Web-Mining Agents

Prof. Dr. Ralf Möller
Dr. Özgür Özçep

Universität zu Lübeck
Institut für Informationssysteme

Tanya Braun (Lab Class)

Structural Causal Models

slides prepared by Özgür Özçep

Part IV: Counterfactuals

Literature

- J.Pearl, M. Glymour, N. P. Jewell: Causal inference in statistics – A primer, Wiley, 2016.
  (Main Reference)
- J. Pearl: Causality, CUP, 2000.

Counterfactuals (Example)

Example (Freeway)

- Came to fork and decided for Sepulveda road ($X=0$) instead of freeway ($X=1$)
- Effect: long driving time of 1 hour ($Y = 1h$)

``If I had taken the free way,

then I would have driven less than 1 hour"

Counterfactuals (Informal Definition)

Definition

A counterfactual is an if-then statement where

- the if-condition, aka antecedens, hypothesizes about an alternative non-actual situation/condition

  (in example: taking freeway) and

- the then-condition, aka succedens, describes some consequence of the hypothetical situation

  (in example: 1h drive)

Counterfactuals ≠ truth-conditional if

- Counterfactuals may be false even if antecedent is false
  - ``If Hamburg is capital of Germany,
    then Schulz is cancellor‘‘ true
  - ``If Hamburg were capital of Germany,
    then Schulz would be cancellor‘‘ false
- Usually, the antecedent in counterfactuals in natural language use is false in actual world
- In natural language distinguished by different modes
  - indicative mode for truth-conditional if-statements vs.
  - conjunctive/subjunctive for counterfactuals

- „Hätte, hätte Fahrradkette....“ https://www.youtube.com/watch?v=qt_ppEL7OLI
- L. Matthäus: „Wäre, wäre, Fahrradkette, so ungefähr – oder wie auch immer“

Counterfactuals Require Minimal Change

- Hypothetical world minimally different from actual world
  - If X=1 were the case (instead of X=0),
    but everything else the same (as far as possible),
    then Y < 1h would be the case

Account for consequences of change (from X= 0 to X = 1).

- Idea of minimal change ubiquitous
  - in particular see discussion in belief revision
  - Lecture „Foundations of Ontologies and Databases“

D. Lewis. Counterfactuals. Harvard University Press, Cambridge, MA, 1973.

D. Makinson. Five faces of minimality. Studia Logica, 52:339–379, 1993.

F. Wolter. The algebraic face of minimality. Logic and Logical Philosophy,6:225 – 240, 1998.

Counterfactuals and Rigidity

- Rigidity as a consequence of minimal change of worlds/ states:

  Objects stay the same in compared worlds

- In example: Driver (characteristics) stays the same: if the driver is a moderate driver, then he will be a moderate driver in the hypothesized world, too

- Rigidity of objects across worlds also debated in early work on foundation of modal logic (work of S. Kripke)

Counterfactuals (Example cont'd)

- **Try:** Formalization with intervention
  - E(driving time |do(freeway), driving time = 1 hour)

    doesn‘t work! Why?
  - There is a clash for RV „driving time“ (Y)
    - $Y = 1$ h in actual world vs.
    - $Y < 1h$ (expected) under hypothesized condition $X = 1$
- **Solution**: Distinguish Y (driving time) under different worlds/conditions $X = 0$ vs. $X = 1$

$$E(Y_{X=1} \mid X = 0, Y_0 = Y = 1)$$

$Y_{X=x}$ formalizes counterfactual

Expected driving time $Y_{X=1}$ if one had chosen freeway $(X=1)$ knowing that other decision $(X=0)$ lead to driving time Y_0 of 1 hour.

Counterfactuals (Definition)

Definition

A counterfactual RV is of the form $Y_{X=x}$ and its semantics is given by

$$Y_{X=x}(u) := Y_{M_x}(u)$$

Note the rigidity assumption: Definition talks about the same ``objects“ u in different worlds

where

- Y, X are (sets of) RVs from an SEM M
- x is an instantiation of X
- M_x is the SEM resulting from M by substituting the equation(s) for (all RVs in) X with value(s) x
- u is an instantiation of all exogenous variables in M

Counterfactuals (consistency rule)

- Consequence of the formal definition of counterfactuals

Consistency rule
If $X = x$, then $Y_{X=x} = Y$

- This case (hypothesized = actual) non-typical in natural language use (Merkel: „If I only would be cancellor..)
- In belief revision the corresponding rule is termed „vacuity“: because there is no reason to change, the change is vacuous.

Counterfactuals (for linear SEMs)

- How to formalize semantics of counterfactuals?
  - Use ideas similar to those of intervention
- Consider linear models
  - Values of all variables determined by values of exogenous variables $U = U_1, \dots ,U_n$
  - So can write $X = X(U)$ for any variable in SEM
  - Example
    - X: Salary, $u = u_1, \dots, u_n$ characterizes individual Joe
    - $X(u) =$ Joe's salary
  - When considering different worlds, the individuals (such as Joe $= (u_1, \dots,u_n)$) stay the same.

Counterfactuals in linear SEMS (Example)

- Linear model M:

$$X = aU \quad ; \quad Y = bX + U$$

- Find $Y_{X=x}(u) = ?$

(value of Y if it were the case that $X = x$ for individual u)

- Algorithm
  1. Identify u under evidence (here: just given)
  2. Consider modified model M_x
     - $X = x$
     - $Y = bX + U$
  3. Calculate $Y_{X=x}(u)$

$$Y_{X=x}(u) = bx + u$$

Counterfactuals in linear SEMs (Example)

- Linear model M:

$$X = aU \quad ; \quad Y = bX + U$$

with $a = b = 1$.

$X_y(U) = ?$

Algorithm

1. $U = u$; 2. $Y = y$; 3. $X = aU = au = u$.

(X unaltered by hypothetical condition $Y = y$)

U	**X(u)**	**Y(u)**	$Y_{X=1}(u)$	$Y_{X=2}(u)$	$Y_{X=3}(u)$	$X_{Y=1}(u)$	$X_{Y=2}(u)$	$X_{y=3}(u)$
1	1	2	2	3	4	1	1	1
2	2	4	3	4	5	2	2	2
3	3	6	4	5	6	3	3	3

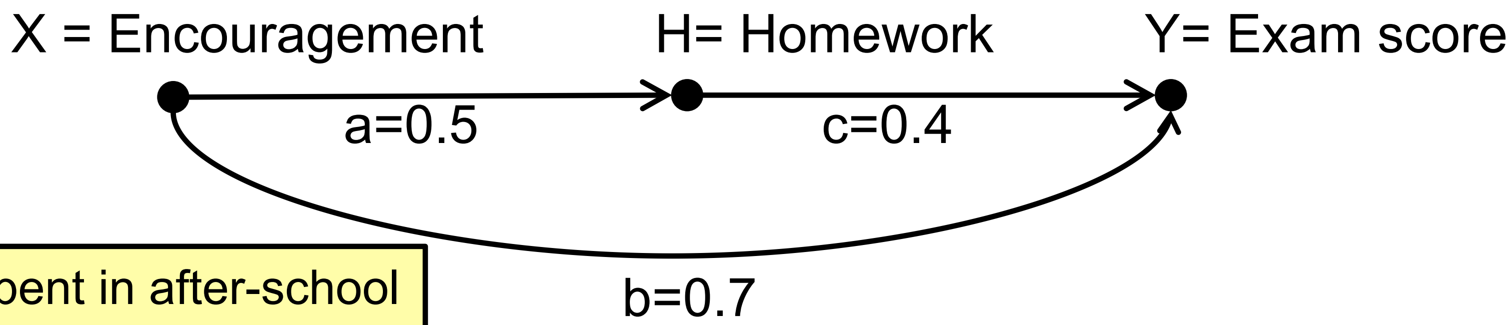
Counterfactuals vs. Intervention with do()

Counterfactual $Y_x(u)$	**Intervention do(X=x)**
Defined locally for each u	Defined globally for whole population/distribution
Can output individual value	Outputs only expectation/ distribution
Allows cross-world speak	Allows single-world speak
Can simulate intervention	Cannot simulate counterfactual

Counterfactuals in linear SEMs (example)

- Linear model M:
  - $X = U_X$
  - $H = aX + U_H$
  - $Y = bX + cH + U_Y$
  - $\sigma_{U_iU_j} = 0$ for all $i,j \in \{X,H,Y\}$ (i.e., U_i, U_j are not linearly correlated/dependent)

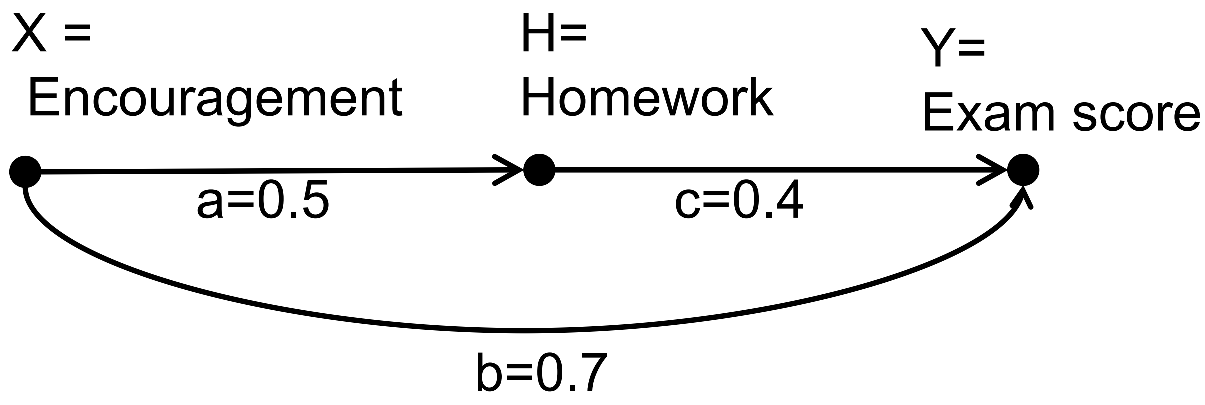
$a = 0.5; \quad b = 0.7; \quad c = 0.4$



X = time spent in after-school remedial program

Counterfactuals in Linear SEMs (Example)

- Linear model M:
  - $X = U_X$
  - $H = aX + U_H$
  - $Y = bX + cH + U_Y$



- Consider an individual Joe given by evidence:

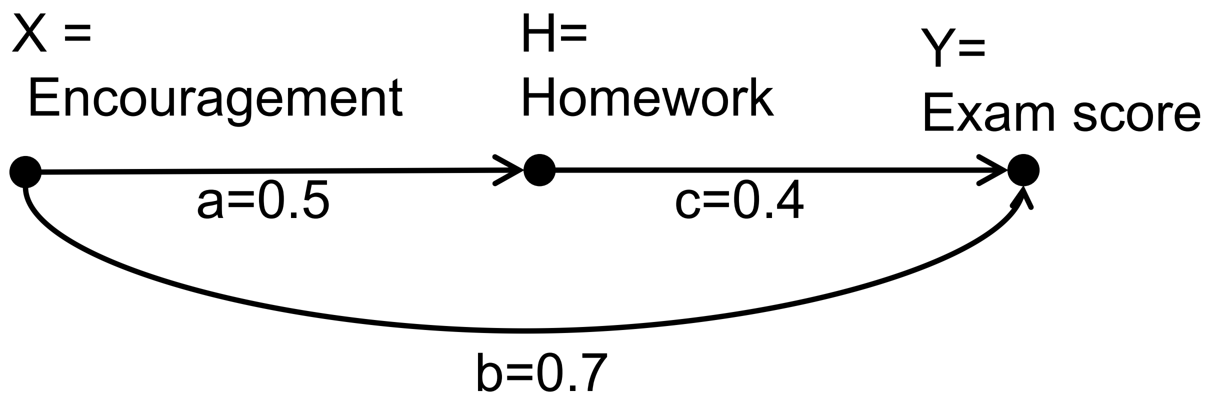
  $X = 0.5, \quad H = 1, \quad Y = 1.5$

- Want to answer counterfactual query:

  „What would Joe‘s exam score be, if he had doubled study time at home?“

Counterfactuals in Linear SEMs (Example)

- Linear model M:
  - $X = U_X$
  - $H = aX + U_H$
  - $Y = bX + cH + U_Y$



- Consider an individual Joe given by evidence:

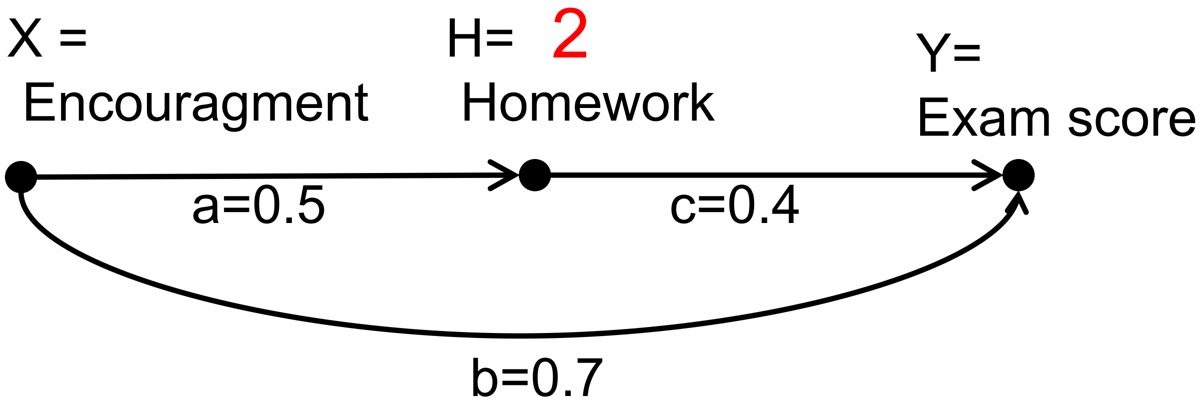
  $X = 0.5, \quad H = 1, \quad Y = 1.5$

- **Step 1**: Determine U-characteristics from evidence
  - $U_X = 0.5$
  - $U_H = 1\text{-}0.5 * 0.5$
  - $U_Y = 1.5 \text{-}0.7 * 0.5 – 04.4 * 1 = 0.75$

The U-characteristics are rigid

Counterfactuals in Linear SEMs (Example)

- Linear model M:
  - $X = U_X$
  - $H = aX + U_H$
  - $Y = bX + cH + U_Y$



- **Step 2**: Simulate hypothetical change (doubling)
  - Set $H = 2$
- **Step 3:** Calculate counterfactual $Y_{H=2}(u)$
  - $Y_{H=2}(U_X = 0.5, U_h = 0.75, U_Y = 0.75)$

    $= 0.7 * 0.5 + 0.4 * 2 + 0.75 = 1.90$

Joe would benefit from doubling homework

($Y = 1.5$ in actual world, $Y = 1.90$ in hypothetical world when doubling H

Deterministic Counterfactuals Algorithm

Algorithm

- Step 1 (Abduction): Use evidence $E = e$ to determine u
- Step 2 (Action): Modify model M to obtain model M_x
- Step 3 (Prediction): Compute counterfactual $Y_{X=x}(u)$ with M_x

- This algorithm considers single individual
- And answers query determined by counterfactual value

- What about classes of individuals and probabilistic counterfactuals?

Nondeterministic Counterfactuals Algorithm

Algorithm

– Step 1 (Abduction): Calculate $P(U|E = e)$

– Step 2 (Action): Modify model M to obtain model M_x

– Step 3 (Prediction): Compute expectation $E(Y_{X=x}|E=e)$ using M_x and $P(U|E=e)$

- Calculate the probabilities of obtaining some individual (step 1)
- Step 2 the same
- Calculate conditional expectation: What is the expected value of Y if one were to change X to x knowing $E = e$

Nondeterministic Counterfactuals (Example)

- Model M: $X = aU$; $Y = bX + U$ (with $a = b = 1$)

  $U = \{1,2,3\}$ represents three types of individuals with prob.

  $P(U = 1) = 1/2$; $P(U = 2) = 1/3$; $P(U=3) = 1/6$

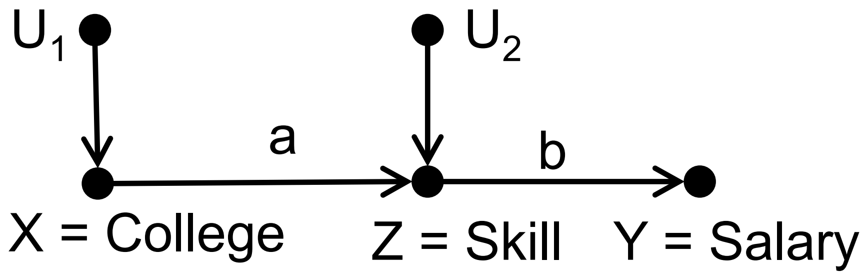
- Examples:
  - $P(Y_{X=2}(u) = 3) = ? = P(U = 1) = 1/2$
  - $P(Y_2 > 3, Y_1 < 4) = P(U=2) = 1/3$
  - $P(Y_1 < Y_2) = 1$

U	**X(u)**	**Y(u)**	**$Y_{X=1}(u)$**	**$Y_{X=2}(u)$**	**$Y_{X=3}(u)$**	**$X_{Y=1}(u)$**	**$X_{Y=2}(u)$**	**$X_{y=3}(u)$**
1	1	2	2	3	4	1	1	1
2	2	4	3	4	5	2	2	2
3	3	6	4	5	6	3	3	3

Counterfactuals More Expressive (Example)

- Counterfactuals more expressive than intervention
- Linear model

  $X = U_1;\ \ Z = aX + U_2; Y = bZ$



  – $E[Y_{X=1} \mid Z = 1] = ?$

  – Not captured by $E[Y|do(X=1), Z=1]$. Why?

  - Gives only the salary Y of all individuals that went to college and since then acquired skill level $Z = 1$.
  - $E[Y|do(X=1), Z=1] = E[Y|do(X=0), Z=1]$

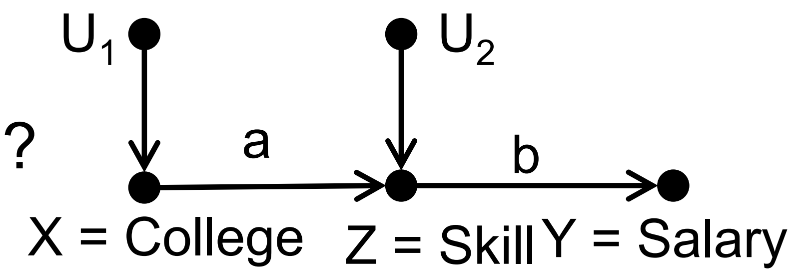
    Talks about postinvention for two different groups

  - In contrast: $E[Y_{X=1} \mid Z = 1]$ captures salary of individuals who in the actual world have skill level $Z = 1$ but might get $Z > 1$
  - $E[Y_{X=0} \mid Z = 1] \neq E[Y_{X=1} \mid Z = 1]$

    Talks about one group acting under different antecedents

Counterfactuals More Expressive (Example)

- $E[Y_{X=0} \mid Z = 1] \neq E[Y_{X=1} \mid Z = 1]$?
  - How is this reflected in numbers?
  - Later: How reflected in graph?



$X = U_1$; $Z = aX + U_2$; $Y = bZ$ (for $a \neq 1$ and $a \neq 0$, $b \neq 0$)

u_1	u_2	$X(u)$	$Z(u)$	$Y(u)$	$Y_{X=0}(u)$	$Y_{X=1}(u)$	$Z_{X=0}(u)$	$Z_{X=1}(u)$
0	0	0	0	0	0	ab	0	a
0	1	0	1	b	b	(a+1)b	1	a+1
1	0	1	a	ab	0	ab	0	a
1	1	1	a+1	(a+1)b	b	(a+1)b	1	a+1

- $E[Y_1|Z=1] = (a+1)b$; $E[Y|do(X=1),Z=1] = b$
- $E[Y_0|Z=1] = b$; $E[Y|do(X=0),Z=1] = b$

In particular: $E[Y_1 - Y_0|Z=1] = ab \neq 0$

Counterfactuals vs. Intervention with do()

Counterfactual $Y_x(u)$	**Intervention do(X=x)**
Defined locally for each u	Defined globally for whole population/distribution
Can output individual value	Outputs only expectation/ distribution
Allows cross-world speak	Allows single-world speak
Can simulate intervention	Cannot simulate counterfactual

$$E[Y|do(X=1), Z=1] = ? = E[Y_{X=1}| Z_{X=1} = 1]$$

Counterfactuals vs. Intervention with do()

Counterfactual $Y_x(u)$	**Intervention do(X=x)**
Defined locally for each u	Defined globally for whole population/distribution
Can output individual value	Outputs only expectation/ distribution
Allows cross-world speak	Allows single-world speak
Can simulate intervention	Cannot simulate counterfactual

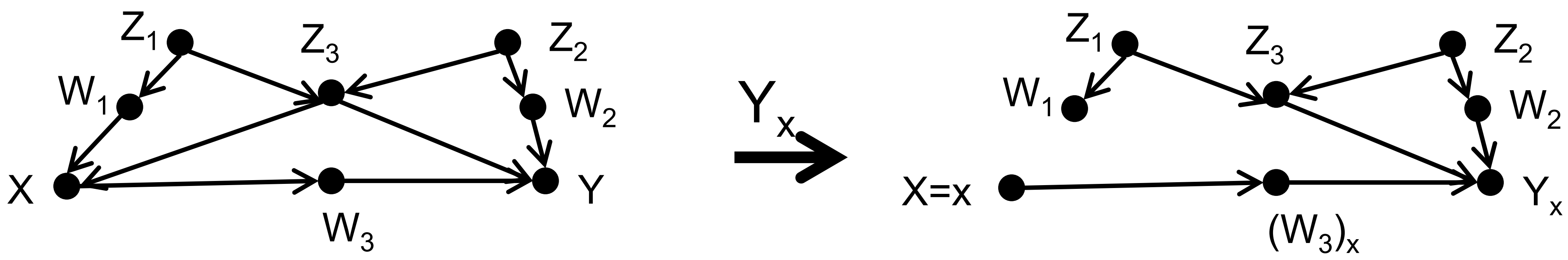
- See road example
- But in non-conditional case we have $E[Y_x=y] = E[Y=y|do(X=x)]$

Graphical representation of counterfactuals

- Rember definition of counterfactual

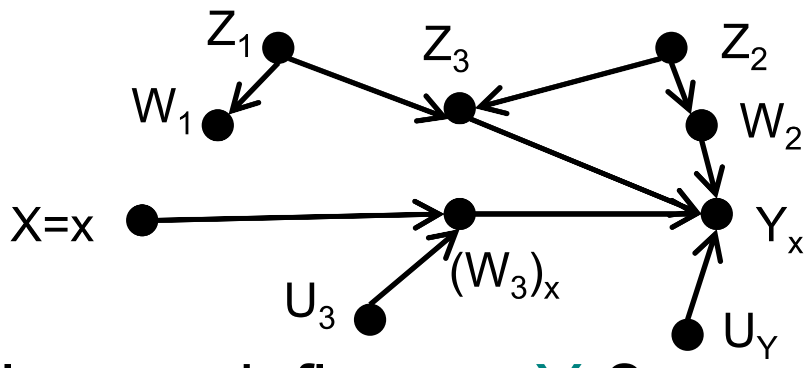
$$Y_{X=x}(u) := Y_{Mx}(u)$$

- Modification as in intervention but with variable change



- Can answer (independence) queries regarding counterfactuals as for any other variable
- Note: Graphs do not show error variables

Independence criterion for counterfactuals



- Which variables can influence Y_x?
  - Parents of Y and parents of nodes on pathway between X and Y (here: $\{Z_3, W_2, U_3, U_y\}$)
- So blocking these with a set of RVs Z renders Y_x independent of X given Z

Theorem (Counterfactual interpretation of backdoor)
If set of RVs Z satisfies backdoor for (X,Y),
then $P(Y_x \mid X,Z) = P(Y_x \mid Z)$ (for all x)

Independence criterion for counterfactuals

Theorem (Counterfactual interpretation of backdoor)
If set of RVs Z satisfies backdoor for (X,Y),
then $P(Y_x \mid X,Z) = P(Y_x \mid Z)$ (for all x)

- Theorem useful for estimating prob. for counterfactuals
- In particular can use adjustment formula

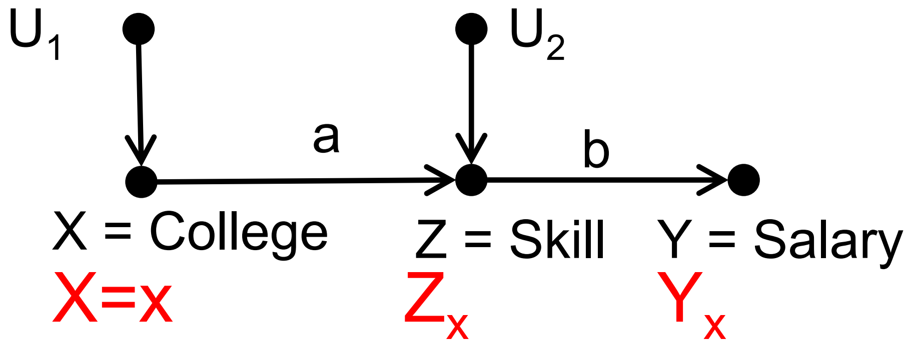
$$
\begin{aligned}
P(Y_x = y) &= \sum_z P(Y_x = y \mid Z = z)P(z) && \text{(summing out)}\\
&= \sum_z P(Y_x = y \mid Z = z, X=x)P(z) && \text{(Thm)}\\
&= \sum_z P(Y=y \mid Z = z, X = x)\, P(z) && \text{(consistency)}
\end{aligned}
$$

- Clear in light of $P(Y_x = y) = P(Y=y \mid do(X=x))$

Independence counterfactuals (example)

- Reconsider linear model

$X = U_1;\ \ Z = aX + U_2;\ Y = bZ$



- Does college education have effect on salary, considering a group of fixed skill level?
- Formally: Is Y_x independent of X, given Z?
  - Is Y_x d-separated from X given Z?
  - No: Z a collider between X and U_2 (as well as X and Y_x)
  - Hence: $E[Y_x \mid X, Z] \neq E[Y_x \mid Z]$
  (hence education has effect for students of given skill)

Counterfactuals in Linear Models

- In linear models any counterfactual identifiable if linear parameters identified.
  - In this case all functions in SEM fully determined
  - Can use $Y_x(u) = Y_{M_x}(u)$ for calculation
- What if some parameters not identified?
  - At least can identify statistical features of form $E[Y_{X=x}|Z=z]$

Theorem (Counterfactual expectation)
Let τ denote slope of total effect of X on Y

$$\tau = E[Y|do(x+1)]-E[Y|do(x)]$$

Then, for any evidence $Z = e$

$$E[Y_{X=x}|Z=e] = E[Y|Z=e] + \tau\,(x-E[X|Z=e])$$

Counterfactuals in Linear Models

Theorem (Counterfactual expectation)
Let τ denote slope of total effect of X on Y

$$\tau = E[Y|do(x+1)]-E[Y|do(x)]$$

Then, for any evidence $Z = e$

$$E[Y_{X=x}|Z=e] = E[Y|Z=e] + \tau\,(x-E[X|Z=e])$$

Current estimate of Y

Expected effect change when x shifted from current best estimate $E[X|Z=e]$

Effect of Treatment on the Treated (ETT)

Theorem (Counterfactual expectation)
Let τ denote slope of total effect of X on Y

$$\tau = E[Y|do(x+1)]-E[Y|do(x)]$$

Then, for any evidence $Z = e$

$$E[Y_{X=x}|Z=e] = E[Y|Z=e] + \tau\,(x-E[X|Z=e])$$

$$\begin{aligned}
ETT &= E[Y_1 - Y_0|X=1] \\
&= E[Y_1|X=1]- E[Y_0|X=1] \\
&= E[Y|X=1]- E[Y|X=1] + \tau\,(1-E[X|X=1]) - \tau\,(0-E[X|X=1])
\end{aligned}$$

(using Thm with $(Z = e) \triangleq (X = 1)$)

$$= \tau$$

Hence, in linear models, effect of treatment on the treated (individual) is the same as total treatment effect on population

Extended Example for ETT

- Job training program (X) for jobless funded by government to increase hiring Y
- Pilot randomized experiment shows:

  Hiring-%(w/ training) > Hiring-%(w/o training) (*)
- Critics
  - (*) not relevant as it might falsely measure effect on those who chose to enroll for program by themselves (these may got job because they are more ambitious)
  - Instead, need to consider ETT

    $E[Y_1 - Y_0 | X=1]$ = causal effect of training X on hiring Y for those who took the training

Extended Example for ETT (cont'd)

- Difficult part: $E[Y_{X=0} \mid X=1]$
  - not given by observational or experimental data
  - but can be reduced to these if appropriate covariates Z (fulfilling backdoor criterion) exist

$P(Y_x = y \mid X = x')$

$= \sum_z P(Y_x = y \mid Z = z, x')P(z|x')$ (by condition on z)

$= \sum_z P(Y_x = y \mid Z = z, x)P(z|x')$ (by Thm on counterfactual backdoor $P(Y_x \mid X,Z) = P(Y_x \mid Z)$)

$= \sum_z P(Y = y \mid Z = z, x)P(z|x')$ (consistency rule)

Contains only observational/testable RVs

- $E[Y_0|X=1] = \sum_z E(Y \mid Z = z, X=0)P(z|X=1)$

(after substitution and commuting sums)

Extended Example Additive Intervention

- Scenario
  - Add amount q of insulin to group of patients (with different insulin levels)
    - $do(X = X+q) = add_X(q)$
    - Different from simple intervention
  - Calculate effect of additive intervention from data where such additions have not been oberved

- Formalization with counterfactual
  - Y = outcome RV = a RV relevant for measuring effect
  - $X = x'$ (previous level of insulin)
  - $Y_{x'+q}$ = outcome after additive intervention with q insul.

Extended Example Additive Intervention

- $E(Y_{x' +q}|x')$ = expected output of additive intervention
  - Part of ETT expression
  - Can be identfied with adjustment formula

    (for backdoor Z such as weight, age, etc.)

- $E[Y|add_X(q)] - E[Y]$

  $= \sum_{x'} E[Y_{x'+q}|X=x']P(X=x') - E[Y]$

  $= \sum_{x'}\sum_z E[Y|X=x'+q,Z=z]P(Z=z|X=x')P(X=x')-E[Y]$

  (using already derived formula

  $E(Y_x \mid X = x') = \sum_z E(Y = y \mid Z = z, x)P(z|x')$

  and substituting $x = x' + q$)

Extended Ex. Additive Intervention (cont'd)

A: = $E[Y|add_X(q)] - E[Y]$ =?=

B: = $\sum_x(E[Y|do(X = x+q)] - E[Y|do(X = x)]P(X=x))$

= $\sum_x(E[Y_{X = x+q}] - E[Y_{X = x}])P(X=x)$

= Average total effect of adding q for each level x

- NO!
  - In A ``nature" choose individuals level of X
  - In A, P(X=x) represents those individuals chosing level X=x by free choice it
  - It could be the case that those highly sensitive to getting dose q addition try to lower X value
  - In B one cuts this natural influence

Extended Example Decision Making (cont'd)

- Scenario 1
  - Cancer patient Ms Jones has to decide between
    1. Lumpectomy alone ($X = 0$)
    2. Lumpectomy with irradiation ($X = 1$)

    hoping for remission of cancer ($Y = 1$)
  - She decides for adding irradiation ($X=1$) and 10 years later the cancer remisses.
  - Is the remission due to her decision?
- Formally: Determine probability of necessity

$$PN = P(Y_{X=0} = 0 \mid X = 1, Y=1)$$

- If you want remission, you have to go for adding irradiation (irradiation necessary for remission)

Extended Example Decision Making (cont'd)

- Scenario 2
  - Cancer patient Mrs Smith had lumpectomy alone ($X=0$) and her tumor reoccurred ($Y=0$).
  - She regrets not having gone for irradiation.

    Is she justified?
- Formally: Determine probability of sufficiency

$$PS = P(Y_{X=1} = 1 \mid X = 0, Y=0)$$

- If you go for adding irradiation, you will achieve cancer remission

Note that, formally, PN and PS are the same.
The distinction comes from interpreting
value 1 = acting
value 0 = omitting an action

Extended Example Decision Making (cont'd)

- Scenario 3
  - Cancer patient Mrs Daily faces same decision as Mrs Jones and argues
    - If my tumor is of type that disappears without irradiation, why should I take irradiation?
    - If my tumor is of type that does not disappear even with irradiation, why even take irradiation?
  - So should she go for irradiation?
- Formally: Determine probability of necessity and sufficiency

$$PNS = P(Y_{X=1} = 1, Y_{X=0} = 0)$$

Extended Example Decision Making (cont'd)

- Formally: Determine probability of necessity and sufficiency

$$PNS = P(Y_{X=1} = 1, Y_{X=0} = 0)$$

- PN (PS and PNS) can be estimated from data under assumption of monotonicity (adding irradiation cannot cause recurrence of tumor)

$$PNS = P(Y=1|do(X=1)) – P(Y=1|do(X=0))$$

= total effect of changing X from no irradiation to irradiation on Y

Extended Example Mediation

- Scenario (Indirect effect of gender on hiring)

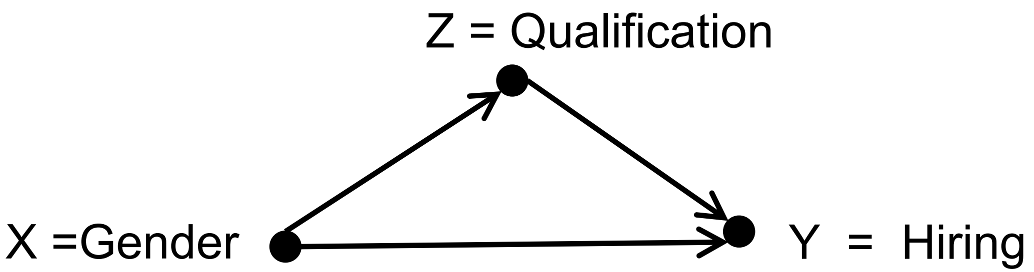
  Policy maker wants to decide whether to

  1. Make hiring procedure gender-blind (direct effect) or
  2. Eliminate gender inequality in education or job trainig (indirect effect)

  – (Controlled) direct effect identifiable with do expression (lecture on interventions)

  – Indirect effect for non-linear system ≠

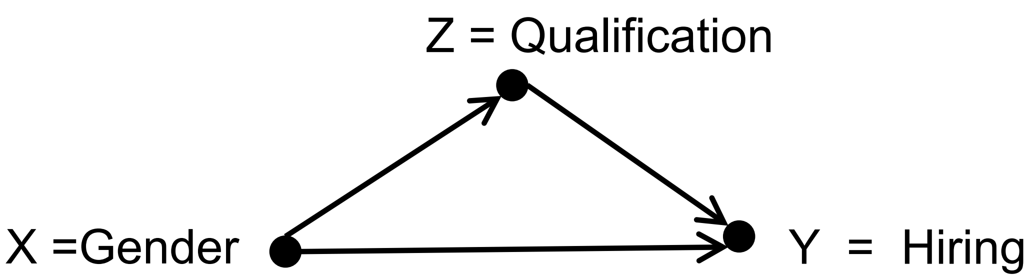
  total effect minus direct effect



Extended Example Mediation (cont'd)

- In order to determine indirect effect of gender:
  - Have to substract outcomes Y in two worlds where
    - gender X is kept fixed to male (X=1)
    - but its mediator (Z) is changed accordingly if one had changed the gender (from male to female)
  - Consider: $E[\, Y_{X=1, Z=Z_{X=0}} \; - \; Y_{X=1, Z=Z_{X=1}} \,]$

> - $Y_{x=1,\, Z = Z_{X=0}(u)}(u) =$
>   Value of Y for u in world where X = 1 and where Z = same value as of Z for u in world where X = 0.
> - Note nesting of quantifiers



Extended Example Mediation (cont'd)

- $Y_{X=1,Z=z}$ = hiring status with qualification $Z = z$ when treated as male ($X=1$)
- Averaging over possible qualifications for females

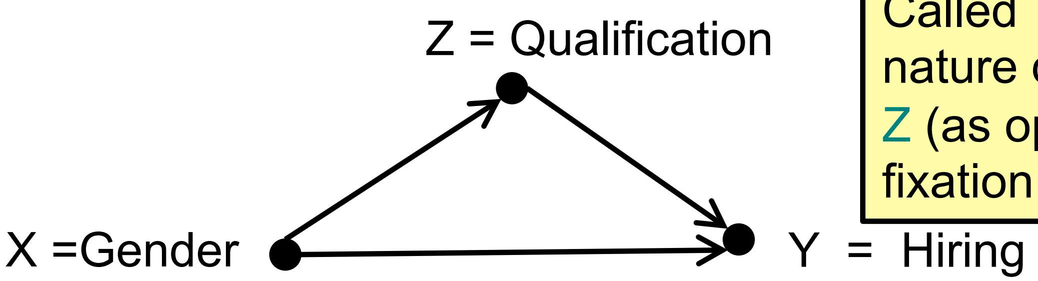
$$\sum_z E[Y_{X=1,Z=z}]P(Z=z|X=0) \qquad (= E[Y_{X=1,Z_{X=0}}])$$

- Averaging over possible qualifications for males

$$\sum_z E[Y_{X=1,Z=z}]P(Z=z|X=1) \qquad (= E[Y_{X=1,Z_{X=1}}])$$

- Natural indirect effect (NIE)

$$\sum_z E[Y_{X=1,Z=z}] \left(P(Z=z|X=0) - P(Z=z|X=1) \right)$$



Called ``natural" because nature determines value of Z (as opposed to controlled fixation in CDE)

Extended Example Mediation

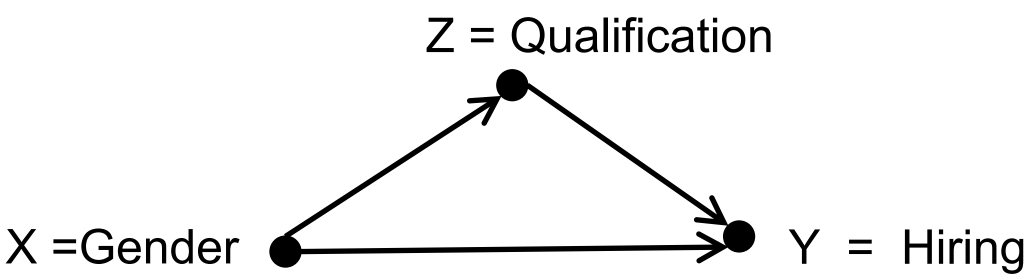
- Natural indirect effect (NIE)

$$\sum_z E[Y_{X=1,Z=z}] \, (P(Z=z|X=0) - P(Z=z|X=1))$$

- NIE identifiable from data in absence of confounding (Pearl 2001)

$$\sum_z E[Y| X=1,Z=z] \, (P(Z=z|X=0) - P(Z=z|X=1))$$

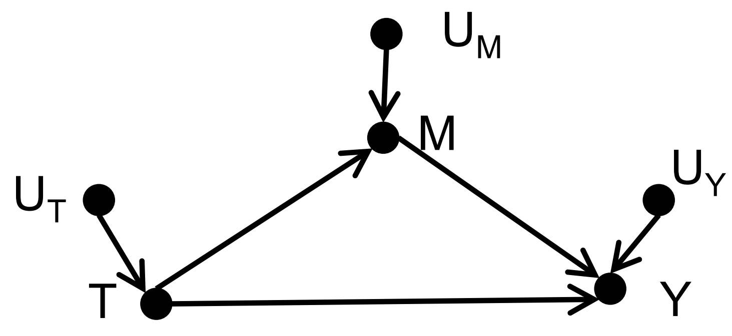
Pearl: Direct and indirect effects. Proceedings of the 7th Conference on Uncertainty in AI. 411-420, 2001



Toolkit for Mediation

Mediation problem

- $T = f(u_T)$;
- $m = f_M(t,u_M)$;
- $y = f_Y(t,m,u_Y)$

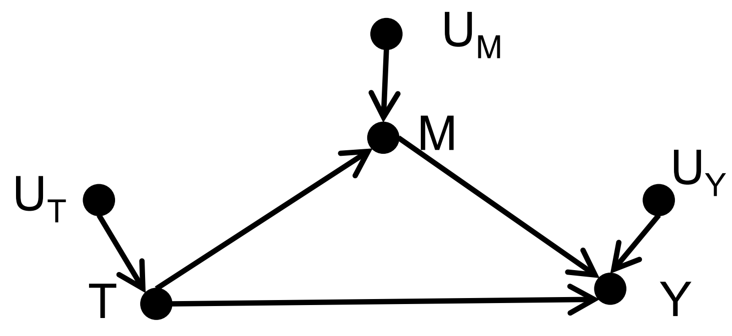


Effect	Formula
Total	$TE = E[Y_1 - Y_0] = E[Y \mid do(T=1)] - E[Y \mid do(T=0)]$
Controlled direct (for fixed mediator M=m)	$CDM(m) = E[Y_{1,m} - Y_{0,m}] = E[Y \mid do(T=1, M=m) - E[Y \mid do(T=0, M=m)]$
Natural direct	$NDE = E[\, Y_{1,M_0} - Y_{0,M_0}]$
Natural indirect	$NIE = E[\, Y_{0,M_1} - Y_{0,M_0}]$

Toolkit for Mediation

Mediation problem

- $T = f(u_T)$;
- $m = f_M(t,u_M)$;
- $y = f_Y(t,m,u_Y)$



Observations

- TE = NDE – NIE_r (for change T from 0 to 1)
  - where NIE_r is NIE under reverse transition of treatment, i.e., T changes from 1 to 0
- TE and CDE(m) are do-expressions, so estimable
  - from experimental data
  - or from observations with backdoor and front-door

Identification for NDE and NIE

- Consider set of covariates W such that
  1. No member of W descendant of T
  2. W blocks all M-Y backdoors after removing T-> M and T -> Y
  3. The W-specific effect is identifiable (using experiments or adjustment)
  4. The W-specific joint effect of {T,M} on Y is identifiable

     (using experiments or adjustment)

Theorem (Identification of NDE)
When 1.and 2. hold, then NDE identifiable by

$$NDE = \sum_m \sum_w [E[Y|do(T=1,M=m),W=w]- E[Y|do(T=0,M=m),W=w]] * P(M = m|do(T=0),W=w)P(W=w)$$

If additionally 3. and 4., then do expressions also identifiable by backdoor or front-door

Outlook: Logic meets ML

- Junction trees
- (Logical) Constraints for constraining ML models
- PAC framework (probably approximately correct)
- PAC learning in logical framework