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

1. The Problems of Causal Inference

Conceptions of Causality

- Constant conjunction
  - Deterministic
- Mechanisms
  - "Physical" causality
- Agency
  - Effects of actions/interventions
- Contrast
  - Variation of effect with changes to cause

Causal Queries

- If I had taken aspirin half an hour ago, would my headache would have gone by now?
  - "Causes of Effects", CoE
  - *Counterfactual*
  - LAW
- If I take aspirin now, will my headache be gone within half an hour?
  - "Effects of Causes", EoC
  - *Hypothetical*
  - **SCIENCE, STATISTICS**

Causal Enquiry

- Experimentation (“doing”)
  - To find out what happens to a system when you interfere with it you have to interfere with it (not just passively observe it) – George Box
- Observation (“seeing”)
  - Cost
  - Ethics
  - Practicality
- No necessary connexion!

Problems of observational studies

An observed association between a “cause” and an “effect” may be spurious:

- Reverse causation
- Regression to mean
- Confounding
  - common cause
  - differential selection
- …

Problems of observational studies

The facts about fuel (Which?, August 2007)

Mr Holloway said that a colleague of his used to drive from London to Leeds and back, using Shell petrol to go up there and BP fuel to drive back. He was convinced the BP petrol gave better fuel economy, but Ray had another explanation: ‘I pointed out that Leeds is at a higher altitude that London: he was going uphill one way and downhill the other!’

Problems of observational studies

Vitamin supplements and mortality

Many observational studies appeared to indicate that antioxidant supplements (vitamins A and E, β-carotene) reduce the risk of disease.

Randomized controlled trials showed that they increase mortality.

Problems of observational studies

Calcium channel blockers

Non-experimental studies suggested an increased risk of myocardial infarction associated with the short-acting calcium channel blocker (CCB) nifedapine.

It took almost a decade to obtain RCT evidence, which showed that long-acting nifedapine is safe.

Simpson's Paradox

	Recovered	Died	Total	Recovery rate
New	200	200	400	50%
Standard	160	240	400	40%

Table 1: Overall results

	Recovered	Died	Total	Recovery rate
New	180	120	300	60%
Standard	70	30	100	70%

Table 2: Male results

	Recovered	Died	Total	Recovery rate
New	20	80	100	20%
Standard	90	220	300	30%

Table 3: Female results

Causal Inference

- Association is not causation!
- Traditionally, Statistics dealt with association
  - Theory of *Statistical Experimental Design and Analysis* does address causal issues
    - but of no real use for observational studies
- How to make inferences about causation?
  - "bold induction", to a novel context
- *Do we need a new formal framework?*

2. Formal Frameworks for Statistical Causality

Some Formal Frameworks

- Probability distributions
- Potential responses
- Functional relationships
- Extended conditional independence
- …
- Structural equations
- Path diagrams
- Directed acyclic graphs
- …

A SIMPLE (??) PROBLEM

- **Randomised experiment**
- Binary (0/1) treatment decision variable T
- Response variable Y

Define/measure "*the effect of treatment*"

Probability Model *(Fisher)*

- Specify/estimate *conditional distributions*

  P_t for Y *given* $T = t$ $(t = 0, 1)$

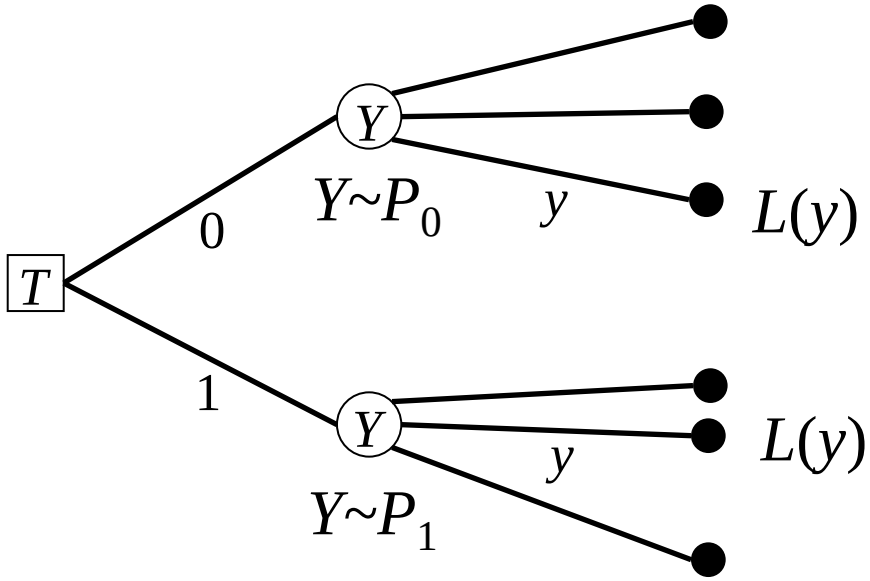
  [*e.g.* $N(\mu_t, \sigma^2)$]

- Measure effect of treatment by change in the distribution of Y: compare P_0 and P_1
  - *e.g.* by change in expected response:

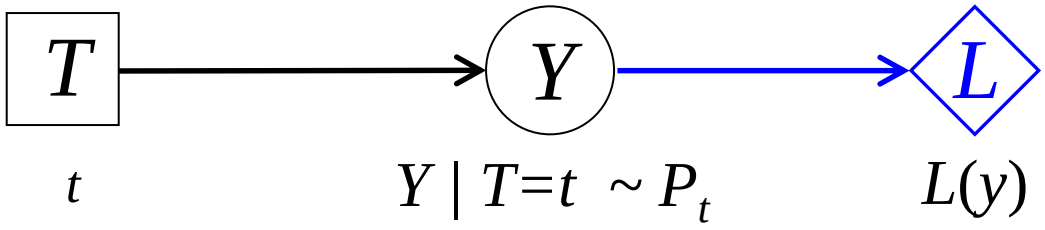
    $\delta = \mu_1 - \mu_0$ (*average causal effect, ACE)*

- Probability model all we need for decision theory
  - choose t to minimise expected loss $\mathrm{E}_{Y \sim P_t}\{L(Y)\}$

Decision Tree



Influence Diagram



Structural Model

$$Y = \mu_T + E$$

[e.g., $E \sim N(0, \sigma^2)$]

(E = "error", "effect of omitted variables",…)

- Deterministic relationship
- Value of E for any unit supposed the same if we were to change T from 0 to 1
- Then value of Y would change by *exactly* $\delta = \mu_1 - \mu_0$
  - *individual causal effect (ICE)*

Potential Response Model (Rubin)

- Split Y in two:

  Y_0 : potential response to $T = 0$

  Y_1 : potential response to $T = 1$

- Consider (for any unit) the *pair* $\mathbf{Y} = (Y_0, Y_1)$
  - *with simultaneous existence and joint distribution*
- Treatment "uncovers" pre-existing response:

  $Y = Y_T$ (determined by $\mathbf{Y}$ and T)
  - other PR unobservable, "counterfactual"
- Unit-level (individual) [random] causal effect

  $Y_1 - Y_0$ – *necessarily unobservable!*

General Functional Model

$$Y = f(T, U)$$

(U = "unit characteristics")

- Value of U supposed the same , independent of value of T
  - and of whether we intervene or just observe
- Formally includes:
  - Structural model: $U = E$, $Y = \mu_T + E$
  - PR model: $U = \mathbf{Y}$, $Y = Y_T$

Potential Response Model

- Any functional model $Y = f(T, U)$ generates a PR model: $Y_t = f(t, U)$
- Any PR model generates a probability model:

  P_t is marginal distribution of Y_t $(t = 0, 1)$

- Distinct PR models can generate the same statistical model
  - e.g., correlation between Y_0 and Y_1 arbitrary
- Cannot be distinguished observationally
- Can have different inferential consequence
  - can be problematic!

Potential Responses: Problems

- *PR* model:

$$\begin{cases} Y_t \sim N(\mu_t, \sigma^2) \quad (t = 0, 1) \\ \qquad \text{corr}(Y_0, Y_1) = \rho \end{cases}$$

- Corresponding *statistical* model:

$$P_t : Y \sim N(\mu_t, \sigma^2)$$

NB: ρ does not enter! – *can never identify* ρ *– does this matter??*

Potential Responses: Problems

Under PR model:

$E(Y_1/Y_0)$ depends on ρ

- *We can not estimate a "ratio" ICE*

$\text{var}(Y_1 - Y_0) = 2(1 - \rho)\,\sigma^2$

- *We can not identify the variance of the ICE*

$E(Y_1 - Y_0 \mid Y_1 = y_1) = (1 - \rho)\,y_1 + (\rho\mu_1 - \mu_0)$

- *We can not identify the (counterfactual) ICE, after observing response to treatment*

OBSERVATIONAL STUDY

- Treatment decision taken may be associated with patient's state of health
- What assumptions are required to make causal inferences?
- When/how can such assumptions be justified?

Functional Model

$$Y = f(T, U)$$

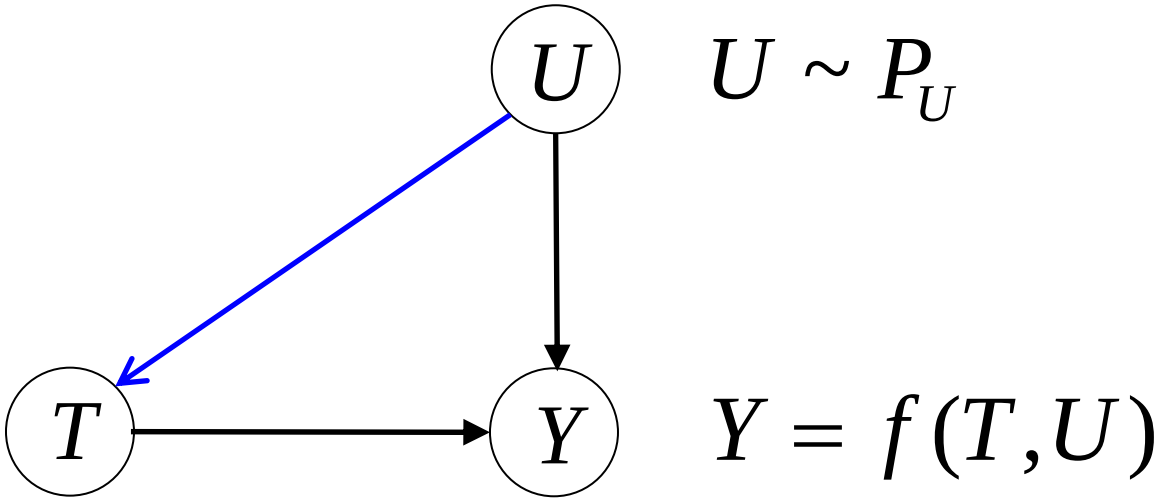
T = treatment received

U = "unit characteristics"

- value supposed unaffected by treatment or how it is applied
- but could influence choice of treatment T
  - observational dependence between T and U

Response to applied treatment t: $Y_t = f(t, U)$.
Observational distribution of Y, *given* $T = t$,
same as distribution of Y_t if $T \perp\!\!\!\perp U$

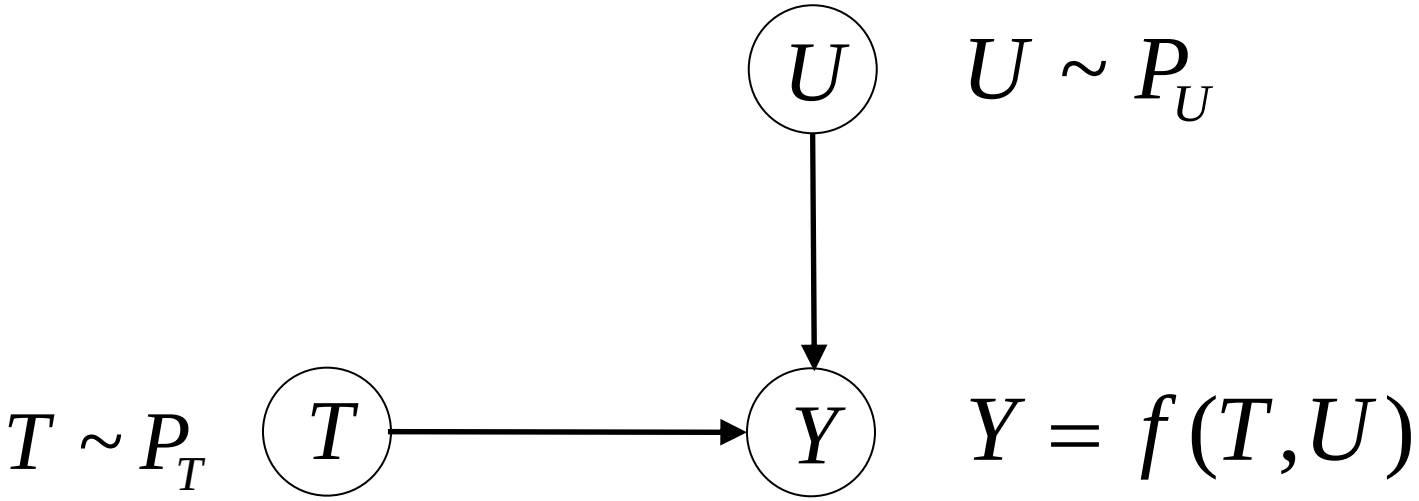
Functional Model



U = "unit characteristics"

- value supposed unaffected by treatment or how it is applied
- but could influence treatment choice

Functional Model

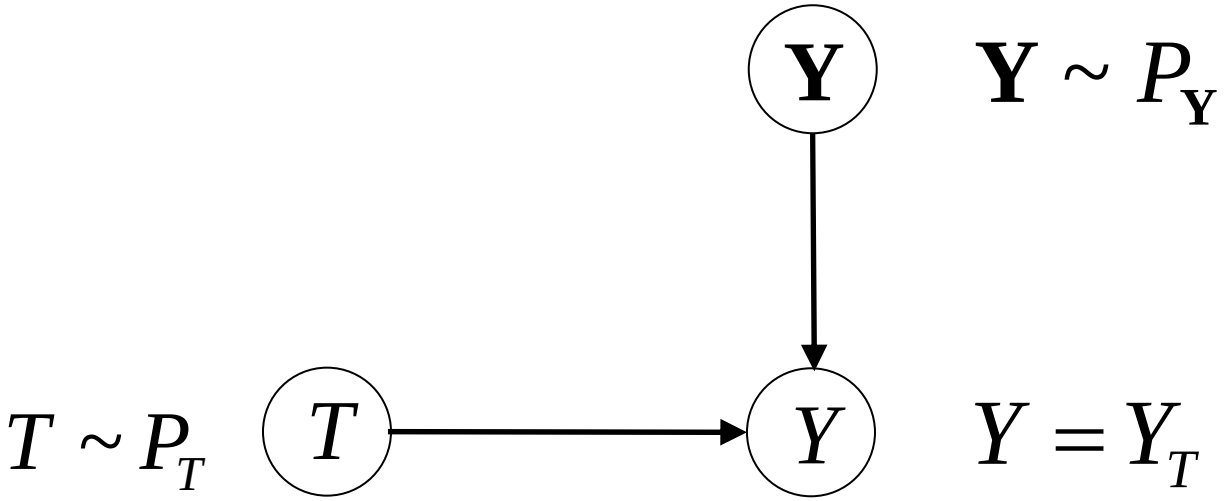


"No confounding" ("ignorable treatment assignment") if

$$T \perp\!\!\!\perp U$$

(treatment independent of "unit characteristics")

PR interpretation ($U = \mathbf{Y}$)



"No confounding" ("ignorable treatment assignment") if

$$T \perp\!\!\!\perp \mathbf{Y}$$

(treatment independent of potential responses)

PR interpretation ($U = \mathbf{Y}$)

- Value of $\mathbf{Y} = (Y_0, Y_1)$ on any unit supposed the same in observational and experimental regimes, as well as for both choices of T
- No confounding: independence of T from PR pair $\mathbf{Y}$

How are we to judge this??

Statistical Decision Model

- "Treatment regime indicator" variable F_T
  - *intervention variable*
  - non-random, parameter

- Values:

$$F_T = 0 : \text{Assign treatment 0} \qquad (\Rightarrow T = 0)$$
$$F_T = 1 : \text{Assign treatment 1} \qquad (\Rightarrow T = 1)$$
$$F_T = \emptyset : \text{Just observe} \qquad (T \text{ random})$$

(*Point intervention*: can generalize)

Statistical Decision Model

- Causal target: comparison of distributions of Y given $F_T = 1$ and given $F_T = 0$
  - e.g., $\mathrm{E}(Y \mid F_T = 1) - \mathrm{E}(Y \mid F_T = 0)$

    average causal effect, ACE

- Causal inference: assess this (if possible) from properties of observational regime, $F_T = \emptyset$

Statistical Decision Model

True ACE is

$$\mathrm{E}(Y \mid T = 1, F_T = 1) - \mathrm{E}(Y \mid T = 0, F_T = 0)$$

Its observational counterpart is:

$$\mathrm{E}(Y \mid T = 1, F_T = \emptyset) - \mathrm{E}(Y \mid T = 0, F_T = \emptyset)$$

"No confounding" (ignorable treatment assignment) when these are equal.

Can strengthen:

$$p(y \mid T = t, F_T = 1) = p(y \mid T = t, F_T = \emptyset)$$

- distribution of $Y \mid T$ the same in observational and experimental regimes

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)

3. Graphical Representations and Applications

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:

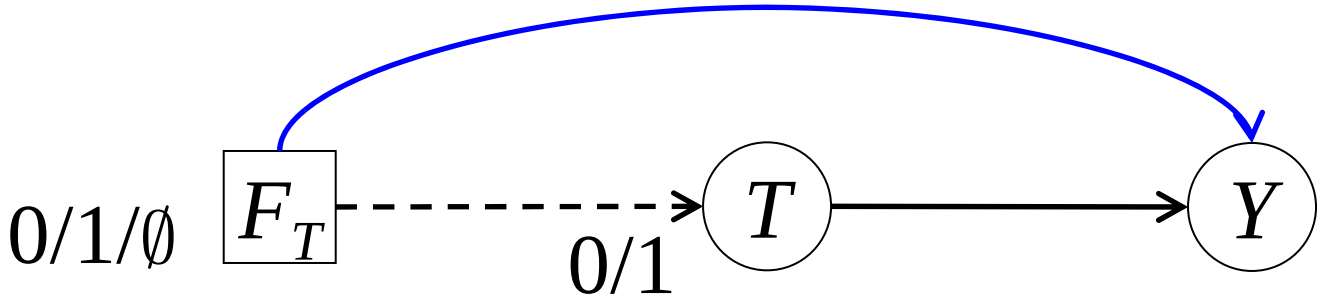
$$Y \perp\!\!\!\perp F_T \mid T$$

(even though F_T is not random)

Augmented DAG

– with random variables and intervention variables

– *probabilistic (not functional) relationships*



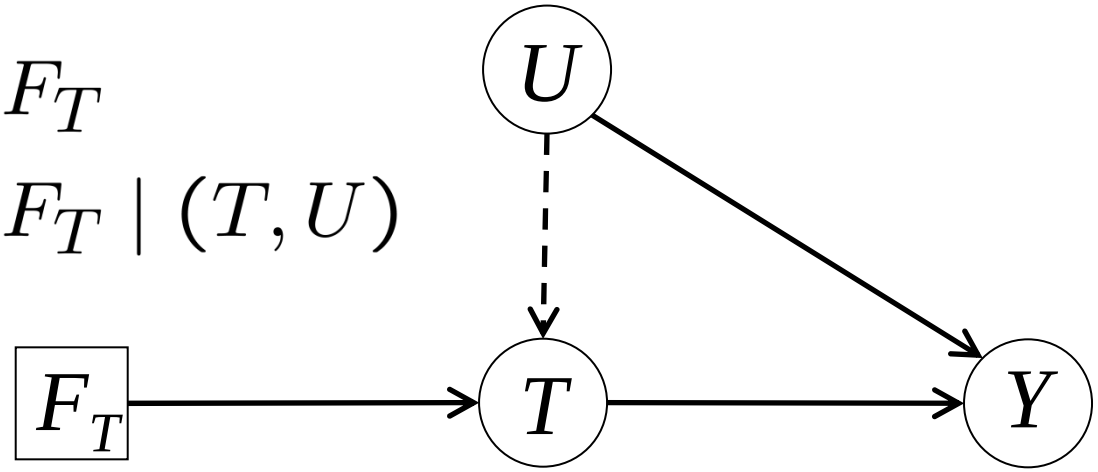
$$T \mid (F_T = \emptyset) \sim P_T \qquad \boldsymbol{Y} \mid (\boldsymbol{T}, F_T)$$

Absence of arrow $F_T \rightarrow Y$ 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"

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

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

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)

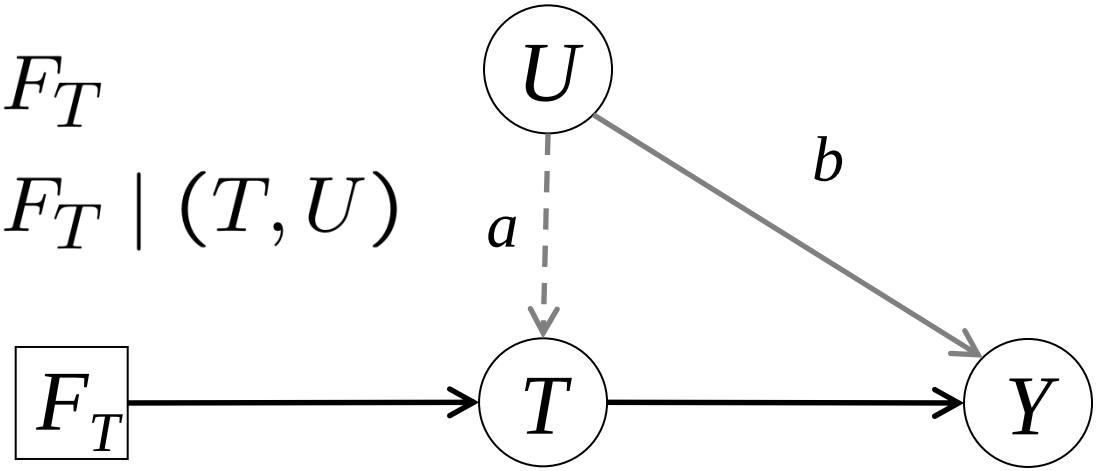
Similarly, whole interventional distribution:

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

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”

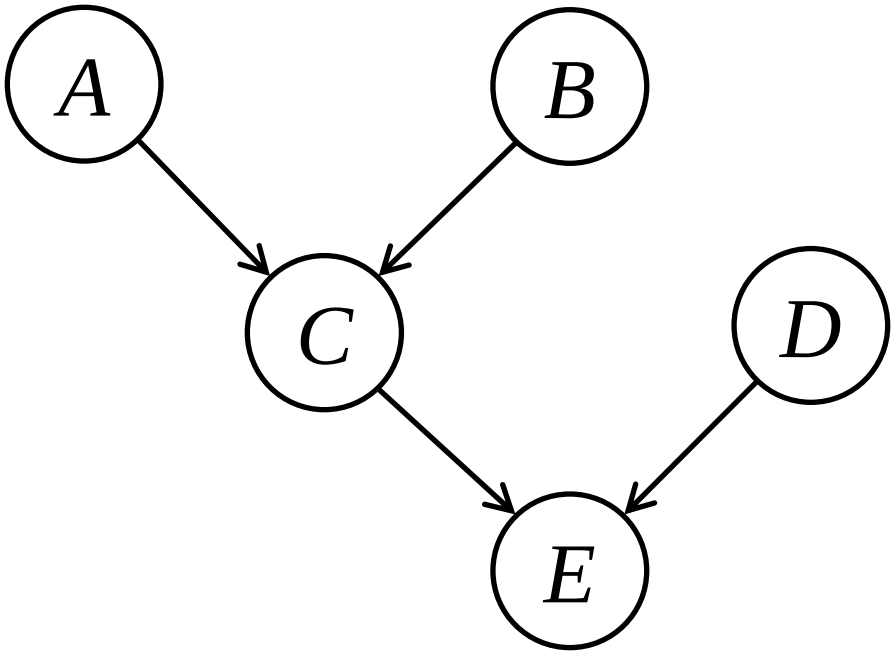
b $Y \perp\!\!\!\perp U \mid T$ “irrelevance”

–then need not even observe U

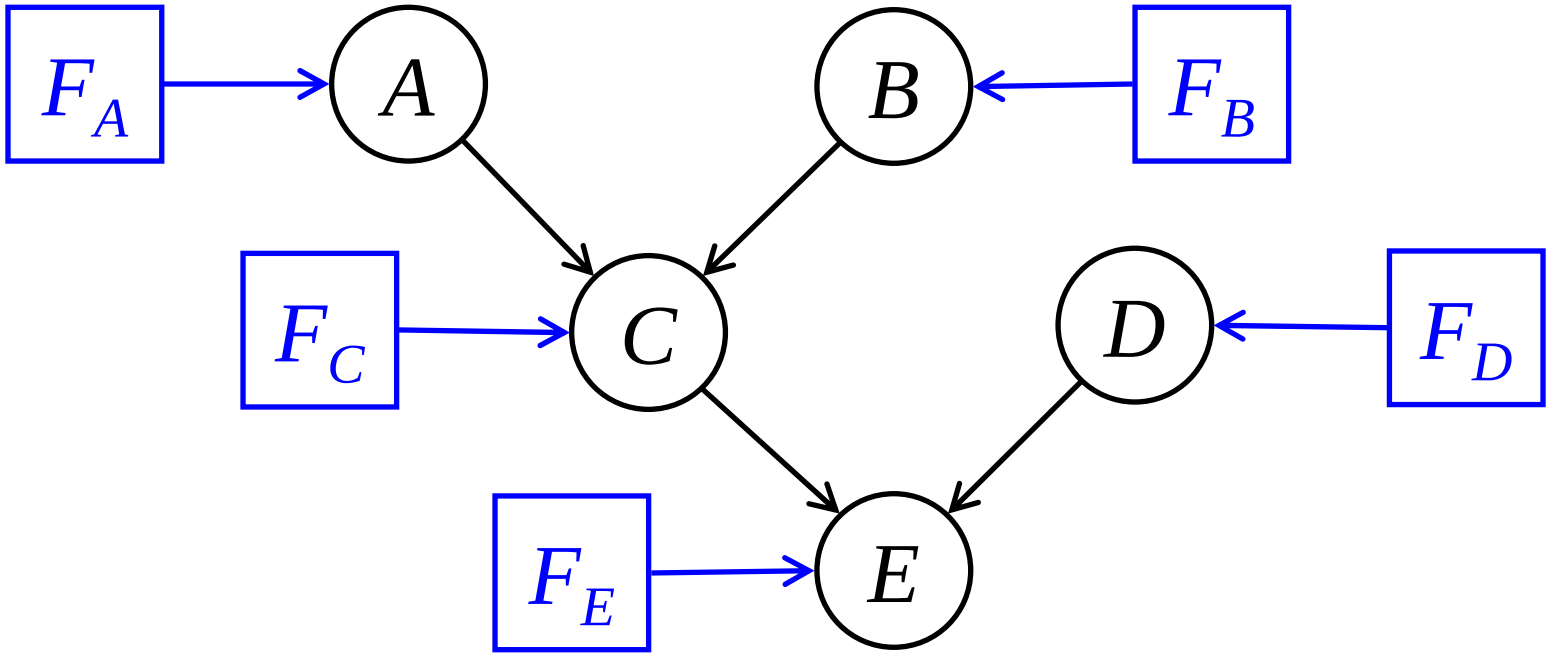
Pearlian DAG

- Envisage intervention on any variable in system
- Augmented DAG model, but with intervention indicators implicit
- Every arrow has a causal interpretation

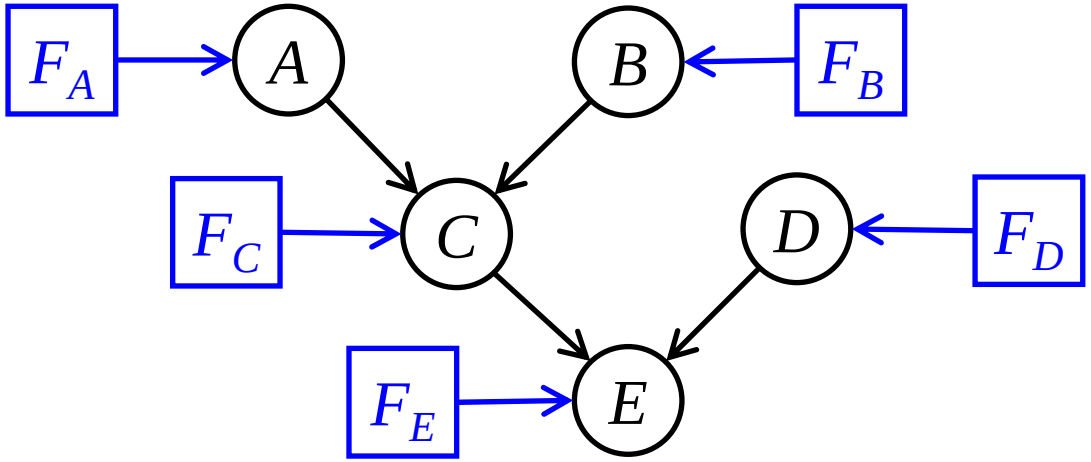
Pearlian DAG



Intervention DAG

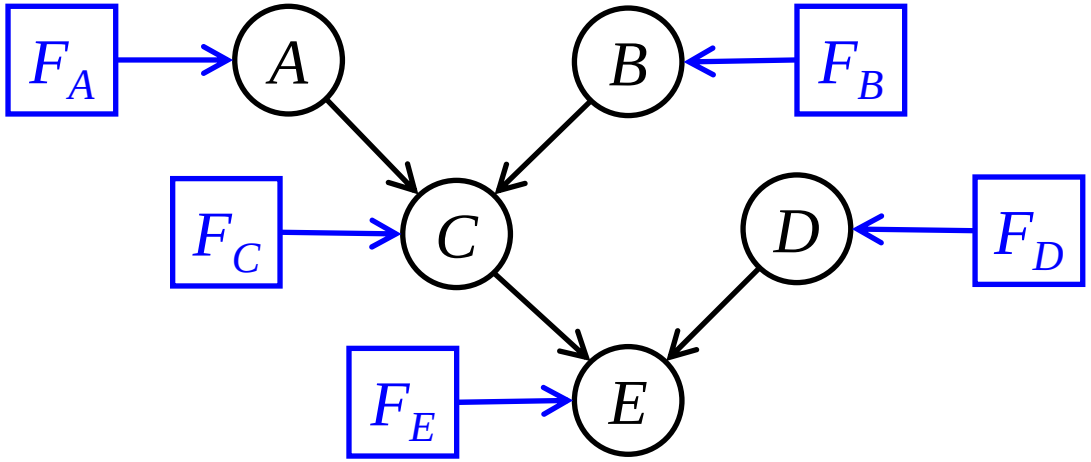


Intervention DAG



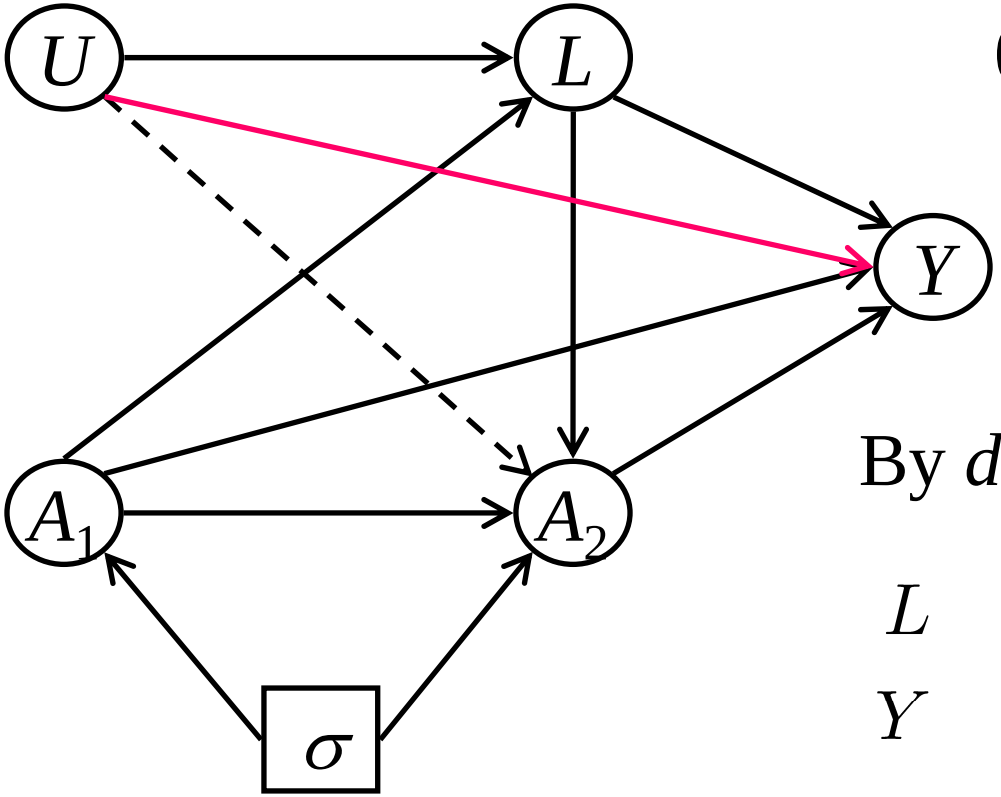
- e.g., $E \perp\!\!\!\perp (A, B, F_A, F_B, F_C, F_D) \mid (C, D, F_E)$

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



(influence diagram)

σ = treatment strategy

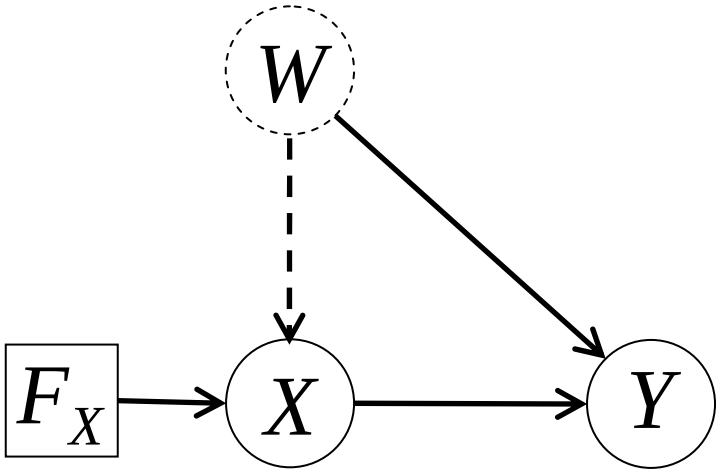
By *d*-separation:

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

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

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

Instrumental Variable



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

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

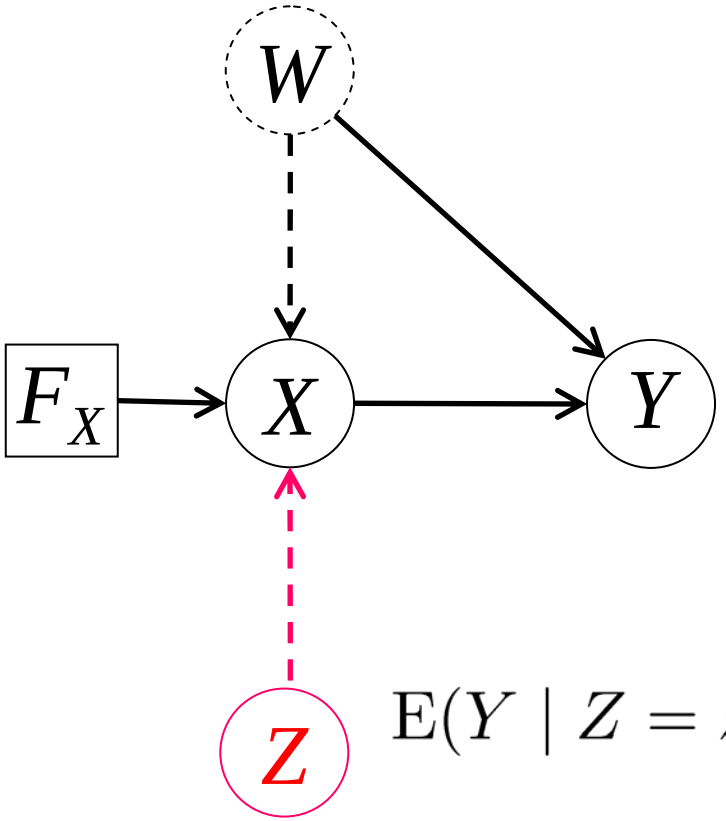
So $E(Y \mid F_X = x) = E(W \mid F_X = x) + \beta x$

$= \alpha + \beta x$

- β is causal regression coefficient

– but not estimable from observational data

Instrumental Variable

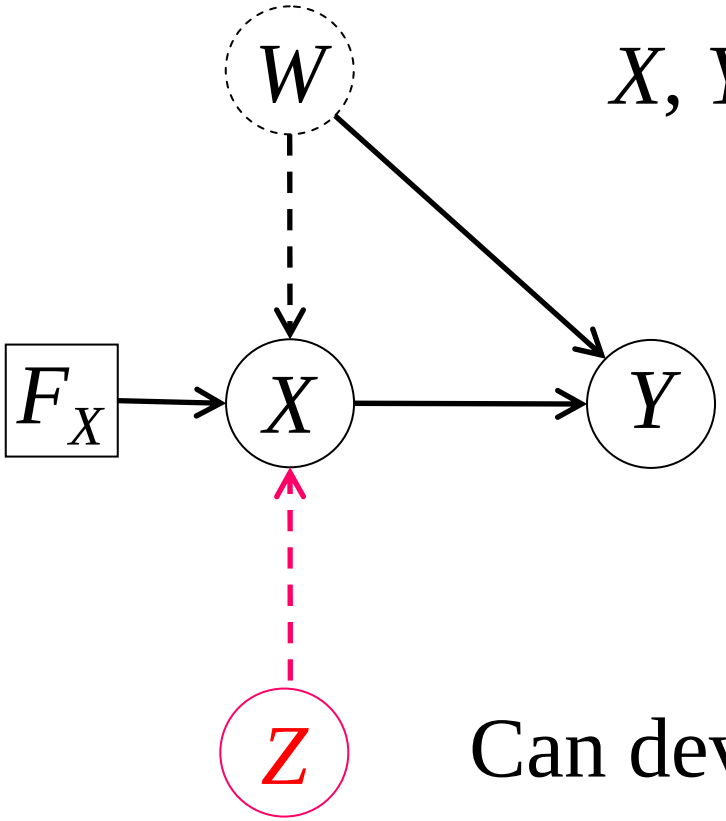


$$
\begin{aligned}
W &\perp\!\!\!\perp F_X \\
Y &\perp\!\!\!\perp F_X \mid (X, W) \\
Z &\perp\!\!\!\perp (W, F_X) \\
Y &\perp\!\!\!\perp Z \mid (X, W; F_X)
\end{aligned}
$$

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

–so can now identify β

Discrete case



X, Y, Z binary

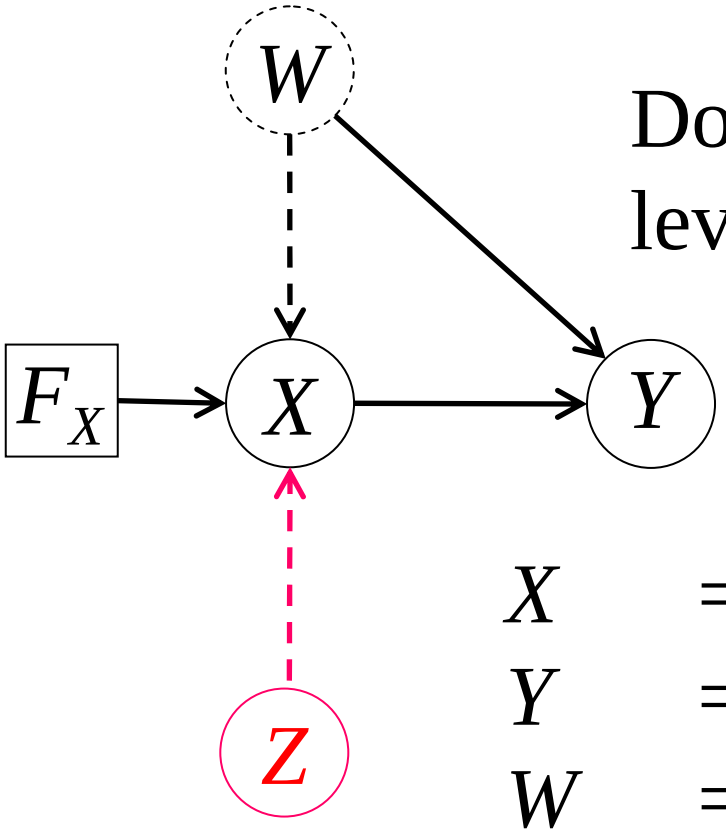
$$W \perp\!\!\!\perp F_X$$
$$Y \perp\!\!\!\perp F_X \mid (X, W)$$
$$Z \perp\!\!\!\perp (W, F_X)$$
$$Y \perp\!\!\!\perp Z \mid (X, W; F_X)$$

Can develop inequalities for ACE

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

in terms of estimable quantities

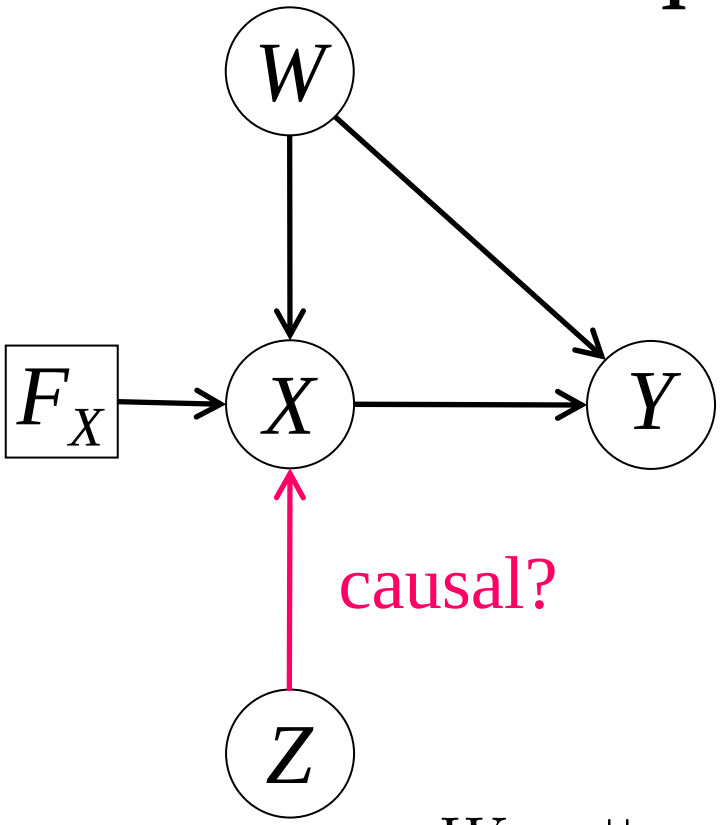
Mendelian Randomisation

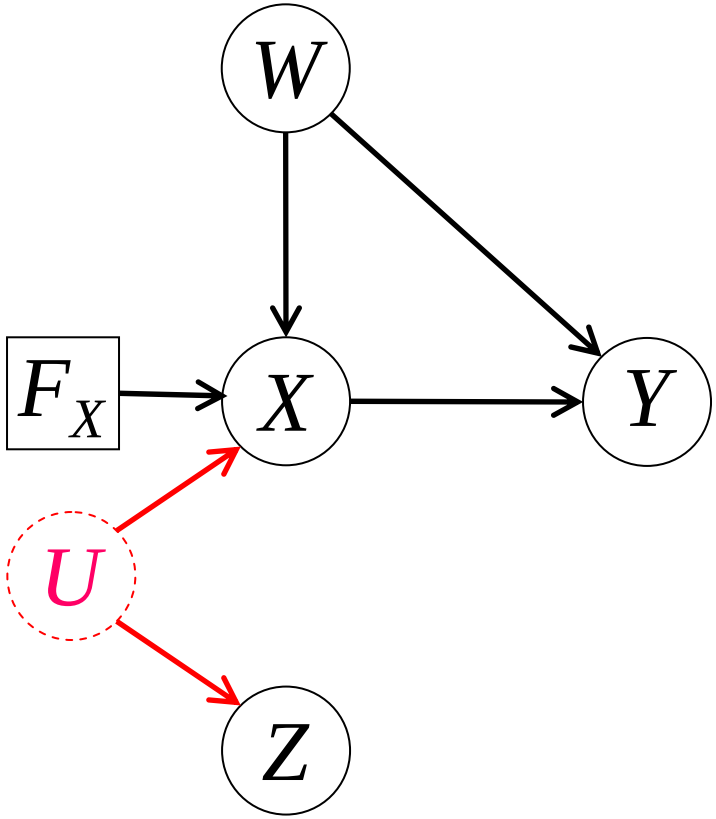


Does low serum cholesterol level increase the risk of cancer?

X = serum cholesterol
Y = cancer
W = diet, smoking,…
Z = APOE gene
(E2 allele induces particularly low serum cholesterol)

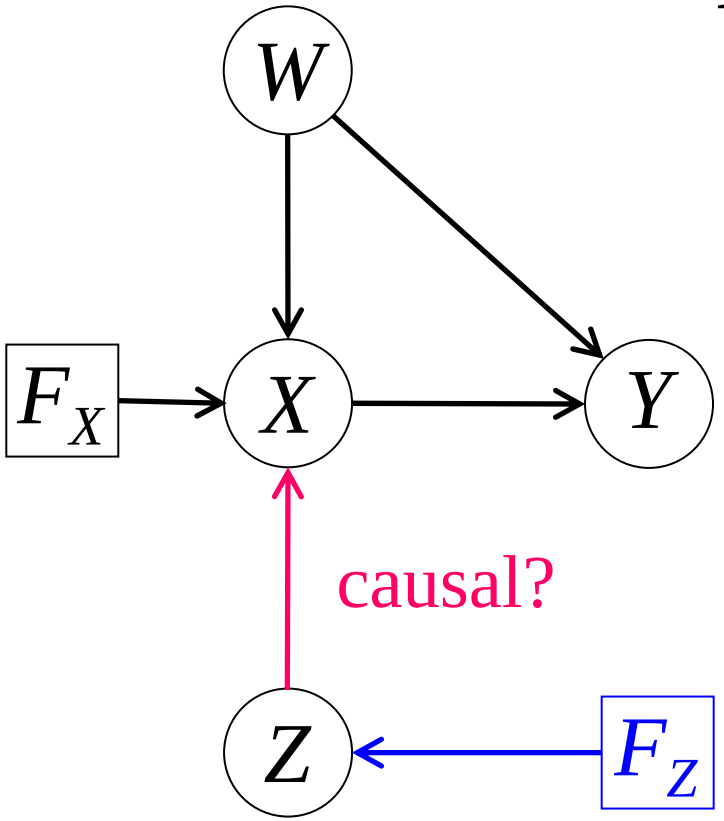
Equivalence

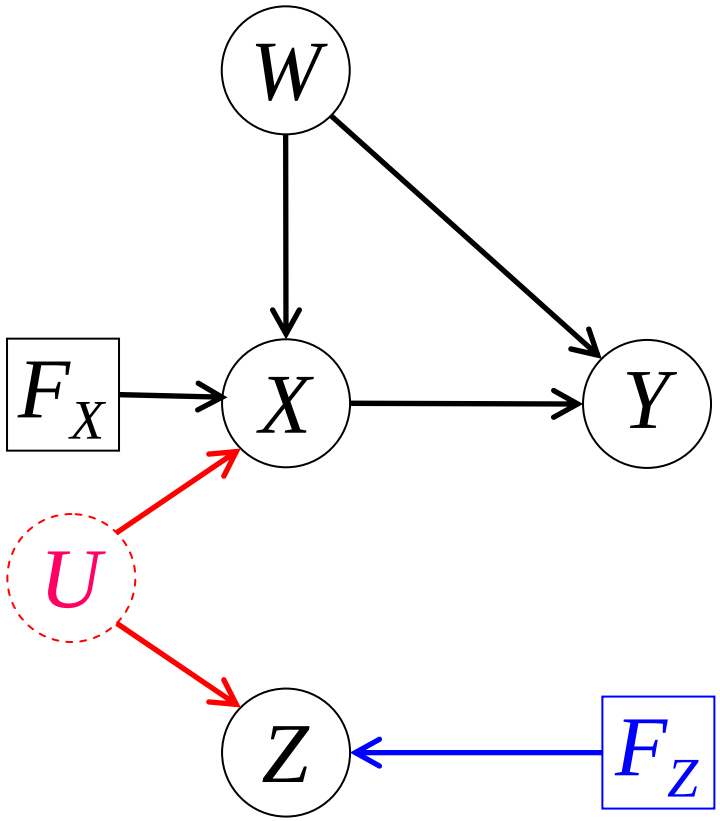




$$
\begin{aligned}
W &\perp\!\!\!\perp F_X \\
Y &\perp\!\!\!\perp F_X \mid (X, W) \\
Z &\perp\!\!\!\perp (W, F_X) \\
Y &\perp\!\!\!\perp Z \mid (X, W; F_X)
\end{aligned}
$$

Non-equivalence





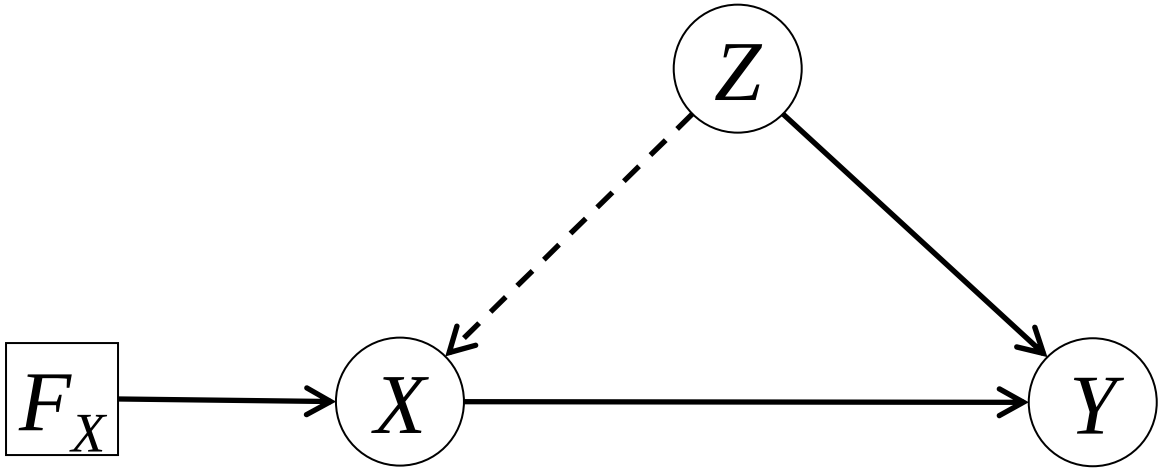
$X \not\perp\!\!\!\perp Z \mid F_Z$

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

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”

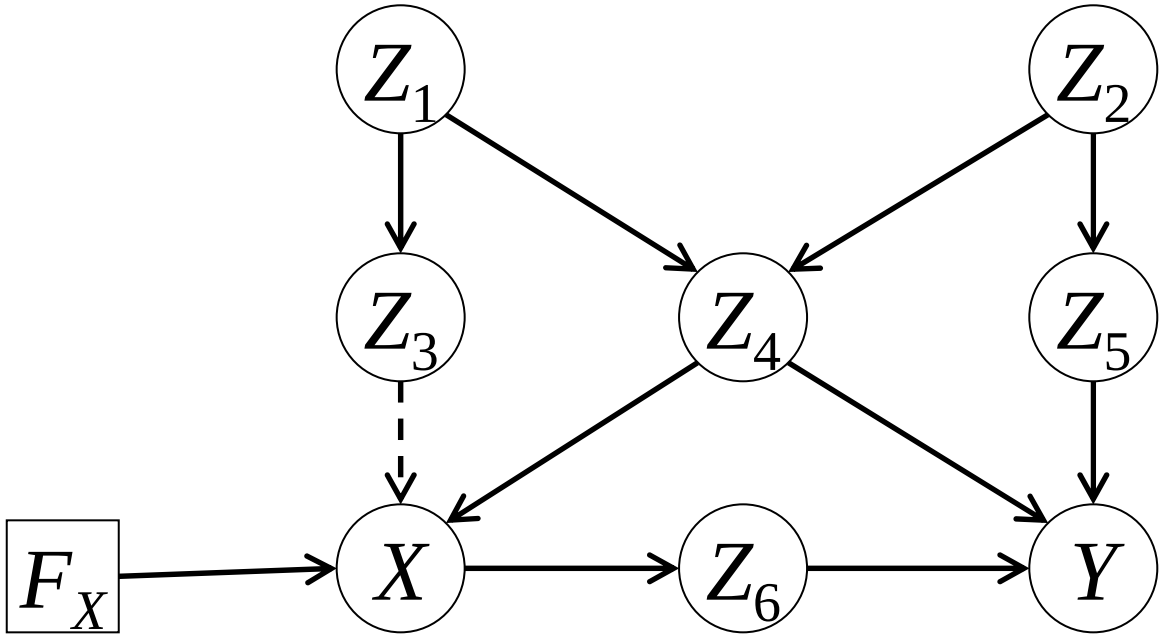


$$Z \quad \perp\!\!\!\perp \quad F_X$$

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

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

Example: “back-door formula”

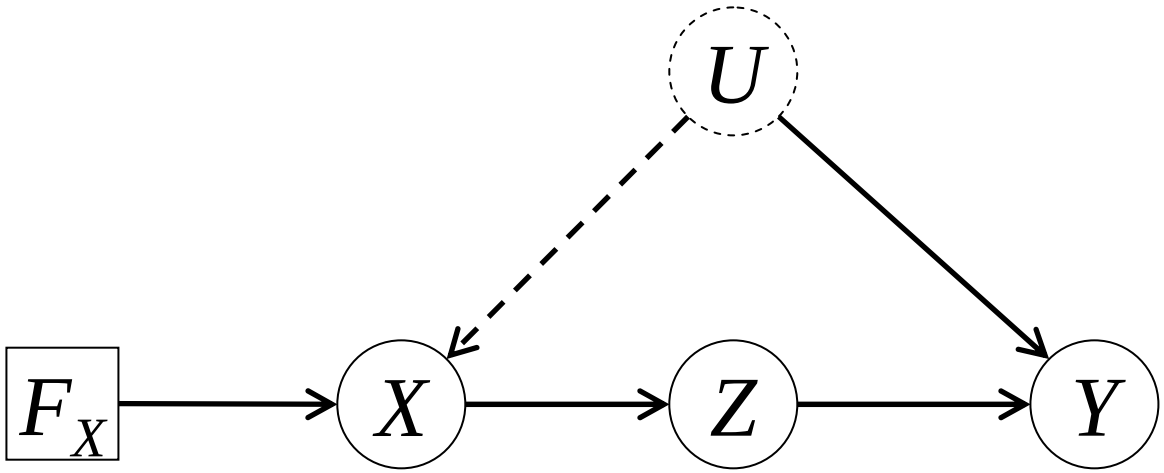


$$Z \perp\!\!\!\perp F_X$$

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

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:

- 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,…, 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,…, $V_{i\text{-}1}$) 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 $V \rightarrow Z \leftarrow W$ 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 $V \rightarrow Z$ –W but not V–W, orient as $V \rightarrow Z \rightarrow W$
   ii. If $V \rightsquigarrow W$ and V–W, orient as $V \rightarrow W$

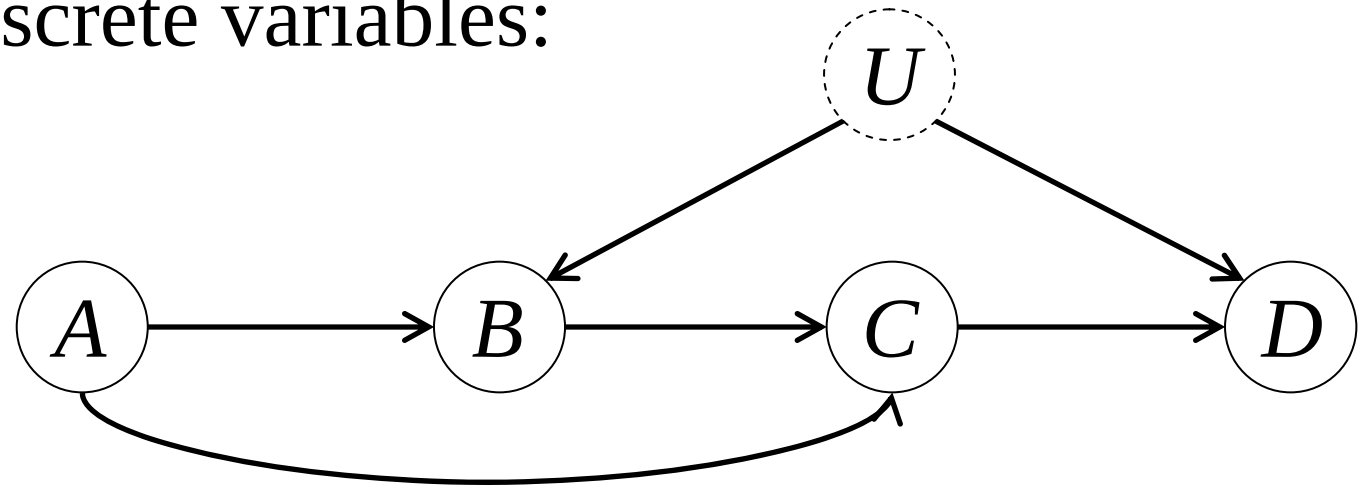
Comments

- 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*.

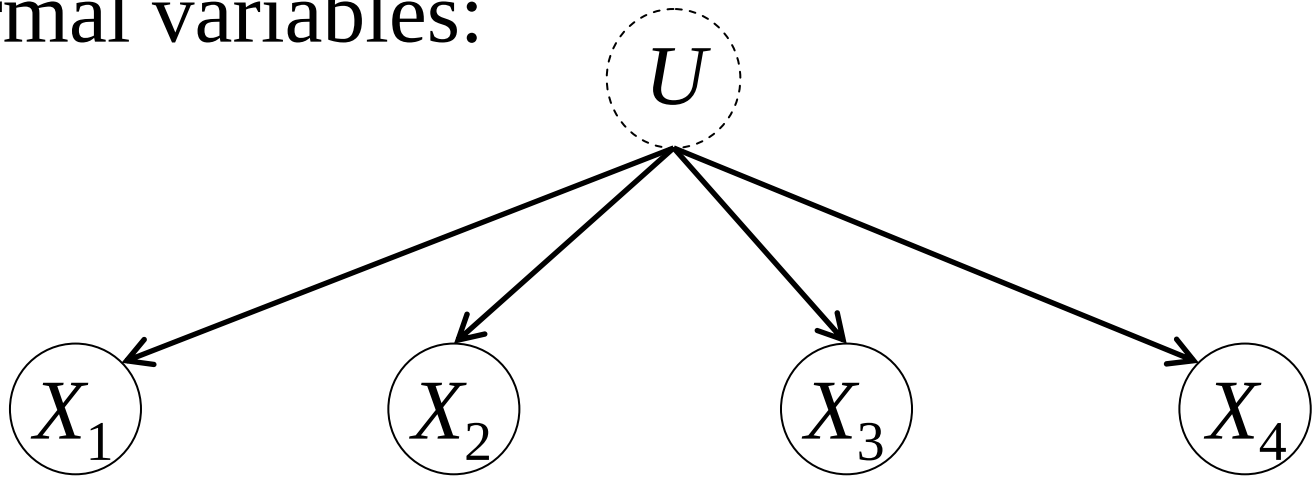
But

$$\sum_b p(b \mid a)p(d \mid a,b,c) = \sum_b p(b \mid a) \sum_u p(d \mid a,b,c,u)p(u \mid a,b,c)$$

$$= \sum_u p(d \mid c,u)p(u \mid a)$$

– 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 $\mathbf{x}$, compute *marginal likelihood* for each model M:

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

- 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.