Causal Identification

The Causal Graph Perspective

Xi Chen

Rotterdam School of Management
Erasmus University Rotterdam

June 14, 2024

Outline

1. The graph terminology
2. Causal graph
3. The do-operator
   - Interventions: the secret sauce
   - Understanding the do $(\cdot)$
4. Understanding do-calculus
5. Causal identification
   - Defining causal identification
   - Four identification strategies
6. Summary

Causation $\neq$ correlation

- Try to explain how causation and correlation are different.
- You know the differences but it is difficult to verbalize.
- We lack a sort of "language" to explain the differences.

Causation $\neq$ correlation

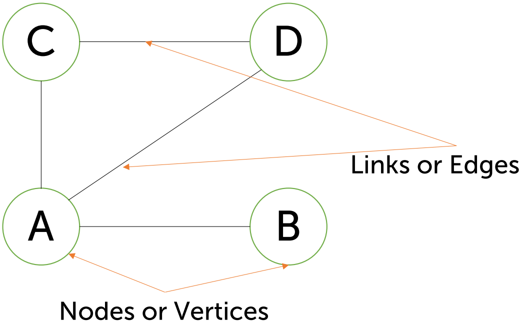
- Try to explain how causation and correlation are different.
- You know the differences but it is difficult to verbalize.
- We lack a sort of “language” to explain the differences.

Causation $\neq$ correlation

- Try to explain how causation and correlation are different.
- You know the differences but it is difficult to verbalize.
- We lack a sort of "language" to explain the differences.

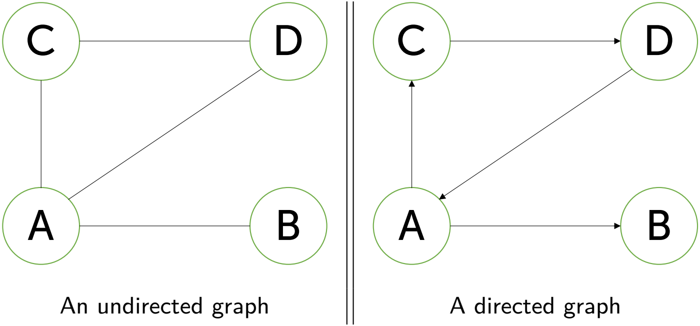
Outline

The atoms of a graph



- **Nodes**: any entity, e.g., people, organizations, charging stations, internet routers
- **Links**: any relationship or connectedness

Directed vs. undirected graphs



- Undirected graph: no directions, e.g., Facebook friends
- Directed graph: with directions, e.g., Twitter followers

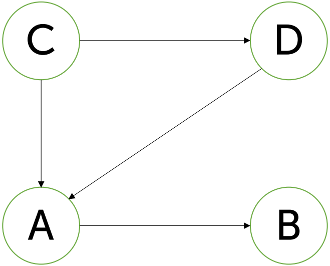
Paths in directed graphs

- Between two nodes X and Y in a directed graph:
  - Adjacent: $X \to Y$ or $Y \to X$
  - Non-adjacent: $X \qquad Y$
- If $X \to Y$, X is a parent of Y and Y is a child of X.
  - If $X \to \cdots \to Y$, X is an ancestor of Y, and Y is a descendant of X.

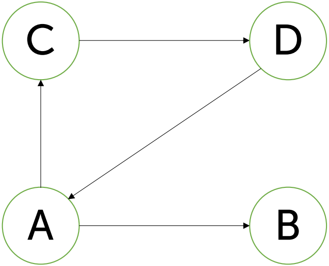
Definition (Paths in directed graphs)

A path in a directed graph is any sequence of adjacent nodes, regardless of the direction of the links that join them.

Cycles in directed graphs



A directed acyclic graph (DAG)



A directed cyclic graph

Definition (Cycles in directed graphs)

A cycle is a particular type of path that starts from a node and links back to the same node, i.e., $X \to \cdots \to X$.

Outline

What is a causal graph?

Simply put it: a causal graph is a directed acyclic graph where

- The nodes represent variables.
- The links represent causal flow.

Definition (Causal Links)

A variable X is said to be the cause of a variable Y, if an ***exogenous*** *change in X can* ***flow*** *on the graph to Y.*

The change in X must be exogenous

It is a parent of X and non-adjacent to Y or other variables in the system.

3 Axioms of causal graphs

3 axioms of causal graphs

For a directed acyclic graph to be a causal graph, we must have:

1. (**Causal Links Assumption or the Existence of Causal Graph**) In a DAG, every parent is a direct cause of all its children.
2. (**Local Markov Assumption**) Given its parents in a DAG, a node is independent of all its non-descendants.
3. (**Faithfulness or No Conspiracy**) The correlation patterns are always implied by the causal Markov property.

Local Markov assumption

The complexity of a causal graph increases exponentially with the no. of variables.

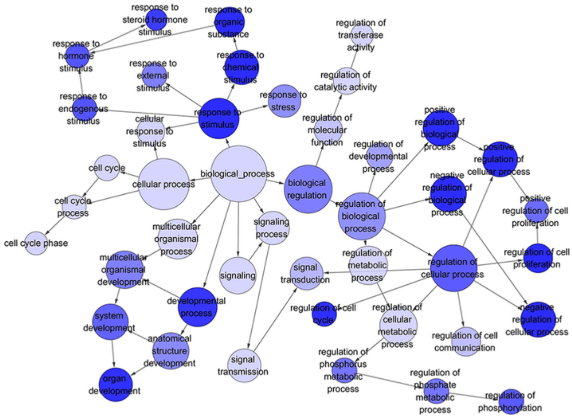
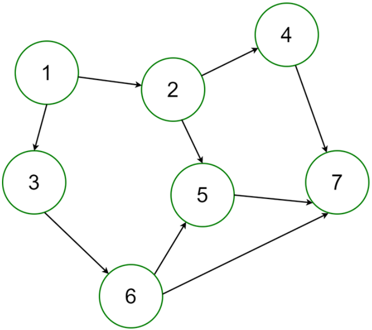


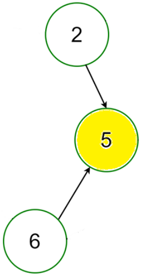
Figure: A complex DAG in medical science

Local Markov assumption

The local Markov assumption allows us to focus on two layers - a node and its parents.



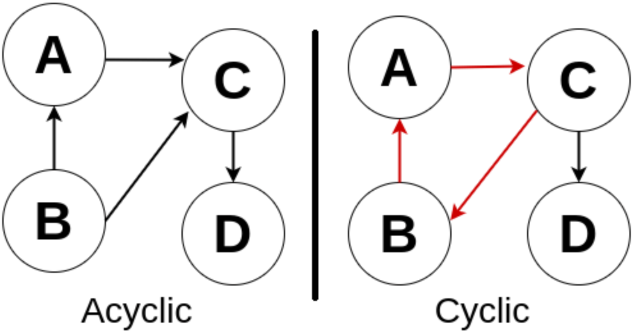
(a) Full DAG



(b) DAG of Node 5

Causal graphs must be acyclic

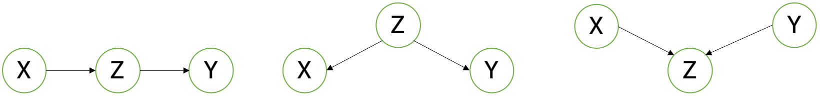
No cycle exists in a causal graph, known as the "**no absorbing states**."
In econometrics, this is called "**no simultaneity**," which implies the policy effects are from a "**partial equilibrium**."



Starting with motifs: 3 motifs

Graphs are complex, and how to analyze it?

- We focus on motifs - repeated structures on graphs.
- Three-node motifs: chains, forks, and collisions (or immorality)

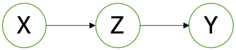


(a) Chain (b) Fork (c) Collision

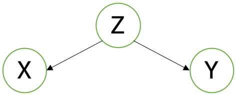
Starting with motifs: some observations

...between X and Y	Chains	Forks	Collisions
Association	> 0	> 0	$= 0$
Causation	> 0	$= 0$	$= 0$

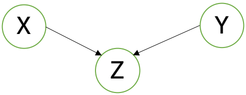
Table: Association vs. causation in 3 motifs



(a) Chain



(b) Fork



(c) Collision

Starting with motifs: blocking paths

- For a fork $X \leftarrow Z \rightarrow Y$, X and Y are correlated but $X \nrightarrow Y$.
- But how do we know this if we only observe data or $\gamma(X, Y)$?
- Solution: **What if we remove the arrow into X ($Z \rightarrow X$)?**

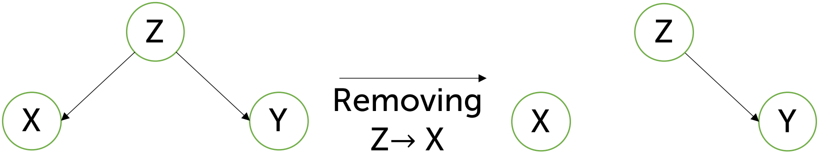


Figure: Blocking incoming paths to X

Starting with motifs: blocking paths

If we want to know the effect of X on Y, we must remove all "redundant paths."

- In formal terms, we "block" all unwanted paths.

How to block paths between X and Y?

- Let's look at the motifs one by one.

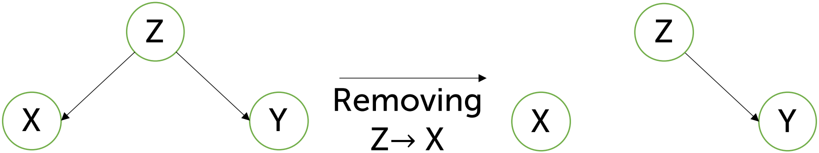
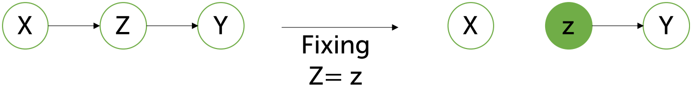


Figure: Blocking incoming paths to X

Blocking paths: chains

Chains (front-door paths)

To block a chain $X \to Z \to Y$, we fix the value of Z (a.k.a. the mediator) or set $Z = z$.



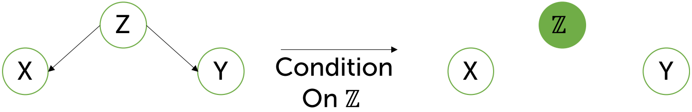
Testing with causal links assumption

A change in ΔX cannot pass through to Y.

Blocking paths: forks

Forks (back-door paths)

To block a fork $X \leftarrow Z \rightarrow Y$, we condition on the full support of $Z \in \mathbb{Z}$ (a.k.a. the moderator).



Testing with causal links assumption

Given Z, the value of X and Y are pre-determined.
A change in ΔX therefore does not pass over to Y.

Blocking paths: collisions

Collisions (immoral paths)

A collision path $X \to Z \leftarrow Y$ is already blocked.

Testing with causal links assumption

A change in ΔX changes Z, but NOT Y.

Blocking paths: more on collisions

Do not condition on collider Z

It is tempting, but conditioning on Z creates spurious correlation between X and Y!

An example

- In the general population, gender and age are independent.
- In the sub-population of CEOs, more older males.
- In the sub-population of nurses, more younger females.

D-separation: the general rules of path-blocking

Definition (Blocked Paths)

A path between nodes X and Y is blocked by a node W along the path if either of the following is true:

1. *For a chain along the path, $\cdots \to W \to \cdots$, W is fixed.*
2. *For a fork, $\cdots \leftarrow W \to \cdots$, W is conditioned on.*
3. *For a collision $\cdots \to W \leftarrow \cdots$, W is not conditioned on, as well as all descents of W.*

D-separation

Two nodes X and Y are d-separated by a set of nodes Z, if all paths between X and Y are blocked by Z.

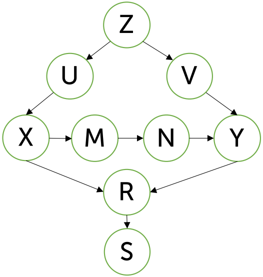
D-separation: an application

First, we find all the three types of paths:

1. Backdoor paths: $X \leftarrow U \leftarrow Z \rightarrow V \rightarrow Y$.
2. Frontdoor paths: $X \rightarrow M \rightarrow N \rightarrow Y$.
3. Collision paths: $X \rightarrow R$ (de(R)) $\leftarrow Y$.

Second, we construct a blocking set that:

1. Includes one or two or three of$\{U, V, Z\}$.
2. Includes one or two of $\{M, N\}$
3. Does NOT include R and S (descendant of R)



Outline

Outline

Interventions / Manipulations

When they are infeasible...

- Randomized control experiments - **the golden standard.**
- However, not everything can be intervened (ethically).
  - Example: age, gender, or some medical treatments etc.
- In business or economics studies:
  - Something is simply too costly to intervene.
  - e.g., lowering price of electricity? Tax cuts?

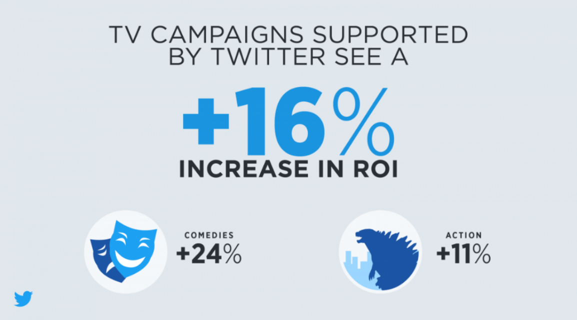
What then?

- People turn to observational data.
  - "Conditional on..." or "adding control variables..."
- Difficult to untangle causation from correlation.
- What's worse: we are causal animals.
  - We see patterns and are drawn towards causality.
  - **It is imperative to evaluate causal claims, but how?**

Twitter example

New movie marketing research reveals Twitter Ads deliver increased ticket sales (LINK)

- "The research team used multivariate regression analysis (a process that measures and predicts the sales impact of various media channels) to understand the effects of changes in Twitter media for movies."



Outline

The do-operator

- Do-operator formalizes manipulations in causal graphs.
  - Manipulating X: $\mathrm{do}(X) = x$ or setting $X = x$.
- It's the first step to evaluate causal claims.
- Graphical interpretation: **to remove all arrows into X.**

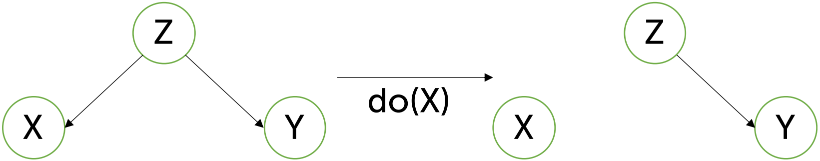


Figure: Do-operation $\Longleftrightarrow$ Manipulation

$P(Y \mid X = x)$ vs. $P(Y \mid do(X) = x)$

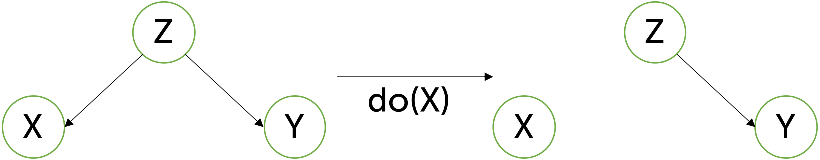
Comparing $P(Y \mid X = x)$ vs. $P(Y \mid do(X) = x)$

$\mathbf{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$	$\mathbf{P}(\mathbf{Y} \mid do(\mathbf{X}) = \mathbf{x})$
Conditioning	Intervening
Purely Statistical	Beyond statistics
Select a subgroup	Create a population
Always calculable	Not always calculable
In the full DAG	In a surgically-altered DAG (subgraph)

How to obtain $P(Y \mid do(X) = x)$?

Procedure to obtain $P(Y \mid do(X) = x)$

1. Do **a surgery on the full DAG** by removing all arrows into X
2. Transform the conditional probability with **the new DAG**
   1. i.e., $P(Y \mid do(X) = x) = P(Y \mid X)$
3. Calculate the conditional probability $P(Y \mid X) = P(Y)$



Outline

The definition of do-calculus

Definition (Do-calculus)

The do-calculus is an axiomatic system for replacing probability formulas containing the do operator with ordinary conditional probabilities.

- The normal reaction to the definition is "**say what?**"
- Let's look at the intuitions behind it!

The intuitions behind do-calculus

- To know causal effects, we need $P(Y \mid \mathrm{do}(X))$.
- E.g., Average Treatment Effect: $P(Y \mid \mathrm{do}(X) = 1) - P(Y \mid \mathrm{do}(X) = 0)$.

- We need to manipulate X to get $P(Y \mid \mathrm{do}(X))$, but in many cases, we cannot.
- We can have $P(Y \mid X)$, but $P(Y \mid X) \neq P(Y \mid \mathrm{do}(X))$.

Question: **under what conditions** $P(Y \mid X) = P(Y \mid \mathbf{do}(X))$**?**

The intuitions behind do-calculus

- To know causal effects, we need $P(Y \mid \text{do}(X))$.
- E.g., Average Treatment Effect: $P(Y \mid \text{do}(X) = 1) - P(Y \mid \text{do}(X) = 0)$.
- We need to manipulate X to get $P(Y \mid \text{do}(X))$, but in many cases, we cannot.
- We can have $P(Y \mid X)$, but $P(Y \mid X) \neq P(Y \mid \text{do}(X))$.

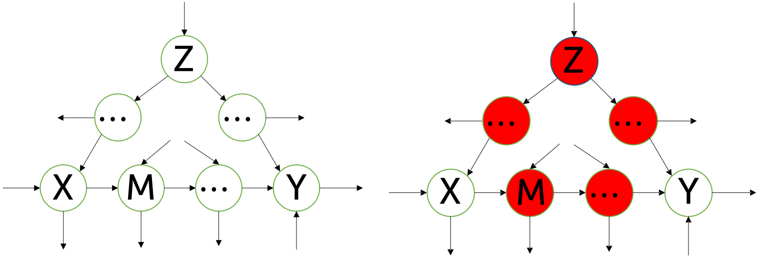
Question: **under what conditions** $P(Y \mid X) = P(Y \mid \mathbf{do}(X))$**?**

Illustrating the intuition

In practice, the DAG may be **much more complex**.

Do-calculus: which do(·)'s for other variables are essential if

$$P\,(Y \mid \mathrm{do}\,(X)) = P\,(Y \mid X)\,.$$



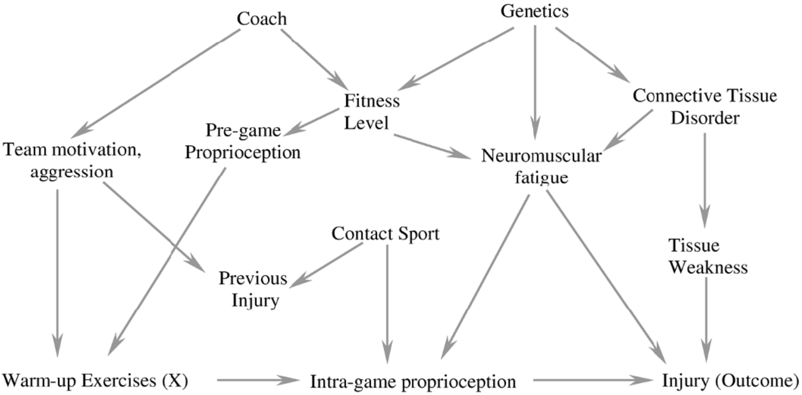
(a) A complex DAG (b) Which one(s) to manipulate?

Three rules of do-calculus

- Omitted here. Please see [1, Pearl 1995] for more details.
- Proved by [2, Shpitser and Pearl 2006], the rules are complete.
  - That is, the three rules are sufficient to transform $P(Y \mid \mathrm{do}(X), \cdots) \mapsto P(Y \mid X, \cdots)$.
  - If the rules are not met, $P(Y \mid \mathrm{do}(X), \cdots)$ may not be known from the DAG.
- For us: **causal identification for business research**.
  - The implications from the three rules of do-calculus.

To sum up: Where do DAGs come from?

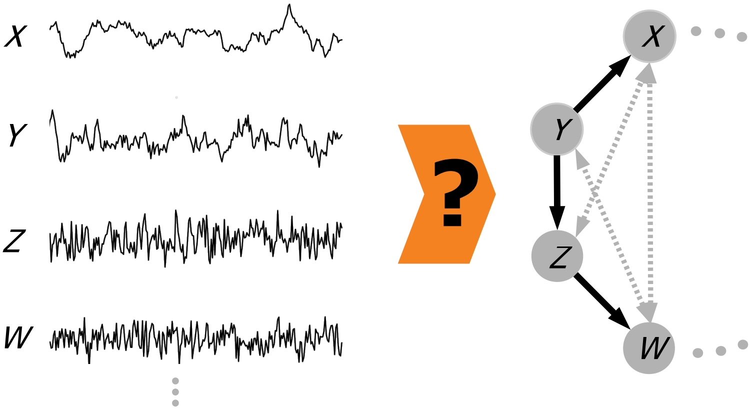
- **From human minds or "theories".**



RSM Erasmus

To sum up: Where do DAGs come from?

- **From data with causal discoveries** (using do-calculus).

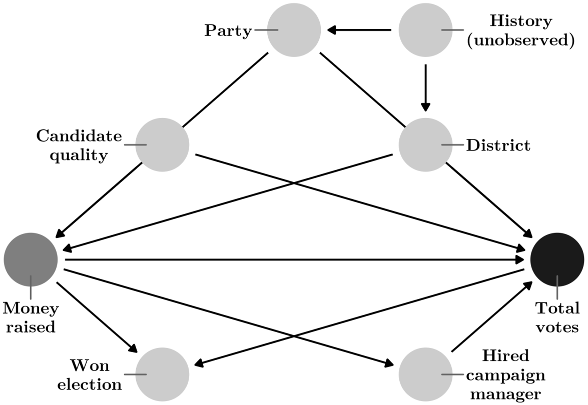


Outline

Outline

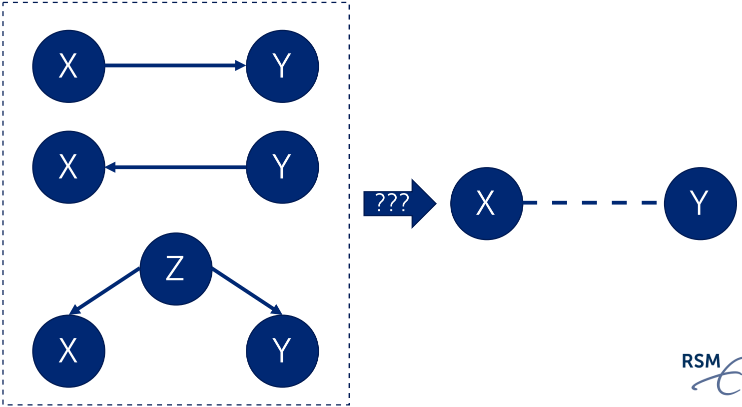
Two types of causal inference

- Type 1: Given data of a system, to find its causal structure.
- **Type 2: Given a causal structure, measure the effect of one variable on another.**



Explaining "Correlation $\neq$ Causation"

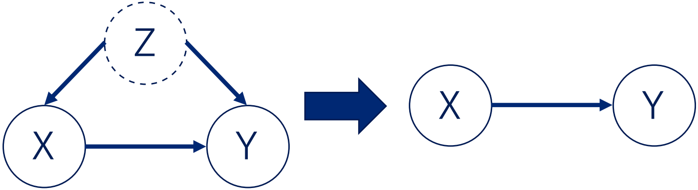
- Correlation is produced by causation
- Multiple causal structures (DAGs) can produce same correlation pattern



The core problem of causal inference

Definition (The Core Problem)

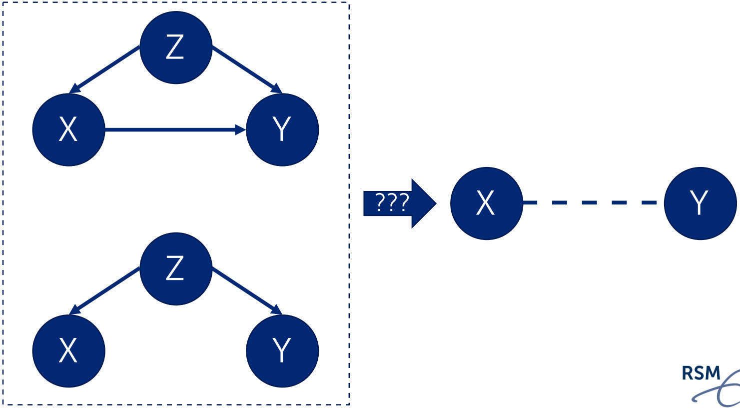
Suppose we want to quantify the effect of X on Y from the DAG below and we perfectly observe X and Y. However, the set of confounders Z is partially or unobserved.



Z has many names: lurking variables, omitted variables, unobservable etc.

The core problem is an identification problem

In almost all business research, obtaining $P(Y \mid do(X))$ boils down to distinguishing two observation-equivalent DAGs.



Defining the identification problem

Definition (The identification problem)

Multiple data-generating processes can result in the same patterns in data.

- **Data-generating processes**: models or theories.
- **Patterns in data**: information in data or statistics of data (e.g., correlations).

Endogeneity issue in linear regression $Y = \beta X + \varepsilon$

$$\underbrace{P(Y \mid X)}_{\text{Patterns in Data}} = P(\beta X + \varepsilon \mid X) = \underbrace{\beta X}_{\text{Cause 1}} + \underbrace{P(\varepsilon \mid X)}_{\text{Cause 2}}$$

Defining the identification problem

Definition (The identification problem)

Multiple data-generating processes can result in the same patterns in data.

- **Data-generating processes**: models or theories.
- **Patterns in data**: information in data or statistics of data (e.g., correlations).

Endogeneity issue in linear regression $Y = \beta X + \varepsilon$

$$\underbrace{P(Y \mid X)}_{\text{Patterns in Data}} = P(\beta X + \varepsilon \mid X) = \underbrace{\beta X}_{\text{Cause 1}} + \underbrace{P(\varepsilon \mid X)}_{\text{Cause 2}}$$

The complex concept of identification

We stick to the core problem of **causal identification** for clarity.

Over-identification

Parametric identification

Partial identification

Identification by construction

Exact identification

Essential identification

Structural identification

First-order identification

Point identification

Identification of bounds

Sampling identification

Identification at infinity

Set identification

Nonparametric identification

Non-robust identification

Causal identification

Strong identification

Bayesian identification

Generic identification

Ill-posed identification

Global identification

Weak identification

Irregular identification

Thin-set identification

Nonstandard weak identification

Eventual identification

Frequentist identification

Local identification

Semi-strong identification

Semiparametric identification

Nearly-weak identification

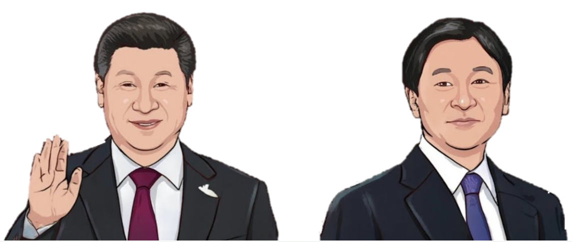
Identification arrangement

Bayesian identification

Causal identification as an axiomatization process

What is an axiomatization process?

Example (Who's Chinese? Left or right?)

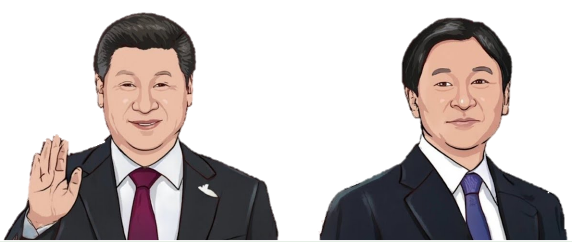


- Assumption 1: Chinese with two eyes and Japanese with two ears.
  - **Conclusion 1: Left? Right?**
- Assumption 2: Chinese with blue ties and Japanese with red ties.
  - **Conclusion 2: the man on the right!**

Causal identification as an axiomatization process

What is an axiomatization process?

Example (Who's Chinese? Left or right?)

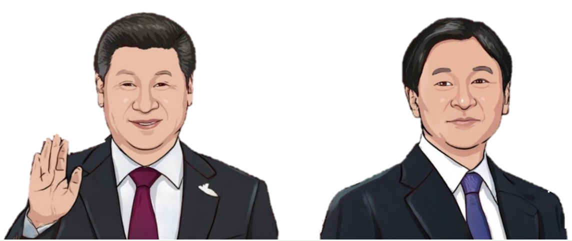


- Assumption 1: Chinese with two eyes and Japanese with two ears.
  - **Conclusion 1: Left? Right?**
- Assumption 2: Chinese with blue ties and Japanese with red ties.
  - **Conclusion 2: the man on the right!**

Causal identification as an axiomatization process

What is an axiomatization process?

Example (Who's Chinese? Left or right?)

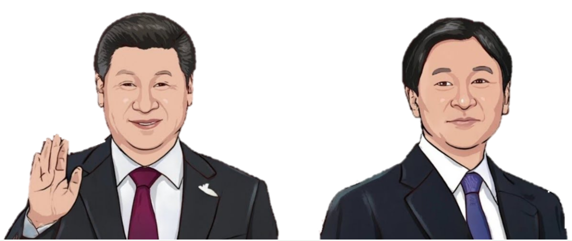


- Assumption 2: Chinese with blue ties and Japanese with red ties.
  - **Conclusion 2: the man on the right!**
- Assumption 3: Chinese with slicked-backs and Japanese with side-parts.
  - **Conclusion 3: the man on the left!**

Causal identification as an axiomatization process

What is an axiomatization process?

Example (Who's Chinese? Left or right?)



- Assumption 2: Chinese with blue ties and Japanese with red ties.
  - **Conclusion 2: the man on the right!**
- Assumption 3: Chinese with slicked-backs and Japanese with side-parts.
  - **Conclusion 3: the man on the left!**

Causal identification as an axiomatization process

Definition (Causal identification)

Making assumptions that are sufficient and learning causal relationships from data based on the assumptions.

- Causal identification requires sufficient assumptions.
- Given the same data, different assumptions may lead to different conclusions.
- Alternative sufficient assumptions exist.

Endogeneity issue in linear regression $Y = \beta X + \varepsilon$

$$\underbrace{P(Y \mid X)}_{\text{Patterns in Data}} = P(\beta X + \varepsilon \mid X) = \underbrace{\beta X}_{\text{Cause 1}} + \underbrace{P(\varepsilon \mid X)}_{\text{Cause 2}}$$

- Assume $P(\varepsilon \mid X) = 0$.

Causal identification as an axiomatization process

Definition (Causal identification)

Making assumptions that are sufficient and learning causal relationships from data based on the assumptions.

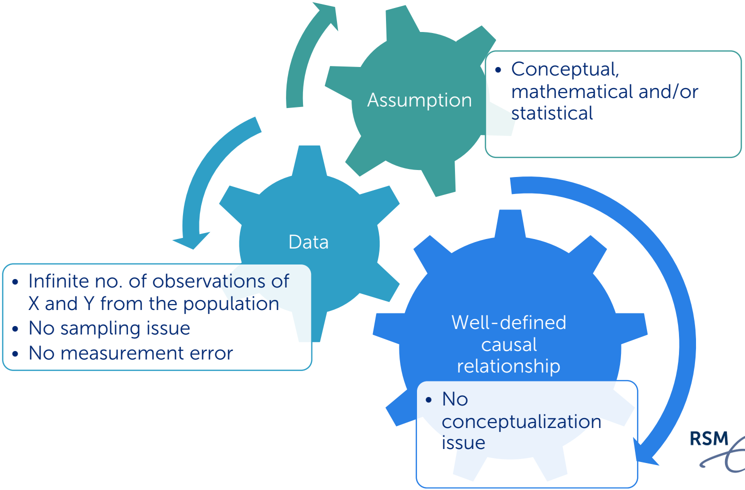
- Causal identification requires sufficient assumptions.
- Given the same data, different assumptions may lead to different conclusions.
- Alternative sufficient assumptions exist.

Endogeneity issue in linear regression $Y = \beta X + \varepsilon$

$$\underbrace{P(Y \mid X)}_{\text{Patterns in Data}} = P(\beta X + \varepsilon \mid X) = \underbrace{\beta X}_{\text{Cause 1}} + \underbrace{P(\varepsilon \mid X)}_{\text{Cause 2}}$$

- Assume $P(\varepsilon \mid X) = 0$.

3 pillars of causal identification



Identification vs. estimation

Proposition (Identification vs. Estimation)

The relationship between identification and estimation is as such: first, identification logically precedes estimation; and second, one does not logically imply the other.

Two examples

Estimable but not identifiable: $\text{Sales} = \beta\text{Price} + \varepsilon$.
Identifiable but not estimable: Static discrete game with multiple equilibria.

Outline

Identification by conditioning

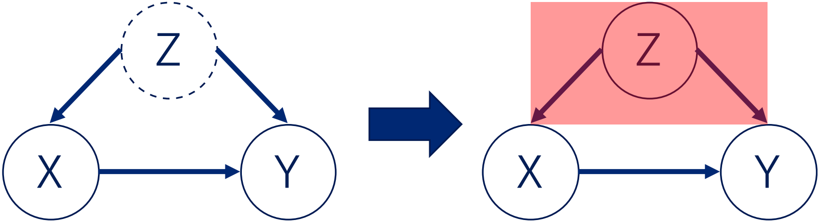


Figure: Backdoor criterion

The set of variables Z satisfies the **backdoor assumptions**:

1. **Exhaustive**: Z block ALL backdoor paths from X to Y.
2. **No colliders**: Z does NOT contain any common descendants (colliders) of X and Y.

RSM Erasmus

Identification by conditioning

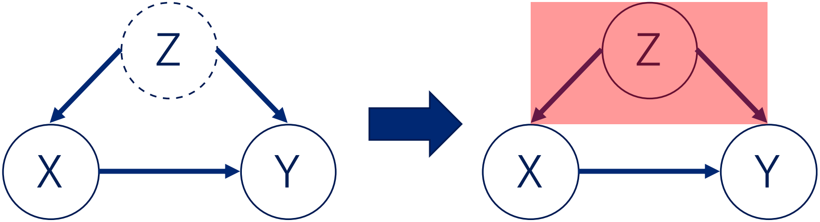


Figure: Backdoor Criterion

Implications of the **backdoor assumptions**:

1. **Exhaustive**: Z contain all common parents of X and Y.
2. **No colliders**: do not include post-treatment variables in Z.

Identification by conditioning

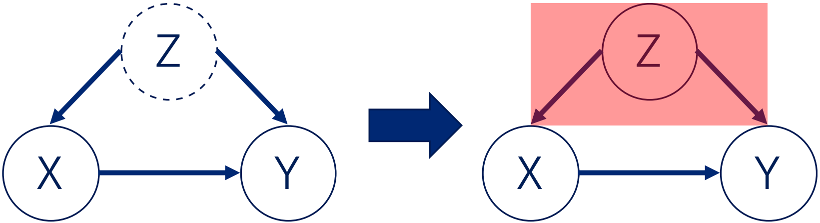


Figure: Backdoor Criterion

Estimation of $P(Y \mid \mathrm{do}(X))$:

- Non-parametric estimation:
  $P(Y \mid \mathrm{do}(X)) = \sum_z P(Y \mid X, z)\, P(z)$.
- In practice, we often use Z as control variables in our regression.

RSM Erasmus

Identification by mechanism

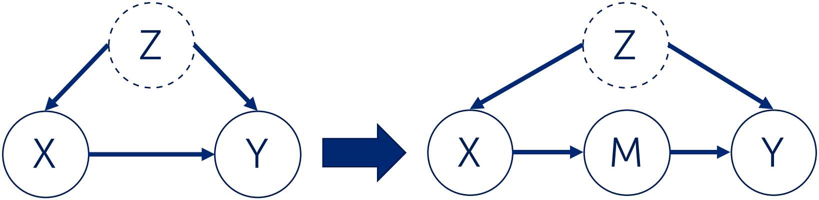


Figure: Front-door criterion

The set of variables M must satisfy the **front-door assumptions**:

- **Exogeneity to** X: All backdoor paths of X and M are blocked.
- **Exogeneity to** Y: All backdoor paths of M and Y are blocked by X.
- **Full Mediation**: M intercept all the front-door paths between X and Y.

Identification by mechanism

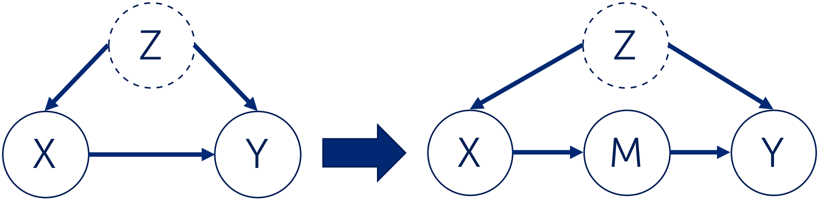


Figure: Front-door criterion

Identification of $P(Y \mid \mathrm{do}(X))$ given **front-door assumptions**:

1. $P(M \mid \mathrm{do}(X))$ is identified.
2. $P(Y \mid \mathrm{do}(M))$ is identified.
3. Combining $P(M \mid \mathrm{do}(X))$ and $P(Y \mid \mathrm{do}(M))$.

Identification by mechanism

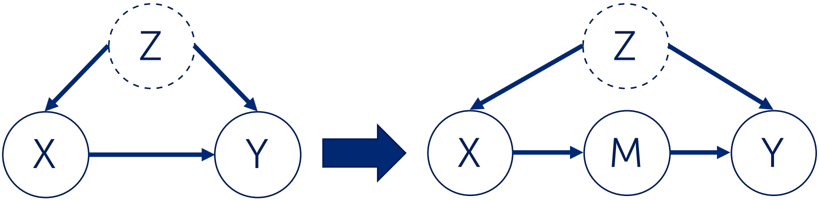


Figure: Front-door criterion

Estimation: Combining $P(M \mid do(X))$ and $P(Y \mid do(M))$ to get $P(Y \mid do(X))$.

$$P(Y \mid do(X)) = \sum_{m} \left[P(M \mid X) \times \sum_{\widetilde{X}} \left(P\left(Y \mid M, \widetilde{X}\right) P\left(\widetilde{X}\right) \right) \right]$$

Identification by mechanism

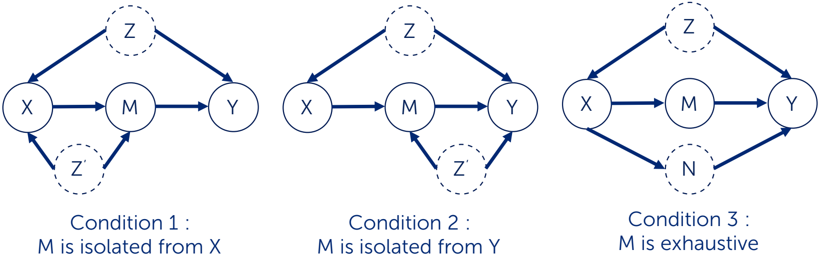
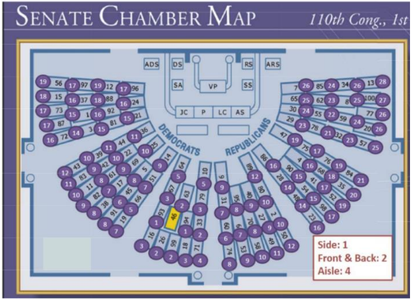


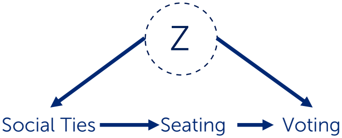
Figure: Understanding the front-door assumptions

In practice, you need an **exogenous super-mechanism** as M.

Identification by mechanism: examples



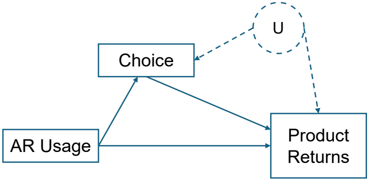
Congressional Seating



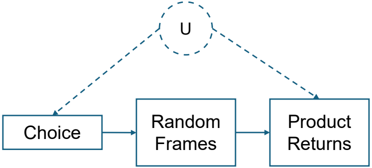
The DAG

Table: Log-rolling in congressional voting (Cohen and Malloy 2014)

Identification by mechanism: examples



(a) AR experiment



(b) Front-door design

- The indirect effect AR $\rightarrow$ Choice $\rightarrow$ Returns is not identified as choices are self-selected.
- Participants are randomly given the frame of their choice or the other one in the set.

Identification by instruments: the origin

"Exogenous shocks/variation"

- An inventions of economists
- Wide adoption outside of economics
- The origin can be traced back to 1920's

One of the early contributors: Jan Tinbergen (1930)

- Bestimmung und Deutung von Angebotskurven Ein Beispiel
- Determination and Interpretation of Supply Curves: An Example

Identification by instruments: how?

Instead of mutilating the DAG, we expand it!

- $P(Y \mid \text{do}(V)) = P(Y \mid V)$: identified.
- $P(X \mid \text{do}(V)) = P(X \mid V)$: identified.
- To have $P(Y \mid \text{do}(X))$, we solve this equation: $P(Y \mid \text{do}(V)) = \sum_X P(Y \mid \text{do}(X))\, P(X \mid \text{do}(V))$.

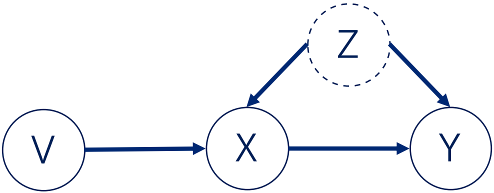
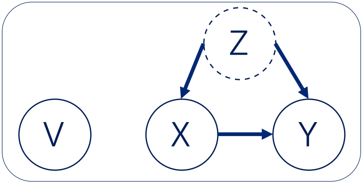
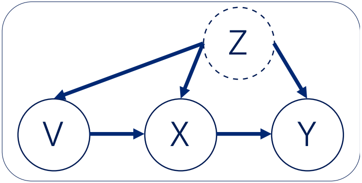


Figure: DAG of the IV approach

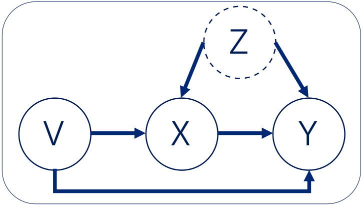
Identification by instruments: assumptions



Relevance



Exogeneity



Exclusion

Is any assumption testable on its own right?

Identification by set

Instead of focusing on a specific value, we try to have a "set".

- "Half-baked" as compared to point identification.
- a.k.a. partial identification, identification by bounds, or interval identification etc.
- Conceptually insightful but empirically less attractive.

Example

An experiment with liking of a brand as outcome
Measured with $0-9$ points scale (0−dislike and 9−like)
Treatment effects must $[-9, +9]$ ("no data bound")

Identification by set

Instead of focusing on a specific value, we try to have a "set".

- "Half-baked" as compared to point identification.
- a.k.a. partial identification, identification by bounds, or interval identification etc.
- Conceptually insightful but empirically less attractive.

Example

An experiment with liking of a brand as outcome
Measured with $0-9$ points scale (0−dislike and 9−like)
Treatment effects must $[-9, +9]$ ("no data bound")

The law of credibility

Quote Manski 2003(Chapter 1):

"Empirical researchers should be concerned with both the logic and the credibility of their inferences. Credibility is a subjective matter, yet I take there to be wide agreement on a principle I shall call: ***The Law of Decreasing Credibility: The credibility of inference decreases with the strength of the assumptions maintained.***
This principle implies that empirical researchers face a dilemma as they decide what assumptions to maintain: Stronger assumptions yield inferences that may be more powerful but less credible."

Application of the law of credibility

Set 1
- No assumptions

Sharper Set 2
- More or stronger assumptions

...Sharper Set...
- ...Even more or stronger assumptions...

Points
- “Minimal sufficient assumptions”

Outline

Summary I

The essence of causal inference is identification.

The data alone cannot reveal the underlying causal structures.

Identification is to find sufficient assumptions to pinpoint a causal structure.

Assumptions + Data = Conclusions

Summary II

The identification-estimation flowchart

The "faithfulness" assumption

- The causal link assumption is not sufficient to have a unique DAG. We need the faithfulness assumption to deal with tricky scenarios.

An example

- Consider a DAG, with $X \to Y \leftarrow Z$, and we further assume $Y = X + Z$, without the loss of generality.
- Conditional on $Y = y$, the covariance between X and Z should be **non-zero**.
  - $cov\left(X, Z \mid Y = y\right) = cov\left(X, y - X\right) = yE\left(X\right) - E\left(X^2\right)$.
- If $yE\left(X\right) - E\left(X^2\right)$ is always 0, then $cov\left(X, Z \mid Y = y\right) = 0$.
- If $cov\left(X, Z \mid Y = y\right)$ is always 0, then the DAG $X \to Y \leftarrow Z$ cannot be right. But it is the true DAG.
- We therefore have the assumption to rule the situation.

Do-operator: some formality

Definition (Do-operator)

If we intervene on a set of nodes S on a DAG, then for all nodes X_i, we have the following:

1. *For $X_i \notin S$, $P(X_i \mid pa(X_i))$ remains unchanged.*
2. *For $X_i \in S$, if $x_i = do(X_i)$, then $P(X_i = x_i \mid pa(X_i)) = 1$, and otherwise $P(X_i = x_i \mid pa(X_i)) = 0$.*

RSM Erasmus

Do-operator: some formality

Corollary (Truncated Factorization)

We assume that P and G satisfy the Markov and causal links assumption. Given a set of intervened nodes S, we have,

$$P\left(X_1, X_2, \cdots, X_n \mid do\left(S = s\right)\right) = \prod_{X_i \notin S} P\left(X_i \mid pa\left(X_i\right)\right)$$

The three rules of do-calculus

For a DAG G, let $G(\nrightarrow_X)$ denote the graph G with all incoming links to X removed. Similarly, $G(\nleftarrow_X)$ denote the graph G with all outgoing links from X removed. Combining these two, the graph G with incoming links to X and outgoing links from Z removed is denoted as $G(\nrightarrow_X \nleftarrow_Z)$. In addition, I will use $\perp_G$ to represent d-separation on the graph.

Theorem (The Rules of Do-calculus)

Given a DAG G, an associated distribution P, and disjoint set of variables Y, X, Z and W, the following rules hold:

1. $P(Y \mid do(X), Z, W) = P(Y \mid do(X), W)$, *if* $Y \perp_{G(\nrightarrow_X)} Z \mid X, W$.
2. $P(Y \mid do(X), do(Z), W) = P(Y \mid do(X), Z, W)$, *if* $Y \perp_{G(\nrightarrow_X \nleftarrow_Z)} Z \mid X, W$.
3. $P(Y \mid do(X), do(Z), W) = P(Y \mid do(X), W)$, *if* $Y \perp_{G\left(\nrightarrow_X \nrightarrow_{Z \setminus Ancestor(W)}\right)} Z \mid X, W$, *where* $Z \setminus Ancestor(W)$ *denotes non-ancestors of* W *in set* Z.

The intuition for the 1st rule

To get the intuition for Rule 1, we may simply remove the intervention $do(X)$. The removal of $do(X)$ reverses the manipulated graph $G(\nrightarrow_X)$ to the original graph G. We also remove any other X from the picture. Thus, we have

$$P(Y \mid Z, W) = P(Y \mid W), \text{ if } Y \perp_G Z \mid W$$

This simply implied if Y and Z are d-separated, then Y is independent from Z, conditional on W. This is from the very definition of d-separation.

The intuition for the 2nd rule

Similar to Intuition for Rule 1, we remove the $do\,(X)$, and get

$$P\,(Y \mid do\,(Z)\,, W) = P\,(Y \mid Z, W)\,, \text{ if } Y \perp_{G(\nleftarrow_Z)} Z \mid W$$

This is the backdoor criterion, where W blocks all backdoor path between Z and Y. The operation $G\,(\nleftarrow_Z)$ serves to block all frontdoor paths from Z, and thus all frontdoor paths between Z and Y are blocked, and only the backdoor paths remain which are all blocked by conditioning on W.

The intuition for the 3rd rule

Again, we remove the $do(X)$, and get

$$P(Y \mid do(Z), W) = P(Y \mid W), \text{ if } Y \perp_{G\left(\nrightarrow_{Z \setminus \text{Ancestor}(W)}\right)} Z \mid W$$

This implies if we remove the do-operation $do(Z)$ by reintroducing the links that come into Z, we would not cause flow of association into Y. Because $do(Z)$ removes all incoming links to Z give us $G(\nrightarrow_Z)$, the main pathways for the flow of association between Z and Y are the frontdoor paths from Z to Y in $G(\nrightarrow_Z)$. If W contain all chain nodes, naturally Y and Z will be d-separated. However, if W also contains colliders (i.e., nodes that are descendants of Z), conditioning would lead to non-causal association flow between Y and Z. Therefore, we have to modify the set of Z and limit the set of manipulated nodes in Z to those that are not ancestors of W. By doing so, W would not contain colliders or descendants of colliders. Simply put, Rule 3 applies the do-separation criterion on colliders.

References I

- Pearl, J. (1995). Causal diagrams for empirical research. Biometrika, 82(4), 669-688.
- Shpitser, I., & Pearl, J. (2006, February). Identification of joint interventional distributions in recursive semi-Markovian causal models. In Proceedings of the National Conference on Artificial Intelligence (Vol. 21, No. 2, p. 1219). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.
- Cohen, L., & Malloy, C. J. (2014). Friends in high places. American Economic Journal: Economic Policy, 6(3), 63-91.
- Manski, C. F. (2003). Partial identification of probability distributions. Springer Science & Business Media.
- Rosenbaum, P. R. (2014). Sensitivity analysis in observational studies. Wiley StatsRef: Statistics Reference Online.