Causal Graphical Methods For Handling Nonignorable Missing Data

Razieh Nabi, Ph.D.
Rollins Assistant Professor
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
razieh.nabi@emory.edu

Daniel Scharfstein, Sc.D.
Department of Population Health Sciences
University of Utah School of Medicine
daniel.scharfstein@hsc.utah.edu

ACIC 2025

Slides are available at: https://raziehnabi.com/files/acic_2025.pdf

Course outline

Motivation ($\sim$10 minutes)

Part I. **Missing data DAG models** ($\sim$75 minutes)

Break (10 minutes)

Part II. **Non-parametric identification** ($\sim$75 minutes)

Break (10 minutes)

Part III. **Non/Semi-parametric estimation** ($\sim$75 minutes)

Wrap-up ($\sim$15 minutes)

Motivation

- We are often interested in a functional of an underlying distribution:
  - Population-level outcome Y: $\psi_1 = \mathbb{E}[Y]$.
  - Outcome mean in a sub-population $X = x$: $\psi_2 = \mathbb{E}[Y \mid X = x]$.
  - Average causal effect of binary treatment T on outcome Y:

$$\psi_3 = \mathbb{E}\Big[\mathbb{E}\big[Y \;\big|\; X, T = 1\big] - \mathbb{E}\big[Y \;\big|\; X, T = 0\big]\Big].$$

(Note that ψ_3 can only be interpreted as a causal effect under certain assumptions, such as *consistency*, *positivity*, and *conditional ignorability*.)

- How to compute these estimands?
  - We need samples from the underlying distribution, $\{X, T, Y\} \sim p(X, T, Y)$

Missing data indicators

- We look at the observed sample and it looks like the following:

X^*	T^*	Y^*
x_1	t_1	?
x_2	?	y_2
x_3	t_3	y_3
$\vdots$	$\vdots$	$\vdots$
?	t_n	?

- Each variable with missing values can have an underlying **missingness/response indicator**:
  - $R_V = 1$ if variable V is observed and $R_V = 0$ if V is "?".

X^*	T^*	Y^*	R_X	R_T	R_Y
x_1	t_1	?	1	1	0
x_2	?	y_2	1	0	1
x_3	t_3	y_3	1	1	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
?	t_n	?	0	1	0

Missing data challenges

- How to use data with missing values to estimate the parameter of interest?
  - Should we ignore rows with missing values?
  - Should we impute the missing values? If so, how?
  - Should we do something else?
- Our choices may affect our data analysis by
  - introducing **bias** due to differences between missing and complete data, and/or
  - losing **efficiency** when we ignore part of the observed sample.
- Before we choose what method to use, we need to know **why we have missing data** in the first place!

Sources of missingness

We encounter missing data for a variety of reasons:

- A survey was conducted and values were just randomly missed when being entered in the computer.
- A respondent chooses not to respond to a question like “Have you ever recreationally used opioids?"
- You decide to start collecting a new variable (due to new actions: like a pandemic) partway through the data collection of a study.
- You want to measure the speed of meteors, and some observations are just “too quick" to be measured properly.

The source of missing values in data leads to three distinct missingness mechanisms: MCAR, MAR, MNAR (Rubin, 1976).

Rubin's hierarchy of missingness

1. **Missing Completely at Random (MCAR)** - the probability of missingness in a variable is the same for all units. Like randomly poking holes in a data set.

2. **Missing at Random (MAR)** - the probability of missingness in a variable depends only on available information (in other predictors).

3. **Missing Not at Random (MNAR)** - the probability of missingness depends on information that has not been recorded and this information also predicts the missing values.

Rubin's hierarchy of missingness

Let Z : variables with no missingness

X : variables that are sometimes missing

X_{obs}: observed entries, X_{miss}: missing entries

R : missingness indicators

---

1. **MCAR**: $R \perp\!\!\!\perp X_{\text{miss}}, X_{\text{obs}}, Z, \quad p(R \mid X, Z) = p(R).$
   - probability that any observation is missing is independent of all data values, regardless of whether they are observed or unobserved.
2. **MAR**: $R \perp\!\!\!\perp X_{\text{miss}} \mid Z, X_{\text{obs}}, \quad p(R \mid X, Z) = p(R \mid Z, X_{\text{obs}})$
   - probability that any observation is missing depends only on elements that are observed.
3. **MNAR**: $R \not\!\perp\!\!\!\perp X_{\text{miss}}$ neither marginally nor conditionally
   - probability that any observation is missing depends on elements that are themselves missing – a missingness mechanism that is neither MCAR nor MAR.

Beyond the traditional missingness hierarchy

- Rubin's categorization, although provides a foundational framework, falls short in several aspects:
  - MCAR and MAR definitions are merely statistical convenience.
  - MNAR definition lacks specificity.
  - It does not determine the best approach for handling missing data in multiple variables.
- In order to better handle missing data, we need to:
  - Have a better understanding of causal relationships between variables and their missingness, and what these relationships imply in terms of identification/recoverability of the target estimand.
- Main takeaway: encourage the use of *missing data DAGs* in data analysis to
  - Make **assumptions** about missingness mechanisms more explicit, and
  - Use **identification** procedures as a guide for **estimation** methods.

Course outline and objectives

Part I. **Missing data DAG models**

- Represent missingness mechanisms graphically; interpret a missing data DAG model as a class of distributions with a set of independence restrictions (Markov properties).

Part II. **Non-parametric identification**

- Given a missing data DAG model, argue for identifiability of a given estimand: write down the estimand as a function of observed data, or prove its non-identifiability.

Part III. **Non/Semi-parametric estimation**

- Given an identified estimand, derive desirable estimators; Derive the efficient influence functions.

Part 1. Missing Data DAG Models

Missing data and causal inference

- Causal inference and missing data are analogous in terminology, theory of identification, and statistical inference.
- Causal inference has been viewed as a missing data problem:
  - Responses to some (hypothetical) treatment interventions are not observed.
  - Given the treatment vs placebo option, we only observe the potential outcome under treatment received or the potential outcome under placebo received, but not both.
- Missing data can be viewed as a causal inference problem:
  - Missingness indicators can be treated as intervenable treatments.
  - We can view variable X as a potential outcome had the missingness indicator R_X been set to 1 (had there been no missingness).
- In this part of the tutorial, we want to use developments in causal graphical models to reason about missing data models.

A causal workflow

1. Define a causal estimand in terms of counterfactuals.
2. Define a causal model that links counterfactuals to factual variables.
   - Impose assumptions on the distribution defined over counterfactual and factual variables.
3. Identify the causal estimand as a function of observed data in the assumed causal model.
4. Define a statistical model to estimate the identified causal estimand.
   - Perform statistical inference which includes testing and estimating the magnitude of a causal estimand given the observed data.
5. Assess assumptions with sensitivity analysis.

Example: a causal workflow

1. **Average causal effect**: $\text{ACE} := \mathbb{E}[Y^{(1)} - Y^{(0)}]$
   - $Y^{(t)}$: potential outcome Y when binary treatment T is assigned to $t = \{0, 1\}$.
2. $\mathcal{M}$: a causal model relating counterfactuals to factuals
   - **Consistency**: observed outcome Y is equal to the potential outcome $Y^{(t)}$ when the treatment received is $T = t$,
   - **Positivity**: $p(T = t \mid X = x) > 0$ for all x in the state space of X,
   - **Conditional ignorability**: $Y^{(t)} \perp T \mid X$.
3. Under the above causal model, we can identify ACE via the following functional, known as **adjustment formula** or **g-formula**:

$$\mathbb{E}[Y^{(1)} - Y^{(0)}] = \mathbb{E}\Big[\mathbb{E}\big[Y \mid T = 1, X\big] - \mathbb{E}\big[Y \mid T = 0, X\big]\Big] .$$

4, 5. Luckily, we are not short of any estimation or sensitivity analysis techniques!

Directed acyclic graph (DAG)

- The second step in the causal workflow is what distinguishes causal inference from traditional statistical inference.
  - Graphical models like directed acyclic graphs (DAGs) are often used to encode assumptions about the causal model.

Directed acyclic graph (DAG)

- The second step in the causal workflow is what distinguishes causal inference from traditional statistical inference.
  - Graphical models like directed acyclic graphs (DAGs) are often used to encode assumptions about the causal model.
- A graph $\mathcal{G}(V, E)$ is a set of vertices/nodes V that correspond to random variables and a set E that contains the set of edges between variables.
- The graph $\mathcal{G}(V, E)$ is said to be **directed** and **acyclic** if:
  - There are only directed edges $(V_i \to V_j)$
  - There are no directed cycles – for any $V_i \in V$ there is no sequence of directed edges in $\mathcal{G}$ such that $V_i \to \ldots \to V_i$

$$\begin{tikzcd} X \arrow[r] \arrow[rr, bend left] & T \arrow[r] & Y \end{tikzcd}$$

✓

$$\begin{tikzcd} X \arrow[r] & T \arrow[r] & Y \arrow[ll, bend right] \arrow[loop above] \end{tikzcd}$$

✗

- For notational convenience, we often refer to $\mathcal{G}(V, E)$ as $\mathcal{G}(V)$ or simply $\mathcal{G}$.

Statistical model of a DAG

- DAG $\mathcal{G}(V)$ encodes a set of independence restrictions on the joint distribution $p(V)$.
- The joint distribution $p(V)$ corresponding to DAG $\mathcal{G}(V)$ has three **equivalent** characterizations:
  - **Factorization** (writes the distribution as a set of small factors.)
  - **Local Markov property** (lists a complete set of independence constraints.)
  - **Global Markov property** (lists all independence constraints in the model.)

Statistical model of a DAG ctd.

The joint distribution $p(V)$ satisfies the **factorization property** wrt DAG $\mathcal{G}(V)$ if:

$$p(V) = \prod_{V_i \in V} p(V_i \mid \mathrm{pa}_{\mathcal{G}}(V_i)),$$

- $\mathrm{pa}_{\mathcal{G}}(V_i) = \{V_j \in V \mid V_j \to V_i\}$ denotes parents of V_i in $\mathcal{G}(V)$.

Statistical model of a DAG ctd.

The joint distribution $p(V)$ satisfies the **factorization property** wrt DAG $\mathcal{G}(V)$ if:

$$p(V) = \prod_{V_i \in V} p(V_i \mid \mathrm{pa}_{\mathcal{G}}(V_i)),$$

- $\mathrm{pa}_{\mathcal{G}}(V_i) = \{V_j \in V \mid V_j \to V_i\}$ denotes parents of V_i in $\mathcal{G}(V)$.

The joint distribution $p(V)$ satisfies the **local Markov property** wrt DAG $\mathcal{G}(V)$ if:

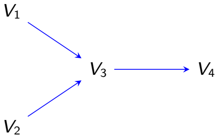
$$V_i \perp\!\!\!\perp \mathrm{nd}_{\mathcal{G}}(V_i) \setminus \mathrm{pa}_{\mathcal{G}}(V_i) \mid \mathrm{pa}_{\mathcal{G}}(V_i), \quad \forall V_i \in V$$

(V_i is independent of its non-descendants non-parents given its parents)

- $\mathrm{nd}_{\mathcal{G}}(V_i) = \{V_j \in V \mid V_j \notin \mathrm{de}_{\mathcal{G}}(V_i)\}$ denotes non-descendants of V_i in $\mathcal{G}(V)$,
- $\mathrm{de}_{\mathcal{G}}(V_i) = \{V_j \in V \mid V_i \to \ldots \to V_j\}$ denotes descendants of V_i in $\mathcal{G}(V)$.

Example: statistical model of a DAG

Consider the following DAG:



- According to the **DAG factorization**, the statistical model of this DAG is a set of distributions $p(V)$, where $V = \{V_1, V_2, V_3, V_4\}$ s.t.,

$$\Big\{ \ p(V) = p(V_1) \times p(V_2) \times p(V_3 \mid V_1, V_2) \times p(V_4 \mid V_3) \ \Big\}.$$

- According to the **local Markov property**, the statistical model of this DAG is a set of distributions $p(V)$ s.t.,

$$\Big\{ \ p(V) \ \text{ s.t. } \ V_1 \perp\!\!\!\perp V_2 \text{ and } \ V_4 \perp\!\!\!\perp V_1, V_2 \mid V_3 \ \Big\}.$$

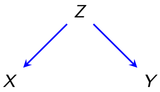
  The above list implies a larger set of independence restrictions, e.g., $V_4 \perp\!\!\!\perp V_1 \mid V_3$ or $V_4 \perp\!\!\!\perp V_1 \mid V_2, V_3$. (graphoid axioms)

Global Markov property: d-separation

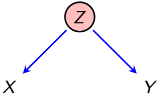
- Given a DAG $\mathcal{G}(V)$, can we answer arbitrary independence queries of the form $X \perp\!\!\!\perp Y \mid Z$ in $p(V)$, where X, Y, Z are disjoint subsets of V?
- **d-separation** (directed-separation) is a graphical criterion that allows one to answer such queries in an automated fashion (Pearl, 2000; Verma and Pearl, 1990).
- Here are the three types of triplets that define d-separation:
  - Forks $X \leftarrow Z \rightarrow Y$
  - Chains $X \rightarrow Z \rightarrow Y$ or $X \leftarrow Z \leftarrow Y$
  - Colliders $X \rightarrow Z \leftarrow Y$

Fork triplets

In a fork $X \leftarrow Z \rightarrow Y$, the variables X and Y are marginally dependent, but conditionally independent given Z.



$X \not\perp\!\!\!\perp Y$



$X \perp\!\!\!\perp Y \mid Z$

- Intuition: X and Y **share a common cause** and thus dependent.
  - Upon observing the common cause Z, the two effects X and Y are no longer related.
- Example: X: shark attacks, Z: warm whether, and Y: ice cream sales.

  Warmer weather draws more people to the beach. It also drives up ice cream sales.

Chain triplets

In a chain $X \to Z \to Y$, the variables X and Y are marginally dependent, but conditionally independent given Z.

$$X \longrightarrow Z \longrightarrow Y \qquad\qquad X \longrightarrow Z \longrightarrow Y$$

$$X \not\perp\!\!\!\perp Y \qquad\qquad X \perp\!\!\!\perp Y \mid Z$$

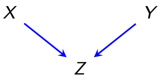
- <u>Intuition</u>: if Z is a noisy version of X and Y is a noisy version of Z, then Y is a noisy version of X.
  - Upon observing Z, X holds no extra information about Y.
  - Z **screens off the effect** of X on Y.

- <u>Example</u>: X: blood sugar, Z: stomach acidity, and Y: hunger.

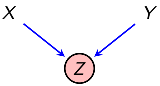
  Blood sugar causes hunger, but only indirectly through increasing the stomach acidity.

Collider triplets

In a collider $X \to Z \leftarrow Y$, the variables X and Y are marginally independent, but conditionally dependent given Z.



$X \perp\!\!\!\perp Y$



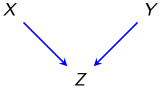
$X \not\!\perp\!\!\!\perp Y \mid Z$

- Intuition: if X and Y only **share a common effect**, they are independent. That is the common effect has two independent sources of causes.
  - Upon observing the common effect, the two causes become dependent.
  - This is often referred to as a **Berkson's paradox**.
- Example: X: battery, Z: car starts, and Y: fuel.

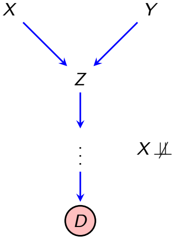
  If we observe that the car fails to start, then knowing something about the fuel status tells us something about the battery status & vice versa

Collider extensions

In a colider $X \to Z \leftarrow Y$, the variables X and Y are marginally independent, but conditionally dependent given a *descendant* of Z.



$X \perp\!\!\!\perp Y$



$X \not\perp\!\!\!\perp Y \mid D$

- Example: X: battery, Z: car starts, Y: fuel, and D: taken to mechanic

  Extend the previous example where the car was taken to a mechanic.

- Conditioning can induce dependence, not just remove it.

Summary of the (in)dependence rules

$$X \longleftarrow Z \longrightarrow Y$$

$$X \not\perp\!\!\!\perp Y$$

$$X \longleftarrow Z \longrightarrow Y$$

$$X \perp\!\!\!\perp Y \mid Z$$

$$X \longrightarrow Z \longrightarrow Y$$

$$X \not\perp\!\!\!\perp Y$$

$$X \longrightarrow Z \longrightarrow Y$$

$$X \perp\!\!\!\perp Y \mid Z$$

$$X \longrightarrow Z \longleftarrow Y$$

$$X \perp\!\!\!\perp Y$$

$$X \longrightarrow Z \longleftarrow Y$$

$$X \not\perp\!\!\!\perp Y \mid Z$$

$$\begin{array}{ccccc} X & \longrightarrow & Z & \longleftarrow & Y \\ & & \downarrow & & \\ & & \vdots & & \\ & & \downarrow & & \\ & & D & & \end{array}$$

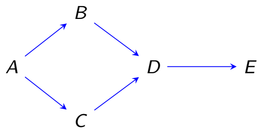
$$X \not\perp\!\!\!\perp Y \mid D$$

- Forks/chains are *open*, but become *blocked* upon conditioning.
- Colliders are *blocked*, but become *open* upon conditioning.

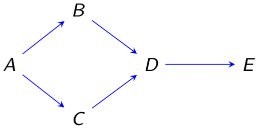
From blocked triplets to d-separation

- A **path** from X to Y is a sequence of consecutive edges connecting X and Y such that no node (and consequently no edge) appears more than once in the sequence.
- A path from X to Y is blocked by Z if there is a **blocking triplet** on the path.
  - There exists a blocked chain or fork on the path, or
  - A collider that is not open.
- Dependence is like water flow and paths are pipes. A single block is enough to block the whole path.
- X and Y are **d-separated** given Z if all paths from X to Y are blocked by Z, and is denoted by $X \perp\!\!\!\perp_d Y \mid Z$.

Examples: d-separation



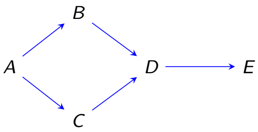
- Is $A \perp\!\!\!\perp_d E \mid C$?



- Is $A \perp\!\!\!\perp_d E \mid C$?

  No! because $A \rightarrow B \rightarrow D \rightarrow E$ is still open.

Examples: d-separation

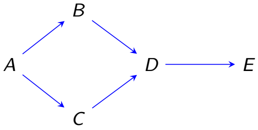


- Is $A \perp\!\!\!\perp_d E \mid C$?

  No! because $A \rightarrow B \rightarrow D \rightarrow E$ is still open.

- Is $B \perp\!\!\!\perp_d C \mid A$?

Examples: d-separation



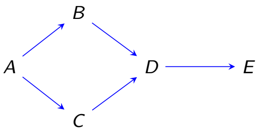
- Is $A \perp\!\!\!\perp_d E \mid C$?

  No! because $A \to B \to D \to E$ is still open.

- Is $B \perp\!\!\!\perp_d C \mid A$?

  Yes! $B \leftarrow A \to C$ and $B \to D \leftarrow C$ are both blocked.

Examples: d-separation



- Is $A \perp\!\!\!\perp_d E \mid C$?

  No! because $A \rightarrow B \rightarrow D \rightarrow E$ is still open.

- Is $B \perp\!\!\!\perp_d C \mid A$?

  Yes! $B \leftarrow A \rightarrow C$ and $B \rightarrow D \leftarrow C$ are both blocked.

- Is $B \perp\!\!\!\perp_d C \mid A, E$?

Examples: d-separation

$$\begin{tikzcd}[row sep=small] & B \arrow[dr] & & \\ A \arrow[ur] \arrow[dr] & & D \arrow[r] & E \\ & C \arrow[ur] & & \end{tikzcd}$$

- Is $A \perp\!\!\!\perp_d E \mid C$?

  No! because $A \to B \to D \to E$ is still open.

- Is $B \perp\!\!\!\perp_d C \mid A$?

  Yes! $B \leftarrow A \to C$ and $B \to D \leftarrow C$ are both blocked.

- Is $B \perp\!\!\!\perp_d C \mid A, E$?

  No! $B \to D \leftarrow C$ is now open (condition on E opens up the collider at D).

Global Markov property

For any distribution $p(V)$ that satisfies the DAG factorization wrt $\mathcal{G}(V)$, the following **Global Markov property** holds: for all disjoint subsets X, Y, Z of V we have,

$$\left(X \perp\!\!\!\perp_d Y \mid Z\right)\Big|_{\mathcal{G}(V)} \implies \left(X \perp\!\!\!\perp Y \mid Z\right)\Big|_{p(V)}$$

where $(\perp\!\!\!\perp_d)\big|_{\mathcal{G}}$ denotes d-separation in $\mathcal{G}$ and $(\perp\!\!\!\perp)\big|_{p}$ denotes independence in p.

- We can apply a purely graphical criterion to a DAG $\mathcal{G}(V)$ to tell us about conditional independence facts in the joint distribution $p(V)$.
- The above is a one way implication!
  - We could indeed have extra independence restrictions in $p(V)$ that cannot be read by d-separation (this occurs in *unfaithful* distributions). (Peters et al., 2014; Peters, 2015; Sadeghi, 2017)

Equivalence of DAG properties

The statistical model of a DAG is characterized with three definitions:

- **Factorization** (writes the distribution as a set of small factors).
- **Local Markov property** (lists a small but complete set of independence constraints).
- **Global Markov property** (lists all independence constraints in the model).

A distribution $p(V)$ factorizes according to a DAG $\mathcal{G}(V)$ *if and only if* it obeys the local Markov property according to $\mathcal{G}(V)$ *if and only if* it obeys the global Markov property according to $\mathcal{G}(V)$ (Verma and Pearl, 1990).

DAG factorization $\iff$ Local Markov property $\iff$ Global Markov property

Causal model of a DAG

- The causal model of a DAG can be formally defined in terms of **Nonparametric Structural Equations Model** (NPSEM).
  - It describes how "nature" assigns values to each variable in the model.

- For every $V_i \in V$:

$$V_i \;\leftarrow\; f_{V_i}\big(\,\mathrm{pa}_{\mathcal{G}}(V_i),\; \epsilon_{V_i}\big)$$

  - ϵ_{V_i} denotes the error term (all external unmeasured causes of V_i).
  - f_{V_i} is nonparametric. It does not constrain the dependence of V_i on its parents and ϵ_{V_i} in any way.
  - This is an imperative assignment, not an equality! which means the model is not "reversible."

Causal model of a DAG

- The causal model of a DAG can be formally defined in terms of **Nonparametric Structural Equations Model** (NPSEM).
  - It describes how "nature" assigns values to each variable in the model.

- For every $V_i \in V$:

$$V_i \;\leftarrow\; f_{V_i}\big(\,\mathrm{pa}_{\mathcal{G}}(V_i),\ \epsilon_{V_i}\big)$$

  - ϵ_{V_i} denotes the error term (all external unmeasured causes of V_i).
  - f_{V_i} is nonparametric. It does not constrain the dependence of V_i on its parents and ϵ_{V_i} in any way.
  - This is an imperative assignment, not an equality! which means the model is not "reversible."

- **NPSEM with independent errors** (NPSEM-IE) (Pearl, 2009)
  - Unmeasured factors are assumed to be independent – a reflection that all common causes have been measured.
  - The explicit assumption is that $\perp\!\!\!\perp \{\epsilon_{V_i},\ \forall V_i \in V\}$, and thus $p(\epsilon) = \prod_{V_i \in V} p(\epsilon_{V_i})$.

Intervention in causal models

- Let T be the variable that we would like to (hypothetically) intervene on and set it to t.
- An intervention that sets $T = t$ entails the following three changes:

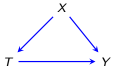
  I. **Structural** changes to the causal model
    - In the corresponding NPSEM, replace $T \leftarrow f_T(\mathrm{pa}_{\mathcal{G}}(T), \epsilon_T)$ with $T \leftarrow t$.
    - The structural equations for other variables may change depending on their genealogical relations to T.

  II. **Graphical** changes to the DAG
    - In the corresponding DAG $\mathcal{G}(V)$, delete all incoming edges into T and switch random T to fixed value t.
    - Nodes on the downstream of T turn into counterfactuals.

  III. **Probabilistical** changes to the joint distribution
    - In the corresponding joint distirbution $p(V)$, drop the factor $p(T \mid \mathrm{pa}_{\mathcal{G}}(T))$ from the factorization of $p(V)$, and evaluate all other factors at $T = t$.
    - Upon the intervention, we end up with a *truncated* factorization.

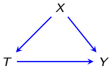
Example: intervention in causal models

- Let T be the treatment of interest in the following DAG:



Example: intervention in causal models

- Let T be the treatment of interest in the following DAG:



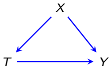
- Structural operation of intervening on T and setting it to t:

$$
\begin{aligned}
X &\leftarrow f_X(\epsilon_X) \\
T &\leftarrow f_T(X, \epsilon_T) \\
Y &\leftarrow f_Y(X, T, \epsilon_Y)
\end{aligned}
\qquad\qquad
\begin{aligned}
X &\leftarrow f_X(\epsilon_X) \\
T &\leftarrow t \\
Y^{(t)} &\leftarrow f_Y(X, t, \epsilon_Y)
\end{aligned}
$$

NPSEM-IE implies the conditional ignorability assumption: $Y^{(t)} \perp\!\!\!\perp T \mid X$.

Example: intervention in causal models

- Let T be the treatment of interest in the following DAG:

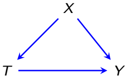


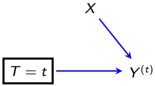
- Structural operation of intervening on T and setting it to t:

$$\begin{aligned} X &\leftarrow f_X(\epsilon_X) \\ T &\leftarrow f_T(X, \epsilon_T) \\ Y &\leftarrow f_Y(X, T, \epsilon_Y) \end{aligned} \qquad\qquad \begin{aligned} X &\leftarrow f_X(\epsilon_X) \\ T &\leftarrow \color{red}{t} \\ Y^{(\color{red}{t})} &\leftarrow f_Y(X, \color{red}{t}, \epsilon_Y) \end{aligned}$$

  NPSEM-IE implies the conditional ignorability assumption: $Y^{(\color{red}{t})} \perp\!\!\!\perp T \mid X$.

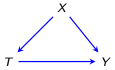
- Graphical operation of this intervention is illustrated as follows:





Example: intervention in causal models

- Let T be the treatment of interest in the following DAG:

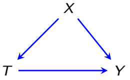


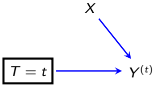
- Structural operation of intervening on T and setting it to t:

$$\begin{aligned} X &\leftarrow f_X(\epsilon_X) \\ T &\leftarrow f_T(X, \epsilon_T) \\ Y &\leftarrow f_Y(X, T, \epsilon_Y) \end{aligned} \qquad\qquad \begin{aligned} X &\leftarrow f_X(\epsilon_X) \\ T &\leftarrow {\color{red}t} \\ Y^{({\color{red}t})} &\leftarrow f_Y(X, {\color{red}t}, \epsilon_Y) \end{aligned}$$

NPSEM-IE implies the conditional ignorability assumption: $Y^{(t)} \perp\!\!\!\perp T \mid X$.

- Graphical operation of this intervention is illustrated as follows:



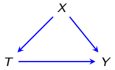


- Probabilistic operation of this intervention is a truncated factorization:

$$p(X, Y^{(t)}) = \left.\frac{p(X, Y, T)}{p(T \mid X)}\right|_{T=t} = \frac{p(X, Y, T = t)}{p(T = t \mid X)} = p(X)\, p(Y \mid T = t, X).$$

Example: intervention in causal models

- Let T be the treatment of interest in the following DAG:

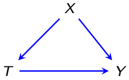


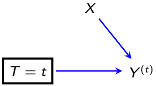
- Structural operation of intervening on T and setting it to t:

$$\begin{aligned} X &\leftarrow f_X(\epsilon_X) \\ T &\leftarrow f_T(X, \epsilon_T) \\ Y &\leftarrow f_Y(X, T, \epsilon_Y) \end{aligned} \qquad\qquad \begin{aligned} X &\leftarrow f_X(\epsilon_X) \\ T &\leftarrow \color{red}{t} \\ Y^{(\color{red}{t})} &\leftarrow f_Y(X, \color{red}{t}, \epsilon_Y) \end{aligned}$$

NPSEM-IE implies the conditional ignorability assumption: $Y^{(\color{red}{t})} \perp\!\!\!\perp T \mid X$.

- Graphical operation of this intervention is illustrated as follows:





- Probabilistic operation of this intervention is a truncated factorization:

$$p(X, Y^{(t)}) = \left.\frac{p(X, Y, T)}{p(T \mid X)}\right|_{T=t} = \frac{p(X, Y, T=t)}{p(T=t \mid X)} = p(X)\; p(Y \mid T=t, X).$$

Do-calculus notation of Pearl: $p(X, Y^{(t)}) \equiv p(X, Y \mid \color{red}{\mathrm{do}(}T = t\color{red}{)})$.

Missing data notation

- $X = (X_1, \ldots, X_K)^T$: a vector of K random variables
- Given a finite sample from $p(X)$:
  - $R = (R_1, \ldots, R_K)^T$: binary missingness indicators
    $R_k = 1$ if X_k is observed, and $R_k = 0$ otherwise
  - $X^* = (X_1^*, \ldots, X_K^*)^T$: coarsened version of X
    $X_k^* = X_k$ if $R_k = 1$, and $X_k^* = ?$ otherwise.

Missing data notation

- $X = (X_1, \ldots, X_K)^T$: a vector of K random variables
- Given a finite sample from $p(X)$:
  - $R = (R_1, \ldots, R_K)^T$: binary missingness indicators

    $R_k = 1$ if X_k is observed, and $R_k = 0$ otherwise
  - $X^* = (X_1^*, \ldots, X_K^*)^T$: coarsened version of X

    $X_k^* = X_k$ if $R_k = 1$, and $X_k^* = ?$ otherwise.
- Causal interpretation of the tuple (X_k, R_k, X_k^*):
  - R_k: a treatment variable that can be intervened on.
  - X_k: a counterfactual – had we intervened and set $R_k = 1$.
  - X_k^*: a factual variable.

Missing data notation

- $X = (X_1, \dots, X_K)^T$: a vector of K random variables
- Given a finite sample from $p(X)$:
  - $R = (R_1, \dots, R_K)^T$: binary missingness indicators

    $R_k = 1$ if X_k is observed, and $R_k = 0$ otherwise
  - $X^* = (X_1^*, \dots, X_K^*)^T$: coarsened version of X

    $X_k^* = X_k$ if $R_k = 1$, and $X_k^* = ?$ otherwise.
- Causal interpretation of the tuple (X_k, R_k, X_k^*):
  - R_k: a treatment variable that can be intervened on.
  - X_k: a counterfactual – had we intervened and set $R_k = 1$.
  - X_k^*: a factual variable.
- Switching notation to emphasize the counterfactual connotation:

$$X,\ R,\ X^* \ \mapsto\ L^{(1)},\ R,\ L$$

$L^{(1)} = (L_1^{(1)}, \dots, L_K^{(1)})^T$ and $L = (L_1, \dots, L_K)^T$.

Missing data notation

- $X = (X_1, \ldots, X_K)^T$: a vector of K random variables
- Given a finite sample from $p(X)$:
  - $R = (R_1, \ldots, R_K)^T$: binary missingness indicators

    $R_k = 1$ if X_k is observed, and $R_k = 0$ otherwise
  - $X^* = (X_1^*, \ldots, X_K^*)^T$: coarsened version of X

    $X_k^* = X_k$ if $R_k = 1$, and $X_k^* = ?$ otherwise.
- Causal interpretation of the tuple (X_k, R_k, X_k^*):
  - R_k: a treatment variable that can be intervened on.
  - X_k: a counterfactual – had we intervened and set $R_k = 1$.
  - X_k^*: a factual variable.
- Switching notation to emphasize the counterfactual connotation:

$$X,\ R,\ X^* \ \mapsto\ L^{(1)},\ R,\ L$$

$L^{(1)} = (L_1^{(1)}, \ldots, L_K^{(1)})^T$ and $L = (L_1, \ldots, L_K)^T$.

- Z: completely observed variables

Missing data models

- A missing data model $\mathcal{M}$ is a set of distributions defined over variables in $\{Z, L^{(1)}, R, L\}$.

- By chain rule of probability, we can factorize $p(Z, L^{(1)}, R, L)$ as follows:

$$\underbrace{\underbrace{p(Z, L^{(1)})}_{\text{target law}} \times \underbrace{p(R \mid L^{(1)}, Z)}_{\text{missingness mechanism}}}_{\text{full law } p(L^{(1)}, Z, R)} \times \underbrace{p(L \mid L^{(1)}, R, \cancel{Z})}_{\text{deterministic terms}}.$$

Missing data models

- A missing data model $\mathcal{M}$ is a set of distributions defined over variables in $\{Z, L^{(1)}, R, L\}$.

- By chain rule of probability, we can factorize $p(Z, L^{(1)}, R, L)$ as follows:

$$\underbrace{\underbrace{p(Z, L^{(1)})}_{\text{target law}} \times \underbrace{p(R \mid L^{(1)}, Z)}_{\text{missingness mechanism}}}_{\text{full law } p(L^{(1)}, Z, R)} \times \underbrace{p(L \mid L^{(1)}, R, \cancel{Z})}_{\text{deterministic terms}}.$$

- Consistency assumption: $L_k = \begin{cases} L_k^{(1)} & \text{if } R_k = 1 \\ ? & \text{if } R_k = 0 \end{cases}$

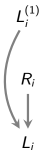
- Observed data law is $p(Z, R, L)$, where counterfactuals are marginalized out.

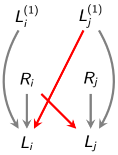
A missing data workflow

1. Define the **estimand** (often done in the absence of missing data).
   - A function of target law $p(Z, L^{(1)})$ or full law $p(Z, L^{(1)}, R)$.
2. Assume a **model** that links the counterfactual, factual, and missingness indicator variables.
   - Use **Directed Acyclic Graphs (DAGs)** to encode the modeling assumptions.
3. Determine whether the estimand is **identifiable** in the assumed model.
   - Focus on **identification of the target and full laws**.
4. If estimand is identifiable, find the best **estimation** strategy, and if it is not, perhaps stronger assumptions are needed (or alternatively obtaining bounds).
5. Conduct **sensitivity analysis** to reflect on the assumptions.

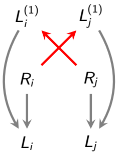
Introducing missing data DAGs

- Define missing data models via restrictions on the full data distribution that can be represented by a DAG (similar to causal inference).
- In missing data DAGs: (Mohan et al., 2013)
  1. Observed and counterfactual variables appear on the same graph
  2. There are certain edge restrictions: (marked in red)





deterministic edges (in gray)

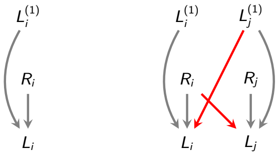


“no interference”

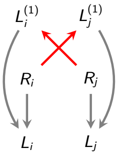
Introducing missing data DAGs

- Define missing data models via restrictions on the full data distribution that can be represented by a DAG (similar to causal inference).

- In missing data DAGs: (Mohan et al., 2013)
  1. Observed and counterfactual variables appear on the same graph
  2. There are certain edge restrictions: (marked in red)



deterministic edges (in gray)



"no interference"

- The "no interference" assumption can be relaxed (Srinivasan et al., 2023).

Missing data DAGs

- Proxy causes ($L_j \to R_i$) were not considered by Mohan et al. (2013), but were considered by Bhattacharya et al. (2019); Nabi et al. (2020).

"counterfactual cause" "proxy/observed cause" (a) (b)

As an example, assume variables in missing data DAGs (a) and (b) are binary:

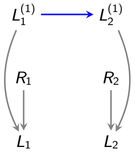
- (a):

$$\underbrace{p(R_i = 1 \mid L_j^{(1)}, R_j)}_{\text{4 parameters}}.$$

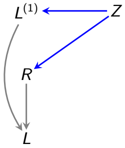
- (b):

$$\underbrace{p(R_i = 1 \mid L_j, R_j = 1)}_{\text{2 parameters}} \quad \text{and} \quad \underbrace{p(R_i = 1 \mid R_j = 0, L_j = ?)}_{\text{1 parameter}}.$$

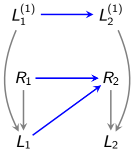
Examples: missing data DAGs



MCAR

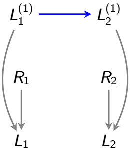


MAR

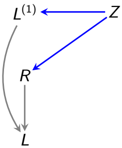


MAR

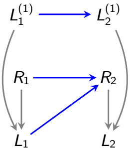
Examples: missing data DAGs



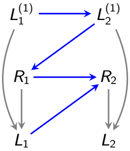
MCAR



MAR

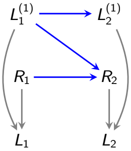


MAR



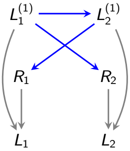
MNAR: permutation

(Robins, 1997)



MNAR: block-conditional

(Zhou et al., 2010)



MNAR: block-parallel

(Mohan et al., 2013)

Examples: missing data DAGs ctd.

How do models differ in telling a story about the missingness mechanisms?

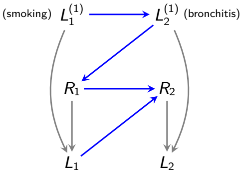
- $L_1^{(1)}$: true smoking status of an individual.
- $L_2^{(1)}$: diagnosis of bronchitis.
- R_1, R_2: encode whether these variables have been measured or not.

Examples: missing data DAGs ctd.

How do models differ in telling a story about the missingness mechanisms?

- $L_1^{(1)}$: true smoking status of an individual.
- $L_2^{(1)}$: diagnosis of bronchitis.
- R_1, R_2: encode whether these variables have been measured or not.

- $L_2^{(1)} \rightarrow R_1$
  Inquiry on patient's smoking status depends on prognosis of bronchitis.
- $R_1 \rightarrow R_2 \leftarrow L_1$
  Whether the true bronchitis status is measured via a diagnostic test depends on the doctor's awareness of the individual's smoking status (R_1) and their observed value of smoking (L_1).



MNAR: permutation

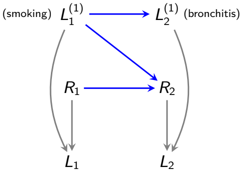
(Robins, 1997)

Examples: missing data DAGs ctd.

How do models differ in telling a story about the missingness mechanisms?

- $L_1^{(1)}$: true smoking status of an individual.
- $L_2^{(1)}$: diagnosis of bronchitis.
- R_1, R_2: encode whether these variables have been measured or not.

- R_1 has no parent
  Inquiry into smoking status is random (e.g., as in random screening programs or surveys).
- $R_1 \rightarrow R_2 \leftarrow L_1^{(1)}$
  Administration of a diagnostic test depends on the inquiry into smoking, as well as the potentially unobserved past history of smoking.



MNAR: block-conditional

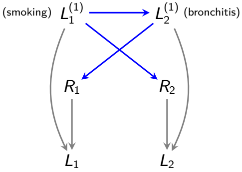
(Zhou et al., 2010)

Examples: missing data DAGs ctd.

How do models differ in telling a story about the missingness mechanisms?

- $L_1^{(1)}$: true smoking status of an individual.
- $L_2^{(1)}$: diagnosis of bronchitis.
- R_1, R_2: encode whether these variables have been measured or not.

- $R_1 \leftarrow L_2^{(1)}$
  Inquiry into smoking status depends on prognosis of bronchitis.
- $R_2 \leftarrow L_1^{(1)}$
  Administration of the diagnostic test depends on the suspected smoking status of an individual.



MNAR: block-parallel

(Mohan et al., 2013)

Missing data DAG models

- Denote the missing data DAG (m-DAG) defined over $V = (Z, L^{(1)}, R, L)$ via $\mathcal{G}(V)$.
- The statistical model of m-DAG $\mathcal{G}(V)$ is a set of distributions that factorize as:

$$\begin{aligned} p(Z, L^{(1)}, R, L) &= \prod_{V_i \in V} p(V_i \mid \text{pa}_{\mathcal{G}}(V_i)) \\ &= \prod_{V_i \in V \setminus L} p(V_i \mid \text{pa}_{\mathcal{G}}(V_i)) \times \prod_{L_i \in L} p(L_i \mid L_i^{(1)}, R_i). \end{aligned}$$

Missing data DAG models

- Denote the missing data DAG (m-DAG) defined over $V = (Z, L^{(1)}, R, L)$ via $\mathcal{G}(V)$.
- The statistical model of m-DAG $\mathcal{G}(V)$ is a set of distributions that factorize as:

$$\begin{aligned} p(Z, L^{(1)}, R, L) &= \prod_{V_i \in V} p(V_i \mid \text{pa}_{\mathcal{G}}(V_i)) \\ &= \prod_{V_i \in V \setminus L} p(V_i \mid \text{pa}_{\mathcal{G}}(V_i)) \times \prod_{L_i \in L} p(L_i \mid L_i^{(1)}, R_i). \end{aligned}$$

- Familiar concepts like d-separation and Markov properties carry over.
  - **Factorization**: probability distribution as a set of small factors.
  - **Local Markov property:** a small but complete set of indep constraints.

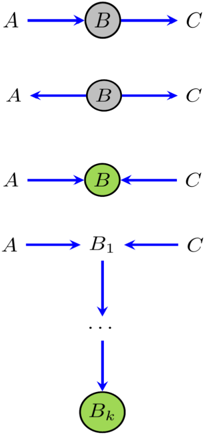
$$V_i \perp\!\!\!\perp \text{nd}_{\mathcal{G}}(V_i) \setminus \text{pa}_{\mathcal{G}}(V_i) \mid \text{pa}_{\mathcal{G}}(V_i), \ \forall V_i \in V.$$

  - **Global Markov property:** all independence constraints in the model.

$$\text{Given } X, Y, Z \in V: \quad (X \perp\!\!\!\perp_{\text{d-sep}} Y \mid Z)_{\mathcal{G}(V)} \implies (X \perp\!\!\!\perp Y \mid Z)_{p(V)}.$$

  - All three properties are equivalent.

d-separation refresher

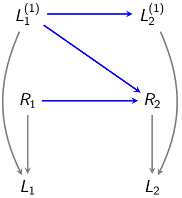


$A \perp\!\!\!\perp C \mid B$

$A \perp\!\!\!\perp C \mid B$

$A \not\perp\!\!\!\perp C \mid B$

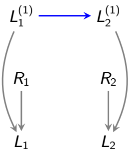
$A \not\perp\!\!\!\perp C \mid B_k$



$R_1 \perp\!\!\!\perp L_1^{(1)}$

$R_1 \not\perp\!\!\!\perp L_1^{(1)} \mid R_2$

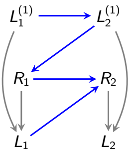
Examples: m-DAG models



$R_1 \perp\!\!\!\perp R_2, L_1^{(1)}, L_2^{(1)}$

$R_2 \perp\!\!\!\perp R_1, L_1^{(1)}, L_2^{(1)}$

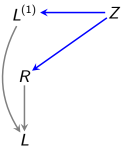
MCAR



$R_1 \perp\!\!\!\perp L_1^{(1)} \mid L_2^{(1)}$

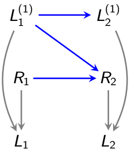
$R_2 \perp\!\!\!\perp L_1^{(1)}, L_2^{(1)} \mid R_1, L_1$

Permutation (Robins, 1997)



$R \perp\!\!\!\perp L^{(1)} \mid Z$

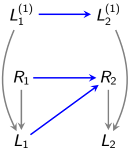
MAR



$R_1 \perp\!\!\!\perp L_1^{(1)}, L_2^{(1)}$

$R_2 \perp\!\!\!\perp L_1^{(1)} \mid R_1, L_1^{(1)}$

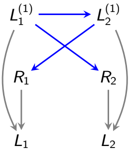
Block-conditional (Zhou et al., 2010)



$R_1 \perp\!\!\!\perp L_1^{(1)}, L_2^{(1)}$

$R_2 \perp\!\!\!\perp L_1^{(1)}, L_2^{(1)} \mid R_1, L_1$

MAR



$R_1 \perp\!\!\!\perp R_2, L_1^{(1)} \mid L_2^{(1)}$

$R_2 \perp\!\!\!\perp R_1, L_2^{(1)} \mid L_1^{(1)}$

Block-parallel (Mohan et al., 2013)

Graphical representations of MCAR, MAR, MNAR mechanisms

Missingness mechanism, $p(R \mid \text{pa}_{\mathcal{G}}(R)) = \prod_{R_k \in R} p(R_k \mid \text{pa}_{\mathcal{G}}(R_k))$, is

- **MCAR**: if $p(R_k \mid \text{pa}_{\mathcal{G}}(R_k))$, $\forall R_k \in R$, is not a function of variables in $\{L^{(1)}, L, Z, R\}$.
  - Graphically speaking, there are no edges that point to variables in R.
- **MAR**: if $p(R_k \mid \text{pa}_{\mathcal{G}}(R_k))$, $\forall R_k \in R$, is not a function of variables in $L^{(1)}$, but could be a function of variables in $\{Z, L, R\}$.
  - Graphically speaking, there are no edges from variables in $L^{(1)}$ to variables in R.
- **MNAR**: if there exists at least one $R_k \in R$ with parents in $L^{(1)}$.

Part 2. Nonparametric Identification

Identification in missing data models

- Let $\psi := \mathbb{E}[h(p(Z, L^{(1)}))]$ denote the parameter (estimand) of interest.
- Let the full law $p(Z, L^{(1)}, R)$ be Markov relative to an m-DAG $\mathcal{G}(V)$.
- To do *inference* on ψ, we first need to argue whether ψ is *identified* as a function of the observed data law in the assumed m-DAG or not?

Identification in missing data models

- Let $\psi := \mathbb{E}[h(p(Z, L^{(1)}))]$ denote the parameter (estimand) of interest.
- Let the full law $p(Z, L^{(1)}, R)$ be Markov relative to an m-DAG $\mathcal{G}(V)$.
- To do *inference* on ψ, we first need to argue whether ψ is *identified* as a function of the observed data law in the assumed m-DAG or not?
- The estimand ψ is identified in the assumed m-DAG $\mathcal{G}$, if it can be expressed as a unique function of the observed data law $p(Z, L, R)$. This means:
  - A parameter is identified under a particular collection of assumptions if these assumptions imply that the distribution of the **observed data is compatible with a single value** of the parameter.
  - If there exists no unique mapping between the counterfactual distribution and the observed data law, then the parameter is **not identified**.
- Instead of ψ, we might be interested in identification of the entire target law $p(Z, L^{(1)})$ or the entire full law $p(Z, L^{(1)}, R)$.

Example: simple identification arguments

Is the target law $p(Z, L^{(1)})$ identified as a function of the observed data law $p(Z, R, L)$?

- Under MCAR missingness, target law is identified:

$$\begin{aligned} p(Z, L^{(1)}) &= p(Z, L^{(1)} \mid R = 1) && \color{blue}{R \perp\!\!\!\perp Z, L^{(1)}} \\ &= p(Z, L \mid R = 1). && \color{blue}{\text{consistency}} \end{aligned}$$

- Under MAR missingness, target law is identified:

$$\begin{aligned} p(Z, L^{(1)}) &= p(Z) \times p(L^{(1)} \mid Z) \\ &= p(Z) \times p(L^{(1)} \mid Z, R = 1) && \color{blue}{R \perp\!\!\!\perp L^{(1)} \mid Z} \\ &= p(Z) \times p(L \mid Z, R = 1). && \color{blue}{\text{consistency}} \end{aligned}$$

- MNAR model:

$$p(Z, L^{(1)}) = ???$$

MNAR models $\equiv$ causal models with unmeasured confounding

- Sometimes we succeed and sometimes we fail to identify MNAR models!

Nonparametric identification theory in causal inference

- Identification questions in causal inference: given an arbitrary DAG with hidden/unmeasured variables, is $p(Y^{(t)})$ identified?
- **Sound** and **complete** algorithms exist for causal effect identification.
  - Soundness: functionals of identified effects are correct.
  - Completeness: no-identifiability of the causal effect is *provable*.

    (Shpitser and Pearl, 2006; Huang and Valtorta, 2006; Bhattacharya et al., 2022; Richardson et al., 2023)

Nonparametric identification theory in causal inference

- Identification questions in causal inference: given an arbitrary DAG with hidden/unmeasured variables, is $p(Y^{(t)})$ identified?
- **Sound** and **complete** algorithms exist for causal effect identification.
  - Soundness: functionals of identified effects are correct.
  - Completeness: no-identifiability of the causal effect is *provable*.

    (Shpitser and Pearl, 2006; Huang and Valtorta, 2006; Bhattacharya et al., 2022; Richardson et al., 2023)
- Similarly, given that assumptions/restrictions in a missing data model are encoded via an m-DAG, we would like to know whether the underlying full/target law is identified or not.
  - Can identification theory in causal DAGs be used to reason about identification in m-DAGs, given their similarities?

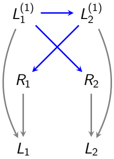
Nonparametric identification theory in causal inference

- Identification questions in causal inference: given an arbitrary DAG with hidden/unmeasured variables, is $p(Y^{(t)})$ identified?
- **Sound** and **complete** algorithms exist for causal effect identification.
  - Soundness: functionals of identified effects are correct.
  - Completeness: no-identifiability of the causal effect is *provable*.

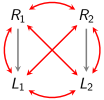
    (Shpitser and Pearl, 2006; Huang and Valtorta, 2006; Bhattacharya et al., 2022; Richardson et al., 2023)
- Similarly, given that assumptions/restrictions in a missing data model are encoded via an m-DAG, we would like to know whether the underlying full/target law is identified or not.
  - Can identification theory in causal DAGs be used to reason about identification in m-DAGs, given their similarities?
- **Causal identification theory is *incomplete* for missing data identification!**
  - There are identified MNAR models where causal identification theory fails.

Incompleteness of causal identification theory for m-DAGs

- Causal identification theory is **incomplete** for missing data identification.



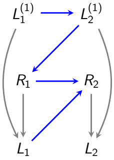
Block-parallel



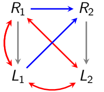
Observed margin

One-line ID (Richardson et al., 2023)

- $Y^* = \{L_1, L_2\}$
- $G_{Y^*} = L_1 \leftrightarrow L_2$
- District: $\{L_1, L_2\}$
- Need to fix R_1, R_2 and fail.



Permutation



Observed margin

One-line ID (Richardson et al., 2023)

- $Y^* = \{L_1, L_2\}$
- $G_{Y^*} = L_1 \leftrightarrow L_2$
- District: $\{L_1, L_2\}$
- Need to fix R_1, R_2 and fail.

Nonparametric identification in m-DAGs

- The target law is identified **if and only if** the missingness mechanism $p(R = 1 \mid L^{(1)}, Z)$ is identified. Using Bayes rule:

$$p(R = 1 \mid L^{(1)}, Z) = \frac{p(Z, L^{(1)}, R = 1)}{p(L^{(1)}, Z)} \quad \rightarrow \quad p(Z, L^{(1)}) = \frac{p(Z, L^{(1)}, R = 1)}{p(R = 1 \mid Z, L^{(1)})}.$$

Nonparametric identification in m-DAGs

- The target law is identified **if and only if** the missingness mechanism $p(R=1 \mid L^{(1)}, Z)$ is identified. Using Bayes rule:

$$p(R=1 \mid L^{(1)}, Z) = \frac{p(Z, L^{(1)}, R=1)}{p(L^{(1)}, Z)} \quad \to \quad p(Z, L^{(1)}) = \frac{p(Z, L^{(1)}, R=1)}{p(R=1 \mid Z, L^{(1)})}.$$

- The full law is identified **if and only if** the missingness mechanism $p(R=r \mid L^{(1)}, Z)$ is identified, for all possible missingness pattern $R=r$. Using chain rule:

$$\begin{aligned} p(Z, L^{(1)}, R=r) &= p(Z, L^{(1)}) \times p(R=r \mid L^{(1)}, Z) \\ &= \frac{p(Z, L^{(1)}, R=1)}{p(R=1 \mid L^{(1)}, Z)} \times p(R=r \mid L^{(1)}, Z). \end{aligned}$$

Nonparametric identification in m-DAGs

- The target law is identified **if and only if** the missingness mechanism $p(R = 1 \mid L^{(1)}, Z)$ is identified. Using Bayes rule:

$$p(R = 1 \mid L^{(1)}, Z) = \frac{p(Z, L^{(1)}, R = 1)}{p(L^{(1)}, Z)} \quad \rightarrow \quad p(Z, L^{(1)}) = \frac{p(Z, L^{(1)}, R = 1)}{p(R = 1 \mid Z, L^{(1)})}.$$

- The full law is identified **if and only if** the missingness mechanism $p(R = r \mid L^{(1)}, Z)$ is identified, for all possible missingness pattern $R = r$. Using chain rule:

$$\begin{aligned} p(Z, L^{(1)}, R = r) &= p(Z, L^{(1)}) \times p(R = r \mid L^{(1)}, Z) \\ &= \frac{p(Z, L^{(1)}, R = 1)}{p(R = 1 \mid L^{(1)}, Z)} \times p(R = r \mid L^{(1)}, Z). \end{aligned}$$

- Game plan: focus on identification of the missingness mechanism $p(R \mid L^{(1)}, Z)$ in a given m-DAG.

Np-identification of missingness mechanisms in m-DAGs

Given an m-DAG, is the missingness mechanism identified or not? We look at two different parameterizations of $p(R \mid L^{(1)}, Z)$:

Np-identification of missingness mechanisms in m-DAGs

Given an m-DAG, is the missingness mechanism identified or not? We look at two different parameterizations of $p(R \mid L^{(1)}, Z)$:

(i) m-DAG factorization: (Bhattacharya et al., 2019)

$$p(R \mid \text{pa}_{\mathcal{G}}(R)) = \prod_{R_k \in R} p(R_k \mid \text{pa}_{\mathcal{G}}(R_k))$$

Identify each **propensity score** $p(R_k \mid \text{pa}_{\mathcal{G}}(R_k))$, for all $R_k \in R$.

Np-identification of missingness mechanisms in m-DAGs

Given an m-DAG, is the missingness mechanism identified or not? We look at two different parameterizations of $p(R \mid L^{(1)}, Z)$:

(i) m-DAG factorization: (Bhattacharya et al., 2019)

$$p(R \mid \text{pa}_{\mathcal{G}}(R)) = \prod_{R_k \in R} p(R_k \mid \text{pa}_{\mathcal{G}}(R_k))$$

Identify each **propensity score** $p(R_k \mid \text{pa}_{\mathcal{G}}(R_k))$, for all $R_k \in R$.

(ii) Odds ratio parameterization: (Chen, 2007; Nabi et al., 2020)

$$\prod_{k=1}^{K} p(R_k \mid R_{-k} = 1, \text{pa}_{\mathcal{G}}(R)) \times \prod_{k=2}^{K} \text{OR}(R_k, R_{\prec k} \mid R_{\succ k} = 1, \text{pa}_{\mathcal{G}}(R)),$$

where $R_{-k} = R \setminus R_k, R_{\prec k} = \{R_1, \ldots, R_{k-1}\}, R_{\succ k} = \{R_{k+1}, \ldots, R_K\}$.

Identify each **univariate conditionals** and **pairwise odds ratio** terms.

Identification arguments

1. m-DAG factorization of the missingness mechanism

2. Odds ratio parameterization of the missingness mechanism

Identification via m-DAG factorization

- Whether and how full/target law is identified in a given m-DAG $\mathcal{G}$.
  - Target law: argue for identification of $p(R = 1 \mid \text{pa}_{\mathcal{G}}(R))$.
  - Full law: argue for identification of $p(R = r \mid \text{pa}_{\mathcal{G}}(R)), \ \forall r \in \{0, 1\}^K$.

Identification via m-DAG factorization

- Whether and how full/target law is identified in a given m-DAG $\mathcal{G}$.
  - Target law: argue for identification of $p(R = 1 \mid \text{pa}_{\mathcal{G}}(R))$.
  - Full law: argue for identification of $p(R = r \mid \text{pa}_{\mathcal{G}}(R)),\ \forall r \in \{0, 1\}^K$.

- Target law is identified if each propensity score $p(R_k \mid \text{pa}_{\mathcal{G}}(R_k))$ is identified only when $R_i = 1$ for $R_i \in \text{pa}_{\mathcal{G}}(R_k)$,

$$p(R = 1 \mid \text{pa}_{\mathcal{G}}(R)) = \prod_{R_k \in R} p(R_k \mid \text{pa}_{\mathcal{G}}(R_k))\Big|_{R=1}.$$

Identification via m-DAG factorization

- Whether and how full/target law is identified in a given m-DAG $\mathcal{G}$.
  - Target law: argue for identification of $p(R = 1 \mid \text{pa}_{\mathcal{G}}(R))$.
  - Full law: argue for identification of $p(R = r \mid \text{pa}_{\mathcal{G}}(R)), \ \forall r \in \{0, 1\}^K$.

- Target law is identified if each propensity score $p(R_k \mid \text{pa}_{\mathcal{G}}(R_k))$ is identified only when $R_i = 1$ for $R_i \in \text{pa}_{\mathcal{G}}(R_k)$,

$$p(R = 1 \mid \text{pa}_{\mathcal{G}}(R)) = \prod_{R_k \in R} p(R_k \mid \text{pa}_{\mathcal{G}}(R_k))\Big|_{R=1}.$$

- Full law is identified if each propensity score $p(R_k \mid \text{pa}_{\mathcal{G}}(R_k))$ is identified at all levels of $R_i \in \text{pa}_G(R_k)$,

$$p(R = r \mid \text{pa}_{\mathcal{G}}(R)) = \prod_{R_k \in R} p(R_k \mid \text{pa}_{\mathcal{G}}(R_k))\Big|_{R=r}$$

Identification via m-DAG factorization

- Whether and how full/target law is identified in a given m-DAG $\mathcal{G}$.
  - Target law: argue for identification of $p(R = 1 \mid \mathrm{pa}_{\mathcal{G}}(R))$.
  - Full law: argue for identification of $p(R = r \mid \mathrm{pa}_{\mathcal{G}}(R))$, $\forall r \in \{0, 1\}^K$.

- Target law is identified if each propensity score $p(R_k \mid \mathrm{pa}_{\mathcal{G}}(R_k))$ is identified only when $R_i = 1$ for $R_i \in \mathrm{pa}_{\mathcal{G}}(R_k)$,

$$p(R = 1 \mid \mathrm{pa}_{\mathcal{G}}(R)) = \prod_{R_k \in R} p(R_k \mid \mathrm{pa}_{\mathcal{G}}(R_k))\Big|_{R=1}.$$

- Full law is identified if each propensity score $p(R_k \mid \mathrm{pa}_{\mathcal{G}}(R_k))$ is identified at all levels of $R_i \in \mathrm{pa}_{G}(R_k)$,

$$p(R = r \mid \mathrm{pa}_{\mathcal{G}}(R)) = \prod_{R_k \in R} p(R_k \mid \mathrm{pa}_{\mathcal{G}}(R_k))\Big|_{R=r}$$

- There are two major ideas for propensity scores identification:
  1. *Associational* irrelevancy: d-separation
  2. *Causal* irrelevancy: invariance property

1. Associational irrelevancy

- In order to identify the propensity score of R_k, $p(R_k \mid \text{pa}_{\mathcal{G}}(R_k))$, we need to **select on** the following missingness indicators:

$$R_k^s = \left\{ R_i \in R \mid L_i^{(1)} \in \text{pa}_{\mathcal{G}}(R_k) \right\} \qquad \text{(selection set for } R_k\text{)}$$

1. Associational irrelevancy

- In order to identify the propensity score of R_k, $p(R_k \mid \mathrm{pa}_{\mathcal{G}}(R_k))$, we need to **select on** the following missingness indicators:

$$R_k^s = \left\{ R_i \in R \mid L_i^{(1)} \in \mathrm{pa}_{\mathcal{G}}(R_k) \right\} \qquad \text{(selection set for } R_k\text{)}$$

- For any $R_i \in R_k^s$, R_i is either a **descendant** of R_k or a **non-descendant** of R_k.

- If R_i is a **non-descendant** of R_k, then we can apply the local Markov property which states $R_k \perp\!\!\!\perp \mathrm{nd}_{\mathcal{G}}(R_k) \setminus \mathrm{pa}_{\mathcal{G}}(R_k) \mid \mathrm{pa}_{\mathcal{G}}(R_k)$.
  - So we can include $R_i = 1$ in the conditioning set and replace $L_i^{(1)}$ with $\{L_i, R_i = 1\}$.

1. Associational irrelevancy

- In order to identify the propensity score of R_k, $p(R_k \mid \text{pa}_{\mathcal{G}}(R_k))$, we need to **select on** the following missingness indicators:

$$R_k^s = \left\{ R_i \in R \mid L_i^{(1)} \in \text{pa}_{\mathcal{G}}(R_k) \right\} \qquad \text{(selection set for } R_k\text{)}$$

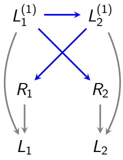
- For any $R_i \in R_k^s$, R_i is either a **descendant** of R_k or a **non-descendant** of R_k.

- If R_i is a **non-descendant** of R_k, then we can apply the local Markov property which states $R_k \perp\!\!\!\perp \text{nd}_{\mathcal{G}}(R_k) \setminus \text{pa}_{\mathcal{G}}(R_k) \mid \text{pa}_{\mathcal{G}}(R_k)$.
  - So we can include $R_i = 1$ in the conditioning set and replace $L_i^{(1)}$ with $\{L_i, R_i = 1\}$.

- Formally, for any $R_i \in R_k^s \cap \text{nd}_{\mathcal{G}}(R_k)$, we can write:

$$p(R_k \mid \text{pa}_{\mathcal{G}}(R_k))\Big|_{R=1} = p(R_k \mid \underbrace{\text{pa}_{\mathcal{G}}(R_k)}_{\text{includes } L_i^{(1)}}, R_i = 1)\Big|_{R=1}.$$

Example 1/2: associational irrelevancy (block-parallel)



- Is the full/target law identified in the **block-parallel** model?

$$p(R \mid \mathrm{pa}_{\mathcal{G}}(R)) = p(R_1 \mid \mathrm{pa}_{\mathcal{G}}(R_1)) \times p(R_2 \mid \mathrm{pa}_{\mathcal{G}}(R_2))$$

Example 1/2: associational irrelevancy (block-parallel)

$$\begin{tikzcd}
L_1^{(1)} \arrow[r] \arrow[dr] \arrow[dd, bend right] & L_2^{(1)} \arrow[dl] \arrow[dd, bend left] \\
R_1 \arrow[d] & R_2 \arrow[d] \\
L_1 & L_2
\end{tikzcd}$$

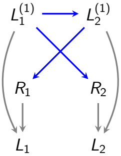
- Is the full/target law identified in the **block-parallel** model?

$$p(R \mid \mathrm{pa}_{\mathcal{G}}(R)) = p(R_1 \mid \mathrm{pa}_{\mathcal{G}}(R_1)) \times p(R_2 \mid \mathrm{pa}_{\mathcal{G}}(R_2))$$

Identification of $p(R_1 \mid \mathrm{pa}_{\mathcal{G}}(R_1))$

$$\begin{aligned}
p(R_1 \mid \mathrm{pa}_{\mathcal{G}}(R_1)) &= p(R_1 \mid L_2^{(1)}) \\
&= p(R_1 \mid R_2 = 1, L_2^{(1)}) \qquad R_1 \perp\!\!\!\perp R_2 \mid L_2^{(1)} \\
&= p(R_1 \mid R_2 = 1, L_2) \qquad \text{consistency}
\end{aligned}$$

Example 1/2: associational irrelevancy (block-parallel)



- Is the full/target law identified in the **block-parallel** model?

$$p(R \mid \mathrm{pa}_{\mathcal{G}}(R)) = p(R_1 \mid \mathrm{pa}_{\mathcal{G}}(R_1)) \times p(R_2 \mid \mathrm{pa}_{\mathcal{G}}(R_2))$$

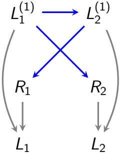
Identification of $p(R_1 \mid \mathrm{pa}_{\mathcal{G}}(R_1))$

$$\begin{aligned} p(R_1 \mid \mathrm{pa}_{\mathcal{G}}(R_1)) &= p(R_1 \mid L_2^{(1)}) \\ &= p(R_1 \mid R_2 = 1, L_2^{(1)}) && R_1 \perp\!\!\!\perp R_2 \mid L_2^{(1)} \\ &= p(R_1 \mid R_2 = 1, L_2) && \text{consistency} \end{aligned}$$

Identification of $p(R_2 \mid \mathrm{pa}_{\mathcal{G}}(R_2))$

$$\begin{aligned} p(R_2 \mid \mathrm{pa}_{\mathcal{G}}(R_2)) &= p(R_2 \mid L_1^{(1)}) \\ &= p(R_2 \mid R_1 = 1, L_1^{(1)}) && R_2 \perp\!\!\!\perp R_1 \mid L_1^{(1)} \\ &= p(R_2 \mid R_1 = 1, L_1) && \text{consistency} \end{aligned}$$

Example 1/2: associational irrelevancy (block-parallel)



- Is the full/target law identified in the **block-parallel** model?

$$p(R \mid \text{pa}_{\mathcal{G}}(R)) = p(R_1 \mid \text{pa}_{\mathcal{G}}(R_1)) \times p(R_2 \mid \text{pa}_{\mathcal{G}}(R_2))$$

Identification of $p(R_1 \mid \text{pa}_{\mathcal{G}}(R_1))$

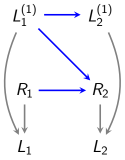
$$\begin{aligned} p(R_1 \mid \text{pa}_{\mathcal{G}}(R_1)) &= p(R_1 \mid L_2^{(1)}) \\ &= p(R_1 \mid R_2 = 1, L_2^{(1)}) && R_1 \perp\!\!\!\perp R_2 \mid L_2^{(1)} \\ &= p(R_1 \mid R_2 = 1, L_2) && \text{consistency} \end{aligned}$$

Identification of $p(R_2 \mid \text{pa}_{\mathcal{G}}(R_2))$

$$\begin{aligned} p(R_2 \mid \text{pa}_{\mathcal{G}}(R_2)) &= p(R_2 \mid L_1^{(1)}) \\ &= p(R_2 \mid R_1 = 1, L_1^{(1)}) && R_2 \perp\!\!\!\perp R_1 \mid L_1^{(1)} \\ &= p(R_2 \mid R_1 = 1, L_1) && \text{consistency} \end{aligned}$$

- So $p(R \mid \text{pa}_{\mathcal{G}}(R))$ is ID, which means the full and target laws are both ID.

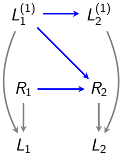
Example 2/2: associational irrelevancy (block-conditional MAR)



- Is the full/target law identified in the **block-conditional MAR** model?

$$p(R \mid \mathrm{pa}_{\mathcal{G}}(R)) = p(R_1 \mid \mathrm{pa}_{\mathcal{G}}(R_1)) \times p(R_2 \mid \mathrm{pa}_{\mathcal{G}}(R_2))$$

Example 2/2: associational irrelevancy (block-conditional MAR)



- Is the full/target law identified in the **block-conditional MAR** model?

$$p(R \mid \mathrm{pa}_{\mathcal{G}}(R)) = p(R_1 \mid \mathrm{pa}_{\mathcal{G}}(R_1)) \times p(R_2 \mid \mathrm{pa}_{\mathcal{G}}(R_2))$$

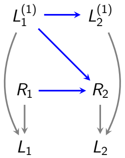
- $p(R_1 \mid \mathrm{pa}_{\mathcal{G}}(R_1)) = p(R_1)$ is identified.

Identification of $p(R_2 \mid \mathrm{pa}_{\mathcal{G}}(R_2))$

$$\begin{aligned} p(R_2 \mid \mathrm{pa}_{\mathcal{G}}(R_2)) &= p(R_2 \mid R_1, L_1^{(1)}) \\ &= ??? \end{aligned}$$

$$\begin{aligned} p(R_2 \mid \mathrm{pa}_{\mathcal{G}}(R_2))\Big|_{R=1} &= p(R_2 = 1 \mid R_1 = 1, L_1^{(1)}) \\ &= p(R_2 = 1 \mid R_1 = 1, L_1) \qquad \text{consistency} \end{aligned}$$

Example 2/2: associational irrelevancy (block-conditional MAR)



- Is the full/target law identified in the **block-conditional MAR** model?

$$p(R \mid \text{pa}_{\mathcal{G}}(R)) = p(R_1 \mid \text{pa}_{\mathcal{G}}(R_1)) \times p(R_2 \mid \text{pa}_{\mathcal{G}}(R_2))$$

- $p(R_1 \mid \text{pa}_{\mathcal{G}}(R_1)) = p(R_1)$ is identified.

Identification of $p(R_2 \mid \text{pa}_{\mathcal{G}}(R_2))$

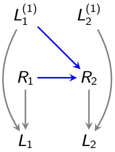
$$\begin{aligned} p(R_2 \mid \text{pa}_{\mathcal{G}}(R_2)) &= p(R_2 \mid R_1, L_1^{(1)}) \\ &= ??? \end{aligned}$$

$$\begin{aligned} p(R_2 \mid \text{pa}_{\mathcal{G}}(R_2))\Big|_{R=1} &= p(R_2 = 1 \mid R_1 = 1, L_1^{(1)}) \\ &= p(R_2 = 1 \mid R_1 = 1, L_1) \qquad \text{consistency} \end{aligned}$$

- Only $p(R_2 \mid R_1 = 1, L_1^{(1)})$ is identified, so the target law is certainly ID.
- It seems the full law might **NOT** be identified. We have to prove that the full law is not identified.

A non-identified structure

- Claim: $p(R_2 = r_2 \mid R_1 = 0, L_1^{(1)})$ is not ID.



- Assume binary data. So the full law $p(R, L^{(1)})$ has 7 parameters.
- 5 identified parameters:
  - $\alpha_1 : \ p(L_1^{(1)} = 1) = p(L_1^{(1)} = 1 \mid R_1 = 1) = p(L_1 = 1 \mid R_1 = 1)$
  - $\alpha_2 : \ p(L_2^{(1)} = 1) = p(L_2^{(1)} = 1 \mid R_2 = 1) = p(L_2 = 1 \mid R_2 = 1)$
  - $\alpha_3 : \ p(R_1 = 1)$
  - $\alpha_{4,5} : \ p(R_2 = 1 \mid R_1 = 1, L_1^{(1)} = l_1) = p(R_2 = 1 \mid R_1 = 1, L_1 = l_1)$.
- 2 unidentified parameters:
  - $\alpha_{6,7} : \ p(R_2 = 1 \mid R_1 = 0, L_1^{(1)} = l_1), \ l_1 \in \{0, 1\}$

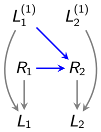
Proving non-identifiability claims

R_1	$p(R_1)$
0	a
1	$1-a$

$L_1^{(1)}$	$p(L_1^{(1)})$
0	b
1	$1-b$

$L_2^{(1)}$	$p(L_2^{(1)})$
0	c
1	c

R_2	R_1	$L_1^{(1)}$	$p(R_2 \mid R_1, L_1^{(1)})$
0	0	0	d
1	0	0	$1-d$
0	1	0	e
1	1	0	$1-e$
0	0	1	f
1	0	1	$1-f$
0	1	1	g
1	1	1	$1-g$



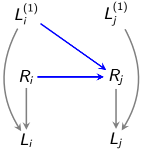
<table>
<tr><th>R_1</th><th>R_2</th><th>$L_1^{(1)}$</th><th>$L_2^{(1)}$</th><th>p(FULL LAW)</th><th>L_1</th><th>L_2</th><th>p(OBSERVED LAW)</th></tr>
<tr><td rowspan="4">0</td><td rowspan="4">0</td><td>0</td><td>0</td><td>$abcd$</td><td rowspan="4">?</td><td rowspan="4">?</td><td rowspan="4">$a\left[db+f(1-b)\right]$</td></tr>
<tr><td>1</td><td>0</td><td>$af(1-b)c$</td></tr>
<tr><td>0</td><td>1</td><td>$adb(1-c)$</td></tr>
<tr><td>1</td><td>1</td><td>$af(1-b)(1-c)$</td></tr>
<tr><td rowspan="4">1</td><td rowspan="4">0</td><td>0</td><td>0</td><td>$(1-a)ebc$</td><td rowspan="2">0</td><td rowspan="4">?</td><td rowspan="2">$(1-a)eb$</td></tr>
<tr><td>1</td><td>0</td><td>$(1-a)g(1-b)c$</td></tr>
<tr><td>0</td><td>1</td><td>$(1-a)eb(1-c)$</td><td rowspan="2">1</td><td rowspan="2">$(1-a)g(1-b)$</td></tr>
<tr><td>1</td><td>1</td><td>$(1-a)g(1-b)(1-c)$</td></tr>
<tr><td rowspan="4">0</td><td rowspan="4">1</td><td>0</td><td>0</td><td>$a(1-d)bc$</td><td rowspan="4">?</td><td rowspan="2">0</td><td rowspan="2">$ac\left[1-\left(db+f(1-b)\right)\right]$</td></tr>
<tr><td>1</td><td>0</td><td>$a(1-f)(1-b)c$</td></tr>
<tr><td>0</td><td>1</td><td>$a(1-d)b(1-c)$</td><td rowspan="2">1</td><td rowspan="2">$a(1-c)\left[1-\left(db+f(1-b)\right)\right]$</td></tr>
<tr><td>1</td><td>1</td><td>$a(1-f)(1-b)(1-c)$</td></tr>
<tr><td rowspan="4">1</td><td rowspan="4">1</td><td>0</td><td>0</td><td>$(1-a)(1-e)bc$</td><td>0</td><td>0</td><td>$(1-a)(1-e)bc$</td></tr>
<tr><td>1</td><td>0</td><td>$(1-a)(1-g)(1-b)c$</td><td>1</td><td>0</td><td>$(1-a)(1-g)(1-b)c$</td></tr>
<tr><td>0</td><td>1</td><td>$(1-a)(1-e)b(1-c)$</td><td>0</td><td>0</td><td>$(1-a)(1-e)b(1-c)$</td></tr>
<tr><td>1</td><td>1</td><td>$(1-a)(1-g)(1-b)(1-c)$</td><td>1</td><td>1</td><td>$(1-a)(1-g)(1-b)(1-c)$</td></tr>
</table>

We can pick any $\{d, f\}$ as long as $bd + (1-b)f$ stays the same.

Colluder: non-identified structure

Definition (Colluder)

If $\exists\ R_i, R_j \in R$ such that $R_i \rightarrow R_j \leftarrow L_i^{(1)}$, then a special collider structure forms at R_j, referred to as colluder.

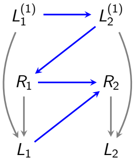


Lemma (Colluder non-identification)

If $\exists\ R_i, R_j \in R$ such that $R_i \rightarrow R_j \leftarrow L_i^{(1)}$ then $p(R_j \mid \mathrm{pa}_{\mathcal{G}}(R_j) \setminus R_i, R_i = 0)$ is not identified (Bhattacharya et al., 2019).

The above result means that whenever we spot a colluder in an m-DAG, we can immediately conclude the underlying full law is not identified.

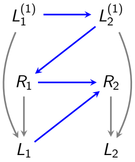
Associational irrelevancy: limitations



- Is the full/target law identified in the **permutation** model?

$$p(R \mid L^{(1)}) = p(R_1 \mid \mathrm{pa}_{\mathcal{G}}(R_1)) \times p(R_2 \mid \mathrm{pa}_{\mathcal{G}}(R_2))$$

Associational irrelevancy: limitations



- Is the full/target law identified in the **permutation** model?

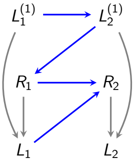
$$p(R \mid L^{(1)}) = p(R_1 \mid \text{pa}_{\mathcal{G}}(R_1)) \times p(R_2 \mid \text{pa}_{\mathcal{G}}(R_2))$$

Identification of $p(R_i \mid \text{pa}_{\mathcal{G}}(R_i))$

$p(R_2 \mid \text{pa}_{\mathcal{G}}(R_2)) = p(R_2 \mid R_1, L_1)$ ✓

$p(R_1 \mid \text{pa}_{\mathcal{G}}(R_1)) = p(R_1 \mid L_2^{(1)})$ $R_1 \not\perp\!\!\!\perp R_2 \mid L_2^{(1)}$

Associational irrelevancy: limitations



- Is the full/target law identified in the **permutation** model?

$$p(R \mid L^{(1)}) = p(R_1 \mid \text{pa}_{\mathcal{G}}(R_1)) \times p(R_2 \mid \text{pa}_{\mathcal{G}}(R_2))$$

Identification of $p(R_i \mid \text{pa}_{\mathcal{G}}(R_i))$

$p(R_2 \mid \text{pa}_{\mathcal{G}}(R_2)) = p(R_2 \mid R_1, L_1)$ ✓

$p(R_1 \mid \text{pa}_{\mathcal{G}}(R_1)) = p(R_1 \mid L_2^{(1)})$ $R_1 \not\perp\!\!\!\perp R_2 \mid L_2^{(1)}$

- What does this mean? Is the propensity score of R_1 not identified? which would then imply the full law is not ID.
- To answer this question, we either need to prove the propensity score is not identified or find a way to identify it.

2. Causal irrelevancy

Invariance property:

- Given the propensity score for $R_k \in R$, the conditioning set $\mathrm{pa}_{\mathcal{G}}(R_k)$ captures all the direct causes of R_k. Hence, it remains invariant to any set of interventions that disrupts other parts of the full law.

- Formally, given $R^* \subseteq R \setminus R_k$, we have

$$p\left(R_k \mid \mathrm{pa}_{\mathcal{G}}(R_k)\right) = p\left(R_k \mid \mathrm{pa}_{\mathcal{G}}(R_k), \mathrm{do}(R^* = 1)\right).$$

2. Causal irrelevancy

Invariance property:

- Given the propensity score for $R_k \in R$, the conditioning set $\text{pa}_{\mathcal{G}}(R_k)$ captures all the direct causes of R_k. Hence, it remains invariant to any set of interventions that disrupts other parts of the full law.

- Formally, given $R^* \subseteq R \setminus R_k$, we have

$$p\left(R_k \mid \text{pa}_{\mathcal{G}}(R_k)\right) = p\left(R_k \mid \text{pa}_{\mathcal{G}}(R_k), \text{do}(R^* = 1)\right).$$

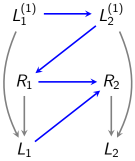
- Due to *invariance* property of the propensity scores, we can sometimes succeed in identifying a propensity score by exploring interventional distributions where a subset of variables are intervened on.

Intervention operation on missingness indicators

An intervention that sets $R_k = 1$, entails the following changes:

- **Graphical** changes to the missing data DAG $\mathcal{G}(V)$
  - Delete all the incoming edges into R_k and fix R_k to take value 1, and
  - Equate the counterfactual variable $L_k^{(1)}$ to L_k (by consistency).
- **Probabilistical** changes to the joint distribution $p(V)$
  - Drop the propensity score $p(R_k \mid \text{pa}_{\mathcal{G}}(R_k))$ from the joint factorization, and evaluate the truncated factorization at $R_k = 1$.

Example: causal irrelevancy



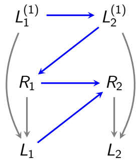
Permutation $\mathcal{G}_0$

$$\mathcal{G}_0 : \quad R_1 \not\perp\!\!\!\perp R_2 \mid L_2^{(1)}$$

Invariance property:

$$p(R_1 \mid L_2^{(1)}) = p(R_1 \mid L_2^{(1)}, \mathrm{do}(R_2 = 1))$$

Example: causal irrelevancy



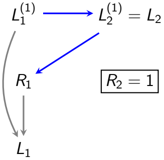
Permutation $\mathcal{G}_0$

$$\mathcal{G}_0 : \quad R_1 \not\perp\!\!\!\perp R_2 \mid L_2^{(1)}$$

Invariance property:

$$p(R_1 \mid L_2^{(1)}) = p(R_1 \mid L_2^{(1)}, \mathrm{do}(R_2 = 1))$$

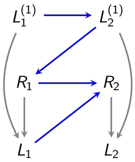
Graphical and probabilistical changes after intervening on R_2:



Permutation $\mathcal{G}_1$

$$p(R_1, L_1^{(1)}, L_2^{(1)}, L_1 \mid \mathrm{do}(R_2 = 1)) = \left. \frac{p(R_1, L_1^{(1)}, L_2^{(1)}, L_1, R_2)}{p(R_2 \mid R_1, L_1)} \right|_{R_2=1}$$

Example: causal irrelevancy



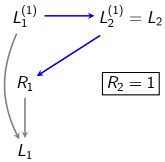
Permutation $\mathcal{G}_0$

$$\mathcal{G}_0 : \quad R_1 \not\perp\!\!\!\perp R_2 \mid L_2^{(1)}$$

Invariance property:

$$p(R_1 \mid L_2^{(1)}) = p(R_1 \mid L_2^{(1)}, \mathrm{do}(R_2 = 1))$$

Graphical and probabilistical changes after intervening on R_2:

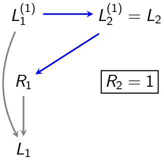


Permutation $\mathcal{G}_1$

$$p(R_1, L_1^{(1)}, L_2^{(1)}, L_1 \mid \mathrm{do}(R_2 = 1)) = \left. \frac{p(R_1, L_1^{(1)}, L_2^{(1)}, L_1, R_2)}{p(R_2 \mid R_1, L_1)} \right|_{R_2=1}$$

- The propensity score of R_1 is identified from the above intervention dist.

Example ctd. causal irrelevancy



Permutation $\mathcal{G}_1$

$$p(R_1, L_1^{(1)}, L_2^{(1)}, L_1 \mid \mathrm{do}(R_2 = 1)) = \frac{p(R_1, L_1^{(1)}, L_2, L_1, R_2 = 1)}{p(R_2 = 1 \mid R_1, L_1)}$$

$$p(R_1, L_2^{(1)} \mid \mathrm{do}(R_2 = 1)) = \sum_{l_1} \frac{p(R_1, l_1, L_2, R_2 = 1)}{p(R_2 = 1 \mid R_1, l_1)}.$$

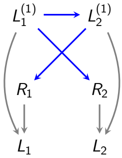
- The propensity score of R_1 is identified from the above intervention dist.

$$\begin{aligned} p(R_1 \mid L_2^{(1)}) &= p(R_1 \mid L_2^{(1)}, \mathrm{do}(R_2 = 1)) \\ &= \frac{p(R_1, L_2^{(1)} \mid \mathrm{do}(R_2 = 1))}{p(L_2^{(1)} \mid \mathrm{do}(R_2 = 1))}. \end{aligned}$$

First equality holds by the invariance property and second holds by Bayes rule.

Order of interventions

- Target law is ID via **parallel/simultaneous interventions** on R_1 and R_2.

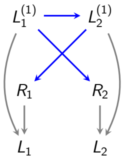


Block-parallel

$$
\begin{aligned}
p(L^{(1)}) &= \frac{p(L^{(1)}, R)}{p(R_1 \mid \mathrm{pa}_{\mathcal{G}}(R_1)) \times p(R_2 \mid \mathrm{pa}_{\mathcal{G}}(R_2))} \Bigg|_{R=1} \\
&= \frac{p(L, R=1)}{p(R_1 = 1 \mid R_2 = 1, L_2) \times p(R_2 = 1 \mid R_1 = 1, L_1)}.
\end{aligned}
$$

Order of interventions

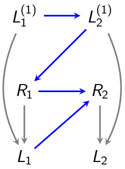
- Target law is ID via **parallel/simultaneous interventions** on R_1 and R_2.



Block-parallel

$$p(L^{(1)}) = \frac{p(L^{(1)}, R)}{p(R_1 \mid \text{pa}_{\mathcal{G}}(R_1)) \times p(R_2 \mid \text{pa}_{\mathcal{G}}(R_2))} \Bigg|_{R=1}$$
$$= \frac{p(L, R=1)}{p(R_1 = 1 \mid R_2 = 1, L_2) \times p(R_2 = 1 \mid R_1 = 1, L_1)}.$$

- Target law ID is obtained via **sequential interventions** on first R_2 and then R_1.

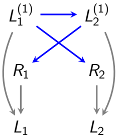


Permutation

$$p(L^{(1)}) = \frac{p(L^{(1)}, R)}{p(R_1 \mid \text{pa}_{\mathcal{G}}(R_1)) \times p(R_2 \mid \text{pa}_{\mathcal{G}}(R_2))} \Bigg|_{R=1}$$
$$= \frac{p(L, R=1)}{p(R_1 = 1 \mid L_2^{(1)}, \text{do}(R_2 = 1)) \times p(R_2 = 1 \mid R_1 = 1, L_1)}.$$

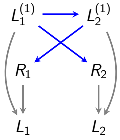
Main identification challenge: selection bias

Can we apply the causal-irrelevancy idea to the block-parallel model?



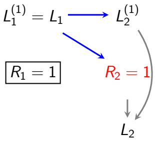
$$
\begin{aligned}
p(R_2 \mid L_1^{(1)}) &= p(R_2 \mid L_1^{(1)}, \mathrm{do}(R_1 = 1)) \\
&= \frac{p(R_2, L_1^{(1)} \mid \mathrm{do}(R_1 = 1))}{p(L_1^{(1)} \mid \mathrm{do}(R_1 = 1))}
\end{aligned}
$$

Can we apply the causal-irrelevancy idea to the block-parallel model?



$$
\begin{aligned}
p(R_2 \mid L_1^{(1)}) &= p(R_2 \mid L_1^{(1)}, \mathrm{do}(R_1 = 1)) \\
&= \frac{p(R_2, L_1^{(1)} \mid \mathrm{do}(R_1 = 1))}{p(L_1^{(1)} \mid \mathrm{do}(R_1 = 1))}
\end{aligned}
$$

An intervention on R_1 implies:



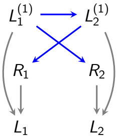
$$
p(R_2, L_1^{(1)}, L_2^{(1)}, L_2 \mid \mathrm{do}(R_1 = 1)) = \frac{p(L_1, R_1 = 1, L_2^{(1)}, L_2, R_2)}{p(R_1 = 1 \mid L_2^{(1)})}
$$

We can only evaluate the above expression when $R_2 = 1$:

$$
p(L_1^{(1)}, R_2 = 1 \mid \mathrm{do}(R_1 = 1)) = \sum_{l_2^{(1)}} \frac{p(L_2^{(1)}, R_1 = 1, R_2 = 1)}{p(R_1 = 1 \mid L_2^{(1)}, R_2 = 1)}
$$

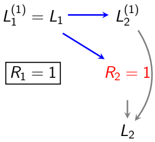
Main identification challenge: selection bias

Can we apply the causal-irrelevancy idea to the block-parallel model?



$$
\begin{aligned}
p(R_2 \mid L_1^{(1)}) &= p(R_2 \mid L_1^{(1)}, \mathrm{do}(R_1 = 1)) \\
&= \frac{p(R_2, L_1^{(1)} \mid \mathrm{do}(R_1 = 1))}{p(L_1^{(1)} \mid \mathrm{do}(R_1 = 1))}
\end{aligned}
$$

An intervention on R_1 implies:



$$
p(R_2, L_1^{(1)}, L_2^{(1)}, L_2 \mid \mathrm{do}(R_1 = 1)) = \frac{p(L_1, R_1 = 1, L_2^{(1)}, L_2, R_2)}{p(R_1 = 1 \mid L_2^{(1)})}
$$

We can only evaluate the above expression when $R_2 = 1$:

$$
p(L_1^{(1)}, \color{red}{R_2 = 1} \mid \mathrm{do}(R_1 = 1)) = \sum_{l_2^{(1)}} \frac{p(L_2^{(1)}, R_1 = 1, \color{red}{R_2 = 1})}{p(R_1 = 1 \mid L_2^{(1)}, \color{red}{R_2 = 1})}
$$

Intervening on R_1 **induces a selection** on R_2.

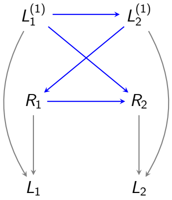
Inevitable selection bias

- The lesson is that **sequential interventions** do not help with identification arguments in the *block-parallel* model. Indeed, we need **parallel interventions** to dodge the selection bias issue.

Inevitable selection bias

- The lesson is that **sequential interventions** do not help with identification arguments in the *block-parallel* model. Indeed, we need **parallel interventions** to dodge the selection bias issue.

- Sometimes we cannot avoid the selection bias and end up with unidentified distributions. An example of this is the so-called *criss-cross* model,



Criss-cross (Nabi and Bhattacharya, 2023)

  - Full law $p(L^{(1)}, R)$ is not identified because of the colluder at R_2.

  - Target law $p(L^{(1)})$ is also **provably** not identified (Nabi and Bhattacharya, 2023; Guo et al., 2023).

Partial orders of interventions

- Sufficient rules for identification: finding **valid partial orders** of interventions that avoid the issue of selection bias.
  - That is a combination of sequential and parallel interventions (as opposed to a *total order* in causal inference).
- **Dodging selection bias** requires:
  - Set interventions
  - Intervening on variables other than R
  - Interventions on margins of $\mathcal{G}$ (pseudo-propensity scores)
- See Bhattacharya et al. (2019) and Nabi et al. (2022) for more discussions.

Identification arguments

1. m-DAG factorization of the missingness mechanism

2. Odds ratio parameterization of the missingness mechanism

Odds ratio parameterization (Chen, 2007)

Given disjoint sets of variables A, B, C, and reference values a_0, b_0 :

$$p(A, B \mid C) = \frac{1}{\mathcal{Z}(C)} \times p(A \mid B = b_0, C) \times p(B \mid A = a_0, C) \times \mathrm{OR}(A, B \mid C),$$

where

$$\begin{aligned}
\mathrm{OR}(A = a, B = b \mid C) &= \frac{p(A = a \mid B = b, C)}{p(A = a_0 \mid B = b, C)} \times \frac{p(A = a_0 \mid B = b_0, C)}{p(A = a \mid B = b_0, C)} \\
&= \frac{p(B = b \mid A = a, C)}{p(B = b_0 \mid A = a, C)} \times \frac{p(B = b_0 \mid A = a_0, C)}{p(B = b \mid A = a_0, C)}
\end{aligned}$$

$$\mathcal{Z}(C) = \sum_{a,b} p(A = a \mid B = b_0, C) \times p(B = b \mid A = a_0, C) \times \mathrm{OR}(A = a, B = b \mid C).$$

Odds ratio parameterization (Chen, 2007)

Given disjoint sets of variables A, B, C, and reference values a_0, b_0 :

$$p(A, B \mid C) = \frac{1}{\mathcal{Z}(C)} \times p(A \mid B = b_0, C) \times p(B \mid A = a_0, C) \times \mathrm{OR}(A, B \mid C),$$

where

$$\begin{aligned} \mathrm{OR}(A = a, B = b \mid C) &= \frac{p(A = a \mid B = b, C)}{p(A = a_0 \mid B = b, C)} \times \frac{p(A = a_0 \mid B = b_0, C)}{p(A = a \mid B = b_0, C)} \\ &= \frac{p(B = b \mid A = a, C)}{p(B = b_0 \mid A = a, C)} \times \frac{p(B = b_0 \mid A = a_0, C)}{p(B = b \mid A = a_0, C)} \end{aligned}$$

$$\mathcal{Z}(C) = \sum_{a,b} p(A = a \mid B = b_0, C) \times p(B = b \mid A = a_0, C) \times \mathrm{OR}(A = a, B = b \mid C).$$

- It is symmetric: $\mathrm{OR}(A, B \mid C) = \mathrm{OR}(B, A \mid C)$
- $\mathrm{OR}(A = a_0, B \mid C) = \mathrm{OR}(A, B = b_0 \mid C) = \mathrm{OR}(A = a_0, B = b_0 \mid C) = 1$

Identification via odds ratio parameterization

- m-DAG factorization view for identification:

$$p(R_i, R_j \mid \text{pa}_{\mathcal{G}}(R_i, R_j)) = p(R_i \mid \text{pa}_{\mathcal{G}}(R_i)) \times p(R_j \mid \text{pa}_{\mathcal{G}}(R_j))$$

  Identify:

  - $p(R_i \mid \text{pa}_{\mathcal{G}}(R_i))$,
  - $p(R_j \mid \text{pa}_{\mathcal{G}}(R_j))$.

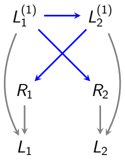
- Odds ratio parameterization view for identification:

$$p(R_i, R_j \mid L^{(1)}) = \frac{1}{\mathcal{Z}(L^{(1)})} \times p(R_i \mid R_j = 1, L^{(1)}) \times p(R_j \mid R_i = 1, L^{(1)}) \times \text{OR}(R_i, R_j \mid L^{(1)})$$

  Identify:

  - $p(R_i \mid R_j = 1, L^{(1)})$,
  - $p(R_j \mid R_i = 1, L^{(1)})$,
  - $\text{OR}(R_i = 0, R_j = 0 \mid L^{(1)})$.

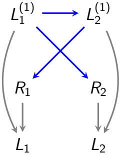
Example: block-parallel model



- Is the full/target law identified in the **block-parallel** model?

$$p(R \mid L_1^{(1)}) = \frac{1}{\mathcal{Z}(L^{(1)})} \times \underbrace{p(R_1 \mid R_2 = 1, L^{(1)})}_{(1)} \times \underbrace{p(R_2 \mid R_1 = 1, L^{(1)})}_{(2)} \times \underbrace{\mathrm{OR}(R_1, R_2 \mid L^{(1)})}_{(3)}$$

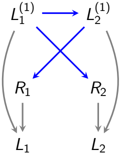
Example: block-parallel model



- Is the full/target law identified in the **block-parallel** model?

$$p(R \mid L_1^{(1)}) = \frac{1}{\mathcal{Z}(L^{(1)})} \times \underbrace{p(R_1 \mid R_2 = 1, L^{(1)})}_{(1)} \times \underbrace{p(R_2 \mid R_1 = 1, L^{(1)})}_{(2)} \times \underbrace{\mathrm{OR}(R_1, R_2 \mid L^{(1)})}_{(3)}$$

Note that $R_1 \perp\!\!\!\perp L_1^{(1)} \mid R_2, L_2^{(1)}$ and $R_2 \perp\!\!\!\perp L_2^{(1)} \mid R_1, L_1^{(1)}$,



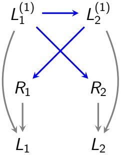
- Is the full/target law identified in the **block-parallel** model?

$$p(R \mid L_1^{(1)}) = \frac{1}{\mathcal{Z}(L^{(1)})} \times \underbrace{p(R_1 \mid R_2 = 1, L^{(1)})}_{(1)} \times \underbrace{p(R_2 \mid R_1 = 1, L^{(1)})}_{(2)} \times \underbrace{\text{OR}(R_1, R_2 \mid L^{(1)})}_{(3)}$$

Note that $R_1 \perp\!\!\!\perp L_1^{(1)} \mid R_2, L_2^{(1)}$ and $R_2 \perp\!\!\!\perp L_2^{(1)} \mid R_1, L_1^{(1)}$,

$$(1) : p(R_1 \mid R_2 = 1, L^{(1)}) = p(R_1 \mid R_2 = 1, L_2^{(1)}) = p(R_1 \mid R_2 = 1, L_2)$$
$$(2) : p(R_2 \mid R_1 = 1, L^{(1)}) = p(R_2 \mid R_1 = 1, L_1^{(1)}) = p(R_2 \mid R_1 = 1, L_1)$$
$$(3) : \text{OR}(R_1 = r_1, R_2 = r_2 \mid L^{(1)}) = 1.$$

Example: block-parallel model



- Is the full/target law identified in the **block-parallel** model?

$$p(R \mid L_1^{(1)}) = \frac{1}{\mathcal{Z}(L^{(1)})} \times \underbrace{p(R_1 \mid R_2 = 1, L^{(1)})}_{(1)} \times \underbrace{p(R_2 \mid R_1 = 1, L^{(1)})}_{(2)} \times \underbrace{\text{OR}(R_1, R_2 \mid L^{(1)})}_{(3)}$$

Note that $R_1 \perp\!\!\!\perp L_1^{(1)} \mid R_2, L_2^{(1)}$ and $R_2 \perp\!\!\!\perp L_2^{(1)} \mid R_1, L_1^{(1)}$,

$$(1) : p(R_1 \mid R_2 = 1, L^{(1)}) = p(R_1 \mid R_2 = 1, L_2^{(1)}) = p(R_1 \mid R_2 = 1, L_2)$$
$$(2) : p(R_2 \mid R_1 = 1, L^{(1)}) = p(R_2 \mid R_1 = 1, L_1^{(1)}) = p(R_2 \mid R_1 = 1, L_1)$$
$$(3) : \text{OR}(R_1 = r_1, R_2 = r_2 \mid L^{(1)}) = 1.$$

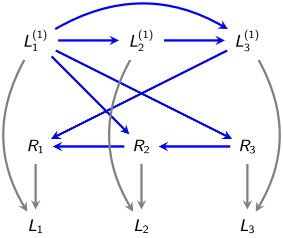
Yes, both full and target laws are identified.

Why the two parameterizations?

There exist identified MANR models where:

- *m-DAG factorization* approach fails to identify the model, but the *odds ratio parameterization* approach succeeds.
- *Odds ratio parameterization* approach fails to identify the model, but the *m-DAG factorization* approach succeeds.

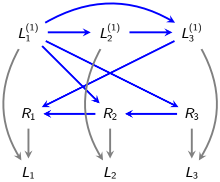
Example: m-DAG factorization fails!



Is the target/full law identified?

$$
\begin{aligned}
p(R \mid L^{(1)}) &= \prod_{R_i \in R} p(R_i \mid \text{pa}_{\mathcal{G}}(R_i)) \\
&= p(R_1 \mid R_2, L_3^{(1)}) \times p(R_2 \mid R_3, L_1^{(1)}) \times p(R_3 \mid L_1^{(1)}).
\end{aligned}
$$

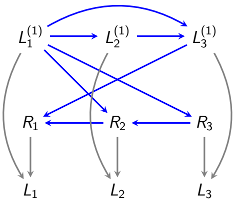
Example ctd. identification of the propensity score of R_1

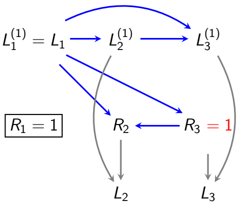


$$\begin{aligned} p(R \mid L^{(1)}) &= \prod_{R_i \in R} p(R_i \mid \mathrm{pa}_{\mathcal{G}}(R_i)) \\ &= p(R_1 \mid R_2, L_3^{(1)}) \times p(R_2 \mid R_3, L_1^{(1)}) \times p(R_3 \mid L_1^{(1)}). \end{aligned}$$

Nonparametric identification of $p(R_1 \mid \mathrm{pa}_{\mathcal{G}}(R_i))$

$$\begin{aligned} p(R_1 \mid \mathrm{pa}_{\mathcal{G}}(R_1)) &= p(R_1 \mid R_2, L_3^{(1)}) \\ &= p(R_1 \mid R_3 = 1, R_2, L_3^{(1)}) && R_1 \perp\!\!\!\perp R_3 \mid R_2, L_3^{(1)} \\ &= p(R_1 \mid R_3 = 1, R_2, L_3) && \text{consistency.} \end{aligned}$$



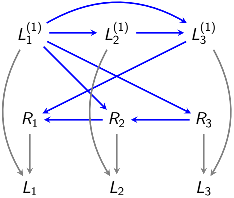


$$p^* = p(L^{(1)}, R_1, R_3 \mid \mathrm{do}(R_1 = 1)) = \frac{p(L^{(1)}, R)}{p(R_1 = 1 \mid R_3 = 1, R_2, L_3^{(1)})}.$$

Nonparametric identification of $p(R_2 \mid \mathrm{pa}_{\mathcal{G}}(R_i))|_{R=1}$

$$\begin{aligned}
p(R_2 = 1 \mid \mathrm{pa}_{\mathcal{G}}(R_2))|_{R=1} &= p(R_2 = 1 \mid R_3 = 1, L_1^{(1)}) \\
&= p(R_2 = 1 \mid R_3 = 1, L_1^{(1)}, \mathrm{do}(R_1 = 1)) \qquad \text{causal irrelevance} \\
&= p^*(R_2 = 1 \mid R_3 = 1, L_1^{(1)}) \\
&= p^*(R_2 = 1 \mid R_1 = 1, R_3 = 1, L_1^{(1)}) \qquad R_2 \perp\!\!\!\perp_{\mathcal{G}^*} R_1 \mid R_3, L_1^{(1)} \\
&= p^*(R_2 = 1 \mid R_1 = 1, R_3 = 1, L_1) \qquad \text{consistency}
\end{aligned}$$

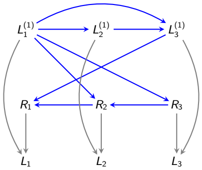
Example ctd. identification of the propensity score of R_3



- Unfortunately, similar tricks do not help with identification of $p(R_3 \mid \mathrm{pa}_{\mathcal{G}}(R_3))$ due to selection bias on R_3 from intervening on either R_1 or R_2.

- It seems that the missingness mechanism is not identified. Thus, it seems neither the full law nor the target law are identified. Can we prove this non-identification claim?!
  - The answer is no, because the model is indeed identified. We can prove identification using odds ratio parameterization of the missingness mechanism.

Example ctd. odds ratio parameterization



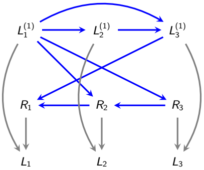
Univariate conditionals:

- $p(R_1 \mid R_2 = 1, R_3 = 1, L^{(1)}) = p(R_1 \mid R_2 = 1, R_3 = 1, L_3)$
    - $R_1 \perp\!\!\!\perp L_1^{(1)}, L_2^{(2)} \mid R_2, R_3, L_3^{(1)}$
- $p(R_2 \mid R_1 = 1, R_3 = 1, L^{(1)}) = p(R_2 \mid R_1 = 1, R_3 = 1, L_1, L_3)$
    - $R_2 \perp\!\!\!\perp L_2^{(1)} \mid R_1, R_3, L_1^{(2)}, L_3^{(1)}$
- $p(R_3 \mid R_1 = 1, R_2 = 1, L^{(1)}) = p(R_3 \mid R_2 = 1, R_1 = 1, L_1)$
    - $R_3 \perp\!\!\!\perp L_2^{(1)}, L_3^{(1)} \mid R_1, R_2, L_1^{(2)}$

Pairwise odds ratios:

- $\mathrm{OR}(R_1, R_2 \mid R_3 = 1, L^{(1)})$
- $\mathrm{OR}(R_2, R_3 \mid R_1 = 1, L^{(1)})$
- $\mathrm{OR}(R_1, R_3 \mid R_2 = 1, L^{(1)})$

Example ctd. identification of the paiwise odds ratio terms



$$
\begin{aligned}
&\mathrm{OR}(R_2 = 0, R_3 = 0 \mid R_1 = 1, L^{(1)}) \\
&\quad = \frac{p(R_3 = 0 \mid R_2 = 0, R_1 = 1, L^{(1)})}{p(R_3 = 1 \mid R_2 = 0, R_1 = 1, L^{(1)})} \times \frac{p(R_3 = 1 \mid R_2 = 1, R_1 = 1, L^{(1)})}{p(R_3 = 0 \mid R_2 = 1, R_1 = 1, L^{(1)})} \\
&\quad = \frac{p(R_3 = 0 \mid R_2 = 0, R_1 = 1, L_1^{(1)})}{p(R_3 = 1 \mid R_2 = 0, R_1 = 1, L_1^{(1)})} \times \frac{p(R_3 = 1 \mid R_2 = 1, R_1 = 1, L_1^{(1)})}{p(R_3 = 0 \mid R_2 = 1, R_1 = 1, L_1^{(1)})} \\
&\quad = \frac{p(R_3 = 0 \mid R_2 = 0, R_1 = 1, L_1)}{p(R_3 = 1 \mid R_2 = 0, R_1 = 1, L_1)} \times \frac{p(R_3 = 1 \mid R_2 = 1, R_1 = 1, L_1)}{p(R_3 = 0 \mid R_2 = 1, R_1 = 1, L_1)}.
\end{aligned}
$$

- First equality holds by definition, second by $R_3 \perp\!\!\!\perp L_2^{(1)}, L_3^{(1)} \mid R_1, R_2, L_1^{(1)}$, and third by consistency.
- $\mathrm{OR}(R_2, R_3 \mid R_1 = 1, L^{(1)})$ and $\mathrm{OR}(R_1, R_3 \mid R_2 = 1, L^{(1)})$ can be similarly identified. Thus, the missingness mechanism, full law, and target law are all identified.

Odds ratio parameterization of $p(R \mid L^{(1)})$

- Without loss of generality assume $Z = \emptyset$.
- Let $R_{-k} = V \setminus R_k$, $R_{\prec k} = \{R_1, \ldots, R_{k-1}\}$, and $R_{\succ k} = \{R_{k+1}, \ldots, R_K\}$.
- The general form of odds ratio parameterization is as follows:

$$p(R \mid L^{(1)}) = \frac{1}{\mathcal{Z}(L^{(1)})} \times \prod_{k=1}^{K} p(R_k \mid R_{-k} = 1, L^{(1)}) \times \prod_{k=2}^{K} \text{OR}(R_k, R_{\prec k} \mid R_{\succ k} = 1, L^{(1)}),$$

$$\text{OR}(R_k, R_{\prec k} \mid R_{\succ k} = 1, L^{(1)}) = \frac{p(R_k \mid R_{\succ k} = 1, R_{\prec k}, L^{(1)})}{p(R_k = 1 \mid R_{\succ k} = 1, R_{\prec k}, X^{(1)})} \times \frac{p(R_k = 1 \mid R_{-k} = 1, X^{(1)})}{p(R_k \mid R_{-k} = 1, L^{(1)})}.$$

Need to identify:

- Univariate conditional distributions: $p(R_k \mid R_{-k} = 1, L^{(1)})$
- Odds ratio terms: $\text{OR}(R_k, R_{\prec k} \mid R_{\succ k} = 1, L^{(1)})$

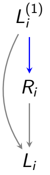
- When can we succeed?
- When do we fail?

Full law identification theory in m-DAGs

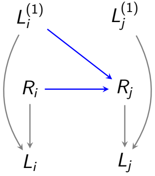
Theorem (Graphical characterization of identified full laws)

Full law $p(R, L^{(1)}, Z)$ that is Markov relative to a missing data DAG $\mathcal{G}$ is identified ***if and only if*** *$\mathcal{G}$ does not contain the following two structures:* (Nabi et al., 2020)

- *self-censoring edge:* $L_i^{(1)} \rightarrow R_i$,
- *colluder:* $L_j^{(1)} \rightarrow R_i \leftarrow R_j$.



self-censoring



colluder

- These graphical conditions are **sound** and **complete** for full law ID.
- Identification functional is given by the odds ratio parameterization of $p(R \mid Z, L^{(1)})$.

Proof sketch

- Absence of self-censoring edges and colluders imply:

$$R_k \perp\!\!\!\perp L_k^{(1)} \mid R_{-k}, L^{(1)} \setminus L_k^{(1)}$$

- Odds ratio parameterization:

$$p(R \mid L^{(1)}) = \frac{1}{\mathcal{Z}(L^{(1)})} \times \prod_{k=1}^{K} p(R_k \mid R_{-k} = 1, L^{(1)}) \times \prod_{k=2}^{K} \mathrm{OR}(R_k, R_{\prec k} \mid R_{\succ k} = 1, L^{(1)}).$$

- $p(R_k \mid R_{-k} = 1, L^{(1)}) = p(R_k \mid R_{-k} = 1, L_{-k})$.

- $\mathrm{OR}(R_i, R_j \mid R_{-\{i,j\}} = 1, L^{(1)})$ is identified via "symmetric argument,"
  - It is not a function of $L_i^{(1)}$ and it is not a function of $L_j^{(1)}$

- DAGs with no self-censoring edges and no colluders are submodels of *Itemwise Conditionally Independence Nonresponse* model
(Sadinle and Reiter, 2017; Shpitser, 2016; Malinsky et al., 2021)

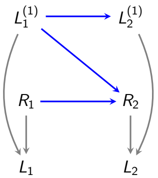
Why the two parameterizations?

There exist identified MANR models where:

- *m-DAG factorization* approach fails to identify the model, but the *odds ratio parameterization* approach succeeds.
- *Odds ratio parameterization* approach fails to identify the model, but the *m-DAG factorization* approach succeeds.

Example: odds ratio parameterization fails!

- If there exists a colluder in the m-DAG, then the full law is not identifiable. However, the target law might still be identified.
- An example of this is the block-conditional MAR model.



- As we saw earlier, $p(R = 1 \mid \text{pa}_{\mathcal{G}}(R))$ is easily identifiable as follows:

$$p(R = 1 \mid \text{pa}_{\mathcal{G}}(R)) = p(R_1 = 1) \times p(R_2 = 1 \mid R_1 = 1, L_1).$$

  Therefore, the target law is identified. However, the odds ratio parameterization approach fails here.
- Even though the odds ratio parameterization led to completeness results for full law identification, the m-DAG factorization is still the only tool we can use for target law identification (in the presence of colluders).

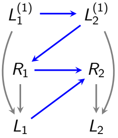
Summary: np-identification in m-DAGs

- Causal identification theories are incomplete for identification of full and target laws in m-DAGs.
- Full law (target law) identification is equivalent to identification of the missingness mechanism for all missingness patters $R = r$ (for all-one pattern $R = 1$).
- Two identification tools are at our disposal:
  - m-DAG factorization of $p(R \mid \mathrm{pa}_{\mathcal{G}}(R))$ (Bhattacharya et al., 2019)
  - odds ratio parameterization of $p(R \mid \mathrm{pa}_{\mathcal{G}}(R))$ (Nabi et al., 2020)
- There does exist sound and complete identification results for full law.
  - Non-identified graphical structures are self-censoring edges and colluders (Nabi et al., 2020)
- The completeness theory of identification for target law remains an open problem.
  - Only known/proven non-identified graphical structures for target law are self-censoring edges and criss-cross structures (Nabi and Bhattacharya, 2023; Guo et al., 2023)
- For comprehensive review of these discussions see Nabi et al. (2022).

Part 3. Non/Semi-parametric Estimation

Estimation: IPW

- Consider the permutation model with two variables:



(Permutation model)

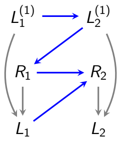
- Let $\mu = \mathbb{E}[h(L_1^{(1)}, L_2^{(1)})]$ denote our parameter of interest, which is identified as:

$$\mu = \mathbb{E}\left[\frac{R_1 \ R_2}{\pi(R_1 = 1 \mid L_2^{(1)}, \color{red}{\mathrm{do}(R_2 = 1)}) \ \pi(R_2 = 1 \mid R_1 = 1, L_1)} \times h(L_1, L_2)\right] \ ,$$

where $\pi(R_k = 1 \mid \mathrm{pa}_{\mathcal{G}}(R_k)) = p(R_k = 1 \mid \mathrm{pa}_{\mathcal{G}}(R_k))$.

Estimation: IPW

- Consider the permutation model with two variables:



(Permutation model)

- Let $\mu = \mathbb{E}[h(L_1^{(1)}, L_2^{(1)})]$ denote our parameter of interest, which is identified as:

$$\mu = \mathbb{E}\left[\frac{R_1 \ R_2}{\pi(R_1 = 1 \mid L_2^{(1)}, \color{red}{\mathrm{do}(R_2 = 1)}) \ \pi(R_2 = 1 \mid R_1 = 1, L_1)} \times h(L_1, L_2)\right] \ ,$$

where $\pi(R_k = 1 \mid \mathrm{pa}_{\mathcal{G}}(R_k)) = p(R_k = 1 \mid \mathrm{pa}_{\mathcal{G}}(R_k))$.

- Given nuisance estimates π_n, h_n, an IPW estimator of μ is:

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{R_{1,i} \ R_{2,i}}{\pi_n(R_1 = 1 \mid L_{2,i}^{(1)}, \color{red}{\mathrm{do}(R_2 = 1)}) \ \pi_n(R_2 = 1 \mid R_1 = 1, L_{1,i})} \times h_n(L_{1,i}, L_{2,i})\right] \ ,$$

- How to obtain $\pi_n(R_1 = 1 \mid L^{(1)})$?

Weighted estimating equations

- How does identification of $\pi(R_1 = 1 \mid L_2^{(1)})$ help with estimation?

$$\pi(R_1 = 1 \mid L_2^{(1)}) = \pi(R_1 = 1 \mid L_2^{(1)}, \text{do}(R_2 = 1)) \ .$$

Weighted estimating equations

- How does identification of $\pi(R_1 = 1 \mid L_2^{(1)})$ help with estimation?

$$\pi(R_1 = 1 \mid L_2^{(1)}) = \pi(R_1 = 1 \mid L_2^{(1)}, \mathrm{do}(R_2 = 1)) \ .$$

- Assume $p(R_1 = 1 \mid L_2^{(1)}) = p(R_1 = 1 \mid L_2^{(1)}; \alpha), \ \alpha \in \mathbb{R}^p$, and consider the following estimating equation for α under the full law:

$$\mathbb{E}[U(R_1, L_2^{(1)}; \alpha)] = 0 \qquad \mapsto \qquad \sum_{i=1}^{n} U(R_{1,i}, L_{2,i}^{(1)}; \alpha) = 0 \ .$$

Weighted estimating equations

- How does identification of $\pi(R_1 = 1 \mid L_2^{(1)})$ help with estimation?

$$\pi(R_1 = 1 \mid L_2^{(1)}) = \pi(R_1 = 1 \mid L_2^{(1)}, \textcolor{red}{\mathrm{do}(R_2 = 1)}) \ .$$

- Assume $p(R_1 = 1 \mid L_2^{(1)}) = p(R_1 = 1 \mid L_2^{(1)}; \alpha), \ \alpha \in \mathbb{R}^p$, and consider the following estimating equation for α under the full law:

$$\mathbb{E}[U(R_1, L_2^{(1)}; \alpha)] = 0 \qquad \mapsto \qquad \sum_{i=1}^{n} U(R_{1,i}, L_{2,i}^{(1)}; \alpha) = 0 \ .$$

- Thus, below an estimating equation for α under the observed data law:

$$\mathbb{E}\left[\frac{R_2}{\pi(R_2 = 1 \mid R_1, L_1)} \times U(R_1, L_2; \alpha)\right] = 0.$$

$$\mapsto \qquad \sum_{i=1}^{n} \frac{R_{2,i}}{\pi_n(R_2 = 1 \mid R_{1,i}, L_{1,i})} \times U(R_{1,i}, L_{2,i}; \alpha) = 0.$$

Weighted estimating equations

- How does identification of $\pi(R_1 = 1 \mid L_2^{(1)})$ help with estimation?

$$\pi(R_1 = 1 \mid L_2^{(1)}) = \pi(R_1 = 1 \mid L_2^{(1)}, \text{do}(R_2 = 1)) \; .$$

- Assume $p(R_1 = 1 \mid L_2^{(1)}) = p(R_1 = 1 \mid L_2^{(1)}; \alpha), \;\; \alpha \in \mathbb{R}^p$, and consider the following estimating equation for α under the full law:

$$\mathbb{E}[U(R_1, L_2^{(1)}; \alpha)] = 0 \qquad \mapsto \qquad \sum_{i=1}^{n} U(R_{1,i}, L_{2,i}^{(1)}; \alpha) = 0 \; .$$

- Thus, below an estimating equation for α under the observed data law:

$$\mathbb{E}\left[\frac{R_2}{\pi(R_2 = 1 \mid R_1, L_1)} \times U(R_1, L_2; \alpha)\right] = 0.$$

$$\mapsto \qquad \sum_{i=1}^{n} \frac{R_{2,i}}{\pi_n(R_2 = 1 \mid R_{1,i}, L_{1,i})} \times U(R_{1,i}, L_{2,i}; \alpha) = 0.$$

- A more desirable estimator for μ can be derived using the nonparametric efficient influence function (EIF).

General theory of estimation

- W.L.G. assume $Z = \emptyset$ and let $O = (R, L)$ denote the observed data.
- Let $\mu = E[h(L^{(1)})]$ be the parameter of interest for a specified function $h(\cdot)$ in a given m-DAG $\mathcal{G}$.
- A regular and asymptotically linear (RAL) estimator μ_n of μ has the property that

$$\sqrt{n} \times (\mu_n - \mu) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \underbrace{\Phi(O_i)}_{\text{Influence Function}} + o_P(1) \ .$$

  Given the influence function $\Phi(O)$, we can construct a RAL estimator.
- Influence functions for estimator of μ live in the orthogonal complement of observed data tangent space, $\Lambda^{O,\perp}$.

General theory of estimation ctd.

- The **tangent space** of the statistical model for an m-DAG $\mathcal{G}(L^{(1)}, R)$ is defined as the collection of score functions over $L^{(1)}, R$, and is given by:

$$\Lambda = \Lambda_1 + \Lambda_2,$$

  - Λ_1 denotes be the collection of all score functions under model for $L^{(1)}$, i.e., mean-zero functions that respect the independence restrictions of the target law $p(L^{(1)})$, and
  - Λ_2 denotes the collection of scores under models for $R \mid L^{(1)}$, i.e., mean-zero functions that respect the independence restrictions of the missingness mechanism $p(R \mid L^{(1)})$.

General theory of estimation ctd.

- The **tangent space** of the statistical model for an m-DAG $\mathcal{G}(L^{(1)}, R)$ is defined as the collection of score functions over $L^{(1)}, R$, and is given by:

$$\Lambda = \Lambda_1 + \Lambda_2,$$

  - Λ_1 denotes be the collection of all score functions under model for $L^{(1)}$, i.e., mean-zero functions that respect the independence restrictions of the target law $p(L^{(1)})$, and
  - Λ_2 denotes the collection of scores under models for $R \mid L^{(1)}$, i.e., mean-zero functions that respect the independence restrictions of the missingness mechanism $p(R \mid L^{(1)})$.

- The **observed data tangent space** is defined as:

$$\Lambda^O = \Lambda_1^O \cup \Lambda_2^O,$$

  - Λ_1^O: any function of observed data that lives in Λ_1.
  - Λ_2^O: any function of observed data that lives in Λ_2.

General theory of estimation ctd.

- The **tangent space** of the statistical model for an m-DAG $\mathcal{G}(L^{(1)}, R)$ is defined as the collection of score functions over $L^{(1)}, R$, and is given by:

$$\Lambda = \Lambda_1 + \Lambda_2,$$

  - Λ_1 denotes be the collection of all score functions under model for $L^{(1)}$, i.e., mean-zero functions that respect the independence restrictions of the target law $p(L^{(1)})$, and
  - Λ_2 denotes the collection of scores under models for $R \mid L^{(1)}$, i.e., mean-zero functions that respect the independence restrictions of the missingness mechanism $p(R \mid L^{(1)})$.

- The **observed data tangent space** is defined as:

$$\Lambda^O = \Lambda_1^O \cup \Lambda_2^O,$$

  - Λ_1^O: any function of observed data that lives in Λ_1.
  - Λ_2^O: any function of observed data that lives in Λ_2.

- The **orthogonal complement of observed data tangent space** is defined as

$$\Lambda^{O,\perp} = \Lambda_1^{O,\perp} \cap \Lambda_2^{O,\perp}.$$

General theory of estimation ctd.

Influence functions for estimator of μ live in the orthogonal complement of observed data tangent space, $\Lambda^{O,\perp}$.

$$\Lambda^{O,\perp} = \Lambda_1^{O,\perp} \cap \Lambda_2^{O,\perp}.$$

General theory of estimation ctd.

Influence functions for estimator of μ live in the orthogonal complement of observed data tangent space, $\Lambda^{O,\perp}$.

$$\Lambda^{O,\perp} = \Lambda_1^{O,\perp} \cap \Lambda_2^{O,\perp}.$$

Scharfstein et al. (1999) have shown:

$$\Lambda_1^{O,\perp} = \left\{ \frac{R}{p(R=1 \mid L^{(1)})} a(L^{(1)}) + b(O) : a(L^{(1)}) \in \Lambda_1^{\perp}, \mathbb{E}[b(O) \mid L^{(1)}] = 0 \right\}$$

$$\Lambda_2^{O,\perp} = \left\{ b(O) : b(O) \in \Lambda_2^{\perp} \right\} \ .$$

General theory of estimation ctd.

Influence functions for estimator of μ live in the orthogonal complement of observed data tangent space, $\Lambda^{O,\perp}$.

$$\Lambda^{O,\perp} = \Lambda_1^{O,\perp} \cap \Lambda_2^{O,\perp}.$$

Scharfstein et al. (1999) have shown:

$$\Lambda_1^{O,\perp} = \left\{ \frac{R}{p(R=1 \mid L^{(1)})} a(L^{(1)}) + b(O) : a(L^{(1)}) \in \Lambda_1^{\perp}, \mathbb{E}[b(O) \mid L^{(1)}] = 0 \right\}$$
$$\Lambda_2^{O,\perp} = \left\{ b(O) : b(O) \in \Lambda_2^{\perp} \right\} .$$

Plan for deriving influence functions:

- Write out an expression for the elements of $\Lambda_1^{O,\perp}$
  - $a(L^{(1)}) \in \Lambda_1^{\perp}$, and
  - $b(O)$ such that $\mathbb{E}[b(O) \mid L^{(1)}] = 0$.
- Find restrictions on these elements to ensure orthogonality to Λ_2

Application of general theory of estimation: $a(L^{(1)})$

- Let $L^{(1)} = (L_1^{(1)}, L_2^{(1)})$ and $R = (R_1, R_2)$.
- Let $\mu = \mathbb{E}[\underbrace{h(L_1^{(1)}, L_2^{(1)})}_{h(L^{(1)})}]$ for a specified function $h(\cdot)$.
- Suppose no restrictions are placed on the distribution of $L^{(1)}$.
  - The elements of $\Lambda_1^{\perp}$ will be proportional to $h(L^{(1)}) - \mu$.
- Thus, we choose: $a(L^{(1)}) = h(L^{(1)}) - \mu$.

Application of general theory of estimation: $b(O)$

- Any observed data random variable can we written as

$$b(O) = R_1 R_2 c_{11}(L^{(1)}) + R_1(1 - R_2)c_{10}(L_1^{(1)}) + (1 - R_1)R_2 c_{01}(L_2^{(1)}) + (1 - R_1)(1 - R_2)c_{00}$$

Application of general theory of estimation: $b(O)$

- Any observed data random variable can we written as

$$b(O) = R_1 R_2 c_{11}(L^{(1)}) + R_1(1 - R_2)c_{10}(L_1^{(1)}) + (1 - R_1)R_2 c_{01}(L_2^{(1)}) + (1 - R_1)(1 - R_2)c_{00}$$

- What restrictions on $c_{11}, c_{10}, c_{01}, c_{00}$ ensure that the $\mathbb{E}[b(O) \mid L^{(1)}] = 0$?

$$c_{11}(L^{(1)}) = \frac{-\pi_{10}(L^{(1)})c_{10}(L_1^{(1)}) - \pi_{01}(L^{(1)})c_{01}(L_2^{(1)}) - \pi_{00}(L^{(1)})c_{00}}{\pi_{11}(L^{(1)})}$$

where $\pi_{ij}(L^{(1)}) = p(R_1 = i, R_2 = j \mid L^{(1)})$.

Application of general theory of estimation: $b(O)$

- Any observed data random variable can we written as

$$b(O) = R_1 R_2 c_{11}(L^{(1)}) + R_1(1-R_2)c_{10}(L_1^{(1)}) + (1-R_1)R_2 c_{01}(L_2^{(1)}) + (1-R_1)(1-R_2)c_{00}$$

- What restrictions on $c_{11}, c_{10}, c_{01}, c_{00}$ ensure that the $\mathbb{E}[b(O) \mid L^{(1)}] = 0$?

$$c_{11}(L^{(1)}) = \frac{-\pi_{10}(L^{(1)})c_{10}(L_1^{(1)}) - \pi_{01}(L^{(1)})c_{01}(L_2^{(1)}) - \pi_{00}(L^{(1)})c_{00}}{\pi_{11}(L^{(1)})}$$

where $\pi_{ij}(L^{(1)}) = p(R_1 = i, R_2 = j \mid L^{(1)})$.

- Any observed data random variable that has mean zero given $L^{(1)}$ can be expressed as

$$\begin{aligned}
&\left\{-\frac{R_1 R_2}{\pi_{11}(L^{(1)})}\pi_{10}(L^{(1)}) + R_1(1-R_2)\right\} c_{10}(L_1^{(1)}) \\
&+ \left\{-\frac{R_1 R_2}{\pi_{11}(L^{(1)})}\pi_{01}(L^{(1)}) + (1-R_1)R_2\right\} c_{01}(L_2^{(1)}) \\
&+ \left\{-\frac{R_1 R_2}{\pi_{11}(L^{(1)})}\pi_{00}(L^{(1)}) + (1-R_1)(1-R_2)\right\} c_{00}
\end{aligned}$$

Application of general theory of estimation: game plan

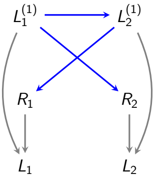
Find restrictions on elements in $\Lambda_1^{O,\perp}$ to ensure orthogonality to Λ_2, where:

$$
\begin{aligned}
\Lambda_1^{O,\perp} = \Bigg\{ & \frac{R_1 R_2}{\pi_{11}(L^{(1)})} \left\{ h(L^{(1)}) - \mu \right\} \\
& + \left\{ -\frac{R_1 R_2}{\pi_{11}(L^{(1)})} \pi_{10}(L^{(1)}) + R_1(1 - R_2) \right\} c_{10}(L_1^{(1)}) \\
& + \left\{ -\frac{R_1 R_2}{\pi_{11}(L^{(1)})} \pi_{01}(L^{(1)}) + (1 - R_1) R_2 \right\} c_{01}(L_2^{(1)}) \\
& + \left\{ -\frac{R_1 R_2}{\pi_{11}(L^{(1)})} \pi_{00}(L^{(1)}) + (1 - R_1)(1 - R_2) \right\} c_{00} \\
& : \text{for any } c_{10}(L_1^{(1)}), c_{01}(L_2^{(1)}), c_{00} \Bigg\}.
\end{aligned}
$$

This means we should find restrictions on $c_{10}(L_1^{(1)}), c_{01}(L_2^{(1)}), c_{00}$ to ensure orthogonality to Λ_2.

- For any $f_1 \in \Lambda_1^{O,\perp}$ and $f_2 \in \Lambda_2$: $\mathbb{E}[f_1 f_2] = 0$.

Example: block-parallel model



$$\pi_{jk}(L^{(1)}) = \pi_1(L_2^{(1)})^j\{1-\pi_1(L_2^{(1)})\}^{1-j}\pi_2(L_1^{(1)})^k\{1-\pi_2(L_1^{(1)})\}^{1-k}$$

$$\mathcal{L} = \pi_1(L_2^{(1)})^{R_1}\{1-\pi_1(L_2^{(1)})\}^{1-R_1}\pi_2(L_1^{(1)})^{R_2}\{1-\pi_2(L_1^{(1)})\}^{1-R_2}$$

$$\begin{aligned}\log\mathcal{L} =& R_1\log\{\pi_1(L_2^{(1)})\} + (1-R_1)\log\{1-\pi_1(L_2^{(1)})\}+\\ &+R_2\log\{\pi_2(L_1^{(1)})\} + (1-R_2)\log\{1-\pi_2(L_1^{(1)})\}\end{aligned}$$

$$\frac{\partial\log\mathcal{L}}{\partial\eta} = \frac{\{R_1-\pi_1(L_2^{(1)})\}}{\pi_1(L_2^{(1)})\{1-\pi_1(L_2^{(1)})\}}\pi_1'(L_2^{(1)}) + \frac{\{R_2-\pi_2(L_1^{(1)})\}}{\pi_2(L_1^{(1)})\{1-\pi_2(L_1^{(1)})\}}\pi_2'(L_1^{(1)})$$

Example: block-parallel model

$$\Lambda_2 = \left\{ \{R_1 - \pi_1(L_2^{(1)})\}g_1(L_2^{(1)}) + \{R_2 - \pi_2(L_1^{(1)})\}g_2(L_1^{(1)})) : g_1(L_2^{(1)}), g_2(L_1^{(1)}) \right\}$$

$$\Lambda_2 = \Lambda_{2,1} \oplus \Lambda_{2,2}$$

, where

$$\Lambda_{2,1} = \left\{ \{R_1 - \pi_1(L_2^{(1)})\}g_1(L_2^{(1)}) : g_1(L_2^{(1)}) \right\}$$

and

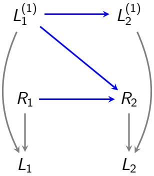
$$\Lambda_{2,2} = \left\{ \{R_2 - \pi_2(L_1^{(1)})\}g_2(L_1^{(1)}) : g_2(L_1^{(1)}) \right\}$$

Example: block-parallel model

What choices of $c_{10}(L_1^{(1)}), c_{01}(L_2^{(1)}), c_{00}$ that index elements of $\Lambda_1^{O,\perp}$ ensure orthogonality to both $\Lambda_{2,1}$ and $\Lambda_{2,2}$

Click here for detailed derivations in the supplement [$\rightarrow$].

Example: block-conditional MAR model



$$
\begin{aligned}
\pi_{jk}(L^{(1)}) =& \pi_1^j\{1-\pi_1\}^{1-j}\times \\
& \pi_2(1, L_1^{(1)})^{jk}\{1-\pi_2(1, L_1^{(1)})\}^{j(1-k)}\pi_2(0, L_1^{(1)})^{(1-j)k}\{1-\pi_2(0, L_1^{(1)})\}^{(1-j)(1-k)}
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{L} =& \pi_1^{R_1}\{1-\pi_1\}^{1-R_1}\pi_2(1, L_1^{(1)})^{R_1R_2}\{1-\pi_2(1, L_1^{(1)})\}^{R_1(1-R_2)}\times \\
& \pi_2(0, L_1^{(1)})^{(1-R_1)R_2}\{1-\pi_2(0, L_1^{(1)})\}^{(1-R_2)(1-R_2)}
\end{aligned}
$$

$$
\begin{aligned}
\log\mathcal{L} =& R_1\log\{\pi_1\} + (1-R_1)\log\{1-\pi_1\} + R_1R_2\log\{\pi_2(1, L_1^{(1)})\} \\
& + R_1(1-R_2)\log\{1-\pi_2(1, L_1^{(1)})\} + (1-R_1)R_2\log\{\pi_2(0, L_1^{(1)})\} \\
& + (1-R_1)(1-R_2)\log\{1-\pi_2(0, L_1^{(1)})\}
\end{aligned}
$$

Example: block-conditional MAR model

$$\frac{\partial \log \mathcal{L}}{\partial \eta} = \frac{\{R_1 - \pi_1\}}{\pi_1\{1-\pi_1\}} + R_1 \frac{\{R_2 - \pi_2(1, L_1^{(1)})\}}{\pi_2(1, L_1^{(1)})\{1 - \pi_2(1, L_1^{(1)})\}} \pi_2'(1, L_1^{(1)}) +$$
$$(1 - R_1) \frac{\{R_2 - \pi_2(0, L_1^{(1)})\}}{\pi_2(0, L_1^{(1)})\{1 - \pi_2(0, L_1^{(1)})\}} \pi_2'(0, L_1^{(1)})$$

$$\Lambda_2 = \Big\{ \{R_1 - \pi_1\} g_1 + R_1\{R_2 - \pi_2(1, L_1^{(1)})\} g_2(1, L_1^{(1)}) +$$
$$(1 - R_1)\{R_2 - \pi_2(0, L_1^{(1)})\} g_2(0, L_1^{(1)}) : g_1, g_2(1, L_1^{(1)}), g_2(0, L_1^{(1)}) \Big\}$$

$$\Lambda_2 = \Lambda_{2,1} \oplus \Lambda_{2,2,1} \oplus \Lambda_{2,2,0}$$

$$\Lambda_{2,1} = \{\{R_1 - \pi_1\} g_1 : g_1\}$$

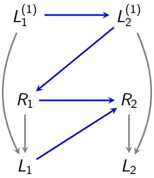
$$\Lambda_{2,2,1} = \Big\{ R_1\{R_2 - \pi_2(1, L_1^{(1)})\} g_2(1, L_1^{(1)}) : g_2(1, L_1^{(1)}) \Big\}$$

$$\Lambda_{2,2,0} = \Big\{ (1 - R_1)\{R_2 - \pi_2(0, L_1^{(1)})\} g_2(0, L_1^{(1)}) : g_2(0, L_1^{(1)}) \Big\}$$

Example: block-conditional MAR model

Click here for detailed derivations in the supplement [$\rightarrow$].

Example: permutation model



$$
\begin{aligned}
\pi_{jk}(L^{(1)}) =& \pi_1(L_2^{(1)})^j \{1 - \pi_1(L_2^{(1)})\}^{1-j} \times \\
& \pi_2(1, L_1^{(1)})^{jk} \{1 - \pi_2(1, L_1^{(1)})\}^{j(1-k)} \pi_2(0)^{(1-j)k} \{1 - \pi_2(0)\}^{(1-j)(1-k)}
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{L} =& \pi_1(L_2^{(1)})^{R_1} \{1 - \pi_1(L_2^{(1)})\}^{1-R_1} \pi_2(1, L_1^{(1)})^{R_1 R_2} \{1 - \pi_2(1, L_1^{(1)})\}^{R_1(1-R_2)} \times \\
& \pi_2(0)^{(1-R_1)R_2} \{1 - \pi_2(0)\}^{(1-R_2)(1-R_2)}
\end{aligned}
$$

$$
\begin{aligned}
\log \mathcal{L} =& R_1 \log\{\pi_1(L_2^{(1)})\} + (1 - R_1) \log\{1 - \pi_1(L_2^{(1)})\} + R_1 R_2 \log\{\pi_2(1, L_1^{(1)})\} \\
& + R_1(1 - R_2) \log\{1 - \pi_2(1, L_1^{(1)})\} + (1 - R_1) R_2 \log\{\pi_2(0)\} \\
& + (1 - R_1)(1 - R_2) \log\{1 - \pi_2(0)\}
\end{aligned}
$$

Example: permutation model

$$\frac{\partial \log \mathcal{L}}{\partial \eta} = \frac{\{R_1 - \pi_1(L_2^{(1)})\}}{\pi_1(L_2^{(1)})\{1 - \pi_1(L_2^{(1)})\}} \pi_1'(L_2^{(1)}) + R_1 \frac{\{R_2 - \pi_2(1, L_1^{(1)})\}}{\pi_2(1, L_1^{(1)})\{1 - \pi_2(1, L_1^{(1)})\}} \pi_2'(1, L_1^{(1)}) + (1 - R_1) \frac{\{R_2 - \pi_2(0)\}}{\pi_2(0\{1 - \pi_2(0\}}$$

$$\Lambda_2 = \Big\{ \{R_1 - \pi_1(L_2^{(1)})\} g_1(L_2^{(1)}) + R_1\{R_2 - \pi_2(1, L_1^{(1)})\} g_2(1, L_1^{(1)}) + (1 - R_1)\{R_2 - \pi_2(0)\} g_2(0) : g_1(L_2^{(1)}), g_2(1, L_1^{(1)}), g_2(0) \Big\}$$

$$\Lambda_2 = \Lambda_{2,1} \oplus \Lambda_{2,2,1} \oplus \Lambda_{2,2,0}$$

$$\Lambda_{2,1} = \Big\{ \{R_1 - \pi_1(L_2^{(1)})\} g_1(L_2^{(1)}) : g_1(L_2^{(1)}) \Big\}$$

$$\Lambda_{2,2,1} = \Big\{ R_1\{R_2 - \pi_2(1, L_1^{(1)})\} g_2(1, L_1^{(1)}) : g_2(1, L_1^{(1)}) \Big\}$$

$$\Lambda_{2,2,0} = \{(1 - R_1)\{R_2 - \pi_2(0)\} g_2(0 : g_2(0)\}$$

Example: permutation model

Click here for detailed derivations in the supplement [$\rightarrow$].

Wrap up

Revisiting course outline and objectives

Part I. **Missing data DAGs**

- Represented missingness mechanisms graphically; interpreted a missing data DAG model as a class of distributions with a set of independence restrictions.

Part II. **Non-parametric identification**

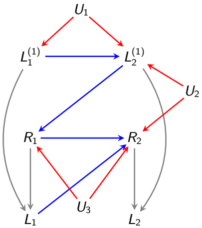
- Discussed identification tricks for full and target laws, and showed how non-identification proofs go.

Part III. **Non/Semi-parametric estimation**

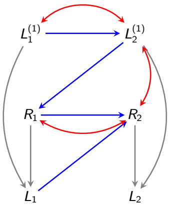
- Given and identified query, derived the non-parametric influence functions; Given three types of m-DAGs with MNAR missingness, derived the tangent space of the underlying full data and observed data distributions.

Provocative question #1: Missing data DAGs with hidden variables

- What if there exist variables that are not just missing but completely unobserved?
- Summarize the observed data distribution with a missing data acyclic directed mixed graph (ADMG).



(a) $\mathcal{G}(V, U)$



(b) $\mathcal{G}(V)$

$L_1^{(1)}$: smoking, $L_2^{(1)}$: lung cancer

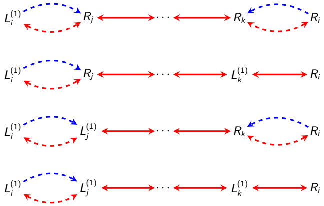
U_1: genotypic traits, U_2: occupation, U_3: ethnicity

- Missing data categorizations depend on whether districts and parents of districts of of missingness indicators contain counterfactuals, not just parents.
  - A district consists of bidirected connected components.

Theorem (Graphical characterization of identified full laws)

Full law $p(R, L^{(1)}, O)$ that is Markov relative to a missing data ADMG $\mathcal{G}$ is identified ***if and only if*** *$\mathcal{G}$ does not contain any colluding paths (Nabi et al., 2020).*

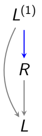
All possible colluding paths between R_i and $L_i^{(1)}$:



- The graphical condition of no colluding paths is **sound** and **complete**
- Identification functional is given by the odds ratio parameterization.

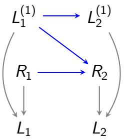
Provocative question #2: Non-identifiability

- What if a parameter of interest is provably not identified in the assumed m-DAG?



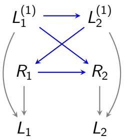
Self-censoring

$p(L^{(1)})$ not ID.



Colluder

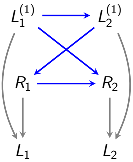
$p(L^{(1)}, R)$ not ID.



Criss-cross

$p(L^{(1)})$ not ID.

- Generally speaking, we have two general options:
  - Restrict the missing data model by posing **extra** assumptions on the full law.
  - Obtain bounds, conduct sensitivity analysis, etc.
    (Rotnitzky et al., 1998; Robins et al., 2000; Scharfstein and Irizarry, 2003)



$p(L_1^{(1)}, L_2^{(1)})$ is not identified.

Criss-cross structure

- Partial identification: $p(L_1^{(1)} \mid L_2^{(1)})$ and $\mathrm{OR}(L_1^{(1)}, L_2^{(1)})$ are identified.
  - We can test $L_1^{(1)} \perp\!\!\!\perp L_2^{(1)}$ without further assumptions.
- Under what conditions $p(L_1^{(1)}, L_2^{(1)})$ is identified? assume $p(L_1^{(1)})$ and $p(L_2^{(1)} \mid L_1^{(1)})$ follow exponential family distributions:

$$L_1^{(1)} \sim \exp\left\{\frac{l_1^{(1)}\eta_{l_1} - b_{l_1}(\eta_{l_1})}{\Phi_{l_1}} + c_{l_1}(l_1;\ \Phi_{l_1})\right\}$$

$$L_2^{(1)} \mid L_1^{(1)} \sim \exp\left\{\frac{l_2^{(1)}\eta - b(\eta)}{\Phi} + c(l_2^{(1)};\ \Phi)\right\}, \quad g(\mu(\eta)) = \alpha + \beta l_1^{(1)}.$$

- What are sufficient conditions for target law ID in the above class of distributions? (Guo et al., 2023)

Suggestion 2: Sensitivity analysis

$$
\begin{tikzcd}
L^{(1)} \arrow[d] \arrow[dd, bend right] \\
R \arrow[d] \\
L
\end{tikzcd}
$$

self-censoring

$$
\begin{aligned}
p(L^{(1)}) &= \sum_{r \in \{0,1\}} p(L^{(1)}, R = r) \\
&= p(L^{(1)} \mid R = 0) \times p(R = 0) + p(L^{(1)} \mid R = 1) \times p(R = 1).
\end{aligned}
$$

$$
p(L^{(1)} \mid R = 0) \;\propto\; p(L^{(1)} \mid R = 1) \times \exp(\gamma \; S(L^{(1)}))
$$

- The relation is controlled by the sensitivity parameter γ.
- $S(L^{(1)})$ is a specified function of $L^{(1)}$.

Provocative question #3: Testable implications in m-DAGs

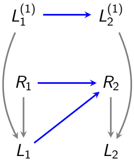
- Similar to DAGs, absence of an edge in an m-DAG implies a restriction of the form $A \perp B \mid C$. Is this restriction testable from observed finite samples?

- If all the restrictions encoded in a missing data DAG are provably untestable (i.e., no restriction on the observed data law), the full law Markov relative to the DAG is said to be **non-parametric saturated** (Robins; 1997)

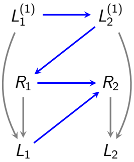
  - An example of a non-parametric saturated model is the **permutation model**.

- Submodels of a non-parametric saturated model can still be tested using partially observed data (Nabi and Bhattacharya, 2023).

Testable implications



MAR model



Permutation supermodel

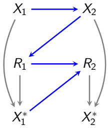
- Is $R_1 \perp\!\!\!\perp L_2^{(1)}$?
- Fit $p(R_1)$ and $p(R_1 \mid L_2^{(1)})$ and compare the goodness of fits.
- Use a weighted estimating equation to fit $p(R_1 \mid L_2^{(1)}; \alpha)$

$$\sum_{i=1}^{n} \frac{R_{2,i}}{\pi_n(R_2 = 1 \mid R_{1,i}, L_{1,i})} \times U(R_{1,i}, L_{2,i}; \alpha) = 0,$$

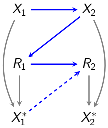
where $\mathbb{E}[U(R_1, L_2^{(1)}; \alpha)] = 0$ with respect to the full law.

Testable implications

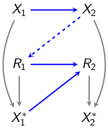
- Nabi and Bhattacharya (2023) have designed empirical tests for restrictions in three broad classes of missing data models via weighted likelihood-ratio tests and odds-ratio parameterizations.



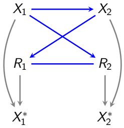
Permuation-model
(Robins; 1997)



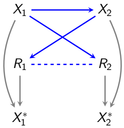
MNAR submodel



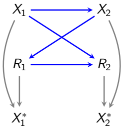
MAR submodel



No self-censoring
(Malinsky et al., 2021)



MNAR submodel



Criss-cross model

Neither full nor target law is ID.

(Nabi and Bhattacharya, 2023)

(Guo et al., 2023)

Many interesting open problems

- Missing data DAGs with or without unmeasured confounding:
  - A concise and precise representation of MNAR mechanisms.
- **Identification:**
  - Complete characterization of target law ID remains an open problem while such characterizations for full law ID exist.
  - Partial identification.
- **Estimation:**
  - Intuitive estimation strategies: IPW-style estimators.
  - An understudied research area: influence-function based estimations in m-DAGs.
- **Testable implications:**
  - Data-driven structure learning approaches.

Acknowledgments

In alphabetical order:

- Rohit Bhattacharya, Ph.D., Williams College
- Anna Guo, Emory University
- James Robins, Ph.D., Harvard School of Public Health
- Dan Scharfstein, Sc.D., University of Utah
- Ilya Shpitser, Ph.D., Johns Hopkins University
- Jiwei Zhao, Ph.D., University of Wisconsin-Madison

- Inspired by works of Judea Pearl, Karthika Mohan, Jin Tian, Eric Tchetgen Tchetgen, Daniel Malinsky, and many more.

References I

R. Bhattacharya, R. Nabi, I. Shpitser, and J. Robins. Identification in missing data models represented by directed acyclic graphs. In *Proceedings of the Thirty Fifth Conference on Uncertainty in Artificial Intelligence (UAI-35th)*. AUAI Press, 2019.

R. Bhattacharya, R. Nabi, and I. Shpitser. Semiparametric inference for causal effects in graphical models with hidden variables. *Journal of Machine Learning Research*, 23: 1–76, 2022.

H. Y. Chen. A semiparametric odds ratio model for measuring association. *biometrics*, 63:413–421, 2007.

A. Guo, J. Zhao, and R. Nabi. Sufficient identification conditions and semiparametric estimation under missing not at random mechanisms. In *Uncertainty in Artificial Intelligence*, pages 777–787. PMLR, 2023.

Y. Huang and M. Valtorta. Pearl's calculus of interventions is complete. In *Twenty Second Conference On Uncertainty in Artificial Intelligence*, 2006.

D. Malinsky, I. Shpitser, and E. J. Tchetgen Tchetgen. Semiparametric inference for nonmonotone missing-not-at-random data: the no self-censoring model. *Journal of the American Statistical Association*, pages 1–9, 2021.

References II

K. Mohan, J. Pearl, and J. Tian. Graphical models for inference with missing data. In *Advances in Neural Information Processing Systems 26*, pages 1277–1285. Curran Associates, Inc., 2013.

R. Nabi and R. Bhattacharya. On testability and goodness of fit tests in missing data models. In *Uncertainty in Artificial Intelligence*, pages 1467–1477. PMLR, 2023.

R. Nabi, R. Bhattacharya, and I. Shpitser. Full law identification in graphical models of missing data: Completeness results. In *Proceedings of the Twenty Seventh International Conference on Machine Learning (ICML-20)*, 2020.

R. Nabi, R. Bhattacharya, I. Shpitser, and J. Robins. **Causal and counterfactual views of missing data models**. *arXiv preprint arXiv:2210.05558*, 2022.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000. ISBN 0-521-77362-8.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009. ISBN 978-0521895606.

J. Peters. On the intersection property of conditional independence and its application to causal discovery. *Journal of Causal Inference*, 3(1):97–108, 2015.

References III

J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(58):2009–2053, 2014. URL `http://jmlr.org/papers/v15/peters14a.html`.

T. S. Richardson, R. J. Evans, J. M. Robins, and I. Shpitser. Nested markov properties for acyclic directed mixed graphs. *The Annals of Statistics*, 51(1):334–361, 2023.

J. M. Robins. Non-response models for the analysis of non-monotone non-ignorable missing data. *Statistics in Medicine*, 16:21–37, 1997.

J. M. Robins, A. Rotnitzky, and D. O. Scharfstein. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 1–94. Springer, 2000.

A. Rotnitzky, J. M. Robins, and D. O. Scharfstein. Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the american statistical association*, 93(444):1321–1339, 1998.

D. B. Rubin. Causal inference and missing data (with discussion). *Biometrika*, 63: 581–592, 1976.

K. Sadeghi. Faithfulness of probability distributions and graphs. *Journal of Machine Learning Research*, 18(148):1–29, 2017.

References IV

M. Sadinle and J. P. Reiter. Itemwise conditionally independent nonresponse modelling for incomplete multivariate data. *Biometrika*, 104(1):207–220, 2017.

D. O. Scharfstein and R. A. Irizarry. Generalized additive selection models for the analysis of studies with potentially nonignorable missing outcome data. *Biometrics*, 59(3):601–613, 2003.

D. O. Scharfstein, A. Rotnitzky, and J. M. Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999.

I. Shpitser. Consistent estimation of functions of data missing non-monotonically and not at random. In *Proceedings of the Thirtieth Annual Conference on Neural Information Processing Systems (NIPS-16)*. Curran Associates, Inc., 2016.

I. Shpitser and J. Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)*. AAAI Press, Palo Alto, 2006.

R. Srinivasan, R. Bhattacharya, R. Nabi, E. L. Ogburn, and I. Shpitser. Graphical models of entangled missingness. *arXiv preprint arXiv:2304.01953*, 2023.

T. S. Verma and J. Pearl. Equivalence and synthesis of causal models. Technical Report R-150, Department of Computer Science, University of California, Los Angeles, 1990.

References V

Y. Zhou, R. J. A. Little, and K. J. D. Block-conditional missing at random models for missing data. *Statistical Science*, 25(4):517–532, 2010.

Sincerely,

Razieh Nabi, Ph.D.
Rollins Assistant Professor
Department of Biostatistics and Bioinformatics
Rollins School of Public Health
Emory University

raziehnabi.bsky.social
@raziehnabi
razieh.nabi@emory.edu
https://raziehnabi.com

Slides are available at: `https://raziehnabi.com/files/acic_2025.pdf`

Upon using the materials, please cite the references appropriately.

Supplementary Materials

Block-Parallel Model

[←]

What choices of $c_{10}(L_1^{(1)}), c_{01}(L_2^{(1)}), c_{00}$ ensure orthogonality with all elements of $\Lambda_{2,1}$?

$$
\begin{aligned}
E\Bigg[& \frac{R_1 R_2}{\pi_{11}(L^{(1)})} \left\{ h(L^{(1)}) - \mu \right\} \{R_1 - \pi_1(L_2^{(1)})\} g_1(L_2^{(1)}) + \\
& \left\{ -\frac{R_1 R_2}{\pi_{11}(L^{(1)})} \pi_{10}(L^{(1)}) + R_1(1 - R_2) \right\} c_{10}(L_1^{(1)}) \{R_1 - \pi_1(L_2^{(1)})\} g_1(L_2^{(1)}) + \\
& \left\{ -\frac{R_1 R_2}{\pi_{11}(L^{(1)})} \pi_{01}(L^{(1)}) + (1 - R_1) R_2 \right\} c_{01}(L_2^{(1)}) \{R_1 - \pi_1(L_2^{(1)})\} g_1(L_2^{(1)}) + \\
& \left\{ -\frac{R_1 R_2}{\pi_{11}(L^{(1)})} \pi_{00}(L^{(1)}) + (1 - R_1)(1 - R_2) \right\} c_{00} \{R_1 - \pi_1(L_2^{(1)})\} g_1(L_2^{(1)}) \Bigg] = 0 \; .
\end{aligned}
$$

for all $g_1(L_2^{(1)})$.

Block-Parallel Model

For this to hold for all $g_1(L_2^{(1)})$, it must be the case that

$$
\begin{aligned}
E\Bigg[& \frac{R_1 R_2}{\pi_{11}(L^{(1)})}\left\{h(L^{(1)})-\mu\right\}\{R_1-\pi_1(L_2^{(1)})\}+ \\
& \left\{-\frac{R_1 R_2}{\pi_{11}(L^{(1)})}\pi_{10}(L^{(1)})+R_1(1-R_2)\right\} c_{10}(L_1^{(1)})\{R_1-\pi_1(L_2^{(1)})\}+ \\
& \left\{-\frac{R_1 R_2}{\pi_{11}(L^{(1)})}\pi_{01}(L^{(1)})+(1-R_1)R_2\right\} c_{01}(L_2^{(1)})\{R_1-\pi_1(L_2^{(1)})\}+ \\
& \left\{-\frac{R_1 R_2}{\pi_{11}(L^{(1)})}\pi_{00}(L^{(1)})+(1-R_1)(1-R_2)\right\} c_{00}\{R_1-\pi_1(L_2^{(1)})\} \Bigg| L_2^{(1)} \Bigg] = 0 \ .
\end{aligned}
\tag{1}
$$

Block-Parallel Model

What choices of $c_{10}(L_1^{(1)}), c_{01}(L_2^{(1)}), c_{00}$ ensure orthogonality with all elements of $\Lambda_{2,2}$?

$$
\begin{aligned}
E\Bigg[& \frac{R_1 R_2}{\pi_{11}(L^{(1)})}\left\{h(L^{(1)})-\mu\right\}\{R_2-\pi_2(L_1^{(1)})\}g_2(L_1^{(1)})+ \\
& \left\{-\frac{R_1 R_2}{\pi_{11}(L^{(1)})}\pi_{10}(L^{(1)})+R_1(1-R_2)\right\}c_{10}(L_1^{(1)})\{R_2-\pi_2(L_1^{(1)})\}g_2(L_1^{(1)})+ \\
& \left\{-\frac{R_1 R_2}{\pi_{11}(L^{(1)})}\pi_{01}(L^{(1)})+(1-R_1)R_2\right\}c_{01}(L_2^{(1)})\{R_2-\pi_2(L_1^{(1)})\}g_2(L_1^{(1)})+ \\
& \left\{-\frac{R_1 R_2}{\pi_{11}(L^{(1)})}\pi_{00}(L^{(1)})+(1-R_1)(1-R_2)\right\}c_{00}\{R_2-\pi_2(L_1^{(1)})\}g_2(L_1^{(1)})\Bigg]=0\ .
\end{aligned}
$$

for all $g_2(L_1^{(1)})$.

Block-Parallel Model

For this to hold for all $g_2(L_1^{(1)})$, it must be the case that

$$
\begin{aligned}
E\Bigg[& \frac{R_1 R_2}{\pi_{11}(L^{(1)})} \left\{ h(L^{(1)}) - \mu \right\} \{ R_2 - \pi_2(L_1^{(1)}) \} + \\
& \left\{ -\frac{R_1 R_2}{\pi_{11}(L^{(1)})} \pi_{10}(L^{(1)}) + R_1(1 - R_2) \right\} c_{10}(L_1^{(1)}) \{ R_2 - \pi_2(L_1^{(1)}) \} + \\
& \left\{ -\frac{R_1 R_2}{\pi_{11}(L^{(1)})} \pi_{01}(L^{(1)}) + (1 - R_1) R_2 \right\} c_{01}(L_2^{(1)}) \{ R_2 - \pi_2(L_1^{(1)}) \} + \\
& \left\{ -\frac{R_1 R_2}{\pi_{11}(L^{(1)})} \pi_{00}(L^{(1)}) + (1 - R_1)(1 - R_2) \right\} c_{00} \{ R_2 - \pi_2(L_1^{(1)}) \} \,\Bigg|\, L_1^{(1)} \Bigg] = 0 \ .
\end{aligned}
\tag{2}
$$

Block-Parallel Model

(1) implies that

$$c_{01}(L_2^{(1)}; c_{00}) = \underbrace{\frac{E\left[h(L^{(1)}) - \mu \,\middle|\, L_2^{(1)}\right]}{E\left[\pi_2(L_1^{(1)}) \,\middle|\, L_2^{(1)}\right]}}_{c_{01}(L_2^{(1)})} - \underbrace{\frac{E\left[1 - \pi_2(L_1^{(1)}) \,\middle|\, L_2^{(1)}\right]}{E\left[\pi_2(L_1^{(1)}) \,\middle|\, L_2^{(1)}\right]}}_{c'_{01}(L_2^{(1)})} c_{00}$$

(2) implies that

$$c_{10}(L_1^{(1)}; c_{00}) = \underbrace{\frac{E\left[h(L^{(1)}) - \mu \,\middle|\, L_1^{(1)}\right]}{E\left[\pi_1(L_2^{(1)}) \,\middle|\, L_1^{(1)}\right]}}_{c_{10}(L_1^{(1)})} - \underbrace{\frac{E\left[1 - \pi_1(L_2^{(1)}) \,\middle|\, L_1^{(1)}\right]}{E\left[\pi_1(L_2^{(1)}) \,\middle|\, L_1^{(1)}\right]}}_{c'_{10}(L_1^{(1)})} c_{00}$$

So, $\Lambda^{O,\perp}$ contains a collection of elements indexed by c_{00}.

Block-Parallel Model

We will work with

$$
\begin{aligned}
&\frac{R_1 R_2}{\pi_{11}(L^{(1)})} \left\{ h(L^{(1)}) - \mu \right\} + \\
&\left\{ -\frac{R_1 R_2}{\pi_{11}(L^{(1)})} \pi_{10}(L^{(1)}) + R_1(1 - R_2) \right\} \left\{ c_{10}(L_1^{(1)}) - c_{10}'(L_1^{(1)}) c_{00} \right\} + \\
&\left\{ -\frac{R_1 R_2}{\pi_{11}(L^{(1)})} \pi_{01}(L^{(1)}) + (1 - R_1) R_2 \right\} \left\{ c_{01}(L_2^{(1)}) - c_{01}'(L_2^{(1)}) c_{00} \right\} + \\
&\left\{ -\frac{R_1 R_2}{\pi_{11}(L^{(1)})} \pi_{00}(L^{(1)}) + (1 - R_1)(1 - R_2) \right\} c_{00}
\end{aligned}
$$

Block-Parallel Model

Need estimators for

- $\pi_1(L_2^{(1)})$
- $\pi_2(L_1^{(1)})$
- $E[\pi_2(L_1^{(1)})|L_2^{(1)}]$
- $E[\pi_1(L_2^{(1)})|L_1^{(1)}]$
- $E[h(L^{(1)})|L_1^{(1)}]$
- $E[h(L^{(1)})|L_2^{(1)}]$

Block-Parallel Model

To find the optimal choice of c_{00}, minimize

$$E\left[\left\{\frac{R_1R_2}{\pi_{11}(L^{(1)})}\left\{h(L^{(1)})-\mu\right\}+\right.\right.$$
$$\left\{-\frac{R_1R_2}{\pi_{11}(L^{(1)})}\pi_{10}(L^{(1)})+R_1(1-R_2)\right\}\left\{c_{10}(L_1^{(1)})-c'_{10}(L_1^{(1)})c_{00}\right\}+$$
$$\left\{-\frac{R_1R_2}{\pi_{11}(L^{(1)})}\pi_{01}(L^{(1)})+(1-R_1)R_2\right\}\left\{c_{01}(L_2^{(1)})-c'_{01}(L_2^{(1)})c_{00}\right\}+$$
$$\left.\left.\left\{-\frac{R_1R_2}{\pi_{11}(L^{(1)})}\pi_{00}(L^{(1)})+(1-R_1)(1-R_2)\right\}c_{00}\right\}^2\right]$$

Block-Parallel Model

Set derivative with respect to c_{00} equal to zero.

$$
\begin{aligned}
E\Bigg[\Bigg\{ & \frac{R_1 R_2}{\pi_{11}(L^{(1)})}\left\{h(L^{(1)})-\mu\right\}+ \\
& \left\{-\frac{R_1 R_2}{\pi_{11}(L^{(1)})}\pi_{10}(L^{(1)})+R_1(1-R_2)\right\}\left\{c_{10}(L_1^{(1)})\right\}+ \\
& \left\{-\frac{R_1 R_2}{\pi_{11}(L^{(1)})}\pi_{10}(L^{(1)})+R_1(1-R_2)\right\}\left\{-c_{10}'(L_1^{(1)})c_{00}\right\}+ \\
& \left\{-\frac{R_1 R_2}{\pi_{11}(L^{(1)})}\pi_{01}(L^{(1)})+(1-R_1)R_2\right\}\left\{c_{01}(L_2^{(1)})\right\}+ \\
& \left\{-\frac{R_1 R_2}{\pi_{11}(L^{(1)})}\pi_{01}(L^{(1)})+(1-R_1)R_2\right\}\left\{-c_{01}'(L_2^{(1)})c_{00}\right\}+ \\
& \left\{-\frac{R_1 R_2}{\pi_{11}(L^{(1)})}\pi_{00}(L^{(1)})+(1-R_1)(1-R_2)\right\}c_{00}\Bigg\}\times \\
\Bigg\{ & \left\{-\frac{R_1 R_2}{\pi_{11}(L^{(1)})}\pi_{10}(L^{(1)})+R_1(1-R_2)\right\}\left\{-c_{10}'(L_1^{(1)})\right\}+ \\
& \left\{-\frac{R_1 R_2}{\pi_{11}(L^{(1)})}\pi_{01}(L^{(1)})+(1-R_1)R_2\right\}\left\{-c_{01}'(L_2^{(1)})\right\}+ \\
& \left\{-\frac{R_1 R_2}{\pi_{11}(L^{(1)})}\pi_{00}(L^{(1)})+(1-R_1)(1-R_2)\right\}\Bigg\}\Bigg]=0
\end{aligned}
$$

Block-Parallel Model

$$c_{00} = \frac{a}{b}$$

where

$$\begin{aligned}
b \quad = \quad & -E\left[\left\{\frac{\pi_{10}(L^{(1)})}{\pi_{11}(L^{(1)})}+1\right\}\pi_{10}(L^{(1)})\left\{c'_{10}(L_1^{(1)})\right\}^2\right] - \\
& E\left[\left\{\frac{\pi_{10}(L^{(1)})\pi_{01}(L^{(1)})}{\pi_{11}(L^{(1)})}\right\}\left\{2c'_{10}(L_1^{(1)})c'_{01}(L_2^{(1)})\right\}\right] + \\
& E\left[\left\{\frac{\pi_{10}(L^{(1)})\pi_{00}(L^{(1)})}{\pi_{11}(L^{(1)})}\right\}\left\{2c'_{10}(L_1^{(1)})\right\}\right] - \\
& E\left[\left\{\frac{\pi_{01}(L^{(1)})}{\pi_{11}(L^{(1)})}+1\right\}\pi_{01}(L^{(1)})\left\{c'_{01}(L_2^{(1)})\right\}^2\right] + \\
& E\left[\left\{\frac{\pi_{01}(L^{(1)})\pi_{00}(L^{(1)})}{\pi_{11}(L^{(1)})}\right\}\left\{2c'_{01}(L_2^{(1)})\right\}\right] + \\
& E\left[\left\{\frac{\pi_{00}(L^{(1)})}{\pi_{11}(L^{(1)})}+1\right\}\pi_{00}(L^{(1)})\right]
\end{aligned}$$

Block-Parallel Model

$$c_{00} = \frac{a}{b}$$

where

$$\begin{aligned}
a \quad = \quad & E\left[\left\{\frac{\pi_{10}(L^{(1)})}{\pi_{11}(L^{(1)})}\left\{h(L^{(1)}) - \mu\right\}\right\}\left\{c'_{10}(L_1^{(1)})\right\}\right] + \\
& E\left[\left\{\frac{\pi_{01}(L^{(1)})}{\pi_{11}(L^{(1)})}\left\{h(L^{(1)}) - \mu\right\}\right\}\left\{c'_{01}(L_2^{(1)})\right\}\right] - \\
& E\left[\left\{\frac{\pi_{00}(L^{(1)})}{\pi_{11}(L^{(1)})}\left\{h(L^{(1)}) - \mu\right\}\right\}\right] - \\
& E\left[\left\{\frac{\pi_{10}(L^{(1)})}{\pi_{11}(L^{(1)})} + 1\right\}\pi_{10}(L^{(1)})\left\{c_{10}(L_1^{(1)})\right\}\left\{c'_{10}(L_1^{(1)})\right\}\right] - \\
& E\left[\left\{\frac{\pi_{10}(L^{(1)})\pi_{01}(L^{(1)})}{\pi_{11}(L^{(1)})}\right\}\left\{c_{10}(L_1^{(1)})c'_{01}(L_2^{(1)}) + c_{01}(L_2^{(1)})c'_{10}(L_1^{(1)})\right\}\right] + \\
& E\left[\left\{\frac{\pi_{10}(L^{(1)})\pi_{00}(L^{(1)})}{\pi_{11}(L^{(1)})}\right\}\left\{c_{10}(L_1^{(1)})\right\}\right] - \\
& E\left[\left\{\frac{\pi_{01}(L^{(1)})}{\pi_{11}(L^{(1)})} + 1\right\}\pi_{01}(L^{(1)})\left\{c_{01}(L_2^{(1)})\right\}\left\{c'_{01}(L_2^{(1)})\right\}\right] + \\
& E\left[\left\{\frac{\pi_{01}(L^{(1)})\pi_{00}(L^{(1)})}{\pi_{11}(L^{(1)})}\right\}\left\{c_{01}(L_2^{(1)})\right\}\right]
\end{aligned}$$

[←]

Block-Conditional Model

What choices of $c_{10}(L_1^{(1)}), c_{01}(L_2^{(1)}), c_{00}$ ensure orthogonality with all elements of $\Lambda_{2,1}$?

$$
\begin{aligned}
E\Bigg[& \frac{R_1 R_2}{\pi_{11}(L^{(1)})}\left\{h(L^{(1)}) - \mu\right\}\{R_1 - \pi_1\}g_1 + \\
& \left\{ -\frac{R_1 R_2}{\pi_{11}(L^{(1)})}\pi_{10}(L^{(1)}) + R_1(1 - R_2) \right\} c_{10}(L_1^{(1)})\{R_1 - \pi_1\}g_1 + \\
& \left\{ -\frac{R_1 R_2}{\pi_{11}(L^{(1)})}\pi_{01}(L^{(1)}) + (1 - R_1)R_2 \right\} c_{01}(L_2^{(1)})\{R_1 - \pi_1\}g_1 + \\
& \left\{ -\frac{R_1 R_2}{\pi_{11}(L^{(1)})}\pi_{00}(L^{(1)}) + (1 - R_1)(1 - R_2) \right\} c_{00}\{R_1 - \pi_1\}g_1 \Bigg] = 0 \; .
\end{aligned}
$$

for all g_1.

Block-Conditional Model

For this to hold for all g_1, it must be the case that

$$
\begin{aligned}
E\Bigg[& \frac{R_1 R_2}{\pi_{11}(L^{(1)})} \left\{ h(L^{(1)}) - \mu \right\} \{R_1 - \pi_1\} + \\
& \left\{ -\frac{R_1 R_2}{\pi_{11}(L^{(1)})} \pi_{10}(L^{(1)}) + R_1(1 - R_2) \right\} c_{10}(L_1^{(1)}) \{R_1 - \pi_1\} + \\
& \left\{ -\frac{R_1 R_2}{\pi_{11}(L^{(1)})} \pi_{01}(L^{(1)}) + (1 - R_1) R_2 \right\} c_{01}(L_2^{(1)}) \{R_1 - \pi_1\} + \\
& \left\{ -\frac{R_1 R_2}{\pi_{11}(L^{(1)})} \pi_{00}(L^{(1)}) + (1 - R_1)(1 - R_2) \right\} c_{00} \{R_1 - \pi_1\} \Bigg] = 0 \qquad (3)
\end{aligned}
$$

Block-Conditional Model

What choices of $c_{10}(L_1^{(1)}), c_{01}(L_2^{(1)}), c_{00}$ ensure orthogonality with all elements of $\Lambda_{2,2,1}$?

$$
\begin{aligned}
E\Bigg[& \frac{R_1 R_2}{\pi_{11}(L^{(1)})} \left\{ h(L^{(1)}) - \mu \right\} R_1\{R_2 - \pi_2(1, L_1^{(1)})\} g_2(1, L_1^{(1)}) + \\
& \left\{ -\frac{R_1 R_2}{\pi_{11}(L^{(1)})} \pi_{10}(L^{(1)}) + R_1(1 - R_2) \right\} c_{10}(L_1^{(1)}) R_1\{R_2 - \pi_2(1, L_1^{(1)})\} g_2(1, L_1^{(1)}) + \\
& \left\{ -\frac{R_1 R_2}{\pi_{11}(L^{(1)})} \pi_{01}(L^{(1)}) + (1 - R_1) R_2 \right\} c_{01}(L_2^{(1)}) R_1\{R_2 - \pi_2(1, L_1^{(1)})\} g_2(1, L_1^{(1)}) + \\
& \left\{ -\frac{R_1 R_2}{\pi_{11}(L^{(1)})} \pi_{00}(L^{(1)}) + (1 - R_1)(1 - R_2) \right\} c_{00} R_1\{R_2 - \pi_2(1, L_1^{(1)})\} g_2(1, L_1^{(1)}) \Bigg] = 0
\end{aligned}
$$

for all $g_2(1, L_1^{(1)})$.

Block-Conditional Model

For this to hold for all $g_2(1, L_2^{(1)})$, it must be the case that

$$
\begin{aligned}
E \Bigg[& \frac{R_1 R_2}{\pi_{11}(L^{(1)})} \left\{ h(L^{(1)}) - \mu \right\} R_1 \{ R_2 - \pi_2(1, L_1^{(1)}) + \\
& \left\{ -\frac{R_1 R_2}{\pi_{11}(L^{(1)})} \pi_{10}(L^{(1)}) + R_1(1 - R_2) \right\} c_{10}(L_1^{(1)}) R_1 \{ R_2 - \pi_2(1, L_1^{(1)}) + \\
& \left\{ -\frac{R_1 R_2}{\pi_{11}(L^{(1)})} \pi_{01}(L^{(1)}) + (1 - R_1) R_2 \right\} c_{01}(L_2^{(1)}) R_1 \{ R_2 - \pi_2(1, L_1^{(1)}) + \\
& \left\{ -\frac{R_1 R_2}{\pi_{11}(L^{(1)})} \pi_{00}(L^{(1)}) + (1 - R_1)(1 - R_2) \right\} c_{00} R_1 \{ R_2 - \pi_2(1, L_1^{(1)}) \Bigg| L_1^{(1)} \Bigg] = 0 \ .
\end{aligned}
\tag{4}
$$

Block-Conditional Model

What choices of $c_{10}(L_1^{(1)}), c_{01}(L_2^{(1)}), c_{00}$ ensure orthogonality with all elements of $\Lambda_{2,2,0}$?

$$
\begin{aligned}
&E\left[\frac{R_1R_2}{\pi_{11}(L^{(1)})}\left\{h(L^{(1)})-\mu\right\}(1-R_1)\{R_2-\pi_2(0,L_1^{(1)})\}g_2(0,L_1^{(1)})+\right.\\
&\left\{-\frac{R_1R_2}{\pi_{11}(L^{(1)})}\pi_{10}(L^{(1)})+R_1(1-R_2)\right\}c_{10}(L_1^{(1)})(1-R_1)\{R_2-\pi_2(0,L_1^{(1)})\}g_2(0,L_1^{(1)})+\\
&\left\{-\frac{R_1R_2}{\pi_{11}(L^{(1)})}\pi_{01}(L^{(1)})+(1-R_1)R_2\right\}c_{01}(L_2^{(1)})(1-R_1)\{R_2-\pi_2(0,L_1^{(1)})\}g_2(0,L_1^{(1)})+\\
&\left.\left\{-\frac{R_1R_2}{\pi_{11}(L^{(1)})}\pi_{00}(L^{(1)})+(1-R_1)(1-R_2)\right\}c_{00}(1-R_1)\{R_2-\pi_2(0,L_1^{(1)})\}g_2(0,L_1^{(1)})\right]\\
&=0
\end{aligned}
$$

for all $g_2(0, L_1^{(1)})$.

Block-Conditional Model

For this to hold for all $g_2(0, L_2^{(1)})$, it must be the case that

$$
\begin{aligned}
E \Bigg[& \frac{R_1 R_2}{\pi_{11}(L^{(1)})} \left\{ h(L^{(1)}) - \mu \right\} (1 - R_1)\{R_2 - \pi_2(0, L_1^{(1)})\} + \\
& \left\{ -\frac{R_1 R_2}{\pi_{11}(L^{(1)})} \pi_{10}(L^{(1)}) + R_1(1 - R_2) \right\} c_{10}(L_1^{(1)})(1 - R_1)\{R_2 - \pi_2(0, L_1^{(1)})\} + \\
& \left\{ -\frac{R_1 R_2}{\pi_{11}(L^{(1)})} \pi_{01}(L^{(1)}) + (1 - R_1)R_2 \right\} c_{01}(L_2^{(1)})(1 - R_1)\{R_2 - \pi_2(0, L_1^{(1)})\} + \\
& \left\{ -\frac{R_1 R_2}{\pi_{11}(L^{(1)})} \pi_{00}(L^{(1)}) + (1 - R_1)(1 - R_2) \right\} c_{00}(1 - R_1)\{R_2 - \pi_2(0, L_1^{(1)})\} \Bigg| L_1^{(1)} \Bigg] = 0
\end{aligned}
\tag{5}
$$

Block-Conditional Model

(4) implies

$$E[c_{01}(L_2^{(1)})|L_1^{(1)}] = c_{00}$$

(5) implies

$$E[\pi_2(0, L_1^{(1)})c_{01}(L_2^{(1)})] + c_{00}E[1 - \pi_2(0, L_1^{(1)})] = 0$$

(4) and (5) imply that

$$c_{00}E[\pi_2(0, L_1^{(1)})] + c_{00}E[1 - \pi_2(0, L_1^{(1)})] = c_{00} = 0$$

Block-Conditional Model

What choices of $c_{01}(L_2^{(1)})$ make $E[c_{01}(L_2^{(1)})|L_1^{(1)}] = 0$?

- Fredholm integral equation of the first kind.
- Obviously, $c_{01}(L_2^{(1)}) = 0$
- Non-trivial choices may or may not exist depending on the conditional distribution of $L_2^{(1)}$ given $L_1^{(1)}$.
- If the conditional distribution of $L_2^{(1)}$ given $L_1^{(1)}$ is from a canonical exponential family, then $c_{01}(L_2^{(1)}) = 0$ a.s.

Block-Conditional Model

With $c_{01}(L_2^{(1)}) = c_{00} = 0$, (3) implies

$$c_{10}(L_1^{(1)}) = \frac{E[h(L^{(1)}) - \mu | L_1^{(1)}]}{\pi_1}$$

Block-Conditional Model

We will work with

$$\frac{R_1 R_2}{\pi_{11}(L^{(1)})}\left\{h(L^{(1)})-\mu\right\}+\left\{-\frac{R_1 R_2}{\pi_{11}(L^{(1)})}\pi_{10}(L^{(1)})+R_1(1-R_2)\right\}\frac{E[h(L^{(1)})-\mu|L_1^{(1)}]}{\pi_1}$$

$$=\frac{R_1}{\pi_1}\left\{\frac{R_2}{\pi_2(1,L_1^{(1)})}h(L^{(1)})+\left(1-\frac{R_2}{\pi_2(1,L_1^{(1)})}\right)E[h(L^{(1)})|L_1^{(1)}]-\mu\right\}$$

Need estimators for

- π_1
- $\pi_2(1, L_1^{(1)})$
- $E[h(L^{(1)})|L_1^{(1)}]$

[←]

Permutation Model

What choices of $c_{10}(L_1^{(1)}), c_{01}(L_2^{(1)}), c_{00}$ ensure orthogonality with all elements of $\Lambda_{2,1}$?

$$
\begin{aligned}
E \Bigg[& \frac{R_1 R_2}{\pi_{11}(L^{(1)})} \left\{ h(L^{(1)}) - \mu \right\} \{R_1 - \pi_1(L_2^{(1)})\} g_1(L_2^{(1)}) + \\
& \left\{ -\frac{R_1 R_2}{\pi_{11}(L^{(1)})} \pi_{10}(L^{(1)}) + R_1(1 - R_2) \right\} c_{10}(L_1^{(1)}) \{R_1 - \pi_1(L_2^{(1)})\} g_1(L_2^{(1)}) + \\
& \left\{ -\frac{R_1 R_2}{\pi_{11}(L^{(1)})} \pi_{01}(L^{(1)}) + (1 - R_1) R_2 \right\} c_{01}(L_2^{(1)}) \{R_1 - \pi_1(L_2^{(1)})\} g_1(L_2^{(1)}) + \\
& \left\{ -\frac{R_1 R_2}{\pi_{11}(L^{(1)})} \pi_{00}(L^{(1)}) + (1 - R_1)(1 - R_2) \right\} c_{00} \{R_1 - \pi_1(L_2^{(1)})\} g_1(L_2^{(1)}) \Bigg] = 0 \ .
\end{aligned}
$$

for all $g_1(L_2^{(1)})$

Permutation Model

For this to hold for all $g_1(L_2^{(1)})$, it must be the case that

$$E\left[\frac{R_1R_2}{\pi_{11}(L^{(1)})}\left\{h(L^{(1)})-\mu\right\}\{R_1-\pi_1(L_2^{(1)})\}+\right.$$
$$\left\{-\frac{R_1R_2}{\pi_{11}(L^{(1)})}\pi_{10}(L^{(1)})+R_1(1-R_2)\right\}c_{10}(L_1^{(1)})\{R_1-\pi_1(L_2^{(1)})\}+$$
$$\left\{-\frac{R_1R_2}{\pi_{11}(L^{(1)})}\pi_{01}(L^{(1)})+(1-R_1)R_2\right\}c_{01}(L_2^{(1)})\{R_1-\pi_1(L_2^{(1)})\}+$$
$$\left.\left\{-\frac{R_1R_2}{\pi_{11}(L^{(1)})}\pi_{00}(L^{(1)})+(1-R_1)(1-R_2)\right\}c_{00}\{R_1-\pi_1(L_2^{(1)})\}\,\middle|\,L_2^{(1)}\right]=0\ . \tag{6}$$

Permutation Model

What choices of $c_{10}(L_1^{(1)}), c_{01}(L_2^{(1)}), c_{00}$ ensure orthogonality with all elements of $\Lambda_{2,2,1}$?

$$
\begin{aligned}
E\Bigg[& \frac{R_1R_2}{\pi_{11}(L^{(1)})}\left\{h(L^{(1)})-\mu\right\} R_1\{R_2-\pi_2(1,L_1^{(1)})\}g_2(1,L_1^{(1)})+ \\
& \left\{-\frac{R_1R_2}{\pi_{11}(L^{(1)})}\pi_{10}(L^{(1)})+R_1(1-R_2)\right\} c_{10}(L_1^{(1)})R_1\{R_2-\pi_2(1,L_1^{(1)})\}g_2(1,L_1^{(1)})+ \\
& \left\{-\frac{R_1R_2}{\pi_{11}(L^{(1)})}\pi_{01}(L^{(1)})+(1-R_1)R_2\right\} c_{01}(L_2^{(1)})R_1\{R_2-\pi_2(1,L_1^{(1)})\}g_2(1,L_1^{(1)})+ \\
& \left\{-\frac{R_1R_2}{\pi_{11}(L^{(1)})}\pi_{00}(L^{(1)})+(1-R_1)(1-R_2)\right\} c_{00}R_1\{R_2-\pi_2(1,L_1^{(1)})\}g_2(1,L_1^{(1)})\Bigg]=0\ .
\end{aligned}
$$

for all $g_2(1, L_1^{(1)})$

Permutation Model

For this to hold for all $g_2(1, L_1^{(1)})$, it must be the case that

$$E\left[\frac{R_1 R_2}{\pi_{11}(L^{(1)})}\left\{h(L^{(1)})-\mu\right\} R_1\{R_2-\pi_2(1, L_1^{(1)})\}+\right.$$
$$\left\{-\frac{R_1 R_2}{\pi_{11}(L^{(1)})}\pi_{10}(L^{(1)})+R_1(1-R_2)\right\} c_{10}(L_1^{(1)}) R_1\{R_2-\pi_2(1, L_1^{(1)})\}+$$
$$\left\{-\frac{R_1 R_2}{\pi_{11}(L^{(1)})}\pi_{01}(L^{(1)})+(1-R_1)R_2\right\} c_{01}(L_2^{(1)}) R_1\{R_2-\pi_2(1, L_1^{(1)})\}+$$
$$\left.\left\{-\frac{R_1 R_2}{\pi_{11}(L^{(1)})}\pi_{00}(L^{(1)})+(1-R_1)(1-R_2)\right\} c_{00} R_1\{R_2-\pi_2(1, L_1^{(1)})\}\,\right|\, L_1^{(1)}\Bigg]=0\ . \tag{7}$$

Permutation Model

What choices of $c_{10}(L_1^{(1)}), c_{01}(L_2^{(1)}), c_{00}$ ensure orthogonality with all elements of $\Lambda_{2,2,0}$?

$$
\begin{aligned}
E\Bigg[& \frac{R_1 R_2}{\pi_{11}(L^{(1)})}\left\{h(L^{(1)}) - \mu\right\}(1 - R_1)\{R_2 - \pi_2(0)\}g_2(0) + \\
& \left\{-\frac{R_1 R_2}{\pi_{11}(L^{(1)})}\pi_{10}(L^{(1)}) + R_1(1 - R_2)\right\} c_{10}(L_1^{(1)})(1 - R_1)\{R_2 - \pi_2(0)\}g_2(0) + \\
& \left\{-\frac{R_1 R_2}{\pi_{11}(L^{(1)})}\pi_{01}(L^{(1)}) + (1 - R_1)R_2\right\} c_{01}(L_2^{(1)})(1 - R_1)\{R_2 - \pi_2(0)\}g_2(0) + \\
& \left\{-\frac{R_1 R_2}{\pi_{11}(L^{(1)})}\pi_{00}(L^{(1)}) + (1 - R_1)(1 - R_2)\right\} c_{00}(1 - R_1)\{R_2 - \pi_2(0)\}g_2(0)\Bigg] = 0
\end{aligned}
$$

for all $g_2(0)$

Permutation Model

For this to hold for all $g_2(0)$, it must be the case that

$$
\begin{aligned}
E\Bigg[& \frac{R_1 R_2}{\pi_{11}(L^{(1)})}\left\{h(L^{(1)})-\mu\right\}(1-R_1)\{R_2-\pi_2(0)\}+ \\
& \left\{-\frac{R_1 R_2}{\pi_{11}(L^{(1)})}\pi_{10}(L^{(1)})+R_1(1-R_2)\right\} c_{10}(L_1^{(1)})(1-R_1)\{R_2-\pi_2(0)\}+ \\
& \left\{-\frac{R_1 R_2}{\pi_{11}(L^{(1)})}\pi_{01}(L^{(1)})+(1-R_1)R_2\right\} c_{01}(L_2^{(1)})(1-R_1)\{R_2-\pi_2(0)\}+ \\
& \left\{-\frac{R_1 R_2}{\pi_{11}(L^{(1)})}\pi_{00}(L^{(1)})+(1-R_1)(1-R_2)\right\} c_{00}(1-R_1)\{R_2-\pi_2(0)\}\Bigg]=0\ .
\end{aligned} \tag{8}
$$

Permutation Model

(8) implies

$$c_{00} = \frac{E[\{1-\pi_1(L_2^{(1)})\}c_{01}(L_2^{(1)})]}{E[\{1-\pi_1(L_2^{(1)})\}]}$$

(6) implies

$$c_{01}(L_2^{(1)}) = \frac{E[\left\{h(L^{(1)})-\mu\right\}|L_2^{(1)}]}{\pi_2(0)} - \frac{1-\pi_2(0)}{\pi_2(0)}c_{00}$$

Together, this implies that

$$c_{00} = \frac{E[h(L^{(1)})\{1-\pi_1(L_2^{(1)})\}]}{E[1-\pi_1(L_2^{(1)})]} - \mu$$

and

$$c_{01}(L_2^{(1)}) = \frac{E[h(L^{(1)})|L_2^{(1)}]}{\pi_2(0)} - \frac{1-\pi_2(0)}{\pi_2(0)}\left\{\frac{E[h(L^{(1)})\{1-\pi_1(L_2^{(1)})\}]}{E[1-\pi_1(L_2^{(1)})]}\right\} - \mu$$

Permutation Model

Adding (7) implies

$$c_{10}(L_1^{(1)}) \quad = \quad \frac{E[h(L^{(1)}) - E[h(L^{(1)})|L_2^{(1)}]|L_1^{(1)}]}{E[\pi_1(L_2^{(1)})|L_1^{(1)}]} + \frac{E[\pi_1(L_2^{(1)})E[h(L^{(1)})|L_2^{(1)}]|L_1^{(1)}]}{E[\pi_1(L_2^{(1)})|L_1^{(1)}]} - \mu$$

Permutation Model

We will work with

$$
\begin{aligned}
&\frac{R_1 R_2}{\pi_{11}(L^{(1)})} h(L^{(1)}) + \\
&\left\{ -\frac{R_1 R_2}{\pi_{11}(L^{(1)})} \pi_{10}(L^{(1)}) + R_1(1 - R_2) \right\} \times \\
&\quad \left\{ \frac{E[h(L^{(1)}) - E[h(L^{(1)})|L_2^{(1)}]|L_1^{(1)}]}{E[\pi_1(L_2^{(1)})|L_1^{(1)}]} + \frac{E[\pi_1(L_2^{(1)}) E[h(L^{(1)})|L_2^{(1)}]|L_1^{(1)}]}{E[\pi_1(L_2^{(1)})|L_1^{(1)}]} \right\} + \\
&\left\{ -\frac{R_1 R_2}{\pi_{11}(L^{(1)})} \pi_{01}(L^{(1)}) + (1 - R_1) R_2 \right\} \times \\
&\quad \left\{ \frac{E[h(L^{(1)})|L_2^{(1)}]}{\pi_2(0)} - \frac{1 - \pi_2(0)}{\pi_2(0)} \left\{ \frac{E[h(L^{(1)})\{1 - \pi_1(L_2^{(1)})\}]}{E[1 - \pi_1(L_2^{(1)})]} \right\} \right\} + \\
&\left\{ -\frac{R_1 R_2}{\pi_{11}(L^{(1)})} \pi_{00}(L^{(1)}) + (1 - R_1)(1 - R_2) \right\} \times \\
&\quad \left\{ \frac{E[h(L^{(1)})\{1 - \pi_1(L_2^{(1)})\}]}{E[1 - \pi_1(L_2^{(1)})]} \right\} - \mu
\end{aligned}
$$

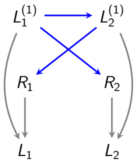
Permutation Model

Need estimators for

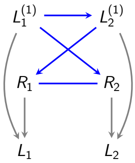
- $\pi_2(0)$
- $\pi_2(1, L_1^{(1)})$
- $\pi_1(L_2^{(1)})$
- Conditional means of functions of $L_2^{(1)}$ given $L_1^{(1)}$ and conditional means of functions of $L_1^{(1)}$ given $L_2^{(1)}$
  - $E[h(L^{(1)})|L_1^{(1)}]$
  - $E[h(L^{(1)})|L_2^{(1)}]$
  - $E[\pi_1(L_2^{(1)})|L_1^{(1)}]$
  - $E[\pi_1(L_2^{(1)})E[h(L^{(1)})|L_2^{(1)}]|L_1^{(1)}]$
  - $E[E[h(L^{(1)})|L_2^{(1)}]|L_1^{(1)}]]$

[←]

No self-censoring chain graph model



MNAR submodel

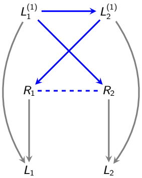


No self-censoring supermodel

- Is $R_1 \perp\!\!\!\perp R_2 \mid L^{(1)}$?
- This translates to whether or not $\mathrm{OR}(R_1, R_2 \mid L^{(1)}) = 1$?
- Identify the odds ratio using symmetric argument:

$$
\begin{aligned}
\mathrm{OR}(R_1 = 0, R_2 = 0 \mid L_1^{(1)}, L_2^{(1)}) &= \frac{p(R_1 = r_1 \mid R_2 = r_2, L_2)}{p(R_1 = 1 \mid R_2 = r_2, L_2)} \times \frac{p(R_1 = 1 \mid R_2 = 1, L_2)}{p(R_1 = r_1 \mid R_2 = 1, L_2)} \\
&= \frac{p(R_2 = r_2 \mid R_1 = r_1, L_1)}{p(R_2 = 1 \mid R_1 = r_1, L_1)} \times \frac{p(R_2 = 1 \mid R_1 = 1, L_1)}{p(R_2 = r_2 \mid R_1 = 1, L_1)} \\
&= f(R_1, R_2).
\end{aligned}
$$

Identification and estimation of odds ratio



$$\mathrm{OR}(R_1 = 0, R_2 = 0 \mid L_1^{(1)}, L_2^{(1)})$$

- Estimating equation that is mean zero under the truth ($\mathbb{E}[U] = 0$)

$$\begin{aligned}
U =& \frac{R_1 R_2}{p(R = 1 \mid X^{(1)})} \times p(R = 0 \mid L^{(1)}) - (1 - R_1)(1 - R_2) \\
=& \frac{R_1 R_2}{p(R_1 = 1 \mid R_2 = 1, L_2^{(1)}) \times p(R_2 = 1 \mid R_1 = 1, L_1^{(1)})} \\
& \times p(R_1 = 0 \mid R_2 = 1, L_2^{(1)}) \times p(R_2 = 0 \mid R_1 = 1, L_1^{(1)}) \times \mathrm{OR}(R_1 = R_2 = 0 \mid L^{(1)}) \\
& - (1 - R_1)(1 - R_2)
\end{aligned}$$