Causal Inference Methods in Data Science
Lecture 6: More on nontrivial inequality constraints, causal compatibility, and quantum-classical gap in causality; Outlook after this course

Lin Liu

July 14, 2022

Announcement

Turn in the paper review before next Wednesday on canvas!

What causal inference is all about?

- Observational data, interventional data, background knowledge $\Rightarrow$ Causal graph/structure

What causal inference is all about?

- Observational data, interventional data, background knowledge $\Rightarrow$ Causal graph/structure

- Is a given causal query like $p(Y(a) = y)$ identifiable as a functional of the observed data? (Tian's ID algorithm, or IV, or proximal, or DiD, or synthetic control, or multiple treatments, or regression discontinuity design, or strong parametric modeling assumptions ...)

What causal inference is all about?

- Observational data, interventional data, background knowledge $\Rightarrow$ Causal graph/structure
- Is a given causal query like $p(Y(a) = y)$ identifiable as a functional of the observed data? (Tian's ID algorithm, or IV, or proximal, or DiD, or synthetic control, or multiple treatments, or regression discontinuity design, or strong parametric modeling assumptions ...)
- Statistics and computation: estimators, confidence intervals

What causal inference is all about?

- Observational data, interventional data, background knowledge ⇒ Causal graph/structure

- If a query is not identifiable, can we construct a nontrivial bound? Causal graph/structure implies (conditional) independences and nested (conditional) independences, which are **equality/algebraic constraints**, but how about **inequality/semi-algebraic constraints**?

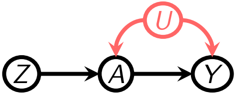
IV inequality

Assuming binary instrumental variable (IV) model:
$Z \in \{0, 1\}$: IV, $A \in \{0, 1\}$: Treatment, $Y \in \{0, 1\}$: outcome

IV inequality

Assuming binary instrumental variable (IV) model:
$Z \in \{0, 1\}$: IV, $A \in \{0, 1\}$: Treatment, $Y \in \{0, 1\}$: outcome

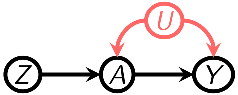
IV DAG:



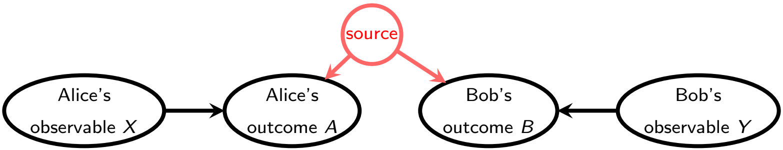
IV inequality

Assuming binary instrumental variable (IV) model:
$Z \in \{0,1\}$: IV, $A \in \{0,1\}$: Treatment, $Y \in \{0,1\}$: outcome

IV DAG:



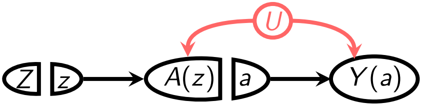
Bell-CHSH experiment:



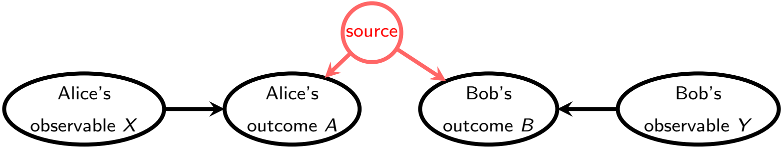
IV inequality

Assuming binary instrumental variable (IV) model:
$Z \in \{0,1\}$: IV, $A \in \{0,1\}$: Treatment, $Y \in \{0,1\}$: outcome

IV SWIG:



Bell-CHSH experiment:



IV inequality

- With IV DAG, it is possible to obtain non-trivial bounds for $\Pr(Y(1) = 1) - \Pr(Y(0) = 1)$,

IV inequality

- With IV DAG, it is possible to obtain non-trivial bounds for $\Pr(Y(1) = 1) - \Pr(Y(0) = 1)$,
- which suggests non-trivial inequalities imposed on the observed margin of IV DAG/SWIG:

Theorem 1 (Pearl UAI 1995)

When Z, A, Y have finite and discrete state spaces $\mathcal{Z}$, $\mathcal{A}$, $\mathcal{Y}$

$$\max_{a \in \mathcal{A}} \sum_{y \in \mathcal{Y}} \max_{z \in \mathcal{Z}} p(a, y|z) \leq 1$$

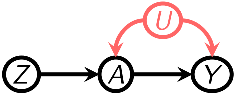
where $p(a, y|z) := \Pr(A = a, Y = y|Z = z)$

Implication: we can falsify IV assumptions by checking if IV inequality holds
But: Pearl's inequality is not sufficient except for binary IV models; see Bonet UAI 2001

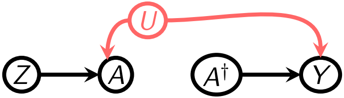
A "coupling" or "cloning" proof due to Robin Evans

Any discrete distribution $\mathbb{P}$ over (Z, A, Y) Markov to IV DAG $\Rightarrow$

For each $\mathfrak{a} \in \mathcal{A}$, able to construct a distribution $\mathbb{Q}_{\mathfrak{a}}$ such that $Y \perp\!\!\!\perp Z$ in $\mathbb{Q}_{\mathfrak{a}}$ and $q_{\mathfrak{a}}(\mathfrak{a}, y|z) = p(\mathfrak{a}, y|z)$ for every $y \in \mathcal{Y}$ and $z \in \mathcal{Z}$



$\mathbb{P}$

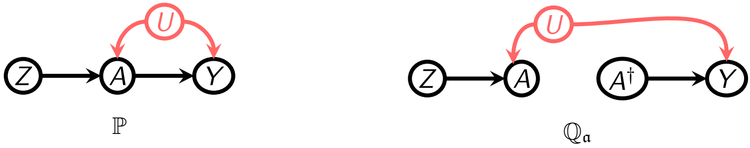


$\mathbb{Q}_{\mathfrak{a}}$

A "coupling" or "cloning" proof due to Robin Evans

Any discrete distribution $\mathbb{P}$ over (Z, A, Y) Markov to IV DAG $\Rightarrow$

For each $\mathfrak{a} \in \mathcal{A}$, able to construct a distribution $\mathbb{Q}_{\mathfrak{a}}$ such that $Y \perp\!\!\!\perp Z$ in $\mathbb{Q}_{\mathfrak{a}}$ and $q_{\mathfrak{a}}(\mathfrak{a}, y|z) = p(\mathfrak{a}, y|z)$ for every $y \in \mathcal{Y}$ and $z \in \mathcal{Z}$



Under $\mathbb{Q}_{\mathfrak{a}}$, $Y \perp\!\!\!\perp Z$ so for $z \neq z'$

$$q_{\mathfrak{a}}(y|z) = q_{\mathfrak{a}}(y|z') \Rightarrow q_{\mathfrak{a}}(y|z) = q_{\mathfrak{a}}(\mathfrak{a}, y|z') + \sum_{a \neq \mathfrak{a}} q_{\mathfrak{a}}(a, y|z')$$

$$\Rightarrow q_{\mathfrak{a}}(y|z) = p(\mathfrak{a}, y|z') + \sum_{a \neq \mathfrak{a}} q_{\mathfrak{a}}(a, y|z')$$

$$\Rightarrow \max_{z \in \mathcal{Z}} p(\mathfrak{a}, y|z) \leq q_{\mathfrak{a}}(y|z) \Rightarrow \max_{\mathfrak{a} \in \mathcal{A}} \sum_{y \in \mathcal{Y}} \max_{z \in \mathcal{Z}} p(\mathfrak{a}, y|z) \leq 1$$

Generalizing IV inequality 1 – More general models

Theorem 2 (Theorem 4.2 of Evans IEEE MLSP 2012)

For any disjoint sets $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}$ of observed vertices of a DAG $\mathcal{G}$, if $\boldsymbol{A}$ and $\boldsymbol{B}$ are d-separated by $\boldsymbol{C}$ in the corresponding SWIG by splitting node $\boldsymbol{D}$, then for any $\boldsymbol{D} = \mathfrak{d}$, there must exist a distribution $\mathbb{Q}_{\mathfrak{d}}$ under which $\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} | \boldsymbol{C}$, compatible with $p(\boldsymbol{a}, \boldsymbol{b}, \mathfrak{d} | \boldsymbol{c})$.

Under $\mathbb{Q}_{\mathfrak{d}}$, we can use $\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} | \boldsymbol{C}$ to derive non-trivial inequalities

Special case: $\boldsymbol{A} : Z, \boldsymbol{B} : Y, \boldsymbol{C} : \emptyset, \boldsymbol{D} : A$ in IV DAG

A sound criterion, but not necessarily complete (i.e. exhausting latent-variable DAGs with inequality constraints)

Generalizing IV inequality – Other related works

- Kédagni & Mourifié Biometrika 2020: discrete treatment but general IV and outcome

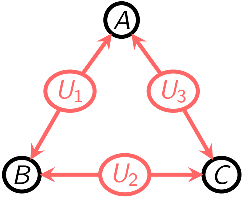
Generalizing IV inequality – Other related works

- Kédagni & Mourifié Biometrika 2020: discrete treatment but general IV and outcome
- Wang, Richardson, Robins, 2017 Biometrika: dealing with baseline covariates (using very basic convex program and analysis)

Generalizing IV inequality – Other related works

- Kédagni & Mourifié Biometrika 2020: discrete treatment but general IV and outcome
- Wang, Richardson, Robins, 2017 Biometrika: dealing with baseline covariates (using very basic convex program and analysis)
- Gunsilius 2020: impossibility results for continuous treatment
- $\cdots$

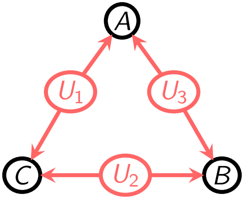
Beyond Bell-CHSH and IV models



"Triangle" model: possessing neither nontrivial (nested conditional) independences, nor inequality constraints using Evans' rule

But there are non-trivial inequality constraints!

Triangle model



"Triangle" model: possessing neither nontrivial (nested conditional) independences, nor inequality constraints using Evans' rule

In fact, it induces the following constraints:

$$\{p(A = C) - p(A \neq C)\} + \{p(B = C) - p(B \neq C)\}$$
$$\leq 1 + \{p(A = 1) - p(A = 0)\}\{p(B = 1) - p(B = 0)\}$$

together with its symmetrized version (over permutations of A, B, C)

The inflation technique: Complete algorithm for finding inequality constraints

- Theoretical physicists Elie Wolfe and colleagues recently solved the completeness problem using an approach generalizing Robin Evans' node-splitting idea, called "the inflation technique"

The inflation technique: Complete algorithm for finding inequality constraints

- Theoretical physicists Elie Wolfe and colleagues recently solved the completeness problem using an approach generalizing Robin Evans' node-splitting idea, called "the inflation technique"

- They initially tried to solve the causal compatibility problem i.e. check if the observed data distribution can be generated from a given latent-variable causal DAG (that is, if violating the necessary conditions implied by a latent-variable causal DAG)

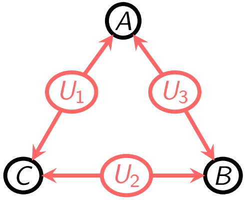
The inflation technique: Complete algorithm for finding inequality constraints

- Theoretical physicists Elie Wolfe and colleagues recently solved the completeness problem using an approach generalizing Robin Evans' node-splitting idea, called "the inflation technique"

- They initially tried to solve the causal compatibility problem i.e. check if the observed data distribution can be generated from a given latent-variable causal DAG (that is, if violating the necessary conditions implied by a latent-variable causal DAG)

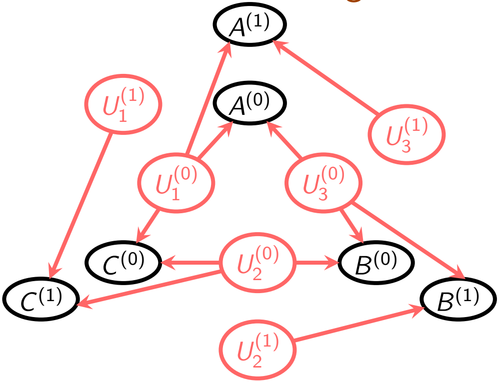
- Opposite to deriving equality (algebraic) constraints for latent-variable causal DAGs by latent projection to obtain ADMG, one can instead augment the graph by introducing copies of observables and latents, but also preserving ancestral relations!

Inflation graph for Triangle model: purely spurious correlation setting



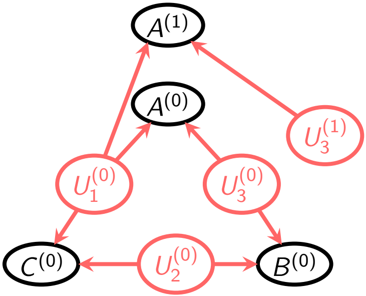
"Triangle" model: possessing neither nontrivial (nested conditional) independences, nor inequality constraints using Evans' rule

Inflation graph for Triangle model: purely spurious correlation setting



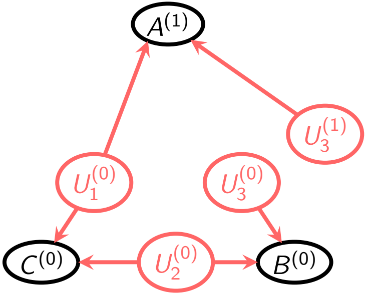
"Spiral" inflation of the "Triangle" model

Inflation graph for Triangle model: purely spurious correlation setting



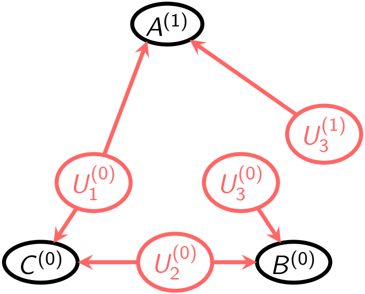
"Capped" inflation of the "Triangle" model

Inflation graph for Triangle model: purely spurious correlation setting



"Cut" inflation of the "Triangle" model

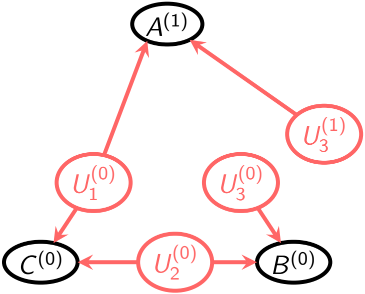
Inflation graph for Triangle model: purely spurious correlation setting



"Cut" inflation of the "Triangle" model

Finding "injectable sets": the subset C' of vertices in the inflated graph $\mathcal{H}$ such that their corresponding original vertex sets C in the original DAG $\mathcal{G}$ sat. $\mathcal{H}_{\text{an}_{\mathcal{H}}(C')} = \mathcal{G}_{\text{an}_{\mathcal{G}}(C)}$

Inflation graph for Triangle model: purely spurious correlation setting



"Cut" inflation of the "Triangle" model

Finding "injectable sets": the subset C' of vertices in the inflated graph $\mathcal{H}$ such that their corresponding original vertex sets C in the original DAG $\mathcal{G}$ sat. $\mathcal{H}_{\text{an}_{\mathcal{H}}(C')} = \mathcal{G}_{\text{an}_{\mathcal{G}}(C)}$

If the subset C' of vertices in the inflated graph $\mathcal{H}$ is injectable, then $\mathbb{P}_{C'} = \mathbb{P}_C$

Trivial inequalities on inflated graph $\Rightarrow$ Non-trivial inequalities on original graph

- For Triangle model, three observed vertices A, B, C

Trivial inequalities on inflated graph $\Rightarrow$ Non-trivial inequalities on original graph

- For Triangle model, three observed vertices A, B, C
- For Boolean random variables, we have trivial inequality (systematic proof by LP):

$$\begin{aligned} &\{\Pr(A = C) - \Pr(A \neq C)\} + \{\Pr(B = C) - \Pr(B \neq C)\} \\ \leq\ & 1 + \{\Pr(A = B) - \Pr(A \neq B)\} \end{aligned}$$

Trivial inequalities on inflated graph $\Rightarrow$ Non-trivial inequalities on original graph

- For Triangle model, three observed vertices A, B, C
- For Boolean random variables, we have trivial inequality (systematic proof by LP):

$$\begin{aligned} &\{\Pr(A = C) - \Pr(A \neq C)\} + \{\Pr(B = C) - \Pr(B \neq C)\} \\ \leq\ &1 + \{\Pr(A = B) - \Pr(A \neq B)\} \end{aligned}$$

- Obviously for "Cut" inflated graph $\mathcal{H}$, we have

$$\begin{aligned} &\{\Pr(A^{(1)} = C^{(0)}) - \Pr(A^{(1)} \neq C^{(0)})\} + \{\Pr(B^{(0)} = C^{(0)}) - \Pr(B^{(0)} \neq C^{(0)})\} \\ \leq\ &1 + \{\Pr(A^{(1)} = B^{(0)}) - \Pr(A^{(1)} \neq B^{(0)})\} \end{aligned}$$

Trivial inequalities on inflated graph $\Rightarrow$ Non-trivial inequalities on original graph

- For Triangle model, three observed vertices A, B, C
- For Boolean random variables, we have trivial inequality (systematic proof by LP):

$$\begin{aligned}&\{\Pr(A = C) - \Pr(A \neq C)\} + \{\Pr(B = C) - \Pr(B \neq C)\} \\ \leq\ & 1 + \{\Pr(A = B) - \Pr(A \neq B)\}\end{aligned}$$

- Obviously for "Cut" inflated graph $\mathcal{H}$, we have

$$\begin{aligned}&\{\Pr(A^{(1)} = C^{(0)}) - \Pr(A^{(1)} \neq C^{(0)})\} + \{\Pr(B^{(0)} = C^{(0)}) - \Pr(B^{(0)} \neq C^{(0)})\} \\ \leq\ & 1 + \{\Pr(A^{(1)} = B^{(0)}) - \Pr(A^{(1)} \neq B^{(0)})\}\end{aligned}$$

- But importantly, $A^{(1)} \perp\!\!\!\perp B^{(0)}$ in $\mathcal{H}$, so

$$\begin{aligned}&\{\Pr(A^{(1)} = C^{(0)}) - \Pr(A^{(1)} \neq C^{(0)})\} + \{\Pr(B^{(0)} = C^{(0)}) - \Pr(B^{(0)} \neq C^{(0)})\} \\ \leq\ & 1 + \{\Pr(A^{(1)} = 1) - \Pr(A^{(1)} = 0)\}\{\Pr(B^{(0)} = 1) - \Pr(B^{(0)} = 0)\}\end{aligned}$$

Trivial inequalities on inflated graph $\Rightarrow$ Non-trivial inequalities on original graph

- For Triangle model, three observed vertices A, B, C
- For Boolean random variables, we have trivial inequality (systematic proof by LP):

$$\begin{aligned} &\{\Pr(A = C) - \Pr(A \neq C)\} + \{\Pr(B = C) - \Pr(B \neq C)\} \\ \leq\ &1 + \{\Pr(A = B) - \Pr(A \neq B)\} \end{aligned}$$

- Obviously for "Cut" inflated graph $\mathcal{H}$, we have

$$\begin{aligned} &\{\Pr(A^{(1)} = C^{(0)}) - \Pr(A^{(1)} \neq C^{(0)})\} + \{\Pr(B^{(0)} = C^{(0)}) - \Pr(B^{(0)} \neq C^{(0)})\} \\ \leq\ &1 + \{\Pr(A^{(1)} = B^{(0)}) - \Pr(A^{(1)} \neq B^{(0)})\} \end{aligned}$$

- Since $\{A^{(1)}, C^{(0)}\}, \{B^{(0)}, C^{(0)}\}, \{A^{(1)}\}, \{B^{(0)}\}$ are injectable in Cut graph: non-trivial inequality

$$\begin{aligned} &\{\Pr(A = C) - \Pr(A \neq C)\} + \{\Pr(B = C) - \Pr(B \neq C)\} \\ \leq\ &1 + \{\Pr(A = 1) - \Pr(A = 0)\}\{\Pr(B = 1) - \Pr(B = 0)\} \end{aligned}$$

Generally speaking: for arbitrary pure-spurious correlation settings

- Draw inflation graphs

Generally speaking: for arbitrary pure-spurious correlation settings

- Draw inflation graphs
- Identify the "AI-expressible sets": sets that are disjoint union of d-separated injectable sets

Generally speaking: for arbitrary pure-spurious correlation settings

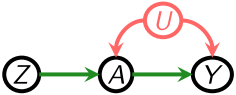
- Draw inflation graphs
- Identify the "AI-expressible sets": sets that are disjoint union of d-separated injectable sets
- Find trivial inequalities satisfied by AI-expressible sets (often by LP) but leverage the d-separations

Generally speaking: for arbitrary pure-spurious correlation settings

- Draw inflation graphs
- Identify the "AI-expressible sets": sets that are disjoint union of d-separated injectable sets
- Find trivial inequalities satisfied by AI-expressible sets (often by LP) but leverage the d-separations
- Map back to the original graph as nontrivial inequality

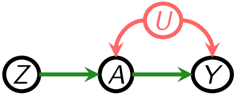
IV model using inflation?

- There are non-spurious correlations among observed vertices for IV model



IV model using inflation?

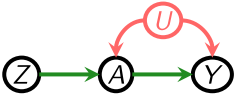
- There are non-spurious correlations among observed vertices for IV model



- Inflation? Only introduced for purely spurious correlation scenarios

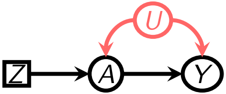
IV model using inflation?

- There are non-spurious correlations among observed vertices for IV model



- Inflation? Only introduced for purely spurious correlation scenarios

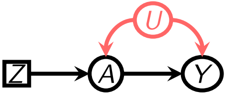
- Fortunately, for all exogenous vertices, we could fix them without changing the essential structure of the latent variable DAG



so we can look at the conditional distributions $p(A, Y|Z = z)$

IV model using inflation?

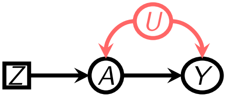
- Fortunately, for all exogenous vertices, we could fix them without changing the essential structure of the latent variable DAG



so we can look at the conditional distributions $p(A, Y | Z = z)$

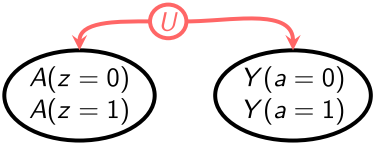
IV model using inflation?

- Fortunately, for all exogenous vertices, we could fix them without changing the essential structure of the latent variable DAG



so we can look at the conditional distributions $p(A, Y|Z = z)$

- Then **unpacking** to eliminate all dependencies among observed vertices

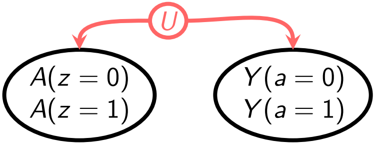


which becomes a purely-correlated scenario after **unpacking** satisfying

$$p(A = a, Y = y|Z = z) = p_{\text{unpacking}}(A(z) = a, Y(a) = y), a, z \in \{0, 1\}^2$$

IV model using inflation?

- Then **unpacking** to eliminate all dependencies among observed vertices

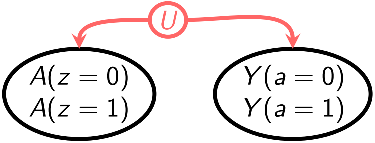


which becomes a purely-correlated scenario after **unpacking** satisfying

$$p(A=a, Y=y|Z=z) = p_{\text{unpack}}(A(z)=a, Y(a)=y), a, z \in \{0,1\}^2$$

IV model using inflation?

- Then **unpacking** to eliminate all dependencies among observed vertices



which becomes a purely-correlated scenario after **unpacking** satisfying

$$p(A = a, Y = y|Z = z) = p_{\text{unpack}}(A(z) = a, Y(a) = y), a, z \in \{0, 1\}^2$$

- Deriving IV inequality? Now easy without even resorting to the inflation technique! See R code "bound.R"

$$\max_a \sum_y \max_z p(A = a, Y = y|Z = z)$$

$$\text{s.t.} \quad \sum_{a_0, a_1, y_0, y_1} p_{\text{unpack}}(A(z = 0) = a_0, A(z = 1) = a_1, Y(a = 0) = y_0, Y(a = 1) = y_1) = 1$$

$$\ldots$$

Completeness?

- The inflation technique is complete, but in the following weak sense: For arbitrary pure-spurious correlation models with L latent variables, there exists an inflation hierarchy indexed by an order m, when $m \to \infty$, if P is not compatible with the causal graph $\mathcal{G}$, then the m-th order inflation can witness it if

$$\inf_{Q \sim \mathcal{G}} \|P - Q\| \gtrsim \sqrt{\frac{L}{m}}$$

Completeness?

- The inflation technique is complete, but in the following weak sense: For arbitrary pure-spurious correlation models with L latent variables, there exists an inflation hierarchy indexed by an order m, when $m \to \infty$, if P is not compatible with the causal graph $\mathcal{G}$, then the m-th order inflation can witness it if

$$\inf_{Q \sim \mathcal{G}} \|P - Q\| \gtrsim \sqrt{\frac{L}{m}}$$

- m-th order inflation?? Clone each latent variable m times and then clone each observed variable L_v^m where L_v is the number of latent variables influencing the observed variable v

Completeness?

- The inflation technique is complete, but in the following weak sense: For arbitrary pure-spurious correlation models with L latent variables, there exists an inflation hierarchy indexed by an order m, when $m \to \infty$, if P is not compatible with the causal graph $\mathcal{G}$, then the m-th order inflation can witness it if

$$\inf_{Q \sim \mathcal{G}} \|P - Q\| \gtrsim \sqrt{\frac{L}{m}}$$

- m-th order inflation?? Clone each latent variable m times and then clone each observed variable L_v^m where L_v is the number of latent variables influencing the observed variable v

- Overkill? Probably yes! For the following trivial case: $U \to A$ with $P(A=1)P(A=0) \leq \frac{1}{4}$, solving m-th order inflation gives us

$$\frac{1}{2}\frac{2m}{4m-2} > \frac{1}{4}$$

Summary

- Inequality constraints induced by causal models are difficult to exhaust

Summary

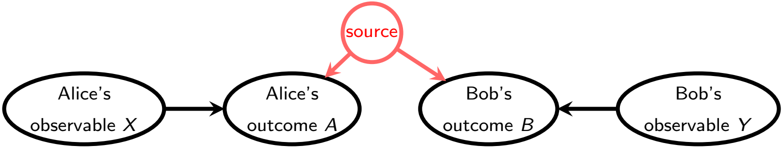
- Inequality constraints induced by causal models are difficult to exhaust
- Inflation techniques are the only approach working for arbitrary causal models in asymptopia

Summary

- Inequality constraints induced by causal models are difficult to exhaust
- Inflation techniques are the only approach working for arbitrary causal models in asymptopia
- Open problem: Could there be a short-cut?

classical-quantum gap?

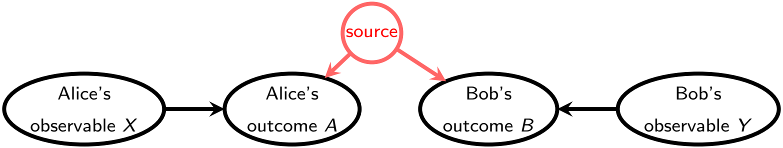
Earliest instance: Tsirelson bound



Bell-CHSH inequality falsifies the hidden-variable locality hypothesis predicted by classical physics; yet it relies on the "freedom of choice" hypothesis (exogeneity of Alice's and Bob's choice of observable); see Chaves et al. 2021

Quantum causal model (at least as formulated in the current literature): latent variables are quantum channels and observed variables are measuring quantum variables (so upon measurement, collapse)

Earliest instance: Tsirelson bound

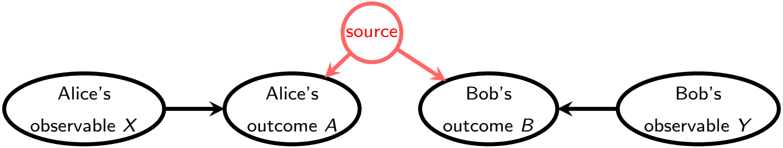


Classical case: For simplicity assume both observable and outcome are $\{-1, +1\}$-valued then (immediate by unpacking, see R code "bound.R")

$$\mathbb{E}[AB|X = Y = 0] + \mathbb{E}[AB|X = 0, Y = 1] + \mathbb{E}[AB|X = 1, Y = 0] - \mathbb{E}[AB|X = Y = 1] \leq 2$$

This is the form of CHSH

Earliest instance: Tsirelson bound



Quantum case: qubits are prepared in the following entangled states in the order of Alice and Bob

$$|\psi\rangle = \frac{1}{\sqrt{2}}\underbrace{|0\rangle\otimes|1\rangle}_{|\psi_1\rangle} - \frac{1}{\sqrt{2}}\underbrace{|1\rangle\otimes|0\rangle}_{|\psi_2\rangle} = \frac{1}{\sqrt{2}}\binom{0}{1}\otimes\binom{1}{0} - \frac{1}{\sqrt{2}}\binom{1}{0}\otimes\binom{0}{1}$$

Alice and Bob measure the outcome with different unitary matrices

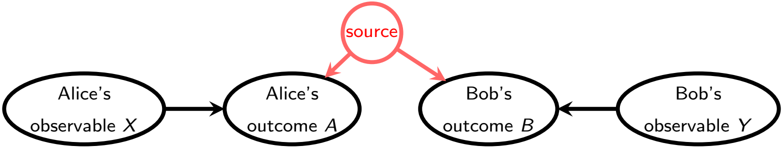
$$\text{Alice: } \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}}_{X=0} \text{ or } \underbrace{\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}}_{X=1} \text{ Bob: } \underbrace{\begin{pmatrix} -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}}_{Y=0} \text{ or } \underbrace{\begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}}_{Y=1}$$

Then $|\psi\rangle$ induces a density matrix (p.s.d.) $\varrho = \frac{1}{2}|\psi_1\rangle\langle\psi_1| + \frac{1}{2}|\psi_2\rangle\langle\psi_2|$, which further gives expectation of observable $C = A \otimes B$ to be $\mathrm{Tr}(\varrho C)$. Following commutativity of trace

$$\langle\psi|A_0\otimes B_0|\psi\rangle = \langle\psi|A_0\otimes B_1|\psi\rangle = \langle\psi|A_1\otimes B_0|\psi\rangle = \frac{1}{\sqrt{2}}, \langle\psi|A_1\otimes B_1|\psi\rangle = -\frac{1}{\sqrt{2}}$$

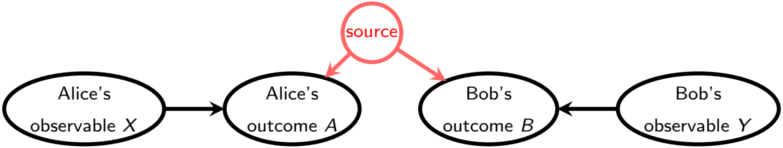
$$\Rightarrow \langle\psi|A_0\otimes B_0|\psi\rangle + \langle\psi|A_0\otimes B_1|\psi\rangle + \langle\psi|A_1\otimes B_0|\psi\rangle - \langle\psi|A_1\otimes B_1|\psi\rangle = 2\sqrt{2} > 2$$

Earliest instance: Tsirelson bound



Any guess on how to prove the quantum CHSH bound $2\sqrt{2}$ in a more first-principled way like LP? See “bound.R”

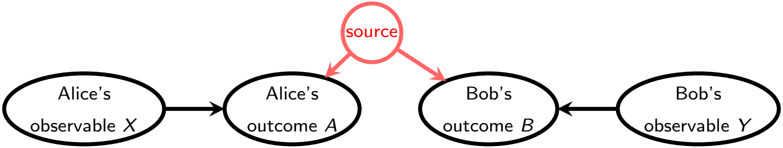
Earliest instance: Tsirelson bound



Any guess on how to prove the quantum CHSH bound $2\sqrt{2}$ in a more first-principled way like LP? See "bound.R"

SemiDefinite Programming (SDP)! See Wehner 2006 PRA

Earliest instance: Tsirelson bound



Any guess on how to prove the quantum CHSH bound $2\sqrt{2}$ in a more first-principled way like LP? See "bound.R"

SemiDefinite Programming (SDP)! See Wehner 2006 PRA

But in general a polynomial programming problem so quantum causal compatibility is computationally hard! Exists SDP relaxation to solve it approximately

Looking ahead

What could be a good attitude towards causal inference?

I quote the following paragraph written by Ogburn, Shpitser, Tchetgen Tchetgen 2019 (a very angry rant because they were super annoyed by a paper written by famous machine learning researcher David Blei):

> *One of the most important roles of causal inference in statistics and data science is to be transparent about the strong, usually untestable assumptions under which causal inference is possible (Pearl, 2000; Robins, 2001). The burden for transparency about assumptions is arguably greater in causal inference than in other areas of statistics, because it is crucial that scientists and consumers of research, e.g. policy makers or doctors, have the tools to reason about whether an association is in fact causal...*
>
> *but unfortunately – we wish this were not the case! – it is impossible to identify causal effects in the presence of unmeasured confounding with nonparametric or empirically verifiable assumptions.*

What could be a good attitude towards causal inference?

Sounds hopeless? But before 1990's, causality is not even allowed to be studied in statistics

If you feel like having better ideas than what people have had, it could be a great thing! But do read a lot of papers before drawing the conclusion that some idea has not been done before (a lesson of David Blei and colleagues)

If you feel like the current methodology cannot solve your applied problem, it could be even a greater thing!

Important aspects not yet covered in this course

- Randomization-based statistical inference

Important aspects not yet covered in this course

- Randomization-based statistical inference
- Interference and peer effect

Important aspects not yet covered in this course

- Randomization-based statistical inference
- Interference and peer effect
- Study design: optimal design, rerandomization, design under interference, large-scale (dynamic) A/B testing

Important aspects not yet covered in this course

- Randomization-based statistical inference
- Interference and peer effect
- Study design: optimal design, rerandomization, design under interference, large-scale (dynamic) A/B testing
- Causal inference with complex data structures: Time series data, Functional data, Data that arises from dynamical systems etc.

Important aspects not yet covered in this course

- Randomization-based statistical inference
- Interference and peer effect
- Study design: optimal design, rerandomization, design under interference, large-scale (dynamic) A/B testing
- Causal inference with complex data structures: Time series data, Functional data, Data that arises from dynamical systems etc.
- Causal invariance from multiple environments

Important aspects not yet covered in this course

- Randomization-based statistical inference
- Interference and peer effect
- Study design: optimal design, rerandomization, design under interference, large-scale (dynamic) A/B testing
- Causal inference with complex data structures: Time series data, Functional data, Data that arises from dynamical systems etc.
- Causal invariance from multiple environments
- How to combine real world evidence with RCT evidence

Important aspects not yet covered in this course

- Randomization-based statistical inference
- Interference and peer effect
- Study design: optimal design, rerandomization, design under interference, large-scale (dynamic) A/B testing
- Causal inference with complex data structures: Time series data, Functional data, Data that arises from dynamical systems etc.
- Causal invariance from multiple environments
- How to combine real world evidence with RCT evidence
- AI system with causality component (covered to some extent in Kun Kuang's guest lecture)

Randomization inference

- Most cases, we view data as a random draw from some unknown probability distribution $\mathbb{P}$

Randomization inference

- Most cases, we view data as a random draw from some unknown probability distribution $\mathbb{P}$

- But in well-designed studies, we can take a different view

Randomization inference

- Most cases, we view data as a random draw from some unknown probability distribution $\mathbb{P}$

- But in well-designed studies, we can take a different view

- For example, if we conduct a randomized trial on an exercise incentive program for college students in China, then the super-population is a finite big sample N where N is the total number of college students in China

Randomization inference

- Most cases, we view data as a random draw from some unknown probability distribution $\mathbb{P}$
- But in well-designed studies, we can take a different view
- For example, if we conduct a randomized trial on an exercise incentive program for college students in China, then the super-population is a finite big sample N where N is the total number of college students in China
- The subset of students that we conduct trial on are a random draw of size n from all N students with $n < N$

Randomization inference

- Most cases, we view data as a random draw from some unknown probability distribution $\mathbb{P}$
- But in well-designed studies, we can take a different view
- For example, if we conduct a randomized trial on an exercise incentive program for college students in China, then the super-population is a finite big sample N where N is the total number of college students in China
- The subset of students that we conduct trial on are a random draw of size n from all N students with $n < N$
- No modeling/distributional assumption on the data at all! All the randomness comes from the sampling process

Randomization inference

- Most cases, we view data as a random draw from some unknown probability distribution $\mathbb{P}$
- But in well-designed studies, we can take a different view
- For example, if we conduct a randomized trial on an exercise incentive program for college students in China, then the super-population is a finite big sample N where N is the total number of college students in China
- The subset of students that we conduct trial on are a random draw of size n from all N students with $n < N$
- No modeling/distributional assumption on the data at all! All the randomness comes from the sampling process
- What is the estimand? What is the estimator? How to show the estimator is good without a reference true distribution $\mathbb{P}$?

Interference

- Renew potential outcomes/counterfactuals by dropping SUTVA

$$Y_i(a_1, \cdots, a_n), i = 1, \cdots, n$$

Interference

- Renew potential outcomes/counterfactuals by dropping SUTVA

$$Y_i(a_1, \cdots, a_n), i = 1, \cdots, n$$

- Even if you can run an RCT, how do you design randomization strategy to preserve the network structure?

Interference

- Renew potential outcomes/counterfactuals by dropping SUTVA

$$Y_i(a_1, \cdots, a_n), i = 1, \cdots, n$$

- Even if you can run an RCT, how do you design randomization strategy to preserve the network structure?

- How to make sense the effect of my own treatment vs. the effect of others' treatment?

Interference

- Renew potential outcomes/counterfactuals by dropping SUTVA

$$Y_i(a_1, \cdots, a_n), i = 1, \cdots, n$$

- Even if you can run an RCT, how do you design randomization strategy to preserve the network structure?

- How to make sense the effect of my own treatment vs. the effect of others' treatment?

- What is the limit? How dense the network can be so that we can still learn something? What if the network is unknown?

RWE

- In exercises, we have seen a made-up scenario where trial might not be ideal for decision making but how do you optimally combine RWE with trial data?

RWE

- In exercises, we have seen a made-up scenario where trial might not be ideal for decision making but how do you optimally combine RWE with trial data?

- In complex and dynamic scenario, how to combine RWE and trial?

Causal invariance

- Initiated in the seminal work Peters, Bühlmann, Meinshausen 2016 JRSS-B (also read the Discussions and Rejoinder)

Causal invariance

- Initiated in the seminal work Peters, Bühlmann, Meinshausen 2016 JRSS-B (also read the Discussions and Rejoinder)

- Axiom: "causal relationship is invariant/stable across different environments"

Causal invariance

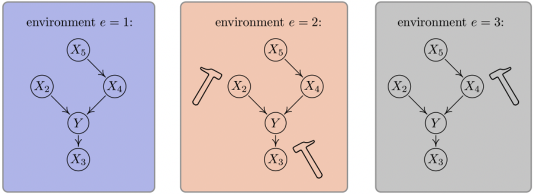
- Initiated in the seminal work Peters, Bühlmann, Meinshausen 2016 JRSS-B (also read the Discussions and Rejoinder)

- Axiom: "causal relationship is invariant/stable across different environments"

- Related to data combination (many different trials + many different observational studies, or more precisely many different interventions + many different passively observed data, e.g. large-scale CRISPR gene-knockdown experiments are a typical instance)

Causal invariance

- Initiated in the seminal work Peters, Bühlmann, Meinshausen 2016 JRSS-B (also read the Discussions and Rejoinder)
- Axiom: "causal relationship is invariant/stable across different environments"
- Math definition: If a subset $S \subseteq V$ is causal for a response Y, then for all different data distributions (environments) $e \in \mathcal{E}$ (a family of environments), $Y^e = g(V_S^e, u^e)$ with g invariant across different environments $\mathcal{E}$ and $u^e \sim F_u$ with F_u invariant across $\mathcal{E}$



The End! I covered quite a lot of grounds of causality. Hope Ya'll learnt something useful here and there!