

# Causal Graphical Methods For Handling Nonignorable Missing Data

**Razieh Nabi, Ph.D.**

Department of Biostatistics and Bioinformatics  
Rollins School of Public Health, Emory University  
✉ [razielh.nabi@emory.edu](mailto:razielh.nabi@emory.edu)

**Daniel Scharfstein, Sc.D.**

Department of Population Health Sciences  
University of Utah School of Medicine  
✉ [daniel.scharfstein@hsc.utah.edu](mailto:daniel.scharfstein@hsc.utah.edu)

Slides are available at: [https://razielnabi.com/files/acic\\_miss.pdf](https://razielnabi.com/files/acic_miss.pdf)  
(Upon using the materials, please cite appropriately.)

# Motivation

- | Missing data or missing values occur when no data value is stored for one or multiple variables in an observational unit.
- | We are often interested in a functional of an underlying distribution:

- | Population-level outcome  $Y$ :

$$\mu_1 = E[Y]$$

- | Outcome mean in a sub-population  $X = x$ :

$$\mu_2 = E[Y \mid X = x]$$

- | Average causal effect of binary treatment  $T$  on outcome  $Y$ :

$$\mu_3 = E[E(Y \mid X, T = 1) - E(Y \mid X, T = 0)]$$

(Note that  $\mu_3$  can only be interpreted as a causal effect under certain assumptions, such as *consistency*, *positivity*, and *conditional ignorability*.)

- | The question is how to compute these estimands?
  - | We need replicates of the underlying distribution, e.g.,  
 $\{X, T, Y\} \sim p(X, T, Y)$

# 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 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 can we use a sample 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** resulting from 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

$X_{\text{obs}}$ : observed entries,  $X_{\text{miss}}$ : missing entries

$R$  : missingness indicators

---

1. **MCAR:**  $R \perp X_{\text{miss}}, X_{\text{obs}}, Z, \quad p(R | X, Z) = p(R).$

| probability that any observation is missing is independent of all data values, regardless of whether they are observed or unobserved.

2. **MAR:**  $R \perp X_{\text{miss}} | Z, X_{\text{obs}}, \quad p(R | X, Z) = p(R | Z, X_{\text{obs}})$

| probability that any observation is missing depends only on elements that are observed.

3. **MNAR:**  $R \not\perp X_{\text{miss}}$  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 does not determine the best approach for handling missing data in multiple variables.
- | How best to handle missing data?
  - | It depends on the assumed 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 underlying **assumptions** more explicit, and
  - | Use **identification** procedures as a guide for **estimation** methods.



# Course outline and objectives

## Part I. Missing data DAGs

- | Representing missingness mechanisms graphically; transferring expert knowledge into a concise graphical representation; interpret a missing data DAG model as a class of distributions with a set of independence restrictions.

## Part II. Non-parametric identification

- | Arguing for identifiability of a given estimand: write it down as a function of observed data, or prove it is not identified without making further assumptions.

## Part III. Non/Semi-parametric estimation

- | Given an identified estimand, derive the class of influence functions; Derive the efficient influence function in a given missing data DAG.

## Part IV. Sensitivity analysis

- | Assessing deviations from assumptions.

## **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: get inspired by 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 about the distribution 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:**  $ACE := E[Y^{(1)} - Y^{(0)}]$

- |  $Y^{(t)}$ : potential outcome  $Y$  when binary treatment  $T$  is assigned to  $t = \{0, 1\}$ .

2.  $\mathcal{M}$ : a causal model relating counterfactuals to factials

- | **Consistency:** observed outcome  $Y$  is equal to the potential outcome  $Y^{(t)}$  when the treatment received is  $T = t$ ,
- | **Positivity:**  $p(T = t / X = x) > 0$  for all  $x$  in the state space of  $X$ ,
- | **Conditional ignorability:**  $Y^{(t)} \perp T / X$

3. Under the above causal model, we can identify ACE via the following functional, known as **adjustment formula** or **g-formula**:

$$E[Y^{(1)} - Y^{(0)}] = E[E(Y | T = 1, X) - E(Y | T = 0, X)]$$

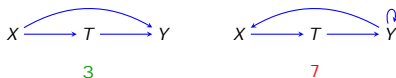
4, 5. Luckily, we are not short of any estimation or sensitivity analysis techniques!

# Directed acyclic graph (DAG)

- | The second step in the causal workflow is what distinguishes causal analyses from traditional statistical analyses.
  - | 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 analyses from traditional statistical analyses.
  - | Graphical models like directed acyclic graphs (DAGs) are often used to encode assumptions about the causal model.
- | A graph  $G(V, E)$  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  $G(V, E)$  is said to be **directed** and **acyclic** if:
  - | There are only directed edges ( $V_i \rightarrow V_j$ )
  - | There are no directed cycles – for any  $V_i \in V$  there is no sequence of directed edges in  $G$  such that  $V_i \rightarrow \dots \rightarrow V_i$



- | For notational convenience, we often refer to  $G(V, E)$  as  $G(V)$  or simply  $G$ .

# Statistical model of a DAG

- | DAG  $G(V)$  encodes a set of independence restrictions on the joint distribution  $p(V)$ .
- | The joint distribution  $p(V)$  corresponding to DAG  $G(V)$  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  $G(V)$  if:

$$p(V) = \prod_{V_i \in V} p(V_i \mid \text{pa}_G(V_i)).$$

|  $\text{pa}_G(V_i) = \{V_j \in V \mid V_j \text{ is a parent of } V_i \text{ in } G(V)\}$ .

## Statistical model of a DAG ctd.

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

$$p(V) = \prod_{V_i \in V} p(V_i \mid \text{pa}_G(V_i)).$$

|  $\text{pa}_G(V_i) = \{V_j \in V \mid V_j \rightarrow V_i\}$  denotes parents of  $V_i$  in  $G(V)$ .

---

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

$$V_i \perp \text{nd}_G(V_i) \setminus \text{pa}_G(V_i) \mid \text{pa}_G(V_i), \quad V_i \in V$$

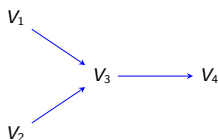
(variable  $V_i$  is independent of its non-descendants non-parents given its parents)

|  $\text{nd}_G(V_i) = \{V_j \in V \mid V_j \not\rightarrow V_i\}$  denotes non-descendants of  $V_i$  in  $G(V)$ ,

|  $\text{de}_G(V_i) = \{V_j \in V \mid V_i \rightarrow \dots \rightarrow V_j\}$  denotes descendants of  $V_i$  in  $G(V)$ .

## 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.,

$$p(V) = p(V_1) \times p(V_2) \times p(V_3 / V_1, V_2) \times p(V_4 / V_3) \quad .$$

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

$$p(V) \text{ s.t. } V_1 \perp V_2 \text{ and } V_4 \perp V_1, V_2 / V_3 \quad .$$

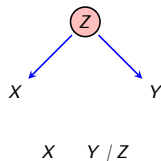
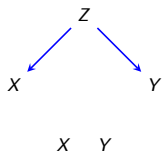
The above list implies a larger set of independence restrictions, e.g.,  $V_4 \perp V_1 / V_3$  or  $V_4 \perp V_1 / V_2, V_3$ . (graphoid axioms)

## Global Markov property: d-separation

- | Given a DAG  $G(V)$ , we need to be able to 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  $X \rightarrow Z \leftarrow Y$
  - | Chains  $X \rightarrow Z \rightarrow Y$  or  $X \leftarrow Z \leftarrow Y$
  - | Colliders  $X \rightarrow Z \rightarrow Y$

# Fork triplets

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



- | 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: warmer weather draw more people to the beach and also drive up ice cream sales.
  - |  $X$ : shark attacks,  $Z$ : warm whether, and  $Y$ : ice cream sales.

# Chain triplets

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



$X \quad Y$

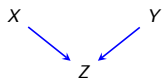


$X \quad Y | Z$

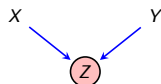
- | Intuition: if  $Z$  is a noisy version of  $X$  and  $Y$  is a noisy version of  $Z$ , then  $Y$  is a noisy version of  $X$ .
  - |  **$Z$  screens off the effect** of  $X$  on  $Y$ .
  - | Upon observing  $Z$ ,  $X$  holds no extra information about  $Y$ .
- | Example: blood sugar causes hunger, but only indirectly through increasing the stomach acidity.
  - |  $X$ : blood sugar,  $Z$ : stomach acidity, and  $Y$ : hunger.

# Collider triplets

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



$X \perp Y$

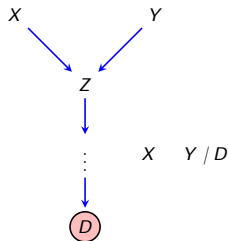
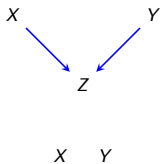


$X \perp Y | Z$

- | 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: 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
  - |  $X$ : battery,  $Z$ : car starts, and  $Y$ : fuel.
- | Conditioning can induce dependence, not just remove the dependence.

# Collider extensions

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



- | Example: extend the previous example where the car was taken to a mechanic.
  - |  $X$ : battery,  $Z$ : car starts,  $Y$ : fuel, and  $D$ : taken to mechanic



# Summary of the (in)dependence rules

$X \leftarrow Z \rightarrow Y$

$X \quad Y$

$X \leftarrow \textcircled{Z} \rightarrow Y$

$X \quad Y / Z$

$X \rightarrow Z \rightarrow Y$

$X \quad Y$

$X \rightarrow \textcircled{Z} \rightarrow Y$

$X \quad Y / Z$

$X \rightarrow Z \leftarrow Y$

$X \quad Y$

$X \rightarrow \textcircled{Z} \leftarrow Y$

$X \quad Y / Z$

$X \rightarrow Z \leftarrow Y$

$\vdots$

$\textcircled{D}$

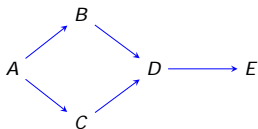
$X \quad Y / D$

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

## From blocked triplets to d-separation

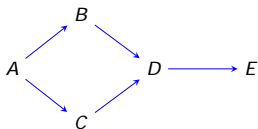
- | A **path** from  $X$  to  $Y$  is a sequence of consecutive edges connecting  $X$  and  $Y$  such that no vertex (and consequently no edge) appears more than once in the sequence.
- | A path from  $X$  to  $Y$  is blocked by  $Z$  if there is a **blocking triplet** on the path.
  - | There exists 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  $X \perp_d Y \mid Z$ .

## Examples: d-separation



| Is  $A \perp_d E \mid C$ ?

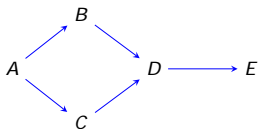
## Examples: d-separation



| Is  $A \perp_d E \mid C$ ?

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

## Examples: d-separation

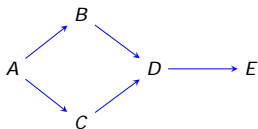


| Is  $A \perp_d E \mid C$ ?

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

| Is  $B \perp_d C \mid A$ ?

## Examples: d-separation



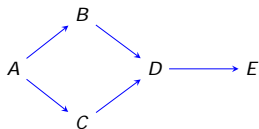
| Is  $A \perp_d E \mid C$ ?

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

| Is  $B \perp_d C \mid A$ ?

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

## Examples: d-separation



| Is  $A \perp_d E \mid C$ ?

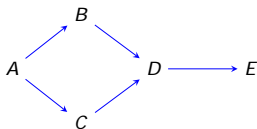
No! because  $A \rightarrow B \rightarrow D \rightarrow E$  is still open.

| Is  $B \perp_d C \mid A$ ?

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

| Is  $B \perp_d C \mid E, A$ ?

## Examples: d-separation



| Is  $A \perp_d E \mid C$ ?

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

| Is  $B \perp_d C \mid A$ ?

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

| Is  $B \perp_d C \mid E, A$ ?

No!  $B \rightarrow D \rightarrow C$  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  $G(V)$ , the following **Global Markov property** holds: for all disjoint subsets  $X, Y, Z$  of  $V$  we have,

$$X \perp_d Y / Z_{G(V)} = X \perp Y / Z_{p(V)}$$

where  $(\perp_d)_G$  denotes d-separation in  $G$  and  $(\perp)_p$  denotes independence in  $p$ .

- | We can apply a purely graphical criterion to a DAG  $G(V)$  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 we cannot read them by d-separation (and that happens in *unfaithful* distributions).

# 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  $G(V)$  if and only if it obeys the local Markov property according to  $G(V)$  if and only if it obeys the global Markov property according to  $G(V)$  (Verma and Pearl, 1990).

DAG factorization

Local Markov property

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  $V_i \in V$ :
$$V_i = f_{V_i}(\text{pa}_G(V_i), v_i)$$
  - |  $v_i$  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  $v_i$  in any way.
  - | This is an imperative assignment, not an equality! which means the model is not “reversible.”

# Causal model of a DAG

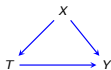
- | The causal model of a DAG can be formally defined in terms of **Nonparametric Structural Equations Model (NPSEM)**.
  - | It describes how “nature” assigns values to each variable in the model.
- | For every  $V_i \in V$ :
$$V_i = f_{V_i}(\text{pa}_G(V_i), v_i)$$
  - |  $v_i$  denotes the error term (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  $v_i$  in any way.
  - | This is an imperative assignment, not an equality! which means the model is not “reversible.”
- | Unmeasured factors are typically assumed to be independent – a reflection of the fact that all unmeasured common causes have been measured.
  - | This is known as the **NPSEM with independent errors (NPSEM-IE)**
  - | The explicit assumption is that  $\{v_i, V_i \in V\}$ , and thus
$$p(\cdot) = \prod_{V_i \in V} p(v_i)$$

# 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  $T = f_T(\text{pa}_G(T), T)$  with  $T = t$ .
    - | 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  $G(V)$ , delete all incoming edges into  $T$  and fix the  $T$  node to take value  $t$ .
    - | The vertices on the downstream of  $T$  turn into counterfactuals.
  - III. **Probabilistical** changes to the joint distribution
    - | In the corresponding joint distribution  $p(V)$ , drop the factor  $p(T | \text{pa}_G(T))$  from the factorization of  $p(V)$ , and evaluate all other factors at  $T = t$ .
    - | Upon the intervention, we end up with a *truncated* factorization.

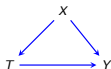
## Example: intervention in causal models

- | 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:



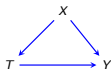
- Structural operation of intervening on  $T$  and setting it to  $t$ :

$X$	$f_X(X)$	$X$	$f_X(X)$
$T$	$f_T(X, T)$	$T$	$t$
$Y$	$f_Y(X, T, Y)$	$Y^{(t)}$	$f_Y(X, t, Y)$

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

## Example: intervention in causal models

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

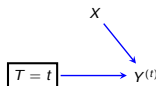
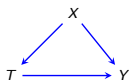


- Structural operation of intervening on  $T$  and setting it to  $t$ :

$X$	$f_X(X)$	$X$	$f_X(X)$
$T$	$f_T(X, T)$	$T$	$t$
$Y$	$f_Y(X, T, Y)$	$Y^{(t)}$	$f_Y(X, t, Y)$

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

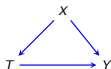
- Graphical operation of this intervention is illustrated as follows:





## 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  $t$ :

$X$	$f_X(X)$	$X$	$f_X(X)$
$T$	$f_T(X, T)$	$T$	$t$
$Y$	$f_Y(X, T, Y)$	$Y^{(t)}$	$f_Y(X, t, Y)$

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

- Graphical operation of this intervention is illustrated as follows:

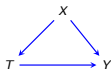


- Probabilistic operation of this intervention is a truncated factorization:

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

## 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  $t$ :

$X$	$f_X(x)$	$X$	$f_X(x)$
$T$	$f_T(X, T)$	$T$	$t$
$Y$	$f_Y(X, T, Y)$	$Y^{(t)}$	$f_Y(X, t, Y)$

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

- Graphical operation of this intervention is illustrated as follows:



- Probabilistic operation of this intervention is a truncated factorization:

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

Do-calculus notation of Pearl:  $p(X, Y^{(t)}) = p(X, Y | \text{do}(T=t))$ .

## 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

# 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

# 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:

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

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

# 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:  
 $X, R, X \quad L^{(1)}, R, L$   
 $L^{(1)} = (L_1^{(1)}, \dots, L_K^{(1)})^T$  and  $L = (L_1, \dots, L_K)^T$ .
- |  $Z$ : 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:

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

full law  $p(L^{(1)}, Z, R)$

# 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:

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

full law  $p(L^{(1)}, Z, R)$

- | 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.

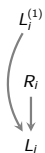


# A missing data workflow

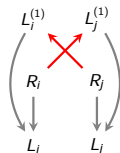
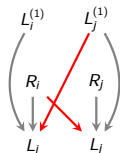
1. Define the **estimand** (often done in the absence of missing data).
  - | A function of target law  $p(Z, L^{(1)})$  or full law  $p(Z, L^{(1)}, R)$ .
2. Assume a **model** that links the counterfactual, factual, and missingness indicator variables.
  - | Use **Directed Acyclic Graphs (DAGs)** to 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)
  - Observed and counterfactual variables appear on the same graph
  - There are certain edge restrictions: (marked in red)



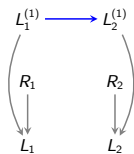
deterministic edges



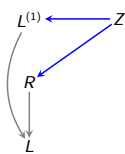
"no interference"

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

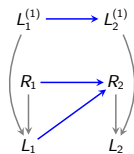
# Examples: missing data DAGs



MCAR

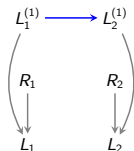


MAR

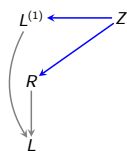


MAR

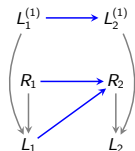
# Examples: missing data DAGs



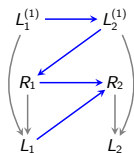
MCAR



MAR

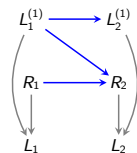


MAR



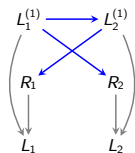
MNAR: permutation

(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.

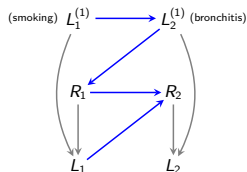
## Examples: missing data DAGs ctd.

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.

|  $L_2^{(1)} \perp R_1$   
A doctor inquires about the patient's smoking status on a suspected diagnosis of bronchitis before administering the test.

|  $R_1 \perp R_2 \perp L_1$   
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)

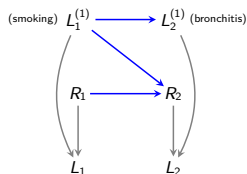
## Examples: missing data DAGs ctd.

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).

- |  $R_1$      $R_2$      $L_1^{(1)}$   
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)

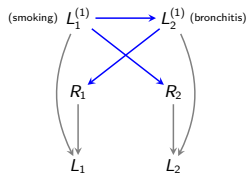
## Examples: missing data DAGs ctd.

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.

|  $R_1 \perp L_2^{(1)}$   
Inquiry into smoking status depends on a suspected diagnosis of bronchitis.

|  $R_2 \perp L_1^{(1)}$   
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  $G(V)$ .
- | The statistical model of m-DAG  $G(V)$  is a set of distributions that factorize as:

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

# Missing data DAG models

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

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

- 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.

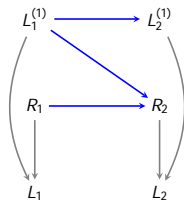
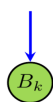
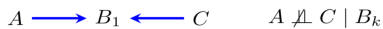
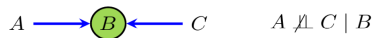
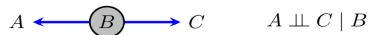
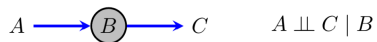
$$V_i \perp\!\!\!\perp \text{nd}_G(V_i) \setminus \text{pa}_G(V_i) \mid \text{pa}_G(V_i), \quad V_i \in V.$$

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

$$\text{Given } X, Y, Z \subseteq V: (X \perp\!\!\!\perp Y \mid Z)_{G(V)} = (X \perp\!\!\!\perp Y \mid Z)_{p(V)}.$$

- All three properties are equivalent.

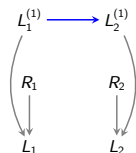
# d-separation refresher



$R_1$      $L_1^{(1)}$

$R_1$      $L_1^{(1)} / R_2$

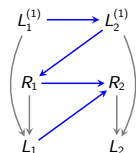
# Examples: m-DAG models



MCAR

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

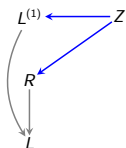
$$R_2 \quad R_1, L_1^{(1)}, L_2^{(1)}$$



Permutation (Robins, 1997)

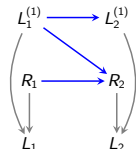
$$R_1 \quad L_1^{(1)} / L_2^{(1)}$$

$$R_2 \quad L_1^{(1)}, L_2^{(1)} / R_1, L_1$$



MAR

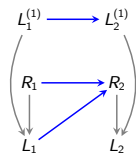
$$R \quad L^{(1)} / Z$$



Block-conditional (Zhou et al., 2010)

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

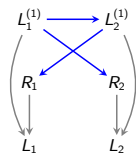
$$R_2 \quad L_1^{(1)} / R_1, L_1^{(1)}$$



MAR

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

$$R_2 \quad L_1^{(1)}, L_2^{(1)} / R_1, L_1$$



Block-parallel (Mohan et al., 2013)

$$R_1 \quad R_2, L_1^{(1)} / L_2^{(1)}$$

$$R_2 \quad R_1, L_2^{(1)} / L_1^{(1)}$$

# Graphical representations of MCAR, MAR, MNAR mechanisms

Missingness mechanism,  $p(R / \text{pa}_G(R)) = \prod_{R_k \in R} p(R_k / \text{pa}_G(R_k))$ , is

- | **MCAR:** if  $p(R_k / \text{pa}_G(R_k))$ ,  $R_k \in R$ , is not a function of variables in  $\{L^{(1)}, L, Z\}$ .
  - | Graphically speaking, there are no edges that point to variables in  $R$ .
- | **MAR:** if  $p(R_k / \text{pa}_G(R_k))$ ,  $R_k \in R$ , is not a function of variables in  $L^{(1)}$ , but could be a function of variables in  $\{Z, L\}$ .
  - | Graphically speaking, there are no edges from variables in  $L^{(1)}$  to variables in  $R$ .
- | **MNAR** otherwise.

## **Part 2. Nonparametric Identification**

# Identification in missing data models

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

# Identification in missing data models

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

$$\begin{aligned} p(Z, L^{(1)}) &= p(Z, L^{(1)} / R = 1) && R \quad Z, L^{(1)} \\ &= p(Z, L / R = 1). && \text{consistency} \end{aligned}$$

- | Under MAR missingness, target law is identified:

$$\begin{aligned} p(Z, L^{(1)}) &= p(Z) \times p(L^{(1)} / Z) \\ &= p(Z) \times p(L^{(1)} / Z, R = 1) && R \quad L^{(1)} / Z \\ &= p(Z) \times p(L / Z, R = 1). && \text{consistency} \end{aligned}$$

- | MNAR model:

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

MNAR models    causal models with unmeasured confounding

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

# Nonparametric identification theory in causal inference

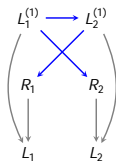
- | Identification questions in causal inference: given an arbitrary DAG with hidden/unmeasured variables, is  $p(Y^{(t)})$  identified?
- | **Sound** and **complete** algorithms exist for causal effect identification.
  - | Soundness: functionals of identified effects are correct.
  - | Completeness: no-identifiability of the causal effect is *provable*.  
(Shpitser and Pearl, 2006; Huang and Valorta, 2006; Richardson et al., 2017; Bhattacharya et al., 2022)

# Nonparametric identification theory in causal inference

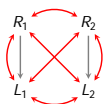
- | Identification questions in causal inference: given an arbitrary DAG with hidden/unmeasured variables, is  $p(Y^{(t)})$  identified?
- | **Sound** and **complete** algorithms exist for causal effect identification.
  - | Soundness: functionals of identified effects are correct.
  - | Completeness: no-identifiability of the causal effect is *provable*.  
(Shpitser and Pearl, 2006; Huang and Valertorta, 2006; Richardson et al., 2017; Bhattacharya et al., 2022)
- | Similarly, given that assumptions/restrictions in a missing data model are encoded via a m-DAG, we would like to know whether the underlying full/target law is identified or not.
- | **Causal identification theory is *incomplete* for missing data identification!**
  - | There are indeed identified MNAR models for which causal identification theory fails.

# Incompleteness of causal identification theory for m-DAGs

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



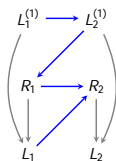
Block-parallel



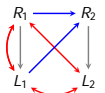
Observed margin

One-line ID (Richardson et al., 2017)

- $Y = \{L_1, L_2\}$
- $G_Y = L_1 \quad L_2$
- District:  $\{L_1, L_2\}$
- Need to fix  $R_1, R_2$  and fail.



Permutation



Observed margin

One-line ID (Richardson et al., 2017)

- $Y = \{L_1, L_2\}$
- $G_Y = L_1 \quad L_2$
- 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  $p(R = 1 / L^{(1)}, Z)$  is identified. Using Bayes rule:

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

# Nonparametric identification in m-DAGs

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

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

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

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

# Nonparametric identification in m-DAGs

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

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

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

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

- | Game plan: focus on identification of the missingness mechanism  $p(R / L^{(1)}, Z)$  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  $p(R / L^{(1)}, Z)$ :



# 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  $p(R / L^{(1)}, Z)$ :

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

$$p(R / \text{pa}_G(R)) = \prod_{R_k \in R} p(R_k / \text{pa}_G(R_k))$$

Identify each **propensity score**  $p(R_k / \text{pa}_G(R_k))$ , 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  $p(R / L^{(1)}, Z)$ :

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

$$p(R / \text{pa}_G(R)) = \prod_{R_k \in R} p(R_k / \text{pa}_G(R_k))$$

Identify each **propensity score**  $p(R_k / \text{pa}_G(R_k))$ , for all  $R_k \in R$ .

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

$$\prod_{k=1}^K p(R_k / R_{-k} = 1, \text{pa}_G(R)) \times \prod_{k=2}^K \text{OR}(R_k, R_{-k} / R_{-k} = 1, \text{pa}_G(R)),$$

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

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  $G$ .
  - | Full law: argue for identification of  $p(R = r / \text{pa}_G(R))$ ,  $r$ .
  - | Target law: argue for identification of  $p(R = 1 / \text{pa}_G(R))$ .

# Identification via m-DAG factorization

- | Whether and how full/target law is identified in a given m-DAG  $G$ .
  - | Full law: argue for identification of  $p(R = r / \text{pa}_G(R))$ ,  $r$ .
  - | Target law: argue for identification of  $p(R = 1 / \text{pa}_G(R))$ .
- | Target law is identified if each propensity score  $p(R_k / \text{pa}_G(R_k))$  is identified only when  $R_i = 1$  for  $R_i \in \text{pa}_G(R_k)$ ,

$$p(R = 1 / \text{pa}_G(R)) = \prod_{R_k \in R} p(R_k / \text{pa}_G(R_k))_{R=1}.$$

# Identification via m-DAG factorization

- | Whether and how full/target law is identified in a given m-DAG  $G$ .
  - | Full law: argue for identification of  $p(R = r / \text{pa}_G(R))$ ,  $r$ .
  - | Target law: argue for identification of  $p(R = 1 / \text{pa}_G(R))$ .
- | Target law is identified if each propensity score  $p(R_k / \text{pa}_G(R_k))$  is identified only when  $R_i = 1$  for  $R_i \in \text{pa}_G(R_k)$ ,

$$p(R = 1 / \text{pa}_G(R)) = \prod_{R_k \in \text{pa}_G(R)} p(R_k / \text{pa}_G(R_k))_{R=1}.$$

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

$$p(R = r / \text{pa}_G(R)) = \prod_{R_k \in \text{pa}_G(R)} p(R_k / \text{pa}_G(R_k))_{R=r}.$$

# Identification via m-DAG factorization

- | Whether and how full/target law is identified in a given m-DAG  $G$ .
  - | Full law: argue for identification of  $p(R = r / \text{pa}_G(R))$ ,  $r$ .
  - | Target law: argue for identification of  $p(R = 1 / \text{pa}_G(R))$ .
- | Target law is identified if each propensity score  $p(R_k / \text{pa}_G(R_k))$  is identified only when  $R_i = 1$  for  $R_i \in \text{pa}_G(R_k)$ ,

$$p(R = 1 / \text{pa}_G(R)) = \prod_{R_k \in R} p(R_k / \text{pa}_G(R_k))_{R=1}.$$

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

$$p(R = r / \text{pa}_G(R)) = \prod_{R_k \in R} p(R_k / \text{pa}_G(R_k))_{R=r}$$

- | 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  $R_k$ ,  $p(R_k / \text{pa}_G(R_k))$ , we need to **select on** the following missingness indicators:

$$R_k^s = R_i \quad R / L_j^{(1)} \quad \text{pa}_G(R_k) \quad (\text{selection set for } R_k)$$



# 1. Associational irrelevancy

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

$$R_k^s = R_i \mid R / L_i^{(1)} \mid \text{pa}_G(R_k) \quad (\text{selection set for } R_k)$$

- | 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  $R_k \perp \text{nd}_G(R_k) \setminus \text{pa}_G(R_k) \mid \text{pa}_G(R_k)$ .
  - | 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  $R_k$ ,  $p(R_k / \text{pa}_G(R_k))$ , we need to **select on** the following missingness indicators:

$$R_k^S = R_i \mid R / L_i^{(1)} \quad \text{pa}_G(R_k) \quad (\text{selection set for } R_k)$$

- 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  $R_k \perp \text{nd}_G(R_k) \setminus \text{pa}_G(R_k) \mid \text{pa}_G(R_k)$ .

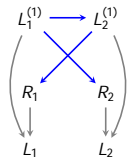
- So we can include  $R_i = 1$  in the conditioning set and replace  $L_i^{(1)}$  with  $\{L_i, R_i = 1\}$ .

- Formally, for any  $R_i \in R_k^S \setminus \text{nd}_G(R_k)$ , we can write:

$$p(R_k / \text{pa}_G(R_k)) \underset{R=1}{=} p(R_k / \text{pa}_G(R_k), R_i = 1) \underset{R=1}{=} .$$

includes  $L_i^{(1)}$

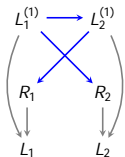
## Example 1/2: associational irrelevancy (block-parallel)



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

$$p(R / pa_G(R)) = p(R_1 / pa_G(R_1)) \times p(R_2 / pa_G(R_2))$$

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



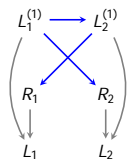
| Is the full/target law identified in the **block-parallel** model?

$$p(R / \text{pa}_G(R)) = p(R_1 / \text{pa}_G(R_1)) \times p(R_2 / \text{pa}_G(R_2))$$

Identification of  $p(R_1 / \text{pa}_G(R_1))$

$$\begin{aligned} p(R_1 / \text{pa}_G(R_1)) &= p(R_1 / L_2^{(1)}) \\ &= p(R_1 / R_2 = 1, L_2^{(1)}) && R_1 \quad R_2 / L_2^{(1)} \\ &= p(R_1 / R_2 = 1, L_2) && \text{consistency} \end{aligned}$$

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



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

$$p(R / pa_G(R)) = p(R_1 / pa_G(R_1)) \times p(R_2 / pa_G(R_2))$$

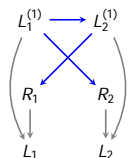
Identification of  $p(R_1 / pa_G(R_1))$

$$\begin{aligned} p(R_1 / pa_G(R_1)) &= p(R_1 / L_2^{(1)}) \\ &= p(R_1 / R_2 = 1, L_2^{(1)}) && R_1 \quad R_2 / L_2^{(1)} \\ &= p(R_1 / R_2 = 1, L_2) && \text{consistency} \end{aligned}$$

Identification of  $p(R_2 / pa_G(R_2))$

$$\begin{aligned} p(R_2 / pa_G(R_2)) &= p(R_2 / L_1^{(1)}) \\ &= p(R_2 / R_1 = 1, L_1^{(1)}) && R_2 \quad R_1 / L_1^{(1)} \\ &= p(R_2 / R_1 = 1, L_1) && \text{consistency} \end{aligned}$$

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



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

$$p(R / pa_G(R)) = p(R_1 / pa_G(R_1)) \times p(R_2 / pa_G(R_2))$$

**Identification of  $p(R_1 / pa_G(R_1))$**

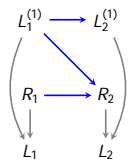
$$\begin{aligned} p(R_1 / pa_G(R_1)) &= p(R_1 / L_2^{(1)}) \\ &= p(R_1 / R_2 = 1, L_2^{(1)}) && R_1 \quad R_2 / L_2^{(1)} \\ &= p(R_1 / R_2 = 1, L_2) && \text{consistency} \end{aligned}$$

**Identification of  $p(R_2 / pa_G(R_2))$**

$$\begin{aligned} p(R_2 / pa_G(R_2)) &= p(R_2 / L_1^{(1)}) \\ &= p(R_2 / R_1 = 1, L_1^{(1)}) && R_2 \quad R_1 / L_1^{(1)} \\ &= p(R_2 / R_1 = 1, L_1) && \text{consistency} \end{aligned}$$

| So  $p(R / pa_G(R))$  is ID, which means the full and target laws are both ID.

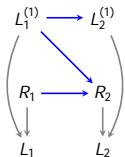
## Example 2/2: associational irrelevancy (block-conditional MAR)



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

$$p(R / \text{pa}_G(R)) = p(R_1 / \text{pa}_G(R_1)) \times p(R_2 / \text{pa}_G(R_2))$$

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



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

$$p(R / \text{pa}_G(R)) = p(R_1 / \text{pa}_G(R_1)) \times p(R_2 / \text{pa}_G(R_2))$$

•  $p(R_1 / \text{pa}_G(R_1)) = p(R_1)$  is identified.

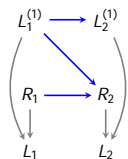
**Identification of  $p(R_2 / \text{pa}_G(R_2))$**

$$p(R_2 / \text{pa}_G(R_2)) = p(R_2 / R_1, L_1^{(1)}) \\ = ???$$

$$p(R_2 / \text{pa}_G(R_2)) \Big|_{R=1} = p(R_2 = 1 / R_1 = 1, L_1^{(1)}) \\ = p(R_2 = 1 / R_1 = 1, L_1) \quad \text{consistency}$$



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



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

$$p(R / \text{pa}_G(R)) = p(R_1 / \text{pa}_G(R_1)) \times p(R_2 / \text{pa}_G(R_2))$$

•  $p(R_1 / \text{pa}_G(R_1)) = p(R_1)$  is identified.

**Identification of  $p(R_2 / \text{pa}_G(R_2))$**

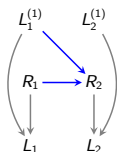
$$p(R_2 / \text{pa}_G(R_2)) = p(R_2 / R_1, L_1^{(1)}) \\ = ???$$

$$p(R_2 / \text{pa}_G(R_2)) \underset{R=1}{=} p(R_2 = 1 / R_1 = 1, L_1^{(1)}) \\ = p(R_2 = 1 / R_1 = 1, L_1) \quad \text{consistency}$$

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



- 
- Assume binary data. So the full law  $p(R, L^{(1)})$  would have 7 parameters.
  - 5 identified parameters:
    - 1 :  $p(L_1^{(1)} = 1) = p(L_1^{(1)} = 1 / R_1 = 1) = p(L_1 = 1 / R_1 = 1)$
    - 2 :  $p(L_2^{(1)} = 1) = p(L_2^{(1)} = 1 / R_2 = 1) = p(L_2 = 1 / R_2 = 1)$
    - 3 :  $p(R_1 = 1)$
    - 4,5 :  $p(R_2 = 1 / R_1 = 1, L_1^{(1)} = l_1) = p(R_2 = 1 / R_1 = 1, L_1 = l_1)$ .
  - 2 unidentified parameters:
    - 6,7 :  $p(R_2 = 1 / R_1 = 0, L_1^{(1)} = l_1), l_1 \in \{0, 1\}$

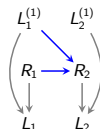
# Proofing non-identifiability claims

$$\begin{array}{c|c} R_1 & p(R_1) \\ \hline 0 & a \\ 1 & 1-a \end{array}$$

$$\begin{array}{c|c} L_1^{(1)} & p(L_1^{(1)}) \\ \hline 0 & b \\ 1 & 1-b \end{array}$$

$$\begin{array}{c|c} L_2^{(1)} & p(L_2^{(1)}) \\ \hline 0 & c \\ 1 & c \end{array}$$

$R_2$	$R_1$	$L_1^{(1)}$	$p(R_2 / R_1, L_1^{(1)})$
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	$1-g$



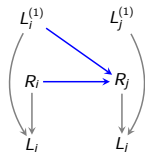
$R_1$	$R_2$	$L_1^{(1)}$	$L_2^{(1)}$	p(FULL LAW)	$L_1$	$L_2$	p(OBSERVED LAW)
0	0	0	0	$abcd$	?	?	$a \quad db + f(1-b)$
		1	0	$af(1-b)c$			
		0	1	$adb(1-c)$			
		1	1	$af(1-b)(1-c)$			
1	0	0	0	$(1-a)ebc$	0	?	$(1-a)eb$
		1	0	$(1-a)g(1-b)c$	1		$(1-a)g(1-b)$
		0	1	$(1-a)eb(1-c)$			
		1	1	$(1-a)g(1-b)(1-c)$			
0	1	0	0	$a(1-d)bc$		?	0
		1	0	$a(1-f)(1-b)c$	1		$a(1-c) \quad 1 - db + f(1-b)$
		0	1	$a(1-d)b(1-c)$			
		1	1	$a(1-f)(1-b)(1-c)$			
1	1	0	0	$(1-a)(1-e)bc$		0	0
		1	0	$(1-a)(1-g)(1-b)c$	1	0	$(1-a)(1-g)(1-b)c$
		0	1	$(1-a)(1-e)b(1-c)$	0	0	$(1-a)(1-e)b(1-c)$
		1	1	$(1-a)(1-g)(1-b)(1-c)$	1	1	$(1-a)(1-g)(1-b)(1-c)$

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

# Collider: non-identified structure

## Definition (Collider)

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

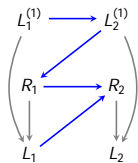


## Lemma (Collider non-identification)

If  $R_i, R_j \in R$  such that  $R_i \perp R_j \mid L_i^{(1)}$  then  $p(R_j \mid \text{pa}_G(R_j) \setminus R_i, R_i = 0)$  is not identified (Bhattacharya et al., 2019).

The above result means that whenever we spot a collider on the m-DAG, we can immediately conclude the full law is not identified.

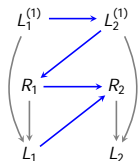
## Associational irrelevancy: limitations



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

$$p(R / L^{(1)}) = p(R_1 / \text{pa}_G(R_1)) \times p(R_2 / \text{pa}_G(R_2))$$

# Associational irrelevancy: limitations



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

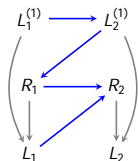
$$p(R / L^{(1)}) = p(R_1 / \text{pa}_G(R_1)) \times p(R_2 / \text{pa}_G(R_2))$$

Identification of  $p(R_i / \text{pa}_G(R_i))$

$$p(R_2 / \text{pa}_G(R_2)) = p(R_2 / R_1, L_1) \times$$

$$p(R_1 / \text{pa}_G(R_1)) = p(R_1 / L_2^{(1)}) \quad R_1 \quad R_2 / L_2^{(1)}$$

# Associational irrelevancy: limitations



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

$$p(R / L^{(1)}) = p(R_1 / \text{pa}_G(R_1)) \times p(R_2 / \text{pa}_G(R_2))$$

## Identification of $p(R_i / \text{pa}_G(R_i))$

$$p(R_2 / \text{pa}_G(R_2)) = p(R_2 / R_1, L_1) \times$$

$$p(R_1 / \text{pa}_G(R_1)) = p(R_1 / L_2^{(1)}) \quad R_1 \quad R_2 / L_2^{(1)}$$

- | 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  $\text{pa}_G(R_k)$  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  $R \in R \setminus R_k$ , we have

$$p(R_k = 1 \mid \text{pa}_G(R_k)) = p(R_k = 1 \mid \text{pa}_G(R_k), \text{do}(R = 1)) .$$



## 2. Causal irrelevancy

### Invariance property:

- | Given the propensity score for  $R_k \in R$ , the conditioning set  $\text{pa}_G(R_k)$  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

$$p(R_k | \text{pa}_G(R_k)) = p(R_k | \text{pa}_G(R_k), \text{do}(R = 1)) .$$

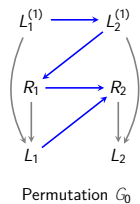
- | Due to *invariance* property of the propensity scores, we can sometimes succeed in identifying a propensity score by exploring interventional distributions where a subset of variables are intervened on.

# Intervention operation on missingness indicators

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

- | **Graphical** changes to the missing data DAG  $G(V)$ 
  - | In the corresponding m-DAG  $G$ , delete all the incoming edges into  $R_k$  and fix  $R_k$  to take value 1, and
  - | Treat the counterfactual variable  $L_k^{(1)}$  as  $L_k$  (by consistency).
- | **Probabilistical** changes to the joint distribution  $p(V)$ 
  - | In the corresponding joint law  $p(V)$ , drop the term  $p(R_k / \text{pa}_G(R_k))$  from the factorization of  $p(V)$ , and evaluate all terms at  $R_k = 1$ .

## Example: causal irrelevancy

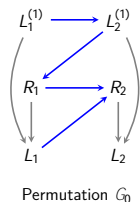


$$G_0 : R_1 \quad R_2 / L_2^{(1)}$$

Invariance property:

$$p(R_1 / L_2^{(1)}) = p(R_1 / L_2^{(1)}, \text{do}(R_2 = 1))$$

# Example: causal irrelevancy

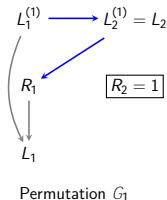


$$G_0 : R_1 \quad R_2 / L_2^{(1)}$$

Invariance property:

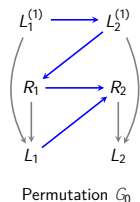
$$p(R_1 / L_2^{(1)}) = p(R_1 / L_2^{(1)}, \text{do}(R_2 = 1))$$

Graphical and probabilistical changes after intervening on  $R_2$ :



$$p(R_1, L_1^{(1)}, L_2^{(1)}, L_1 / \text{do}(R_2 = 1)) = \frac{p(R_1, L_1^{(1)}, L_2^{(1)}, L_1, R_2)}{p(R_2 / R_1, L_1)} \quad R_2=1$$

## Example: causal irrelevancy

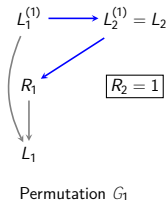


$$G_0 : R_1 \quad R_2 / L_2^{(1)}$$

Invariance property:

$$p(R_1 / L_2^{(1)}) = p(R_1 / L_2^{(1)}, \text{do}(R_2 = 1))$$

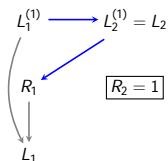
Graphical and probabilistical changes after intervening on  $R_2$ :



$$p(R_1, L_1^{(1)}, L_2^{(1)}, L_1 / \text{do}(R_2 = 1)) = \frac{p(R_1, L_1^{(1)}, L_2^{(1)}, L_1, R_2)}{p(R_2 / R_1, L_1)} \quad R_2=1$$

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

## Example ctd. causal irrelevancy



Permutation  $G_1$

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

$$p(R_1, L_2^{(1)} / \text{do}(R_2 = 1)) = \int_{l_1} \frac{p(R_1, l_1, L_2, R_2 = 1)}{p(R_2 = 1 / R_1, l_1)}.$$

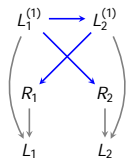
- The propensity score of  $R_1$  is identified from the above intervention dist.

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

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  $R_2$ .

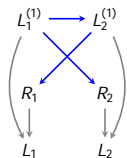


Block-parallel

$$\begin{aligned} p(L^{(1)}) &= \frac{p(L^{(1)}, R)}{p(R_1 / \text{pa}_G(R_1)) \times p(R_2 / \text{pa}_G(R_2))} \quad R=1 \\ &= \frac{p(L, R=1)}{p(R_1=1 / R_2=1, L_2) \times p(R_2=1 / R_1=1, L_1)}. \end{aligned}$$

# Order of interventions

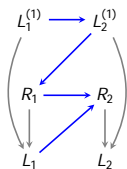
- Target law is ID via **parallel/simultaneous interventions** on  $R_1$  and  $R_2$ .



Block-parallel

$$\begin{aligned}
 p(L^{(1)}) &= \frac{p(L^{(1)}, R)}{p(R_1 / \text{pa}_G(R_1)) \times p(R_2 / \text{pa}_G(R_2))} \quad R=1 \\
 &= \frac{p(L, R = 1)}{p(R_1 = 1 / R_2 = 1, L_2) \times p(R_2 = 1 / R_1 = 1, L_1)}.
 \end{aligned}$$

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



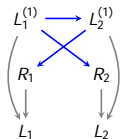
Permutation

$$\begin{aligned}
 p(L^{(1)}) &= \frac{p(L^{(1)}, R)}{p(R_1 / \text{pa}_G(R_1)) \times p(R_2 / \text{pa}_G(R_2))} \quad R=1 \\
 &= \frac{p(L, R = 1)}{p(R_1 = 1 / L_2^{(1)}, \text{do}(R_2 = 1)) \times p(R_2 = 1 / R_1 = 1, L_1)}.
 \end{aligned}$$



# Main identification challenge: selection bias

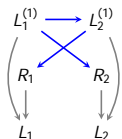
Can we apply the causal-irrelevancy idea to the block-parallel model?



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

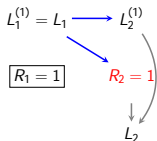
# Main identification challenge: selection bias

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



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

An intervention on  $R_1$  implies:



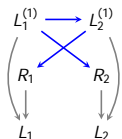
$$p(R_2, L_1^{(1)}, L_2^{(1)}, L_2 / \text{do}(R_1 = 1)) = \frac{p(L_1, R_1 = 1, L_2^{(1)}, L_2, R_2)}{p(R_1 = 1 / L_2^{(1)})}$$

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

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

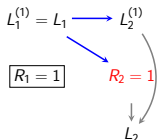
# Main identification challenge: selection bias

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



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

An intervention on  $R_1$  implies:



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

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

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

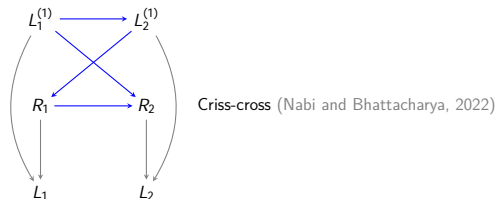
Intervening on  $R_1$  induces a selection on  $R_2$ .

## 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,



- | Full law  $p(L^{(1)}, R)$  is not identified because of the collider at  $R_2$ .
- | Target law  $p(L^{(1)})$  is also **provably** not identified (Nabi and Bhattacharya, 2022).

# 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  $G$  (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  $a_0, b_0$  :

$$p(A, B | C) = \frac{1}{Z(C)} \times p(A | B = b_0, C) \times p(B | A = a_0, C) \times \text{OR}(A, B | C),$$

where

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

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



# Odds ratio parameterization (Chen, 2007)

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

$$p(A, B / C) = \frac{1}{Z(C)} \times p(A / B = b_0, C) \times p(B / A = a_0, C) \times \text{OR}(A, B / C),$$

where

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

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

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

# Identification via odds ratio parameterization

- | m-DAG factorization view for identification:

$$p(R_i, R_j / \text{pa}_G(R_i, R_j)) = p(R_i / \text{pa}_G(R_i)) \times p(R_j / \text{pa}_G(R_j))$$

Identify:

- |  $p(R_i / \text{pa}_G(R_i))$ ,
- |  $p(R_j / \text{pa}_G(R_j))$ .

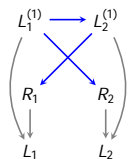
- | Odds ratio parameterization view for identification:

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

Identify:

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

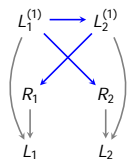
## Example: block-parallel model



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

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

## Example: block-parallel model

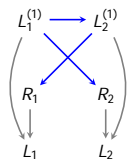


| Is the full/target law identified in the **block-parallel** model?

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

Note that  $R_1 \quad L_1^{(1)} / R_2, L_2^{(1)}$  and  $R_2 \quad L_2^{(1)} / R_1, L_1^{(1)}$ ,

## Example: block-parallel model



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

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

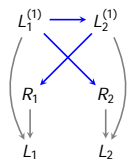
Note that  $R_1 \quad L_1^{(1)} / R_2, L_2^{(1)}$  and  $R_2 \quad L_2^{(1)} / R_1, L_1^{(1)}$ ,

$$(1) : p(R_1 / R_2 = 1, L^{(1)}) = p(R_1 / R_2 = 1, L_2^{(1)}) = p(R_1 / R_2 = 1, L_2)$$

$$(2) : p(R_2 / R_1 = 1, L^{(1)}) = p(R_2 / R_1 = 1, L_1^{(1)}) = p(R_2 / R_1 = 1, L_1)$$

$$(3) : \text{OR}(R_1 = r_1, R_2 = r_2 / L^{(1)}) = 1.$$

## Example: block-parallel model



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

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

Note that  $R_1 \quad L_1^{(1)} / R_2, L_2^{(1)}$  and  $R_2 \quad L_2^{(1)} / R_1, L_1^{(1)}$ ,

$$(1) : p(R_1 / R_2 = 1, L^{(1)}) = p(R_1 / R_2 = 1, L_2^{(1)}) = p(R_1 / R_2 = 1, L_2)$$

$$(2) : p(R_2 / R_1 = 1, L^{(1)}) = p(R_2 / R_1 = 1, L_1^{(1)}) = p(R_2 / R_1 = 1, L_1)$$

$$(3) : \text{OR}(R_1 = r_1, R_2 = r_2 / L^{(1)}) = 1.$$

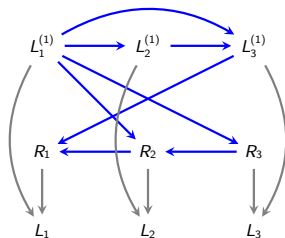
Yes, both full and target laws are identified.

# 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!

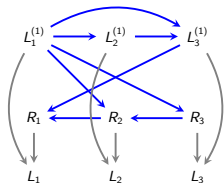


Is the target/full law identified?

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



# Example ctd. identification of the propensity score of $R_1$

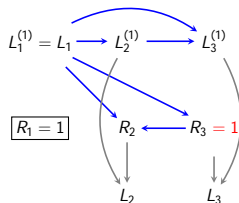
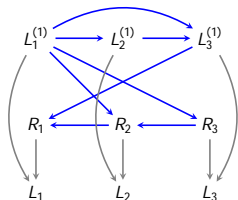


$$\begin{aligned}
 p(R / L^{(1)}) &= \prod_{R_i \in R} p(R_i / \text{pa}_G(R_i)) \\
 &= p(R_1 / R_2, L_3^{(1)}) \times p(R_2 / R_3, L_1^{(1)}) \times p(R_3 / L_1^{(1)}).
 \end{aligned}$$

## Nonparametric identification of $p(R_1 / \text{pa}_G(R_1))$

$$\begin{aligned}
 p(R_1 / \text{pa}_G(R_1)) &= p(R_1 / R_2, L_3^{(1)}) \\
 &= p(R_1 / R_3 = 1, R_2, L_3^{(1)}) \quad R_1 \quad R_3 / R_2, L_3^{(1)} \\
 &= p(R_1 / R_3 = 1, R_2, L_3) \quad \text{consistency.}
 \end{aligned}$$

## Example ctd. identification of the propensity score of $R_2$

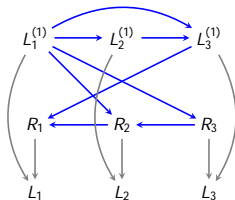


$$p = p(L^{(1)}, R_1, R_3 / \text{do}(R_1 = 1)) = \frac{p(L^{(1)}, R)}{p(R_1 = 1 / R_3 = 1, R_2, L_3^{(1)})}$$

Nonparametric identification of  $p(R_2 / \text{pa}_G(R_i))_{R=1}$

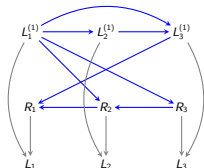
$$\begin{aligned} p(R_2 = 1 / \text{pa}_G(R_2))_{R=1} &= p(R_2 = 1 / R_3 = 1, L_1^{(1)}) \\ &= p(R_2 = 1 / R_3 = 1, L_1^{(1)}, \text{do}(R_1 = 1)) && \text{causal irrelevance} \\ &= p(R_2 = 1 / R_3 = 1, L_1^{(1)}) \\ &= p(R_2 = 1 / R_1 = 1, R_3 = 1, L_1^{(1)}) && R_2 \perp_G R_1 / R_3, L_1^{(1)} \\ &= p(R_2 = 1 / R_1 = 1, R_3 = 1, L_1) && \text{consistency} \end{aligned}$$

## Example ctd. identification of the propensity score of $R_3$



- Unfortunately, similar tricks do not help with identification of  $p(R_3 / \text{pa}_G(R_3))$  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 a different parameterization of the missingness mechanism (via odds ratio parameterization).

## Example ctd. odds ratio parameterization



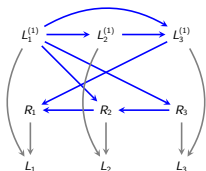
Univariate conditionals:

- |  $p(R_1 / R_2 = 1, R_3 = 1, L^{(1)}) = p(R_1 / R_2 = 1, R_3 = 1, L_3)$ 
  - |  $R_1 \quad L_1^{(1)}, L_2^{(2)} / R_2, R_3, L_3^{(1)}$
- |  $p(R_2 / R_1 = 1, R_3 = 1, L^{(1)}) = p(R_2 / R_1 = 1, R_3 = 1, L_1, L_3)$ 
  - |  $R_2 \quad L_2^{(1)} / R_1, R_3, L_1^{(2)}, L_3^{(1)}$
- |  $p(R_3 / R_1 = 1, R_2 = 1, L^{(1)}) = p(R_3 / R_2 = 1, R_1 = 1, L_1)$ 
  - |  $R_3 \quad L_2^{(1)}, L_3^{(1)} / R_1, R_2, L_1^{(2)}$

Pairwise odds ratios:

- |  $OR(R_1, R_2 / R_3 = 1, L^{(1)})$
- |  $OR(R_2, R_3 / R_1 = 1, L^{(1)})$
- |  $OR(R_1, R_3 / R_2 = 1, L^{(1)})$

## Example ctd. identification of the pairwise odds ratio terms



OR( $R_2 = 0, R_3 = 0 / R_1 = 1, L^{(1)}$ )

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

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

# Odds ratio parameterization of $p(R / L^{(1)})$

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

$$p(R / L^{(1)}) = \frac{1}{Z(L^{(1)})} \times \prod_{k=1}^K p(R_k / R_{-k} = 1, L^{(1)}) \times \prod_{k=2}^K \text{OR}(R_k, R_{> k} / R_{< k} = 1, L^{(1)}),$$

$$\text{OR}(R_k, R_{> k} / R_{< k} = 1, L^{(1)}) = \frac{p(R_k / R_{< k} = 1, R_{> k}, L^{(1)})}{p(R_k = 1 / R_{< k} = 1, R_{> k}, X^{(1)})} \times \frac{p(R_k = 1 / R_{< k} = 1, X^{(1)})}{p(R_k / R_{< k} = 1, L^{(1)})}.$$

Need to identify:

- | Univariate conditional distributions:  $p(R_k / R_{-k} = 1, L^{(1)})$
- | Odds ratio terms:  $\text{OR}(R_k, R_{> k} / R_{< k} = 1, L^{(1)})$
  
- | When can we succeed?
- | When do we fail?

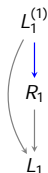
# Full law identification theory in m-DAGs

## Theorem (Graphical characterization of identified full laws)

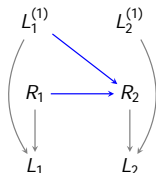
Full law  $p(R, L^{(1)}, Z)$  that is Markov relative to a missing data DAG  $G$  is identified **if and only if**  $G$  does not contain the following two structures: (Nabi et al., 2020)

| self-censoring edge:  $L_j^{(1)} \rightarrow R_j$

| collider:  $L_j^{(1)} \rightarrow R_i \rightarrow R_j$



self-censoring



colluder

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

# Proof sketch

- | Absence of self-censoring edges and colluders imply:

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

- | Odds ratio parameterization:

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

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

- |  $\text{OR}(R_i, R_j / R_{-\{i,j\}} = 1, L^{(1)})$  is identified via “symmetric argument,”

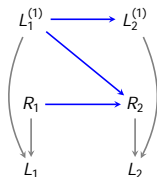
- | It is not a function of  $L_i^{(1)}$  and it is not a function of  $L_j^{(1)}$

- | 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)



## Example: odds ratio parameterization fails!

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



- | As we saw earlier,  $p(R = 1 / \text{pa}_G(R))$  is easily identifiable as follows:

$$p(R = 1 / \text{pa}_G(R)) = p(R_1 = 1) \times p(R_2 = 1 / R_1 = 1, L_1).$$

Therefore, the target law is identified. However, the odds ratio parameterization approach fails here.

## **Part 3. Non/Semi-parametric Estimation**

# General Theory of Estimation

- | Let  $L^{(1)}$  be the *full* data
- | Let  $R$  be missingness indicators ( $R = 1$  is no missingness)
- |  $C = (R, L^{(1)})$  be the *complete* data
- | Let  $L_r(L^{(1)})$  be the components of  $L^{(1)}$  when  $R = r$ 
  - |  $L_1(L^{(1)}) = L^{(1)}$
- | Let  $O = (R, L_R(L^{(1)}))$  be the *observed* data
- | We assume  $P[R = 1/L^{(1)}] > 0$
- | Let  $F_{L^{(1)}}$  represent the model for  $L^{(1)}$ , indexed by parameters  $\mu$  (scalar target parameter) and  $\nu$  (nuisance)
- | Let  $F_{R/L^{(1)}}$  represent the model for  $R/L^{(1)}$ , indexed by parameters  $\nu$  (nuisance)

# General Theory of Estimation

- | Let  $O_1 = (F_{L(1)})$  be the collection of nuisance scores for  $\mu$  based on observation of  $C$
- | Let  $O_2 = (F_{R/L(1)})$  be the collection of nuisance scores for  $\mu$  based on observation of  $C$
- | Let  $O_j = \overline{R(g \cdot \cdot)_j}$  where
  - |  $\Pi_j$  is a projection operator that maps from  $C$  to  $\Lambda_j$  and
  - |  $g(\cdot) = E[\cdot/O]$ .
- |  $O = \overline{O_1 + O_2}$
- |  $O_i = \overline{O_{i1} + O_{i2}}$
- | Influence functions (up to a normalizing constant) for regular and asymptotically linear (RAL) estimator of  $\mu$  live in  $O_i$ .
- | A RAL estimator  $\hat{\mu}$  of  $\mu$  has the property that

$$\sqrt{n}(\hat{\mu} - \mu) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \underbrace{(O_i; \mu, \cdot, \cdot)}_{\text{Influence Function}} + o_P(1)$$

# General Theory of Estimation

- Consider  $W(O_i; \mu, \dots)$
- Let  $\hat{\mu}_n$  and  $\hat{\theta}_n$  be estimators for  $\mu$  and  $\theta$ , respectively.
- Estimate  $\mu$  as the solution to

$$\sum_{i=1}^n W(O_i; \mu, \dots) = 0$$

- Under regularity conditions on the rate of convergence of  $\hat{\mu}_n$  and  $\hat{\theta}_n$ ,  $\hat{\mu}_n$  will be RAL with influence function

$$\frac{E[W(O_i; \mu, \dots)]}{\mu}^{-1} W(O_i; \mu, \dots)$$

- Could also use profile estimators  $\hat{\mu}(\hat{\theta}_n)$  and  $\hat{\theta}_n$  and estimate  $\mu$  as the solution to

$$\sum_{i=1}^n W(O_i; \mu, \hat{\theta}_n) = 0$$

# General Theory of Estimation

$$O_1 = \frac{I(R=1)}{P[R=1/L^{(1)}]} a(L^{(1)}) + b(O) : a(L^{(1)}) \quad E[b(O)/L^{(1)}] = 0$$
$$O_2 = b(O) : b(O)$$

Plan:

- | Write out an expression for the elements of  $O_1$
- | Find restrictions on these elements to ensure orthogonality to  $O_2$

# Application of General Theory of Estimation

- | Let  $L^{(1)} = (L_1^{(1)}, L_2^{(1)})$
- | Let  $R = (R_1, R_2)$
- | Let  $p_{ij}(L^{(1)}) = P[R_1 = i, R_2 = j/L^{(1)}]$

Any observed data random variable can be 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 are  $c_{11}(L_1^{(1)}, L_2^{(1)})$ ,  $c_{10}(L_1^{(1)})$ ,  $c_{01}(L_2^{(1)})$  and  $c_{00}$  ensure that the  $E[b(O)/L^{(1)}] = 0$

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

# Application of General Theory of Estimation

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

$$\begin{aligned} & -\frac{R_1 R_2}{11(L^{(1)})} {}_{10}(L^{(1)}) + R_1(1 - R_2) {}_{c10}(L_1^{(1)}) + \\ & -\frac{R_1 R_2}{11(L^{(1)})} {}_{01}(L^{(1)}) + (1 - R_1)R_2 {}_{c01}(L_2^{(1)}) + \\ & -\frac{R_1 R_2}{11(L^{(1)})} {}_{00}(L^{(1)}) + (1 - R_1)(1 - R_2) {}_{c00} \end{aligned}$$



# Application of General Theory of Estimation

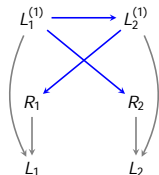
- | Let  $\mu = E[h(L^{(1)})]$  for a specified function  $h(\cdot)$
- | Suppose no restrictions are placed on the distribution of  $L^{(1)}$
- | Let  $\mathcal{T}_{1,t}$  be the nuisance tangent space under no-restrictions.
- | The elements of  $\mathcal{T}_{1,t}$  will be proportional to  $h(L^{(1)}) - \mu$
- | If there are restrictions placed on the distribution of  $L^{(1)}$ , then  $\mathcal{T}_{1,t} \subset \mathcal{T}_1$ .
- | Thus,  $h(L^{(1)}) - \mu \in \mathcal{T}_1$ .

# Application of General Theory of Estimation

We will work with

$$\begin{aligned} o_{1,t} &= \frac{R_1 R_2}{11(L^{(1)})} h(L^{(1)}) - \mu \\ &\quad - \frac{R_1 R_2}{11(L^{(1)})} {}_{10}(L^{(1)}) + R_1(1 - R_2) c_{10}(L_1^{(1)}) + \\ &\quad - \frac{R_1 R_2}{11(L^{(1)})} {}_{01}(L^{(1)}) + (1 - R_1)R_2 c_{01}(L_2^{(1)}) + \\ &\quad - \frac{R_1 R_2}{11(L^{(1)})} {}_{00}(L^{(1)}) + (1 - R_1)(1 - R_2) c_{00} : \\ &\quad c_{10}(L_1^{(1)}), c_{01}(L_2^{(1)}), c_{00} \end{aligned}$$

# Block-Parallel Model



$$g_{jk}(L^{(1)}) = g_{1j}(L_2^{(1)}) \{1 - g_{1j}(L_2^{(1)})\}^{1-j} g_{2k}(L_1^{(1)}) \{1 - g_{2k}(L_1^{(1)})\}^{1-k}$$

$$g_2 = \{R_1 - g_{1j}(L_2^{(1)})\}g_1(L_2^{(1)}) + \{R_2 - g_{2k}(L_1^{(1)})\}g_2(L_1^{(1)}) : g_1(L_2^{(1)}), g_2(L_1^{(1)})$$

# Block-Parallel Model

What choices of  $c_{10}(L_1^{(1)})$ ,  $c_{01}(L_2^{(1)})$ ,  $c_{00}$  ensure orthogonality with all elements of  $\mathcal{L}_2^{(1)}$ ?

$$\begin{aligned}
 E \frac{R_1 R_2}{11(L^{(1)})} h(L^{(1)}) - \mu \{R_1 - 1(L_2^{(1)})\} + \\
 - \frac{R_1 R_2}{11(L^{(1)})} c_{10}(L_1^{(1)}) \{R_1 - 1(L_2^{(1)})\} + \\
 - \frac{R_1 R_2}{11(L^{(1)})} c_{01}(L_2^{(1)}) \{R_1 - 1(L_2^{(1)})\} + \\
 - \frac{R_1 R_2}{11(L^{(1)})} c_{00} \{R_1 - 1(L_2^{(1)})\} \Big|_{\mathcal{L}_2^{(1)}} = 0 \quad (1)
 \end{aligned}$$

$$\begin{aligned}
 E & \frac{R_1 R_2}{11(L^{(1)})} h(L^{(1)}) - \mu \{R_2 - {}_2(L_1^{(1)})\} + \\
 & - \frac{R_1 R_2}{11(L^{(1)})} {}_{10}(L^{(1)}) + R_1(1 - R_2) c_{10}(L_1^{(1)}) \{R_2 - {}_2(L_1^{(1)})\} + \\
 & - \frac{R_1 R_2}{11(L^{(1)})} {}_{01}(L^{(1)}) + (1 - R_1)R_2 c_{01}(L_2^{(1)}) \{R_2 - {}_2(L_1^{(1)})\} + \\
 & - \frac{R_1 R_2}{11(L^{(1)})} {}_{00}(L^{(1)}) + (1 - R_1)(1 - R_2) c_{00} \{R_2 - {}_2(L_1^{(1)})\} \Big| L_1^{(1)} = 0 \quad (2)
 \end{aligned}$$

# Block-Parallel Model

(1) implies that

$$c_{01}(L_2^{(1)}; c_{00}) = \frac{E \ h(L^{(1)}) - \mu \Big| L_2^{(1)}}{E \ 2(L_1^{(1)}) \Big| L_2^{(1)}} - \frac{E \ 1 - 2(L_1^{(1)}) \Big| L_2^{(1)}}{E \ 2(L_1^{(1)}) \Big| L_2^{(1)}} c_{00}$$

$\underbrace{\hspace{10em}}_{c_{01}(L_2^{(1)})} \qquad \underbrace{\hspace{10em}}_{c_{01}(L_2^{(1)})}$

(2) implies that

$$c_{10}(L_1^{(1)}; c_{00}) = \frac{E \ h(L^{(1)}) - \mu \Big| L_1^{(1)}}{E \ 1(L_2^{(1)}) \Big| L_1^{(1)}} - \frac{E \ 1 - 1(L_2^{(1)}) \Big| L_1^{(1)}}{E \ 1(L_2^{(1)}) \Big| L_1^{(1)}} c_{00}$$

$\underbrace{\hspace{10em}}_{c_{10}(L_1^{(1)})} \qquad \underbrace{\hspace{10em}}_{c_{10}(L_1^{(1)})}$

So,  $\mathcal{O}_1$  contains a collection of elements indexed by  $c_{00}$ .

# Block-Parallel Model

We will work with

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

Need estimators for

$$| \quad \beta_1(L_2^{(1)})$$

$$| \quad \beta_2(L_1^{(1)})$$

$$| \quad E[\beta_2(L_1^{(1)})/L_2^{(1)}]$$

$$| \quad E[\beta_1(L_2^{(1)})/L_1^{(1)}]$$

$$| \quad E[h(L^{(1)})/L_1^{(1)}]$$

$$| \quad E[h(L^{(1)})/L_2^{(1)}]$$



# Block-Parallel Model

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

$$\begin{aligned}
 E & \frac{R_1 R_2}{11(L^{(1)})} h(L^{(1)}) - \mu + \\
 & - \frac{R_1 R_2}{11(L^{(1)})} {}_{10}(L^{(1)}) + R_1(1 - R_2) \quad c_{10}(L_1^{(1)}) - c_{10}(L_1^{(1)})c_{00} + \\
 & - \frac{R_1 R_2}{11(L^{(1)})} {}_{01}(L^{(1)}) + (1 - R_1)R_2 \quad c_{01}(L_2^{(1)}) - c_{01}(L_2^{(1)})c_{00} + \\
 & - \frac{R_1 R_2}{11(L^{(1)})} {}_{00}(L^{(1)}) + (1 - R_1)(1 - R_2) \quad c_{00}^2
 \end{aligned}$$

# Block-Parallel Model

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

$$\begin{aligned}
 E & \frac{R_1 R_2}{_{11}(L^{(1)})} h(L^{(1)}) - \mu + \\
 & - \frac{R_1 R_2}{_{11}(L^{(1)})} {}_{10}(L^{(1)}) + R_1(1 - R_2) c_{10}(L_1^{(1)}) + \\
 & - \frac{R_1 R_2}{_{11}(L^{(1)})} {}_{10}(L^{(1)}) + R_1(1 - R_2) - c_{10}(L_1^{(1)})c_{00} + \\
 & - \frac{R_1 R_2}{_{11}(L^{(1)})} {}_{01}(L^{(1)}) + (1 - R_1)R_2 c_{01}(L_2^{(1)}) + \\
 & - \frac{R_1 R_2}{_{11}(L^{(1)})} {}_{01}(L^{(1)}) + (1 - R_1)R_2 - c_{01}(L_2^{(1)})c_{00} + \\
 & - \frac{R_1 R_2}{_{11}(L^{(1)})} {}_{00}(L^{(1)}) + (1 - R_1)(1 - R_2) c_{00} \times \\
 & - \frac{R_1 R_2}{_{11}(L^{(1)})} {}_{10}(L^{(1)}) + R_1(1 - R_2) - c_{10}(L_1^{(1)}) + \\
 & - \frac{R_1 R_2}{_{11}(L^{(1)})} {}_{01}(L^{(1)}) + (1 - R_1)R_2 - c_{01}(L_2^{(1)}) + \\
 & - \frac{R_1 R_2}{_{11}(L^{(1)})} {}_{00}(L^{(1)}) + (1 - R_1)(1 - R_2) = 0
 \end{aligned}$$

# Block-Parallel Model

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

where

$$\begin{aligned}
 b = & -E \frac{{}_{10}(L^{(1)})}{{}_{11}(L^{(1)})} + 1 \quad {}_{10}(L^{(1)}) \quad c_{10}(L_1^{(1)})^2 - \\
 & E \frac{{}_{10}(L^{(1)}) \quad {}_{01}(L^{(1)})}{{}_{11}(L^{(1)})} \quad 2c_{10}(L_1^{(1)})c_{01}(L_2^{(1)}) + \\
 & E \frac{{}_{10}(L^{(1)}) \quad {}_{00}(L^{(1)})}{{}_{11}(L^{(1)})} \quad 2c_{10}(L_1^{(1)}) - \\
 & E \frac{{}_{01}(L^{(1)})}{{}_{11}(L^{(1)})} + 1 \quad {}_{01}(L^{(1)}) \quad c_{01}(L_2^{(1)})^2 + \\
 & E \frac{{}_{01}(L^{(1)}) \quad {}_{00}(L^{(1)})}{{}_{11}(L^{(1)})} \quad 2c_{01}(L_2^{(1)}) + \\
 & E \frac{{}_{00}(L^{(1)})}{{}_{11}(L^{(1)})} + 1 \quad {}_{00}(L^{(1)})
 \end{aligned}$$

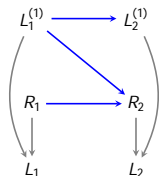
# Block-Parallel Model

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

where

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

# Block-Conditional Model



$$g_{jk}(L^{(1)}) = \frac{j}{1} \{1 - \frac{1}{2}\}^{1-j} \times \frac{1}{2} (1, L_1^{(1)})^{jk} \{1 - \frac{1}{2} (1, L_1^{(1)})\}^{j(1-k)} \frac{1}{2} (0, L_1^{(1)})^{(1-j)k} \{1 - \frac{1}{2} (0, L_1^{(1)})\}^{(1-j)(1-k)}$$

$$\frac{1}{2} = \{R_1 - \frac{1}{2}\} g_1 + R_1 \{R_2 - \frac{1}{2} (1, L_1^{(1)})\} g_2(1, L_1^{(1)}) + (1 - R_1) \{R_2 - \frac{1}{2} (0, L_1^{(1)})\} g_2(0, L_1^{(1)}) : g_1, g_2(1, L_1^{(1)}), g_2(0, L_1^{(1)})$$

# Block-Conditional Model

What choices of  $c_{10}(L_1^{(1)})$ ,  $c_{01}(L_2^{(1)})$ ,  $c_{00}$  ensure orthogonality with all elements of  $\mathcal{L}_2$ ?

$$\begin{aligned}
 E & \frac{R_1 R_2}{11(L^{(1)})} h(L^{(1)}) - \mu R_1 \{R_2 - 2(1, L_1^{(1)}) + \\
 & - \frac{R_1 R_2}{11(L^{(1)})} c_{10}(L_1^{(1)}) + R_1(1 - R_2) c_{10}(L_1^{(1)}) R_1 \{R_2 - 2(1, L_1^{(1)}) + \\
 & - \frac{R_1 R_2}{11(L^{(1)})} c_{01}(L_2^{(1)}) + (1 - R_1) R_2 c_{01}(L_2^{(1)}) R_1 \{R_2 - 2(1, L_1^{(1)}) + \\
 & - \frac{R_1 R_2}{11(L^{(1)})} c_{00} + (1 - R_1)(1 - R_2) c_{00} R_1 \{R_2 - 2(1, L_1^{(1)}) \} \Big|_{L_1^{(1)}} = 0
 \end{aligned} \tag{3}$$

# Block-Conditional Model

$$\begin{aligned}
 E & \frac{R_1 R_2}{11(L^{(1)})} h(L^{(1)}) - \mu (1 - R_1) \{R_2 - {}_2(0, L_1^{(1)})\} + \\
 & - \frac{R_1 R_2}{11(L^{(1)})} {}_{10}(L^{(1)}) + R_1(1 - R_2) c_{10}(L_1^{(1)})(1 - R_1) \{R_2 - {}_2(0, L_1^{(1)})\} + \\
 & - \frac{R_1 R_2}{11(L^{(1)})} {}_{01}(L^{(1)}) + (1 - R_1)R_2 c_{01}(L_2^{(1)})(1 - R_1) \{R_2 - {}_2(0, L_1^{(1)})\} + \\
 & - \frac{R_1 R_2}{11(L^{(1)})} {}_{00}(L^{(1)}) + (1 - R_1)(1 - R_2) c_{00}(1 - R_1) \{R_2 - {}_2(0, L_1^{(1)})\} \Big| L_1^{(1)} = 0
 \end{aligned} \tag{4}$$

## Block-Conditional Model

$$\begin{aligned}
 E & \frac{R_1 R_2}{11(L^{(1)})} h(L^{(1)}) - \mu \{R_1 - 1\} + \\
 & - \frac{R_1 R_2}{11(L^{(1)})} c_{10}(L_1^{(1)}) \{R_1 - 1\} + \\
 & - \frac{R_1 R_2}{11(L^{(1)})} c_{01}(L_2^{(1)}) \{R_1 - 1\} + \\
 & - \frac{R_1 R_2}{11(L^{(1)})} c_{00} \{R_1 - 1\} = 0 \quad (5)
 \end{aligned}$$



# Block-Conditional Model

(4) implies

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

(5) implies

$$E[\alpha_2(0, L_1^{(1)})c_{01}(L_2^{(1)})] + c_{00}E[1 - \alpha_2(0, L_1^{(1)})] = 0$$

(4) and (5) imply that

$$c_{00}E[\alpha_2(0, L_1^{(1)})] + c_{00}E[1 - \alpha_2(0, L_1^{(1)})] = c_{00} = 0$$

# 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.

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

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

# Block-Conditional Model

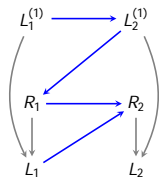
We will work with

$$\begin{aligned} & \frac{R_1 R_2}{11(L^{(1)})} h(L^{(1)}) - \mu + -\frac{R_1 R_2}{11(L^{(1)})} 10(L^{(1)}) + R_1(1 - R_2) \frac{E[h(L^{(1)}) - \mu/L_1^{(1)}]}{1} \\ &= \frac{R_1}{1} \frac{R_2}{2(1, L_1^{(1)})} h(L^{(1)}) + 1 - \frac{R_2}{2(1, L_1^{(1)})} E[h(L^{(1)})/L_1^{(1)}] - \mu \end{aligned}$$

Need estimators for

- |  $1$
- |  $2(1, L_1^{(1)})$
- |  $E[h(L^{(1)})/L_1^{(1)}]$

# Permutation Model



$$j_k(L^{(1)}) = {}_1(L_2^{(1)})^j \{1 - {}_1(L_2^{(1)})\}^{1-j} \times \\ {}_2(1, L_1^{(1)})^{jk} \{1 - {}_2(1, L_1^{(1)})\}^{j(1-k)} {}_2(0)^{(1-j)k} \{1 - {}_2(0)\}^{(1-j)(1-k)}$$

$${}_2 = \{R_1 - {}_1(L_2^{(1)})\} g_1(L_2^{(1)}) + R_1 \{R_2 - {}_2(1, L_1^{(1)})\} g_2(1, L_1^{(1)}) + \\ (1 - R_1) \{R_2 - {}_2(0)\} g_2(0) : g_1(L_2^{(1)}), g_2(1, L_1^{(1)}), g_2(0)$$

# Permutation Model

What choices of  $c_{10}(L_1^{(1)})$ ,  $c_{01}(L_2^{(1)})$ ,  $c_{00}$  ensure orthogonality with all elements of  $\mathcal{L}_2^{(1)}$ ?

$$\begin{aligned}
 E \frac{R_1 R_2}{11(L^{(1)})} h(L^{(1)}) - \mu \{R_1 - 1(L_2^{(1)})\} + \\
 - \frac{R_1 R_2}{11(L^{(1)})} 10(L^{(1)}) + R_1(1 - R_2) c_{10}(L_1^{(1)}) \{R_1 - 1(L_2^{(1)})\} + \\
 - \frac{R_1 R_2}{11(L^{(1)})} 01(L^{(1)}) + (1 - R_1)R_2 c_{01}(L_2^{(1)}) \{R_1 - 1(L_2^{(1)})\} + \\
 - \frac{R_1 R_2}{11(L^{(1)})} 00(L^{(1)}) + (1 - R_1)(1 - R_2) c_{00} \{R_1 - 1(L_2^{(1)})\} \Big|_{L_2^{(1)}} = 0 \quad (6)
 \end{aligned}$$

# Permutation Model

$$\begin{aligned}
 E & \frac{R_1 R_2}{11(L^{(1)})} h(L^{(1)}) - \mu R_1 \{R_2 - {}_2(1, L_1^{(1)})\} + \\
 & - \frac{R_1 R_2}{11(L^{(1)})} {}_{10}(L^{(1)}) + R_1(1 - R_2) c_{10}(L_1^{(1)}) R_1 \{R_2 - {}_2(1, L_1^{(1)})\} + \\
 & - \frac{R_1 R_2}{11(L^{(1)})} {}_{01}(L^{(1)}) + (1 - R_1)R_2 c_{01}(L_2^{(1)}) R_1 \{R_2 - {}_2(1, L_1^{(1)})\} + \\
 & - \frac{R_1 R_2}{11(L^{(1)})} {}_{00}(L^{(1)}) + (1 - R_1)(1 - R_2) c_{00} R_1 \{R_2 - {}_2(1, L_1^{(1)})\} \Big|_{L_1^{(1)}} = 0
 \end{aligned} \tag{7}$$

# Permutation Model

$$\begin{aligned}
 E & \frac{R_1 R_2}{11(L^{(1)})} h(L^{(1)}) - \mu (1 - R_1) \{R_2 - {}_2(0)\} + \\
 & - \frac{R_1 R_2}{11(L^{(1)})} {}_{10}(L^{(1)}) + R_1(1 - R_2) c_{10}(L_1^{(1)})(1 - R_1) \{R_2 - {}_2(0)\} + \\
 & - \frac{R_1 R_2}{11(L^{(1)})} {}_{01}(L^{(1)}) + (1 - R_1)R_2 c_{01}(L_2^{(1)})(1 - R_1) \{R_2 - {}_2(0)\} + \\
 & - \frac{R_1 R_2}{11(L^{(1)})} {}_{00}(L^{(1)}) + (1 - R_1)(1 - R_2) c_{00}(1 - R_1) \{R_2 - {}_2(0)\} = 0
 \end{aligned} \tag{8}$$



# Permutation Model

(8) implies

$$c_{00} = \frac{E[\{1 - {}_1(L_2^{(1)})\}c_{01}(L_2^{(1)})]}{E[\{1 - {}_1(L_2^{(1)})\}]}$$

(6) implies

$$c_{01}(L_2^{(1)}) = \frac{E[h(L^{(1)}) - \mu | L_2^{(1)}]}{{}_2(0)} - \frac{1 - {}_2(0)}{{}_2(0)} c_{00}$$

Together, this implies that

$$c_{00} = \frac{E[h(L^{(1)})\{1 - {}_1(L_2^{(1)})\}]}{E[1 - {}_1(L_2^{(1)})]} - \mu$$

and

$$c_{01}(L_2^{(1)}) = \frac{E[h(L^{(1)})/L_2^{(1)}]}{{}_2(0)} - \frac{1 - {}_2(0)}{{}_2(0)} \frac{E[h(L^{(1)})\{1 - {}_1(L_2^{(1)})\}]}{E[1 - {}_1(L_2^{(1)})]} - \mu$$

Adding (7) implies

$$c_{10}(L_1^{(1)}) = \frac{E[h(L^{(1)}) - E[h(L^{(1)})/L_2^{(1)}]/L_1^{(1)}]}{E[\text{ }_1(L_2^{(1)})/L_1^{(1)}]} + \frac{E[\text{ }_1(L_2^{(1)})E[h(L^{(1)})/L_2^{(1)}]/L_1^{(1)}]}{E[\text{ }_1(L_2^{(1)})/L_1^{(1)}]} - \mu$$

# Permutation Model

We will work with

$$\begin{aligned}
 & \frac{R_1 R_2}{11(L^{(1)})} h(L^{(1)}) + \\
 & - \frac{R_1 R_2}{11(L^{(1)})} {}_{10}(L^{(1)}) + R_1(1 - R_2) \times \\
 & \frac{E[h(L^{(1)}) - E[h(L^{(1)})/L_2^{(1)}]/L_1^{(1)}]}{E[{}_1(L_2^{(1)})/L_1^{(1)}]} + \frac{E[{}_1(L_2^{(1)})E[h(L^{(1)})/L_2^{(1)}]/L_1^{(1)}]}{E[{}_1(L_2^{(1)})/L_1^{(1)}]} + \\
 & - \frac{R_1 R_2}{11(L^{(1)})} {}_{01}(L^{(1)}) + (1 - R_1)R_2 \times \\
 & \frac{E[h(L^{(1)})/L_2^{(1)}]}{{}_2(0)} - \frac{1 - {}_2(0)}{{}_2(0)} \frac{E[h(L^{(1)})\{1 - {}_1(L_2^{(1)})\}]}{E[1 - {}_1(L_2^{(1)})]} + \\
 & - \frac{R_1 R_2}{11(L^{(1)})} {}_{00}(L^{(1)}) + (1 - R_1)(1 - R_2) \times \\
 & \frac{E[h(L^{(1)})\{1 - {}_1(L_2^{(1)})\}]}{E[1 - {}_1(L_2^{(1)})]} - \mu
 \end{aligned}$$

# Permutation Model

Need estimators for

- |  $\mu_2(0)$
- |  $\mu_2(1, L_1^{(1)})$
- |  $\mu_1(L_2^{(1)})$
- | Conditional means of functions of  $L_2^{(1)}$  given  $L_1^{(1)}$  and conditional means of functions of  $L_1^{(1)}$  given  $L_2^{(1)}$ 
  - |  $E[h(L^{(1)})/L_1^{(1)}]$
  - |  $E[h(L^{(1)})/L_2^{(1)}]$
  - |  $E[\mu_1(L_2^{(1)})/L_1^{(1)}]$
  - |  $E[\mu_1(L_2^{(1)})E[h(L^{(1)})/L_2^{(1)}]/L_1^{(1)}]$
  - |  $E[E[h(L^{(1)})/L_2^{(1)}]/L_1^{(1)}]$

**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.

## Part IV. Sensitivity analysis

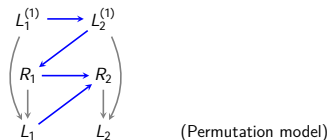
- | Assessing deviations from assumptions.

# General questions and thoughts

1. How can identification results guide estimation strategies?
2. What if there exist variables that are not just missing but completely unobserved?
3. Similar to DAGs, absence of an edge in an m-DAG implies a restriction of the form  $A \perp B / C$ . Is this restriction testable from observed finite samples?
4. What if the model is not nonparametrically identified?

# I. Estimation strategies (weighted estimating equations)

- How identification results can guide estimation strategies?



- Let our parameter of interest be  $h = E[h(L_1^{(1)}, L_2^{(1)})]$ , which can be rewritten as:

$$h = E \left[ \frac{R_1 R_2}{p(R_1 = 1 / L_2^{(1)}) p(R_2 = 1 / R_1 = 1, L_1)} \times h(L_1^{(1)}, L_2^{(1)}) \right].$$

$p(R_1 = 1 / L_2^{(1)})$  is identified from an intervention distribution where  $R_2$  is intervened on:

$$p(R_1 = 1 / L_2^{(1)}) = p(R_1 = 1 / L_2, do(R_2 = 1)).$$

- How does identification of  $p(R_1 = 1 / L_2^{(1)})$  help with estimation?

Assume  $p(R_1 = 1 / L_2^{(1)}) = p(R_1 = 1 / L_2^{(1)}; \cdot)$ ,  $R^p$  and  $E[U(R_1, L_2^{(1)}; \cdot)] = 0$ , then:

$$E \left[ \frac{R_2}{p(R_2 / R_1, L_1)} \times U(R_1, L_2^{(1)}; \cdot) \right] = 0.$$



# I. Estimation strategies (EIF derivation)

$$\begin{aligned}h &= E h(L_1^{(1)}, L_2^{(1)}) \\ &= E \frac{R_1 R_2}{p(R_1 = 1 / L_2^{(1)}) p(R_2 = 1 / R_1 = 1, L_1)} \times h(L_1^{(1)}, L_2^{(1)})\end{aligned}$$

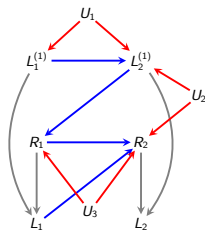
- | A more reasonable estimator for  $h$  is the estimator derived based on the nonparametric efficient influence function (EIF).
- | The core idea of deriving the EIF for  $h$  is to use an intermediate variable that first takes care of the missingness of  $L_1^{(1)}$ , and then  $L_2^{(1)}$  in a sequential manner

$$\begin{aligned}h(L_2^{(1)}) &= \frac{R_1}{p(R_1 = 1 / L_2^{(1)})} \times h(L_1^{(1)}, L_2^{(1)}), \\ h &= E \frac{R_2}{p(R_2 = 1 / R_1 = 1, L_1)} \times h(L_2^{(1)}) .\end{aligned}$$

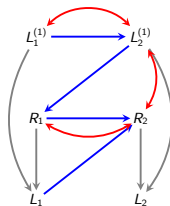
(Robins, 1997)

## II. 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)  $G(V, U)$



(b)  $G(V)$

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

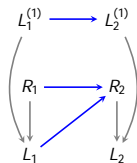
$U_1$ : genotypic traits,  $U_2$ : occupation,  $U_3$ : ethnicity

- | Identification results are extended to missing data ADMGs, and results remain complete for full law identification (Nabi et al., 2020).

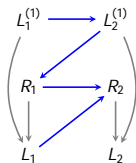
### III. Testable implications in m-DAGs

- | Similar to DAGs, absence of an edge in an m-DAG implies a restriction of the form  $A \perp\!\!\!\perp B \mid C$ . 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)
  - | A self-censoring mechanism imposes no restriction on the observed data law
  - | Another 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, 2022).

### III. Testable implications



MAR model



Permutation supermodel

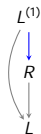
- | Is  $R_1 \perp L_2^{(1)}$ ?
- | Fit  $p(R_1)$  and  $p(R_1 / L_2^{(1)})$  and compare the goodness of fits
- | Use a weighted estimating equation to fit  $p(R_1 / L_2^{(1)}; )$

$$P_n \frac{R_2}{p(R_2 / R_1, L_1)} \times U(.; ) = 0,$$

where  $P_n[U(.; )] = 0$  with respect to the full law.

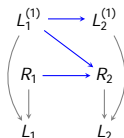
## IV. Missing data non-identification

- What if a parameter of interest is not identified from the observed data law?



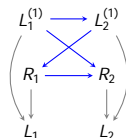
Self-censoring

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



Collider

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

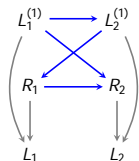


Criss-cross

$p(L^{(1)})$  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)

## IV. Partial/parametric identification



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

Cross-cross structure

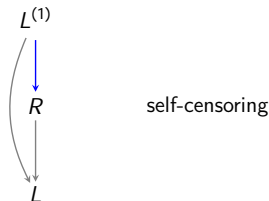
- Partial identification:  $p(L_1^{(1)} / L_2^{(1)})$  and  $\text{OR}(L_1^{(1)}, L_2^{(1)})$  are identified.
  - We can test  $L_1^{(1)} \perp L_2^{(1)}$  without further assumptions.
- Under what conditions  $p(L_1^{(1)}, L_2^{(1)})$  is identified? assume  $p(L_1^{(1)})$  and  $p(L_2^{(1)} / L_1^{(1)})$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

$$L_1^{(1)} \sim \exp \left\{ \frac{l_1^{(1)} \eta_1 - b_{\eta_1}(\eta_1)}{\Phi_{\eta_1}} + c_{\eta_1}(\eta_1; \Phi_{\eta_1}) \right\}$$

$$L_2^{(1)} / L_1^{(1)} \sim \exp \left\{ \frac{l_2^{(1)} - b(\eta_2)}{\Phi} + c(l_2^{(1)}; \Phi) \right\}, \quad g(\mu(\eta_2)) = \eta_2 + l_1^{(1)}.$$

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

## IV. Alternative handling of non-identified models



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

$$p(L^{(1)} / R = 0) = p(L^{(1)} / R = 1) \times \exp(-S(L^{(1)}))$$

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

# 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.



# Ananke: Software for causal inference

## ***ananke-causal:***

A python package for causal inference using the language of graphical models.

## **Highlights:**

- Identification algorithms
- Semiparametric inference
- Surrogate experiments
- Missing data



## **Links:**

- | Documentation: <https://ananke.readthedocs.io/en/latest/index.html>
- | Gitlab repository: <https://gitlab.com/causal/ananke>
- | Python package index: <https://pypi.org/project/ananke-causal/>

**Contributors:** Rohit Bhattacharya, Jaron Lee, and RN

# 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. Semiparametric identifiability and estimation under missing not at random mechanisms, 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.
- 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.

## References II

- R. Nabi and R. Bhattacharya. On testability and goodness of fit tests in missing data models. *arXiv preprint arXiv:2203.00132*, 2022.
- 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.
- T. S. Richardson, R. J. Evans, J. M. Robins, and I. Shpitser. Nested markov properties for acyclic directed mixed graphs. *arXiv preprint arXiv:1701.06686*, 2017.
- 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.

## References III

- 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.
- 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.
- 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.
- 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.

**Razieh Nabi, Ph.D.**

Rollins Assistant Professor  
Department of Biostatistics and Bioinformatics  
Rollins School of Public Health  
Emory University

🐦 @razielnabi

✉ razieh.nabi@emory.edu

🏠 <https://razielnabi.com>

**Daniel Scharfstein, Sc.D.**

Chief of the Division of Biostatistics  
Department of Population Health Sciences  
School of Medicine  
University of Utah

🐦 @dscharf3

✉ Daniel.Scharfstein@hsc.utah.edu

🏠 <https://medicine.utah.edu/dscharf>

Slides are available at: [https://razielnabi.com/files/acic\\_miss.pdf](https://razielnabi.com/files/acic_miss.pdf)  
Upon using the materials, please cite appropriately.