Causal Inference

Qingyuan Zhao

University of Cambridge

MT2023 Part III

Outline

Welcome

- Main course webpage: `http://www.statslab.cam.ac.uk/~qz280/teaching/causal-2023/`.
- Moodle page is not regularly maintained but will be used for announcements: `https://www.vle.cam.ac.uk/user/index.php?id=252866`.
- Example class:
  - Location: MR11.
  - Time: Thursday at 3:30pm on 2nd Nov, 23th Nov, 18th Jan.
  - Instructor: Jieru (Hera) Shi.
- Past lecture notes and literature can be found on course webpage.

What is causal inference?

- Infer causal relationships from experimental or observational data.
- This field has come under spotlight after several pioneers won major awards in recent years:
  - Judea Pearl (Turing Award, 2011);
  - Joshua Angrist & Guido Imbens (Nobel Memorial Prize in Economics, 2021);
  - James Robins, Miguel Hernán, Thomas Richardson, Andrea Rotnitzky, & Eric Tchetgen Tchetgen (Rousseeuw Prize for Statistics, 2022).

Why is causal inference important?

1. Ubiquitous in many scientific disciplines:
   - Medicine & biological sciences;
   - Economics, psychology, & social sciences;
   - Computer science & artificial intelligence;
   - Policy & business decisions.
2. Connects mathematical theory with the real world:
   - Classical statistics: models $\rightarrow$ inference.
   - Machine learning: data $\rightarrow$ prediction.[1]
   - Causal inference: models and inference $\leftrightarrow$ reality.

Useful quote by George Barnard[2]: "in statistical inference, as distinct from mathematical inference, there is a world of difference between the two statements "X is true" and "X is known to be true"."

[1] Recommended reading: "Statistical Modeling: The Two Cultures" by Leo Breiman, *Statistical Science*, 2001. See also the recent reprint and discussion in *Observational Studies*.

[2] Paraphrased, see `https://artowen.su.domains/links/` for the original quote.

Motivating example 1: Smoking and lung cancer

By the mid-1940s, it had been observed that lung cancer cases had tripled over the previous three decades. Possible explanations included

- Changes in air quality due to the introduction of automobile;
- Widespread expansion of paved roads that contained many carcinogens;
- Aging of the population;
- The advent of radiography;
- Better clinical awareness of lung cancer and better diagnostic methods;
- Smoking.

Advertisement for cigarette smoking (1950)

"I'm going to grow a hundred years old!"

. . . and possibly she may—for the amazing strides of medical science have added years to life expectancy

● It's a fact—a warm, wonderful fact—that this five-year-old child, or your own child, has a life expectancy almost a whole decade longer than was her mother's, and a good 18 to 20 years longer than that of her grandmother. Not only the expectation of a longer life, but of a life by far healthier.

Thank medical science for that. Thank your doctor and thousands like him . . . toiling ceaselessly . . . that you and yours may enjoy a longer, better life.

According to a recent Nationwide survey:

More Doctors smoke Camels

than any other cigarette!

NOT ONE but three outstanding independent research organizations conducted this survey. And they asked not just a few thousand, but 113,597, doctors from coast to coast to name the cigarette they themselves preferred to smoke.

Answers came in by the thousands . . . from general physicians, diagnosticians, surgeons, nose and throat specialists too. The most-named brand was Camel.

If you are not now smoking Camels, try them. Let your "T-Zone" tell you *(see right)*.

R. J. Reynolds Tobacco Co., Winston-Salem, N. C.

CAMELS *Costlier Tobaccos*

CAMEL
TURKISH & DOMESTIC BLEND CIGARETTES
CHOICE QUALITY

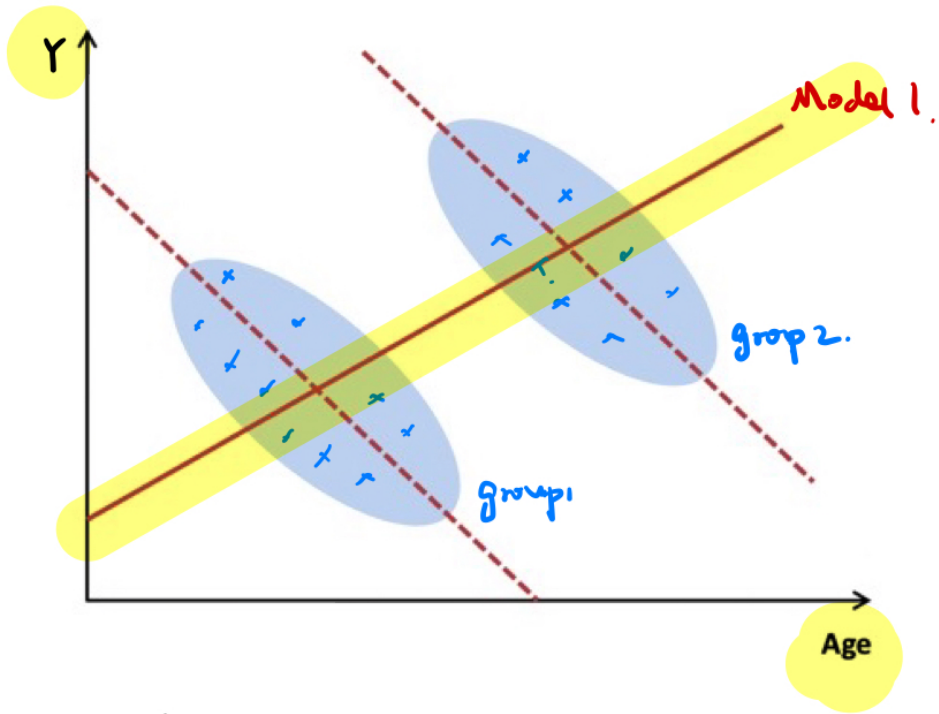
THE "T-ZONE" TEST WILL TELL YOU

The "T-Zone"—T for taste and T for throat—is your own proving ground for any cigarette. Only your taste and throat can decide which cigarette tastes best to you . . . how it affects your throat.

Correlation = causation?

- A series of observational studies since 1950 reported overwhelmingly strong association between smoking and lung cancer.
- Some prominent statisticians including R A Fisher objected to the idea that this implies that smoking causes lung cancer, but no compelling competing hypothesis could be found.
- This eventual led to one of the biggest public health intervention to reduce tobacco consumption.

Yule-Simpson paradox in linear regression analysis



Model 1 $\quad Y = \beta_1 + \beta_2 \cdot AGE + noise$

$\Rightarrow \beta_2 > 0$

Model 2 $\quad Y = \beta_1 + \beta_2 \cdot AGE + \beta_3 \cdot GROUP + noise$

$\Rightarrow \beta_2 < 0.$

Motivating example 2: Policy evaluation

- Do job training programmes increase the participants' earnings?
- Do gun control policies actually decrease gun homicides?
- Do mask mandates decrease the chance of getting COVID?
- Do new web designs increase a company's revenue?

Two more philosophical points

Cycle of scientific research (George Box)

hypothesis $\rightarrow$ **design** $\rightarrow$ data collection $\rightarrow$ **analysis** $\rightarrow$ hypothesis $\rightarrow \ldots$

Three principles of causal inference

Randomization $\subset$ Identification $\subset$ Elaboration

Outline

Randomized experiments: Introduction

- Originated from Fisher's seminal work on agricultural experiments in 1920-30s.
- Remains the "gold standard" for causal inference.
- Example: mRNA vaccines for COVID-19.

Notation

- Upper-case (lower-case) letters: random (fixed) quantities;
- Subscript: experimental unit.
- $[n] = \{1, \ldots, n\}$;

Problem setup

See blackboard:

- Variables: covariates, outcome, treatment, exposure.

Design of experiments

The *treatment assignment mechanism* or *randomization scheme* refers to the probability distribution

$$\pi(\boldsymbol{z} \mid \boldsymbol{x}) = \mathbb{P}(\boldsymbol{Z} = \boldsymbol{z} \mid \boldsymbol{X} = \boldsymbol{x}).$$

Examples

See blackboard:

1. Bernoulli trials.
2. Sampling without replacement.
3. Randomized complete block design.

Towards a formal theory

"Implicit" causal inference

Often, experimental data are analyzed using linear regression:

$$Y_i = \alpha + \beta A_i + \boldsymbol{\gamma}^T \boldsymbol{X}_i + \epsilon_i.$$

- Does $\beta \neq 0$ mean there is a causal effect?
- Or $\mathbb{E}[Y \mid A = 1] \neq \mathbb{E}[Y \mid A = 0]$?

Formal theory: Neyman-Rubin causal model

See blackboard:

- Neyman's potential outcome model.
- Consistency of potential outcomes.
- Fundamental problem of causal inference.[a]
- Validity of exposure mapping.

[a] Recommended reading: "Statistics and causal inference" by Paul Holland in *JASA*, 1986.

Example: Interference

- We will almost always assume the identity exposure mapping is valid.
- But this may not be true in many problems: vaccine studies, experiments on social networks (e.g. Instagram).
- See blackboard: n units interacting via a network.

Example: Lady tasting tea

Description of the experiment

A young lady claimed that she is able to tell whether the tea or the milk was added first to a cup.
The experiment provides the lady with eight randomly ordered cups of tea—four prepared by pouring milk and then tea, four by pouring tea and then milk. The lady attempts to select the four cups prepared by one method or the other, and may compare cups directly against each other as desired. The method employed in the experiment is fully disclosed to the lady.

See blackboard:

- Potential outcomes schedule.
- 2×2 contigency table.
- Fisher's exact test.

Randomization inference: General tests

- Randomization provides the "reasoned basis" for Fisher's exact test.

See blackboard:

- Exogeneity of randomization.
- Fisher's sharp null $\Rightarrow$ imputation of p. o. schedule $\boldsymbol{W} = (\boldsymbol{Y}(\boldsymbol{z}))_{\boldsymbol{z}\in\mathcal{Z}}$.
- Randomization p-value (p for probability):

$$\begin{aligned} P &= \mathbb{P}(T(\boldsymbol{Z}', \boldsymbol{X}, \boldsymbol{W}) \leq T(\boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{W}) \mid \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{W}) \\ &= \sum_{\boldsymbol{z}} 1_{\{T(\boldsymbol{z},\boldsymbol{X},\boldsymbol{W})\leq T(\boldsymbol{Z},\boldsymbol{X},\boldsymbol{W})\}} \pi(\boldsymbol{z} \mid \boldsymbol{x}). \end{aligned}$$

- Fisher's exact test: Neyman-Rubin model, $\mathcal{A} = \mathcal{Y} = \{0, 1\}$, sampling without replacement.
- Randomization test vs. permutation test vs. Monte-Carlo approximation.

Randomization inference: Estimation

- Setting: Neyman-Rubin model, $\mathcal{A} = \{0, 1\}$, sampling without replacement.

Notation

- Sample average treatment effect: $\beta_n = \frac{1}{n} \sum_{i=1}^{n} Y_i(1) - Y_i(0)$.
- Difference-in-means estimator: $\hat{\beta}_{\text{DIM}} = \bar{Y}_1 - \bar{Y}_0$.
- Denote $\bar{Y}(a) = \frac{1}{n} \sum_{i=1}^{n} Y_i(a)$, $S^2(a) = \frac{1}{n-1} \sum_{i=1}^{n} \{Y_i(a) - \bar{Y}(a)\}^2$,

  $S^2(0,1) = \frac{1}{n-1} \sum_{i=1}^{n} \{Y_i(1) - Y_i(0) - \beta_n\}^2$.

See blackboard:

- Mean and variance of the randomization distribution of $\hat{\beta}_{\text{DIM}}$.
- Variance estimator and finite-sample CLT.

Randomization inference: F-test

- Setting: Neyman-Rubin model, $\mathcal{A} = \{0, 1, \ldots, k-1\}$.
- Normal linear model: $Y_i = \sum_{j=0}^{k-1} \mu_j 1_{\{A_i=j\}} + \epsilon_i$, $\epsilon_i \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2)$.
- OLS: $\hat{\mu}_j = \bar{Y}_j = \frac{1}{N_j} \sum_{i=1}^{n} Y_i 1_{\{A_i=j\}}$, where $N_j = \sum_{i=1}^{n} 1_{\{A_i=j\}}$.
- ANOVA: $S_A = \sum_{i=1}^{n} (\hat{\mu}_{A_i} - \bar{Y})^2$, $S_E = \sum_{i=1}^{n} (Y_i - \hat{\mu}_{A_i})^2$. Then under $H_0 : \mu_0 = \cdots = \mu_{k-1}$, we have

$$
\begin{aligned}
& S_A/\sigma^2 \sim \chi^2_{k-1},\ S_E/\sigma^2 \sim \chi^2_{n-k} \\
\Rightarrow & \frac{\mathbb{E}(S_A)}{\mathbb{E}(S_E)} = \frac{k-1}{n-k},\ F = \frac{S_A/(k-1)}{S_E/(n-k)} \sim F_{k-1,n-k}.
\end{aligned}
$$

See blackboard:

- Randomization distribution of S_A and S_E.

Repeated sampling

- Setting: Neyman-Rubin model, $\mathcal{A} = \{0, 1\}$, $(X_i, A_i, (Y_i(a))_{a \in \mathcal{A}})$ are i.i.d.
- Simplifies mathematical calculations.
- Usually gives good approximations to sampling without replacement.
- Notation: (A, X, Y) for a generic random variable.

See blackboard:

- Positivity/overlap assumption.
- Causal identification in randomized experiments.
- OR and IPW estimators of the average treatment effect (ATE).

M-estimation

- Suppose $D_i \overset{i.i.d.}{\sim} \mathbb{P}$, $i = 1, \ldots, n$.
- An m-estimator of $\theta = \theta(\mathbb{P})$ can be obtained by minimizing a loss function ℓ:

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, D_i).$$

- Example: maxmimum likelihood estimator.
- Denote $\theta_0 = \arg\min_{\theta} \mathbb{E}[\ell(\theta; D)]$ and $\psi(\theta, D) = \frac{\partial}{\partial \theta} \ell(\theta, D)$.

See blackboard:

- Taylor expansion for $\frac{1}{n} \sum_{i=1}^{n} \psi(\hat{\theta}, D_i) = 0$ at θ_0.
- Asymptotic distribution of $\hat{\theta}$.
- Special case: linear regression with heteroscedastic noise.

Regression adjustment

- Further assume $A \perp\!\!\!\perp X$ (simple Bernoulli trial) and $\mathbb{E}[X] = 0$.
- Three OR estimators of the ATE:

$$(\hat{\alpha}_1, \hat{\beta}_1) = \arg\min_{\alpha,\beta} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \alpha - \beta A_i)^2,$$

$$(\hat{\alpha}_2, \hat{\beta}_2, \hat{\gamma}_2) = \arg\min_{\alpha,\beta,\gamma} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \alpha - \beta A_i - \gamma X_i)^2,$$

$$(\hat{\alpha}_3, \hat{\beta}_3, \hat{\gamma}_3, \hat{\delta}_3) = \arg\min_{\alpha,\beta,\gamma,\delta} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \alpha - \beta A_i - \gamma X_i - \delta A_i X_i)^2.$$

See blackboard:

- Asymptotic distribution of $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$: all unbiased, $\hat{\beta}_3$ is the most efficient.
- What happens if $\mathbb{E}[X] \neq 0$?

Randomization experiments: Different paradigms

	Randomization test	Neyman's method	Regression
Population	Finite	Finite	Super-population
Randomness	Only A	Only A	$A, \boldsymbol{X}, Y$
Point estimator	Hodges-Lehmann (example sheet)	Difference-in-means	Least squares
Inference	Exact	Exact variance formula, conservative & asymptotic inference	Asymptotic
Covariate adjustment	Possible	Possible	Yes

Outline

Introduction to linear SEMs

- Linear structural equation models (SEMs) were first developed by Sewall Wright around 1920 to study the heredity of guinea pigs.[3]

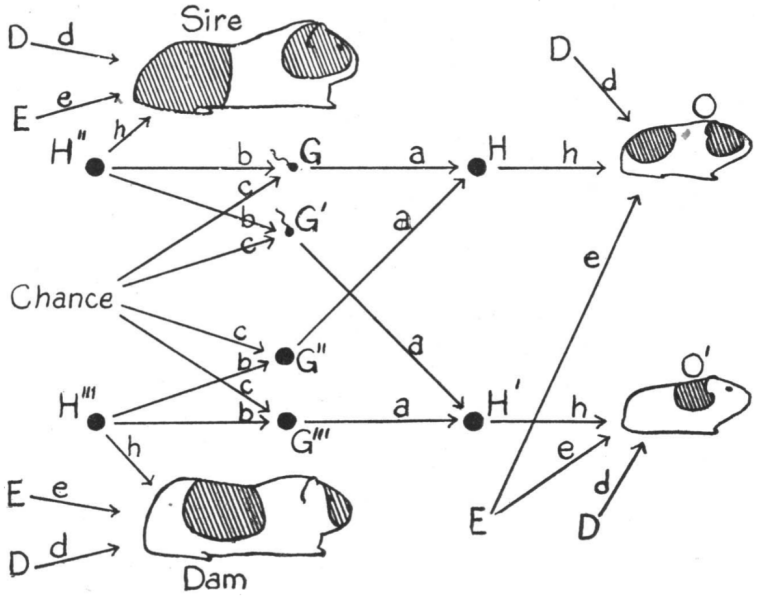


FIG. 5.

Diagram illustrating the casual relations between litter mates (O, O′) and between each of them and their parents. H, H′, H″, H,‴ represent the genetic constitutions of the four individuals, G, G′, G″, and G‴ that of four germ cells. E represents such environmental factors as are common to litter mates. D represents other factors, largely ontogenetic irregularity. The small letters stand for the various path coefficients.

- It is a precursor of the general theory of (causal) graphical models developed since 1980/90s and remains widely used in practice.

[3] Interestingly, Fisher's work on randomized experiments are also very much related to his work on genetics; see Tudball, Davey Smith, and Zhao, "Almost exact Mendelian randomization", arXiv:2208.14035.

Graph basics

We will be primarily interested in using acyclic directed mixed graphs (ADMGs).

- $\mathcal{G} = (\mathcal{V}, \mathcal{D}, \mathcal{B})$: finite vertex set $\mathcal{V}$, directed edges $\mathcal{D} \subseteq \mathcal{V} \times \mathcal{V}$, and bidirected edges $\mathcal{B} \subseteq \mathcal{V} \times \mathcal{V}$.
- A graphical model means a bijection from $\mathcal{V}$ to a collection of random variables $\{V_1, \ldots, V_p\}$, usually either $j \mapsto V_j$ or $V_j \mapsto V_j$ (we will use both).
- Write $(V_j, V_k) \in \mathcal{D}$ as $V_j \to V_k$, representing *direct causal effect*.
- Write $(V_j, V_k) \in \mathcal{B}$ as $V_j \leftrightarrow V_k$, representing *hidden common causes*.
- We say $\mathcal{G}$ is *acyclic* if it contains no directed cycle like $V_1 \to \cdots \to V_l \to V_1$.
- A *walk* on $\mathcal{G}$ is a sequence of adjacent edges of any type or orientation.
- A *path* is a walk without repeated vertices.

Graph basics (cont.)

Collider

- A key concept in probabilistic graphical models is *collider*—colliding arrowheads on a walk like $\rightarrow V_j \leftarrow$, $\rightarrow V_j \leftrightarrow$, $\leftrightarrow V_j \leftarrow$ or $\leftrightarrow V_j \leftrightarrow$.
- A walk without colliders will be called an *arc* and denoted by a squiggly line ($\rightsquigarrow$), with matching end-point arrowheads.

Familial terminology

Walk notation	V_j is a ... of V_k	Familial notation
$V_j \rightarrow V_k$	parent	$V_j \in \mathrm{pa}(V_k)$
$V_j \leftarrow V_k$	child	$V_j \in \mathrm{ch}(V_k)$
$V_j \rightsquigarrow V_k$	ancestor	$V_j \in \mathrm{an}(V_k)$
$V_j \leftsquigarrow V_k$	descendant	$V_j \in \mathrm{de}(V_k)$

- We will often consider *directed acyclic graphs* (DAGs), which are ADMGs with no bidirected edges.

Linear structural equation models

We say a random vector $\boldsymbol{V}$ satisfies a linear *structural equation model* (SEM) with respect to an ADMG $\mathcal{G} = (\mathcal{V} = [p], \mathcal{D}, \mathcal{B})$ and parameters $(\boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Lambda})$ if $\boldsymbol{V}$ solves

$$\boldsymbol{V} = \boldsymbol{\alpha} + \boldsymbol{B}^T \boldsymbol{V} + \boldsymbol{E}, \tag{1}$$

- $\boldsymbol{\alpha} \in \mathbb{R}^{|\mathcal{V}|}$ is the intercept;
- $\boldsymbol{B} = (\beta_{jk} : j, k \in \mathcal{V})$ is a weighted adjancency matrix of $(\mathcal{V}, \mathcal{D})$;
- $\boldsymbol{E}$ is a random vector with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{\Lambda} = (\lambda_{jk} : j, k \in \mathcal{V})$ that is positive semidefinite and a weighted adjancency matrix of $(\mathcal{V}, \mathcal{B})$

Structural?

- The above definition is incomplete.
- "Structural" (aka "causal") means that these equations "still hold" if we set $\boldsymbol{V}_{\mathcal{J}}$ to $\boldsymbol{v}_{\mathcal{J}}$ in an intervention/experiment.

See blackboard:

- Example.
- Definition of total causal effect in linear SEMs.

Path analysis

- In a linear SEM, any edge h is associated with a coefficient:

$$\sigma(h) = \begin{cases} \beta_{jj'}, & \text{if } h = j \to j', \\ \lambda_{jj'}, & \text{if } h = j \leftrightarrow j'. \end{cases}$$

- For any walk $\pi = h_1 \cdot \cdots \cdot h_q$, define

$$\sigma(\pi) = \begin{cases} 1, & \text{if } q = 0 \text{ (i.e. } \pi \text{ is empty)}, \\ \prod_{m=1}^{q} \sigma(h_m), & \text{if } q \geq 1. \end{cases}$$

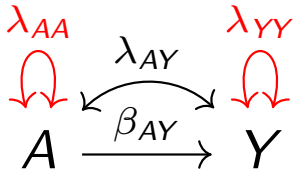
- Convention: $\cdot$ for concatenation; $+$ for disjoint union (of sets of walks).

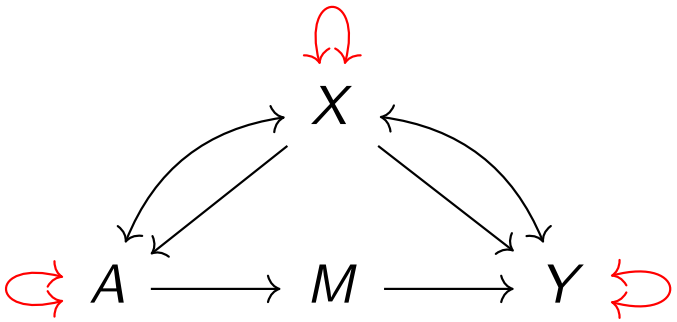
See blackboard:

- Trek rule: $\mathrm{Cov}(\boldsymbol{V}) = \sigma(\boldsymbol{W}(\mathcal{V} \overset{t}{\leftsquigarrow} \mathcal{V}))$;
- (Unconditional) m-connected walks/paths.
- Wright's path analysis: if $\boldsymbol{V}$ is standardized so that $\mathrm{var}(V_j) = 1$ for all j, then $\mathrm{cov}(V_j, V_k) = \sigma(\mathcal{P}(j \rightsquigarrow k))$ for all j, k.

Two examples

Find $\mathrm{Var}(A)$, $\mathrm{Var}(Y)$, $\mathrm{Cov}(A, Y)$, and the causal effect of A on Y.





Correlation versus causation

For the rest of this course, we will maintain the following assumptions:

- The causal diagram $\mathcal{G}$ is acyclic;
- All vertices of $\mathcal{G}$ has bidirected self-loops (which will be omitted).

Suppose further that $\boldsymbol{V}$ is standardized: $\mathrm{Var}(V_j = 1)$ for all j. Then

$$\text{Causal effect of } V_j \text{ on } V_k = \sigma(\mathcal{P}(j \rightsquigarrow k)),$$
$$\mathrm{Cov}(V_j, V_k) = \sigma(\mathcal{P}(j \leftsquigarrow k)).$$

See blackboard: Discussion on three examples:

- Confounder: $A \leftarrow X \rightarrow Y$;
- Mediator: $A \rightarrow M \rightarrow Y$;
- Collider: $A \rightarrow C \leftarrow Y$.

Partial correlation

- There's much interest in conditional independence, a notion of irrelevance.
- In linear models, this is usually measured by partial correlation.
- Consider the partition $\boldsymbol{V} = (V_1, V_2, \boldsymbol{V}_3)$. Let

$$\mathrm{Cov}(\boldsymbol{V}) = \boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \boldsymbol{\Sigma}_{13} \\ \Sigma_{21} & \Sigma_{22} & \boldsymbol{\Sigma}_{23} \\ \boldsymbol{\Sigma}_{31} & \boldsymbol{\Sigma}_{32} & \boldsymbol{\Sigma}_{33} \end{pmatrix}, \quad \boldsymbol{\Sigma}^{-1} = \boldsymbol{\Omega} = \begin{pmatrix} \Omega_{11} & \Omega_{12} & \boldsymbol{\Omega}_{13} \\ \Omega_{21} & \Omega_{22} & \boldsymbol{\Omega}_{23} \\ \boldsymbol{\Omega}_{31} & \boldsymbol{\Omega}_{32} & \boldsymbol{\Omega}_{33} \end{pmatrix}.$$

- The *partial correlation* of V_1 and V_2 given $\boldsymbol{V}_3$ is defined as

$$\begin{aligned} \mathrm{Cor}(V_1, V_2 \mid \boldsymbol{V}_3) &= \mathrm{Cor}(V_1 - \boldsymbol{\Sigma}_{13}\boldsymbol{\Sigma}_{33}^{-1}\boldsymbol{V}_3, V_2 - \boldsymbol{\Sigma}_{23}\boldsymbol{\Sigma}_{33}^{-1}\boldsymbol{V}_3) \\ &= -\frac{\Omega_{12}}{\sqrt{\Omega_{11}\Omega_{22}}}. \end{aligned}$$

- When $\boldsymbol{V}$ is multivariate normal, this suggests that $V_1 \perp\!\!\!\perp V_2 \mid \boldsymbol{V}_3$ iff $\Omega_{12} = 0$.

Roadmap

Suppose $\boldsymbol{V}$ follows a linear SEM w.r.t. an ADMG $\mathcal{G}$, and $\{j\}, \{k\}, \mathcal{L}$ are disjoint subsets of $\mathcal{V} = [p]$.

Central question

When can we use the graph $\mathcal{G}$ to conclude that $\mathrm{Cor}(V_j, V_k \mid \boldsymbol{V}_{\mathcal{L}}) = 0$?

- Let $\tilde{\mathcal{V}} = \{j, k, \mathcal{L}\}$, and $\mathcal{U} = \mathcal{V} \setminus \tilde{\mathcal{V}}$.

Solution

1. Understand the "projection graph" $\tilde{\mathcal{G}}$ that describes $\tilde{\boldsymbol{V}} = (V_j, V_k, \boldsymbol{V}_{\mathcal{L}})$.
2. Understand how $\tilde{\boldsymbol{\Omega}} = (\mathrm{Cov}(\tilde{\boldsymbol{V}}))^{-1}$ relates to $\tilde{\mathcal{G}}$.

Marginalization and latent projection

- Let Σ be the covariance matrix of $\boldsymbol{V} = (\tilde{\boldsymbol{V}}, \boldsymbol{U})$. Then

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

- Useful formula for block matrix inversion:

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{E} & \boldsymbol{F} \\ \boldsymbol{G} & \boldsymbol{H} \end{pmatrix}, \ \boldsymbol{A}^{-1} = \begin{pmatrix} \boldsymbol{S}^{-1} & -\boldsymbol{S}^{-1}\boldsymbol{F}\boldsymbol{H}^{-1} \\ -\boldsymbol{H}^{-1}\boldsymbol{G}\boldsymbol{S}^{-1} & \boldsymbol{H}^{-1} + \boldsymbol{H}^{-1}\boldsymbol{G}\boldsymbol{S}^{-1}\boldsymbol{F}\boldsymbol{H}^{-1} \end{pmatrix},$$

  where $\boldsymbol{S} = \boldsymbol{E} - \boldsymbol{F}\boldsymbol{H}^{-1}\boldsymbol{G}$ is the Schur complement.

See blackboard:

- Block matrix inversion for $\boldsymbol{I} - \boldsymbol{B}$.
- Expressing $\tilde{\Sigma}$ using treks in the *latent projection* graph $\tilde{\mathcal{G}}$ with vertex set $\tilde{\mathcal{V}}$ and the following edges:

$$a \left\{ \begin{matrix} \to \\ \leftarrow \\ \leftrightarrow \end{matrix} \right\} b \ [\tilde{\mathcal{G}}] \quad \Longleftrightarrow \quad a \left\{ \begin{matrix} \text{via } \mathcal{U} \\ \rightsquigarrow \\ \text{via } \mathcal{U} \\ \leftsquigarrow \\ \text{via } \mathcal{U} \\ \leftrightsquigarrow \end{matrix} \right\} b \ [\mathcal{G}], \quad \text{for all } a, b \in \tilde{\mathcal{V}}.$$

- Corollary: Latent projection preserves $\rightsquigarrow$ and $\leftrightsquigarrow$.
- Example.

A graphical criterion for conditional independence

- Now consider the precision matrix $\tilde{\boldsymbol{\Omega}} = \tilde{\boldsymbol{\Sigma}}^{-1} = (\boldsymbol{I} - \tilde{\boldsymbol{B}})\tilde{\boldsymbol{\Lambda}}^{-1}(\boldsymbol{I} - \tilde{\boldsymbol{B}})^T$.
- Notation: $*$ is a wildcard character meaning any number of colliders.

See blackboard:

- Definitions: m*-connected (walk) and confounding.
- A simple graphical criterion for *unconfoundedness*:

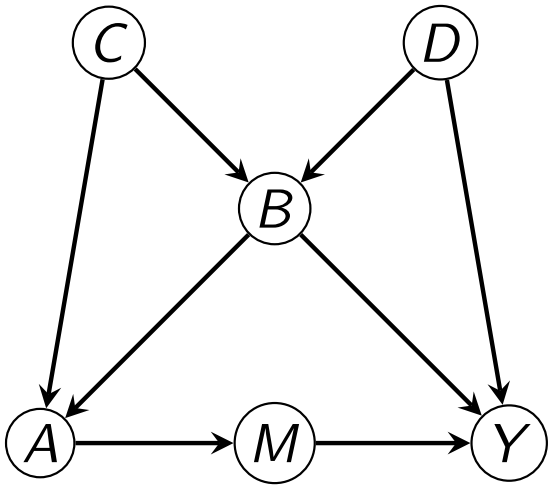
$$j \leftrightsquigarrow * \not\leftrightsquigarrow k \mid \mathcal{L} \ [\mathcal{G}] \ \Leftrightarrow \ j \leftrightarrow * \not\leftrightarrow k \ [\tilde{\mathcal{G}}] \ \Rightarrow \ (\tilde{\boldsymbol{\Lambda}}^{-1})_{jk} = 0.$$

- *m-separation* as a graphical criterion for conditional independence:

$$j \rightsquigarrow * \not\leftsquigarrow k \mid \mathcal{L} \ [\mathcal{G}] \ \Leftrightarrow \ j \longrightarrow * \not\longleftarrow k \ [\tilde{\mathcal{G}}]$$
$$\Downarrow$$
$$(\tilde{\boldsymbol{\Omega}})_{jk} = 0 \ \Leftrightarrow \ \mathrm{Cor}(V_j, V_k \mid \boldsymbol{V}_{\mathcal{L}}) = 0.$$

- Comments on completeness.
- Definition: m-connected (path).

Examples



Is the following true?

1. $A \not\rightsquigarrow Y$.
2. $A \rightsquigarrow \not* \rightsquigarrow Y \mid B$.
3. $A \rightsquigarrow \not* \rightsquigarrow Y \mid B, M$.

Another example: Open `https://www.dagitty.net/dags.html`, then load Shrier & Platt, 2008 under the Examples menu.

Identifiability problems in linear SEMs

Given a graph $\mathcal{G}$, the trek rule/Wright's path analysis defines a map

$$(\boldsymbol{B}, \boldsymbol{\Lambda}) \mapsto \boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{V}).$$

The general identifiability problem

Under what graphical conditions is this map *invertible*?

- There are different notions of invertibility.

See blackboard:

- Instrumental variable graph and generic identifiability.
- Factor analysis and measurement models in psychometrics (example paper).

Review: the squiggly line notation

This new notation is introduced this year to visualize graphical arguments.

- A walk without colliders is called an arc and denoted by a squiggly line ($\rightsquigarrow$), with matching end-point arrowheads.
- There are three kinds of arcs: $\rightsquigarrow$, $\leftsquigarrow$, $\leftrightsquigarrow$ (bidirected arc or confounding arc) (Q: why isn't there a arc like $\rightsquigarrow$?).
- Half arrowhead means unrestricted. For example, the set of all arcs from j to k is denoted as

$$\mathcal{W}(j \leftrightsquigarrow k) = \mathcal{W}(j \rightsquigarrow k) + \mathcal{W}(j \leftsquigarrow k) + \mathcal{W}(j \leftrightsquigarrow k).$$

- $*$ means any number of colliders, so $\mathcal{W}(j \leftrightsquigarrow * \leftrightsquigarrow k)$ contains all walks from j to k. Similarly, $\mathcal{W}(j \leftrightsquigarrow * \leftrightsquigarrow k)$ contains all walks from j to k consisting of any number of bidirected arcs.
- $\mathcal{W}(j \leftrightsquigarrow * \leftrightsquigarrow k \mid \mathcal{L})$ contains all walks from j to k that are m*-connected given $\mathcal{L}$. Similarly for $\mathcal{W}(j \leftrightsquigarrow * \leftrightsquigarrow k \mid \mathcal{L})$.
- When these sets are non-empty, we say the corresponding type of connection is true. For example, $\mathcal{W}(j \leftrightsquigarrow * \leftrightsquigarrow k \mid \mathcal{L}) \neq \emptyset$ means j and k are m-connected given $\mathcal{L}$.

A review and look forward

Summary of this chapter

- Basic terminology and notation for directed mixed graphs.
- Definition of linear SEMs.
- Trek rule and Wright's path analysis.
- Marginalization and latent projection.
- Partial correlation/conditional independence, m-separation.
- Identifiability problems.

Beyond linearity

We will consider different "structures" between random variables that can be modelled by graphs:

1. Factorization (probabilistic);
2. Conditional independences/Markov properties (probabilistic);
3. Causality.

Outline

Factorization

f is joint density function of $\boldsymbol{V} = (V_1, \ldots, V_p)$ (w.r.t. a product measure).

DAG model (Bayesian statistics, causality, ...)

We say the distribution of $\boldsymbol{V}$ *factorizes* according to a DAG $\mathcal{G} = (\mathcal{V} = [p], \mathcal{D})$ if

$$f(\boldsymbol{v}) = \prod_{j=1}^{p} f_{j|\mathrm{pa}(j)}(v_j \mid v_{\mathrm{pa}(j)}).$$

Undirected graphical model (statistical physics, contigency tables, ...)

We say the distribution of $\boldsymbol{V}$ *factorizes* according to an undirected graph $\mathcal{G} = (\mathcal{V} = [p], \mathcal{E})$ if there exists $\{\psi_{\mathcal{C}}(\cdot) : \mathcal{C} \subseteq \mathcal{V}\}$ such that

$$f(\boldsymbol{v}) = \prod_{\mathcal{C} \subseteq \mathcal{V}} \psi_{\mathcal{C}}(\boldsymbol{v}_{\mathcal{C}}),$$

where the product is over all *cliques* (complete subgraphs) of $\mathcal{G}$.

See blackborad: Examples.

Conditional independence

- Conditioning is a key concept in probability theory.
- Let $f_{\mathcal{J}|\mathcal{K}}$ denote the conditional density function of $\boldsymbol{V}_{\mathcal{J}}$ given $\boldsymbol{V}_{\mathcal{K}}$:

$$f_{\mathcal{J}|\mathcal{K}}(\boldsymbol{v}_{\mathcal{J}} \mid \boldsymbol{v}_{\mathcal{K}}) = \frac{f_{\mathcal{J}\cup\mathcal{K}}(\boldsymbol{v}_{\mathcal{J}}, \boldsymbol{v}_{\mathcal{K}})}{f_{\mathcal{K}}(\boldsymbol{v}_{\mathcal{K}})} \text{ when } f_{\mathcal{K}}(\boldsymbol{v}_{\mathcal{K}}) > 0.$$

- The subscript of f will often be omitted.

Definition

For disjoint $\mathcal{J}, \mathcal{K}, \mathcal{L} \subseteq \mathcal{V}$, we say $\boldsymbol{V}_{\mathcal{J}}$ is *conditionally independent* of $\boldsymbol{V}_{\mathcal{K}}$ given $\boldsymbol{V}_{\mathcal{L}}$ and write $\boldsymbol{V}_{\mathcal{J}} \perp\!\!\!\perp \boldsymbol{V}_{\mathcal{K}} \mid \boldsymbol{V}_{\mathcal{L}}$ if

$$f(\boldsymbol{v}_{\mathcal{J}}, \boldsymbol{v}_{\mathcal{K}} \mid \boldsymbol{v}_{\mathcal{L}}) = f(\boldsymbol{v}_{\mathcal{J}} \mid \boldsymbol{v}_{\mathcal{L}}) f(\boldsymbol{v}_{\mathcal{K}} \mid \boldsymbol{v}_{\mathcal{L}}) \text{ whenever } f(\boldsymbol{v}_{\mathcal{L}}) > 0.$$

- If $\mathcal{L} = \emptyset$, we just write $\boldsymbol{V}_{\mathcal{J}} \perp\!\!\!\perp \boldsymbol{V}_{\mathcal{K}}$.

Graphoid axioms

As a notion of irrelevance, conditional independence satisfies several axiomatic properties: for all disjoint $\mathcal{J}, \mathcal{K}, \mathcal{L}, \mathcal{M} \subset \mathcal{V}$,

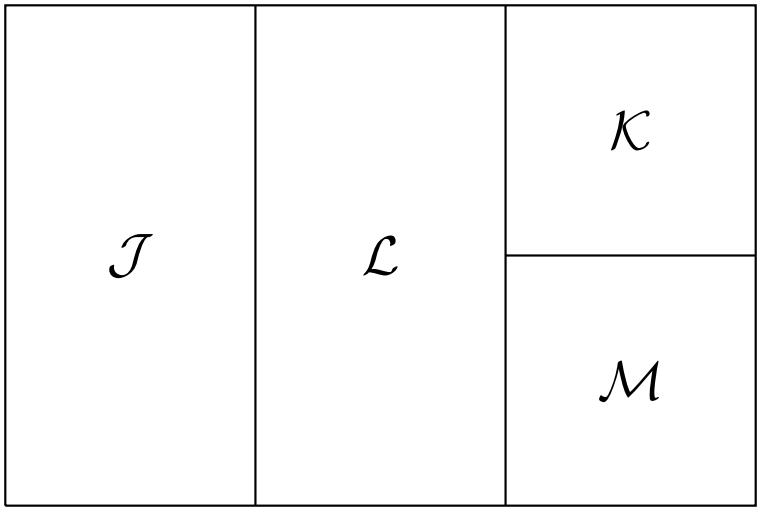
Symmetry $\boldsymbol{V}_{\mathcal{J}} \perp\!\!\!\perp \boldsymbol{V}_{\mathcal{K}} \mid \boldsymbol{V}_{\mathcal{L}} \Leftrightarrow \boldsymbol{V}_{\mathcal{K}} \perp\!\!\!\perp \boldsymbol{V}_{\mathcal{J}} \mid \boldsymbol{V}_{\mathcal{L}}$;

Chain rule $\boldsymbol{V}_{\mathcal{J}} \perp\!\!\!\perp \boldsymbol{V}_{\mathcal{K}} \mid \boldsymbol{V}_{\mathcal{L} \cup \mathcal{M}}$ and $\boldsymbol{V}_{\mathcal{J}} \perp\!\!\!\perp \boldsymbol{V}_{\mathcal{M}} \mid \boldsymbol{V}_{\mathcal{L}} \Leftrightarrow \boldsymbol{V}_{\mathcal{J}} \perp\!\!\!\perp \boldsymbol{V}_{\mathcal{K} \cup \mathcal{M}} \mid \boldsymbol{V}_{\mathcal{L}}$.

If $f(\boldsymbol{v}) > 0$ for all $\boldsymbol{v}$, we further have

Intersection $\boldsymbol{V}_{\mathcal{J}} \perp\!\!\!\perp \boldsymbol{V}_{\mathcal{K}} \mid \boldsymbol{V}_{\mathcal{L} \cup \mathcal{M}}$ and $\boldsymbol{V}_{\mathcal{J}} \perp\!\!\!\perp \boldsymbol{V}_{\mathcal{M}} \mid \boldsymbol{V}_{\mathcal{L} \cup \mathcal{K}} \Rightarrow \boldsymbol{V}_{\mathcal{J}} \perp\!\!\!\perp \boldsymbol{V}_{\mathcal{K} \cup \mathcal{M}} \mid \boldsymbol{V}_{\mathcal{L}}$

- Ternary relations that satisfy these axioms are called *graphoids*.



Global Markov property in undirected graphical models

Fix an undirected graph $\mathcal{G} = (\mathcal{D}, \mathcal{E})$.

Definition

For disjoint $\mathcal{J}, \mathcal{K}, \mathcal{L} \subset \mathcal{V}$, we say $\mathcal{J}$ and $\mathcal{K}$ are *separated* by $\mathcal{L}$ and write $\mathcal{J} - \not{*} - \mathcal{K} \mid \mathcal{L}$ if every path from $j \in \mathcal{J}$ to $k \in \mathcal{K}$ in $\mathcal{G}$ goes through a node in $\mathcal{L}$.

Theorem (Hammersley-Clifford)

Suppose $f(\boldsymbol{v}) > 0$ for all $\boldsymbol{v}$. Then the distribution $\boldsymbol{V}$ factorizes according to $\mathcal{G}$ iff

$$\mathcal{J} - \not{*} - \mathcal{K} \mid \mathcal{L} \Longrightarrow \boldsymbol{V}_{\mathcal{J}} \perp\!\!\!\perp \boldsymbol{V}_{\mathcal{K}} \mid \boldsymbol{V}_{\mathcal{L}}, \ \forall\, \mathcal{J}, \mathcal{K}, \mathcal{L}\,. \qquad \text{(Global Markov)}$$

- Statistical physics: Gibbs random field $\Leftrightarrow$ Markov random field.

Global Markov property in DAG models

Now fix an DAG $\mathcal{G} = (\mathcal{V}, \mathcal{D})$.

Definition

For disjoint $\mathcal{J}, \mathcal{K}, \mathcal{L} \subset \mathcal{V}$, we say $\mathcal{J}$ and $\mathcal{K}$ are *d-separated*[4] by $\mathcal{L}$ and write $\mathcal{J} \rightsquigarrow \not\ast \not\leftsquigarrow \mathcal{K} \mid \mathcal{L}$ if there exists no walk π from $j \in \mathcal{J}$ to $k \in \mathcal{K}$ in $\mathcal{G}$ such that

- All non-colliders on π are not in $\mathcal{L}$;
- All colliders on π are in $\mathcal{L}$.

Theorem

The distribution of $\boldsymbol{V}$ *factorizes according to* $\mathcal{G}$ *iff*

$$\mathcal{J} \rightsquigarrow \not\ast \not\leftsquigarrow \mathcal{K} \mid \mathcal{L} \Longrightarrow \boldsymbol{V}_{\mathcal{J}} \perp\!\!\!\perp \boldsymbol{V}_{\mathcal{K}} \mid \boldsymbol{V}_{\mathcal{L}}, \ \forall\, \mathcal{J}, \mathcal{K}, \mathcal{L}\,. \qquad \text{(Global Markov)}$$

See blackboard:

- Proof of this theorem by induction.

[4] d/m-separation: d means "directed" and m means "mixed"

Structure learning

Goal: infer the graphical structure from data.

Gaussian graphical models

- For undirected graphs and Gaussian data, it suffices to estimate the inverse covariance matrix.
- In particular, this is a model selection problem. A common algorithm called the "graphical lasso" is covered in Modern Statistical Methods.

DAG models: Faithfulness

- We say a distribution of $\boldsymbol{V}$ is *faithful* to a DAG $\mathcal{G}$ if the conditional independences are exactly those implied by the global Markov property.
- Question: Given faithfulness, can we learn the DAG $\mathcal{G}$ from a sample from the distribution of $\boldsymbol{V}$?

Markov equivalence of DAG models

Observation: we clearly cannot distinguish $V_1 \to V_2$ from $V_1 \leftarrow V_2$ based on just a sample of (V_1, V_2).

- Two DAGs are said to be *Markov equivalent* if they imply the same d-separations.
- We can only hope to recover a *Markov equivalence class*, i.e. a maximal set of Markov equivalent DAGs.

Theorem

Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent iff the following conditions are true:

1. *$\mathcal{G}_1$ and $\mathcal{G}_2$ have the same "skeleton" (set of edges ignoring their directions);*
2. *$\mathcal{G}_1$ and $\mathcal{G}_2$ have the same set of "unshielded colliders" ($j \to \ell \leftarrow k$ but j and k are not adjacent).*

See blackboard: Example.

Algorithms based on conditional independence testing

IC/SGS algorithm

Step 0 Start with a fully connected undirected graph.

Step 1 Remove $j - k$ if $V_j \perp\!\!\!\perp V_k \mid V_{\mathcal{L}}$ for some $\mathcal{L} \subset \mathcal{V}$.

Step 2 Orient $j - \ell - k$ as $j \to \ell \leftarrow k$ if j and k are not adjacent and $V_j \not\perp\!\!\!\perp V_k \mid V_{\mathcal{L}}$ for all subsets $\mathcal{L}$ containing ℓ.

Step 3 Orient some of the other edges so that the graph contains no cycles or new unshielded colliders.

- The more widely used PC algorithm accelerates Step 1 by gradually increasing the size of $\mathcal{L}$, because if j and k are d-separated by $\mathcal{L}$, they are also d-separated by any superset of $\mathcal{L}$.
- Some difficulties: conditional independence testing; latent variables; causal interpretation.

Outline

Nonparametric structural equation model (NPSEM)

Fix an ADMG $\mathcal{G} = (\mathcal{V} = [p], \mathcal{D}, \mathcal{B})$. We may generalize linear SEMs by using a general nonlinear relationship for each equation:

$$V_j = f_j(\boldsymbol{V}_{\mathrm{pa}(j)}, E_j),\ j \in \mathcal{V},$$

where $(E_1, \ldots, E_p)$ "obeys" the bidirected graph $(\mathcal{V}, \mathcal{B})$.

Structural?

- Recall the remark on slide 28: what makes a system of equations "structural" or "causal"?
- To formally answer this question, we need to consider potential outcomes of the system under different interventions.

See blackboard:

- Definition via recursive substitution.
- "Natural counterfactual" of the variables being intervened on.
- Simplification of potential outcomes and the consistency property.
- Examples.

Markov properties of basic potential outcomes

Two views

- Functional/Multiple-world: the random errors $(E_1, \ldots, E_p)$ satisfy the global Markov property w.r.t. $(\mathcal{V}, \mathcal{B})$.
- Counterfactual/Single-world: the basic p. o. $(V_1(\boldsymbol{v}_{[p]}), \ldots, V_p(\boldsymbol{v}_{[p]}))$ satisfy the global Markov property w.r.t. $(\mathcal{V}, \mathcal{B})$ for all $\boldsymbol{v}_{[p]}$.

Definition

We say $\boldsymbol{V}$ and all its counterfactuals satisfy the (single/multiple-world) causal model w.r.t. an ADMG $\mathcal{G} = (\mathcal{V}, \mathcal{D}, \mathcal{E})$ if

- All the potential outcomes satisfy the consistency property w.r.t. $(\mathcal{V}, \mathcal{D})$;
- The basic potential outcomes satisfy the (single/multiple-world) Markov property w.r.t. $(\mathcal{V}, \mathcal{B})$.

See blackboard:

- An example.
- Defining a NPSEM through basic potential outcomes.

Representing recursive substitution as a graph operation

- Suppose $\boldsymbol{V}_{\tilde{\mathcal{I}}}$ has already been intervened on.
- Key idea: We may view an additional intervention on V_i as splitting it into two: a natural counterfactual $V_i(\boldsymbol{v}_{\tilde{\mathcal{I}}})$ and a fixed value v_i.

Definition

Given an ADMG $\mathcal{G} = (\mathcal{V}, \mathcal{B}, \mathcal{D})$, the *single-world intervention graph* (SWIG) $\mathcal{G}(\mathcal{I})$ is an ADMG with vertex set $\boldsymbol{V}(\boldsymbol{v}_{\mathcal{I}}) \cup \boldsymbol{v}_{\mathcal{I}}$, and

- For $j \notin \mathcal{I}$, $V_j(\boldsymbol{v}_{\mathcal{I}})$ inherits all edges of V_j in $\mathcal{G}$.
- For $i \in \mathcal{I}$, $V_i(\boldsymbol{v}_{\mathcal{I}})$ inherits all incoming edges and v_i inherits all outgoing edges of V_i in $\mathcal{G}$.

See blackboard:

- Relationship with recursive substitution.
- Example.

Recursive substitution preserves global Markov properties

Theorem

Suppose $\boldsymbol{V}$ *and its counterfactuals satisfy the single-world causal model w.r.t. an ADMG* $\mathcal{G} = (\mathcal{V}, \mathcal{D}, \mathcal{B})$*. Then for any* $\mathcal{I} \subseteq \mathcal{V}$*,* $\boldsymbol{V}(\boldsymbol{v}_{\mathcal{I}})$ *satisfies the global Markov property w.r.t.* $\mathcal{G}(\mathcal{I})$*.*

- Any fixed vertex v_i, $i \in \mathcal{I}$ in $\mathcal{G}(\mathcal{I})$ does not introduce (statistical) dependence.
- Thus, it is convenient to consider the graph $\mathcal{G}^*(\mathcal{I})$ without $\boldsymbol{v}_{\mathcal{I}}$.

See blackboard: proof using the following observations (let $\mathcal{I}' = \mathcal{I} \cup \{j\}$).

Lemma 1 For any $\mathcal{I} \subset \mathcal{V}$, one can always find $j \notin \mathcal{J}$ such that $\mathrm{de}(j) \subseteq \mathcal{I}$.

Lemma 2 If $k \notin \mathrm{ch}(j)$, then $V_k(\boldsymbol{v}_{\mathcal{I}}) = V_k(\boldsymbol{v}_{\mathcal{I}'})$.

Lemma 3 Compared to $\mathcal{G}^*(\mathcal{I}')$, $\mathcal{G}^*(\mathcal{I})$ has the new edges $V_j(\boldsymbol{v}_{\mathcal{I}}) \to \boldsymbol{V}_{\mathrm{ch}(j)}(\boldsymbol{v}_{\mathcal{I}})$.

Lemma 4 Any m-separation in $\mathcal{G}^*(\mathcal{I})$ also holds in $\mathcal{G}^*(\mathcal{I}')$.

DAG causal models

Now suppose $\mathcal{G}$ is a DAG.

- An immediate corollary is that $\boldsymbol{V}(\boldsymbol{v}_{\mathcal{I}})$ should factorize w.r.t. $\mathcal{G}(\mathcal{I})$.
- But we can say more. Identification results below are stated with discrete $\boldsymbol{V}$.

Theorem (g-computation formula)

Suppose $\boldsymbol{V}$ and its counterfactuals satisfy the single-world causal model w.r.t. a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{D}, \mathcal{B})$. Then for any $\mathcal{I} \subseteq \mathcal{V}$, we have

$$\mathbb{P}(\boldsymbol{V}(\boldsymbol{v}_{\mathcal{I}}) = \tilde{\boldsymbol{v}}) = \prod_{j=1}^{p} \mathbb{P}(V_j = \tilde{v}_j \mid \boldsymbol{V}_{\mathrm{pa}(j)\cap\mathcal{I}} = \boldsymbol{v}_{\mathrm{pa}(j)\cap\mathcal{I}}, \boldsymbol{V}_{\mathrm{pa}(j)\setminus\mathcal{I}} = \tilde{\boldsymbol{v}}_{\mathrm{pa}(j)\setminus\mathcal{I}}).$$

- This matches our intuition of *modularity* of causal models.

See blackborad: example.

Back-door criterion

Next: queries about causal identifiability in ADMG models. Let $\mathcal{G} = (\mathcal{V}, \mathcal{D}, \mathcal{B})$.

Theorem

Let $\boldsymbol{X}, \{A\}, \{Y\} \subset \mathcal{V}$ be disjoint. Suppose

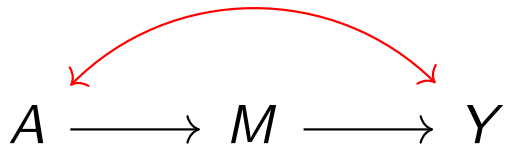
1. *$\boldsymbol{X}$ contains no descendant of A;*
2. *$A \overset{}{\leftsquigarrow\!\!\!\!\!\!\not\;\;\leftsquigarrow} Y \mid \boldsymbol{X}$ (no m-connected "backdoor" paths).*

Then $Y(a) \perp\!\!\!\perp A \mid \boldsymbol{X}$. Under the positivity assumption $\mathbb{P}(A = a \mid \boldsymbol{X} = \boldsymbol{x}) > 0$ for all a and $\boldsymbol{x}$, we have

$$\mathbb{P}(Y(a) = y) = \sum_{\boldsymbol{x}} \mathbb{P}(Y = y \mid A = a, \boldsymbol{X} = \boldsymbol{x}) \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}), \text{ for all } y.$$

See blackboard: examples and proof.

Front-door criterion



Proposition

For the above causal graph, the causal effect of A on Y is identified by

$$\begin{aligned}&\mathbb{P}(Y(a)=y)\\&=\sum_m\left\{\sum_{a'}\mathbb{P}(Y=y\mid M=m,A=a')\mathbb{P}(A=a')\right\}\mathbb{P}(M=m\mid A=a).\end{aligned}$$

See blackboard: proof and interpretation.

The fixing operator

For more general queries about causal identification, we need to answer the following question: Does the g-formula hold when $\mathcal{G}$ is not a DAG but an ADMG?

Definition

- For an ADMG $\mathcal{G}$, a vertex i is called *fixable* if there exists no j such that $i \rightsquigarrow j$ and $i \leftrightarrow * \leftrightarrow j$.
- The *Markov blanket* of $\mathcal{I} \subseteq \mathcal{V}$ is defined as $\mathrm{mb}(\mathcal{I}) = \{j \notin \mathcal{I} : j \hookrightarrow * \hookleftarrow \mathcal{I}\}$.

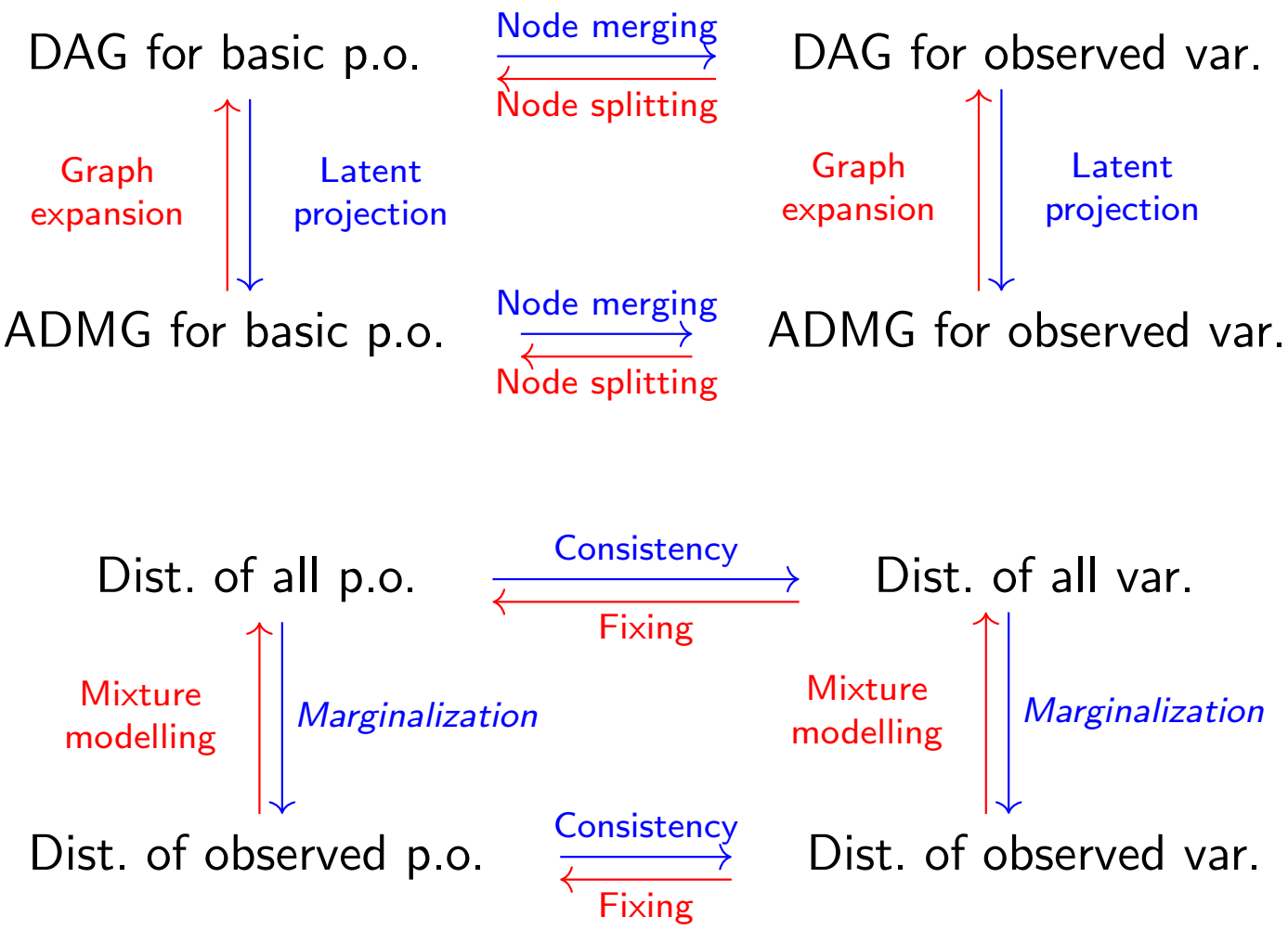
Theorem

Suppose indicies in $\mathcal{I}$ can be arranged into a sequence $(i_1, \ldots, i_m)$ such that $V_{i_k}(\boldsymbol{v}_{\mathcal{I}_{[k-1]}})$ is fixable in $\mathcal{G}(\mathcal{I}_{[k-1]})$, $k = 1, \ldots, m$. Then for all $\boldsymbol{v}$ and $\tilde{\boldsymbol{v}}$ we have

$$\frac{\mathbb{P}(\boldsymbol{V}_{\mathcal{I}}(\boldsymbol{v}_{\mathcal{I}}) = \tilde{\boldsymbol{v}}_{\mathcal{I}}, \boldsymbol{V}_{\mathcal{V}\setminus\mathcal{I}}(\boldsymbol{v}_{\mathcal{I}}) = \tilde{\boldsymbol{v}}_{\mathcal{V}\setminus\mathcal{I}})}{\mathbb{P}(\boldsymbol{V}_{\mathcal{I}} = \boldsymbol{v}_{\mathcal{I}}, \boldsymbol{V}_{\mathcal{V}\setminus\mathcal{I}} = \tilde{\boldsymbol{v}}_{\mathcal{V}\setminus\mathcal{I}})} = \frac{\mathbb{P}(\boldsymbol{V}_{\mathcal{I}} = \tilde{\boldsymbol{v}}_{\mathcal{I}} \mid \boldsymbol{V}_{\mathrm{mb}(\mathcal{I})} = \tilde{\boldsymbol{v}}_{\mathrm{mb}(\mathcal{I})})}{\mathbb{P}(\boldsymbol{V}_{\mathcal{I}} = \boldsymbol{v}_{\mathcal{I}} \mid \boldsymbol{V}_{\mathrm{mb}(\mathcal{I})} = \tilde{\boldsymbol{v}}_{\mathrm{mb}(\mathcal{I})})}.$$

See blackboard: proof (when $\mathcal{I} = \{i\}$) and examples.

Big picture


DAG for basic p.o.
Node merging
Node splitting
DAG for observed var.
Graph expansion
Latent projection
Graph expansion
Latent projection
ADMG for basic p.o.
Node merging
Node splitting
ADMG for observed var.
Dist. of all p.o.
Consistency
Fixing
Dist. of all var.
Mixture modelling
Marginalization
Mixture modelling
Marginalization
Dist. of observed p.o.
Consistency
Fixing
Dist. of observed var.

Outline

Big picture

Causal Inference $\approx$ **Causal Identification** $+$ **Statistical Inference**.

Design trumps analysis

Let $\boldsymbol{V}$ be the observed data, $\boldsymbol{F}$ be the full data (incl. all p.o.), and η be some nuisance parameters.

$$\overbrace{\beta(V;\hat{\eta}) - \beta(\mathbb{P}_F)}^{\text{Error of causal estimator}}$$
$$= \underbrace{\beta(\mathbb{P}_V) - \beta(\mathbb{P}_F)}_{\text{Design bias}} + \underbrace{\beta(\eta(\mathbb{P}_V)) - \beta(\mathbb{P}_V)}_{\text{Model bias}} + \underbrace{\beta(V;\hat{\eta}) - \beta(\eta(\mathbb{P}_V))}_{\text{Statistical noise}}.$$

For the rest of this course, we will shift our focus to the statistical aspect, i.e. the tradeoff between the last two terms in this decomposition.

Setting

- Full data $F_i = (X_i, A_i, Y_i(\cdot))$.
- Observed data $V_i = (X_i, A_i, Y_i)$.

No unmeasured confounders[5]

For the rest of this Chapter, we will assume i.i.d. data and $A \perp\!\!\!\perp Y(a) \mid X, a \in \mathcal{A}$.

- We will work with binary exposure ($\mathcal{A} = \{0, 1\}$), although many things below easily extend to exposures with multiple levels.
- We will discuss two approaches to statistical inference:
  1. Matching and randomization inference;
  2. Semiparametric inference.

[5] This assumption dismisses the important practical question about how the confounders should be selected. For a recent review on confounder selection, see Guo, Lundborg, Zhao, "Confounder Selection: Objectives and Approaches", arxiv:2208.13871.

Matching

- Basic idea: use observational data to "mimic" a randomized experiment by grouping treated and control units with similar covariates.
- This requires a "distance". One example is the Mahalanobis distance:

$$d_{\text{MA}}(x, \tilde{x}) = (x - \tilde{x})^T \hat{\Sigma} (x - \tilde{x}),$$

where $\hat{\Sigma}$ is some estimated covariance matrix of x.

Propensity score

One useful concept is the propensity score $\pi(X) = \mathbb{P}(A = 1 \mid X)$.

- This "summarizes" the confounders in the sense that $A \perp\!\!\!\perp Y(a) \mid \pi(X), a \in \mathcal{A}$.
- From a graphical perspective, this is essentially the fixing operator.
- This motivates the following choice of distance:

$$d_{\text{PS}}(x, \tilde{x}) = \left\{ \text{logit}(\hat{\pi}(x)) - \text{logit}(\hat{\pi}(\tilde{x})) \right\}^2,$$

where $\text{logit}(\pi) = \log(\pi/(1-\pi))$ is the logistic function and $\hat{\pi}$ is an estimator of the propensity score.

Some matching algorithms

When discussing matching algorithms, we assume $A_1 = \cdots = A_{n_1} = 1$ and $A_{n_1+1} = \cdots = A_n = 0$.

Nearest-neighbour (i.e. "greedy") matching

Sequentially match unit $1 \le i \le n_1$ to $\underset{n_1+1 \le j \le n}{\arg\min}\ d(X_i, X_j)$.

Optimal matching

Solve the following optimization problem:

$$\text{minimize} \quad \sum_{i=1}^{n_1} d\Big(X_i, \sum_{j=n_1+1}^{n} M_{ij} X_j\Big)$$

$$\text{subject to} \quad M_{ij} \in \{0,1\},\ \sum_{j=1}^{n_0} M_{ij} = 1,\ \sum_{i=1}^{n_1} M_{ij} \le 1,\ 1 \le i \le n_1,\ 1 \le j \le n_0.$$

This can be recasted as a network flow problem and solved efficiently.

How to check if matching is satisfactory?

Heuristic: in randomized Bernoulli trials, all pre-treatment covariates are approximately "balanced" in the treated and control groups.

- We can measure covariate imbalance as

$$B_k(\boldsymbol{M}) = \frac{\frac{1}{n_1}\sum_{i=1}^{n_1}(X_{ik} - \sum_{j=1}^{n_0} M_{i,j+n_1} X_{jk})}{\sqrt{(S_{1k}^2 + S_{0k}^2)/2}}, \ k = 1, \ldots, p.$$

- Some practical guidelines require $B_k(\boldsymbol{M}) < 0.1$ for all k.
- One can also incorporate such constraints in the mixed integer program and solve it using modern optimizers.

Inference after matching

- One option: use the average treated-minus-control difference to estimate the *average treatment effect on the treated*: $\mathbb{E}\{Y(1) - Y(0) \mid A = 1\}$.
- Here we explore randomization inference after matching. To simplify the notation, suppose unit i is matched to unit $i + n_1$.
- Denote $\mathcal{M} = \{\boldsymbol{a} \in \{0,1\}^{2n_1} : a_i + a_{i+n_1} = 1 \text{ for all } i \in [n_1]\}$.

Assumption: Matching recreates a pairwise randomized experiment

$$\mathbb{P}(\boldsymbol{A} = \boldsymbol{a} \mid \boldsymbol{X}, \boldsymbol{Y}(\cdot), \boldsymbol{A} \in \mathcal{M}) = \begin{cases} 2^{-n_1}, & \text{if } \boldsymbol{a} \in \mathcal{M}, \\ 0, & \text{otherwise.} \end{cases}$$

- This is true if there are no unmeasured confounders and the propensity scores are exactly matched.[6]

- One can then use any randomization test (e.g. with the signed rank/score statistic; see Example Sheet 1) to test a sharp null hypothesis.

[6] There is one more caveat: the set of matched units may still depend on the realized exposures.

Semiparametric inference

- The general problem: estimating a statistical functional $\beta = \beta(\mathbb{P}_V)$.
- Example: $\beta = \mathbb{E}\{Y(1) - Y(0)\} = \mathbb{E}\{\mathbb{E}(Y \mid A = 1, X) - \mathbb{E}(Y \mid A = 0, X)\}$.
- This is a *semiparametric* problem because it involves infinite-dimensional nuisance parameters (two regression functions).
- We will discuss some general solutions to this problem before applying them to functionals in causal inference.

General setup

$V_1, \ldots, V_n \overset{\text{i.i.d.}}{\sim} \mathbb{P} \in \mathcal{P}$, where $\mathcal{P}$ is a *statistical model*. We say $\mathcal{P}$ is

- *parametric* if it is indexed by finite-dimensional parameters;
- *nonparametric* if it is dense in all probability distributions of V;
- *semiparametric* if it is in between.

- As this is a course on causal inference instead of asymptotic statistics, we will gloss over regularity conditions and technical details below.

von-Mises expansion

Definition: Gateaux derivative

Consider a one-dimensional model $\{\mathbb{P}_\epsilon = (1-\epsilon)\mathbb{P} + \epsilon\mathbb{Q} \in \mathcal{P} : \epsilon \in (-\delta, \delta)\}$. The Gateaux derivative of β at $\mathbb{P}$ in the direction $\mathbb{Q} - \mathbb{P}$ is given by

$$\beta'_{\mathbb{P}}(\mathbb{Q} - \mathbb{P}) = \frac{d}{d\epsilon}\beta(\mathbb{P}_\epsilon)\Big|_{\epsilon=0},$$

if the derivative exists at $\epsilon = 0$.

The (first order) *von-Mises expansion* of β is given by

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

where $R(\mathbb{P}_n, \mathbb{P}) = o_{\mathbb{P}}(1/\sqrt{n})$ is a negligible remainder term.

See blackboard:

- Definition of the empirical distribution $\mathbb{P}_n$.
- Heuristics for the von-Mises expansion.

Influence function

- Typically, $\beta'_{\mathbb{P}}$ is a continuous linear map, so

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

- Denote $\phi_{\mathbb{P}}(v) = \beta'_{\mathbb{P}}(\delta_v - \mathbb{P})$, the *influence function* of β at $\mathbb{P}$.

Theorem (Functional delta method)

Suppose β is smooth[7] and $\beta(\mathbb{P}_n)$ is well defined. Then

$$\sqrt{n}\{\beta(\mathbb{P}_n) - \beta(\mathbb{P})\} \xrightarrow{d} \mathrm{N}(0, \mathrm{Var}(\phi_{\mathbb{P}}(V))).$$

See blackboard:

- Properties of the influence function.
- Examples: population mean and Z-estimation.

[7] The "right" notion of smoothness is Hardamard differentiability. See Sec. 20.2 of van der Vaart, *Asymptotic Statistics*, CUP.

Bias correction using the influence function

- Often $\beta(\mathbb{P}_n)$ is not well defined (e.g. if β depends on the density function).
- An alternative is to use $\beta(\hat{\mathbb{P}})$, where $\mathbb{P}$ is a parametric or smooth estimator of $\mathbb{P}$. However, $\hat{\mathbb{P}}$ may not converge to $\mathbb{P}$ at a $1/\sqrt{n}$ rate.
- The *one-step correction* of $\beta(\hat{\mathbb{P}})$ is defined as

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

Theorem (Asymptotic normality of the one-step estimator)

Suppose β is smooth and the following conditions are satisfied:

1. $\sqrt{n}\mathbb{E}_{\mathbb{P}_n - \mathbb{P}}(\phi_{\hat{\mathbb{P}}}(V) - \phi_{\mathbb{P}}(V)) \to 0$ *in probability;*
2. $\sqrt{n}R(\mathbb{P}, \hat{\mathbb{P}}) \to 0$ *in probability (e.g. if* $n^{1/4}\|\mathbb{P} - \hat{\mathbb{P}}\| \to 0$*).*

Then we have $\sqrt{n}\{\hat{\beta}_{\text{1-step}} - \beta(\mathbb{P})\} \xrightarrow{d} \mathrm{N}(0, \mathrm{Var}(\phi_{\mathbb{P}}(V)))$.

See blackboard:

- Heuristics for the one-step estimator.
- Expansion of $\hat{\beta}_{\text{1-step}} - \beta(\mathbb{P})$.
- Using cross-fitting to remove condition 1.

Calculus of influence functions

- In general, the influence function is defined as the Riesz representation of the derivative of β. This requires us to solve an integral equation.
- More intuitive alternative: calculate $\phi_{\mathbb{P}}(v) = \beta'_{\mathbb{P}}(\delta_v - \mathbb{P})$ using the chain rule for differentiation.[8]
- In applying this calculus, one may pretend that V has a finite support.

See blackboard:

- Influence function of $\mathbb{P}(X = x)$ is given by $\delta_x - \mathbb{P}(X = x)$.
- Influence function of $\mathbb{E}(Y \mid X = x)$ is $\{Y - \mathbb{E}(Y \mid X = x)\}\dfrac{\delta_x}{\mathbb{P}(X = x)}$.
- Influence function of $\beta = \mathbb{E}\{\mathbb{E}(Y \mid A = 1, X)\}$ is given by

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

- The augmented inverse probability weighted (AIPW) aka the doubly robust (DR) estimator of the ATE and its properties.

[8] The main issue we are neglecting here is that $\beta'_{\mathbb{P}}$ is generally not defined for a singular direction like $\delta_v - \mathbb{P}$.

Inverse probability weighting (IPW)

- Let $\mu_a(X) = \mathbb{E}(Y \mid X, A = a)$ and $\pi(X) = \mathbb{P}(A = 1 \mid X)$.
- Motivation: $\mathbb{E}\{\mu_1(X)\}$ is a linear functional of μ_1, so it has an Riesz representation:

$$\mathbb{E}\{\mu_1(X)\} = \mathbb{E}\Big\{\frac{A}{\pi(X)}\mu_1(X)\Big\} = \mathbb{E}\Big\{\frac{A}{\pi(X)}Y\Big\}.$$

- This may also be viewed as an application of the fixing operator.
- This motivates the IPW estimator of the ATE

$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n} \frac{A_i}{\hat{\pi}(X_i)}Y_i - \frac{1 - A_i}{1 - \hat{\pi}(X_i)}Y_i$$

- In practice, one often gets better finite-sample properties by directly estimating $1/\pi(X)$ and $1/(1 - \pi(X))$.

Case study: Entropy Balancing

- Problem: assuming no unmeasured confounders, estimate

$$\beta = \mathbb{E}(AY(0)) = \mathbb{E}(A\mu_0(X)) = \mathbb{E}(\pi(X)\mu_0(X)) = \mathbb{E}\Big\{(1-A)w(X)\mu_0(X)\Big\},$$

  where $w(X) = \pi(X)/\{1-\pi(X)\}$.
- The influence function of β is given by

$$\phi_\beta(V) = (1-A)w(X)\{Y - \mu_0(X)\} + A\mu_0(X) - \beta.$$

- Suppose $A_i = 1$ for $1 \le i \le n_1$ and $A_i = 0$ for $n_1 + 1 \le i \le n$.
- Entropy balancing estimates $w_i = w(X_i), i = n_1 + 1, \ldots, n$ by solving

$$\begin{aligned} \text{maximize} \quad & -\sum_{i=n_1}^{n} w_i \log w_i \\ \text{subject to} \quad & \sum_{i=1}^{n_1} X_i = \sum_{i=n_1+1}^{n} w_i X_i, \\ & w_i > 0, \ \ i = 1, \ldots, n. \end{aligned}$$

See blackboard:

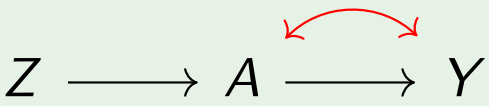
- Lagrangian dual problem and its statistical interpretation.
- Double robustness of entropy balancing.

Outline

Introduction

- Instrumental variable (IV) is the oldest and most well developed approach to deal with unmeasured confounders.

IV in linear SEMs



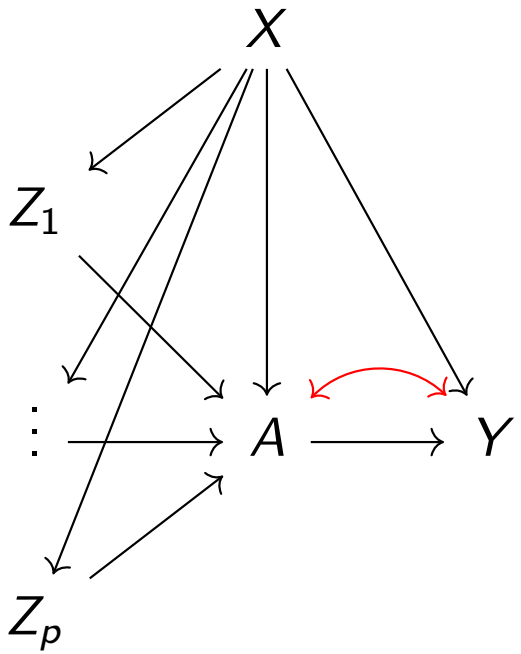
When a linear SEM is assumed, the causal effect of A on Y is generically identified by

$$\beta_{AY} = \frac{\mathrm{Cov}(Z, Y)}{\mathrm{Cov}(Z, A)}.$$

See handout:

- Example: returns to schooling.

Linear SEMs: Multiple IVs & exogenous covariates

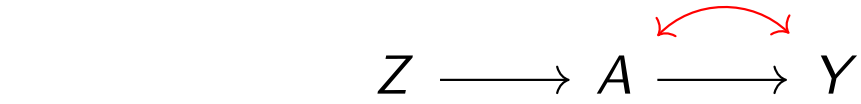


See blackboard:

- Linear SEM for this graph.
- Expression of $\mathbb{E}(Y \mid \boldsymbol{Z}, X)$.
- The two-stage least squares estimator.

Counterfactual definition of IV

Now consider the simplest case:



This entails three assumptions:

Assumption	Graph	Counterfactual
1. Relevance	$Z \to A$	$Z \not\perp\!\!\!\perp A$
2. Exogeneity	$Z \dashleftarrow\!\!\!\not\dashrightarrow Y$	$Z \perp\!\!\!\perp Y(z, a)$
3. Exclusion restriction	$Z \not\to Y$	$Y(z, a) = Y(a)$

- Any Z that satisfies these assumptions is called a *valid* instrumental variable.
- However, this generally only provides *partial identification* of the causal effect (ES2 Q13).
- Additional assumptions are needed for *point identification*. One example is effect homogeneity:

$$Y(a) - Y(a') = \beta(a - a') \text{ for all } a, a' \in \mathcal{A}.$$

Semiparametric estimation with IV

- Assume effect homogeneity and $0 \in \mathcal{A}$.
- By assumption, $Z \perp\!\!\!\perp Y(0) = Y - \beta A$. This defines a semiparametric model.
- Denote $\alpha = \mathbb{E}(Y - \beta A)$. Then (α, β) solves

$$\mathbb{E}\{(Y - \alpha - \beta A)g(Z)\} = 0, \text{ for all } g(\cdot).$$

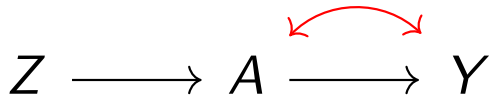
- Now suppose we have an i.i.d. sample (Z_i, A_i, Y_i), $i = 1, \ldots, n$.

See blackboard:

- Plug-in estimator $\hat{\beta}_g$ and its asymptotic distribution.
- Optimal choice of g ("optimal instrument").
- Further comments on IV-based estimators.

Beyond effect homogeneity: Motivating example

Suppose the simple IV graph describes a randomized experiment with *non-compliance*:



- Z is the treatment assignment (randomized);
- A is the actual treatment (not randomized).

There are two ways to analyze this data:

1. Ignore A (intention-to-treat analysis);
2. Use Z as an IV for A.

IV identification: complier average causal effect

- Assume $Z, A \in \{0, 1\}$. Define the compliance classes as

$$C = \begin{cases} \text{always taker (at)}, & \text{if } A(0) = 0, A(1) = 1, \\ \text{never taker (nt)}, & \text{if } A(0) = 1, A(1) = 0, \\ \text{complier (co)}, & \text{if } A(0) = 0, A(1) = 1, \\ \text{defier (de)}, & \text{if } A(0) = 1, A(1) = 0. \end{cases}$$

Theorem (Identification under monotonicity)

Assuming the multiple-world causal model w.r.t. the IV graph, $A \not\perp\!\!\!\perp Z$, and $\mathbb{P}(A(1) \geq A(0)) = 1$. Then

$$\beta_{CACE} = \mathbb{E}\{Y(1) - Y(0) \mid C = co\} = \frac{\mathrm{Cov}(Z, Y)}{\mathrm{Cov}(Z, A)}.$$

See blackboard:

- Proof of this result.