TECHNISCHE UNIVERSITÄT
BERGAKADEMIE FREIBERG
Die Ressourcenuniversität. Seit 1765.

Kevin Bitterlich
Faculty of Mathematics and Computer Science
Institute of Stochastics

Counterfactuals, Markov Property, Faithfulness and Causal Minimality

Research seminar on causality, May 29th 2024

Overview

1. Counterfactuals
2. Markov Property
3. Faithfulness and Causal Minimality

1. Counterfactuals

Definition (Counterfactuals)

Consider SCM $\mathfrak{C} := (\mathbf{S}, P_{\mathbf{N}})$ over nodes $\mathbf{X}$. Given some observations $\mathbf{x}$, we define a counterfactual SCM by replacing the distribution of noise variables:

$$\mathfrak{C}_{\mathbf{X}=\mathbf{x}} := (\mathbf{S}, P_{\mathbf{N}}^{\mathfrak{C}|\mathbf{X}=\mathbf{x}}), \quad P_{\mathbf{N}}^{\mathfrak{C}|\mathbf{X}=\mathbf{x}} := P_{\mathbf{N}|\mathbf{X}=\mathbf{x}}$$

The new set of noise variables need not to be jointly independent anymore. Counterfactual statements can now be seen as do-statements in the new counterfactual SCM.

- We restrict counterfactuals to the discrete case, that is, when the noise distribution has a probability mass function.
- The definition can be generalized such that we observe not the full vector $X = x$ but only some of the variables.
- Counterfactual statements depend strongly on the structure of the SCM

Example : Consider the following SCM:

$$X := N_X$$
$$Y := X^2 + N_Y$$
$$Z := 2 \cdot Y + X + N_Z$$

with $N_X, N_Y, N_Z \sim U(\{-5, -4, \ldots, 4, 5\})$ iid. Now, assume that we observe $(X, Y, Z) = (1, 2, 4)$.

Then $P_{\mathbf{N}}^{\mathfrak{C}|\mathbf{X}=\mathbf{x}}$ puts a point mass on $(N_X, N_Y, N_Z) = (1, 1, -1)$ because here all noise terms can be uniquely reconstructed from the observations.

We therefore have the counterfactual statement (in the context of $(X, Y, Z) = (1, 2, 4)$): "Z would have been 11 had X been (set to) 2." Mathematically, this means that $P_Z^{\mathfrak{C}|\mathbf{X}=\mathbf{x};do(X:=2)}$ has a point mass on 11.
In the same way, we obtain "Y would have been 5, had X been 2," and "Z would have been 10, had Y been 5."

Example : Let $N_1, N_2 \sim$ Ber(0.5) and $N_3 \sim$ U$(\{0, 1, 2\})$, such that the three variables are jointly independent. We define two different SCMs.

$\mathfrak{C}_A$:

$$
\begin{aligned}
X_1 &:= N_1 \\
X_2 &:= N_2 \\
X_3 &:= (\mathbb{1}_{N_3>0} \cdot X_1 + \mathbb{1}_{N_3=0} \cdot X_2) \cdot \mathbb{1}_{X_1 \neq X_2} + N_3 \cdot \mathbb{1}_{X_1=X_2}
\end{aligned}
$$

$\mathfrak{C}_B$:

$$
\begin{aligned}
X_1 &:= N_1 \\
X_2 &:= N_2 \\
X_3 &:= (\mathbb{1}_{N_3>0} \cdot X_1 + \mathbb{1}_{N_3=0} \cdot X_2) \cdot \mathbb{1}_{X_1 \neq X_2} + (2 - N_3) \cdot \mathbb{1}_{X_1=X_2}
\end{aligned}
$$

Both SCMs induce the same graph and entail the same observational distribution as well as the same intervention distributions (for any possible intervention). **But** the two models differ in a counterfactual statement.

Suppose, we have an observation $(X_1, X_2, X_3) = (1, 0, 0)$ and we are interested in the counterfactual question: What would X_3 have been if X_1 had been 0? Then $\mathfrak{C}_A$ and $\mathfrak{C}_B$ predict different values for X_3 (0 and 2, resp.).

Remark:

1. Counterfactual statements are not transitive. Consider first example of this talk. Given observation $(X, Y, Z) = (1, 2, 4)$:

   "Y would have been 5, had X been 2",
   "Z would have been 10, had Y been 5",

   But

   "Z would have not been 10, had X been 2".

2. Humans often think in counterfactuals: "Do you remember our flight to New York on September 11, 2000? Imagine if we would have taken the flight one year later!"

2. Markov Property

Definition (Markov property)

Given a DAG $\mathcal{G}$ and a joint distribution P_X, this distribution is said to satisfy

(i) the global Markov property with respect to the DAG $\mathcal{G}$ if

$$\forall \text{ disjoint vertex sets } \mathbf{A}, \mathbf{B}, \mathbf{C} : \quad \mathbf{A} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{B} | \mathbf{C} \implies \mathbf{A} \perp\!\!\!\perp \mathbf{B} | \mathbf{C}$$

(ii) the local Markov property with respect to the DAG $\mathcal{G}$ if each variable is independent of its non-descendants (without the parents of the variable) given the parents of the variable

(iii) the Markov factorization property with respect to the DAG $\mathcal{G}$ if

$$p(x) = p(x_1, \ldots, x_d) = \prod_{j=1}^{d} p(x_j | \mathrm{pa}_j^{\mathcal{G}})$$

For this, we have to assume that P_X has a density p.

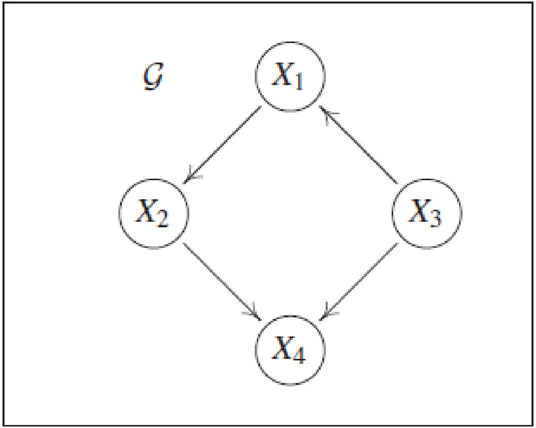
Theorem (Equivalence of Markov properties)

If P_X has a density p, then all Markov properties in the definition above are equivalent.

Example : A distribution $P_{X1,X2,X3,X4}$ is Markovian with respect to the following graph $\mathcal{G}$

$$X_1 := f_1(X_3, N_1)$$
$$X_2 := f_2(X_1, N_2)$$
$$X_3 := f_3(N_3)$$
$$X_4 := f_4(X_2, X_3, N_4)$$

- $N_1, \ldots, N_4$ jointly independent
- $\mathcal{G}$ is acyclic



if, according to (i) or (ii),

$$X_2 \perp\!\!\!\perp X_3 | X_1 \quad \text{and} \quad X_4 \perp\!\!\!\perp X_1 | X_2, X_3$$

or, according to (iii),

$$p(x_1, x_2, x_3, x_4) = p(x_3)p(x_1|x_3)p(x_2|x_1)p(x_4|x_2, x_3).$$

The Markov condition relates statements about graph separation to conditional independences.
We will now see,in which case different graphs encode the exact same set of conditional independences.

Definition (Markov equivalence of graphs)

We denote by $\mathcal{M}(\mathcal{G})$ the set of distributions that are Markovian with respect to $\mathcal{G}$:

$$\mathcal{M}(\mathcal{G}) := \{P : P \text{ satisfies the global (or local) Markov property with respect to } \mathcal{G}\}.$$

Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent if $\mathcal{M}(\mathcal{G}_1) = \mathcal{M}(\mathcal{G}_2)$. This is the case if and only if $\mathcal{G}_1$ and $\mathcal{G}_2$ satisfy the same set of d-separations.

The set of all DAGs that are Markov equivalent to some DAG is called Markov equivalence class of $\mathcal{G}$. It can be represented by a completed PDAG that is denoted by CPDAG$(\mathcal{G}) = (\mathbf{V}, \mathcal{E})$.

Definition

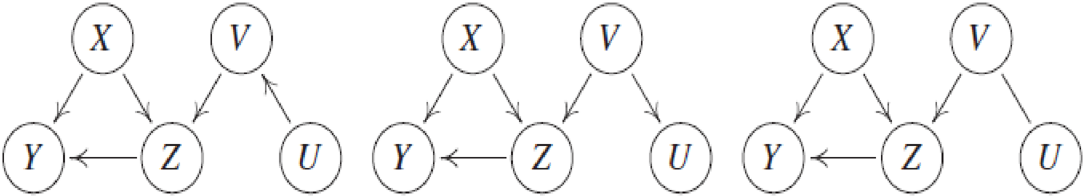
Let $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ be a graph with nodes $\mathbf{V}$ and edges $\mathcal{E} \subset \mathbf{V}^2$ with $(v, v) \notin \mathcal{E}$ for any $v \in \mathbf{V}$.

- Three nodes are called an immorality or a v-structure if one node is a child of the two others that themselves are not adjacent.
- The skeleton of $\mathcal{G}$ does not take the directions of the edges into account. It is the graph $(\mathbf{V}, \tilde{\mathcal{E}})$ with $(i, j) \in \tilde{\mathcal{E}}$, if $(i, j) \in \mathcal{E}$ or $(j, i) \in \mathcal{E}$.

Lemma (Markov equivalence of graphs)

Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent if and only if they have the same skeleton and the same immoralities.

Example of two Markov equivalent graphs (left and middle) and corresponding CPDAG (right):



Definition (Markov blanket)

Consider a DAG $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ and a target node Y. The Markov blanket of Y is the smallest set M such that

$$Y \perp\!\!\!\perp_{\mathcal{G}} \mathbf{V} \setminus (\{Y\} \cup M) \text{ given } M.$$

If P_X is Markovian with respect to $\mathcal{G}$, then

$$Y \perp\!\!\!\perp \mathbf{V} \setminus (\{Y\} \cup M) \text{ given } M.$$

Proposition (Markov blanket)

Consider a DAG $\mathcal{G}$ and a target node Y. Then, the Markov blanket M of Y includes its parents, its children, and the parents of its children

$$M = \mathbf{PA}_Y \cup \mathbf{CH}_Y \cup \mathbf{PA}_{\mathbf{CH}_Y}$$

Example : Consider the follwing graph

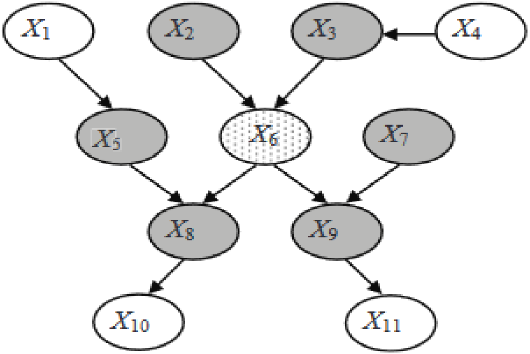


Figure 1: Visweswaran, Cooper, *Learning Instance-Specific Predictive Models*, JMLR, 2010

$Y = X_6$, $\mathbf{PA}_Y = \{X_2, X_3\}$, $\mathbf{CH}_Y = \{X_8, X_9\}$ $\mathbf{PA}_{\mathbf{CH}_Y} = \{X_5, X_7\}$

$\implies M = \{X_2, X_3, X_5, X_7, X_8, X_9\}$

Recall **Reichenbach's common cause principle**: When X and Y are dependent, there must be a "causal explanation" for this dependence:

(i) X is causing Y, or

(ii) Y is causing X, or

(iii) there is a (possibly unobserved) common cause Z that causes both X and Y.

But, we have no further specified the meaning of the word "causing". In the following proposition we use a weak notion of "causing", namely the existence of a directed path.

Proposition (Reichenbach's common caus principle)

Assume that any pair of variables X and Y can be embedded into a larger system in the following sense. There exists a correct SCM over the collection $\mathbf{X}$ of random variables that contains X and Y with graph $\mathcal{G}$.

If X and Y are (unconditionally) dependent, then there is

(i) either a directed path from X to Y, or

(ii) from Y to X, or

(iii) there is a node Z with a directed path from Z to X and from Z to Y.

Berkson's paradox : "Why are handsome men such jerks?" (Ellenberg example).

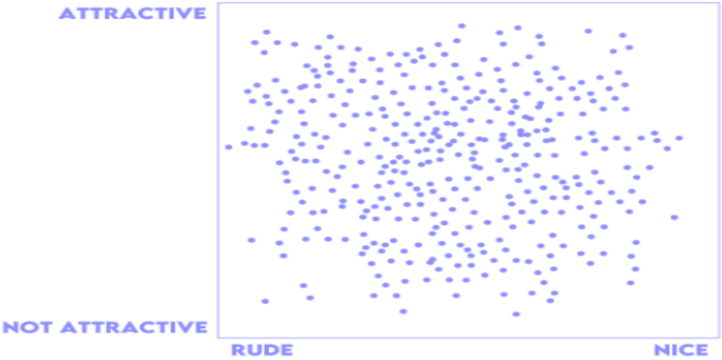


Figure 2: linkedin.com

Berkson's paradox : "Why are handsome men such jerks?" (Ellenberg example).

Figure 3: linkedin.com

Proposition (SCMs imply Markov property)

Assume that $P_{\mathbf{X}}$ is induced by an SCM with graph $\mathcal{G}$. Then, $P_{\mathbf{X}}$ is Markovian with respect to $\mathcal{G}$.

- The assumption that a distribution is Markovian w.r.t. the causal graph is sometimes called the causal Markov condition. For us, causal graphs are induced by the underlying SCM.
- For defining intervention distributions, it usually suffices to have knowledge of the observational distribution and the graph structure (next talk).

Therefore, we define a causal graphical model as a pair that consists of a graph and an observational distribution s.t. the distribution is Markovian w.r.t. the graph (causal Markov condition).

Definition (Causal graphical model)

A causal graphical model over random variables $\mathbf{X} = (X_1, \ldots, X_d)$ contains a graph $\mathcal{G}$ and a collection of functions $f_j(x_j, x_{\mathbf{PA}_j^{\mathcal{G}}})$ that integrate to 1:

$$\int f_j(x_j, x_{\mathbf{PA}_j^{\mathcal{G}}})\, dx_j = 1.$$

These functions induce a distribution $P_{\mathbf{X}}$ over $\mathbf{X}$ via

$$p(x) = p(x_1, \ldots, x_d) = \prod_{j=1}^{d} f_j(x_j, x_{\mathbf{PA}_j^{\mathcal{G}}})$$

and thus play the role of conditionals: $f_j(x_j, x_{\mathbf{PA}_j^{\mathcal{G}}}) = p(x_j | x_{\mathbf{PA}_j^{\mathcal{G}}})$.

If a distribution $P_{\mathbf{X}}$ over $\mathbf{X}$ is Markovian with respect to a graph $\mathcal{G}$ and allows for a strictly positive, continuous denisty p, the pair $(\mathcal{G}, P_{\mathbf{X}})$ defines a causal graphical model by $f_j(x_j, x_{\mathbf{PA}_j^{\mathcal{G}}}) := p(x_j | x_{\mathbf{PA}_j^{\mathcal{G}}})$.

Why primarily work with SCMs and not just with causal graphical models? Because SCMs contain strictly more information than their corresponding graph and law (e.g. counterfactual statements).

3. Faithfulness and Causal Minimality

Definition (Faithfulness and causal minimality)

Consider a distribution $P_{\mathbf{X}}$ and a DAG $\mathcal{G}$.

(i) $P_{\mathbf{X}}$ is faithful to the DAG $\mathcal{G}$ if

$$\forall \text{ disjoint vertex sets } \mathbf{A}, \mathbf{B}, \mathbf{C} : \quad \mathbf{A} \perp\!\!\!\perp \mathbf{B} | \mathbf{C} \implies \mathbf{A} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{B} | \mathbf{C}$$

(ii) A distribution satisfies causal minimality w.r.t. $\mathcal{G}$ if it is Markovian w.r.t. $\mathcal{G}$, but not to any proper subgraph of $\mathcal{G}$.

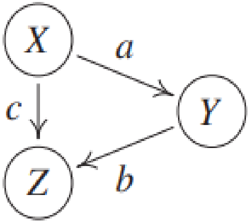
Part (i) posits an implication that is the opposite of the global Markov condition

$$\mathbf{A} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{B} | \mathbf{C} \implies \mathbf{A} \perp\!\!\!\perp \mathbf{B} | \mathbf{C}$$

There might be a distribution that is Markovian but not faithful w.r.t. a given DAG (see next example).

Violation of faithfulness : Consider the follwing figure.

$$\mathcal{G}_1: \quad X := N_X, \quad Y := aX + N_Y, \quad Z := cX + bY + N_z,$$



$\mathcal{G}_1$

with $N_X \sim \mathcal{N}(0, \sigma_X^2)$, $N_Y \sim \mathcal{N}(0, \sigma_Y^2)$ and $N_Z \sim \mathcal{N}(0, \sigma_Z^2)$ jointly independent. Now if

$$a \cdot b + c = 0,$$

the distribution is not faithful with respect to $\mathcal{G}_1$ since we obtain $X \perp\!\!\!\perp Z$, but $X \not\perp\!\!\!\perp_{\mathcal{G}} Z \mid \emptyset$.

In general, causal minimality is weaker than faithfulness.

Proposition (Faithfulness implies causal minimality)

If $P_{\mathbf{X}}$ is faithful and Markovian w.r.t. $\mathcal{G}$, then causal minimality is satisfied.

We can also find a statement with equvialence for causal minimality. This is the case, if there is no node, that is conditionally independent of any of its parents, given the remaining parents.

Proposition (Equivalence of causal minimality)

Consider $\mathbf{X} = (X_1, \ldots, X_d)$ and assume that the joint distribution has a density w.r.t. a product measure. Suppose, $P_{\mathbf{X}}$ is Markovian w.r.t. $\mathcal{G}$. Then: $P_{\mathbf{X}}$ satisfies causal minimality w.r.t. $\mathcal{G}$ if and only if

$$\forall X_j \forall Y \in \mathbf{PA}_j^{\mathcal{G}} : \quad X_j \not\perp\!\!\!\perp Y | \mathbf{PA}_j^{\mathcal{G}} \setminus \{Y\}.$$