Directed Mixed Graphs for Latent Variables

Qing Zhou

UCLA Department of Statistics

Stats 212 Graphical Models
Lecture Notes

Outline

Acyclic directed mixed graphs

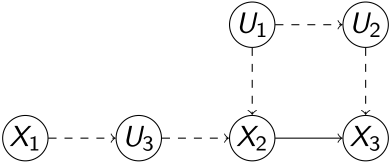
Latent projection of a DAG (Tian and Pearl 2002b):
Given a DAG with latent variables $\mathcal{G}(V \cup L)$, where V is observed and L latent, the *latent projection* $\mathcal{G}(V)$ is constructed as follows:

1. $\mathcal{G}(V)$ contains an edge $a \to b$ if there is a directed path $a \to \cdots \to b$ in $\mathcal{G}(V \cup L)$ with all intermediate vertices in L.
2. $\mathcal{G}(V)$ contains an edge $a \leftrightarrow b$ if there is a collider-free path $a \leftarrow \cdots \to b$ with all intermediate vertices in L.

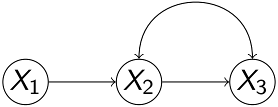
Note: Step 1 adds all directed edges $a \to b$ in $\mathcal{G}(V \cup L)$ to $\mathcal{G}(V)$.

Acyclic directed mixed graphs

DAG $\mathcal{G}(V \cup L)$, $V = \{X_1, X_2, X_3\}$ and $L = \{U_1, U_2, U_3\}$:



Latent projection $\mathcal{G}(V)$ is an acyclic directed mixed graph (ADMG):



Acyclic directed mixed graphs

Definitions. Let $\mathcal{G} = (V, E)$ be a directed mixed graph, i.e. a graph with two types of edges: directed ($\rightarrow$) or bidirected ($\leftrightarrow$).

- A *path* is a sequence of distinct adjacent edges, of any type or orientation, between distinct vertices.
  directed path: $a \rightarrow \cdots \rightarrow b$. bidirected path: $a \leftrightarrow \cdots \leftrightarrow b$.
- If $a \rightarrow b$, then a is a parent of b and b is a child of a.
- If there is a directed path from a to d or $a = d$, we say a is an ancestor of d and d is a descendant of a. Accordingly define non-descendant.
- If $a \leftrightarrow b$, then a is a sibling of b.
- notation: $\mathrm{pa}_{\mathcal{G}}(a)$, $\mathrm{ch}_{\mathcal{G}}(a)$, $\mathrm{an}_{\mathcal{G}}(a)$, $\mathrm{de}_{\mathcal{G}}(a)$, $\mathrm{nd}_{\mathcal{G}}(a)$, and $\mathrm{sib}_{\mathcal{G}}(a)$.

Acyclic directed mixed graphs

- A *directed cycle* is a path of the form $v \to \cdots \to w$ along with an edge $w \to v$.
- An acyclic directed mixed graph (ADMG) is a mixed graph containing no directed cycles.
- A topological sort of an ADMG is defined in the same way as for a DAG: $a \to b$ implies $a \prec b$.

Acyclic directed mixed graphs

m-separation:

- A vertex z is a collider on a path if $\rightarrow z \leftarrow$, $\leftrightarrow z \leftrightarrow$, $\rightarrow z \leftrightarrow$, or $\leftrightarrow z \leftarrow$; otherwise, z is a non-collider.
- *m*-connection: A path between a and b is *m*-connecting given C if (i) every non-collider on the path is not in C and (ii) every collider on the path is an ancestor of C $(\text{an}(C) := \cup_{a \in C} \text{an}(a))$.
- *m*-separation: If there is no path *m*-connecting a and b given C, then a and b are *m*-separated given C.
- If $\mathcal{G}$ is a DAG, *m*-separation is identical to *d*-separation.

Acyclic directed mixed graphs

m-separation:

Proposition 1 (Richardson et al. (2023))

Let $\mathcal{G}(V \cup L)$ be a DAG and $\mathcal{G}(V)$ be its latent projection. For disjoint subsets $A, B, C \subset V$, A and B are *d*-separated given C in $\mathcal{G}(V \cup L)$ if and only if A and B are *m*-separated given C in $\mathcal{G}(V)$.

- On every path between $a, b \in V$ in $\mathcal{G}(V \cup L)$, colliders (resp. non-colliders) in V are also colliders (resp. non-colliders) on a path in $\mathcal{G}(V)$.
- ADMG $\mathcal{G}(V)$ captures all conditional independence constraints among the observed variables V in the DAG $\mathcal{G}(V \cup L)$ with latent variables.

Acyclic directed mixed graphs

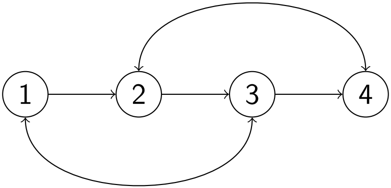
Districts in ADMG $\mathcal{G}(V)$:

- The *district* of vertex v, denoted $\text{dis}_{\mathcal{G}}(v)$, is the set of vertices that are connected to v by a bidirected path (including v itself).
- A district of $\mathcal{G}$ is a maximal bidirected-connected set of vertices.
- A district corresponds to a confounded component (c-component) (Tian and Pearl 2002b).
- Districts specify variable partitions that define terms in the factorization of $\mathbb{P}(V)$.

Denote districts by $\mathcal{D}(\mathcal{G}) = \{D : D \text{ is a district of } \mathcal{G}\}$.
Define $\text{pa}_{\mathcal{G}}(D) := (\cup_{a \in D} \text{pa}_{\mathcal{G}}(a)) \setminus D$.

Factorizations on ADMGs

District factorization:



- Districts of $\mathcal{G}$: $D_1 = \{1, 3\}, D_2 = \{2, 4\}$.
- $\text{pa}_{\mathcal{G}}(D_1) = \{2\}$, $\text{pa}_{\mathcal{G}}(D_2) = \{1, 3\}$.

Using $a \leftrightarrow b \Leftrightarrow a \leftarrow u \rightarrow b$:

$$p(x_1, \dots, x_4) = \left[\sum_{u_1} p(x_1 \mid u_1)p(x_3 \mid x_2, u_1)p(u_1)\right] \times \left[\sum_{u_2} p(x_2 \mid x_1, u_2)p(x_4 \mid x_3, u_2)p(u_2)\right]$$
$$= q_{1,3}(x_1, x_3 \mid x_2) \times q_{2,4}(x_2, x_4 \mid x_1, x_3).$$

Factorizations on ADMGs

$$p(x_1, \ldots, x_4) = q_{1,3}(x_1, x_3 \mid x_2) \times q_{2,4}(x_2, x_4 \mid x_1, x_3)$$
$$= q_{D_1}(x_{D_1} \mid \mathrm{pa}_{\mathcal{G}}(D_1)) \times q_{D_2}(x_{D_2} \mid \mathrm{pa}_{\mathcal{G}}(D_2)).$$

For general case, district factorization:

$$\mathbb{P}(V) = \prod_{D \in \mathcal{D}(\mathcal{G})} q_D(x_D \mid \mathrm{pa}_{\mathcal{G}}(D)). \qquad (1)$$

- Each factor $q_Y(y \mid W)$ is called a *kernel*, i.e. a probability density of Y with W being a parameter: $\sum_y q_Y(y \mid W = w) = 1$, $\forall w$.
- $q_Y(y \mid W = w) = \mathbb{P}(Y = y \mid do(w))$ and thus, in general $q_Y(y \mid W) \neq \mathbb{P}(Y = y \mid W = w)$.

Factorizations on ADMGs

Express $q_D(x_D \mid \mathrm{pa}_{\mathcal{G}}(D))$ as $\prod_{i \in D} p(x_i \mid \cdots)$:

- The Markov blanket of $a \in V$ in ADMG $\mathcal{G}$ is

$$\mathrm{mb}(a, \mathcal{G}) := \mathrm{pa}_{\mathcal{G}}(D) \cup (D \setminus \{a\}),$$

  where $D = \mathrm{dis}_{\mathcal{G}}(a)$. We have $a \perp \mathrm{nd}(a) \mid \mathrm{mb}(a)$.

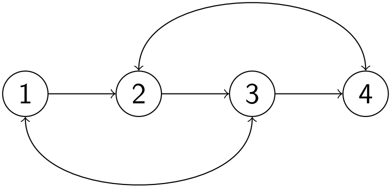
- Suppose that $1 \prec \cdots \prec p = |V|$ is a topological sort of $\mathcal{G}$. Let $V_i = \{1, \dots, i\}$ and $\mathcal{G}_i$ be the induced subgraph on V_i. Then $X_i \perp X_k \mid \mathrm{mb}(i, \mathcal{G}_i)$, $k < i$:

$$q_D(x_D \mid \mathrm{pa}_{\mathcal{G}}(D)) = \prod_{i \in D} p(x_i \mid \mathrm{mb}(i, \mathcal{G}_i)). \tag{2}$$

- Putting together into (1), we get

$$\mathbb{P}(V) = \prod_{i \in V} p(x_i \mid \mathrm{mb}(i, \mathcal{G}_i)). \tag{3}$$

Factorizations on ADMGs



Sort: $1 \prec 2 \prec 3 \prec 4$.
$\mathrm{mb}(1, \mathcal{G}_1) = \varnothing$,
$\mathrm{mb}(2, \mathcal{G}_2) = \{1\}$,
$\mathrm{mb}(3, \mathcal{G}_3) = \{1, 2\}$,
$\mathrm{mb}(4, \mathcal{G}_4) = \{1, 2, 3\}$.

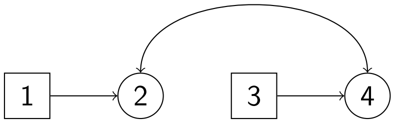
$$q_{1,3}(x_1, x_3 \mid x_2) = p(x_1)p(x_3 \mid x_1, x_2), \tag{4}$$
$$q_{2,4}(x_2, x_4 \mid x_1, x_3) = p(x_2 \mid x_1)p(x_4 \mid x_1, x_2, x_3). \tag{5}$$
$$\Rightarrow \; p(x) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1, x_2)p(x_4 \mid x_1, x_2, x_3).$$

This does NOT imply any conditional independence among $X_1, \cdots, X_4$.
In particular, $X_1 \not\perp X_4 \mid S$ for any $S \subseteq \{X_2, X_3\}$ (m-connected) even though no edge between X_1 and X_4.

Generalized CI constraints

No edge between X_1 and X_4 encodes a generalized conditional independence a.k.a. Verma constraint (Verma and Pearl 1990).

Represent $q_{2,4}(x_2, x_4 \mid x_1, x_3) = p(x_2, x_4 \mid do(x_1, x_3))$ by a conditional ADMG (CADMG) with graph $\mathcal{G}^{|W}$ ($W = \{1, 3\}$) by cutting all edges with an arrow into W:



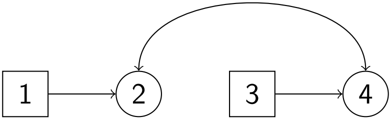
- Two types of vertices in a CADMG $\mathcal{G}(V, W)$: (i) Random $V = \{2, 4\}$; (ii) Fixed $W = \{1, 3\}$.
- Kernel $q_V(x_V \mid x_W)$ is an (intervention) distribution for V after fixing W.
- We may further fix other random vertices if they are *fixable*.

Generalized CI constraints

Definition 1

The set of *fixable* vertices in a CADMAG $\mathcal{G}(V, W)$ is $F(\mathcal{G}) := \{v \in V : \text{dis}_{\mathcal{G}}(v) \cap \text{de}_{\mathcal{G}}(v) = \{v\}\}$.

v is fixable if none of its descendants is in the same district.



Fixable vertices $=\{2, 4\}$.

Fix vertex 2: (i) $\mathcal{G}(V = \{4\}, W = \{1, 2, 3\})$

(ii) New kernel district-factorized according to $\mathcal{G}(\{4\}, \{1, 2, 3\})$:

$$q_4(x_4 \mid x_2, x_1, x_3) = f_4(x_4 \mid x_3). \quad \text{nested factorization} \tag{6}$$

Generalized CI constraints

The new kernel $q_4(x_4 \mid x_2, x_1, x_3)$ is defined by the fixing operator:

Definition 2

Given a kernel $q_V(x_V \mid W)$ associated with a CADMG $\mathcal{G} = \mathcal{G}(V, W)$, for any fixable vertex $r \in F(\mathcal{G})$, the fixing operator ϕ_r yields a new kernel

$$q_{V\setminus r}(x_{V\setminus r} \mid r, W) = \phi_r(q_V; \mathcal{G}) := \frac{q_V(x_V \mid W)}{q_V(x_r \mid \mathrm{mb}(r, \mathcal{G}), W)}. \tag{7}$$

- $q_V(x_r \mid \mathrm{mb}(r, \mathcal{G}), W)$ is a conditional distribution calculated from $q_V(x_V \mid W)$.
- If r is fixable, then r can be sorted as the last vertex in its district and its causal effect $\mathbb{P}(V \setminus r \mid do(r); \mathcal{G})$ on $V \setminus r$ can be calculated by (7).

Generalized CI constraints

Apply ϕ_2 on $q_{2,4}(x_2, x_4 \mid x_1, x_3)$ ($\text{mb}(2, \mathcal{G}) = \{1, 4, 3\}$):

$$\begin{aligned} q_4(x_4 \mid x_2, x_1, x_3) = \phi_2(q_{2,4}; \mathcal{G}) &= \frac{q_{2,4}(x_2, x_4 \mid x_1, x_3)}{q_{2,4}(x_2 \mid x_4, x_1, x_3)} \\ &= q_{2,4}(x_4 \mid x_1, x_3) \\ &= \sum_{x_2'} q_{2,4}(x_2', x_4 \mid x_1, x_3) \\ &= \sum_{x_2'} p(x_2' \mid x_1) p(x_4 \mid x_1, x_2', x_3). \quad \text{by (5)} \end{aligned}$$

By nested factorization (6):

$$\sum_{x_2'} p(x_2' \mid x_1) p(x_4 \mid x_1, x_2', x_3) = f_4(x_4 \mid x_3)$$

does not depend on x_1, which is a GCI constraint.

Generalized CI constraints

Nested factorization:

- Suppose $p(x)$ factorizes by a DAG $\mathcal{G}(V \cup L)$ and $\mathcal{G} = \mathcal{G}(V)$ is the ADMG defined by latent projection.
- For a valid fixing sequence $w = (w_1, \ldots, w_r)$, let $\phi_w(\mathcal{G})$ be the CADMG after fixing w sequentially and $\mathcal{D}_w = \mathcal{D}(\phi_w(\mathcal{G}))$ be the districts of (random vertices) in $\phi_w(\mathcal{G})$.

Theorem 1 (Richardson et al. (2023))

For any valid fixing sequence w,

$$\phi_w(p(x_V); \mathcal{G}) = \prod_{D \in \mathcal{D}_w} f_D^w(x_D \mid \mathrm{pa}_{\mathcal{G}}(D))$$

for some kernels $f_D^w(x_D \mid \mathrm{pa}_{\mathcal{G}}(D))$.

Generalized CI constraints

Algorithm to find systematically CI and GCI constraints implied by ADMG: Tian and Pearl (2002b).

Input: ADMG $\mathcal{G}(V)$; assume V is sorted, $1 \prec \ldots \prec p$.
Output: CI and GCI constraints on $p(x_V)$ implied by $\mathcal{G}(V)$.

For $i = 1$ to p,
Part 1: CI constraints $X_i \perp X_k \mid \text{mb}(i, \mathcal{G}_i)$, $k < i$, $k \notin \text{mb}(i, \mathcal{G}_i)$.
Part 2: $S \leftarrow \text{dis}_{\mathcal{G}_i}(i)$ and $G \leftarrow \phi_{[i]\setminus S}(\mathcal{G}_i)$ ($[i] = \{1, \ldots, i\}$).
For each descendent set $D \subset S$ s.t. $i \notin D$: Let $D' = S \setminus D$.

1. $\sum_{x_D} q_S = q_{D'}$ (fixing D); $G' = \phi_D(G)$.
2. If G' has 2 or more districts, $E \leftarrow \text{dis}_{G'}(i)$ and $q_{D'}/\sum_{x_i} q_{D'}$ is a function of $\text{mb}(i, G') = E \cup \text{pa}_{G'}(E)$.
3. Repeat part 2 with $S \leftarrow E$ and $G \leftarrow \phi_{S\setminus E}(G)$.

Identification of causal effects

Identification of causal effects given an ADMG $\mathcal{G}(V)$:

- Let $k \in V$ be a single variable and $S \subset V$.
- The causal effect of X_k on S is identifiable (from observational data) if $\mathbb{P}(S \mid do(X_k))$ can be computed from the joint distribution $\mathbb{P}(V)$.

Theorem 2 (Tian and Pearl (2002a))

If there is no bidirected path connecting X_k to any of its children in $\mathcal{G}_{\text{an}(S)}$, then the causal effect of X_k on S is identifiable.

- Recent results: Theorem 48 in Richardson et al. (2023), Corollary 16 in Bhattacharya et al. (2022).

Constructive proof of Theorem 2:

1. Let $V = \mathrm{an}(S)$, $\mathcal{G} = \mathcal{G}_{\mathrm{an}(S)}$ and $M = V \setminus \{S \cup k\}$. Then

$$p(x_S \mid do(x_k)) = \sum_{x_M} p(x_{V\setminus k} \mid do(x_k)).$$

2. Let $D = \mathrm{dis}_{\mathcal{G}}(k) \in \mathcal{D} = \mathcal{D}(\mathcal{G})$. Since $\mathrm{ch}(k) \cap D = \varnothing$,

$$p(x_{V\setminus k} \mid do(x_k)) = \sum_{x_k'} q_D(x_D \mid \mathrm{pa}_{\mathcal{G}}(D)) \prod_{D' \in \mathcal{D}} q_{D'}(x_{D'} \mid \mathrm{pa}_{\mathcal{G}}(D')).$$

If X_k is fixable, we may instead apply fixing operator:

$$p(x_{V\setminus k} \mid do(x_k)) = \phi_k(p(x); \mathcal{G}) = \frac{p(x_V)}{p(x_k \mid \mathrm{mb}(k, \mathcal{G}))}.$$

Identification of causal effects

The identify algorithm by Tian and Pearl (2002a) reformulated with fixing operators: Theorem 48 in Richardson et al. (2023).

Let $\mathcal{G} = \mathcal{G}(V)$. For $A, Y \subset V$, want to identify $\mathbb{P}(Y \mid do(A))$.

- Let $Y^* = \text{an}_{\mathcal{G}_{V \setminus A}}(Y) \supseteq Y$: there is a directed path from every $v \in Y^*$ to Y not blocked by A.

  Since $V \setminus (A \cup Y) = [V \setminus (A \cup Y^*)] \cup (Y^* \setminus Y)$,

$$\begin{aligned}\mathbb{P}(Y \mid do(A)) &= \sum_{V \setminus (A \cup Y)} \mathbb{P}(V \setminus A \mid do(A)) \\ &= \sum_{Y^* \setminus Y} \sum_{V \setminus (A \cup Y^*)} \mathbb{P}(V \setminus A \mid do(A)) \\ &= \sum_{Y^* \setminus Y} \mathbb{P}(Y^* \mid do(A)), \quad (Y^* \text{ is ancestral}).\end{aligned}$$

Identification of causal effects

- Let $\mathcal{D}^* = \mathcal{D}(\mathcal{G}_{Y^*})$. District factorization on $\mathcal{G}_{Y*}$ shows

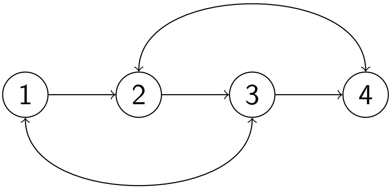
$$\mathbb{P}(Y^* \mid do(A)) = \prod_{D \in \mathcal{D}^*} \mathbb{P}[D \mid do(\mathrm{pa}_{\mathcal{G}}(D))].$$

- If every D is intrinsic (i.e. $V \setminus D$ is fixable), then $\mathbb{P}[D \mid do(\mathrm{pa}_{\mathcal{G}}(D))] = \phi_{V \setminus D}(\mathbb{P}(V); \mathcal{G})$, and

$$\therefore \quad \mathbb{P}(Y \mid do(A)) = \sum_{Y^* \setminus Y} \prod_{D \in \mathcal{D}^*} \phi_{V \setminus D}(\mathbb{P}(V); \mathcal{G}). \tag{8}$$

  Otherwise, $\mathbb{P}[D \mid do(\mathrm{pa}_{\mathcal{G}}(D))]$ is not identifiable for some D, and $\mathbb{P}(Y \mid do(A))$ is not identifiable.

Identification of causal effects



Find $p(x_4 \mid do(x_2))$.
$Y = \{4\}, A = \{2\}$
$Y^* = \{3, 4\}$
$\mathcal{D}^* = \{D_1, D_2\} = \{\{3\}, \{4\}\}$

$$p(x_3 \mid do(x_2)) = \phi_{1,2,4}(p(x_V); \mathcal{G}) = \phi_1(q_{1,3}(x_1, x_3 \mid x_2); \mathcal{G}^{|2,4})$$
$$= \sum\nolimits_{x_1} p(x_1)p(x_3 \mid x_1, x_2).$$
$$p(x_4 \mid do(x_3)) = \phi_{2,1,3}(p(x_V); \mathcal{G}) = \phi_2(q_{2,4}(x_2, x_4 \mid x_1, x_3); \mathcal{G}^{|1,3})$$
$$= \sum\nolimits_{x_2'} p(x_2' \mid x_1)p(x_4 \mid x_1, x_2', x_3).$$

$$\therefore \quad p(x_4 \mid do(x_2)) = \sum\nolimits_{x_3} p(x_3 \mid do(x_2))p(x_4 \mid do(x_3)).$$

Linear SEM associated with ADMG

Given an ADMG $\mathcal{G}$ with directed edge set E_d and bidirected edge set E_b, define linear SEM

$$X_j = \sum_{i \in \mathrm{pa}_{\mathcal{G}}(j)} \beta_{ij} X_i + \varepsilon_j, \quad j = 1, \ldots, p. \tag{9}$$

$$(\varepsilon_1, \ldots, \varepsilon_p) \sim \mathcal{N}_p(0, \Omega).$$

- $B \in \mathcal{B}(E_d) := \{(\beta_{ij})_{p \times p} : \beta_{ij} = 0 \text{ if } i \to j \notin E_d\}$.
- $\Omega \in \mathcal{P}(E_b) := \{(\omega_{ij})_{p \times p} : \omega_{ij} = 0 \text{ if } i \leftrightarrow j \notin E_b\}$.

The linear SEM (9) defines a family of multivariate Gaussian distributions $\mathcal{N}_p(0, \Sigma)$ with

$$\Sigma = \Sigma_{\mathcal{G}}(B, \Omega) := (\mathbf{I} - B)^{-\mathsf{T}} \Omega (\mathbf{I} - B)^{-1}.$$

Linear SEM associated with ADMG

Definition 3 (Identifiability)

The linear SEM for an ADMG $\mathcal{G}$ is said to be identifiable if $\Sigma_{\mathcal{G}}(B, \Omega)$ is an *injective* (one-to-one) map from $\mathcal{B}(E_d) \times \mathcal{P}(E_b)$ to the set of positive definite matrices.

Reachable closure (Shpitser et al. 2018).

Definition 4

For a CADMG $\mathcal{G}(V, W)$, a reachable subset $C \subseteq V$ is called a reachable closure for $S \subseteq C$ if the set of fixable vertices in $\phi_{V \setminus C}(\mathcal{G})$ is a subset of S.

- Reachable closure is unique for any $S \subseteq V$, denoted $\langle S \rangle$.
- $\langle S \rangle$ is the set of random vertices in $\phi_{\neg S}(\mathcal{G})$ (fixing as many vertices in $V \setminus S$ as possible).

Linear SEM associated with ADMG

Graphical criterion for identifiability:

Theorem 3 (Drton et al. (2011))

The linear SEM for an ADMG $\mathcal{G}(V)$ is identifiable if and only if $\langle v \rangle = \{v\}$ for all $v \in V$.

- Identifiability means that given $\mathcal{G}(V)$ and Σ, there is a unique set of parameters (B, Ω) for the linear SEM. Thus, given $\mathcal{G}(V)$ and data, one may estimate (B, Ω).
- Example: $a \rightarrow s \leftarrow b$ and $b \leftrightarrow a \leftrightarrow s$.
  - $\langle s \rangle = \{a, b, s\}$ (a, b are not fixable in $V \setminus s$).
  - $\mathcal{G}_{a,b,s}$ contains a sink node s and its parents a, b in the same district.
  - Linear SEM is *not* identifiable.

Ancestral graphs

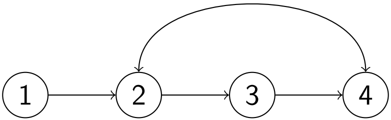
Motivations.

- A class of ADMGs that represents conditional independences among V in a DAG $\mathcal{G}(V, L)$ with latent variables L.
- Retains the ancestral relationships and hence causal relations among V.
- Its equivalence class can be constructed from CI relations learned from observational data.
- Does *not* preserve all confounding structures in $\mathcal{G}(V, L)$, i.e. bidirected edges in the latent projection.
- Does *not* represent GCI (Verma) constraints: potential loss of efficiency.

Ancestral graphs

Definitions. Let $\mathcal{G} = (V, E)$ be an ADMG.

- An *almost directed cycle* occurs when $a \leftrightarrow b$ and $a \in \mathrm{an}_{\mathcal{G}}(b)$ (removing the arrowhead at b results in a directed cycle).
- Let $L \subset V$. An *inducing path relative to L* is a path on which every intermediate vertex $\notin L$ is a collider and every collider is an ancestor of an endpoint. If $L = \varnothing$, call it an inducing path.



Almost directed cycle: $(2, 3, 4, 2)$.
Inducing path: $1 \to 2 \leftrightarrow 4$
$\Rightarrow$ 1 and 4 not m-separated by any subsets.

Ancestral graphs

Definition 5 (MAG)

A mixed graph is a maximal ancestral graph (MAG) if

(i) it does not contain any directed or almost directed cycles (ancestral);

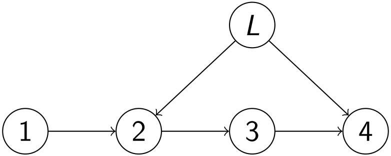
(ii) there is no inducing path between any two non-adjacent vertices (maximal).

Constructing MAG $\mathcal{M}$ from DAG $\mathcal{G} = \mathcal{G}(V, L)$:

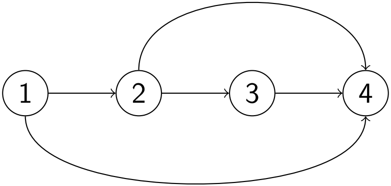
1. For each pair $a, b \in V$, a and b are adjacent in $\mathcal{M}$ iff there is an inducing path between them relative to L in $\mathcal{G}$.
2. For each adjacent pair (a, b) in $\mathcal{M}$,
   orient $a \to b$ in $\mathcal{M}$ if $a \in \mathrm{an}_{\mathcal{G}}(b)$; orient $b \to a$ in $\mathcal{M}$ if $b \in \mathrm{an}_{\mathcal{G}}(a)$; orient $a \leftrightarrow b$ otherwise.

Ancestral graphs

DAG $\mathcal{G}(V \cup L)$



MAG



Every edge among V in a DAG (trivial inducing path) is an edge in MAG.

Inducing paths relative to L:

$1 \to 2 \leftarrow L \to 4 \Rightarrow 1 \to 4$ in $\mathcal{M}$

$2 \leftarrow L \to 4 \Rightarrow 2 \to 4$ in $\mathcal{M}$

$1, 2$ are ancestors of 4.

Ancestral graphs

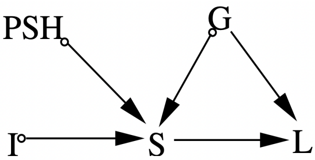
Equivalence class of a MAG:

- Two MAGs are Markov equivalent if they have the same set of *m*-separations.
  Sufficient and necessary conditions: same skeleton and *v*-structures, and share some covered colliders (Proposition 2, Zhang (2008b)).
- The equivalence class $[\mathcal{M}]$ of a MAG $\mathcal{M}$ is represented by a partial ancestral graph $\mathcal{P}$:
  i. $\mathcal{P}$ has the same adjacencies (skeleton) as $\mathcal{M}$;
  ii. A mark of arrowhead is in $\mathcal{P}$ iff it is shared by all MAGs in $[\mathcal{M}]$;
  iii. A mark of tail is in $\mathcal{P}$ iff it is shared by all MAGs in $[\mathcal{M}]$.

  Edge marks in (ii) and (iii) are invariant across $[\mathcal{M}]$; other variable marks are represented by $\circ$ in $\mathcal{P}$.

Ancestral graphs

Example PAG (Zhang 2008a)



I: income, S: smoking, PSH: parent smoking habits, G: genotype, L: lung cancer

- $I \circ\!\!\rightarrow S = I \rightarrow S$ or $I \leftrightarrow S$.
- preserve the 3 v-structures at the collider S.
- no directed or almost directed cycles among G, S, L.

The FCI algorithm

Constraint-based learning of MAGs by the FCI (fast causal inference) algorithm (Spirtes et al. 1999):

Use CI constraints learned from observational data to construct the equivalence class of a MAG represented by a PAG:

- skeleton;
- invariant marks (arrowheads and tails).

The FCI algorithm

Algorithm outline

1: $E \leftarrow$ edge set of the complete undirected graph on V. Every edge is $\circ\!\!-\!\!\circ$.

2: **for** $(i, j) \in E$ **do**

3: Search for a subset S_{ij} such that $X_i \perp X_j \mid S_{ij}$. If found, $E \leftarrow E \setminus \{(i, j), (j, i)\}$ and store S_{ij}.

4: **end for**

5: Orient edges in v-structures based on E and $\{S_{ij}\}$.

6: Apply orientation rules R1 to R4 (Zhang 2008b) until none of them applies.

7: Apply orientation rules R8 to R10 (Zhang 2008b) until none of them applies.

The FCI algorithm

Suppose $\mathcal{M}$ is the true MAG, and assume we have CI oracle.

- Line 1 to 5: similar to the PC algorithm.
- After Line 4: correctly construct the skeleton $sk(\mathcal{M})$.
- After Line 6: identify all and only invariant arrowheads in $[\mathcal{M}]$.
- After Line 7: identify all and only invariant tails in $[\mathcal{M}]$.

Theorem 4 (Theorem 4, Zhang (2008b))

Given a perfect conditional independence oracle, the FCI algorithm returns the PAG for the true MAG $\mathcal{M}$.

References I

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.

Mathias Drton, Rina Foygel, and Seth Sullivant. Global identifiability of linear structural equation models. *The Annals of Statistics*, 39(2):865–886, 2011.

Thomas Richardson, R.J. Evans, J.M. Robins, and I. Shpitser. Nested Markov properties for acyclic directed mixed graphs. *Annals of Statistics*, to appear, 2023.

I. Shpitser, R.J. Evans, and Thomas S. Richardson. Acyclic linear SEMs obey the nested markov property. *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*, 2018.

References II

Peter Spirtes, Christopher Meek, and Thomas S Richardson. An algorithm for causal inference in the presence of latent variables and selection bias. *Computation, Causation, and Discovery*, pages 211–252, 1999.

Jin Tian and Judea Pearl. A general identification condition for causal effects. *Proceedings of the AAAI*, pages 567–573, 2002a.

Jin Tian and Judea Pearl. On the testable implications of causal models with hidden variables. *Proceedings of the 18th Conferences on Uncertainty in Artificial Intelligence*, pages 519–527, 2002b.

Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, pages 220–227, 1990.

References III

Jiji Zhang. Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 9:1437–1474, 2008a.

Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172:1873–1896, 2008b.