Causal Structure Discovery from Distributions Arising from Mixtures of DAGs

Basil Saeed, Snigdha Panigrahi and Caroline Uhler

Presented By:

Vinay Verma

Aug 21, 2020
Outline

- Definitions and Preliminaries
- Mixture DAG and Markov Property
- Learning from mixture data
- Experiments and results
Definitions and Preliminaries

Directed Acyclic Graph (DAG): A graph $\mathcal{D} = (V, E)$ defined over directed ($\rightarrow$) edge and does not contains cycle. Here $V$ is set of vertices and $E$ is set of edges in the graph $\mathcal{D}$.

Mixed Graph: A graph $\mathcal{M} = (V, D, B)$ defined over directed ($\rightarrow$) and bi-directed ($\leftrightarrow$) edges. $D : \text{Directed edge, } B : \text{Bi-directed edges}$.

Ancestral Graph: A mixed graph is ancestral if it does not contains any directed cycle and whenever there is a bi-directed edge $u \leftrightarrow v$, then there is no directed path from $u$ to $v$.

D-Separation: For any disjoint $A, B, C \subseteq V$, $A$ and $B$ are D-separated by $C$ if all the path are blocked wrt. $C$. 
Maximal Ancestral Graph (MAG): An ancestral graph where any non-adjacent pair of nodes is d-separated given some subset of nodes is called maximal ancestral graph.

Markov Property: Given a graph $\mathcal{M}$ with nodes $V$, we associate to each node $v \in V$ a random variable $X_v$ and denote the joint distribution of $X_V := (x_v : v \in V)$ by $p_{X_V}$.

Distribution $p_{X_V}$ is said to satisfy the Markov property with respect to $\mathcal{M}$: if for any disjoint $A, B, C \subseteq V$ such that $A$ and $B$ are d-separated given $C$ in $\mathcal{M} \Rightarrow X_A \perp \perp X_B | X_C$ in $p_{X_V}$

Faithfulness Assumption: For any disjoint $A, B, C \subseteq V$: if $X_A \perp \perp X_B | X_C$ in $p_{X_V} \Rightarrow A$ and $B$ are d-separated given $C$ in $\mathcal{M}$
Motivation:

- In various applications, data used for causal structure discovery is heterogeneous and it arise from different causal models (mixture) on the same set of variables.
- The causal effects of the mixture distribution will be difficult to represent by a single DAG.
- Single DAG inferred from such samples cannot identify differences between the component DAGs in the mixture, which may be critical for personalized biomedical interventions.
Mixture DAG

Consider $K$ DAGs $\{\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(K)}\}$ with $\mathcal{D}^{(j)} = (V, E^{(j)})$ for $1 \leq j \leq K$, i.e., these $K$ DAGs are defined on the same set of nodes.

**Definition:** Let $v^{(j)}$ denote vertex $v$ in DAG $j$ and let $[V] := \cup_{1 \leq j \leq K} V^{(j)}$ denote the vertices of the $K$ component DAGs. The mixture DAG, denoted by $\mathcal{D}_\mu$, has nodes $[V] \cup \{y\}$ and edges $E_\mu$:

- $E_\mu$ consisting of edges in each component DAG.
- $E_\mu$ contains additional edge $\bigcup_{j=1}^{K} \{y \rightarrow v^{(j)} : v \in V \setminus V_{\text{INV}}\}$.

Where:

$$V_{\text{INV}} = \left\{ v \in V : p^{(j)}(x_v | x_{\text{pa}_{\mathcal{D}(j)}(v)}) = p^{(k)}(x_v | x_{\text{pa}_{\mathcal{D}(k)}(v)}) \right\}$$  \hspace{1cm} (1)
Figure 1: Example of mixture DAG
Let $J$ be a discrete variable taking values in $\{1, \ldots, K\}$ with probabilities $p_J(j)$ for each $j \in \{1, \ldots, K\}$. Defining a joint distribution $p_\mu$ over $X_V \cup J$ by

$$p_\mu(x_V, j) := p_J(j) \cdot p^{(j)}(x_V),$$

(2)

The observed mixture distribution is obtained by marginalizing $p_\mu$ over the unobserved index variable $J$. 

Basil Saeed, Snigdha Panigrahi and Caroline Uhler  
Presented By: Vinay Verma  
Causal Structure Discovery from Distributions  
Aug 21, 2020 8 / 30
Theorem (Markov Property)

Let $A, B, C \subseteq V$ be disjoint. If $[A]$ and $[B]$ are d-separated given $[C]$ in the mixture DAG $D_\mu$, then $X_A \perp \perp X_B | X_C$ in the mixture distribution $p_\mu$.

Where: $[A] = \bigcup_{1 \leq j \leq K} A^{(j)}$

**Example-2:** Consider Example-1 since $[1] = \{1^{(1)}, 1^{(2)}\}$ and $[4] = \{4^{(1)}, 4^{(2)}\}$ are d-separated given $\emptyset$ in the mixture DAG, then the mixture distribution $p_\mu(x_1, x_2, x_3, x_4)$ satisfies $X_1 \perp \perp X_4$. 
Lemma

Let $A, B, C \subseteq V$ be disjoint. If for all $1 \leq j \leq K$ it holds that

(a) $A^{(j)}$ and $B^{(j)}$ are d-separated given $C^{(j)}$, and;

(b) $A^{(j)}$ and $y$ are d-separated given $C^{(j)}$ in $\tilde{D}^{(j)}$,

then $X_A \perp \perp X_B \mid X_C$ in $p_\mu$, implying the factorization

$$p^{(j)}(x_A, x_B \mid x_C) = p^{(1)}(x_A \mid x_C)p^{(j)}(x_B \mid x_C)$$

for all $1 \leq j \leq K$. 

Markov Property: Proof Sketch

By definition of $p_\mu$ in (2),

$$p_\mu(x_A, x_B|x_C) = \sum_{j=1}^K p^{(j)}(x_A, x_B|x_C)p_J(j),$$

and hence as a consequence of above Lemma we obtain

$$p_\mu(x_A, x_B|x_C) = \sum_{j=1}^K p^{(1)}(x_A|x_C)p^{(j)}(x_B|x_C)p_J(j)$$

$$= p^{(1)}(x_A|x_C) \sum_{j=1}^K p^{(j)}(x_B|x_C)p_J(j),$$

providing a factorization of the desired form.
Definition (Mixture Faithfulness)

The mixture distribution $p_\mu$ is faithful with respect to a mixture DAG $D_\mu$ if for any disjoint $A, B, C \subseteq V$ with $X_A \perp\!\!\!\!\perp X_B | X_C$ in $p_\mu$ it holds that $[A]$ and $[B]$ are d-separated given $[C]$. 
Figure 2: Example of mixture DAG
Consider the distributions $p^{(1)}(x_V), p^{(2)}(x_V)$ on $V = \{1, 2, 3, 4\}$ that factor according to the DAGs $\mathcal{D}^{(1)}, \mathcal{D}^{(2)}$, respectively, shown in Figure 2. Namely

$$p^{(1)}(x_V) = p^{(1)}(x_1)p^{(1)}(x_2|x_1)p^{(1)}(x_3)p^{(1)}(x_4),$$

$$p^{(2)}(x_V) = p^{(2)}(x_1)p^{(2)}(x_2)p^{(2)}(x_3|x_4)p^{(2)}(x_4),$$

Where:

$$p^{(1)}(x_1) = \mathcal{N}(x_1; 0, 1), \quad p^{(2)}(x_1) = \mathcal{N}(x_1; 0, 1),$$

$$p^{(1)}(x_2|x_1) = \mathcal{N}(x_2; x_1, 1), \quad p^{(2)}(x_2) = \mathcal{N}(x_2; 0, 2),$$

$$p^{(1)}(x_3) = \mathcal{N}(x_3; 0, 1), \quad p^{(2)}(x_3|x_4) = \mathcal{N}(x_3; x_4, 1),$$

$$p^{(1)}(x_4) = \mathcal{N}(x_4; 0, 1), \quad p^{(2)}(x_4) = \mathcal{N}(x_4; 0, 1).$$
Then, defining \( p_\mu(x_V) := \sum_{j=1}^{2} p^{(j)}(x_V)p_J(j) \), for some \( J \sim p_J(j) \), we obtain that

\[
p_\mu(x_2, x_3) = \int p_\mu(x_V) dx_1 dx_4
= \int p_J(1)p^{(1)}(x_1)p^{(1)}(x_2|x_1)p^{(1)}(x_3)p^{(1)}(x_4)dx_1 dx_4
+ \int p_J(2)p^{(2)}(x_1)p^{(2)}(x_2)p^{(2)}(x_3|x_4)p^{(2)}(x_4)dx_1 dx_4
= p_J(1) \mathcal{N}(x_2; 0, 2) \mathcal{N}(x_3; 0, 1)
+ p_J(2) \mathcal{N}(x_2; 0, 2) \mathcal{N}(x_3; 0, 2)
= \mathcal{N}(x_2; 0, 2) \left(p_J(1)\mathcal{N}(x_3; 0, 1) + p_J(2)\mathcal{N}(x_3; 0, 2)\right)
= f(x_2)g(x_3),
\]

which implies that \( X_2 \perp \perp X_3 \) in \( p_\mu \), although in the mixture DAG corresponding to \( p_\mu \) shown in Figure 2 the nodes 2 and 3 are d-connected via the path through \( y \).
This example was carefully crafted; even a slight perturbation such as choosing \( p^{(2)}(x_2) = \mathcal{N}(x_2; 0, 2.001) \) would have meant that \( p_\mu(x_2, x_3) \) does not factor, indicating that mixture-faithfulness violations are rare.

**Proposition (Realizability of \( D_\mu \))**

*For any mixture DAG \( D_\mu \), there exists (almost sure) a mixture distribution \( p_\mu \) that is faithful with respect to \( D_\mu \).*
Without knowing the membership of each sample to a component DAG, we cannot generally learn the structure of $D^{(j)}$, for each $j$ from the data.

Since the mixing variable is latent, an intuitive approach is to apply Fast Causal Inference (FCI) to learn a MAG representation of $p_\mu$.

**Problem:** Under the mixture-faithfulness assumption the conditional independence relations in a mixture distribution $p_\mu$ may not be representable by any MAG.
For the above example no MAG exists, that satisfies: A d-sep from B given C in $\tilde{\mathcal{M}}$ if and only if \([A] \text{ d-sep from } [B] \text{ given } [C] \text{ in } \mathcal{D}_\mu\).

- figure-(b) is only possible skeleton of MAG, otherwise it violate d-sep statement in $\mathcal{D}_\mu$.
- $\tilde{\mathcal{M}}$ would also need to contain the colliders $4 \rightarrow 5 \leftarrow 2$ and $1 \rightarrow 2 \leftarrow 5$ to respect the d-separation relations resulting from $4^{(2)} \rightarrow 5^{(2)} \leftarrow y \rightarrow 2^{(2)}$ and $1^{(1)} \rightarrow 2^{(1)} \leftarrow y \rightarrow 5^{(1)} \implies 2 \leftrightarrow \tilde{\mathcal{M}} 5$

- Hence, $2 \leftrightarrow \tilde{\mathcal{M}} 5$ and $2 \in \text{ an} \tilde{\mathcal{M}} (5)$, violating the ancestral property.
POSET Compatibility

Let $\mathcal{M}^{(j)}$ be the MAG constructed from the induced sub-DAG $\bar{\mathcal{D}}^{(j)}$. The MAGs $\mathcal{M}^{(1)}, \ldots, \mathcal{M}^{(K)}$ are said to be compatible with the same poset if there exists a partial order $\pi$ on $V$ such that for all $1 \leq j \leq K$ it holds that:

- $u \in \text{an}_{\mathcal{M}^{(j)}}(v) \Rightarrow u \prec_\pi v$
- $u \leftrightarrow_{\mathcal{M}^{(j)}} v \Rightarrow u \not\preceq_\pi v.$

Union Graph:

(a) $\mathcal{M}^{(1)}$

(b) $\mathcal{M}^{(2)}$

(c) $\mathcal{M}_\cup$
Property of $\mathcal{M}_\cup$

- Under the assumption that $\mathcal{M}^{(1)} \ldots, \mathcal{M}^{(K)}$ are compatible with the same poset, $\mathcal{M}_\cup$ is a MAG.
- The d-separations in the mixture DAG are equivalent to d-separation statements in a MAG ($\mathcal{M}_\cup$).

**Theorem**

Let $A, B, C \subseteq V$ be disjoint. If the component MAGs satisfy the poset compatibility assumption, then $A$ and $B$ are d-separated given $C$ in $\mathcal{M}_\cup$ if and only if $[A]$ and $[B]$ are d-separated given $[C]$ in $\mathcal{D}_\mu$. 

Basil Saeed, Snigdha Panigrahi and Caroline Uhler
Presented By: Vinay Verma
Causal Structure Discovery from Distributions
Aug 21, 2020 20 / 30
Corollary

If the distribution $p_\mu$ is faithful with respect to a mixture DAG whose component MAGs satisfy the poset compatibility assumption, then FCI outputs the Markov equivalence class of the corresponding union MAG $\mathcal{M}_\cup$.

Proposition

A bidirected edge $u \leftrightarrow v$ in the union graph $\mathcal{M}_\cup$ implies that $u \in V \setminus V_{\text{INV}}$. Additionally, this implies that $p^{(j)}(x_u | x_{\text{pa}_\mathcal{D}(j)}(u)) \neq p^{(i)}(x_u | x_{\text{pa}_\mathcal{D}(i)}(u))$. 
Experiment and Result

Synthetic Data:
The synthetic data are generated as follows:

- $K$ component DAGs each with $|V| = 10$ nodes.
- The expected degree for each DAG is $d = 1.5/K$ so that the nodes in the $\mathcal{M}_\cup$ have expected degree less than 1.5.
- From these DAGs, the corresponding MAGs $\mathcal{M}^{(j)}$ were computed using Algorithm-1.
- If the MAGs were not compatible with the same poset, the DAGs were discarded to ensure poset-compatibility (2 out of 270 graphs were discarded).
Algorithm 1: Construction of the marginal ancestral graph

**Input:** DAG $\mathcal{D} = (V \cup \{y\}, E)$, where $y$ has in-degree 0.

**Output:** the marginal ancestral graph of $\mathcal{D}$ w.r.t. $y$.

1. Initialize $D = \emptyset$, $B = \emptyset$
2. For $u, v \in \text{ch}_\mathcal{D}(y)$: add $u \leftrightarrow v$ to $B$.
3. For $t, u, v$ such that $(t \rightarrow u) \in E$ and $(u \leftrightarrow v) \in B$:
   - if $u \in \text{an}_\mathcal{D}(v)$, then add $t \rightarrow v$ to $D$.
4. For $u, v$ such that $u \leftrightarrow v \in B$: if $u \in \text{an}_\mathcal{D}(v)$, then
   - remove $u \leftrightarrow v$ from $B$ and add $u \rightarrow v$ to $D$.
5. Return the ancestral graph $\mathcal{M} = (V, D, B)$. 

On the synthetic data run FCI algorithm with threshold $\alpha$. The output is $\hat{P}_U$ representing the Markov equivalence class of the union graph.

Compute the true union graph $M_U$ based on the MAGs $M^{(j)}$, generated $n$ samples from this graph and ran FCI on these samples to obtain an estimate $\tilde{P}_U$ on union graph.

The difference between $\hat{P}_U$ and $\tilde{P}_U$ was measured via a normalized structural Hamming distance.
Result: Learning the Union MAG

$|V| = 10, n = 5000, K = 4$

Average Normalized SHD

$\alpha$

0.0001 0.0002 0.0005 0.001 0.002 0.005 0.01 0.02 0.05 0.1
To evaluate the invariant proposition, $V \setminus V_{INV}$ are calculated by determining all nodes incident to bidirected edges in $\hat{P}_U$ estimated using FCI.

This set was compared to the ground truth.
Under mixture-faithfulness, $X_{V \setminus V_{\text{INV}}}$ represents the set of nodes whose conditionals vary across the component DAGs. This motivates using the nodes $X_{V \setminus V_{\text{INV}}}$ and their descendents as features for clustering.
Real Data: Ovarian Cancer

- Applied on gene-expression data from ovarian cancer ($K=2$ with 93 and 168 observation)

Figure 6: (a) shows the output of FCI on genes in the apoptosis pathway using mixture data without knowledge of the cluster membership for each sample, while (b) shows the difference graph of on the same genes learned when cluster membership of each sample is known (Wang et al., 2018 (NeurIPS)).
Conclusions

- The paper propose mixture of graph (Mixture-DAG) to model the causal graph defined over same set of variable and shows that it is realizable and satisfy the Markov property.

- It introduce the union graph and shows that, under a faithfulness and partial ordering assumption on the DAGs in the mixture, the FCI algorithm applied to data from a mixture of DAGs outputs the union graph.

- It shows that the union graph can be used to identify variables whose conditional distribution across the component DAGs changes.

- Paper demonstrate the implication of their result for identifying critical nodes and for clustering samples.
Thank You