ETH zürich

Constraint-based causal structure learning

Causality

Christina Heinze-Deml

Spring 2019

Announcements

- Series 5 will be uploaded later today
- No class on May 29

Last week

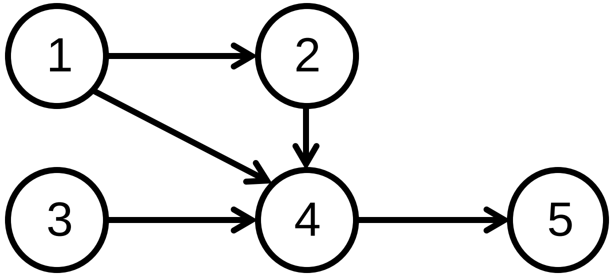
- Estimation
- Markov properties
- Causal minimality

Today

- Faithfulness
- Markov equivalence
- CPDAGs
- SGS algorithm
- PC algorithm

Graph terminology

- A triple (i, j, k) in a DAG G is called a v-structure (or immorality or unshielded collider) if
  - $i \to j \leftarrow k$ in G, and
  - i and k are not adjacent in G.



- A triple (i, j, k) in a DAG G is called a an unshielded triple if
  - i and j are adjacent in G, and
  - j and k are adjacent in G, and
  - i and k are not adjacent in G.

Global Markov property

- Given a DAG $G = (\boldsymbol{V}, E)$, a distribution P with density p on $X_{\boldsymbol{V}}$ is said to satisfy:
  - The global Markov property wrt G if for all pairwise disjoint subsets $\boldsymbol{A}, \boldsymbol{B}$ and $\boldsymbol{S}$ of $\boldsymbol{V}$:

$$\boldsymbol{A} \text{ and } \boldsymbol{B} \text{ are d-separated by } \boldsymbol{S} \text{ in } G \Rightarrow X_{\boldsymbol{A}} \perp\!\!\!\perp X_{\boldsymbol{B}} | X_{\boldsymbol{S}} \text{ in } P$$

Faithfulness and perfect maps

- Given a DAG $G = (\boldsymbol{V}, \boldsymbol{E})$, a distribution P on $X_{\boldsymbol{V}}$ is said to be faithful with respect to G if for all pairwise disjoint subsets $\boldsymbol{A}, \boldsymbol{B}$ and $\boldsymbol{S}$ of $\boldsymbol{V}$:

$$X_{\boldsymbol{A}} \perp\!\!\!\perp X_{\boldsymbol{B}} | X_{\boldsymbol{S}} \text{ in } P \Rightarrow \boldsymbol{A} \text{ and } \boldsymbol{B} \text{ are d-separated by } \boldsymbol{S} \text{ in } G$$

- If a distribution P is Markov and faithful with respect to a DAG G, then G is said to be a perfect map of P. In this case, we have that for all pairwise disjoint subsets $\boldsymbol{A}, \boldsymbol{B}$ and $\boldsymbol{S}$ of $\boldsymbol{V}$:

$$X_{\boldsymbol{A}} \perp\!\!\!\perp X_{\boldsymbol{B}} | X_{\boldsymbol{S}} \text{ in } P \Leftrightarrow \boldsymbol{A} \text{ and } \boldsymbol{B} \text{ are d-separated by } \boldsymbol{S} \text{ in } G$$

Faithfulness and perfect maps

- If a distribution P is Markov and faithful with respect to a DAG G, then G is said to be a perfect map of P. In this case, we have that for all pairwise disjoint subsets $\boldsymbol{A}, \boldsymbol{B}$ and $\boldsymbol{S}$ of $\boldsymbol{V}$:

$$X_{\boldsymbol{A}} \perp\!\!\!\perp X_{\boldsymbol{B}} | X_{\boldsymbol{S}} \text{ in } P \Leftrightarrow \boldsymbol{A} \text{ and } \boldsymbol{B} \text{ are d-separated by } \boldsymbol{S} \text{ in } G$$

- Combination of the Markov and faithfulness assumptions creates one-to-one link between d-separation in the DAG and conditional independence in P
- This will turn out to be very convenient for structure learning
- Not every distribution has a perfect map

Examples

Clicker question – Faithfulness

- Let (X_1, X_2, X_3) with joint distribution P be generated from the following SEM with DAG G

$$
\begin{aligned}
X_1 &\leftarrow \epsilon_1 \\
X_2 &\leftarrow 2X_1 + \epsilon_2 \\
X_3 &\leftarrow 4X_1 - 2X_2 + \epsilon_3
\end{aligned}
$$

with $\epsilon_1, \epsilon_2, \epsilon_3$ jointly independent.

Perfect maps and Markov equivalence

- Example: $X \to Y$, $X \leftarrow Y$ imply the same d-separation: $X \perp Y$

- Definition: Two DAGs G_1 and G_2 are Markov equivalent if they describe the same set of d-separation relationships, i.e., for all pairwise disjoint subsets $\boldsymbol{A}, \boldsymbol{B}$ and $\boldsymbol{S}$ of $\boldsymbol{V}$, we have:

  $\boldsymbol{A}$ and $\boldsymbol{B}$ are d-separated by $\boldsymbol{S}$ in $G_1 \Leftrightarrow \boldsymbol{A}$ and $\boldsymbol{B}$ are d-separated by $\boldsymbol{S}$ in G_2

- A perfect map (if it exists) is unique up to Markov equivalence

Constraint-based structure learning

- Problem:
  - Suppose a distribution P is generated from a SEM with DAG G
  - We only get to see the distribution P – can we learn the DAG G?

- P is Markov wrt G
- If we also assume that P is faithful wrt G, then G is a perfect map of P
  - d-separation relationships in G correspond exactly to conditional independencies in P
- Note that we also assume “no hidden variables” (causal sufficiency)

- Main idea: given all conditional independence relationships in the observational distribution, we should be able to infer things about G

Example

Markov equivalence

- In general, we cannot identify G from observational data
- But we can identify the Markov equivalence class of G
- Example:

		$X_1 \perp\!\!\!\perp X_3$	$X_1 \perp\!\!\!\perp X_3 \mid X_2$	
no v-structure	$1 \rightarrow 2 \rightarrow 3$	false	true	Markov equivalent
	$1 \leftarrow 2 \leftarrow 3$	false	true	
	$1 \leftarrow 2 \rightarrow 3$	false	true	
v-structure	$1 \rightarrow 2 \leftarrow 3$	true	false	Markov equivalent

Markov equivalence and CPDAGs

- Theorem (Verma & Pearl, 1990): All DAGs in a Markov equivalence class have the same skeleton and the same v-structures.

- A Markov equivalence class of DAGs can be uniquely represented by a Completed Partially Directed Acyclic Graph (CPDAG):
  - $i \rightarrow j$ iff $i \rightarrow j$ in all DAGs in the Markov equivalence class (direct causal effect)
  - $i - j$ iff there is a DAG in the Markov equivalence class with $i \rightarrow j$ and one with $i \leftarrow j$ (unidentifiable orientations)

CPDAGs

Example 1

CPDAGs:

$1 - 2 - 3$

$(1 \circ\!\!-\!\!\circ 2 \circ\!\!-\!\!\circ 3)$

$1 \rightarrow 2 \leftarrow 3$

DAG

$1 \rightarrow 2 \rightarrow 3$

$1 \leftarrow 2 \rightarrow 3$

$1 \leftarrow 2 \leftarrow 3$

$1 \rightarrow 2 \leftarrow 3$

CPDAGs

Example 2

Determine CPDAG of

$$\begin{tikzcd} & 2 \arrow[dr] & & \\ 1 \arrow[ur] \arrow[dr] & & 4 \arrow[r] & 5 \\ & 3 \arrow[ur] & & \end{tikzcd}$$

(1) MEC

$$\begin{tikzcd} & 2 \arrow[dr] & & \\ 1 \arrow[ur] \arrow[dr] & & 4 \arrow[r] & 5 \\ & 3 \arrow[ur] & & \end{tikzcd}$$

$$\begin{tikzcd} & 2 \arrow[dr] & & \\ 1 \arrow[ur] & & 4 \arrow[r] & 5 \\ & 3 \arrow[ul] \arrow[ur] & & \end{tikzcd}$$

$$\begin{tikzcd} & 2 \arrow[dl] \arrow[dr] & & \\ 1 \arrow[dr] & & 4 \arrow[r] & 5 \\ & 3 \arrow[ur] & & \end{tikzcd}$$

would introduce a new v-structure ←

$$\begin{tikzcd} & 2 \arrow[dl] \arrow[dr] & & \\ 1 & & 4 \arrow[r] & 5 \\ & 3 \arrow[ul] \arrow[ur] & & \end{tikzcd}$$

(2) CPDAG

$$\begin{tikzcd} & 2 \arrow[dl, dash] \arrow[dr] & & \\ 1 & & 4 \arrow[r] & 5 \\ & 3 \arrow[ul, dash] \arrow[ur] & & \end{tikzcd}$$

Constraint-based structure learning

- Let $G = (\boldsymbol{V}, \boldsymbol{E})$ be DAG. Let $i, j \in V$ such that $i \neq j$. Then the following hold:
  - If i and j are adjacent in G, they cannot be d-separated by any subset of the remaining nodes
  - If i and j are not adjacent in G, then they are d-separated by $\mathrm{pa}(i)$ or by $\mathrm{pa}(j)$

- Hence, i and j are adjacent if and only if they cannot be d-separated by any subset of the remaining nodes
- Moreover, if they can be d-separated by some subset of the remaining nodes, they can be d-separated by $\mathrm{pa}(i)$ or $\mathrm{pa}(j)$

Verma and Pearl (1991), Lemma 1

SGS algorithm

- Assuming Markov and faithfulness, a CPDAG can be estimated by the SGS-algorithm of Spirtes, Glymour and Scheines:
  - Determine the skeleton
  - Determine the v-structures
  - Direct as many of the remaining edges as possible

SGS algorithm

- Assuming Markov and faithfulness, a CPDAG can be estimated by the SGS-algorithm of Spirtes, Glymour and Scheines:
  - Determine the skeleton
    - No edge between i and j
      $\Leftrightarrow$
      i and j are d-separated by some subset $\boldsymbol{S}$ of the remaining nodes
      $\Leftrightarrow$
      $X_i \perp\!\!\!\perp X_j | X_{\boldsymbol{S}}$ for some subset $\boldsymbol{S}$ of the remaining nodes
    - Start with the complete graph
    - For all pairs $i \neq j$ assess conditional independence of X_i and X_j given $X_{\boldsymbol{S}}$ for all subsets $\boldsymbol{S}$ of the remaining nodes and remove an edge if a conditional independence is found.
  - …

SGS algorithm

- Assuming Markov and faithfulness, a CPDAG can be estimated by the SGS-algorithm of Spirtes, Glymour and Scheines:
  - Determine the skeleton
  - Determine the v-structures
    - By checking for conditional dependence
  - Direct as many of the remaining edges as possible
    - By consistency with already directed edges

Example

PC algorithm

- Assuming Markov and faithfulness, a CPDAG can be estimated by the PC-algorithm of Peter Spirtes and Clark Glymour:
  - Determine the skeleton
    - No edge between i and j
      $\Leftrightarrow$
      i and j are d-separated by $\mathrm{pa}(i, G)$ or $\mathrm{pa}(j, G)$
      $\Leftrightarrow$
      i and j are d-separated by a subset $\boldsymbol{S}$' of $\mathrm{adj}(i, G)$ or of $\mathrm{adj}(j, G)$
      $\Leftrightarrow$
      $X_i \perp\!\!\!\perp X_j | X_{\boldsymbol{S}'}$ for a subset $\boldsymbol{S}'$ of $\mathrm{adj}(i, G)$ or of $\mathrm{adj}(j, G)$
    - Start with the complete graph
    - For $k = 0, 1, \dots, p-2$
      - Consider all pairs of adjacent vertices (i, j), and remove edge if X_i and X_j are conditionally independent given some subset of size k of $\mathrm{adj}(i, G)$ or of $\mathrm{adj}(j, G)$

PC algorithm

- Assuming Markov and faithfulness, a CPDAG can be estimated by the PC-algorithm of Peter Spirtes and Clark Glymour:
  - Determine the skeleton
  - Determine the v-structures
    - By checking for conditional dependence
  - Direct as many of the remaining edges as possible
    - By consistency with already directed edges

PC algorithm – sample version

- Instead of a conditional independence “oracle”, we perform conditional independence tests
- In the multivariate Gaussian setting, this is equivalent to testing for zero partial correlation: $H_0 : \rho_{ij|\mathbf{S}} = 0$ versus $H_A : \rho_{ij|\mathbf{S}} \neq 0$

- The significance level α serves as a tuning parameter for the PC algorithm
  - Do not necessarily want to treat type I error as in traditional testing

Partial correlation

- We call X and Y uncorrelated if

$$\rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = 0$$

- We say that X and Y are partially uncorrelated given Z if

$$\rho_{X,Y|Z} = \frac{\rho_{X,Y} - \rho_{X,Z}\rho_{Z,Y}}{\sqrt{(1-\rho_{X,Z}^2)(1-\rho_{Z,Y}^2)}} = 0$$

- $\rho_{X,Y|Z}$ equals correlation between residuals after linearly regressing X on Z and Y on Z

Partial correlation

- In general,

$$\rho_{X,Y|Z} \nRightarrow X \perp\!\!\!\perp Y|Z \qquad \text{and} \qquad \rho_{X,Y|Z} \nLeftarrow X \perp\!\!\!\perp Y|Z$$

  - See Elements of Causal Inference, Example 7.9

PC algorithm

- Assume that P has a perfect map G. Verify the following statements about the oracle version of the PC algorithm:
  - If an edge $i - j$ is removed at some point in the PC algorithm, then i and j are not adjacent in G.
  - At any point in the algorithm, the current skeleton is a supergraph of the skeleton of G.
  - If an edge $i - j$ is not removed in the PC algorithm, then i and j are adjacent in G.
  - The output of the skeleton phase of the PC algorithm is the skeleton of G.
  - [See Series 5.]

Recap

- Concepts to know:
  - Faithfulness
  - Markov equivalence
  - CPDAGs
  - SGS algorithm
  - PC algorithm

References and acknowledgments

- Slides adapted from M. Maathuis
- Markov properties, faithfulness and causal minimality
  - Peters, Janzing and Schölkopf (2017). Elements of Causal Inference. Chapter 6.5
- Constraint-based structure learning
  - Shalizi (2019). Chapter 24.
  - Peters, Janzing and Schölkopf (2017). Elements of Causal Inference. Chapter 7.2.1