mid X, A=1) \right) + \mathbb{E}(Y \mid A=1, X) - \beta.$$

$$= \frac{A}{\pi(X)} \cdot \left(Y - \mu_1(X) \right) + \mu_1(X) - \beta.$$

- AIPW / DR estimator

Initial estimator: $\hat{\beta}_{OR} = \frac{1}{n} \sum_{i=1}^{n} \hat{\mu}_1(X_i)$

One-step estimator (AIPW or DR): $\hat{\beta}_{DR} = \hat{\beta}_{OR} + \frac{1}{n} \sum_{i=1}^{n} IC(\beta; \hat{\pi}, \hat{\mu}_1)$

$$= \frac{1}{n} \sum_{i=1}^{n} \frac{A_i}{\hat{\pi}(X_i)} \cdot \left(Y_i - \hat{\mu}_1(X_i) \right) + \hat{\mu}_1(X_i)$$

Remainder term:

$$
\begin{aligned}
R(\mathbb{P}, \hat{\mathbb{P}}) &= \beta(\mathbb{P}) - \beta(\hat{\mathbb{P}}) - \mathbb{E}_{\mathbb{P}}(\phi_{\hat{\mathbb{P}}}(V)) \\
&= \mathbb{E}_{\mathbb{P}}(\mu_1(X)) - \mathbb{E}_{\mathbb{P}}\left(\frac{A}{\hat{\pi}(X)} \cdot (Y - \hat{\mu}_1(X)) + \hat{\mu}_1(X)\right) \\
&= \mathbb{E}_{\mathbb{P}}(\mu_1(X) - \hat{\mu}_1(X)) - \mathbb{E}_{\mathbb{P}}\left(\frac{\pi(X)}{\hat{\pi}(X)} \cdot (\mu_1(X) - \hat{\mu}_1(X))\right) \\
&= -\mathbb{E}_{\mathbb{P}}\left(\left(\frac{1}{\hat{\pi}(X)} - \frac{1}{\pi(X)}\right)(\hat{\mu}_1(X) - \mu_1(X)) \cdot \pi(X)\right).
\end{aligned}
$$

Define: $\mathrm{MSE}(\hat{\pi}) = \mathbb{E}_{\mathbb{P}}((\hat{\pi}(X) - \pi(X))^2)$

So the remaider term $\xrightarrow{P} 0$ if $\hat{\pi} \to \pi$ or $\hat{\mu}_1 \to \mu_1$

"double robustness"

— —— ——— is negligible if $\mathrm{MSE}(\hat{\pi}) \cdot \mathrm{MSE}(\hat{\mu}_1) = o(\frac{1}{\sqrt{n}})$.

and $\pi(X) \geq C > 0$ for some C.

This theory can be easily extended to estimating the ATE, ATT...

Slide 72 (Entropy balancing)

- Lagrangian dual

$$
L(w, \lambda, \nu) = \sum_{i=n_1+1}^{n} w_i \log w_i + \sum_{i=n_1+1}^{n} \lambda_i w_i + \nu^T \left\{ \sum_{i=1}^{n_1} X_i - \sum_{i=n_1+1}^{n} w_i X_i \right\}
$$

$\lambda_i \leq 0$.

$$
\frac{\partial}{\partial w_i} L = \log w_i + 1 + \lambda_i - (\nu^T X_i)
$$

$$\Rightarrow \quad w_i^* = e^{(\nu^T X_i) - \lambda_i - 1}$$

Lagrangian dual function $g(\lambda, \nu) = \inf_w L(w, \lambda, \nu)$

$$= L(w^*, \lambda, \nu)$$

$$\frac{\partial}{\partial \lambda_i} g = (\log w_i^* + 1) \cdot \frac{\partial}{\partial \lambda_i} w_i^* + w_i^* + (\lambda_i - \nu^T X_i) \cdot \frac{\partial}{\partial \lambda_i} w_i^*$$

$$= - w_i^* (\log w_i^* + \lambda_i - \nu^T X_i)$$

$$= w_i^* > 0.$$

So optimal dual $\lambda_i^* = 0$.

Lagrangian dual optimization: Denote $\pi_i = \dfrac{e^{\nu^T X_i - 1}}{1 + e^{\nu^T X_i - 1}}$

$$\sup_\nu L(w^*, \lambda^*, \nu)$$

$$= \sup_\nu \sum_{i=n_1+1}^{n} \frac{\pi_i}{1-\pi_i} (\nu_i^T X_i - 1) + \nu^T \left\{ \sum_{i=1}^{n_1} X_i - \sum_{i=n_1+1}^{n} \frac{\pi_i}{1-\pi_i} X_i \right\}$$

$$= \sup_\nu - \sum_{i=n_1+1}^{n} \frac{\pi_i}{1-\pi_i} + \nu^T \sum_{i=1}^{n_1} X_i$$

$$= \inf_\nu \sum_{i=1}^{n} (1-A_i) \frac{\pi_i}{1-\pi_i} - A_i \log \frac{\pi_i}{1-\pi_i}.$$

Compare to the maximum likelihood problem:

$$\max_\nu \sum_{i=1}^{n} (1-A_i) \log(1-\pi_i) + A_i \log \pi_i.$$

Thus, the dual problem is fitting a logistic regression with a different loss function.

- Double robusteness

A benefit of this is that when one we do the one-step correction to $\hat{\beta} = \sum_{i=1}^{n} (1-A_i) w_i Y_i$, we get

$$\hat{\beta}_{DR} = \cancel{\frac{1}{n}\sum_{i=1}^{n}(1-A_i) w_i Y_i}$$

$$+ \frac{1}{n}\sum_{i=1}^{n}(1-A_i) w_i (Y_i - \hat{\mu}_0(X_i)) + A_i \hat{\mu}_0(X_i) - \cancel{\frac{1}{n}\sum_{i=1}^{n}(1-A_i) w_i Y_i}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(1-A_i) w_i Y_i + \underbrace{\frac{1}{n}\sum_{i=1}^{n} A_i \hat{\mu}_0(X_i) - (1-A_i) w_i \hat{\mu}_0(X_i)}_{=0 \text{ if } \hat{\mu}_0(X_i) \text{ is linear in } X_i.}$$

$$= \hat{\beta}$$

Slide 75 (Linear SEMs: Multiple IVs)

- Linear SEM: Assume all variables have mean 0.

$$A = Z^T \beta_{ZA} + X \beta_{XA} + \varepsilon_A$$

$$Y = A \beta_{AY} + X \beta_{XY} + \varepsilon_Y.$$

ε_A, ε_Y may be correlated. $(X, Z) \perp\!\!\!\perp (\varepsilon_A, \varepsilon_Y)$.

Then $\mathbb{E}(Y|Z,X) = \beta_{AY}\,\mathbb{E}(A|Z,X) + X\beta_{XY}$

- This motivates the two-stage least squares estimator:

1. Regress A on Z, X to obtain $\hat{\mathbb{E}}(A|Z,X)$.

2. Regress Y on $\hat{\mathbb{E}}(A|Z,X)$.

Slide 77 (Semiparametric estimation with IV)

- Plug-in estimator: Let $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and $\bar{A} = \frac{1}{n}\sum_{i=1}^{n} A_i$.

So $\hat{\alpha}(\beta) = \bar{Y} - \beta\bar{A}$. By solving the empirical estimating equation, we obtain

$$\hat{\beta}_g = \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})\,g(Z_i)}{\frac{1}{n}\sum_{i=1}^{n}(A_i - \bar{A})\,g(Z_i)} \longrightarrow \frac{\mathrm{Cov}(Y, g(Z))}{\mathrm{Cov}(A, g(Z))} = \beta.$$

For fixed $g(\cdot)$ and under regularity conditions, $\hat{\beta}_g$ is asymptotically linear with the following IF:

$$\boxed{\text{ES}}\quad \psi_g(Z,A,Y) = \frac{\left[\{Y-\mathbb{E}(Y)\} - \beta\{A-\mathbb{E}(A)\}\right]\left[g(Z) - \mathbb{E}\{g(Z)\}\right]}{\mathrm{Cov}(A, g(Z))}.$$

So $\sqrt{n}(\hat{\beta}_g - \beta) \xrightarrow{d} N(0, \sigma_g^2)$ as $n \to \infty$.

where $$\sigma_g^2 = \frac{\mathrm{Var}(Y-\beta A)\cdot\mathrm{Var}(g(Z))}{\mathrm{Cov}(A, g(Z))^2}$$

By Cauchy-Schwarz inequality, this is minimised at

$$g(Z) = g^*(Z) = \mathbb{E}(A \mid Z) \qquad \text{"optimal instrument"}.$$

To improve efficiency, it is common to estimate $g^*(\cdot)$ first and then plug it in $\hat{\beta}_g$. Let the resulting estimator be $\hat{\beta}_{\hat{g}}$. This is essentially the two-stage least squares.

Rem 1. As long as $\hat{g}$ is not too complex, empirical process theory suggests $\hat{\beta}_g - \hat{\beta}_{\hat{g}}$ is negligible (g is the limit of $\hat{g}$), so $\hat{\beta}_{\hat{g}}$ is still asymptotically normal.

2. A robustness property: $\hat{\beta}_g$ is $\sqrt{n}$-consistent as long as $\mathrm{Cov}(A, g(Z)) > 0$. Do not need a consistent estimator of $g^*(\cdot)$. This is similar to regression adjustment methods for randomised experiments.

Slide 79 (IV identification: Complier average causal effect)

Pf $$\beta = \frac{\mathrm{Cov}(Z, Y)}{\mathrm{Cov}(Z, A)} = \frac{\mathbb{E}(Y \mid Z=1) - \mathbb{E}(Y \mid Z=0)}{\mathbb{E}(A \mid Z=1) - \mathbb{E}(A \mid Z=0)}.$$

By assumption, $Z \perp\!\!\!\perp C$ and $\mathbb{P}(C = de) = 0$.

$$\mathbb{E}(Y \mid Z=1) = \sum_{c \in \{at, nt, co, de\}} \mathbb{E}\{Y \mid Z=1, C=c\} \, \mathbb{P}(C=c \mid Z=1)$$

$$\begin{aligned}
&= \mathbb{E}\{Y(1) \mid C=at\} \, \mathbb{P}(C=at) \\
&+ \mathbb{E}\{Y(0) \mid C=nt\} \, \mathbb{P}(C=nt) \\
&+ \mathbb{E}\{Y(1) \mid C=co\} \, \mathbb{P}(C=co) \\
&+ \mathbb{E}\{Y(0) \mid C=de\} \, \mathbb{P}(C=de)
\end{aligned}$$

Similarly $\mathbb{E}(Y \mid Z=0) = \mathbb{E}\{Y(1) \mid C=at\} \, \mathbb{P}(C=at)$

$$\begin{aligned}
&+ \mathbb{E}\{Y(0) \mid C=nt\} \, \mathbb{P}(C=nt) \\
&+ \mathbb{E}\{Y(0) \mid C=co\} \, \mathbb{P}(C=co) \\
&+ \mathbb{E}\{Y(1) \mid C=de\} \, \mathbb{P}(C=de)
\end{aligned}$$

So $\mathbb{E}(Y \mid Z=1) - \mathbb{E}(Y \mid Z=0) = \mathbb{E}\{Y(1) - Y(0) \mid C=co\} \, \mathbb{P}(C=co)$.

Similarly, $\mathbb{E}(A \mid Z=1) - \mathbb{E}(A \mid Z=0) = \mathbb{P}(C=co)$.

The desired conclusion then follows. $\square$