Title: Tutorial 2

Speakers: Ilya Shpitser

Collection: Causal Inference & Quantum Foundations Workshop

Date: April 17, 2023 - 2:00 PM

URL: https://pirsa.org/23040106

An Introduction to Causal Inference

Ilya Shpitser

Causal Inference And Quantum Foundations Workshop

April 17, 2023

Overview

- Statistical versus causal models of a Directed Acyclic Graph (DAG).
- One graph, many causal models.
- Hidden variables in causal inference.
- Nomenclature / glossary (throughout).

Statistical Models

- Statisticians think about joint probability distributions $p(\vec{V})$ on a set of random variables $\vec{V}$.
- Glossary: a statistical model is a set of distributions on a particular set of random variables.
- For example, if $\vec{V} = \{Y\} \cup \vec{W}$, where Y is an outcome variable, and $\vec{W}$ is a vector of feature variables, a linear regression model is the following set of distributions on $p(Y, \vec{W})$:

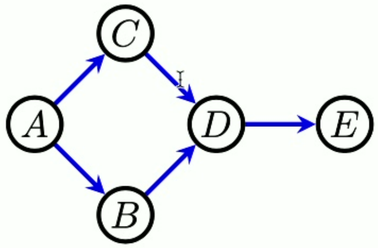
$$\left\{ p(Y, \vec{W}) : Y = \beta_0 + \vec{\beta}^T \cdot \vec{W} + \epsilon \right\},$$

  where ϵ is typically a Gaussian random variable.
- This is generally not how the word "model" is used in physics, but the above definition is important to keep in mind when talking to statisticians or reading their literature.

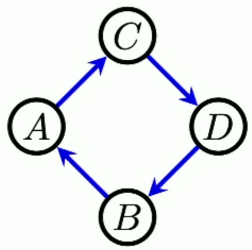
2/41

Directed Acyclic Graphs

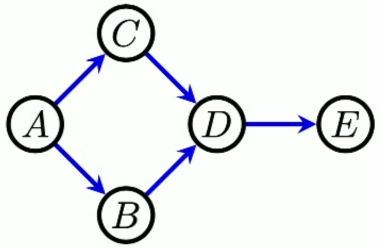
- Directed Acyclic Graphs (DAGs) have vertices and (directed) edges connecting vertex pairs.
- DAGs do not allow directed cycles.
- Positive example:



- Negative example:



Genealogic Relations Among Vertices In A DAG



- In graph theory, vertex relations in a graph are described using genealogic terms.
- For example, in the graph above, we have:
  - A is a parent of B, B is a child of A.
  - A is an ancestor of E, E is a descendant of A.
- By convention every vertex is both an ancestor and a descendant of itself.
- We define the following notation for sets of vertices related to any vertex V:

$$\text{parents of } V : \mathrm{pa}_{\mathcal{G}}(V);$$
$$\text{children of } V : \mathrm{ch}_{\mathcal{G}}(V);$$
$$\text{ancestors of } V : \mathrm{an}_{\mathcal{G}}(V);$$
$$\text{descendants of } V : \mathrm{de}_{\mathcal{G}}(V).$$

The Statistical Model Of A DAG

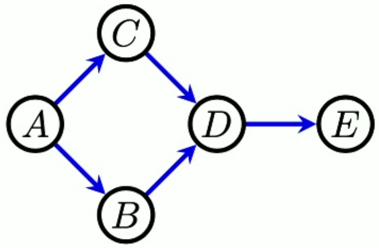
- A graphical model is a statistical model associated with a graph in a particular way.
- Random variables in a distribution in a graphical model correspond to vertices in the associated graph.
- Often, notation for vertices and random variables is the same.
- Three definitions (all involve the graph):
  - Factorization (probability distribution as a set of small factors).
  - Local Markov property (a small set of independence constraints).
  - Global Markov property (all independence constraints in the model). Will skip this.
- Example: the statistical model of a DAG $\mathcal{G}(\vec{V})$ is the set of distributions:

$$\left\{ p(\vec{V}) : p(\vec{V}) = \prod_{V \in \vec{V}} p(V \mid \mathrm{pa}_{\mathcal{G}}(V)) \right\}.$$

- This representation of $p(\vec{V})$ is called the DAG factorization.
- Another name for the statistical model of a DAG is the Bayesian network model.

DAG Factorization: An Example

Given the DAG

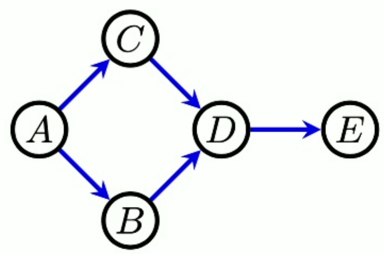


the corresponding statistical model is the set of all distributions $p(A, B, C, D, E)$ which can be written as:

$$p(A, B, C, D, E) = p(A)p(B \mid A)p(C \mid A)p(D \mid A, C)p(E \mid D).$$

Local Markov Property

- Graph implies a small list of independences that imply the rest.
- Every X is independent of non-parental non-descendants, conditional on parents.
- Example:



- $(C \perp\!\!\!\perp B \mid A)$, $(D \perp\!\!\!\perp A \mid B, C)$, $(E \perp\!\!\!\perp A, B, C \mid D)$.

Aside: Observational Equivalence

- Consider the following two DAGs:

- Local Markov property gives same independence: $(A \perp\!\!\!\perp C \mid B)$.
- In fact, the only independence in this model.
- If one graph is causal, the other isn't...
- These graphs are called observationally equivalent.
- This creates problems for model selection and model compatibility.

Aside: Statistical Inference

- Statistical models are used to formulate learning from data.
- One formulation goes like this:
  - We consider a statistical model $\mathcal{P}$, with one distribution $p_0(\vec{V}) \in \mathcal{P}$ (we don't know which) the "true distribution."
  - Nature generates a set of n samples $[\vec{V}] = (\vec{v}_1, \ldots, \vec{v}_n)$ which are independent draws from $p_0(\vec{V})$.
  - We are interested in learning values of a set of *target parameters* $\vec{\beta}$ in $p_0(\vec{V})$ from $[\vec{V}]$.
  - A function that maps possible $[\vec{V}]$ to a guess for $\vec{\beta}$ is called an *estimator*, with its output written as $\hat{\vec{\beta}}$.
- Statistical inference is the process of constructing and using this function to make a guess for $\vec{\beta}$ using data.
- Lots of problems may be formulated in this way: predictive modeling in machine learning, parameter estimation, image analysis, text and speech processing, model selection, etc.

Statistical Inference (Continued)

- We can write target parameters as a function $\vec{\beta}(\vec{\eta})$ of $\vec{\eta}$.
- A common approach to statistical inference in parametric $\mathcal{P}$ is to:
  - Posit a *likelihood function*

$$\mathcal{L}_{[\vec{V}]}(\vec{\eta}) = \prod_{i=1}^{n} p(\vec{v}_i; \vec{\eta}),$$

  - Choose $\vec{\eta}^*$ that maximize this function, and
  - Let $\vec{\beta}(\vec{\eta}^*)$ be our guess for $\vec{\beta}$ based on $[\vec{V}]$.
- Other approaches: minimize a loss tailored to our application, solve an estimating equation we know should hold, etc.

Causal Models: Two Approaches

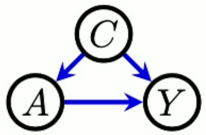
- The *random variable approach* (closer to statistics, originating with Jerzy Neyman).
- The *causal mechanism approach* (closer to econometrics, and computer science, originating with Sewall Wright).
- These approaches are closely connected.

The Potential Outcome Approach

- A potential outcome (or counterfactual) $Y(a)$ reads "the outcome Y if A were set, possibly contrary to fact, to value a."
- $Y(a)$ is a *random variable*.
- One conception of causal models is as statistical models of joint distributions of counterfactual random variables.
- Many ways to do so, we will describe causal models of a DAG.
- In such a causal model, directed edges in the DAG represent "direct causal relationship" between two variables.
- This is cashed out in different ways.

Graphical Causal Model (Counterfactual View)

- Given a DAG $\mathcal{G}(\vec{V})$, for every $V \in \vec{V}$, assume counterfactuals $V(\vec{a}_V)$ exist, for all values $\vec{a}_V$ of $\text{pa}_{\mathcal{G}}(V)$.

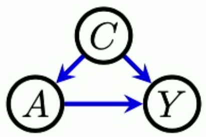


- Example (for binary C, A, Y, and the graph above): C, $A(c=0)$, $A(c=1)$, $Y(c=0,a=0)$,$Y(c=0,a=1)$, $Y(c=1,a=0)$, $Y(c=1,a=1)$ exist.
- Recall: $\text{pa}_{\mathcal{G}}(V)$ are "direct causes" of V.
- $V(\vec{a}_V)$ described the behavior of V in response to direct causes assuming particular values.

Defining General Counterfactuals

- $V(\vec{a}_V)$ exist, for all values $\vec{a}_V$ of $\text{pa}_{\mathcal{G}}(V)$ are called *one-step-ahead counterfactuals*.
- We use them to construct other counterfactuals (inductively) via recursive substitution:

$$V(\vec{a}) = V(\vec{a}_{\vec{A}\cap \text{pa}_{\mathcal{G}}(V)}, \{W(\vec{a}) : W \in \text{pa}_{\mathcal{G}}(V) \setminus A\})$$



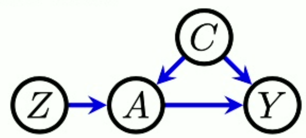
- Examples (for graph above):

$$Y(a) \equiv Y(a, C)$$
$$Y(c) \equiv Y(c, A(c)).$$

- Interpret $Y(a, C)$ to mean "Y if A were set to a, and C were set to whatever value it would naturally take."

Consequences of Recursive Substitution

- Recursive substitution has a number of important implications.
- Causal irrelevance: given a set of interventions, a counterfactual outcome is only influenced by those interventions that appear in the recursive substitution definition.



- Example: in the graph above, $Y(a, z) \equiv Y(a, C)$ is not a function of z.
- These constraints are sometimes called exclusion restrictions.
- There are other interpretations of this: will come back to this later.
- Exclusion restrictions correspond to missing edges in a graph (missing $Z \to Y$ edges).
- Consistency: states that if $\vec{W}(\vec{a}) = \vec{w}$ then $\vec{Y}(\vec{a}, \vec{w}) = \vec{Y}(\vec{a})$.
- Example: in the graph above, if $A = a$, $Y(a) = Y$.
- Consistency allows us to link counterfactual and observed variables.
- Sometimes phrased as coarsening: $Y = Y(A) = Y(a = 1)A + Y(a = 0)(1 - A)$ (for a binary A).

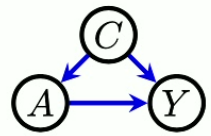
15/41

A Generic Causal Model

- Recall: a model is a set of distributions.
- Given a DAG $\mathcal{G}$, and a set of one-step-ahead counterfactuals $V(\vec{a}_V)$ on $\vec{V}$, the set of distributions

$$p(\{V(\vec{a}_V) : \vec{a}_V \in \mathfrak{X}_{\text{pa}_{\mathcal{G}}(V)}\} \cup \{V(\vec{a}_{\vec{A} \cap \text{pa}_{\mathcal{G}}(V)}, \{W(\vec{a}) : W \in \text{pa}_{\mathcal{G}}(V) \backslash \vec{A}\} : \vec{a} \in \mathfrak{X}_{\vec{A}}; \})$$

is called the *non-parametric structural equation model (NPSEM)*, or *structural causal model (SCM)*.



- Example: the NPSEM (for all binary variables) for the above graph is the set of all distributions of the form:

$$p(C, A(C), Y(A(C), C), \{A(c_1), Y(c_2, a_1), Y(a_2), Y(c_3) : c_1, c_2, c_3, a_1, a_2 \in \{0, 1\}\})$$

where $Y(a_2) = Y(a_2, C)$, $Y(c_3) = Y(A(c_3), c_3)$.

Identification

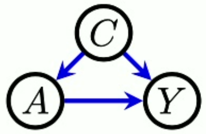
- In classical statistics, we want to learn about a parameter β of the observed data distribution $p(\vec{V})$ given samples from $p(\vec{V})$.
- In causal inference, the observed data distribution is still $p(\vec{V})$, but we want counterfactual parameters.
- This yields a question of identification: is a parameter such as $\mathbb{E}[Y(a)] = \int Y(a)p(Y(a))dY(a)$ a function of $p(\vec{V})$?
- In general, no.
- Causal models may give us assumptions under which parameters may be identified.

Graphical Causal Model (Structural Equation View)

- Given a DAG $\mathcal{G}(\vec{V})$, for every $V \in \vec{V}$, the values of V are determined by means of a *structural equation*:

$$f_V : \mathfrak{X}_{\mathrm{pa}_{\mathcal{G}}(V) \cup \{\epsilon_V\}} \mapsto \mathfrak{X}_V$$

  and an exogenous random variable ϵ_V.



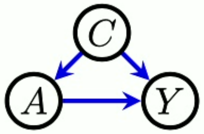
- Example (for binary C, A, Y, and the graph above):

$$\begin{aligned} C &= f_C(\epsilon_C) \\ A &= f_A(C, \epsilon_A) \\ Y &= f_Y(C, A, \epsilon_Y). \end{aligned}$$

- Recall: $\mathrm{pa}_{\mathcal{G}}(V)$ are "direct causes" of V.
- f_V described the behavior of V in response to direct causes assuming particular values.
- ϵ_V is needed since behavior of V may still be random, even if all direct causes are specified.

Interventions And Structural Equation Replacement

- In the structural equation view, counterfactuals are defined by equation replacement.
- An intervention that sets A to a is implemented by replacing f_A by f_A^* that always outputs a constant a.
- Counterfactual variables are defined using this new set of structural equations.



- Example: in the graph above, an intervention that sets A to a yields the following structural equations:

$$
\begin{aligned}
C &= f_C(\epsilon_C) \\
A^* &= f_A^* = a \\
Y(a) &= f_Y(C, A^*, \epsilon_Y) = f_Y(C, a, \epsilon_Y).
\end{aligned}
$$

The Structural Equation View of the NPSEM / SCM

- Given a DAG $\mathcal{G}$, and a set of structural equations and exogenous variables $\{f_V, \epsilon_V : V \in \vec{V}\}$, where each f_V maps from $\mathfrak{X}_{\text{pa}_\mathcal{G}(V)\cup\{\epsilon_V\}} \mapsto \mathfrak{X}_V$, an NPSEM or SCM is the set of the distributions of all variables under all possible interventions, such that the joint distribution $p(\{\epsilon_V : V \in \vec{V}\})$ is unrestricted.
- The potential outcome view and the structural equation view are equivalent: they use different notation and emphasize different things to describe the same object.

$$Y(\vec{a}) = f_Y(\{W : \text{pa}_\mathcal{G}(Y) \setminus \vec{A}\}, \vec{a}_{\text{pa}_\mathcal{G}(Y)\cap\vec{A}}, \epsilon_Y)$$

$$W(\vec{a}) = f_W(\{Z : \text{pa}_\mathcal{G}(Z) \setminus \vec{A}\}, \vec{a}_{\text{pa}_\mathcal{G}(W)\cap\vec{A}}, \epsilon_W)$$

$$\ldots$$

- The potential outcome view emphasizes the output (as a random variable), the structural equation view emphasizes the mechanism, and the intervention operation itself.

The Use Of The Term "Model"

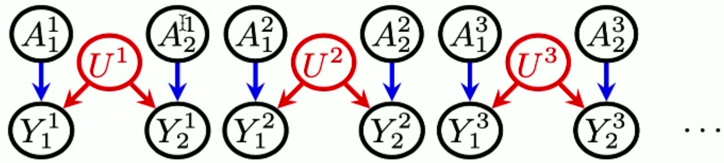
- Statisticians say "model" to mean "a set of distributions on a sample space."
- By contrast, in model/set theory, a "model" is some mathematical object about which we want to build a "theory" (a set of tautologies in some formal language).
- Some authors (Pearl, for example), use "model" in the sense closer to the latter.
- In particular, a "structural causal model" may be viewed as a particular set of structural equations f_V and exogenous variables ϵ_V associated with a particular DAG $\mathcal{G}$.
- This would be a "model" in the model-theoretic sense (a mathematical object we want to build a theory about).
- To a statistician, that same object would correspond to an element of the "model."

Individuals, Distributions, and Interference

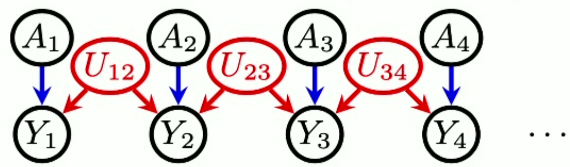
- In the standard view of statistical inference, "individuals" or "experimental units" correspond to data samples from the true distribution in the statistical model.
- Data samples are usually considered to be *independent, identically distributed (i.i.d.)*.
- Often not true in practice (social networks, infectious disease, spatial proximity).
- In causal inference, dependent samples are studied in *interference problems*.
- Glossary: interference variables of one "experimental unit" influences variables of another.

Two Types Of Interference

- Two types of interference:
  - Partial interference: data samples may be partitioned into independent blocks, with units in a blocks dependent.



  - Full interference: data samples are all pairwise dependent.



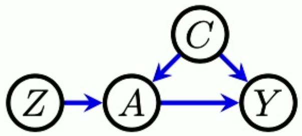
23/41

Aside: SUTVA

- A common assumption in the literature is SUTVA: Stable Unit Treatment Value Assumption.
- Two part assumption:
  1. $Y = Y(A)$ (consistency).
  2. Lack of interference.
- Being able to write $Y(a = 1)$ as a random variables, with samples $Y_i(a_i)$ corresponding to an experimental unit i implicitly assumes no dependence of Y_i on A of another unit j: $Y_i(a_i, a_j) = Y_i(a_i, a'_j)$.

Interpretation Of Exclusion Restrictions

- Two interpretations of causal irrelevance: $Y(a, z) = Y(a, z')$ for all z, z' in the graph below:



- Individual level: for every unit i, $Y_i(a_i, z_i) = Y_i(a_i, z_i')$ for all z_i, z_i'.
- Distribution/population level: the distribution $p(Y(a, z))$ is not a function of z for every a.
- Recursive substitution imposes an individual level restriction, but distribution/population level restrictions are sometimes discussed as well.

Defining Causal Models

- A causal model associated with a DAG $\mathcal{G}$ that assumes nothing beyond recursive substitution is an NPSEM or SCM.
- We may impose additional restrictions to obtain models where, in some sense, unobserved confounding is absent.
- Two important models:
  - NPSEM-IE (NPSEM with independent errors):

    sets $\{V(\vec{a}_V) : \vec{a}_V \in \mathfrak{X}_{\mathrm{pa}_{\mathcal{G}}(V)}\}$ are mutually independent $(\forall V \in \vec{V})$.

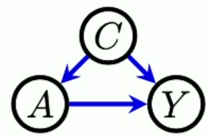
  - FFRCISTG (finest fully randomized causally interpretable structured tree graph):

    for $\vec{v} \in \mathfrak{X}_{\vec{V}}$, variables $V(\vec{v}_{\mathrm{pa}_{\mathcal{G}}(V)})$ are mutually independent $(\forall V \in \vec{V})$

- FFRCISTG is a historic name. I use 'multiple worlds model' for NPSEM-IE and 'single world model' for FFRCISTG.

Single World Versus Multiple Worlds Models

- Example:



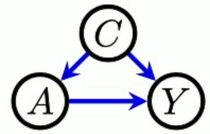
- FFRCISTG: $C \perp\!\!\!\perp A(c) \perp\!\!\!\perp Y(a,c)$ for all a, c.
- NPSEM-IE: $C \perp\!\!\!\perp A(c) \perp\!\!\!\perp Y(a,c')$ for all a, c, c'.
- NPSEM-IE is a strong model, e.g. is a submodel of the FFRCISTG.
- Observation 1: single graph may correspond to different models!
- Observation 2: unclear how to check if $A(c) \perp\!\!\!\perp Y(a,c')$ holds.
- Glossary: cross-world assumption: an assumption on counterfactuals that do not correspond to a single consistent assignment of interventions.

Identification Via The G-formula

- Given $\vec{A} \subseteq \vec{V}$, $p(\{Y(a) : Y \in \vec{V} \setminus \vec{A}\})$ is identified by the g-formula, a modified factorization of the DAG, as follows:

$$p(\{Y(a) : Y \in \vec{V} \setminus \vec{A}\}) = \prod_{Y \in \vec{V} \setminus \vec{A}} p(Y \mid \mathrm{pa}_{\mathcal{G}(Y)} \setminus \vec{A}, \vec{a}_{\vec{A} \cap \mathrm{pa}_{\mathcal{G}}(Y)}).$$

- Example:



$$\begin{aligned} p(C, A, Y) &= p(Y|A, C)p(A|C)p(C) \\ p(C, A(c), Y(c)) &= p(Y|A, c)p(A|c)p(C) \\ p(C, A, Y(a)) &= p(Y|a, C)p(A|C)p(C) \\ p(Y(a)) &= \sum_{C,A} p(Y|a, C)p(A|C)p(C) \\ &= \sum_{C} p(Y|a, C)p(C). \end{aligned}$$

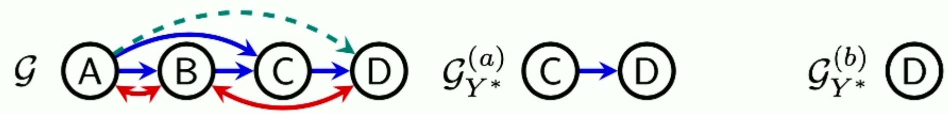
- Obvious corollary: a causal model of a DAG implies the statistical model of a DAG.

28/41

Central Analogy For The Remainder Of This Talk

- Fully observed model story:
  - Causal models imply statistical directed acyclic graph (DAG) models on the observed law.
  - Statistical DAG models admit factorizations.
  - Identification of counterfactual laws is via modified DAG factorizations (g-formula and friends).
- Hidden variable model story:
  - Causal models imply statistical acyclic directed mixed graph (ADMG) models on the observed law.
  - Statistical ADMG models admit factorizations.
  - Identification of counterfactual laws is via modified ADMG factorizations (ID algorithm and friends).

Examples Of Identification



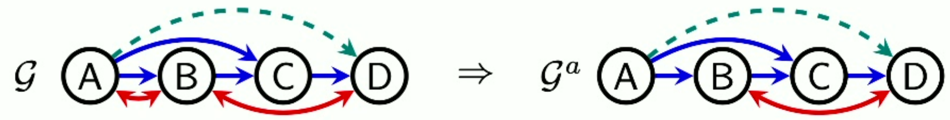
- Basic pieces: $\begin{cases} q_A(A) = p(A) \\ q_{A,B}(A,B) = p(B,A) \\ q_C(C|B,A) = p(C|B,A) \\ q_{B,D}(B,D,A|C) = p(D|C,B,A)p(B,A) \\ q_D(D|C,A) = \sum_B p(D|C,B,A)p(B|A) \end{cases}$
- Examples of identified counterfactual laws:

$$(a) : p(D(b,a)) = \sum_C q_C(C|a,b) q_D(D|C) = \sum_C p(C|a,b) \left(\sum_B p(D|C,B,A)p(B|A) \right)$$

$$(b) : p(D(c,a)) = q_D(D|c,a) = \sum_B p(D|c,B,a)p(B|a).$$

- Examples of non-identified counterfactual laws: $p(D(a))$.
- This is a complete procedure for any hidden variable model: failure implies non-identification (S and Pearl, 2006).

Example Of Maximum Likelihood Estimation



- Basic pieces: $\begin{cases} p(A;\eta_A) \\ p(B|A;\eta_{A,B}) \\ p(C|A,B;\eta_C) \\ q_{B,D}(B,D|A,C;\eta_{B,D}) = p(D|C,B,A)p(B|A) \\ q_D(D|C,A,\eta_D) = \sum_B p(D|C,B,A)p(B|A) \end{cases}$

$$p(D(b,a)) = \sum_C p(C|a,b)q_D(D|C) \Rightarrow \sum_C p(C|a,b;\widehat{\eta_C})q_D(D|C;\widehat{\eta_D})$$

$$p(D(c,a)) = q_D(D|c,a) \Rightarrow q_D(D|c,a;\widehat{\eta_{B,D}})$$

- Discrete data: parameters are tables, the parameter map is via a generalized Möbius transform.
- The multivariate normal nested Markov model of $\mathcal{G}$ is the linear SEM model for the arid projection graph $\mathcal{G}^a$ of $\mathcal{G}$ (S et al, 2018).
- Linear SEMs of arid graphs are everywhere identified (Drton et al).
- Gaussian nested likelihood in terms of linear SEM path coefficients.

Summary

- Important to distinguish statistical and causal graphical models. (The latter imply the former).
- Two complementary views of causal DAG models: structural equations and counterfactual random variables.
- A given DAG may correspond to multiple causal models.
- Causal effects are conceptualized as parameters in distributions defined over counterfactual r.v.s.
- The causal inference workflow is:
  - Posit a (causal) model. Or maybe learn it from data...
  - Formulate a parameter of interest.
  - Check if identified.
  - If identified, obtain an estimation strategy (maximum likelihood, etc.)
  - Quantify uncertainty (confidence intervals), sensitivity analysis.
- In fully observed DAGs identification is via the g-formula.
- In hidden variable DAGs, identification is not always possible, but is given by the ID algorithm if it is.
- Both the g-formula and the ID algorithm may be viewed as modified factorizations of a graphical model.

To Think About

- Quantum physics and causal inference have evolved in parallel, but considered similar topics.
- What can we do to accelerate progress?
  - Term glossary: interference, consistency, exclusion restrictions, faithfulness, etc.
  - Problems of common interest: model selection/compatibility, model descriptions/factorizations, others?
  - Bounds on non-identified effects?
- How to read each other's papers?
- Hoping to make progress at this event!

40/41

Thank you for listening!

Contact info:

Ilya Shpitser `ilyas@cs.jhu.edu`

41/41