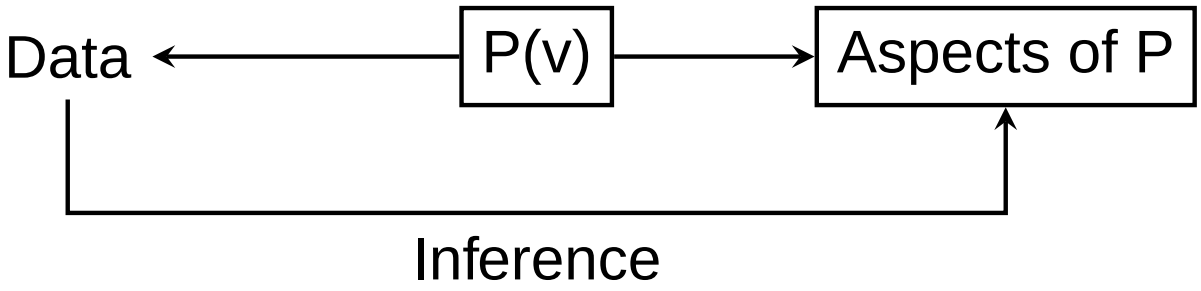
Causal Inference Tutorial (part 1) Graphical Causal Models and Effects of Interventions

Ilya Shpitser
ishpitse@hsph.harvard.edu

Causal Inference Group, Department of Epidemiology
Harvard School of Public Health

July 19, 2011

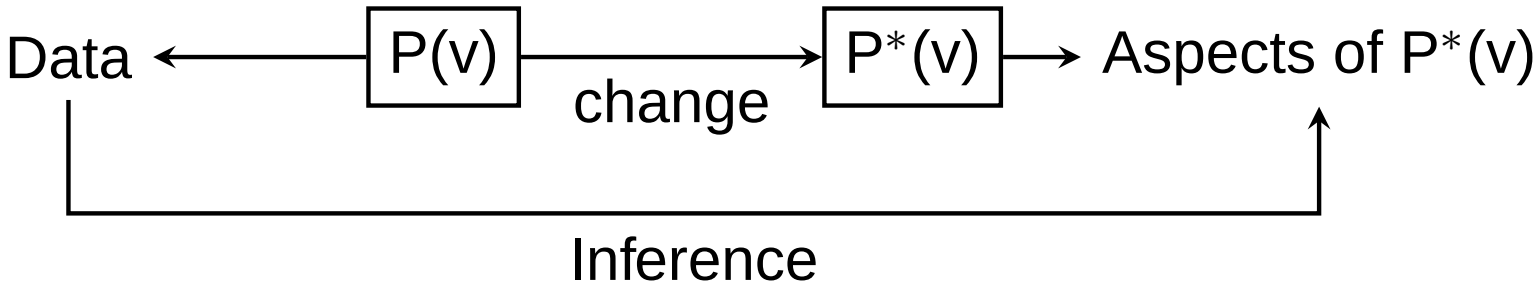
Statistical Inference: an Overview



- Classification problems, probability of X given evidence e: $P(x|e)$
- Disease correlates (for diagnosis), data mining, intelligent agents, etc.
- The Netflix problem (million dollar problem!)

Causal Inference: an Overview

Causal inference is reasoning about change



What happens when P(v) changes due to “outside forces” or experimentation?

- key question: what changes, and what stays invariant.
- causes of disease (etiology), treatment effects, gene regulation, scientific theories (econometrics, social science), etc.

Why Should We Care About Causal Inference?

- “I would rather discover one causal law than be king of Persia!” – Democritus
- Human beings understand the world in terms of causes and effects
- There is *consensus* on the meaning of causal statements
- Empirical science is about establishing cause
- Causal inference gives a mathematical language for causal statements, and tools to solve causal problems formally

Why Should We Care About Causal Inference?

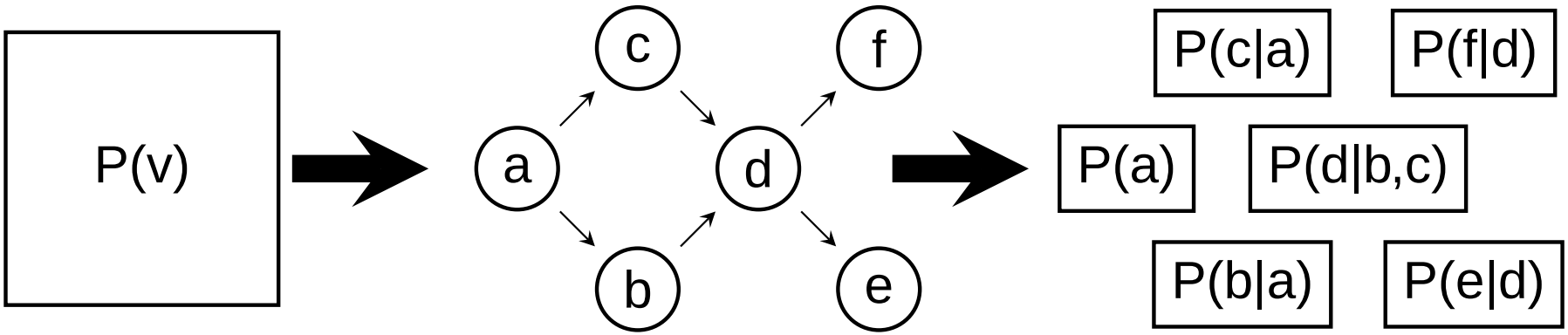
- “I would rather discover one causal law than be king of Persia!” – Democritus
- Human beings understand the world in terms of causes and effects
- There is *consensus* on the meaning of causal statements
- Empirical science is about establishing cause
- Causal inference gives a mathematical language for causal statements, and tools to solve causal problems formally

Why Should We Care About Graphical Models?

- Computing things from $P(v)$ can be difficult, because $P(v)$ can be large: in a binary model if |V| = n, |P(v)| = 2^n
- Graphical models tame this complexity with conditional independence constraints (tractable learning/inference)
- Graphical models give a visual interpretation for dependence/independence
- Graphical models admit a *causal* interpretation

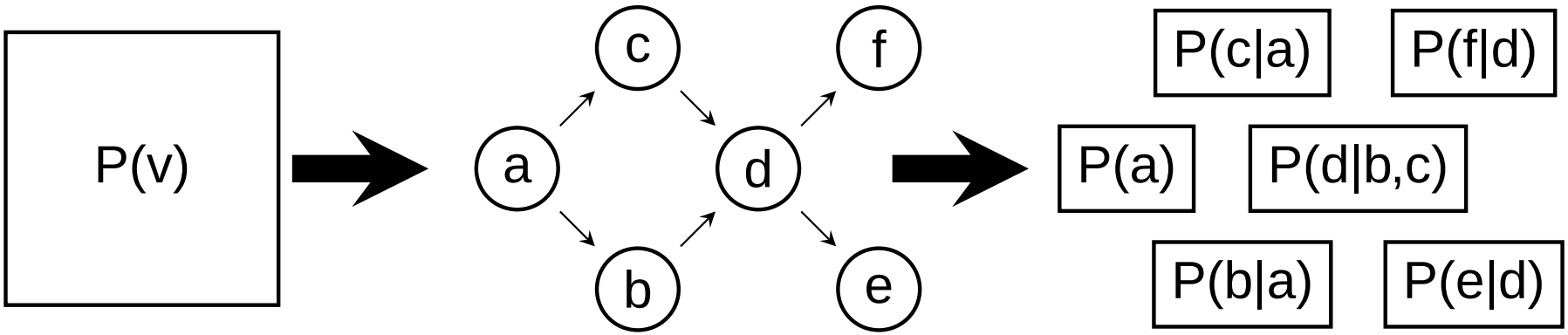
Why Should We Care About Graphical Models?

- Computing things from $P(v)$ can be difficult, because $P(v)$ can be large: in a binary model if |V| = n, |P(v)| = 2^n
- Graphical models tame this complexity with conditional independence constraints (tractable learning/inference)
- Graphical models give a visual interpretation for dependence/independence
- Graphical models admit a *causal* interpretation



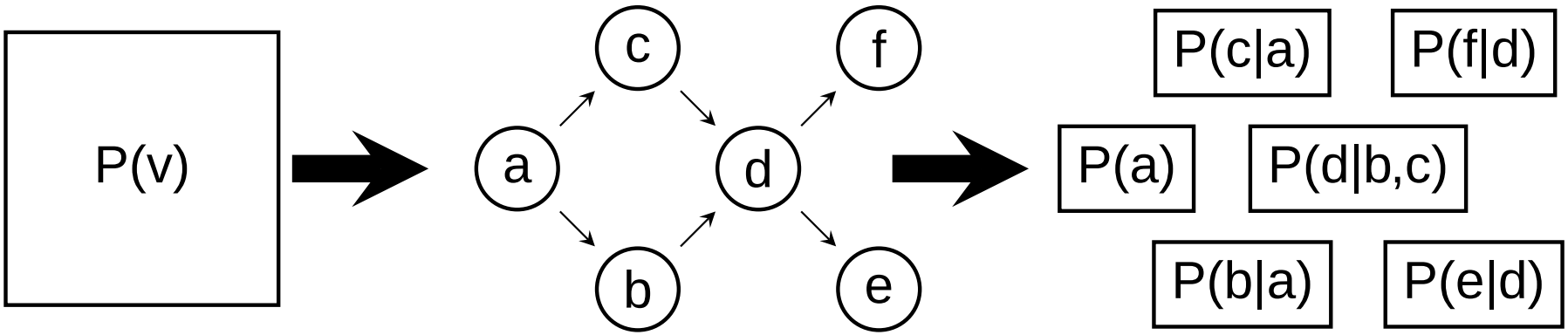
Why Should We Care About Graphical Models?

- Computing things from $P(v)$ can be difficult, because $P(v)$ can be large: in a binary model if |V| = n, |P(v)| = 2^n
- Graphical models tame this complexity with conditional independence constraints (tractable learning/inference)
- Graphical models give a visual interpretation for dependence/independence
- Graphical models admit a *causal* interpretation



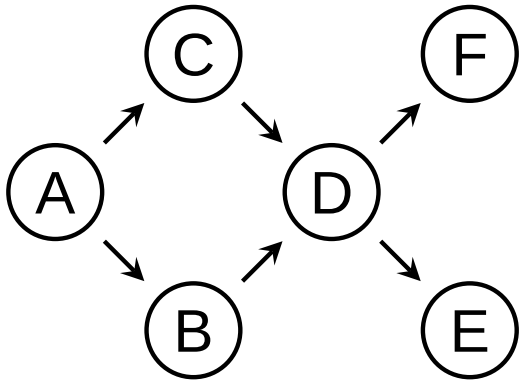
Why Should We Care About Graphical Models?

- Computing things from $P(v)$ can be difficult, because $P(v)$ can be large: in a binary model if |V| = n, |P(v)| = 2^n
- Graphical models tame this complexity with conditional independence constraints (tractable learning/inference)
- Graphical models give a visual interpretation for dependence/independence
- Graphical models admit a *causal* interpretation



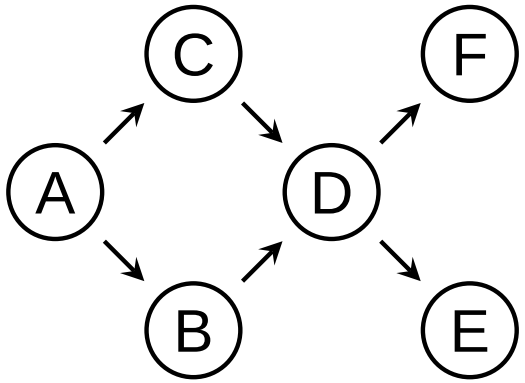
Graphical Models

- $P(v) = \prod_i P(x_i | parents(x_i))$
- $P(a, b, c, d, e, f) = P(a)P(b|a)P(c|a)P(d|b, c)P(e|d)P(f|d)$
- Conditional independence constraints: $(X_i \perp\!\!\!\perp \text{Non-descendants}(X_i) \mid \text{Parents}(X_i))$
- Other conditional independences implied by this, can be read off via d-separation



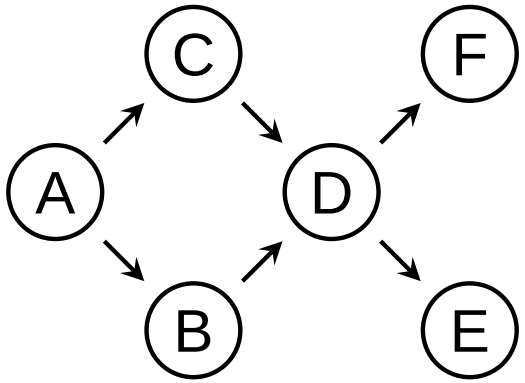
Graphical Models

- $P(v) = \prod_i P(x_i | parents(x_i))$
- $P(a, b, c, d, e, f) = P(a)P(b|a)P(c|a)P(d|b, c)P(e|d)P(f|d)$
- Conditional independence constraints: $(X_i \perp\!\!\!\perp \text{Non-descendants}(X_i) \mid \text{Parents}(X_i))$
- Other conditional independences implied by this, can be read off via d-separation



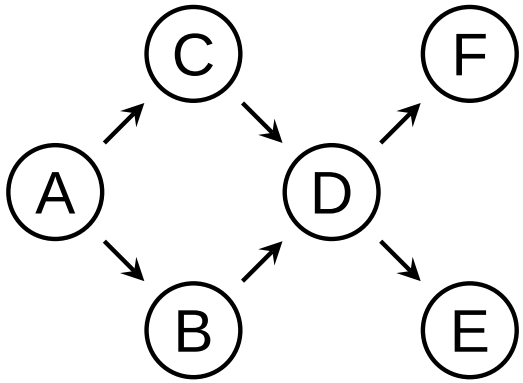
Graphical Models

- $P(v) = \prod_i P(x_i | parents(x_i))$
- $P(a, b, c, d, e, f) = P(a)P(b|a)P(c|a)P(d|b, c)P(e|d)P(f|d)$
- Conditional independence constraints: $(X_i \perp\!\!\!\perp \text{Non-descendants}(X_i) \mid \text{Parents}(X_i))$
- Other conditional independences implied by this, can be read off via d-separation



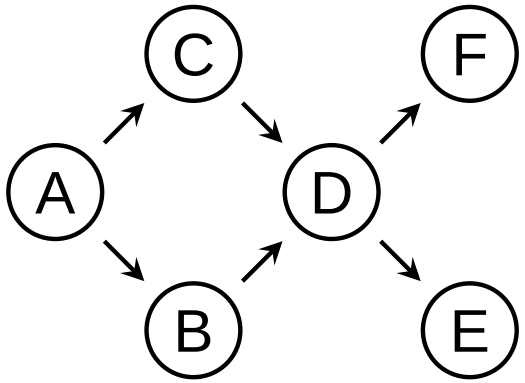
Graphical Models

- $P(v) = \prod_i P(x_i | parents(x_i))$
- $P(a, b, c, d, e, f) = P(a)P(b|a)P(c|a)P(d|b, c)P(e|d)P(f|d)$
- Conditional independence constraints: $(X_i \perp\!\!\!\perp \text{Non-descendants}(X_i) \mid \text{Parents}(X_i))$
- Other conditional independences implied by this, can be read off via d-separation



Graphical Models

- $P(v) = \prod_i P(x_i | parents(x_i))$
- $P(a, b, c, d, e, f) = P(a)P(b|a)P(c|a)P(d|b, c)P(e|d)P(f|d)$
- Conditional independence constraints: $(X_i \perp\!\!\!\perp \text{Non-descendants}(X_i) \mid \text{Parents}(X_i))$
- Other conditional independences implied by this, can be read off via d-separation

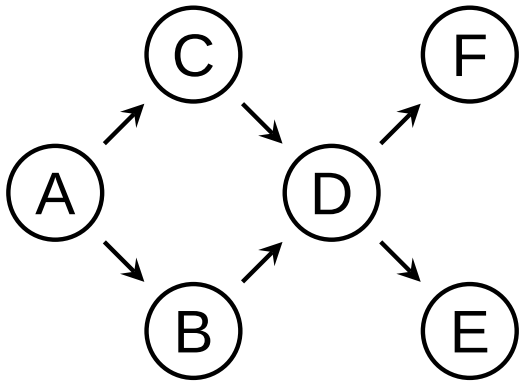


D-separation (Examples)

- D-separation metaphor: dependence as flow of influence along paths. All paths "blocked" implies independence.
- A set Z^* blocks a path from X to Y if one of these triples occurs on the path:

$$\circ \rightarrow Z \rightarrow \circ \quad \circ \leftarrow Z \rightarrow \circ \quad \circ \rightarrow W \leftarrow \circ$$

  ($Z \in Z^*$, $De(W)_G$ does not intersect Z^*)
- $(A \not\perp\!\!\!\perp F | B)$ due to an open path $A \rightarrow C \rightarrow D \rightarrow F$.
- $(C \perp\!\!\!\perp B | A)$. Both the path $C \leftarrow A \rightarrow B$ and $C \rightarrow D \leftarrow B$ are blocked

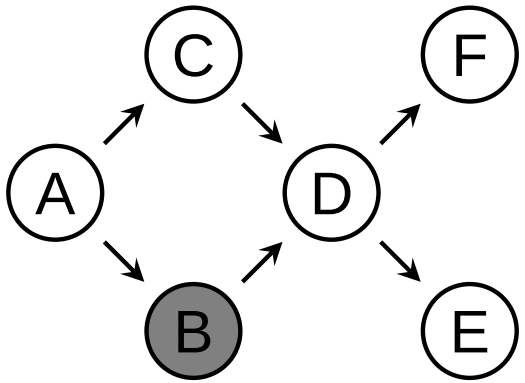


D-separation (Examples)

- D-separation metaphor: dependence as flow of influence along paths. All paths "blocked" implies independence.
- A set Z^* blocks a path from X to Y if one of these triples occurs on the path:

$$\circ \rightarrow Z \rightarrow \circ \quad \circ \leftarrow Z \rightarrow \circ \quad \circ \rightarrow W \leftarrow \circ$$

  ($Z \in Z^*$, $De(W)_G$ does not intersect Z^*)
- $(A \not\perp\!\!\!\perp F | B)$ due to an open path $A \rightarrow C \rightarrow D \rightarrow F$.
- $(C \perp\!\!\!\perp B | A)$. Both the path $C \leftarrow A \rightarrow B$ and $C \rightarrow D \leftarrow B$ are blocked

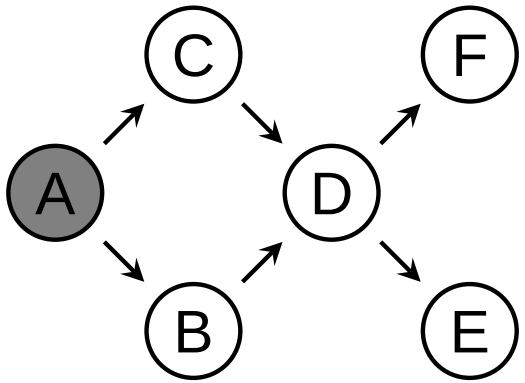


D-separation (Examples)

- D-separation metaphor: dependence as flow of influence along paths. All paths "blocked" implies independence.
- A set Z^* blocks a path from X to Y if one of these triples occurs on the path:

$$\circ \rightarrow Z \rightarrow \circ \quad \circ \leftarrow Z \rightarrow \circ \quad \circ \rightarrow W \leftarrow \circ$$

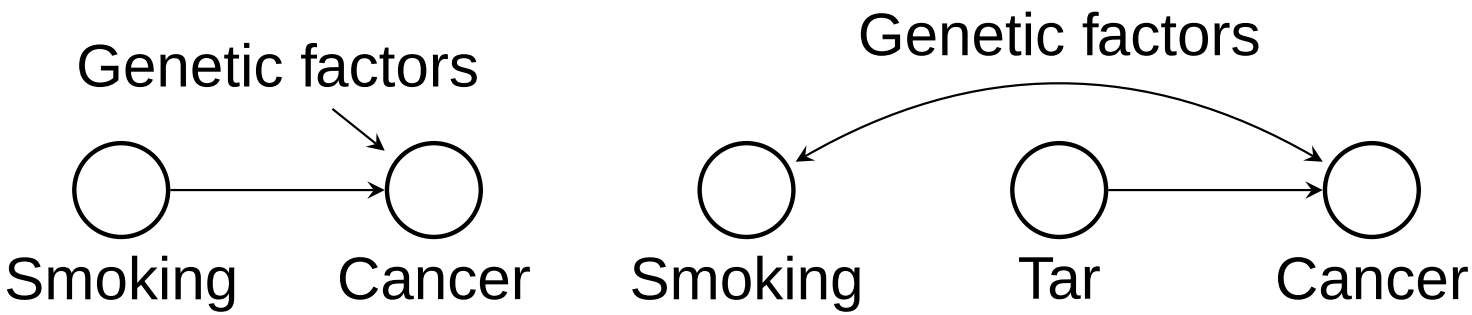
  ($Z \in Z^*$, $De(W)_G$ does not intersect Z^*)
- $(A \not\perp\!\!\!\perp F|B)$ due to an open path $A \rightarrow C \rightarrow D \rightarrow F$.
- $(C \perp\!\!\!\perp B|A)$. Both the path $C \leftarrow A \rightarrow B$ and $C \rightarrow D \leftarrow B$ are blocked



Interventions in Causal Models

An action $\text{do}(x)$ sets X to the value x regardless of the natural influences on X.

- The causal effect of $\text{do}(x)$ on $P(v)$ is an interventional distribution $P_x(v)$ or $P(v \mid \text{do}(x))$.
- A variable Y under $\text{do}(x)$ is sometimes written as Y_x.
- $\text{do}(x)$ removes all arrows (causal influences) incoming to X in model M to create a submodel M_x.



An Example From Medicine

Causal effect of a time-varying treatment on a patient outcome from longitudinal data collected in an observational study.

- Domain: HIV.
- CD4 is the immune cell destroyed by the HIV virus. Counts: 600-1000 is normal, 200 is a high risk of opportunistic infection.
- HAART is highly active anti-retroviral therapy, three drug cocktail.
- Medical question: what is the optimal CD4 count at which to start HAART in HIV infected patients.

HIV Example (cont.)

- HAART turned HIV into a chronic disease, increases CD4 counts, virus undetectable in blood (until resistance).
- Starting too late: chance of opportunistic infection, irreversible immune system damage, etc.
- Starting too early: resistance, side effects.
- Randomized trial data only on patients with CD4 < 200 (cuts death rate in half).
- Must use observational data.

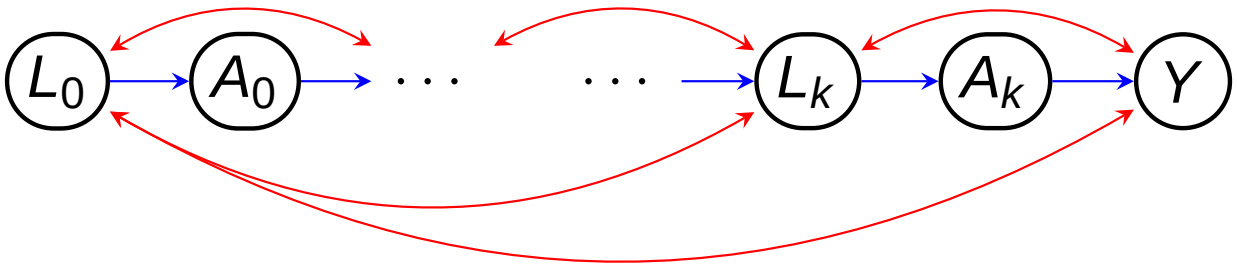
Conditioning vs Interventions

- Have to be careful with observational data, easy to get silly conclusions.
- People who get HAART tend to die a lot more.
- Does this mean HAART should not be used?
- No! HAART is prescribed to people who are *already very sick*: $p(\text{death} \mid \text{HAART}) \neq p(\text{death} \mid \text{do}(\text{HAART}))$.
- Medical question: how would outcome of patients differ between HAART treatment and no HAART treatment.
- Want to estimate $p(\text{death} \mid \text{do}(\text{HAART}))$ from observational data.

Conditioning vs Interventions

- Have to be careful with observational data, easy to get silly conclusions.
- People who get HAART tend to die a lot more.
- Does this mean HAART should not be used?
- No! HAART is prescribed to people who are *already very sick*: $p(\text{death} \mid \text{HAART}) \neq p(\text{death} \mid \text{do}(\text{HAART}))$.
- Medical question: how would outcome of patients differ between HAART treatment and no HAART treatment.
- Want to estimate $p(\text{death} \mid \text{do}(\text{HAART}))$ from observational data.

HIV Example (cont.)

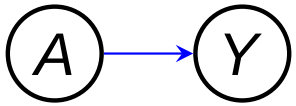


- CD4 ($L_0, \ldots, L_k$) is a "time-varying confounder".
- Affects both treatment (doctors prescribe HAART based on CD4 counts, confounding by indication) and outcome (cause of clinical AIDS).
- Earlier treatment affects CD4 counts also.
- Unspecified latent variables may affect both outcome and CD4, but treatments ($A_0, \ldots A_k$) are only affected by CD4 measurements.
- Given $p(l_0, a_0, \ldots, l_k, a_0, y)$. Want: $p(y \mid \text{do}(a_0, \ldots, a_k))$.

Fundamental Questions in Causal Inference

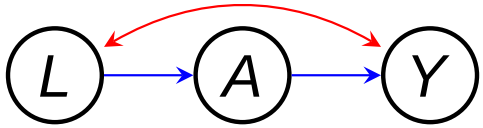
- Representation: formalizing intuitive causal notions (actual cause, direct effect, discrimination, etc.)
- Identification: expressing a causal quantity in terms of observational data (using causal assumptions).
- Estimation: given an identifiable causal quantity, what's the best way of estimating it from data?
- Structure learning: if the causal graph is not available from experts, which parts of this graph can we infer from observational data?

Causal Effect Identification



- Need to link observational and interventional data.
- Crucial assumption (consistency): $(A = a) \Rightarrow (Y_a = Y)$. ($Y_a$ is Y after do(a)).
- Untestable assumption, but needed to link observations and interventions.
- Simple example with no confounding: $p(y \mid \text{do}(a)) = p(y \mid a)$.
- Intuition: model postulates no common causes of A and Y – therefore observed dependence of A and Y must be causal.

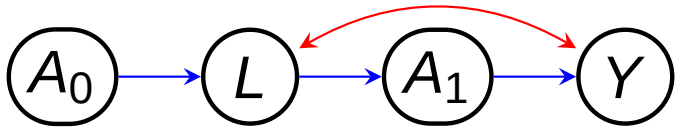
Identification With Confounding



- More complex example: L a common cause of A and Y.
- Observed dependence of A and Y could be due to non-causal influence via L.
- Solution: "adjust for" L. Conditioning on L d-separates all non-causal paths from A and Y. Average over levels of L.

$$
\begin{aligned}
p(y \mid \mathrm{do}(a)) &= \sum_l p(y, l \mid \mathrm{do}(a)) \\
&= \sum_l p(y \mid \mathrm{do}(a), l) p(l \mid \mathrm{do}(a)) \\
&= \sum_l p(y \mid a, l) p(l)
\end{aligned}
$$

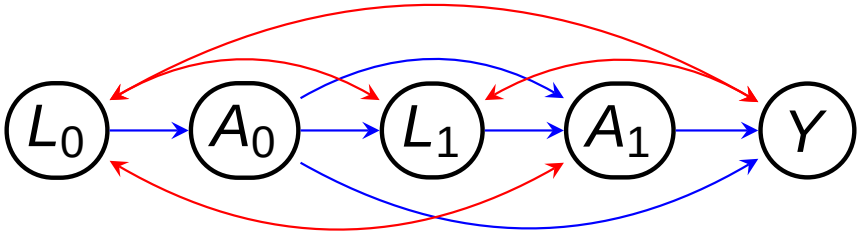
Identification With Time-varying Confounding



- Time varying confounding: L is both a child of treatment, and a cause of treatment.
- Adjust for confounders recursively, conditioned on the past.

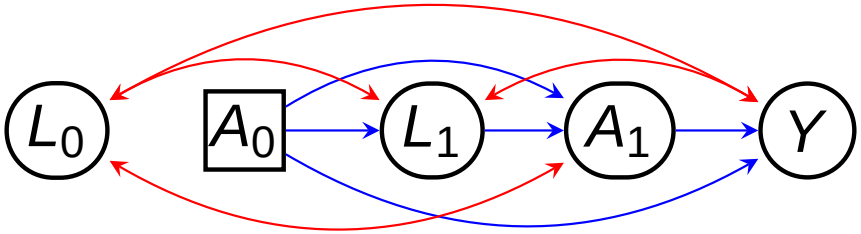
$$
\begin{aligned}
p(y \mid \mathrm{do}(a_0, a_1)) &= \sum_l p(y \mid l, \mathrm{do}(a_0, a_1)) p(l \mid \mathrm{do}(a_0, a_1)) \\
&= \sum_l p(y \mid l, a_1, \mathrm{do}(a_0)) p(l \mid \mathrm{do}(a_0)) \\
&= \sum_l p(y \mid l, a_1, a_0) p(l \mid a_0)
\end{aligned}
$$

General Identification Algorithm



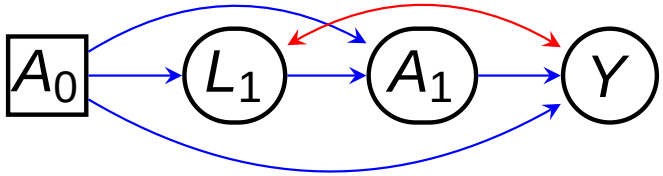
- Want $p(y \mid \text{do}(a_0, a_1))$.
- First compute
  $p(l_0, l_1, a_1, y \mid \text{do}(a_0)) = p(y, a_1, l_1 \mid a_0, l_0)p(l_0)$
- Then marginalize L_0:
  $p^*_{a_0}(y, a_1, l_1) = \sum_{l_0} p(y, a_1, l_1 \mid a_0, l_0)p(l_0)$
- Then compute $p^*_{a_0}(y, l_1 \mid \text{do}(a_1)) = p^*_{a_0}(y \mid a_1, l_1)p^*_{a_0}(l_1)$
- Finally, marginalize L_1: $p^{**}_{a_0,a_1}(y) = \sum_{l_1} p^*_{a_0}(y \mid a_1, l_1)p^*_{a_0}(l_1)$.
- Can show that $p(y \mid \text{do}(a_0, a_1)) = p^{**}_{a_0,a_1}(y)$

General Identification Algorithm



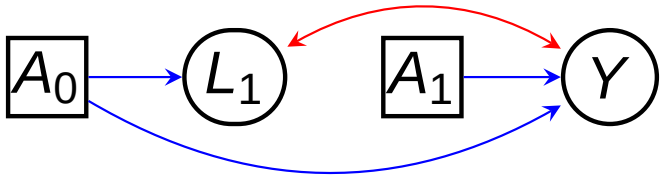
- Want $p(y \mid \text{do}(a_0, a_1))$.
- First compute
  $p(l_0, l_1, a_1, y \mid \text{do}(a_0)) = p(y, a_1, l_1 \mid a_0, l_0)p(l_0)$
- Then marginalize L_0:
  $p^*_{a_0}(y, a_1, l_1) = \sum_{l_0} p(y, a_1, l_1 \mid a_0, l_0)p(l_0)$
- Then compute $p^*_{a_0}(y, l_1 \mid \text{do}(a_1)) = p^*_{a_0}(y \mid a_1, l_1)p^*_{a_0}(l_1)$
- Finally, marginalize L_1: $p^{**}_{a_0,a_1}(y) = \sum_{l_1} p^*_{a_0}(y \mid a_1, l_1)p^*_{a_0}(l_1)$.
- Can show that $p(y \mid \text{do}(a_0, a_1)) = p^{**}_{a_0,a_1}(y)$

General Identification Algorithm



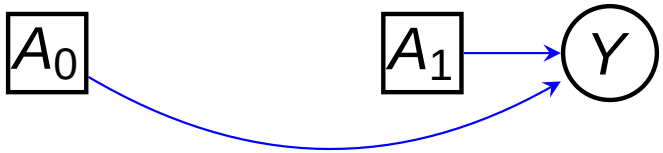
- Want $p(y \mid \text{do}(a_0, a_1))$.
- First compute
  $p(l_0, l_1, a_1, y \mid \text{do}(a_0)) = p(y, a_1, l_1 \mid a_0, l_0)p(l_0)$
- Then marginalize L_0:
  $p^*_{a_0}(y, a_1, l_1) = \sum_{l_0} p(y, a_1, l_1 \mid a_0, l_0)p(l_0)$
- Then compute $p^*_{a_0}(y, l_1 \mid \text{do}(a_1)) = p^*_{a_0}(y \mid a_1, l_1)p^*_{a_0}(l_1)$
- Finally, marginalize L_1: $p^{**}_{a_0,a_1}(y) = \sum_{l_1} p^*_{a_0}(y \mid a_1, l_1)p^*_{a_0}(l_1)$.
- Can show that $p(y \mid \text{do}(a_0, a_1)) = p^{**}_{a_0,a_1}(y)$

General Identification Algorithm



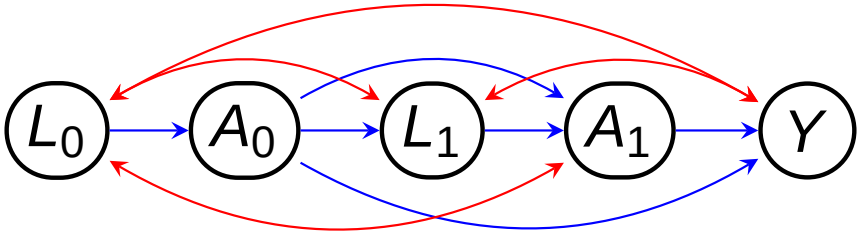
- Want $p(y \mid \text{do}(a_0, a_1))$.
- First compute
  $p(l_0, l_1, a_1, y \mid \text{do}(a_0)) = p(y, a_1, l_1 \mid a_0, l_0)p(l_0)$
- Then marginalize L_0:
  $p^*_{a_0}(y, a_1, l_1) = \sum_{l_0} p(y, a_1, l_1 \mid a_0, l_0)p(l_0)$
- Then compute $p^*_{a_0}(y, l_1 \mid \text{do}(a_1)) = p^*_{a_0}(y \mid a_1, l_1)p^*_{a_0}(l_1)$
- Finally, marginalize L_1: $p^{**}_{a_0,a_1}(y) = \sum_{l_1} p^*_{a_0}(y \mid a_1, l_1)p^*_{a_0}(l_1)$.
- Can show that $p(y \mid \text{do}(a_0, a_1)) = p^{**}_{a_0,a_1}(y)$

General Identification Algorithm



- Want $p(y \mid \text{do}(a_0, a_1))$.
- First compute
  $p(l_0, l_1, a_1, y \mid \text{do}(a_0)) = p(y, a_1, l_1 \mid a_0, l_0)p(l_0)$
- Then marginalize L_0:
  $p^*_{a_0}(y, a_1, l_1) = \sum_{l_0} p(y, a_1, l_1 \mid a_0, l_0)p(l_0)$
- Then compute $p^*_{a_0}(y, l_1 \mid \text{do}(a_1)) = p^*_{a_0}(y \mid a_1, l_1)p^*_{a_0}(l_1)$
- Finally, marginalize L_1: $p^{**}_{a_0,a_1}(y) = \sum_{l_1} p^*_{a_0}(y \mid a_1, l_1)p^*_{a_0}(l_1)$.
- Can show that $p(y \mid \text{do}(a_0, a_1)) = p^{**}_{a_0,a_1}(y)$

General Identification Algorithm



- Want $p(y \mid \text{do}(a_0, a_1))$.
- First compute
  $p(l_0, l_1, a_1, y \mid \text{do}(a_0)) = p(y, a_1, l_1 \mid a_0, l_0)p(l_0)$
- Then marginalize L_0:
  $p^*_{a_0}(y, a_1, l_1) = \sum_{l_0} p(y, a_1, l_1 \mid a_0, l_0)p(l_0)$
- Then compute $p^*_{a_0}(y, l_1 \mid \text{do}(a_1)) = p^*_{a_0}(y \mid a_1, l_1)p^*_{a_0}(l_1)$
- Finally, marginalize L_1: $p^{**}_{a_0,a_1}(y) = \sum_{l_1} p^*_{a_0}(y \mid a_1, l_1)p^*_{a_0}(l_1)$.
- Can show that $p(y \mid \text{do}(a_0, a_1)) = p^{**}_{a_0,a_1}(y)$

General Identification Scheme

$$p(y \mid \text{do}(a_0, a_1)) = \sum_{l_1} p^*_{a_0}(y \mid a_1, l_1) p^*_{a_0}(l_1)$$

where

$$p^*_{a_0}(y, a_1, l_1) = \sum_{l_0} p(y, a_1, l_1 \mid a_0, l_0) p(l_0)$$

1. Divide by $p(a_0 \mid l_0)$
2. Marginalize L_0
3. Divide by $p^*_{a_0}(a_1 \mid l_1)$
4. Marginalize L_1

General approach by Tian (2002), proved complete by Shpitser (2006), also Huang and Valtorta (2006).
Complete: if the algorithm gives no answer, there is no answer.

The Curse of Dimensionality in Causal Inference

- Many causal effects are identified by covariate adjustment:

$$p(y \mid \text{do}(a)) = \sum_l p(y \mid a, l)p(l)$$

- In fact, lots of causal effects in practice are of the form:

$$E[y \mid \text{do}(a)] = E[E[y \mid a, l]] = E[y \cdot \frac{1}{p(a \mid l)}]$$

- L is "a lot" of covariates.
- Curse of dimensionality applies to $p(y \mid a, l)$ and/or $p(a \mid l)$.
- Have to model without a lot of data (but what if modeling assumptions are wrong?)
- Handling conditional distributions may be computationally intractable.

Addressing the Curse

- The curse is a statistical problem (need lots of samples) and a computational problem (big graphs are intractable)
- Statistical way out: modeling. Danger: mis-specification!
- Solution: there is an estimator $\hat{\theta}$ for $E[y \mid \text{do}(a)]$ that combines models for $p(y \mid a, l)$ and $p(a \mid l)$ such that:
  - If $p(a \mid l)$ is correct, $\hat{\theta}$ is consistent (even if model for $p(y \mid a, l)$ is wrong!)
  - If $p(y \mid a, l)$ is correct, $\hat{\theta}$ is consistent (even if model for $p(a \mid l)$ is wrong!)
- So called "doubly robust" (or multiply robust) estimators (papers by Robins, Rotnitzky, van der Laan, others)
- Computational way out: exploiting Markov factorization via belief propagation, etc.
- Plug: see my talk for how to do this in latent variable models.

Addressing the Curse

- The curse is a statistical problem (need lots of samples) and a computational problem (big graphs are intractable)
- Statistical way out: modeling. Danger: mis-specification!
- Solution: there is an estimator $\hat{\theta}$ for $E[y \mid \text{do}(a)]$ that combines models for $p(y \mid a, l)$ and $p(a \mid l)$ such that:
  - If $p(a \mid l)$ is correct, $\hat{\theta}$ is consistent (even if model for $p(y \mid a, l)$ is wrong!)
  - If $p(y \mid a, l)$ is correct, $\hat{\theta}$ is consistent (even if model for $p(a \mid l)$ is wrong!)
- So called "doubly robust" (or multiply robust) estimators (papers by Robins, Rotnitzky, van der Laan, others)
- Computational way out: exploiting Markov factorization via belief propagation, etc.
- Plug: see my talk for how to do this in latent variable models.

Addressing the Curse

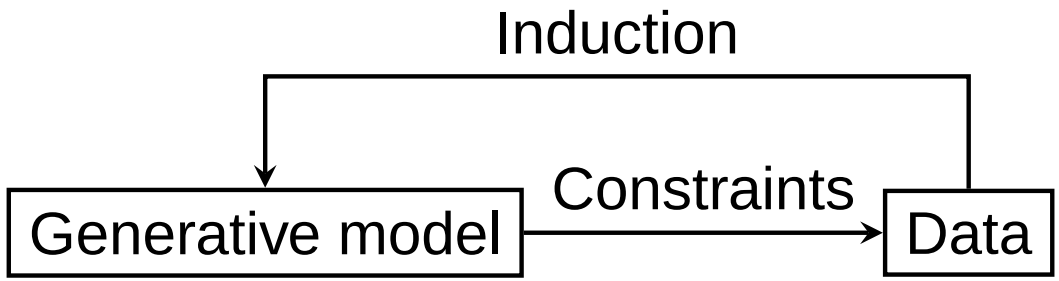
- The curse is a statistical problem (need lots of samples) and a computational problem (big graphs are intractable)
- Statistical way out: modeling. Danger: mis-specification!
- Solution: there is an estimator $\hat{\theta}$ for $E[y \mid \text{do}(a)]$ that combines models for $p(y \mid a, l)$ and $p(a \mid l)$ such that:
  - If $p(a \mid l)$ is correct, $\hat{\theta}$ is consistent (even if model for $p(y \mid a, l)$ is wrong!)
  - If $p(y \mid a, l)$ is correct, $\hat{\theta}$ is consistent (even if model for $p(a \mid l)$ is wrong!)
- So called "doubly robust" (or multiply robust) estimators (papers by Robins, Rotnitzky, van der Laan, others)
- Computational way out: exploiting Markov factorization via belief propagation, etc.
- Plug: see my talk for how to do this in latent variable models.

Addressing the Curse

- The curse is a statistical problem (need lots of samples) and a computational problem (big graphs are intractable)
- Statistical way out: modeling. Danger: mis-specification!
- Solution: there is an estimator $\hat{\theta}$ for $E[y \mid \text{do}(a)]$ that combines models for $p(y \mid a, l)$ and $p(a \mid l)$ such that:
  - If $p(a \mid l)$ is correct, $\hat{\theta}$ is consistent (even if model for $p(y \mid a, l)$ is wrong!)
  - If $p(y \mid a, l)$ is correct, $\hat{\theta}$ is consistent (even if model for $p(a \mid l)$ is wrong!)
- So called “doubly robust” (or multiply robust) estimators (papers by Robins, Rotnitzky, van der Laan, others)
- Computational way out: exploiting Markov factorization via belief propagation, etc.
- Plug: see my talk for how to do this in latent variable models.

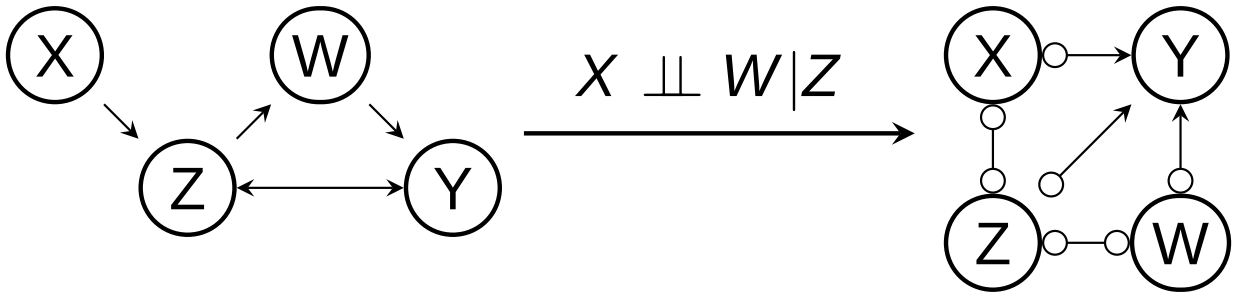
Induction: Inferring (Causal) Theories From Data.

- Central problem of empirical science
- Very important problem in AI (vast literature)
- In AI, causal theories are represented as directed graphs using the graphical models formalism
- The causal induction problem is to infer the directed graph from the constraints it places on the joint probability distribution over observables



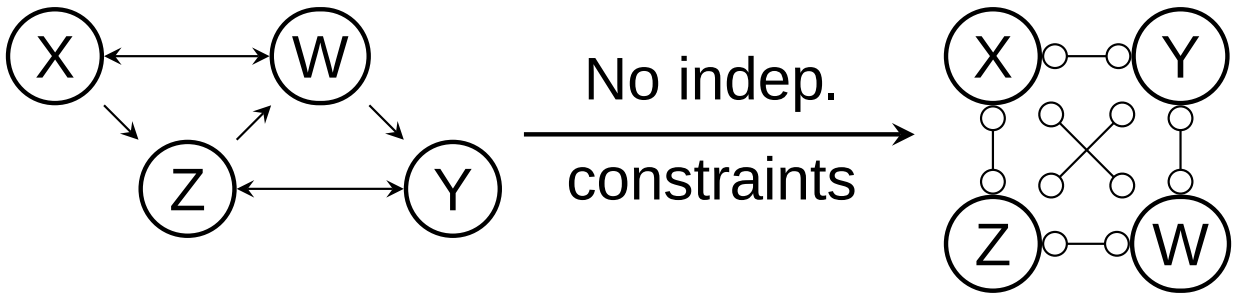
Induction: Inferring (Causal) Theories From Data.

- Central problem of empirical science
- Very important problem in AI (vast literature)
- In AI, causal theories are represented as directed graphs using the graphical models formalism
- The causal induction problem is to infer the directed graph from the constraints it places on the joint probability distribution over observables



Induction: Inferring (Causal) Theories From Data.

- Central problem of empirical science
- Very important problem in AI (vast literature)
- In AI, causal theories are represented as directed graphs using the graphical models formalism
- The causal induction problem is to infer the directed graph from the constraints it places on the joint probability distribution over observables



Current Approaches

- Constraint based: FCI algorithm (Spirtes et al)
  - search for constraints in the data, rule out graphs without these constraints
  - Issues: multiple hypothesis testing, doing many independence tests is intractable.
- Search and score: GES algorithm (Chickering) using BIC (Schwartz)
- Big issue: both approaches (so far!) only exploit conditional independences. Latent models have other constraints..

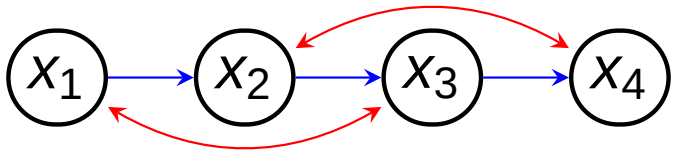
Current Approaches

- Constraint based: FCI algorithm (Spirtes et al)
  - search for constraints in the data, rule out graphs without these constraints
  - Issues: multiple hypothesis testing, doing many independence tests is intractable.
- Search and score: GES algorithm (Chickering) using BIC (Schwartz)
  - Score each graph by how well it explains the data, penalize big graphs (Occam's razor). Search for a high scoring graph.
  - Issues: BIC only has asymptotic guarantees, highest scoring model not always informative.
- Big issue: both approaches (so far!) only exploit conditional independences. Latent models have other constraints..

Current Approaches

- Constraint based: FCI algorithm (Spirtes et al)
  - search for constraints in the data, rule out graphs without these constraints
  - Issues: multiple hypothesis testing, doing many independence tests is intractable.
- Search and score: GES algorithm (Chickering) using BIC (Schwartz)
  - Score each graph by how well it explains the data, penalize big graphs (Occam's razor). Search for a high scoring graph.
  - Issues: BIC only has asymptotic guarantees, highest scoring model not always informative.
- Big issue: both approaches (so far!) only exploit conditional independences. Latent models have other constraints..

“Odd” Constraints in Latent Models.



- Consider a binary model pictured above.
- No conditional independences, according to current theory this model is saturated ($2^4 - 1 = 15$ parameters).
- There is a dimension-reducing constraint in this model (!)

$$\frac{\partial}{\partial x_1} \sum_{x_2} p(x_4 \mid x_1, x_2, x_3) p(x_2 \mid x_1) = 0$$

- Stay tuned for the second part, where I will discuss what these types of constraints buy us, and their relationship to causal effect identification.

Conclusions (part 1)

- Causal inference is about identification and estimation of causal effects, and learning causal structure from data.
- Important in medicine, social sciences, computational biology, etc.
- Many published studies contain causal inference problems.
- Graphical models are a useful tool for representing causality.
- Big problems:

Conclusions (part 1)

- Causal inference is about identification and estimation of causal effects, and learning causal structure from data.
- Important in medicine, social sciences, computational biology, etc.
- Many published studies contain causal inference problems.
- Graphical models are a useful tool for representing causality.
- Big problems:
  - What to do when quantity of interest is not identifiable? (instruments, bounds, pseudo-randomization)
  - Dealing with the curse of dimensionality.
  - Dealing with latent models (more on this in part 2).

References

Where to learn more:

- Causality: Models, Reasoning, and Inference. J. Pearl, Cambridge University Press, 2009.
- Causation, Prediction, and Search. P. Spirtes, C. Glymour, R. Scheines, MIT Press 2000.
- James Robin's papers: http://www.biostat.harvard.edu/~robins/research.html In particular: "A new approach to causal inference..." (1986).
- "The Method of Path Coefficients" S. Wright, Annals of Mathematical Statistics (1934). Earliest reference in 1918.
- UAI!

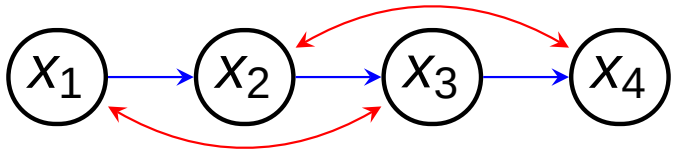
Causal Inference Tutorial (part 2) Latent Variable Models of Post-Truncation Independence

Ilya Shpitser
ishpitse@hsph.harvard.edu

Causal Inference Group, Department of Epidemiology
Harvard School of Public Health

July 19, 2011

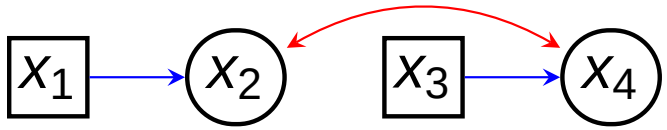
“Odd” Constraints: What’s Going On?



$$\frac{\partial}{\partial x_1} \sum_{x_2} p(x_4 \mid x_1, x_2, x_3) p(x_2 \mid x_1) = 0$$

- What is the interpretation of this constraint?
- Assume this is a causal graph, and consider identifying $p(x_4 \mid \text{do}(x_3, x_1))$.
- Identifiable and equal to $\sum_{x_2} p(x_4 \mid x_1, x_2, x_3) p(x_2 \mid x_1)$.
- Constraint: $p(x_4 \mid do(x_3, x_1)) = p(x_4 \mid do(x_3))$, or X_1 is independent of X_4 if we intervene on (not condition on!) X_3.
- Graphical interpretation: intervention on X_3 cuts arcs pointing to X_3, this can create new d-separation.

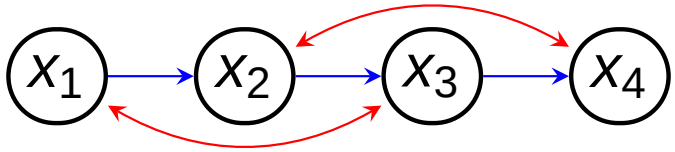
“Odd” Constraints: What’s Going On?



$$\frac{\partial}{\partial x_1} \sum_{x_2} p(x_4 \mid x_1, x_2, x_3) p(x_2 \mid x_1) = 0$$

- What is the interpretation of this constraint?
- Assume this is a causal graph, and consider identifying $p(x_4 \mid \text{do}(x_3, x_1))$.
- Identifiable and equal to $\sum_{x_2} p(x_4 \mid x_1, x_2, x_3) p(x_2 \mid x_1)$.
- Constraint: $p(x_4 \mid do(x_3, x_1)) = p(x_4 \mid do(x_3))$, or X_1 is independent of X_4 if we intervene on (not condition on!) X_3.
- Graphical interpretation: intervention on X_3 cuts arcs pointing to X_3, this can create new d-separation.

Generalized Independence Constraints



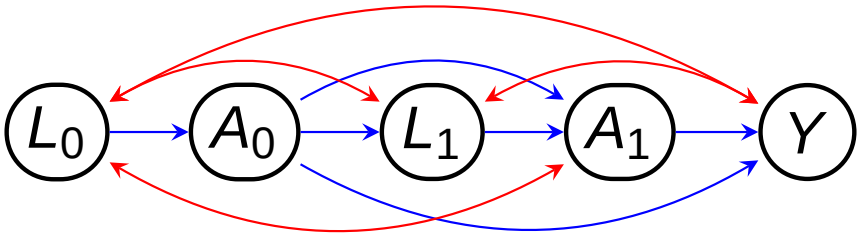
$$\frac{\partial}{\partial x_1} \sum_{x_2} p(x_4 \mid x_1, x_2, x_3) p(x_2 \mid x_1) = 0$$

- Another interpretation: $X_1 \perp\!\!\!\perp X_4$ in $\frac{p(x_1,x_2,x_3,x_4)}{p(x_3|x_1,x_2)}$ ("post-truncation independence")
- Usual conditional independences: $X_1 \perp\!\!\!\perp X_4$ in $\frac{p(x_1,x_2,x_3,x_4)}{p(x_3)}$.
- Recall: in DAG models we have $p(v) = \prod_i p(x_i \mid \text{pa}(x_i))$, which implies $X_i \perp\!\!\!\perp$ Non-descendants(X_i) in $\frac{p(v)}{p(\text{pa}(x_i))}$.
- Want something similar for these generalized independences in graphical models with latents.

A New Factorization: What Are The Pieces?

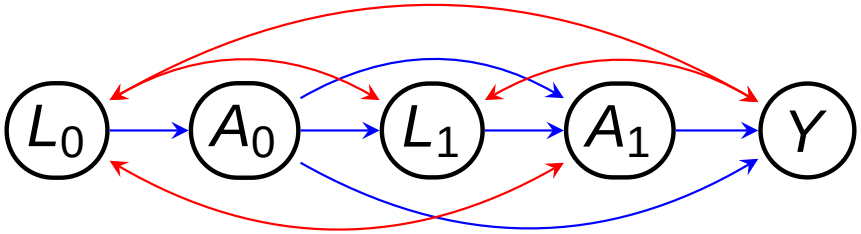
- In the Markov factorization of DAG models, $p(v \mid \text{pa}(v))$ are the building blocks.
- What should the building blocks be for us?
- Since we have latent variables, some variables cannot be made independent no matter what we do.
- Thus: our building blocks will involve sets.
- We want to represent $X_i \perp\!\!\!\perp X_j$ in $\frac{p(v)}{p(x_k|x_m)}$, which "corresponds to" independence after $\text{do}(x_k)$.
- Thus: our building blocks will involve interventional distributions.

Intrinsic Sets



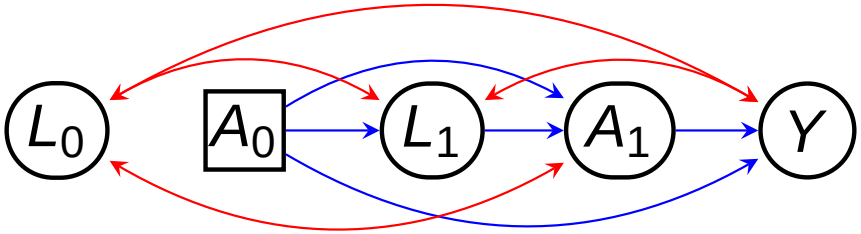
- Intrinsic set S: $P(s \mid \text{do}(\text{pa}(s) \setminus s))$ identifiable, nodes in S are a spanning tree in $\leftrightarrow$ subgraph.
- Intrinsic sets: $\{L_0\}$, $\{A_0\}$, $\{A_1\}$, $\{L_0, L_1\}$, $\{L_0, L_1, A_1\}$, $\{L_0, L_1, A_1, Y\}$, $\{L_1, Y\}$, $\{Y\}$.
- Not intrinsic: $\{L_0, A_0\}$ (not a spanning tree in $\leftrightarrow$ subgraph), $\{L_0, A_1\}$ ($p(l_0, a_1 \mid \text{do}(l_1))$ not identifiable)
- Note: unlike the DAG case, intrinsic sets overlap.

Recall: General Identification



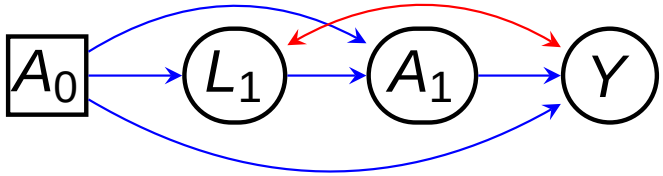
- Want $p(y \mid \text{do}(a_0, a_1))$.
- First compute
  $p(l_0, l_1, a_1, y \mid \text{do}(a_0)) = p(y, a_1, l_1 \mid a_0, l_0)p(l_0)$
- Then marginalize L_0:
  $p^*_{a_0}(y, a_1, l_1) = \sum_{l_0} p(y, a_1, l_1 \mid a_0, l_0)p(l_0)$
- Then compute $p^*_{a_0}(y, l_1 \mid \text{do}(a_1)) = p^*_{a_0}(y \mid a_1, l_1)p^*_{a_0}(l_1)$
- Finally, marginalize L_1: $p^{**}_{a_0,a_1}(y) = \sum_{l_1} p^*_{a_0}(y \mid a_1, l_1)p^*_{a_0}(l_1)$.
- Can show that $p(y \mid \text{do}(a_0, a_1)) = p^{**}_{a_0,a_1}(y)$

Recall: General Identification



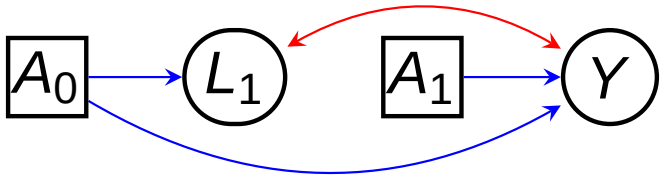
- Want $p(y \mid \text{do}(a_0, a_1))$.
- First compute
  $p(l_0, l_1, a_1, y \mid \text{do}(a_0)) = p(y, a_1, l_1 \mid a_0, l_0)p(l_0)$
- Then marginalize L_0:
  $p^*_{a_0}(y, a_1, l_1) = \sum_{l_0} p(y, a_1, l_1 \mid a_0, l_0)p(l_0)$
- Then compute $p^*_{a_0}(y, l_1 \mid \text{do}(a_1)) = p^*_{a_0}(y \mid a_1, l_1)p^*_{a_0}(l_1)$
- Finally, marginalize L_1: $p^{**}_{a_0,a_1}(y) = \sum_{l_1} p^*_{a_0}(y \mid a_1, l_1)p^*_{a_0}(l_1)$.
- Can show that $p(y \mid \text{do}(a_0, a_1)) = p^{**}_{a_0,a_1}(y)$

Recall: General Identification



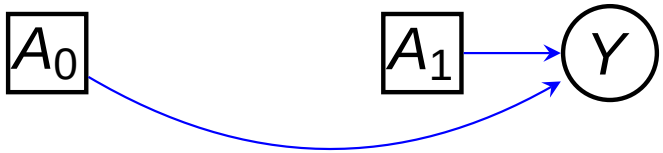
- Want $p(y \mid \text{do}(a_0, a_1))$.
- First compute
  $p(l_0, l_1, a_1, y \mid \text{do}(a_0)) = p(y, a_1, l_1 \mid a_0, l_0)p(l_0)$
- Then marginalize L_0:
  $p^*_{a_0}(y, a_1, l_1) = \sum_{l_0} p(y, a_1, l_1 \mid a_0, l_0)p(l_0)$
- Then compute $p^*_{a_0}(y, l_1 \mid \text{do}(a_1)) = p^*_{a_0}(y \mid a_1, l_1)p^*_{a_0}(l_1)$
- Finally, marginalize L_1: $p^{**}_{a_0,a_1}(y) = \sum_{l_1} p^*_{a_0}(y \mid a_1, l_1)p^*_{a_0}(l_1)$.
- Can show that $p(y \mid \text{do}(a_0, a_1)) = p^{**}_{a_0,a_1}(y)$

Recall: General Identification



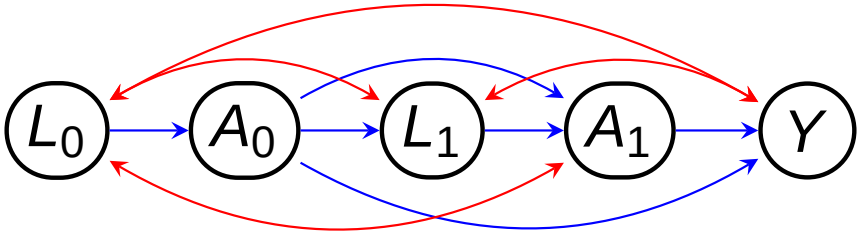
- Want $p(y \mid \text{do}(a_0, a_1))$.
- First compute
  $p(l_0, l_1, a_1, y \mid \text{do}(a_0)) = p(y, a_1, l_1 \mid a_0, l_0)p(l_0)$
- Then marginalize L_0:
  $p^*_{a_0}(y, a_1, l_1) = \sum_{l_0} p(y, a_1, l_1 \mid a_0, l_0)p(l_0)$
- Then compute $p^*_{a_0}(y, l_1 \mid \text{do}(a_1)) = p^*_{a_0}(y \mid a_1, l_1)p^*_{a_0}(l_1)$
- Finally, marginalize L_1: $p^{**}_{a_0,a_1}(y) = \sum_{l_1} p^*_{a_0}(y \mid a_1, l_1)p^*_{a_0}(l_1)$.
- Can show that $p(y \mid \text{do}(a_0, a_1)) = p^{**}_{a_0,a_1}(y)$

Recall: General Identification



- Want $p(y \mid \text{do}(a_0, a_1))$.
- First compute
  $p(l_0, l_1, a_1, y \mid \text{do}(a_0)) = p(y, a_1, l_1 \mid a_0, l_0)p(l_0)$
- Then marginalize L_0:
  $p^*_{a_0}(y, a_1, l_1) = \sum_{l_0} p(y, a_1, l_1 \mid a_0, l_0)p(l_0)$
- Then compute $p^*_{a_0}(y, l_1 \mid \text{do}(a_1)) = p^*_{a_0}(y \mid a_1, l_1)p^*_{a_0}(l_1)$
- Finally, marginalize L_1: $p^{**}_{a_0,a_1}(y) = \sum_{l_1} p^*_{a_0}(y \mid a_1, l_1)p^*_{a_0}(l_1)$.
- Can show that $p(y \mid \text{do}(a_0, a_1)) = p^{**}_{a_0,a_1}(y)$

Recall: General Identification



- Want $p(y \mid \text{do}(a_0, a_1))$.
- First compute
  $p(l_0, l_1, a_1, y \mid \text{do}(a_0)) = p(y, a_1, l_1 \mid a_0, l_0)p(l_0)$
- Then marginalize L_0:
  $p^*_{a_0}(y, a_1, l_1) = \sum_{l_0} p(y, a_1, l_1 \mid a_0, l_0)p(l_0)$
- Then compute $p^*_{a_0}(y, l_1 \mid \text{do}(a_1)) = p^*_{a_0}(y \mid a_1, l_1)p^*_{a_0}(l_1)$
- Finally, marginalize L_1: $p^{**}_{a_0,a_1}(y) = \sum_{l_1} p^*_{a_0}(y \mid a_1, l_1)p^*_{a_0}(l_1)$.
- Can show that $p(y \mid \text{do}(a_0, a_1)) = p^{**}_{a_0,a_1}(y)$

General Identification Scheme

$$p(y \mid \mathrm{do}(a_0, a_1)) = \sum_{l_1} p^*_{a_0}(y \mid a_1, l_1) p^*_{a_0}(l_1)$$

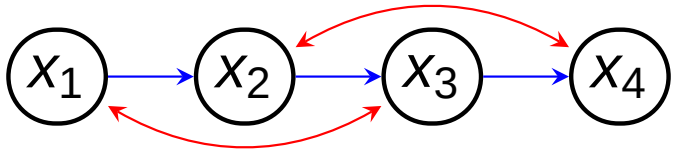
where

$$p^*_{a_0}(y, a_1, l_1) = \sum_{l_0} p(y, a_1, l_1 \mid a_0, l_0) p(l_0)$$

1. Divide by $p(a_0 \mid l_0)$
2. Marginalize L_0
3. Divide by $p^*_{a_0}(a_1 \mid l_1)$
4. Marginalize L_1

Identification is *recursive*. At any stage of the recursion, we might have an independence constraint. Thus our factorization must itself be recursive.

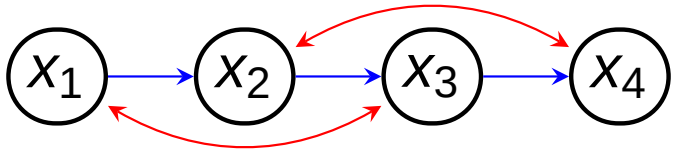
Recursive Factorization (Preliminaries)



- Ancestral set A: $x \in A \Rightarrow \text{an}(x) \subseteq A$. All ancestral sets in $\mathcal{G}$: $\mathcal{A}(\mathcal{G})$.
- $\mathcal{A}(\mathcal{G}) = \{X_1\}, \{X_1, X_2\}, \{X_1, X_2, X_3\}, \{X_1, X_2, X_3, X_4\}$.
- A district of $\mathcal{G}$ is a set of nodes forming a maximal spanning tree in $\leftrightarrow$ subgraph of $\mathcal{G}$.
- A set of districts $\mathcal{D}(\mathcal{G})$ in $\mathcal{G}$ always forms a unique partition of nodes in $\mathcal{G}$. $\mathcal{D}(\mathcal{G}) = \{X_1, X_3\}, \{X_2, X_4\}$.
- Property of causal models (district factorization):

$$p(v) = \prod_{d \in \mathcal{D}(\mathcal{G})} p(d \mid \text{do}(\text{pa(d)} \setminus d))$$

Recursive Factorization (Preliminaries)



- Ancestral set A: $x \in A \Rightarrow \text{an}(x) \subseteq A$. All ancestral sets in $\mathcal{G}$: $\mathcal{A}(\mathcal{G})$.
- $\mathcal{A}(\mathcal{G}) = \{X_1\}, \{X_1, X_2\}, \{X_1, X_2, X_3\}, \{X_1, X_2, X_3, X_4\}$.
- A district of $\mathcal{G}$ is a set of nodes forming a maximal spanning tree in $\leftrightarrow$ subgraph of $\mathcal{G}$.
- A set of districts $\mathcal{D}(\mathcal{G})$ in $\mathcal{G}$ always forms a unique partition of nodes in $\mathcal{G}$. $\mathcal{D}(\mathcal{G}) = \{X_1, X_3\}, \{X_2, X_4\}$.
- Property of causal models (district factorization):

$$p(x_1, x_2, x_3, x_4) = p(x_3, x_1 \mid \text{do}(x_2))p(x_2, x_4 \mid \text{do}(x_1, x_3))$$

Recursive Factorization

Definition

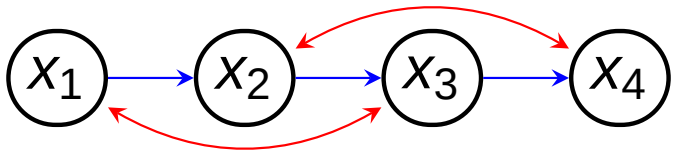
$p(v)$ recursively factorizes according to $\mathcal{G}$ if district factorization holds for $p(v)$ at "the outer level," and any interventional distribution corresponding to a district itself factorizes according to the appropriate subgraph of $\mathcal{G}$.

Formally: $p(v)$ recursively factorizes with respect to $\mathcal{G}$ and a set $\{p(s \mid \text{do}(\text{pa}(s) \setminus s)) \mid S \in \mathcal{I}(\mathcal{G})\}$ if for every $A \in \mathcal{A}(\mathcal{G})$:

- $p(a) = \prod_{d \in \mathcal{D}(\mathcal{G}_A)} p(d \mid \text{do}(pa(d) \setminus d))$
- if $|\mathcal{D}(\mathcal{G}_A)| > 1$, then for every $D \in \mathcal{G}_A$, and every assignment ν to $\text{pa}(D) \setminus D$:
  - $p(d \mid \text{do}(\nu))$ r-factorizes according to $\mathcal{G}_D$ and $\{p(s \mid \text{do}(\text{pa}(s) \setminus s)) \mid S \in \mathcal{I}(\mathcal{G}_D)\}$.

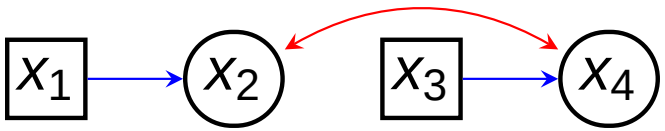
where $\mathcal{G}_A$ is a restriction of $\mathcal{G}$ to A.

Example



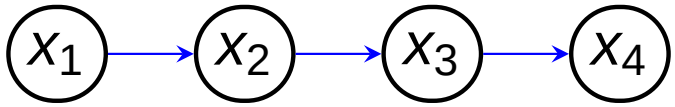
- Intrinsic sets: $\{X_1\}, \{X_2\}, \{X_1, X_3\}, \{X_3\}, \{X_2, X_4\}, \{X_4\}$.
- R-factors: $p(x_1)$, $p(x_2 \mid \mathrm{do}(x_1))$, $p(x_1, x_3 \mid \mathrm{do}(x_2))$, $p(x_3 \mid \mathrm{do}(x_2))$, $p(x_2, x_4 \mid \mathrm{do}(x_1, x_3))$, $p(x_4 \mid \mathrm{do}(x_3))$.
- "Outer" factorization: $p(x_1, x_2, x_3, x_4) = p(x_2, x_4 \mid \mathrm{do}(x_1, x_3))p(x_1, x_3 \mid \mathrm{do}(x_2))$.
- "Inner" factorization of $p(x_2, x_4 \mid \mathrm{do}(x_1, x_3))$: ancestral sets of $\{X_2, X_4\}$ in the subgraph: $\{X_2\}, \{X_4\}$.
- $p(x_4 \mid \mathrm{do}(x_1, x_3)) = p(x_4 \mid \mathrm{do}(x_3))$.
- All r-factors $p(s \mid \mathrm{do}(\mathrm{pa}(s) \setminus s))$ are identifiable from $p(v)$, so the factorization is "about" $p(v)$. This is a *statistical* model!

Example



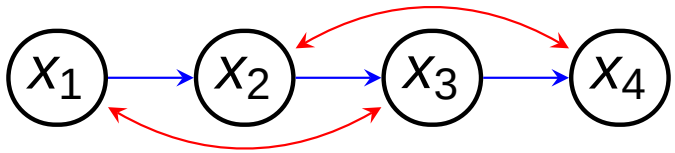
- Intrinsic sets: $\{X_1\}, \{X_2\}, \{X_1, X_3\}, \{X_3\}, \{X_2, X_4\}, \{X_4\}$.
- R-factors: $p(x_1)$, $p(x_2 \mid \mathrm{do}(x_1))$, $p(x_1, x_3 \mid \mathrm{do}(x_2))$, $p(x_3 \mid \mathrm{do}(x_2))$, $p(x_2, x_4 \mid \mathrm{do}(x_1, x_3))$, $p(x_4 \mid \mathrm{do}(x_3))$.
- "Outer" factorization: $p(x_1, x_2, x_3, x_4) = p(x_2, x_4 \mid \mathrm{do}(x_1, x_3))p(x_1, x_3 \mid \mathrm{do}(x_2))$.
- "Inner" factorization of $p(x_2, x_4 \mid \mathrm{do}(x_1, x_3))$: ancestral sets of $\{X_2, X_4\}$ in the subgraph: $\{X_2\}, \{X_4\}$.
- $p(x_4 \mid \mathrm{do}(x_1, x_3)) = p(x_4 \mid \mathrm{do}(x_3))$.
- All r-factors $p(s \mid \mathrm{do}(\mathrm{pa}(s) \setminus s))$ are identifiable from $p(v)$, so the factorization is "about" $p(v)$. This is a *statistical* model!

Obtaining The Model Dimension



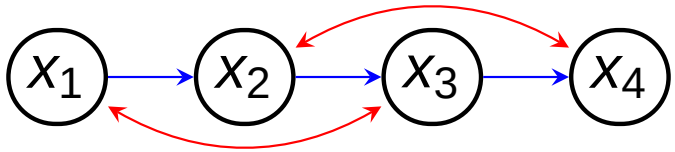
- Binary DAG model. 15 parameters without Markov factorization.
- But we know $p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_2)p(x_4 \mid x_3)$.
- Thus model dimension is $1 + 2 + 2 + 2 = 7$ (parameters: $p(x_i = 0 \mid \text{pa}(x_i))$).
- Binary latent model. No conditional independence constraints: 15 parameters.
- We know $p(x_4 \mid \text{do}(x_1, x_3)) = p(x_4 \mid \text{do}(x_3))$.
- Can we parameterize to take advantage of this?

Obtaining The Model Dimension



- Binary DAG model. 15 parameters without Markov factorization.
- But we know $p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_2)p(x_4 \mid x_3)$.
- Thus model dimension is $1 + 2 + 2 + 2 = 7$ (parameters: $p(x_i = 0 \mid \text{pa}(x_i))$).
- Binary latent model. No conditional independence constraints: 15 parameters.
- We know $p(x_4 \mid \text{do}(x_1, x_3)) = p(x_4 \mid \text{do}(x_3))$.
- Can we parameterize to take advantage of this?

Parameters For Binary R-Factorizing Models



- Intrinsic sets: $\{X_1\}, \{X_2\}, \{X_1, X_3\}, \{X_3\}, \{X_2, X_4\}, \{X_4\}$.
- R-factors: $p(x_1), p(x_2 \mid \mathrm{do}(x_1)), p(x_1, x_3 \mid \mathrm{do}(x_2))$, $p(x_3 \mid \mathrm{do}(x_2)), p(x_2, x_4 \mid \mathrm{do}(x_1, x_3)), p(x_4 \mid \mathrm{do}(x_3))$.
- Parameters: $p(x_1 = 0), p(x_2 = 0 \mid \mathrm{do}(x_1))$, $p(x_1 = 0, x_3 = 0 \mid \mathrm{do}(x_2)), p(x_3 = 0 \mid \mathrm{do}(x_2))$, $p(x_2 = 0, x_4 = 0 \mid \mathrm{do}(x_1, x_3)), p(x_4 = 0 \mid \mathrm{do}(x_3))$.
- Model dimension: $1 + 2 + 2 + 2 + 4 + 2 = 13$.

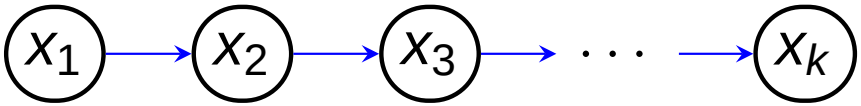
General Parameterization

- For an intrinsic set S in $\mathcal{G}$, let the recursive head $\mathrm{rh}(S)$ be the subset of S with no children in S.
- For an r-factor $p(s \mid \mathrm{do}(\mathrm{pa}(s) \setminus s))$, let its parameters be

$$q_{\mathrm{rh}(s)}(\mathrm{pa}(s)) = p(\mathrm{rh}(S) = 0 \mid s \setminus \mathrm{rh}(s), \mathrm{do}(\mathrm{pa}(s) \setminus s))$$

- Note: intrinsic sets overlap, so parameters are derived from overlapping marginals.
- Parameterization is variation dependent.
- Möbius transform can be used to obtain $p(v)$ from parameters.
- Models are smooth, in the curved exponential family.

What Is This Good For: Efficient Inference



- Say we want to compute $p(x_k) = \sum_{x_1,\ldots,x_{k-1}} p(x_1, \ldots, x_k)$
- Naive algorithm for summing is intractable (big table).
- Algorithms for DAG models exploit Markov factorization to perform sums efficiently (belief propagation, variable elimination, sampling methods, variational methods, etc.):

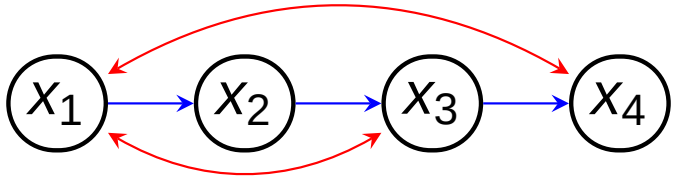
$$p(x_k) = \sum_{x_{k-1}} p(x_k \mid x_{k-1}) \sum_{x_{k-2}} p(x_{k-1} \mid x_{k-2}) \cdots \sum_{x_1} p(x_1) p(x_2 \mid x_1)$$

- See my talk for how to do this with these new parameters.

What Is This Good For: Variance Minimizing Estimators

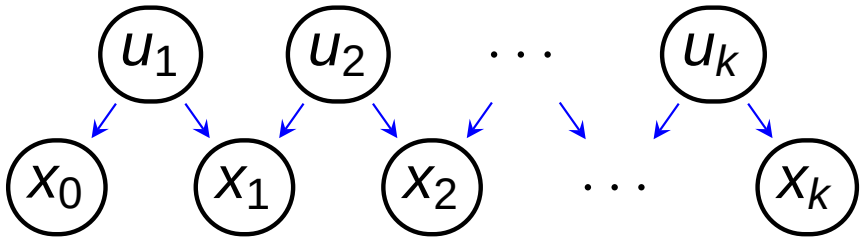
- An identifiable causal effect $p(y \mid \mathrm{do}(x))$ has multiple functionals in terms of $p(v)$. Which one minimizes the variance?
- For an r-factorizing model: fit parameters by ML, express $p(y \mid \mathrm{do}(x))$ in terms of ML parameters.
- The resulting estimator minimizes variance among all estimators in the model.
- See my talk for more on this.

What Is This Good For: Causal Discovery



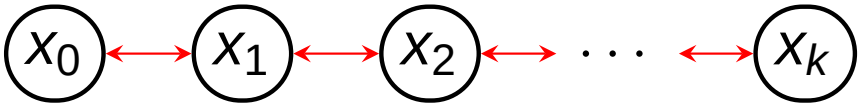
- Say we want to reconstruct this graph from data.
- FCI sees no conditional independences, returns a complete graph.
- GES can't score latent models.
- Structural EM: slow, infinite search space, singularities.
- An alternative: write down a likelihood using new parameters, use local search with BIC (code exists).
- A single independence in this model ($X_4 \perp\!\!\!\perp X_2$) in $\frac{p(x_1,x_2,x_3,x_4)}{p(x_2|x_1)}$ is enough to narrow things down to *one* graph.
- Can distinguish this model from a complete graph.

Caveats



- R-factorizing models do not make *any* assumptions on latents.
- Assume binary DAG model, we only observe $x_1, ..., x_k$.
- This model is parameterized by $p(u_1 = 0), \ldots, p(u_k = 0)$, $p(x_0 = 0 \mid u_1)$, $p(x_i = 0 \mid u_i, u_{i+1})$ ($O(k)$ parameters).
- An R-factorizing model has $O(k^2)$ parameters.
- Difference: DAG model does not represent all DAG marginal distributions with above graph (what if u_i nodes have more than 2 states?)

Caveats



- R-factorizing models do not make *any* assumptions on latents.
- Assume binary DAG model, we only observe $x_1, ..., x_k$.
- This model is parameterized by $p(u_1 = 0), \ldots, p(u_k = 0)$, $p(x_0 = 0 \mid u_1)$, $p(x_i = 0 \mid u_i, u_{i+1})$ ($O(k)$ parameters).
- An R-factorizing model has $O(k^2)$ parameters.
- Difference: DAG model does not represent all DAG marginal distributions with above graph (what if u_i nodes have more than 2 states?)

Why Not DAGs With Latents?

- DAGs with latents already in use (HMMs, Kalman filters, many others).
- Algorithms exist: (structural) EM.
- Why new models?
- Marginal DAG models are “nasty” (distributions not in nice form, singularities)
- Infinite search space for learning graphs (how many latents to add?)
- If latent state spaces not big enough there will be misspecification bias.

Why Not DAGs With Latents?

- DAGs with latents already in use (HMMs, Kalman filters, many others).
- Algorithms exist: (structural) EM.
- Why new models?
- Marginal DAG models are “nasty” (distributions not in nice form, singularities)
- Infinite search space for learning graphs (how many latents to add?)
- If latent state spaces not big enough there will be misspecification bias.

Advantages Of R-factorizing Models

- Smooth.
- In the curved exponential family.
- Finite search space for causal discovery.
- Makes no assumptions on latents
- Explicitly incorporates “native” latent model constraints: $(X_i \perp\!\!\!\perp X_j)$ in $\frac{p(v)}{p(x_k|x_m)}$.

Conclusions (part 2)

- Latent variable models contain constraints of the form $(X_i \perp\!\!\!\perp X_j)$ in $\frac{p(v)}{p(x_k|x_m)}$.
- A new factorization (and parameterization) exists which takes advantage of these constraints for learning and inference.
- Factorization fairly complex, but can be interpreted in terms of pieces corresponding to certain interventions (causal effects).
- In DAG models the pieces are $p(x_i \mid \text{pa}(x_i))$ for every node X_i. In the new models, the pieces are $p(s \mid \text{do}(\text{pa}(s) \setminus s))$ for certain sets S.
- Models are a generalization of Bayesian networks to the latent variable case.

Future Work

- d-separation is the global Markov property for DAGs. What is the equivalent for r-factorizing models? (difficult to do on the original graph, Lauritzen's chain graphs involved, etc.)
- How to do belief propagation on these models? (Variable elimination already known – see my talk).
- What do continuous r-factorizing models look like? (Copulas?)
- Can it be shown there are no more dimension reducing constraints in marginal DAG models other than

$$(X_i \perp\!\!\!\perp X_j) \text{ in } \frac{p(v)}{p(x_k \mid x_m)}$$

- Causal discovery, applications.
- Finding a job :).

Future Work

- d-separation is the global Markov property for DAGs. What is the equivalent for r-factorizing models? (difficult to do on the original graph, Lauritzen's chain graphs involved, etc.)
- How to do belief propagation on these models? (Variable elimination already known – see my talk).
- What do continuous r-factorizing models look like? (Copulas?)
- Can it be shown there are no more dimension reducing constraints in marginal DAG models other than

$$(X_i \perp\!\!\!\perp X_j) \text{ in } \frac{p(v)}{p(x_k \mid x_m)}$$

- Causal discovery, applications.
- Finding a job :).