Causal inference and control

Week 5:

Causal discovery

Søren Wengel Mogensen
Department of Automatic Control
soren.wengel_mogensen@control.lth.se

Recap

Structural causal model

Definition (Structural causal model)

We consider a set of ‘endogenous’ variables $\{X_1, \dots, X_d\}$ and a set of independent noise variables $\{N_1, \dots, N_d\}$. For each $j = 1, \dots, d$, $PA_j \subseteq \{X_1, \dots, X_d\} \setminus \{X_j\}$. A *structural causal model*, $\mathcal{C} = (S, P_N)$, consists of a distribution P_N over the noise variables and a set, S, of acyclic structural assignments,

$$X_j = f_j(PA_j, N_j), j = 1, \dots, d.$$

Graph of a structural causal model

We define a graph, (V, E), from an SCM with $V = \{X_1, X_2, \dots, X_d\}$. For each j, we include an edge $X_i \to X_j$ for each $X_i \in PA_j$.

We assume throughout that the graph of an SCM is a DAG.

Interventions

Definition (Interventional distribution)

Let $C = (S, P_N)$ be SCM and let $\tilde{S}$ be a set of structural assignments indexed by $I \subseteq \{1, \dots, d\}$,

$$X_k = \tilde{f}_k(\tilde{PA}_k, \tilde{N}_k), k \in I.$$

We define an SCM by using the corresponding structural assignment in $\tilde{S}$ if $j \in I$, and otherwise the structural assignment in S, and require that the resulting set of assignments is acyclic. This gives a new SCM, and we say that its distribution is the *interventional* distribution defined by $\tilde{S}$, denoted by $P^{do(X_k = \tilde{f}_k(\tilde{PA}_k, \tilde{N}_k), k \in I)} = P^{do(X_k)}$. We assume that the corresponding density exists and denote this by $p^{do(X_k)}$.

The book uses the notation $P^{C; do(X_k = \tilde{f}_k(\tilde{PA}_k, \tilde{N}_k), k \in I)}$, but we omit the causal model in the notation. The *observational distribution*, P_{X}, corresponds to $\tilde{S} = \emptyset$ (no intervention). The noise terms are jointly independent, also in the interventional SCM.

SCM, 'procedural' explanation

We assume that the structural causal models are acyclic. This means that there is an *order* of the variables such that for each j and $k > j$, X_k is not a parent of X_j.

$$
\begin{aligned}
X_1 &= f_1(PA_1, N_1) = \bar{f}_1(N_1) \\
X_2 &= f_2(PA_2, N_2) = \bar{f}_2(X_1, N_2) \\
X_3 &= f_3(PA_3, N_3) = \bar{f}_3(X_1, X_2, N_3) \\
&\dots \\
X_{d-1} &= f_{d-1}(PA_{d-1}, N_{d-1}) = \bar{f}_{d-1}(X_1, X_2, \dots, X_{d-2}, N_{d-1}) \\
X_d &= f_d(PA_d, N_d) = \bar{f}_d(X_1, X_2, \dots, X_{d-1}, N_d)
\end{aligned}
$$

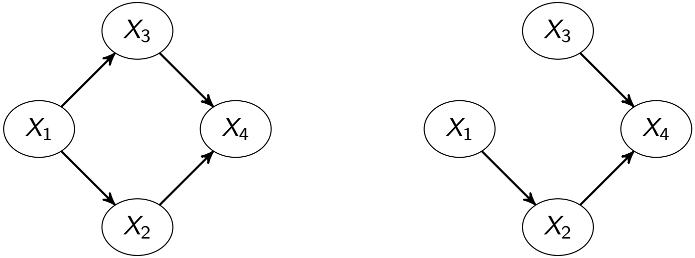
Interventions, 'procedural' explanation

If we have an (atomic) intervention $do(X_j = a)$, we obtain,

$$
\begin{aligned}
X_1 &= f_1(PA_1, N_1) = \bar{f}_1(N_1) \\
X_2 &= f_2(PA_2, N_2) = \bar{f}_2(X_1, N_2) \\
X_3 &= f_3(PA_3, N_3) = \bar{f}_3(X_1, X_2, N_3) \\
&\dots \\
X_j &= a \\
&\dots \\
X_{d-1} &= f_{d-1}(PA_{d-1}, N_{d-1}) = \bar{f}_{d-1}(X_1, X_2, \dots, X_{d-2}, N_{d-1}) \\
X_d &= f_d(PA_d, N_d) = \bar{f}_d(X_1, X_2, \dots, X_{d-1}, N_d)
\end{aligned}
$$

Interventions, graph

An interventional SCM corresponding to an atomic intervention, $do(X_k = a)$, is represented by a graph where all edges 'into' X_j are removed.



$do(X_3 = \tilde{N}_3)$

Adjustment

Definition (Valid adjustment set)

We consider a structural causal model on nodes $\boldsymbol{V}$. Let $X, Y \in \boldsymbol{V}$ such that $Y \notin \boldsymbol{PA}_X$. We say that $\boldsymbol{Z} \subseteq \boldsymbol{V} \setminus \{X, Y\}$ is a *valid adjustment set* for (X, Y) if

$$p^{do(x)}(y) = \sum_{z} p(y|x, \boldsymbol{z})p(\boldsymbol{z}).$$

A valid adjustment set allows a simple type of identification.

Valid adjustment sets

Proposition (ECI, Proposition 6.41)

Assume that $Y \notin \boldsymbol{PA}_X$. The following sets are valid adjustment sets for (X, Y).

- $\boldsymbol{Z} = \boldsymbol{PA}_X$.
- $\boldsymbol{Z} \subseteq \boldsymbol{V} \setminus \{X, Y\}$ *such that* $\boldsymbol{Z}$ *contains no descendants of* X *and blocks all back-door paths between* X *and* Y.
- *no member of* $\boldsymbol{Z}$ *is a descendant of any* $W \in \boldsymbol{V} \setminus \{X\}$ *which lies on a directed path from* X *to* Y, *and* $\boldsymbol{Z}$ *blocks all nondirected paths between* X *and* Y.

We say that W is a *descendant* of X if there exists a directed path $X \to \ldots \to W$. We say that a path between X and Y is a *back-door path* if $X \leftarrow \ldots Y$. We say that $\boldsymbol{Z} \subseteq \boldsymbol{X} \setminus \{X, Y\}$ *blocks* a path between X and Y if the path is not d-connecting given $\boldsymbol{Z}$.

Do-calculus

The *do-calculus* is a set of three rules that connect interventional and observational distributions. Say we have a causal graph $\mathcal{G}$ and disjoint node sets X, Y, Z, W.

Do-calculus is *complete* in the sense that all identifiable interventional distributions can be computed by repeatedly applying the three rules of do-calculus [Huang and Valtorta, 2006, Shpitser and Pearl, 2006].

Causal discovery/graphical structure learning

Adjustment and do-calculus allow us to identify causal effects (interventional distributions) from observational data and a *known* graph.

Causal discovery/graphical structure learning comprise methods for learning (about) the causal graph from data.

Causal discovery/graphical structure learning

There is a wealth of research in this subfield. Some research exploits restrictions on the SCM like

- additive noise, $X_j = f_j(\boldsymbol{PA}_j) + N_j$,
- linear/nonlinear functions,
- Gaussian/non-Gaussian noise.

Some research considers combinations of observational and interventional data.

We will start from a completely 'nonparametric' point of view and consider only the observed conditional independences.

Markov equivalence of DAGs

We say that two DAGs, $\mathcal{G}_1 = (V, E_1)$ and $\mathcal{G}_2 = (V, E_2)$, are *Markov equivalent* if they agree on all *d*-separations. That is, if for all $i, j \in V$ and $C \subseteq V \setminus \{i, j\}$ we have

$$i \perp\!\!\!\perp_{\mathcal{G}_1} j \mid C \Leftrightarrow i \perp\!\!\!\perp_{\mathcal{G}_2} j \mid C.$$

In the Week 1 exercises, we proved EIC Lemma 6.25.

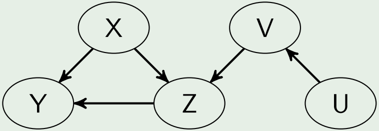
Theorem (Markov equivalence of DAGs, EIC Lemma 6.25)

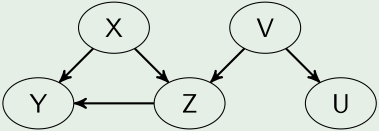
DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent if and only if they have the same skeleton and the same set of unshielded colliders.

The *skeleton* is the undirected graph obtained by replacing every directed edge in the DAG with an undirected edge. Three nodes (i, j, and k) are an *unshielded collider*, or *v-structure*, if $i \to k \leftarrow j$ and there is no edge between i and j.

Markov equivalence of DAGs

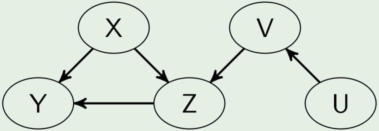
Example

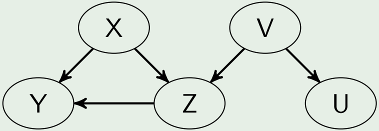




Markov equivalence of DAGs

Example

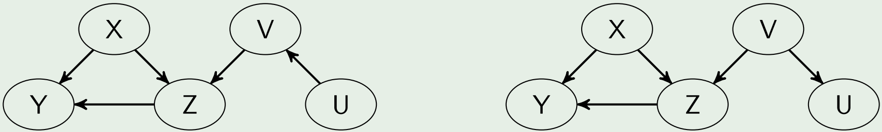




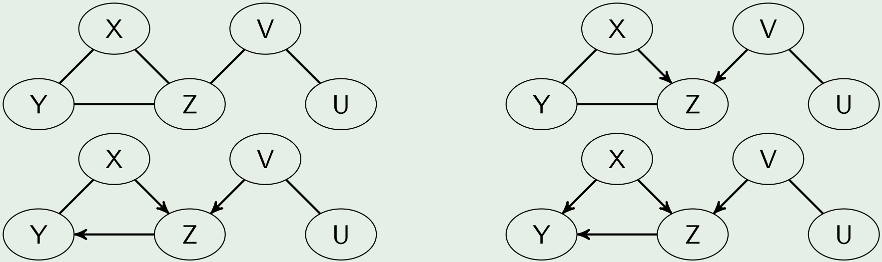
Let us argue that these constitute an equivalence class:

Markov equivalence of DAGs

Example



Let us argue that these constitute an equivalence class:



Markov equivalence of DAGs

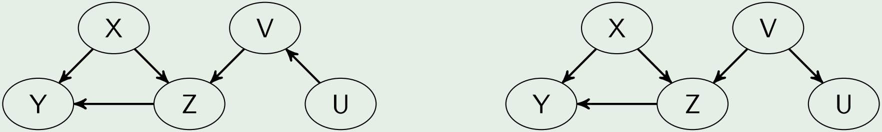
Markov equivalence defines an equivalence relation on the set of DAGs with node set V. Let $\mathcal{G} = (V, E)$ be a DAG. We say that

$$\{\tilde{\mathcal{G}} = (V, \tilde{E}) : \tilde{\mathcal{G}} \text{ and } \mathcal{G} \text{ are Markov equivalent}\}$$

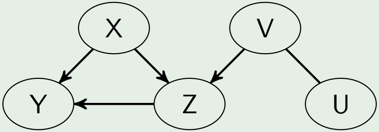
is the *Markov equivalence class* of $\mathcal{G}$. It is useful to have a graphical representation of an entire Markov equivalence class. For this purpose, we define the *completed partially directed acyclic graph* (CPDAG) on nodes V by including $i \to j$ in the CPDAG if $i \to j$ in every graph in the Markov equivalence class and $i - j$ if $i \to j$ in some graph in the equivalence class and $i \leftarrow j$ in another.

Markov equivalence of DAGs

Example



Equivalence class from before, and its CPDAG.



Markov equivalence of DAGs

Lemma (Meek [1995])

Let $\mathcal{C}$ be a CPDAG. If $i \to k - j$ in $\mathcal{C}$, then $i \to j$ in $\mathcal{C}$.

Markov equivalence of DAGs

Lemma (Meek [1995])

Let $\mathcal{C}$ be a CPDAG. If $i \to k - j$ in $\mathcal{C}$, then $i \to j$ in $\mathcal{C}$.

Theorem (Markov equivalence of DAGs, EIC Lemma 6.25)

DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent if and only if they have the same skeleton and the same set of unshielded colliders.

The *skeleton* is the undirected graph obtained by replacing every directed edge in the DAG with an undirected edge. Three nodes (i, j, and k) are an *unshielded collider*, or *v-structure*, if $i \to k \leftarrow j$ and there is no edge between i and j.

Faithfulness

Definition (Faithfulness, ECI Definition 6.33)

The distribution P_X is *faithful* to the DAG $\mathcal{G} = (\boldsymbol{V}, E)$ if

$$\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{C} \Rightarrow \boldsymbol{A} \perp\!\!\!\perp_{\mathcal{G}} \boldsymbol{B} \mid \boldsymbol{C}$$

for all disjoint $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C} \subseteq \boldsymbol{V}$, that is, conditional independence implies d-separation.

Faithfulness

Definition (Faithfulness, ECI Definition 6.33)

The distribution P_X is *faithful* to the DAG $\mathcal{G} = (\boldsymbol{V}, E)$ if

$$\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{C} \Rightarrow \boldsymbol{A} \perp\!\!\!\perp_{\mathcal{G}} \boldsymbol{B} \mid \boldsymbol{C}$$

for all disjoint $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C} \subseteq \boldsymbol{V}$, that is, conditional independence implies d-separation.

Under faithfulness (and the global Markov property) the set of d-separations is exactly the same as the set of conditional independences. This means that the distribution identifies the Markov equivalence class of the underlying DAG.

Independence-based methods/constraint-based methods

Assuming faithfulness, one can test conditional independences in the observed data and output a CPDAG which corresponds to the observed conditional independences.

We first assume access to an *independence oracle*, a magical device that tells us if a conditional independence holds or not in the observational distribution.

Again, the *skeleton* of a DAG is the undirected graph with the same adjacencies.

Independence-based methods/constraint-based methods

Lemma (ECI Lemma 7.8)

(i) Nodes X and Y in a DAG are adjacent if and only if they cannot be d-separated by any subset of $\boldsymbol{V} \setminus \{X, Y\}$.
(ii) If X and Y are not adjacent then they are d-separated by $\boldsymbol{PA}_X$ or by $\boldsymbol{PA}_Y$.

PC algorithm Spirtes et al. [2000]

[whiteboard]

PC algorithm Spirtes et al. [2000]

[whiteboard]

Orientation of edges using Meek's rules Meek [1995] which are complete.

PC algorithm Spirtes et al. [2000]

The PC algorithm is order-dependent and it does not necessarily output a CPDAG.

There is a number of adaptations of this basic algorithm.

Score-based methods

We may instead search for a graph which allows a good fit to data, that is,

$$\hat{\mathcal{G}} = \arg\max_{\mathcal{G}} S(\mathcal{D}, \mathcal{G})$$

for data $\mathcal{D}$ and a scoring function S. For instance, the BIC (assuming a parametrization, θ)

$$S(\mathcal{D}, \mathcal{G}) = \log p(\mathcal{D} \mid \hat{\theta}, \mathcal{G}) - n_p (\log n)/2$$

where $\hat{\theta}$ is the maximum-likelihood estimator, n_p is the number of parameters, and n is the number of data points. In most cases, the search space is so vast that heuristic/greedy search is needed.

Known causal ordering

The search space is really vast, even for node sets of moderate size (d).

d	Number of DAGs on d nodes
4	543
5	29281
6	3781503
7	1138779265
8	783702329343
9	1213442454842881

Note that if the causal ordering (topological order) is *known*, then we can use standard variable selection methods to decide which arguments f_j depends on,

$$X_j = f_j(\boldsymbol{X}_{i<j}, N_j). \tag{1}$$

Graphical marginalization

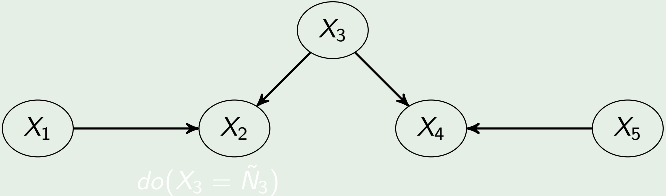
In causal modeling, the idea of *hidden* variables is central. In Week 4, we looked at identification methods that do not require full observation.

In causal discovery, we may be interested in learning a 'marginal' of the causal graph when there are hidden variables, $\boldsymbol{H}$, as well as observed variables, $\boldsymbol{O}$. One natural requirement is that the implied conditional independences are the same when restricting to $\boldsymbol{O}$.

Graphical marginalization

Example (DAGs are not closed under marginalization, Richardson and Spirtes [2002])

Assume X_3 is unobserved. There is no DAG on nodes $\{X_1, X_2, X_4, X_5\}$ that encodes the same conditional independences (using d-separation).



Acyclic directed mixed graphs

We say that a graph is a *directed acyclic mixed graph* (ADMG) if every edge is either *directed*, $\rightarrow$, or *bidirected*, $\leftrightarrow$.

The extension of *d*-separation to ADMGs is known as *m-separation*.

Latent projection

Let $\mathcal{G} = (\boldsymbol{V}, E)$ be an ADMG, $\boldsymbol{V} = \boldsymbol{O} \cup \boldsymbol{H}$. We define the following transformation.

Definition (Latent projection)

We define $m(\mathcal{G}, \boldsymbol{O})$ as the graph such that for $X, Y \in \boldsymbol{O}$

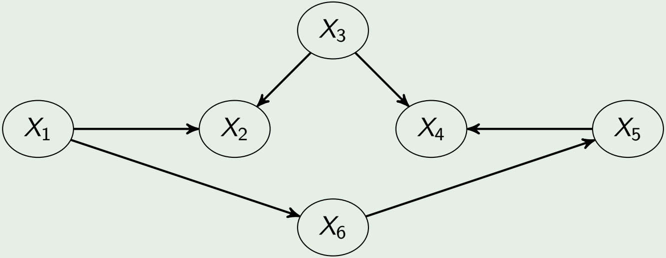
- $X \rightarrow Y$ in $m(\mathcal{G}, \boldsymbol{O})$ if there is a directed path $X \rightarrow \ldots \rightarrow Y$ in $\mathcal{G}$ such that every non-endpoint node is in $\boldsymbol{H}$,
- $X \leftrightarrow Y$ in $m(\mathcal{G}, \boldsymbol{O})$ if there is a path between X and Y such that all non-endpoint nodes are in $\boldsymbol{H}$, all non-endpoint nodes are non-colliders, and there are arrowheads at both X and Y.

The latent projection is also an ADMG!

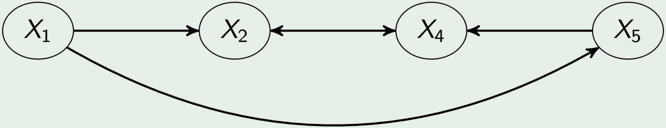
Latent projection

Example

Let $\boldsymbol{V} = \{X_1, X_2, X_3, X_4, X_5, X_6\}$ and $\boldsymbol{O} = \{X_1, X_2, X_4, X_5\}$



$do(X_3 = \tilde{N}_3)$



Latent projection as a marginal

Let $\mathcal{G} = (\boldsymbol{O} \cup \boldsymbol{H}, E)$ be a DAG and let $\mathcal{M} = m(\mathcal{G}, \boldsymbol{O})$.

Proposition

Let $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C} \subseteq \boldsymbol{O}$. We have that $\boldsymbol{A}$ and $\boldsymbol{B}$ are d-separated by $\boldsymbol{C}$ in $\mathcal{G}$ if and only if $\boldsymbol{A}$ and $\boldsymbol{B}$ are m-separated by $\boldsymbol{C}$ in $m(\mathcal{G}, \boldsymbol{O})$.

References I

Yimin Huang and Marco Valtorta. Pearl's calculus of intervention is complete. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, 2006.

Christopher Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 1995.

Thomas S. Richardson and Peter Spirtes. Ancestral graph markov models. *The Annals of Statistics*, 30(4):962–1030, 2002.

Ilya Shpitser and Judea Pearl. Identification of conditional interventional distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, 2006.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2000.