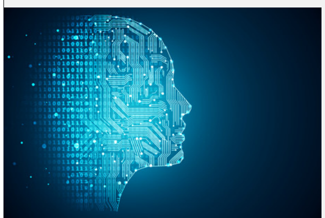


Data Science for Researchers and Scholars

Vasant G. Honavar
Dorothy Foehr Huck and J. Lloyd Huck Chair in Biomedical Data Sciences and Artificial Intelligence
Professor of Data Sciences, Informatics, Computer Science and Engineering, Bioinformatics & Genomics, Public Health Sciences and Neuroscience
Director, Center for Artificial Intelligence Foundations and Scientific Applications
Associate Director, Institute for Computational and Data Sciences
Pennsylvania State University

vhonavar@psu.edu
http://faculty.ist.psu.edu/vhonavar
http://ailab.ist.psu.edu

Pearl's do-calculus is complete for many more problems

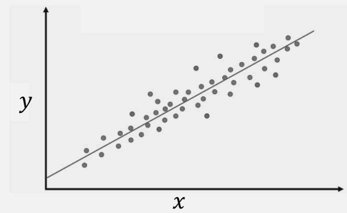
- Identifiability using surrogate variables Z when X is not experimentally manipulable was solved in 2012 by Bareinboim and Pearl
- Causal effect transportability – solved by Pearl and Bareinboim, Lee and Honavar, Bareinboim and Pearl, Bareinboim, Lee, Honavar, Pearl (2012-2013 AAAI, UAI, NeurIPS)
- Identifying the intervention cover of a causal graph (Kandasamy, Bhattacharya, and Honavar, AAAI 2019)
- Variants of do-calculus for relational causal models (Lee and Honavar, UAI 2016, Lee and Honavar, AAAI 2020)

Do-calculus is for causal inference what Newton's laws of motion are for classical physics

Linear Structural Causal Models

- Linear Regression
- Introduction to Linear Structural Causal Models
- When regression can and cannot be used to find causal effects.
- Identification in linear SCM

Regression



- Predict the value of Y based on X
- Supervised machine learning is often just regression on steroids
- How do we fit a regression line?
  - Given a data set of X, Y pairs, we fit them to $y = mx + b$ so as to minimize

$$\sum_i (y_i - b - mx_i)^2$$

  - m denotes the slope and b the intercept along the Y axis

Regression Coefficient

- R_{YX} is slope of regression line of Y *on* X
- $m = R_{YX} = \sigma_{XY}/{\sigma_X}^2$
- Slope gives correlation
  - Positive slope → positive correlation
  - Negative slope → negative correlation
  - Zero slope → X and Y are independent or non-linearly correlated

Variance of X, i.e., ${\sigma_X}^2 = \mathbb{E}[(X - \mathbb{E}[X])^2]$

Covariance $\sigma_{XY} \triangleq \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$

Correlation coefficient $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\, \sigma_Y}$

Multiple Regression

- $y = r_0 + r_1 \cdot x + r_2 \cdot z$
- How do we visualize?: a plane
- What happens if we fix X at some value?
  - $r_1 \cdot x$ becomes a constant
- r_2 is now the slope of slice along X–axis
- What happens if we fix Z at some value?
  - $r_2 \cdot z$ becomes a constant
- r_1 is now the slope of slice along Z–axis

Interpreting regression coefficients

Example: If $y = 1 + 2x_1 + 3x_2$

- Do not interpret the coefficients unless they are statistically significant.
- It *is NOT* accurate to say "For each change of 1 unit in x_1, y changes 2 units".
- What *is* correct to say is "If x_2 is fixed, then for each change of 1 unit in x_1, y changes 2 units."

Linear Structural Causal Models

Linear SCM are defined as a system of linear equations representing ground-truth:

$$Y := \sum_i \lambda_{x_i y} X_i + \mathcal{E}_y$$

1. All correlations between $\mathcal{E}$ are explicitly specified.
2. X_i are the direct causes of Y, and $\lambda_{x_i y}$ is the change in Y per X_i.
3. WLOG assume normalized data ($\mathbf{E}[X] = 0$ and $\mathbf{E}[XX] = 1$) to simplify math
4. Assume $\mathcal{E}_y \sim \mathcal{N}$, meaning that the distribution is fully specified by covariance matrix Σ (σ_{ij}).

Causal Inference In Linear Systems

- What is the effect of salt intake on blood pressure after adjusting for confounders; or the total effect of an after-school study program on test scores;
- What is the direct effect or the unmediated by other variables, of the program on test scores.
- What is the effect of enrollment in an optional work training program on future earnings, when enrollment and earnings are confounded by a common cause (e.g., motivation).
- Continuous variables
  - We need to model with continuous variables.
  - We will assume linear relationships and Normal distributions of errors.

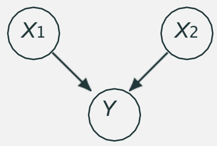
Non-Parametric to Linear

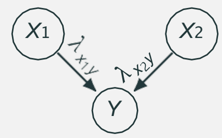
The only substantive change we are making is that the function f becomes linear:

$$V_i \leftarrow f_i(pa_i, U_i) \quad \Rightarrow \quad V_i \leftarrow \sum_{j|V_j \in pa_i} \lambda_{ji} V_j + \mathcal{E}_i$$

1. λ_{ji} is called the "Structural Coefficient".
2. Instead of using U_i, we rename it to $\mathcal{E}_i$ by convention.
3. If we know all λ_{ji}, we can find the causal effect of V_j on V_i.

Example: linear structural causal model





$$X_1 = f_{x_1}(U_{x_1})$$
$$X_2 = f_{x_2}(U_{x_2})$$
$$Y = f_y(X_1, X_2, U_y)$$

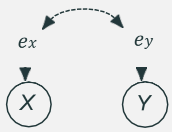
$$X_1 = \varepsilon_{x_1}$$
$$X_2 = \varepsilon_{x_2}$$
$$Y = \lambda_{x_1 y} X_1 + \lambda_{x2y} X_2 + \varepsilon_y$$

We can draw the structural coefficients directly on the graph, which then fully specifies the model.

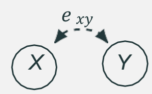
Example: linear structural causal model

The covariance between e_i and e_j is represented by e_{ij}, and is used as the value of a bidirected edge:

Latent Confounding



$\Rightarrow$



$$e_{xy} \equiv \mathbb{E}[e_x e_y]$$

- e_{xy} is unobserved, since it is covariance of latent variables. It is mathematically useful, however, so we draw it on the graph just like structural coefficients.

Linear SCM: Interventions

$$X \xrightarrow{\lambda} Y$$

$$\mathbb{E}[Y \mid do(X = x)] = ?$$

Linear SCM: Interventions

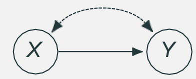
$$X \xrightarrow{\lambda} Y$$

$$\begin{aligned}\mathbb{E}[Y|do(X=x)] &= \mathbb{E}[\lambda x + e_Y] \\ &= \lambda x + \mathbb{E}[e_Y] \\ &= \lambda x\end{aligned}$$

Note that *x* is a value of *X*

Linear SCM

- Graph: We are assuming that you have a hypothesized causal graph structure. In other words, you think you know what causes what, and which variables have an unknown common cause.
- Observational Data: You have a set of data samples with measurements of all of the observable variables.
- Goal: Find Structural Coefficients You do NOT have knowledge of the underlying structural coefficients. These represent the actual causal effects that we want to find.

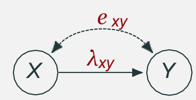


(x_1, y_1)

(x_2, y_2)

…

(x_n, y_n)



Linear SCM: Interventions

Remember that we assumed $e \sim N$, meaning that the distribution is fully specified by covariance matrix Σ (σ_{xy}).

$$X \xrightarrow{\lambda} Y$$

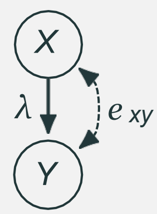
$$\begin{aligned}
\sigma_{xy} &= \mathbb{E}[XY] \\
&= \mathbb{E}[X(\lambda X + e_y)] \\
&= \mathbb{E}[\lambda XX + Xe_y] \\
&= \lambda \mathbb{E}[XX] + \mathbb{E}[Xe_y] \\
&= \lambda 1 + 0 \\
&= \lambda
\end{aligned}$$

Remember, we normailize
The mean to 0 and variance to 1

Connecting Observed with Unobserved

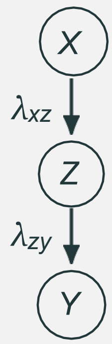
Solve for σ_{xy} in terms of the structural coefficients λ and e_{xy}

$$\sigma_{xy} = \mathbb{E}[XY]$$



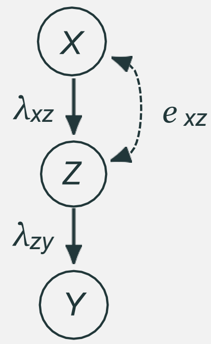
$$\sigma_{xy} = \mathbb{E}[XY] \\ = \lambda + e_{xy}$$

A Curious Property of Linear Causal Models



$$
\begin{aligned}
\sigma_{xy} &= \mathbb{E}[XY] \\
&= \lambda_{zy}\lambda_{xz}
\end{aligned}
$$

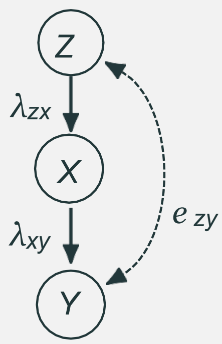
A Curious Property of Linear Causal Models



$$\sigma_{xy} = \mathbb{E}[XY]$$
$$= \lambda_{zy}\lambda_{xz} + \lambda_{zy}\, e_{xz}$$

Paths and Covariances

There is a relationship between covariances and paths in the graph.

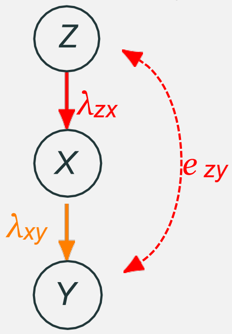


$$\sigma_{xy} = \mathbb{E}[XY] = \mathbb{E}[X(\lambda_{xy}X + e_y)]$$
$$= \lambda_{xy}\ \mathbb{E}[XX] + \mathbb{E}[Xe_y]$$
$$= \lambda_{xy} + \mathbb{E}[(\lambda_{zx}Z + e_x)e_y]$$
$$= \lambda_{xy} + \lambda_{zx}\ \mathbb{E}[e_z e_y] + \mathbb{E}[e_x e_y]$$
$$= \lambda_{xy} + \lambda_{zx}\, e_{zy}$$

e_x and e_y are uncorrelated
$\mathbb{E}[e_z e_y] = e_{zy}$ by definition

Paths and Covariances

There is a relationship between covariances and paths in the graph.

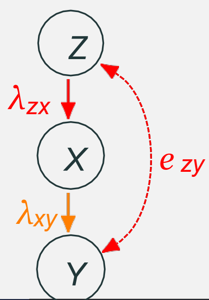


$$\sigma_{xy} = \lambda x y + \lambda_{zx}\, e_{zy}$$

The resulting terms correspond to paths between X and Y in the causal graph

Wright's Rule

The covariance between variables X and Y is the sum of the contributions of the paths between them in the causal graph, i.e. any non-self-intersecting path without colliding arrowheads (→←)

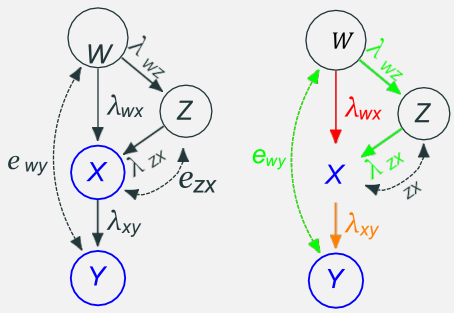


$$\sigma_{xy} = Assoc(X \rightarrow Y) + Assoc(X \leftarrow Z \leftarrow Y)$$

$$\sigma_{xy} = \lambda_{xy} + \lambda_{zx} e_{zy}$$

Reading Covariances off the Graph

The covariance between variables X and Y is the sum of open paths between them in the causal graph, so paths with no colliding arrowheads ($\rightarrow\leftarrow$)



$$\sigma_{xy} = \lambda_{xy} + \lambda_{wx}\, e_{wy} + \lambda_{zx}\lambda_{wz} e_{wy}$$

Wright's Rules

σ_{xy} = Sum of products of path coefficients along all open paths between X and Y

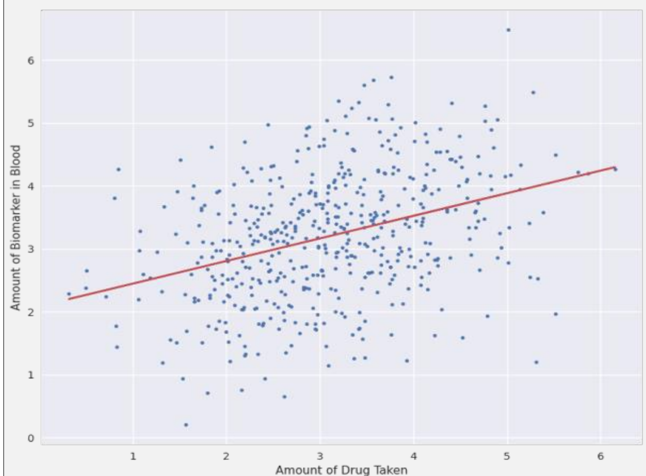
Wright's Rules (1921)

- σ_{xy} is 0 only when X and Y are d-separated.
- If there is an edge $X \xrightarrow{\alpha} Y$ in the model, then $\sigma_{xy} = \alpha +$ contributions of other paths between X and Y.
  - $\sigma_{xy} = \alpha$ if X and Y are d-separated in G_α (G with edge α removed)
- Wright's rules are defined for acyclic models (DAG)

Linear Regression

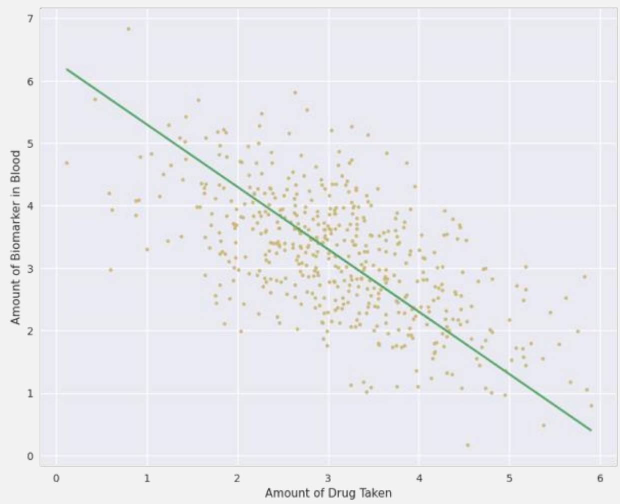
- Suppose you want to determine if a new drug is helpful for curing a disease

Perform regression



- Perform a regression $Y = \beta X + e$ on the data, with X being drug dosage, and Y biomarker measured giving $\beta = 0.375$
- Drug seems helpful, so you recommend it

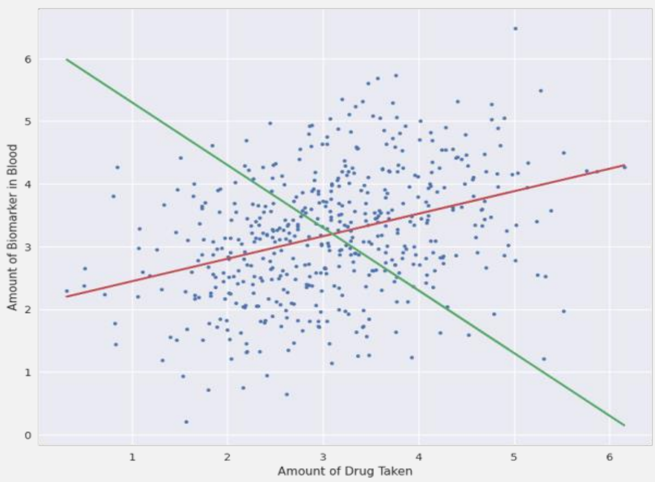
What happens when the drug is given to everyone?



- When the drug is given to everyone in the population, you find a clear negative association between drug dosage and blood antibodies, with slope −1.
- This drug actually seems to hurt people!

Why did regression mislead us here?

What's Happening Here?

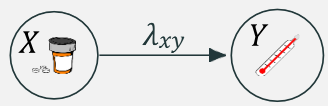


- Why was this negative effect (green line) not apparent from regression on the original dataset?
- Association ≠ causation!
- Can we get causation from the original dataset?

Why did regression mislead us here?

Key Assumption: Lack of Confounding

The following world model is implicitly assumed when attributing causal meaning to the regression coefficient:

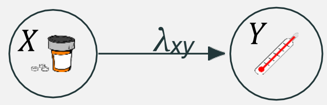


$$X := e_x$$
$$Y := \lambda_{xy} X + e_y$$

e_x, e_y independent

Why did regression mislead us here?

The following world model is implicitly assumed when attributing causal meaning to the regression coefficient:



$$X := e_x$$
$$Y := \lambda_{xy} X + e_y$$

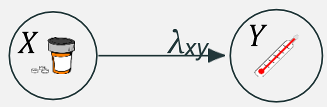
e_x, e_y independent

Regression $Y = \beta X + e$ gives correct $\beta = \lambda_{xy}$

The key assumption is lack of confounding!

Why did regression mislead us here?

The following world model (lack of confounding) is implicitly assumed when attributing causal meaning to the regression coefficient:



$$X := e_x$$
$$Y := \lambda_{xy} X + e_y$$

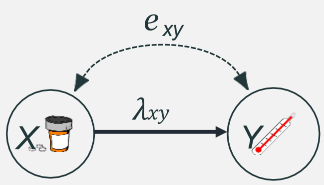
e_x, e_y independent

Covariance gives the same answer:

$$\sigma_{xy} = \mathbb{E}[XY] = \mathbb{E}[X(\lambda_{xy} X + e_y)] = \lambda_{xy}\,\mathbb{E}[XX] + \mathbb{E}[Xe_y] = \lambda_{xy}$$

The True Scenario

If one is unable to ascertain the assumption of no confounding between X and Y, this is the corresponding graphical model



$$X := e_x$$
$$Y := \lambda_{xy}X + e_y$$

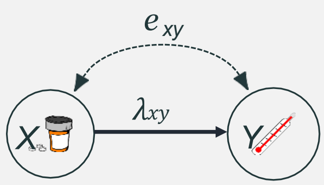
e_x, e_y correlated

May be

- The drug is expensive so mostly rich people are getting it.
- Rich people also tend to get better care overall and hence have a better chance of recovery
- But data about financial status not gathered

The True Scenario

If one is unable to ascertain the assumption of no confounding between X and Y, this is the corresponding graphical model



$X := e_x$
$Y := \lambda_{xy} X + e_y$

e_x, e_y correlated

- Regression $Y = \beta X + e$ gives a biased answer

  $\sigma_{xy} = \lambda_{xy}\mathbb{E}[XX] + \mathbb{E}[e_x e_y]$

  $\sigma_{xy} = \lambda_{xy} + e_{xy}$

- In this case, the causal effect of the drug X on blood antibodies Y is provably unidentifiable from observational data
- What can you do? Run an RCT!

What does Regression Compute?

$$Y = \beta X + e$$

We want to minimize the square of the error between Y and βX

$$\begin{aligned}\mathbb{E}[(Y - \beta X)^2] &= \mathbb{E}[YY - 2\beta XY + \beta^2 XX] \\ &= \mathbb{E}[YY] - 2\beta\,\mathbb{E}[XY] + \beta^2\,\mathbb{E}[XX] \\ &= 1 + \beta^2 - 2\beta\,\mathbb{E}[XY] \\ &= 1 + \beta^2 - 2\beta\sigma_{xy}\end{aligned}$$

Solving $\frac{\partial}{\partial\beta}(1 + \beta^2 - 2\beta\sigma_{xy}) = (2\beta - 2\sigma_{xy}) = 0$

We get: $\beta = \sigma_{xy}$

The regression coefficient is just the covariance between X and Y!

What does Regression Compute?

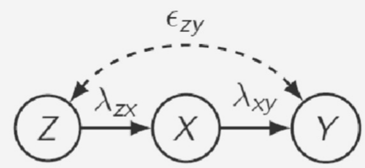
- The regression equation $Y = \beta X + e$ assumes $e \coprod X$
- The solution of the regression equation is: $\beta = \sigma_{xy}$.
- We will call this value r_{yx} (solved value of linear regression of Y on X)
- Knowledge of r_{yx} supports no causal claims.

- In contrast, the structural causal model $X \xrightarrow{\lambda} Y$

- Corresponds to the structural equation $Y = \lambda X + e_y$
- which implies $\mathbb{E}[Y|do(X)] = \lambda X$

- The structural model makes causal claims, that is, claims about the interventional distribution which can be tested, and can be falsified.
- The SCM and regression equation look similar but have different interpretations.

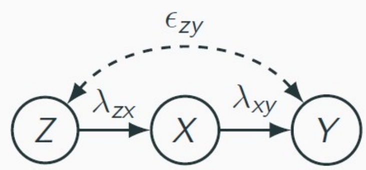
Equations for Causal Effect Identification in Linear Causal Models



$$\begin{bmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_{zz} \end{bmatrix} = \begin{bmatrix} 1 & \lambda_{xy} + \lambda_{zx}\epsilon_{zy} & \lambda_{zx} \\ \lambda_{xy} + \lambda_{zx}\epsilon_{zy} & 1 & \lambda_{zx}\lambda_{xy} + \epsilon_{zy} \\ \lambda_{zx} & \lambda_{zx}\lambda_{xy} + \epsilon_{zy} & 1 \end{bmatrix}$$

- Note that the sigmas can be expressed in terms of lambdas using techniques previously introduced (path analysis)

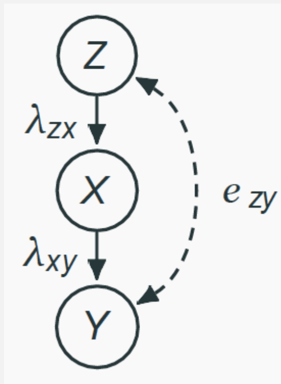
Equations for Causal Effect Identification in Linear Causal Models



$$\begin{bmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_{zz} \end{bmatrix} = \begin{bmatrix} 1 & \lambda_{xy} + \lambda_{zx}\epsilon_{zy} & \lambda_{zx} \\ \lambda_{xy} + \lambda_{zx}\epsilon_{zy} & 1 & \lambda_{zx}\lambda_{xy} + \epsilon_{zy} \\ \lambda_{zx} & \lambda_{zx}\lambda_{xy} + \epsilon_{zy} & 1 \end{bmatrix}$$

- Covariance matrix Σ is symmetric
- Only the entries in the lower or upper triangle need to be considered

Causal Effect Identification in Linear Causal Models

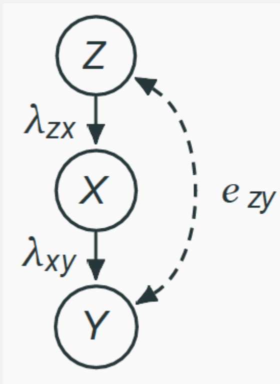


- Given a SCM and an observational dataset, is it possible to uniquely determine λ_{xy}?
- Can λ_{xy} be solved in terms of Σ?

$$\sigma_{xz} = \lambda_{zx}$$
$$\sigma_{xy} = \lambda_{xy} + \lambda_{zx}\, e_{zy}$$
$$\sigma_{zy} = \lambda_{zx}\lambda_{xy} + e_{zy}$$

- Σ can be estimated from the observational data (and hence known)
- The Λ need to be solved for

Causal Effect Identification in Linear Causal Models

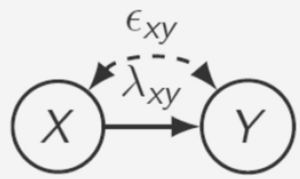


- Can Λ be solved in terms of Σ?

$$\sigma_{xz} = \lambda_{zx}$$
$$\sigma_{xy} = \lambda_{xy} + \lambda_{zx} e_{zy}$$
$$\sigma_{zy} = \lambda_{zx}\lambda_{xy} + e_{zy}$$

- λ_{zx} can be solved from the first equation
- Substituting λ_{zx} into the remaining 2 equations, we get 2 equations in 2 unknowns
- Hence, we can solve for Λ from Σ
- The given linear causal model can be identified from observational data

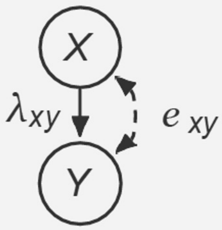
Causal Effect Identification in Linear Causal Models



$$\begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{bmatrix} = \begin{bmatrix} 1 & \lambda_{xy} + \epsilon_{xy} \\ \lambda_{xy} + \epsilon_{xy} & 1 \end{bmatrix}$$

- Can Λ e solved in terms of Σ?

Causal Effect Identification in Linear Causal Models



- Can Λ e solved in terms of Σ?

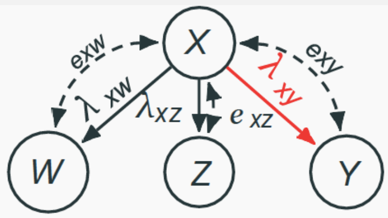
$$\begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{bmatrix} = \begin{bmatrix} 1 & \lambda_{xy} + \epsilon_{xy} \\ \lambda_{xy} + \epsilon_{xy} & 1 \end{bmatrix}$$

- We have one equation in 2 unknowns

$$\sigma_{xy} = \lambda_{xy} + e_{xy}$$

- There is no unique solution for λ_{xy} or e_{xy}

Causal Effect Identification in Linear Causal Models



$$\sigma_{xw} = \lambda_{xw} + e_{xw} \qquad \sigma_{wz} = \lambda_{xw}\lambda_{xz} + \lambda_{xz}\, e_{xw} + \lambda_{xw}\, e_{xz}$$

$$\sigma_{xz} = \lambda_{xz} + e_{xz} \qquad \sigma_{wy} = \lambda_{xw}\lambda_{xy} + \lambda_{xw}\, e_{xy} + \lambda_{xy}\, e_{xw}$$

$$\sigma_{xy} = \lambda_{xy} + e_{xy} \qquad \sigma_{zy} = \lambda_{xz}\lambda_{xy} + \lambda_{xz}\, e_{xy} + \lambda_{xy}\, e_{xz}$$

- Can we identify λ_{XY}?
- Yes, by solving the system of equations

Causal Effect Identification in Linear Causal Models

- $P(Y|do(X))$ Identifiable: Unique value of λ_{XY} consistent with observational data
- $P(Y|do(X))$ NOT identifiable: Infinite set of possible solutions for λ_{XY} consistent with observational data
- $P(Y|do(X))$ finite identifiable: if there is only a finite number of solutions for λ_{XY} that are consistent with observational data

Causal transportability[1]

- Suppose we have run a study in Chicago and learned a causal relationship, say between poverty and obesity
- Suppose we want to see if the relationship is true in some form in Los Angeles
  - Los Angeles is different from Chicago in some respects, e.g., demographics
- We now have tools to answer if the causal relationship which we learned from a study in Chicago can be tweaked in some way so that it applies to Los Angeles

[1] Bareinboim and Pearl, 2012; Lee and Honavar, 2013a; 2013b, Bareinboim, Lee, Honavar, and Pearl, 2013, Bareinboim and Pearl, 2016; Lee et al., 2019.

Transportability of Causal Effects Across Populations

Source Population Π

Target Population Π^*

$P(y, t, x \cdots)$

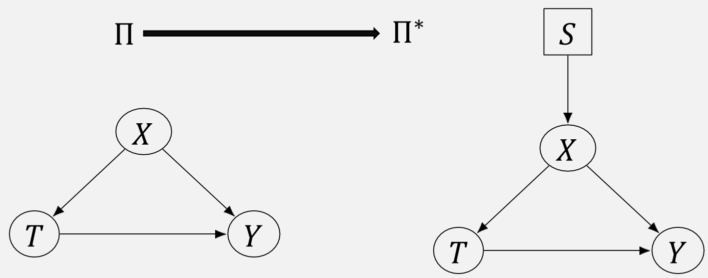
$P^*(y, t, x \cdots)$

Given $P(y \mid do(t), x)$

$P(y \mid do(t), x) \stackrel{?}{=} P^*(y \mid do(t), x)$

Selection Diagrams

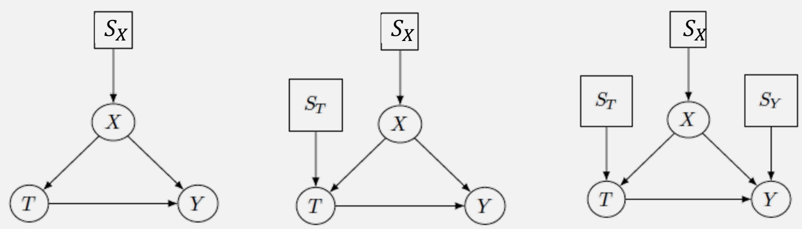
- Represent different causal mechanisms across the source and target distributions (Π and Π^*)



Selection Diagrams

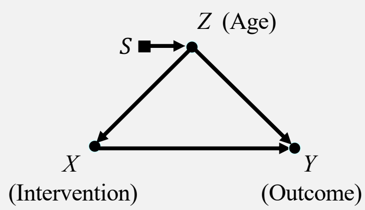
Selection diagrams

- Allow for different causal mechanisms across the source and target distributions (Π and Π^*)



Π_1^* Π_2^* Π_3^*

Causal transportability



- Source Π and target Π^* differ with respect to the distribution of Z (Age)
- Π includes all ages, Π^* includes only young
- Indicated by the "selection" arrow into Z

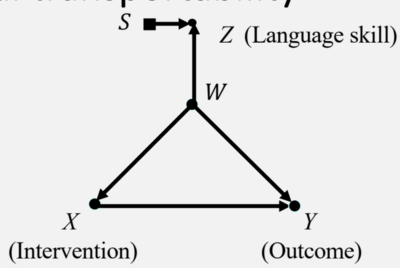
Experimental study in LA
Measured: $P(x, y, z), P(y|do(x), z)$

Observational study in NYC
Measured: $P^*(x, y, z)$
$P^*(z) \neq P(z)$

Needed:
$$Q = P^*(y|do(x)) = \sum_z P(y|do(x), z)P^*(z)$$

Transport Formula: $F(P, P_{do}, P^*)$

Causal transportability



- Source Π and target Π^* differ with respect to the distribution of Z
- For example, Π and Π^* differ through "selection" on Z (language skill)

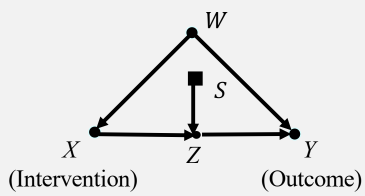
Experimental study in LA
Measured: $P(x, y, w, z), P(y|do(x), w)$
Needed:
$Q = P^*(y|do(x)) = P(y|do(x))$

Transport Formula: $F(P, P_{do}, P^*)$

Observational study in NYC
Measured: $P^*(x, y, w, z)$
$P^*(z) \neq P(z)$

- Not all differences between source and target matter!

Causal transportability



- Source Π and target Π^* differ with respect to the distribution of Z
- For example, Π and Π^* differ through "selection" on Z (language skill)

Experimental study in LA
Measured: $P(x, y, w, z), P(y|do(x), w)$

Observational study in NYC
Measured: $P^*(x, y, w, z)$
$P^*(z) \neq P(z)$

Needed:
$Q = P^*(y|do(x)) = P(y|do(x), z)\ P^*(z|x)$

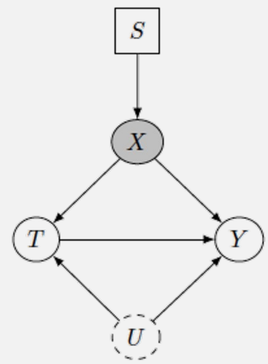
Transport Formula: $F(P, P_{do}, P^*)$

Causal transportability reduced to do-calculus

- Theorem: A causal relation $\boldsymbol{R}$ is transportable from a source domain Π to a target domain Π^*
  - if and only if it is reducible, using the rules of *do*-calculus, to an expression in which the selection variable(s) $\boldsymbol{S}$ is(are) separated from *do*( ).

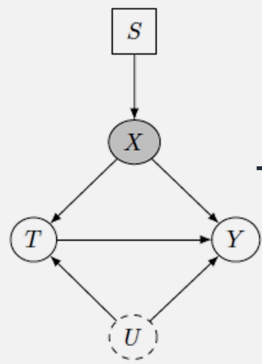
Transportability

$$P(y \mid do(t), x) \stackrel{?}{=} P^*(y \mid do(t), x)$$

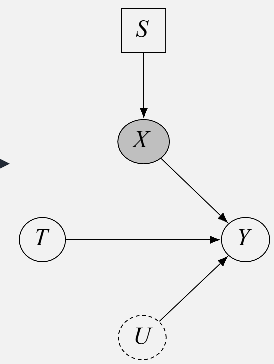


Transportability and do-calculus

$$P(y \mid do(t), x) \stackrel{?}{=} P^*(y \mid do(t), x)$$



Intervention on T



Proof:

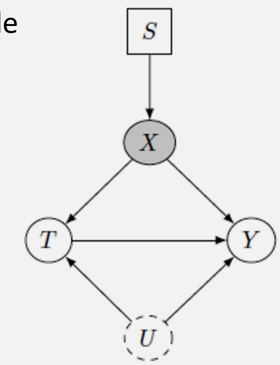
$$\begin{aligned} P^*(y \mid do(t), x) &= P(y \mid do(t), x, s^*) \\ &= P(y \mid do(t), x) \end{aligned}$$

External validity (direct transportability)

- We say that the causal effect $P(y|do(t),x)$ is directly transportable from source domain Π to a target domain Π^* if

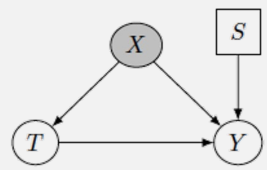
$$P(y \mid do(t), x) = P^*(y \mid do(t), x)$$

- Such a causal effect is said to have external validity (or generalizability)



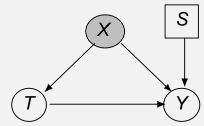
Transportability

$$P(y \mid do(t), x) \stackrel{?}{=} P^*(y \mid do(t), x)$$



$$P(y \mid do(t), x) \neq P^*(y \mid do(t), x)$$

Trivial Transportability



- We clearly don't have direct transportability
  - $P(y \mid do(t), x) \neq P^*(y \mid do(t), x)$
- Suppose we have access to observational data from the target population: $P^*(y, t, x)$
- Then we can identify $P^*(y \mid do(t), x)$ using only target data
  - $P^*(y \mid do(t), x) = P^*(y \mid t, x)$
- If a causal effect is identifiable from observational data in the target domain,
  - We do not need any information from the source domain to estimate it
  - It is trivially transportable from any source domain

Causal transportability – general version

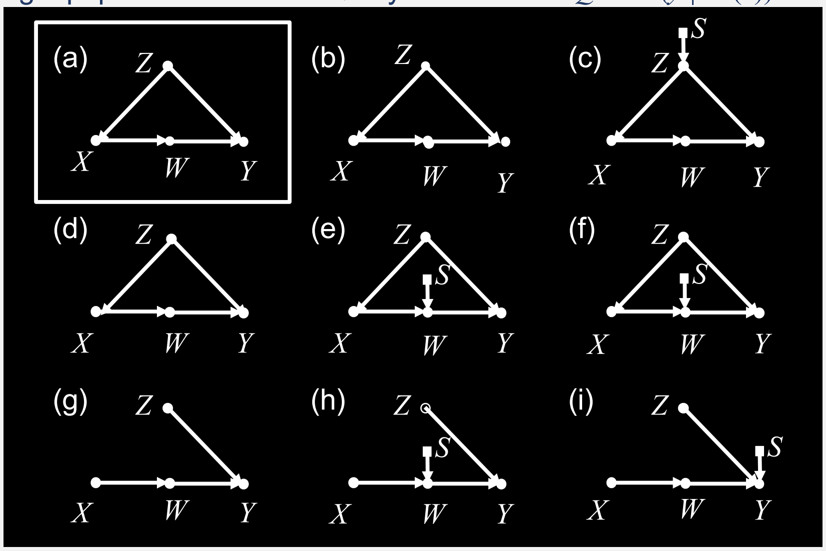
- How to combine results of
  - several experimental and observational studies,
  - each conducted on a different population and under a different set of conditions,
  - to construct a valid estimate a causal effect of interest,
  - in a new (target) population,
  - that may be different from any of the ones studied

Causal transportability

Target population Π^* Query of interest: $Q = P^*(y \mid do(x))$

(a) Arkansas Only survey data	(b) New York Only survey data Resembling target	(c) Los Angeles Only survey data Young fashionistas
(d) Boston Age not recorded Mostly educated scholars	(e) San Francisco Mostly techies	(f) Texas Mostly Hispanics
(g) State College RCT College students	(h) Utah RCT paid volunteers	(i) Wyoming RCT, young athletes

Target population Π^* Query of interest: $Q = P^*(y \mid do(x))$



Summary

- Given the commonalities and differences between one or more source domains and a target domain encoded in selection diagrams, transportability of a causal effect of interest from the source domain(s) to a target domain can be determined using do-calculus
- When an effect is transportable, the transport formula can be derived in time that is polynomial in the size of the formula
- The algorithm is sound and complete
- Corollary do-calculus is complete for causal transportability

Further generalizations

- mz-transportability
  - Identification from proxy experiments
  - Multiple transportability
- Meta analysis

Do calculus for causal inference

Do calculus is complete for

- ✓Causal transportability
  - Bareinboim & Pearl, 2012
- ✓Causal m-transposability
  - Bareinboim & Pearl, 2013; Lee and Honavar, 2013
- ✓Causal z transposability
  - Bareinboim & Pearl, 2013; Lee & Honavar, 2013
- ✓Causal mz-transportability
  - Bareinboim, Lee, Honavar & Pearl, 2013
- ✓Meta analysis
  - Bareinboim et al., 2016; Lee et al., 2019

Analyses have been extended to non IID setting (Lee and Honavar, 2015, 2016, 2020)

Do-calculus is for causal inference what Newton's laws of motion are for classical physics