Statistical Causality

Philip Dawid
Statistical Laboratory
University of Cambridge

Statistical Causality

1. The Problems of Causal Inference
2. Formal Frameworks for Statistical Causality
3. Graphical Representations and Applications
4. Causal Discovery

3. Graphical Representations and Applications

Graphical Representation

- Certain collections of CI properties can be described and manipulated using a DAG representation
  - *very far from complete*
- Each CI property *is represented by* a graphical separation property
  - *d-separation*
  - *moralization*

Moralization: 1

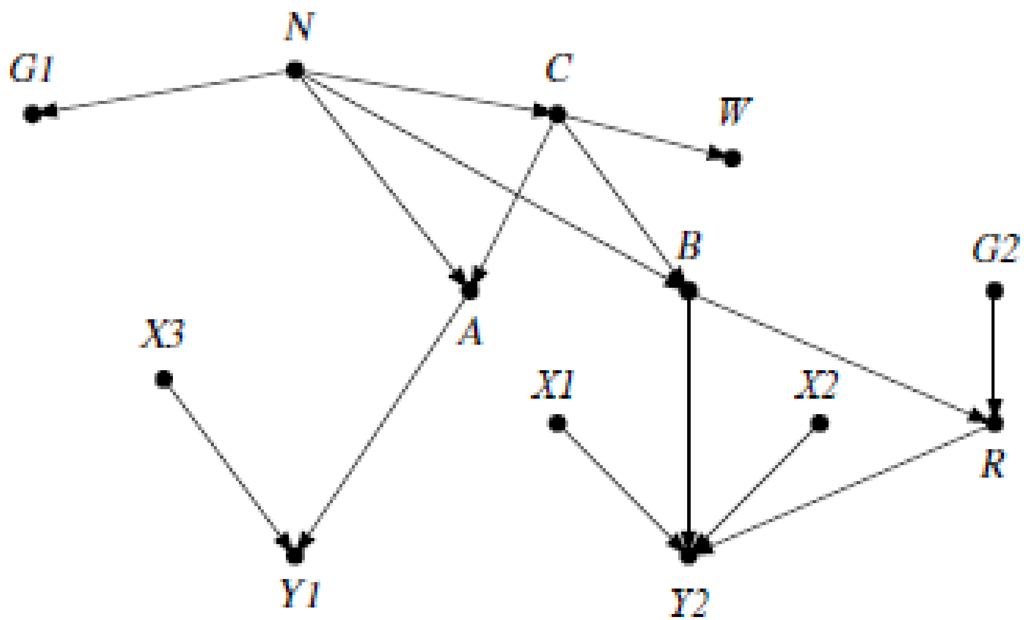


Figure 6.1: Directed graph $\mathcal{D}$ for criminal evidence

Moralization: 1

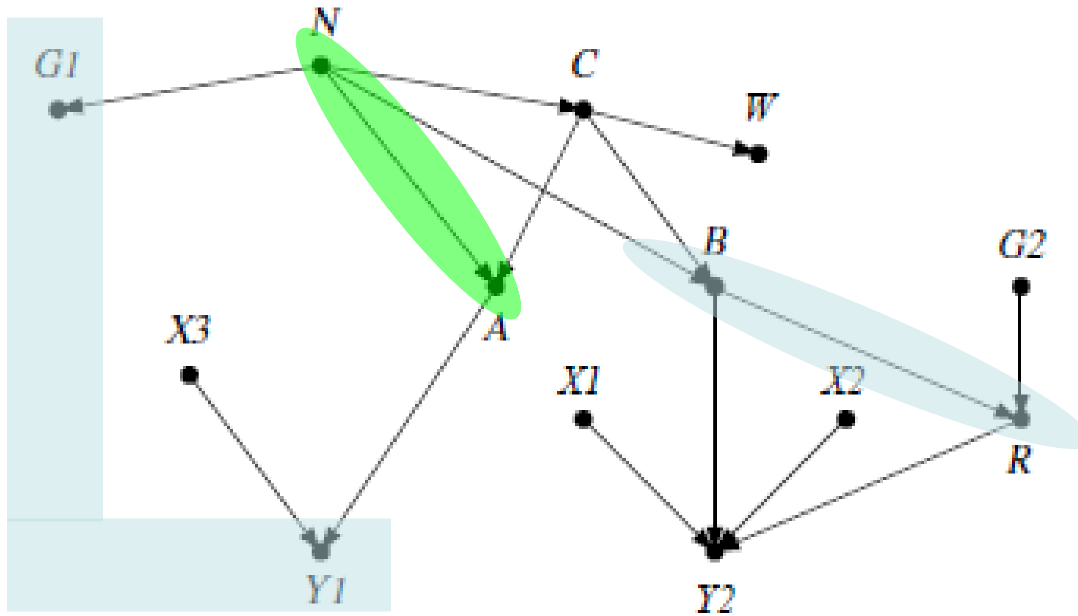


Figure 6.1: Directed graph $\mathcal{D}$ for criminal evidence

$$(B, R) \perp\!\!\!\perp (G1, Y1) \mid (A, N) \ ?$$

Moralization: 2

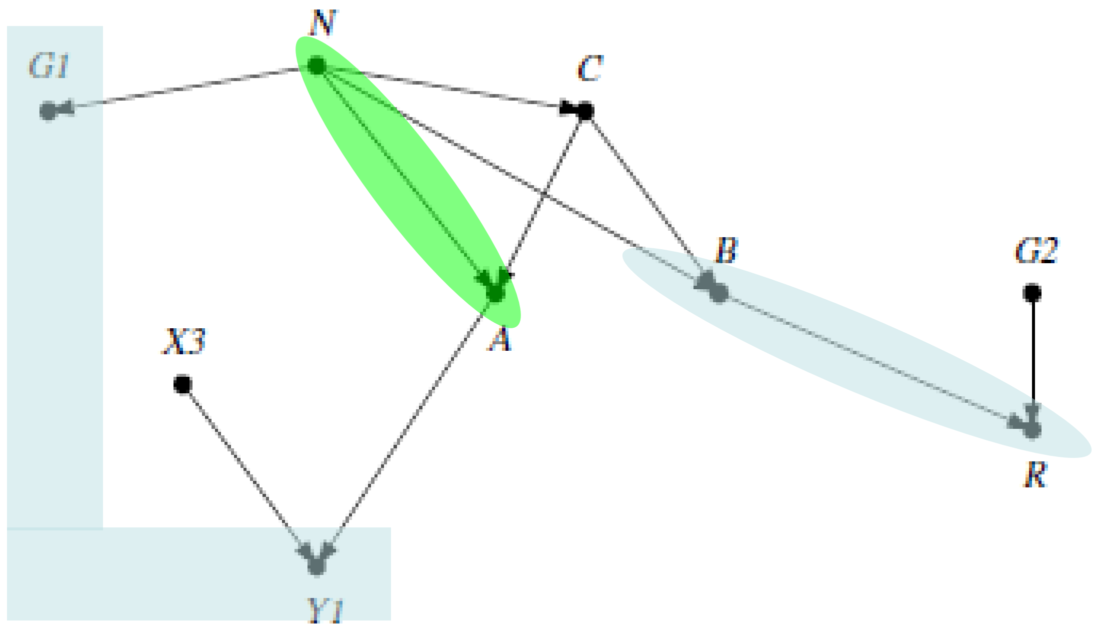


Figure 6.2: Ancestral subgraph $\mathcal{D}'$

$$(B, R) \perp\!\!\!\perp (G1, Y1) \mid (A, N) \ ?$$

Moralization: 3

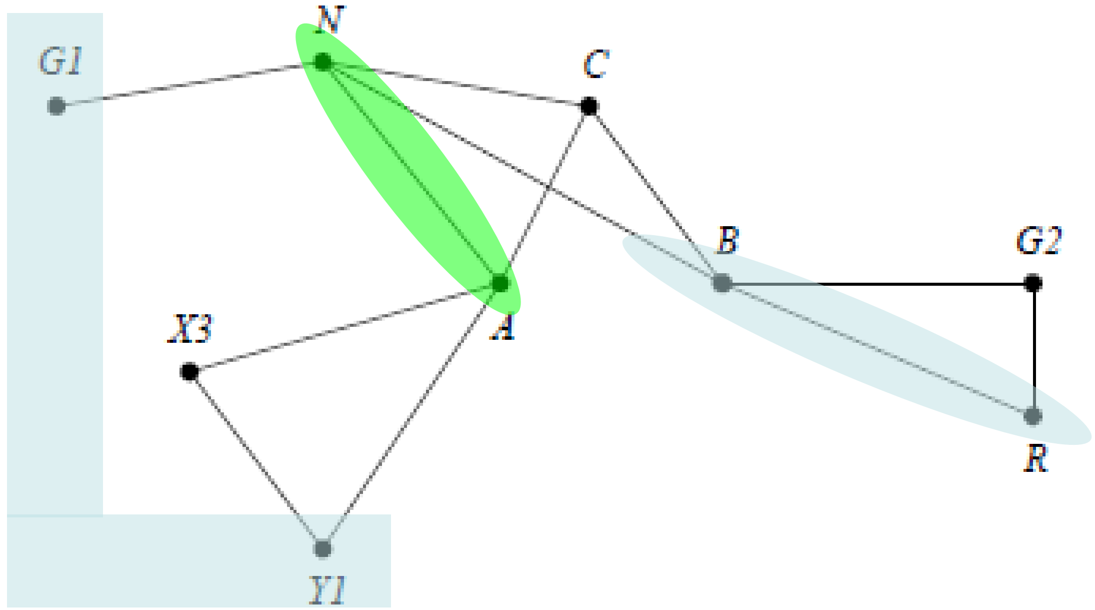


Figure 6.3: Moralized ancestral subgraph $\mathcal{G}'$

$$(B, R) \perp\!\!\!\perp (G1, Y1) \mid (A, N) \ ?$$

Extended Conditional Independence

Distribution of $Y \mid T$ the same in observational and experimental regimes:

$Y \mid (F_T, T)$ does not depend on value of F_T

Can express and manipulate using notation and theory of conditional independence:

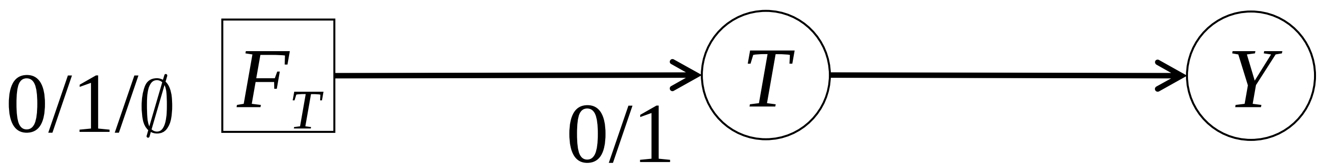
$$Y \perp\!\!\!\perp F_T \mid T$$

(even though F_T is not random)

Augmented DAG

– with random variables and intervention variables

– *probabilistic (not functional) relationships*

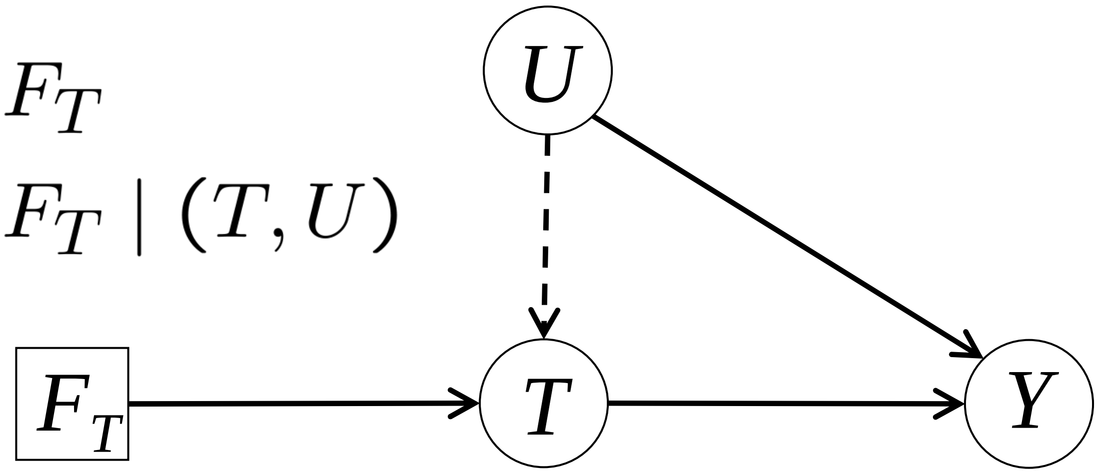


$T \mid (F_T = \emptyset) \sim P_T$ $\qquad$ $Y \mid T$

Absence of arrow $F_T \to Y$ expresses $Y \perp\!\!\!\perp F_T \mid T$

Sufficient Covariate
"(un)confounder"

$$U \perp\!\!\!\perp F_T$$
$$Y \perp\!\!\!\perp F_T \mid (T, U)$$



- Treatment assignment ignorable given U
  - (generally) *not* marginally ignorable
- If U is observed, can fit model (e.g. regression) for dependence of Y on (T,U)
  - causally meaningful

$$\mathrm{ACE}(u) := \mathrm{E}(Y \mid T = 1, U = u) - \mathrm{E}(Y \mid T = 0, U = u)$$

Sufficient covariate “(un)confounder”

$$U \perp\!\!\!\perp F_T$$

$$Y \perp\!\!\!\perp F_T \mid (T, U)$$

Can estimate ACE:

$$\begin{aligned}
\mathrm{E}(Y \mid F_T = t) &= \mathrm{E}\{\mathrm{E}(Y \mid U, F_T = t) \mid F_T = t)\} \\
&= \mathrm{E}\{\mathrm{E}(Y \mid U, F_T = t, T = t) \mid F_T = t)\} \\
&= \mathrm{E}\{\mathrm{E}(Y \mid U, T = t)\} \\
\mathrm{ACE} &= \mathrm{E}\{\mathrm{ACE}(U)\}
\end{aligned}$$

(“back-door” formula)

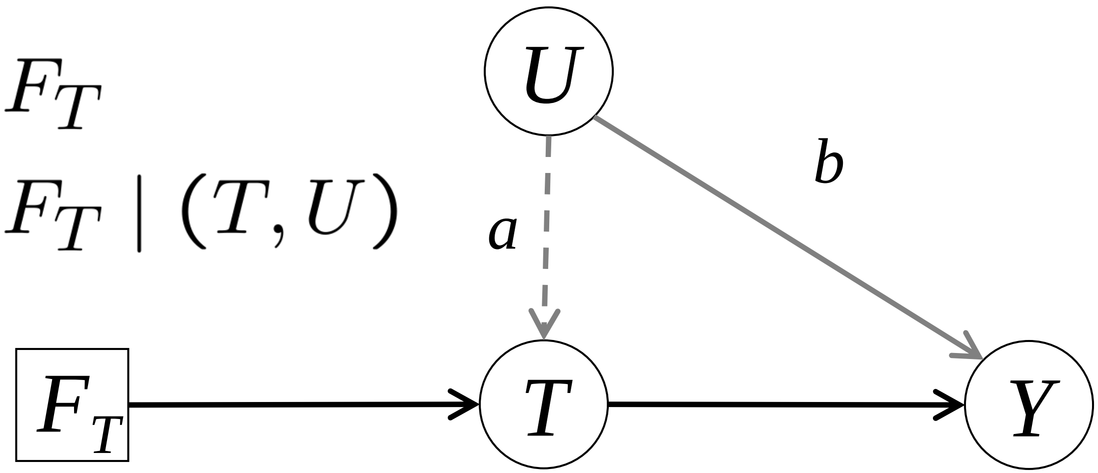
Similarly, whole interventional distribution:

$$p(y \mid F_T = t) = \int p(y \mid u, T = t)\, p(u)\, du$$

Non-confounding

$$U \perp\!\!\!\perp F_T$$

$$Y \perp\!\!\!\perp F_T \mid (T, U)$$



Treatment assignment ignorable given U

Ignorable marginally if either *a* or *b* is absent:

a $T \perp\!\!\!\perp U \mid F_T$

“randomization”

b $Y \perp\!\!\!\perp U \mid T$

“irrelevance”

–then need not even observe U

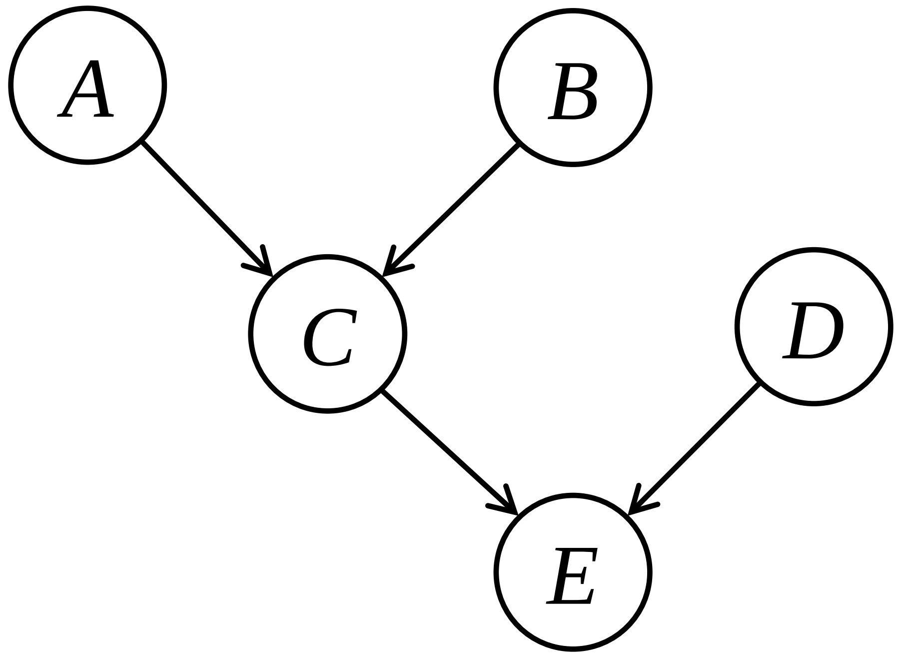
Pearlian DAG

- Envisage intervention on every variable in the system

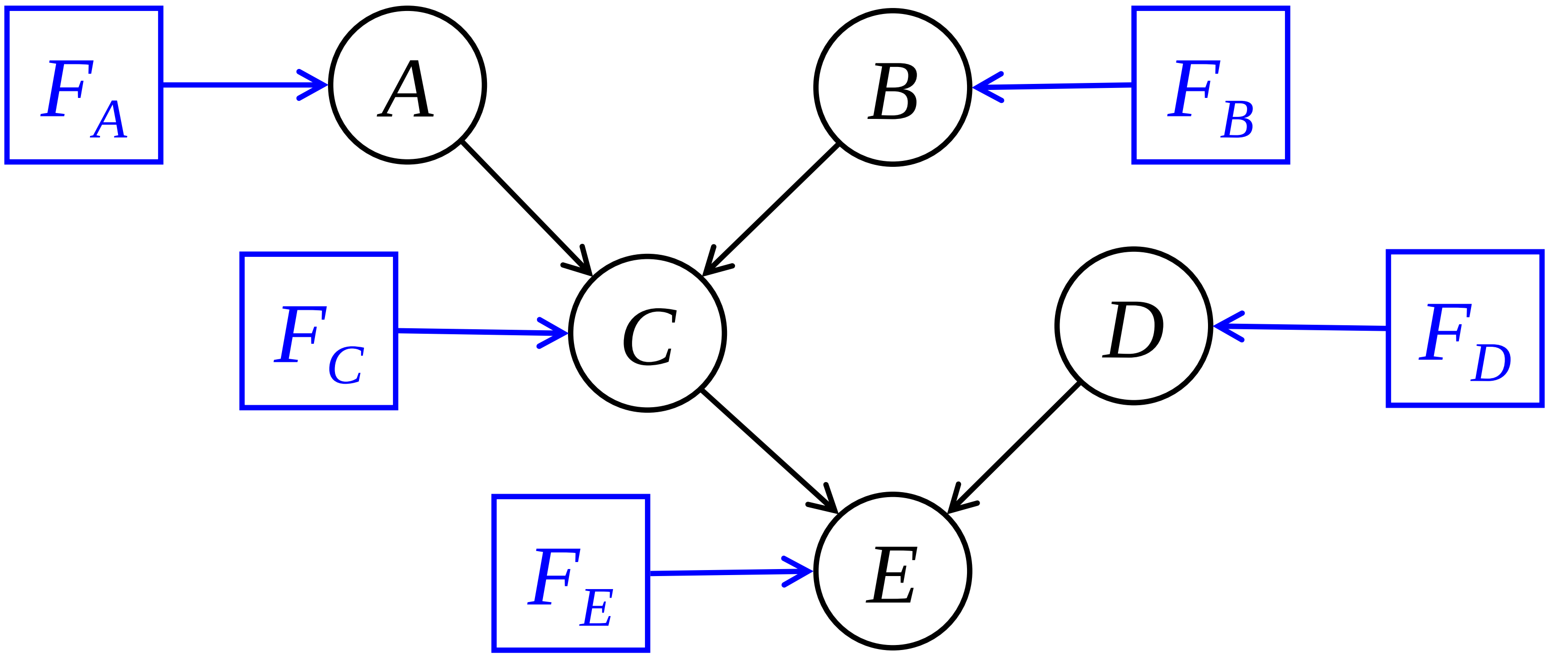
- Augmented DAG model
  - but with intervention indicators implicit

- Every arrow has a causal interpretation

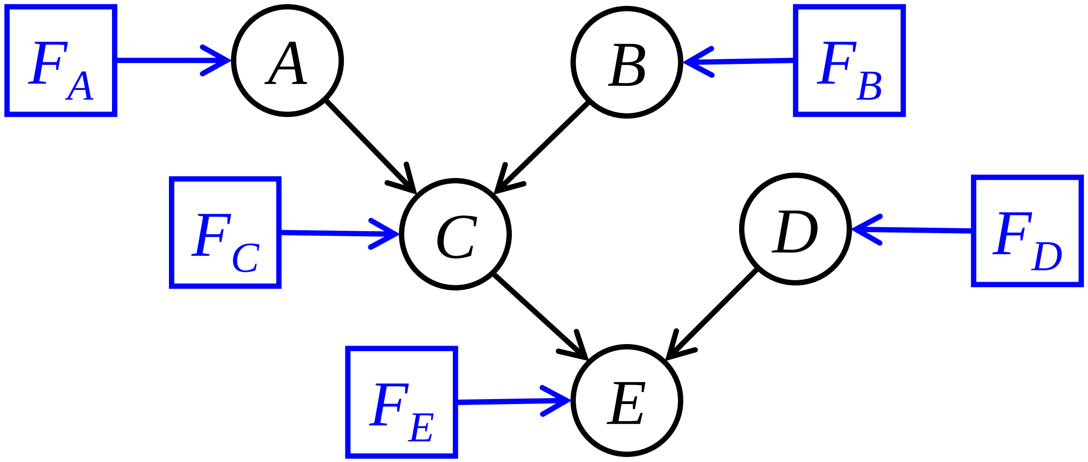
Pearlian DAG



Intervention DAG

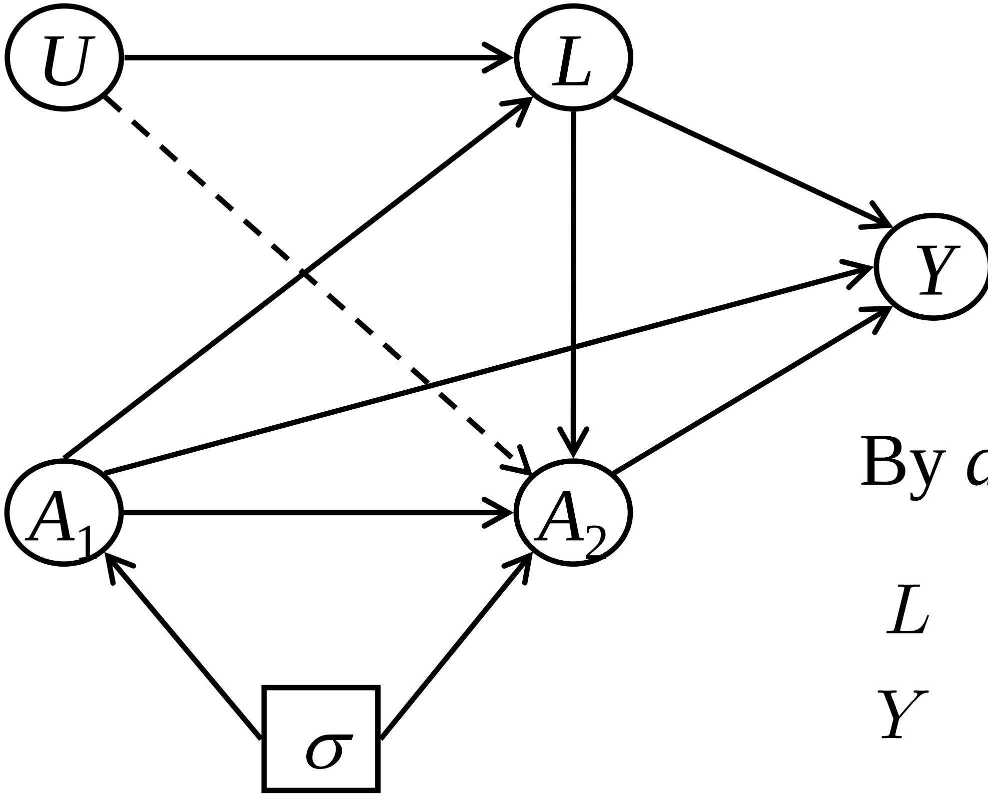


Intervention DAG



- *e.g.*, $E \perp\!\!\!\perp (A, B, F_A, F_B, F_C, F_D) \mid (C, D, F_E)$
- When E is not manipulated, its conditional distribution, given its parents C, D is unaffected by the values of A, B and by whether or not any of the other variables is manipulated
  - modular component

More complex DAGs



σ = treatment strategy

(influence diagram)

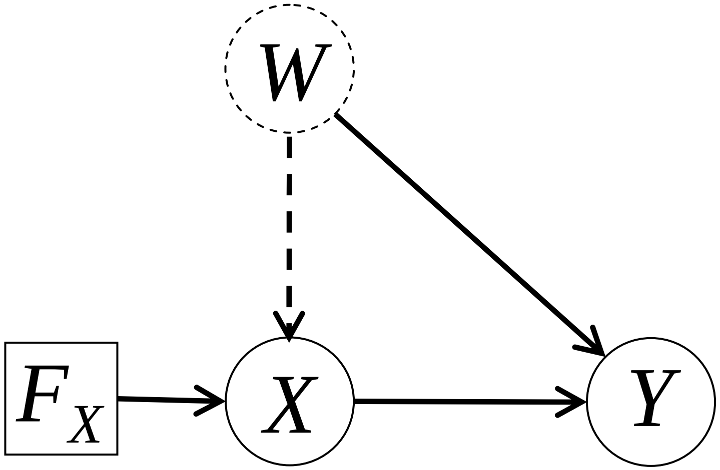
By *d*-separation:

$$L \perp\!\!\!\perp \sigma \mid A_1$$
$$Y \perp\!\!\!\perp \sigma \mid A_1, A_2, L$$

(would fail if *e.g.* $U \rightarrow Y$)

$$p(y \mid \sigma) = \int da_1\, dl\, da_2\; p_\sigma(a_1) p(l \mid a_1) p_\sigma(a_2 \mid a_1, l) p(y \mid a_2, a_2, l)$$

Instrumental Variable



$$W \perp\!\!\!\perp F_X$$
$$Y \perp\!\!\!\perp F_X \mid (X, W)$$

Linear model: $\mathrm{E}(Y \mid X{=}x,\, W,\, F_X = x\,) = f(W) + \beta\, x$

So $\mathrm{E}(Y \mid F_X = x) = \mathrm{E}\{f(W) \mid F_X = x\} + \beta\, x$

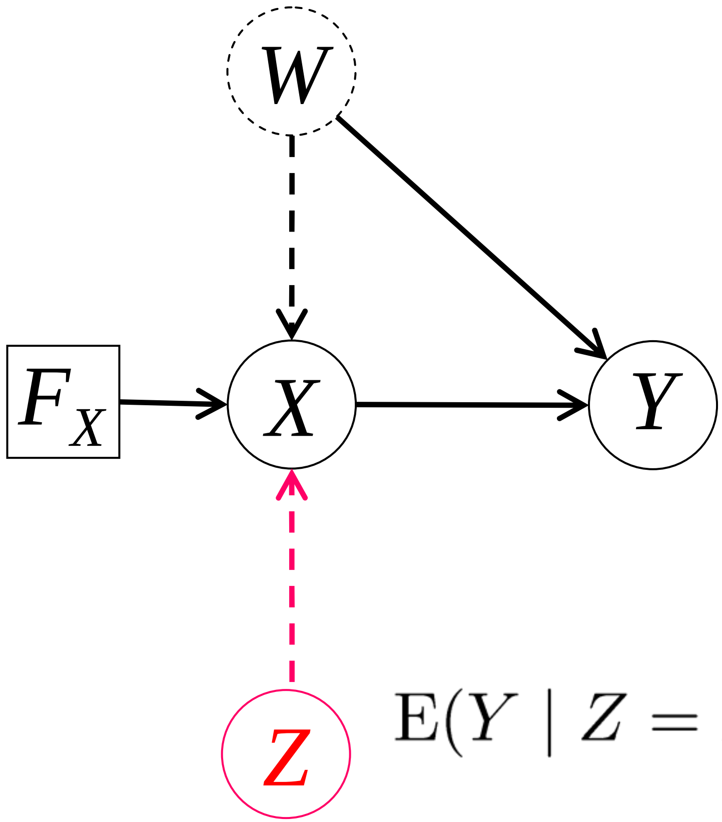
$$= \alpha + \beta\, x$$

- β is causal regression coefficient

– but not estimable from observational data:

$$\mathrm{E}(Y \mid X{=}x) = \mathrm{E}\{f(W) \mid X = x\} + \beta\, x$$

Instrumental Variable



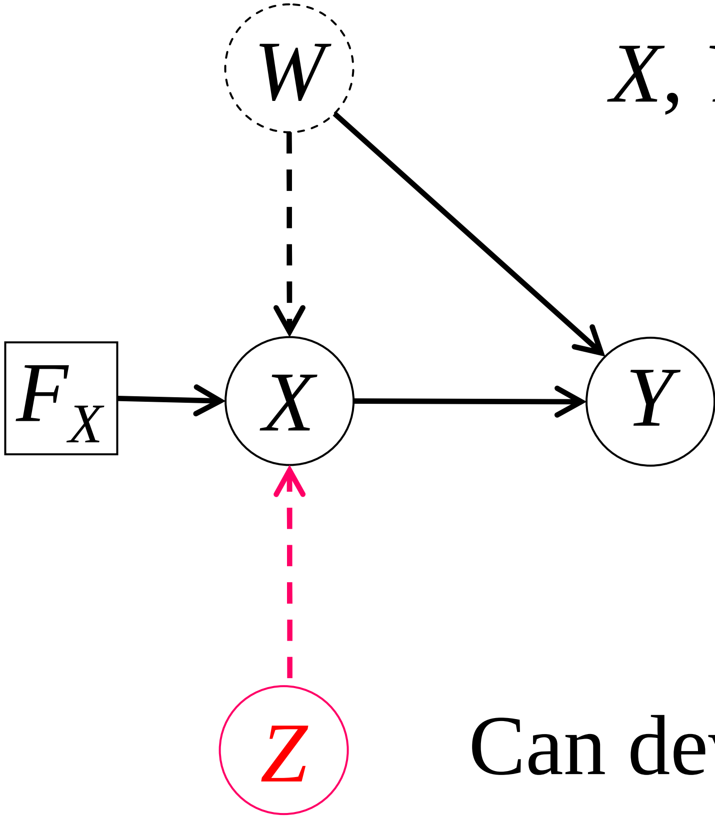
$$
\begin{array}{rcl}
W & \perp\!\!\!\perp & F_X \\
Y & \perp\!\!\!\perp & F_X \mid (X, W) \\
Z & \perp\!\!\!\perp & (W, F_X) \\
Y & \perp\!\!\!\perp & Z \mid (X, W; F_X)
\end{array}
$$

$$
\begin{aligned}
\mathrm{E}(Y \mid Z = z) &= \mathrm{E}\{\mathrm{E}(Y \mid Z, X, W) \mid Z = z\} \\
&= \mathrm{E}\{f(W) + \beta X \mid Z = z\} \\
&= \alpha + \beta\, \mathrm{E}(X \mid Z = z)
\end{aligned}
$$

–so can now identify β

Discrete case

X, Y, Z binary

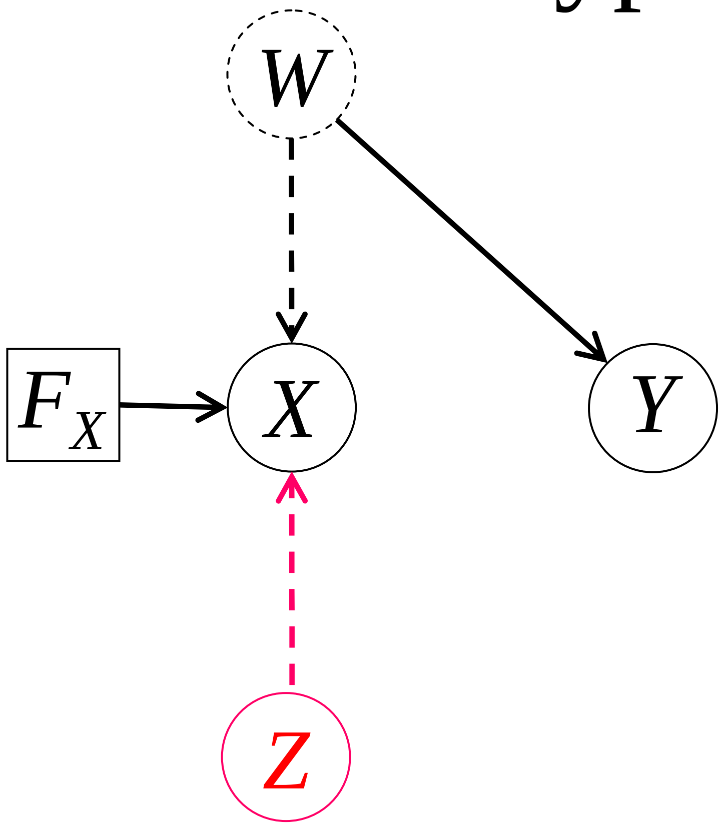


Can develop inequalities for ACE

$$E(Y \mid F_X = 1) - E(Y \mid F_X = 0)$$

in terms of estimable quantities

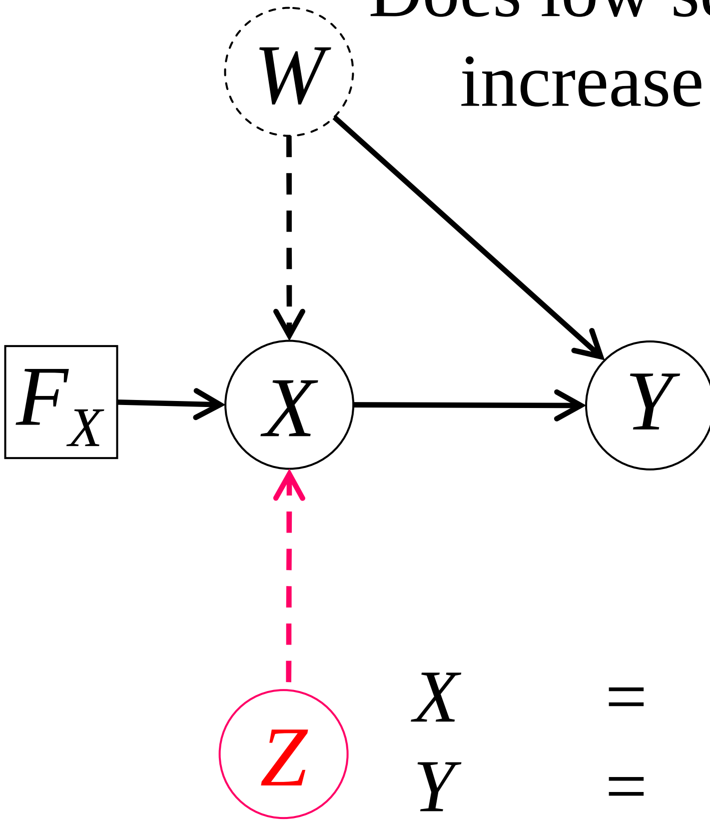
Hypothesis Test



If arrow $X \to Y$ missing, then $Y \perp\!\!\!\perp F_X$
(X has no causal effect on Y).
In this case $Y \perp\!\!\!\perp Z \mid F_X = \emptyset$—can test.

Mendelian Randomisation

Does low serum cholesterol level increase the risk of cancer?

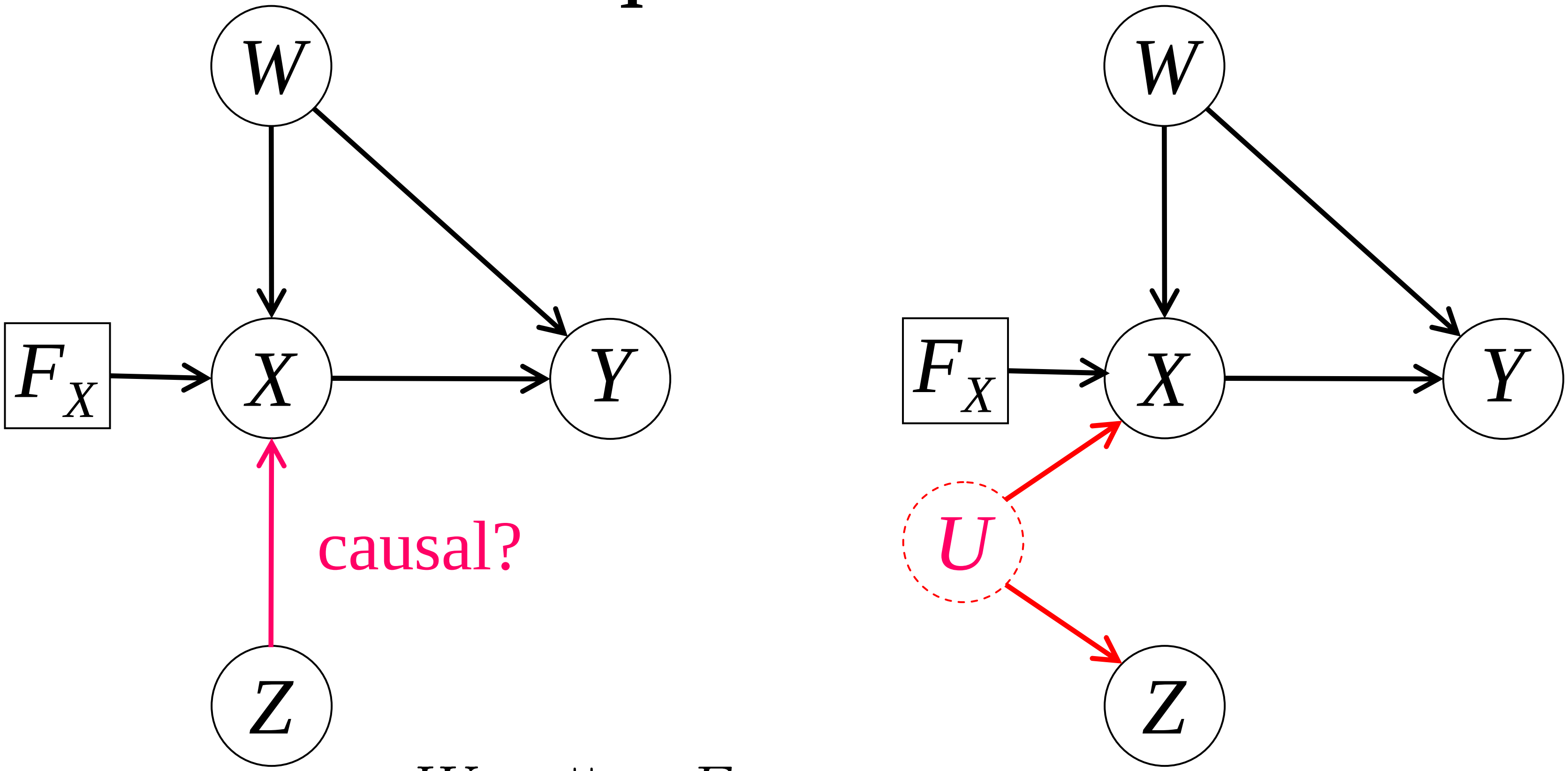


$$
\begin{aligned}
W &\perp\!\!\!\perp F_X \\
Y &\perp\!\!\!\perp F_X \mid (X, W) \\
{\color{red}Z} &{\color{red}\perp\!\!\!\perp (W, F_X)} \\
{\color{red}Y} &{\color{red}\perp\!\!\!\perp Z \mid (X, W; F_X)}
\end{aligned}
$$

X	=	serum cholesterol
Y	=	cancer
W	=	diet, smoking, hidden tumour,…
Z	=	APOE gene

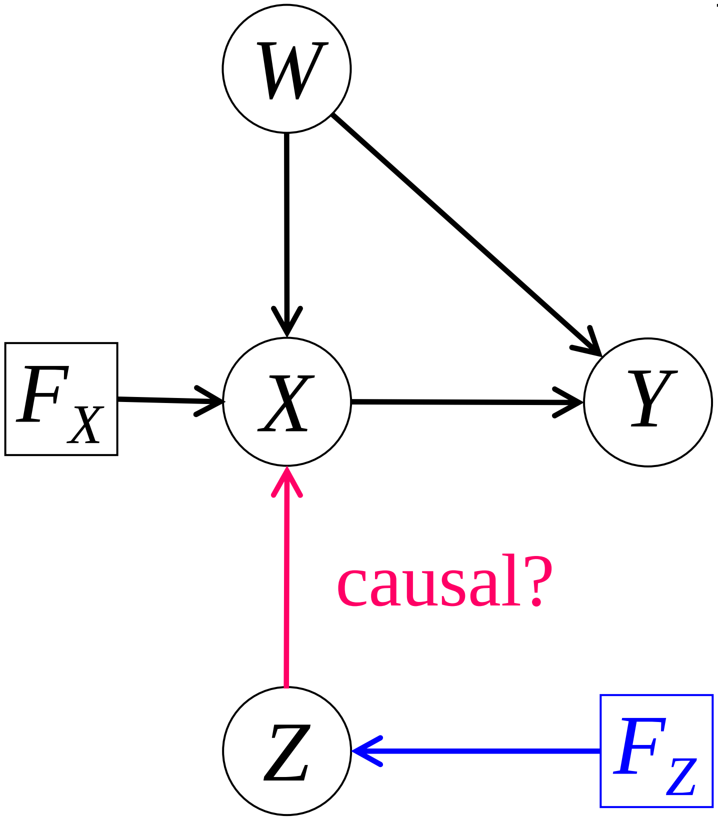
(E2 allele induces particularly low serum cholesterol)

Equivalence

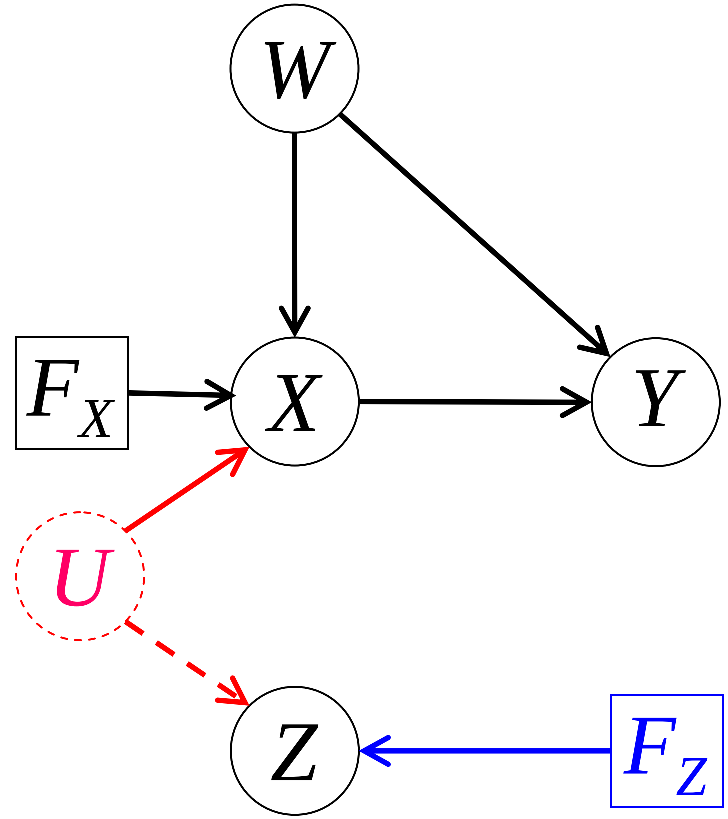


$$
\begin{aligned}
W &\perp\!\!\!\perp F_X \\
Y &\perp\!\!\!\perp F_X \mid (X, W) \\
Z &\perp\!\!\!\perp (W, F_X) \\
Y &\perp\!\!\!\perp Z \mid (X, W; F_X)
\end{aligned}
$$

Non-equivalence



$$X \not\perp\!\!\!\perp Z \mid F_Z \neq \emptyset$$

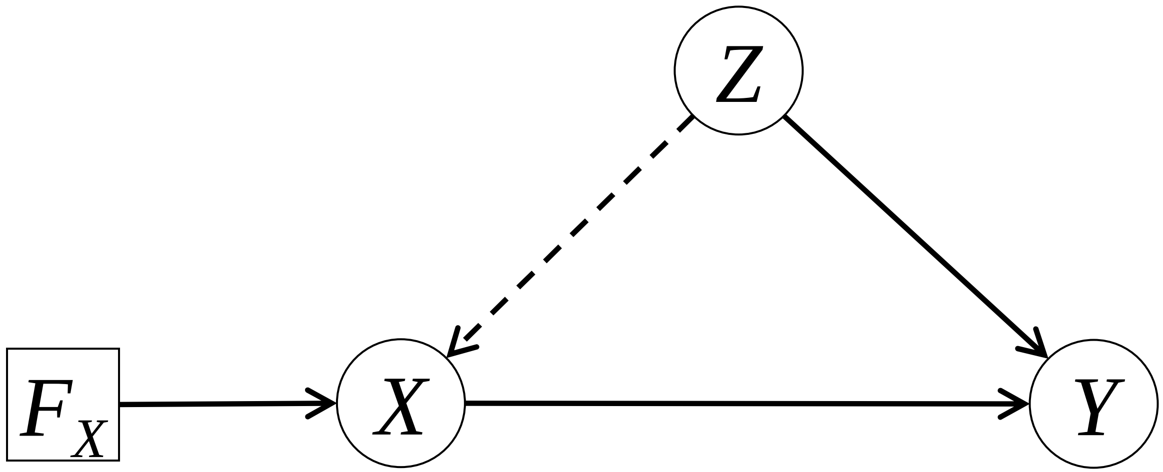


$$X \perp\!\!\!\perp Z \mid F_Z \neq \emptyset$$

Can we identify a causal effect from observational data?

- Model with domain and (explicit or implicit) intervention variables, specified ECI properties
  - e.g. augmented DAG, Pearlian DAG
- Observed variables $\mathcal{V}$, unobserved variables $\mathcal{U}$
- Can identify observational distribution over $\mathcal{V}$
- Want to answer causal query, e.g. $p(y \mid F_X = x)$
  - write as $p(y \mid \check{x})$
- When/how can this be done?

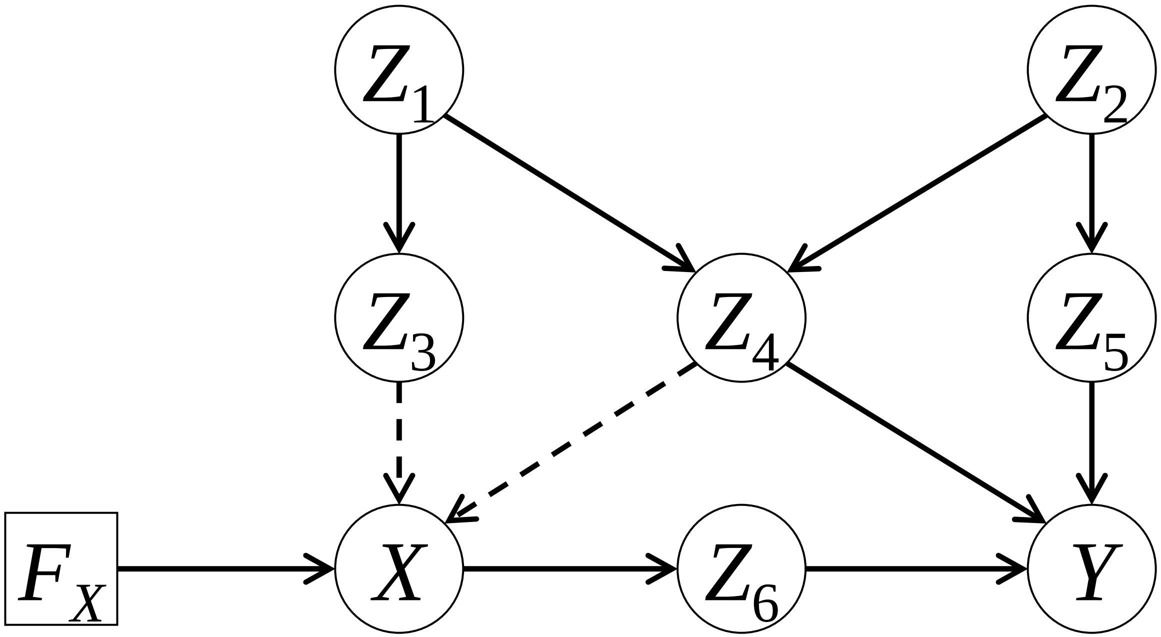
Example: “back-door formula”



$$Z \perp\!\!\!\perp F_X$$
$$Y \perp\!\!\!\perp F_X \mid (X, Z)$$

$$p(y \mid \check{x}) = \sum_z p(y \mid x, z)\, p(z)$$

Example: “back-door formula”

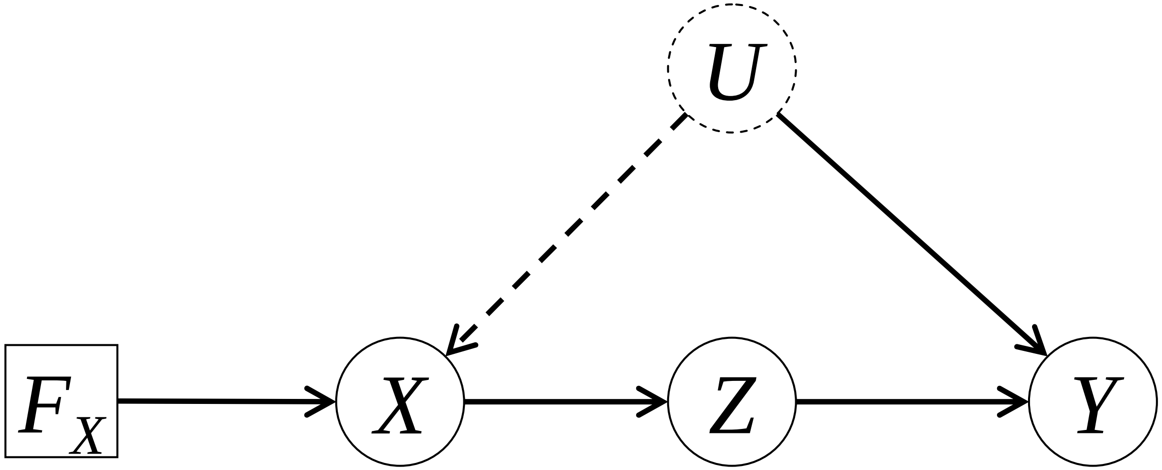


$$Z \perp\!\!\!\perp F_X$$

$$Y \perp\!\!\!\perp F_X \mid (X, Z)$$

Works for $Z = (Z_3, Z_4)$, and also for $Z = (Z_4, Z_5)$

Example: “front-door formula”



$$
\begin{aligned}
U &\perp\!\!\!\perp F_X \\
Z &\perp\!\!\!\perp (U, F_X) \mid X \\
Y &\perp\!\!\!\perp (F_X, X) \mid (Z, U)
\end{aligned}
$$

$$p(y \mid \check{x}) = \sum_z p(z \mid x) \sum_{x'} p(y \mid x', z)\, p(x').$$

do-calculus

Rule 1 (Insertion/deletion of observations)

If $Y \perp\!\!\!\perp Z \mid (X, F_X \neq \emptyset, W)$ then

$$p(y \mid \check{x}, z, w) = p(y \mid \check{x}, w)$$

Rule 2 (Action/observation exchange)

If $Y \perp\!\!\!\perp F_Z \mid (X, F_X \neq \emptyset, Z, W)$, then

$$p(y \mid \check{x}, \check{z}, w) = p(y \mid \check{x}, z, w)$$

Rule 3 (Insertion/deletion of actions)

If $Y \perp\!\!\!\perp F_Z \mid (X, F_X \neq \emptyset, W)$, then

$$p(y \mid \check{x}, \check{z}, w) = p(y \mid \check{x}, w)$$

do-calculus

For a problem modelled by a Pearlian DAG, the *do*-calculus is complete:

- We can tell whether a given causal effect is computable (from the observational distribution)
- Any computable causal effect can be computed by successive applications of rules 2 and 3
  - together with probability calculus, and property $F_T = t \Rightarrow T = t$ (delete dotted arrows)
- There exist algorithms to accomplish this

4. Causal Discovery

Probabilistic Causality

- *Intuitive concepts* of “cause”, “direct cause”,…
- Principle of the common cause:

  “Variables are independent, given their common causes”

- Assume *causal DAG* representation:
  - direct causes of *V* are its DAG parents
  - all “common causes” included

Probabilistic Causality

CAUSAL MARKOV CONDITION

– The causal DAG also represents the observational conditional independence properties of the variables

- WHEN??
- WHY??

- **CAUSAL FAITHFULNESS CONDITION**

– No extra conditional independencies

- WHY??

Causal Discovery

- An attempt to learn causal relationships from observational data
- Assume there is an underlying *causal DAG* (possibly including unobserved variables) satisfying the (faithful) Causal Markov Condition
- Use data to search for a DAG representing the observational independencies
  - ➢ *model selection*
- Give this a causal interpretation

Causal Discovery

Two main approaches:

- “Constraint-based”
  - Qualitative
  - Infer (patent or latent) conditional independencies between variables
  - Fit conforming DAG model(s)
- Statistical model selection
  - Quantitative
  - General approach, applied to DAG models
  - Need not commit to one model (model uncertainty)

Constraint-Based Methods (complete data)

- Identify/estimate conditional independencies holding between observed variables
- Assume sought-for causal DAG does not involve any variables other than those observed

Wermuth-Lauritzen algorithm

- Assume variables are "causally ordered" *a priori*:

  $(V_1, V_2, \ldots, V_N)$, s.t arrows can only go from lower to higher

- For each i, identify (smallest) subset S_i of

  $V^{i\text{-}1} := (V_1, V_2, \ldots, V_{i\text{-}1})$ such that

$$V_i \perp\!\!\!\perp V^{i-1} \mid S_i$$

- Draw arrow from each member of S_i to V_i

SGS algorithm (no prior ordering)

1. Start with complete undirected graph over V^N
2. Remove edges V–W s.t., for some S, $V \perp\!\!\!\perp W \mid S$
3. Orient any V–Z–W as $V \to Z \leftarrow W$ if:
   - no edge V–W
   - for each $S \subseteq V^N$ with $Z \in S$, $V \not\perp\!\!\!\perp W \mid S$
4. Repeat while still possible:

   i. if $V \to Z$ –W but not V–W, orient as $V \to Z \to W$

   ii. If $V \rightsquigarrow W$ and V–W, orient as $V \to W$

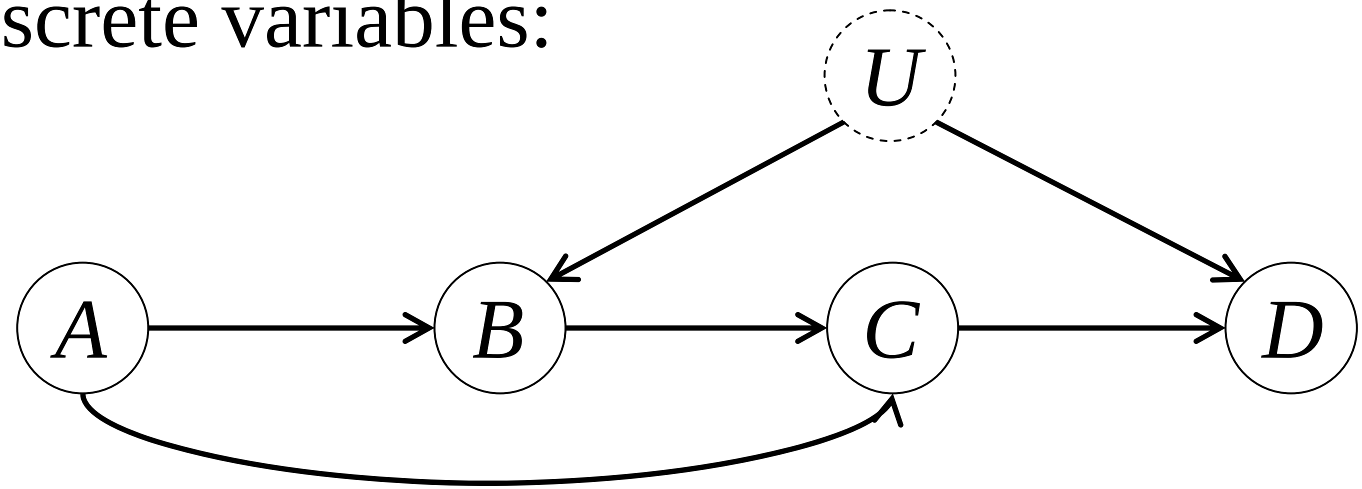
Comments

- Wermuth-Lauritzen algorithm
  - always finds a valid DAG representation
  - need not be faithful
  - depends on prior ordering
- SGS algorithm
  - may not succeed if there is no faithful DAG representation
  - output may not be fully oriented
  - computationally inefficient (too many tests)
  - better variations: PC, PC*

Constraint-Based Methods (incomplete data)

- Allow now for unobserved (latent) variables
- Can modify previous algorithms to work just with conditional independencies between observed variables
- But latent CI has other (quantitative) implications too…

Discrete variables:



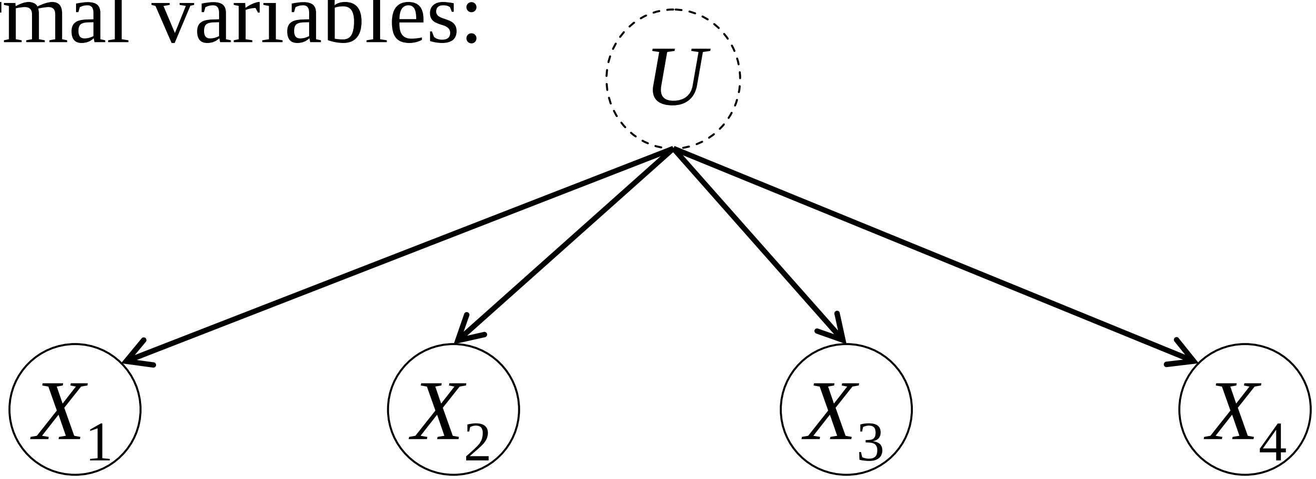
No CI properties between observables *A*, *B*, *C*, *D*.

But

$$\begin{aligned}\sum_b p(b \mid a)p(d \mid a, b, c) &= \sum_b p(b \mid a) \sum_u p(d \mid a, b, c, u)p(u \mid a, b, c) \\ &= \sum_u p(d \mid c, u)p(u \mid a)\end{aligned}$$

– does not depend on *a*

Normal variables:



No CI properties between observables X_1, X_2, X_3, X_4.

But $\rho_{13}\rho_{24} = \rho_{14}\rho_{23} = \rho_{12}\rho_{34}$

Such properties form basis of TETRAD II program

Bayesian Model Selection

- Consider collection $\mathcal{M} = \{M\}$ of models
- Have prior distribution $\pi_M(\boldsymbol{\theta}_M)$ for parameter $\boldsymbol{\theta}_M$ of model M
- Based on data $\boldsymbol{x}$, compute *marginal likelihood* for each model M:

$$L_M = \int p(\boldsymbol{x} \mid \boldsymbol{\theta}_M)\, \mathrm{d}\boldsymbol{\theta}_M$$

- Use as score for comparing models, or combine with prior distribution $\{w_M\}$ over models to get posterior:

$$w^*_M \propto w_M\, L_M$$

Bayesian Model Selection

- Algebraically straightforward for discrete or Gaussian DAG models, parametrised by parent-child conditional distributions, having conjugate priors (with local and global independence)
  - Zoubin Ghahramani's lectures
- Can arrange hyperparameters so that indistinguishable (Markov equivalent) models get same score

Mixed data

- Data from experimental and observational regimes
- Model-selection approach:
  - assume Pearlian DAG
  - ignore local likelihood contribution when the response variable is set
- Constraint-based approach?
  - base on ECI properties, e.g. $X \perp\!\!\!\perp F_Y \mid (W, F_Z)$

A Parting Caution

- We have powerful statistical methods for attacking causal problems
- But to apply them we have to make strong assumptions (e.g. ECI assumptions, relating distinct regimes)
- Important to consider and justify these in context
  - *e.g.*, Mendelian randomisation

"NO CAUSES IN, NO CAUSES OUT"

Thank you!

Further Reading

- A. P. Dawid (2007). *Fundamentals of Statistical Causality*. Research Report 279, Department of Statistical Science, University College London. 94 *pp*. http://www.ucl.ac.uk/Stats/research/reports/abs07.html#279
- R. E. Neapolitan (2003). *Learning Bayesian Networks*. Prentice Hall, Upper Saddle River, New Jersey.
- J. Pearl (2009). *Causality: Models, Reasoning and Inference* (second edition). Cambridge University Press.
- D. B. Rubin (1978). Bayesian inference for causal effects: the role of randomization. *Annals of Statistics* **6**, 34–68.
- P. Spirtes, C. Glymour, and R. Scheines (2000). *Causation, Prediction and Search* (second edition). Springer-Verlag, New York.
- P. Suppes (1970). *A Probabilistic Theory of Causality*. North Holland, Amsterdam.