Causal inference with hidden variables

Stefan Franssen

Leiden University College

8 April 2019

Overview

- Recap and notation

Overview

- Recap and notation
- Causal substructures

Overview

- Recap and notation
- Causal substructures
- Simpsons paradox

Overview

- Recap and notation
- Causal substructures
- Simpsons paradox
- Instrumental variables

Overview

- Recap and notation
- Causal substructures
- Simpsons paradox
- Instrumental variables
- Graphical representations in case of unobserved causes

Overview

- Recap and notation
- Causal substructures
- Simpsons paradox
- Instrumental variables
- Graphical representations in case of unobserved causes
- Learning the graphical structure

Overview

- Recap and notation
- Causal substructures
- Simpsons paradox
- Instrumental variables
- Graphical representations in case of unobserved causes
- Learning the graphical structure
- Other constraints

Structural causal models: Notation

Definition

A structural causal model is a structure $(\Omega, F, \mathbb{P}, \mathcal{C}, G, \mathcal{N}, \mathcal{F})$ *where*

- $(\Omega, F, \mathbb{P})$ *is a probability space.*

Structural causal models: Notation

Definition

A structural causal model is a structure $(\Omega, F, \mathbb{P}, \mathcal{C}, G, \mathcal{N}, \mathcal{F})$ *where*

- $(\Omega, F, \mathbb{P})$ *is a probability space.*
- $\mathcal{C}$ *is a collection of random variables, not necessarily taking values in the same spaces.*

Structural causal models: Notation

Definition

A structural causal model is a structure $(\Omega, F, \mathbb{P}, \mathcal{C}, G, \mathcal{N}, \mathcal{F})$ *where*

- $(\Omega, F, \mathbb{P})$ *is a probability space.*
- $\mathcal{C}$ *is a collection of random variables, not necessarily taking values in the same spaces.*
- G *is a DAG with vertex set* $\mathcal{C}$ *and edges* E*.*

Structural causal models: Notation

Definition

A structural causal model is a structure $(\Omega, F, \mathbb{P}, \mathcal{C}, G, \mathcal{N}, \mathcal{F})$ *where*

- $(\Omega, F, \mathbb{P})$ *is a probability space.*
- $\mathcal{C}$ *is a collection of random variables, not necessarily taking values in the same spaces.*
- G *is a DAG with vertex set* $\mathcal{C}$ *and edges* E*.*
- $\mathcal{N}$ *is a set of independent random variables.*

Structural causal models: Notation

Definition

A structural causal model is a structure $(\Omega, F, \mathbb{P}, \mathcal{C}, G, \mathcal{N}, \mathcal{F})$ *where*

- $(\Omega, F, \mathbb{P})$ *is a probability space.*
- $\mathcal{C}$ *is a collection of random variables, not necessarily taking values in the same spaces.*
- G *is a DAG with vertex set* $\mathcal{C}$ *and edges* E*.*
- $\mathcal{N}$ *is a set of independent random variables.*
- $\mathcal{F}$ *is a collection of functions such that for all* $X \in \mathcal{C}$ *there exists an* $f_X \in \mathcal{F}$ *and* $N_X \in \mathcal{N}$ *such that*

$$X = f_X(pa(X), N_X)$$

Causal Substructures

Definition

Let $\mathbf{SCM} = (\Omega, F, \mathbb{P}, \mathcal{C}, G, \mathcal{N}, \mathcal{F})$ *be a structural causal model. Let* $\mathcal{C}' \subset \mathcal{C}$*. We call* $\mathbf{SCM}' = (\Omega, F, \mathbb{P}, \mathcal{C}', G, \mathcal{N}', \mathcal{F}')$ *a structural causal substructure of* **SCM** *if* $\mathbf{SCM}'$ *is a structural causal structure.*

Causally Sufficient substructures

With some abuse of definition we call a node without any parents a root.

Definition

We call a causal substructure SCM' of SCM causally sufficient if all collections of ancestors A_j in SCM of roots X_j in SCM' are disjoint.

This means that there is no confounding going on.

Interventional sufficiency

Definition

We call a causal substructure SCM' of SCM interventionally sufficient if it cannot be falsified.

Here we use falsification in the sense that the substructure generates all the correct interventions.

Causal sufficiency implies interventional sufficiency

Theorem

Let SCM' be a causal substructure of SCM, then it is also an interventional substructure.

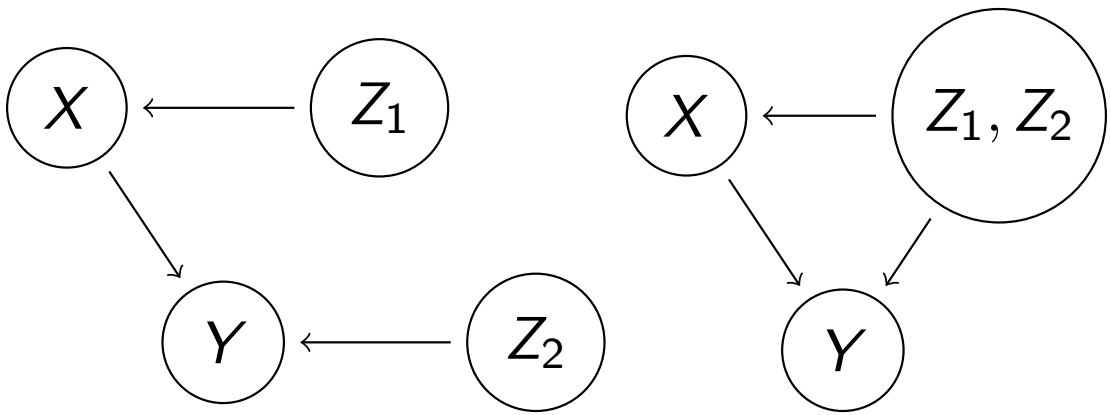
Interventional sufficiency does not imply causal sufficiency

Theorem

The reverse implication does not hold.

Interventional sufficiency does not imply causal sufficiency

If we have three random variables X, Y, Z, where $Z = (Z_1, Z_2)$ such that Z_1 is independent of Z_2. Then we can construct counterexamples since we cannot interventionally distinguish the following two structural causal models.



Simpson's paradox

Simpson's paradox follows from model misspecification. In the next example you see that one treatment has better results in every situation, but on average it performs worse.

	Overal	Small stones	large stones
Treatment A	78% (273/350)	93% (81/87)	73% (192/263)
Treatment B	83% (289/350)	87% (234/270)	69% (55/80)

In this situation it follows from Treatment A and treatment B not being applied in the same frequency over both situations, where treatment A is applied more often in the harder situation.

Simpson's paradox to the extreme

Proposition

There exists $X, Y, Z_1, Z_2, \cdots$ *random variables taking values in* $\{0, 1\}$ *such that for all* n *and* $z_1, z_2, \cdots$

$$\begin{aligned}&\mathbb{P}(Y = 1 | X = 1, Z_1 = z_1, Z_2 = z_2, \cdots, Z_{2n} = z_{2n}) \\ &> \mathbb{P}(Y = 1 | X = 0, Z_1 = z_1, Z_2 = z_2, \cdots, Z_{2n} = z_{2n})\end{aligned}$$

But

$$\begin{aligned}&\mathbb{P}(Y = 1 | X = 1, Z_1 = z_1, Z_2 = z_2, \cdots, Z_{2n+1} = z_{2n+1}) \\ &> \mathbb{P}(Y = 1 | X = 0, Z_1 = z_1, Z_2 = z_2, \cdots, Z_{2n+1} = z_{2n+1})\end{aligned}$$

So our causal estimate if setting $X = 1$ increases or decreases the chance of $Y = 1$ changes with every random variable we add.

Simpson's paradox to the extreme: proof sketch

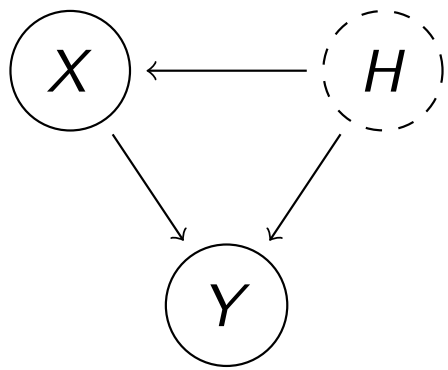
The idea is repeat the following procedure inductively:

- We specify the conditional probability of $Y = 1$ in such a way to satisfy the requirements
- We specify the conditional probability of $Z_n = z_n$ given the rest is such a way to make the probabilities add up

We have enough degrees of freedom to do this.

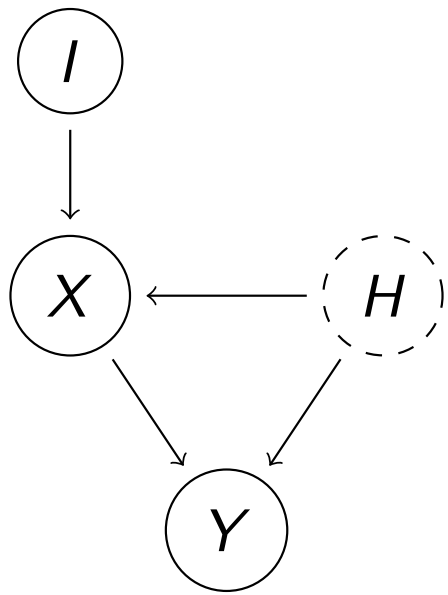
Instrumental variables

In case of missing confounders, we might be able to find instruments which are independent of the noise terms and which cause the explanatory variable of interest.



Instrumental variables

In case of missing confounders, we might be able to find instruments which are independent of the noise terms and which cause the explanatory variable of interest.



Instrumental variables: Identification

Theorem

Let $\mathcal{F}$ be a class of measureable functions. Let X, ϵ, Z be random variables. Let $f_0 \in \mathcal{F}$. Then, in the model $Y = f_0(X) + \epsilon$ with $\mathbb{E}[\epsilon|Z] = 0$, we can identify f_0 if and only if $f = f_0$ is the only solution in $\mathcal{F}$ of

$$\mathbb{E}[f_0(X) - f(X)|Z] = 0 \qquad a.s.$$

Instrumental variables: Identification example

Suppose

$$
\begin{aligned}
H &:= N_H \\
Z &:= N_Z \\
X &:= f(Z) + g(H) + N_X \\
Y &:= aX + j(H) + N_Y
\end{aligned}
$$

and suppose we observe the joint distribution over Z, X and Y. Then we can identify a using two stage least squares and nonparametric regression of X on Z.

Instrumental variables: estimation

The paper on nonparametric instrumental variables by Newey in 2013 he provides a series estimation by penalization method. In this paper, Newey also talks a bit about instrument strength.

Graphical representations: Problems with DAGs

If we do not restrict to the number of hidden variables the search space is to big to search over:

- Infinite candidates
- No longer an exponential family

Marginalized DAGs

We could instead try to marginalize over the missing variables, however, we run into the following two problems

- Representing the set of conditional independences can lead to causal misinterpretation.
- The set of distributions whose pattern of independences correspond to the d-seperation statements in a DAG is not closed under marginalization.

m-seperation

Definition

Suppose G is an ancestral graph. For given source and target nodes s, t and a set Z of nodes in $G \setminus \{s, t\}$. Consider a path from s to t. An intermediate node on the path is called a collider if both edges on the path touching it are directed toward the node. The path is said to m-connect the nodes s and t, given Z, iff

- *every non-collider on the path is outside Z,*

m-seperation

Definition

Suppose G is an ancestral graph. For given source and target nodes s, t and a set Z of nodes in $G \setminus \{s, t\}$. Consider a path from s to t. An intermediate node on the path is called a collider if both edges on the path touching it are directed toward the node. The path is said to m-connect the nodes s and t, given Z, iff

- *every non-collider on the path is outside Z,*
- *for each collider c on the path, either c is in Z or there is a directe path from c to an element of Z.*

m-seperation

Definition

Suppose G is an ancestral graph. For given source and target nodes s, t and a set Z of nodes in $G \setminus \{s, t\}$. Consider a path from s to t. An intermediate node on the path is called a collider if both edges on the path touching it are directed toward the node. The path is said to m-connect the nodes s and t, given Z, iff

- *every non-collider on the path is outside Z,*
- *for each collider c on the path, either c is in Z or there is a directe path from c to an element of Z.*

Definition

We call s and t m-seperated relative to Z if it is not m connected relative to Z for any path.

Alternative: MAG

Definition

A Maximal Ancestral Graph is the smallest superclass of DAGs that is closed under marginalization.

A MAG is a mixed graph and can contain direct, bidirected and undirected edges. We also need to replace d-separation by m-separation.

Alternative: PAG

Definition

A Partially ancestral graph is an equivalance class of MAGs representing the same set of m-separations.

Alternative: IPG and POIPG

As an alternative induced path graphs and partially oriented induced path graphs have been used. These were initially used to represent the output of the fast causal inference algorithm.

Alternative: ADMG

Yet another alternative is to start with the original DAG containing hidden variables and then apply a latent projection as defined by Pearl and Verma, 1991 and Pearl, 2009.

Alternative: ADMG

Yet another alternative is to start with the original DAG containing hidden variables and then apply a latent projection as defined by Pearl and Verma, 1991 and Pearl, 2009. This takes a graph with both observed and unobserved hidden variables and constructs a new graph over the observed variables.

Alternative: ADMG

Yet another alternative is to start with the original DAG containing hidden variables and then apply a latent projection as defined by Pearl and Verma, 1991 and Pearl, 2009. This takes a graph with both observed and unobserved hidden variables and constructs a new graph over the observed variables. The resulting structure is called an acyclic directed mixed graph. The separation notion is again *m*-separation.

Learning the graphical structure

It is possible to modify the PC algorithm to output a PAG, however, this algorithm does not always output the correct PAG.

Fast causal inference

There has been work on providing an algorithm which provides a correct graphical object. The resulting algorithm was FCI. It has been proven that a modification of FCI gives the maximally informative output.

Fast causal inference +

The FCI algorithm has been replaced by the FCI+ algorithm which is both complete and quick.

Constraints beyond conditional independence

Models with hidden variables can lead to additional constrains beyond conditional independence. The book shows a couple of examples but again, does not show details or derivations and just refers to the literature.