SCM and hidden variables

Seminar on causality and causal inference

Anna Chekhanova

Institute of Stochastics, TU Bergakademie Freiberg

July 3, 2024

Overview

1. Interventional Sufficiency
2. Simpson's paradox
3. Instrumental variables
4. Conditional independences and Graphical representations
5. References

Interventional Sufficiency

A set of variables **X** is usually said to be **causally sufficient** if there is no hidden common cause $C \notin \mathbf{X}$ that is causing more than one variable in **X**.

We propose a small modification of causal sufficiency that we call *interventional sufficiency*, a concept that is based on falsifiability of SCMs (Section 6.8, Konstantin Ibadullaev).

Definition (Interventional sufficiency)

We call a set **X** of variables **interventionally sufficient** if there exists an SCM over **X** that cannot be falsified as an interventional model.

That is, it induces observational and intervention distributions that coincide with what we observe in practice.

Sometimes, it can be possible to compute the correct intervention distributions even in the presence of latent confounding.

Example 1. (Interventionally sufficiency and causally insufficiency) Consider the following SCM

$$
\begin{aligned}
Z &:= N_Z \\
X &:= \mathbf{1}_{Z \geq 2} + N_X \\
Y &:= Z \bmod 2 + X + N_Y
\end{aligned}
$$

with $N_Z \sim \mathcal{U}(\{0, 1, 2, 3\})$ and $N_X, N_Y \sim \mathcal{N}(0, 1)$.

Interventional Sufficiency

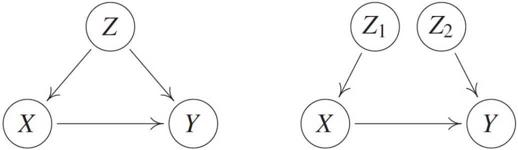


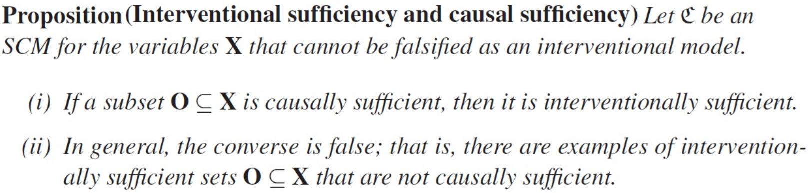
Figure 1: Both graphs represent interventionally equivalent SCMs for the model described in Example 1. While only the second representation renders X and Y causally sufficient, X and Y are interventionally sufficient independently of the representation.

Obvious the variables X and Y *causally insufficient* while the variables X and Y are *interventionally sufficient* and the reason is that the „confounder“Z consists of **two independent** parts:

1. $Z_1 := \mathbf{1}_{Z \geq 2}$ is the first bit of the binary representation of Z, and
2. $Z_2 := Z\text{mod}2$ is the second bit.

In this sense, the „confounder“can be separated into the independent variables Z_1 and Z_2, with Z_1 influencing X, and Z_2 influencing Y. (Fig.1)

In general we have:



Proposition (Interventional sufficiency and causal sufficiency) *Let $\mathfrak{C}$ be an SCM for the variables $\mathbf{X}$ that cannot be falsified as an interventional model.*

(i) If a subset $\mathbf{O} \subseteq \mathbf{X}$ is causally sufficient, then it is interventionally sufficient.

(ii) In general, the converse is false; that is, there are examples of interventionally sufficient sets $\mathbf{O} \subseteq \mathbf{X}$ that are not causally sufficient.

For many SCMs with a structure similar to Figure 1 (left), X and Y are interventionally insufficient.

Remark 1. (omitting an „intermediate“variable preserves interventional sufficiency)

We have the following three statements:

(i) Assume that there is an SCM over X, Y, Z with graph $X \to Y \to Z$ and $X \not\perp\!\!\!\perp Z$ that induces the correct interventions. Then X and Z are *interventionally sufficient* due to the SCM over X, Z satisfying $X \to Z$.

(ii) Assume that there is an SCM $\mathfrak{C}$ over X, Y, Z that induces the correct interventions with graph $X \to Y \to Z$ and additional $X \to Z$ and assume further that $P^{\mathfrak{C}}_{X,Y,Z}$ is faithful with respect to this graph; see also (iii). Then, again, X and Z are *interventionally sufficient* due to the SCM over X, Z satisfying $X \to Z$.

(iii) If the situation is the same as in (ii) with the difference that

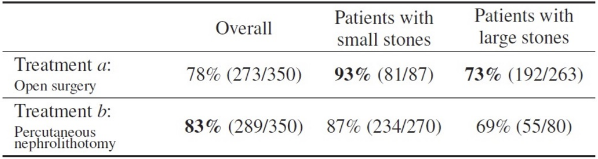
$$P^{\mathfrak{C}}_{Z|X=x} = P^{\mathfrak{C};do(X:=x)}_{Z} = P^{\mathfrak{C}}_{Z}$$

for all x (in particular, $P^{\mathfrak{C}}_{X,Y,Z}$ is not faithful with respect to the graph). Then, X and Z are *interventionally sufficient* due to the SCM over X, Z with the empty graph. Note that the counterfactuals may not be represented correctly.

Simpson's paradox

An SCM over the two observed variables that ignores confounding does not only entail the wrong intervention distributions, it can even reverse the sign of the causal effect: a treatment can look beneficial although it is harmful.

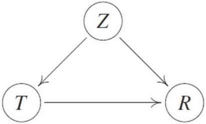
Example 2. (Kidney stone data)



	Overall	Patients with small stones	Patients with large stones
Treatment *a*: Open surgery	78% (273/350)	**93%** (81/87)	**73%** (192/263)
Treatment *b*: Percutaneous nephrolithotomy	**83%** (289/350)	87% (234/270)	69% (55/80)

Table 6.1: A classic example of Simpson's paradox. The table reports the success rates of two treatments for kidney stones [Bottou et al., 2013, Charig et al., 1986, tables I and II]. Although the overall success rate of treatment *b* seems better (any bold number is largest in its column), treatment *b* performs worse than treatment *a* on both patients with small kidney stones and patients with large kidney stones (see Examples 6.37 and Section 9.2).

Assume that the true underlying SCM allows for the graph



Here, Z is the size of the stone, T the treatment, and R the recovery (all binary). We see that the recovery is influenced by the treatment and the size of the stone.

We have

$$P^{\mathfrak{C}}(R=1\,|\,T=A) < P^{\mathfrak{C}}(R=1\,|\,T=B) \qquad \text{but}$$
$$P^{\mathfrak{C};do(T:=A)}(R=1) > P^{\mathfrak{C};do(T:=B)}(R=1);$$

Suppose that we have not measured the variable Z (size of the stone) and furthermore that we do not even know about its existence. We might then hypothesize that $T \to R$ is the correct graph. If we denote this (wrong) SCM by $\widetilde{\mathfrak{C}}$, we can rewrite equation above as

$$P^{\tilde{\mathfrak{C}};do(T:=A)}(R=1) < P^{\tilde{\mathfrak{C}};do(T:=B)}(R=1) \text{ but}$$
$$P^{\mathfrak{C};do(T:=A)}(R=1) > P^{\mathfrak{C};do(T:=B)}(R=1).$$

Due to the model misspecification, the causal statement *gets reversed*. Although A is the more effective drug, we propose to use B.
Summarizing, the *Simpson's paradox* is about warning of how *sensitive causal* reasoning can be with respect to model misspecifications.

Instrumental variables

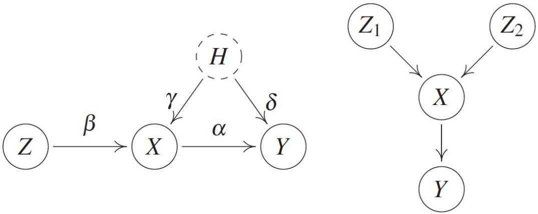


Figure 2: *Left*: setting of an instrumental variable. A famous example is a randomized clinical trial with non-compliance: Z is the treatment assignment, X the treatment and Y the outcome. *Right*: Y-structure.

Consider a linear Gaussian SCM with the graph (Fig. 2, left):

$$Y := \alpha X + \delta H + N_Y,$$

coefficient α in the structural assignment is the quantity of interest. Sometimes it is called the **average causal effect** (ACE). It is not directly accessible, because of the hidden common cause H. Simply regressing Y on X and taking the regression coefficient generally results in a biased estimator for α:

$$\frac{\operatorname{cov}[X,Y]}{\operatorname{var}[X]} = \frac{\alpha \operatorname{var}[X] + \delta\gamma \operatorname{var}[H]}{\operatorname{var}[X]} = \alpha + \frac{\delta\gamma \operatorname{var}[H]}{\operatorname{var}[X]}.$$

Definition

Formally, we call a variable Z in an SCM an **instrumental variable** for (X, Y) if

(a) Z is independent of H;

(b) Z is not independent of X („relevance“); and

(c) Z effects Y only through X („exclusion restriction“).

In Figure 2 (left) all of these assumptions satisfies. Note, however, that other structures do, too. For example, one can allow for a hidden common cause between Z and X. In practice, one usually uses domain knowledge to argue why conditions (a), (b), and (c) hold.

In the linear case, we can exploit the existence of Z in the following way:

(i) because (H, N_X) is *independent* of Z, we can regard $\gamma H + N_X$ in

$$X := \beta Z + \gamma H + N_X$$

as noise. It becomes apparent that we can therefore consistently estimate the coefficient β and therefore have access to βZ (which, in the case of finitely many data, is approximated by fitted values of Z).

(ii) because of

$$Y := \alpha X + \delta H + N_Y = \alpha(\beta Z) + (\alpha\gamma + \delta)H + N_Y$$

we can then consistently estimate α by regressing Y on βZ.

Summarizing, we first regress X on Z and then regress Y on the predicted values $\widehat{\beta}Z$ (predicted from the first regression). The average causal effect α becomes identifiable in the limit of infinite data. This method is commonly referred to as „two-stage least squares."

However, identifiability is not restricted to the linear setting.

Conditional independences and Graphical representations

In causal learning,

- we are trying to reconstruct the causal model from observational data.
- identifiability allow us to identify the graph structure of an SCM over variables $\mathbf{X}$ from the observational distribution $P_{\mathbf{X}}$.

Consider: SCM $\mathfrak{C}$ over variables $\mathbf{X} = (\mathbf{O}, \mathbf{H})$, where $\mathbf{O}$ observed variables and $\mathbf{H}$ are hidden variables.
To find: Is the graph structure of $\mathfrak{C}$ becomes identifiable from the distribution $P_{\mathbf{O}}$ over the observed variables? If so, how can we identify it?

- In the case *without* hidden variables, we discussed in Section 7.2.1 how to learn (parts of) the causal structure under the *Markov condition* and *faithfulness* (Hanyue Gu, Methods for Structure Identifiability).
- For causal learning *with hidden variables*, we search over the space of DAGs with latent variables.

Difficulties, if there are hidden variables:

- unknown the size of **H**; if do not restrict the number of hidden variables, there is an *infinite number of graphical* candidates.
- statistical models that are Markovian and faithful to a DAG with latent variables *do not form* a curved exponential family, which justify using the Bayesian Information Criterion.

Can we represent each DAG with latent variables by a *marginalized graph* (over the observed variables), using more than one type of edge?

- This approach also faces challenges because the marginalized graph must accurately reflect the underlying SCM, including the effects of hidden variables.

More detail: Bongers et al. [2016].

Thus,

- Instead of determining the complete distribution from a specific DAG with latent variables, we can focus on dealing with distributions that satisfy certain conditional independences over the observed variables **O** (implicitly assuming the Markov condition and faithfulness).
- Then we ask how this set of conditional independences can be represented.
- This leads to exploring graphical representations like Directed Acyclic Graphs (DAGs), Maximal Ancestral Graphs (MAGs), and Partial Ancestral Graphs (PAGs).

Representing the conditional independence structure P_O with a DAG has two well-known drawbacks:

(1) Representing the set of conditional independences with a DAG over the observed variables can lead to causal misinterpretations, and

(2) the set of distributions whose pattern of independences correspond to the *d*-separation statements in a DAG is not closed under marginalization

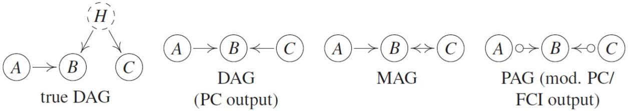


Figure 3: Starting with an SCM on the left-hand side, the three graphs on the right encode the set of conditional independences ($A \perp\!\!\!\perp C$). Due to an erroneous causal interpretation, the DAG is not desirable as an output of a causal learning method. In this example, the IPG and the latent projection (ADMG) are equal to the MAG.

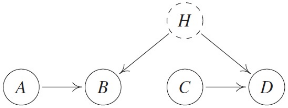
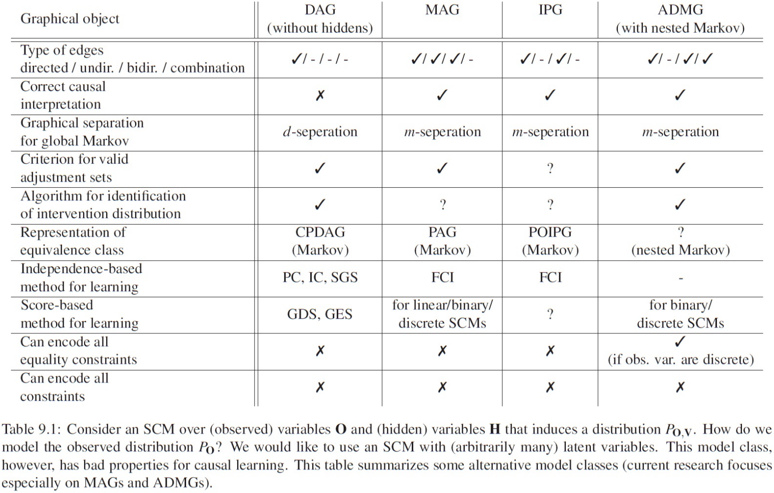


Figure 4: This example is taken from Richardson and Spirtes [2002, Figure 2(i)]. It shows that DAGs are not closed under marginalization. There is no DAG over nodes $\mathbf{O} = \{A, B, C, D\}$ that encodes all conditional independences from the graph including H.

The following table discusses some ideas that suggest new graphical representations (over **O**) in causal inference:



Graphical object	DAG (without hiddens)	MAG	IPG	ADMG (with nested Markov)
Type of edges directed / undir. / bidir. / combination	✓/ - / - / -	✓/✓/✓/ -	✓/ - /✓/ -	✓/ - /✓/✓
Correct causal interpretation	✗	✓	✓	✓
Graphical separation for global Markov	*d*-seperation	*m*-seperation	*m*-seperation	*m*-seperation
Criterion for valid adjustment sets	✓	✓	?	✓
Algorithm for identification of intervention distribution	✓	?	?	✓
Representation of equivalence class	CPDAG (Markov)	PAG (Markov)	POIPG (Markov)	? (nested Markov)
Independence-based method for learning	PC, IC, SGS	FCI	FCI	-
Score-based method for learning	GDS, GES	for linear/binary/ discrete SCMs	?	for binary/ discrete SCMs
Can encode all equality constraints	✗	✗	✗	✓ (if obs. var. are discrete)
Can encode all constraints	✗	✗	✗	✗

Table 9.1: Consider an SCM over (observed) variables **O** and (hidden) variables **H** that induces a distribution $P_{\mathbf{O},\mathbf{V}}$. How do we model the observed distribution $P_{\mathbf{O}}$? We would like to use an SCM with (arbitrarily many) latent variables. This model class, however, has bad properties for causal learning. This table summarizes some alternative model classes (current research focuses especially on MAGs and ADMGs).

DAG – Directed Acyclic Graph without hidden variables; **MAG** – Maximal Ancestral Graph; **IPG** – Induced Path Graph; **ADMG** – Acyclic Directed Mixed Graphs.

References

References

1. Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Elements of causal inference: foundations and learning algorithms. The MIT Press, 2017.

2. Judea Pearl, Madelyn Glymour, Nicholas P. Jewell. Causal inference in statistics. John Wiley & Sons Ltd, 2016.