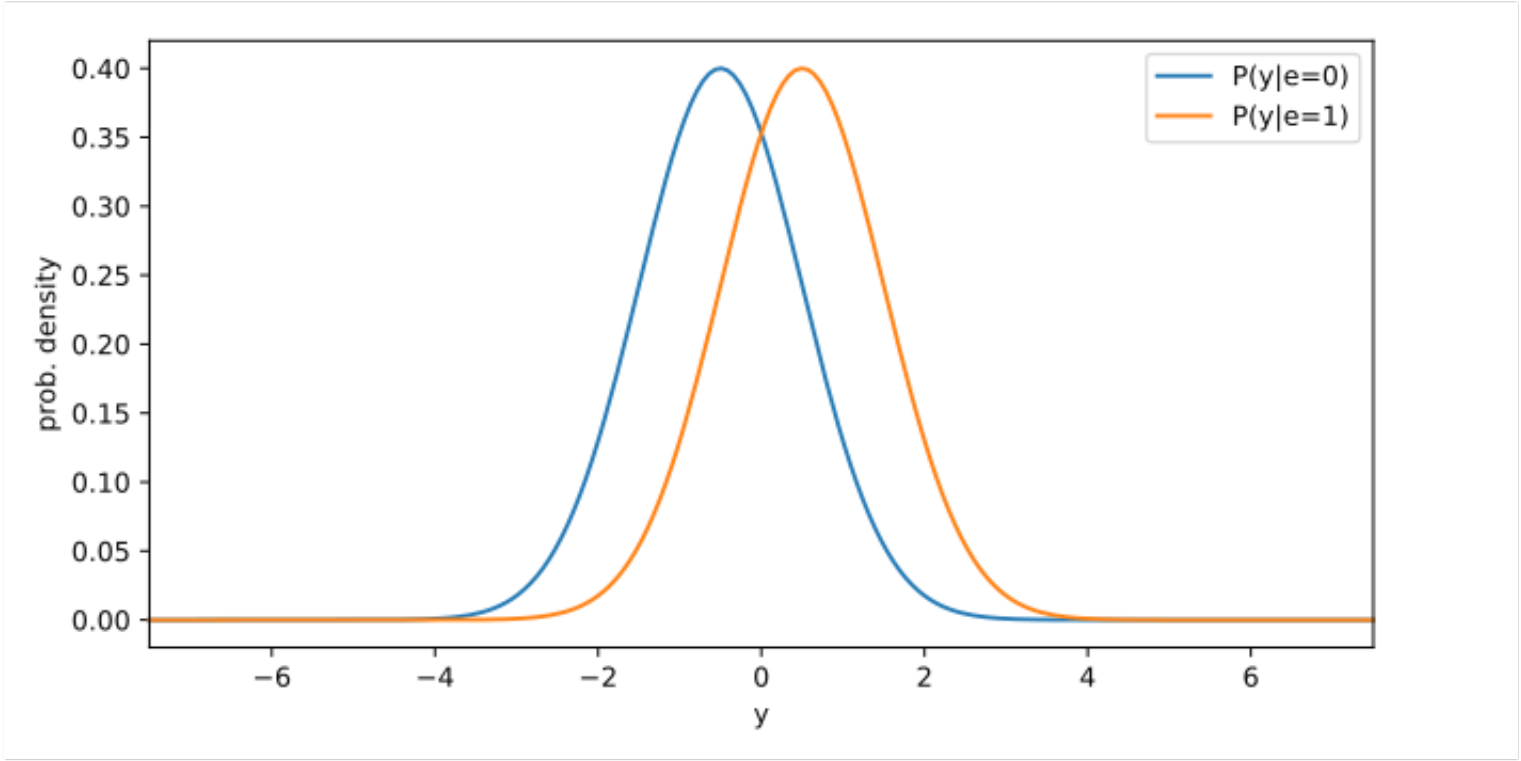
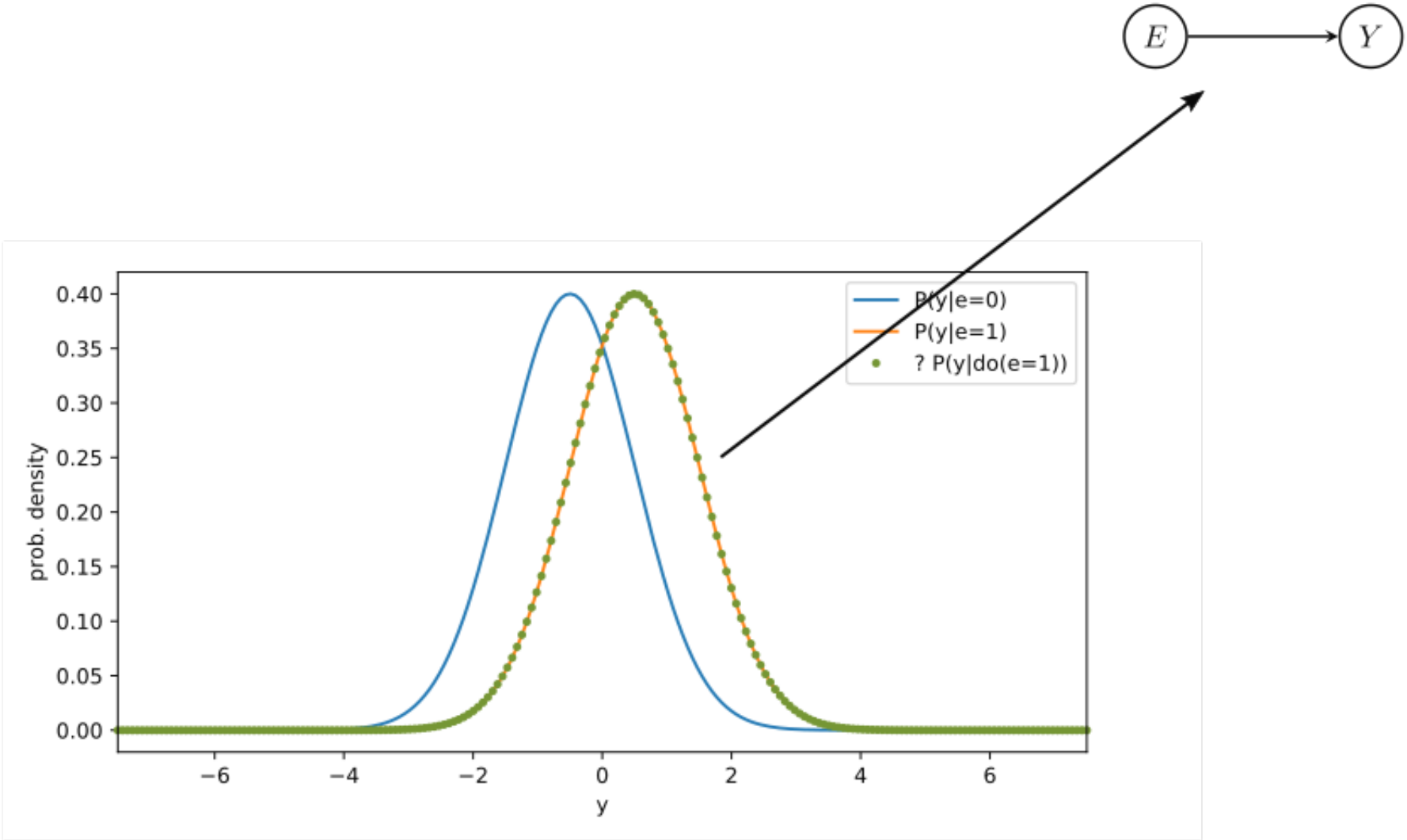
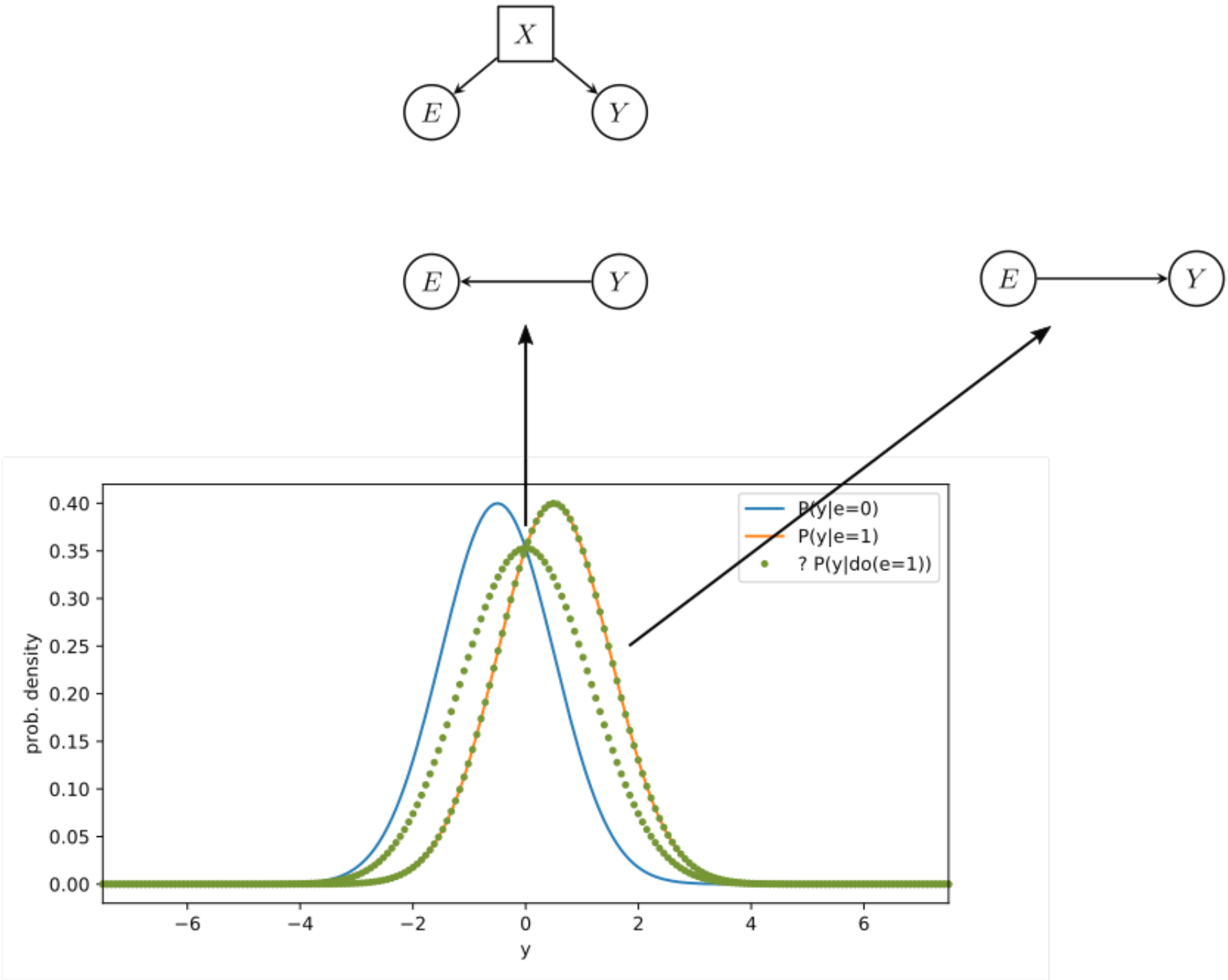
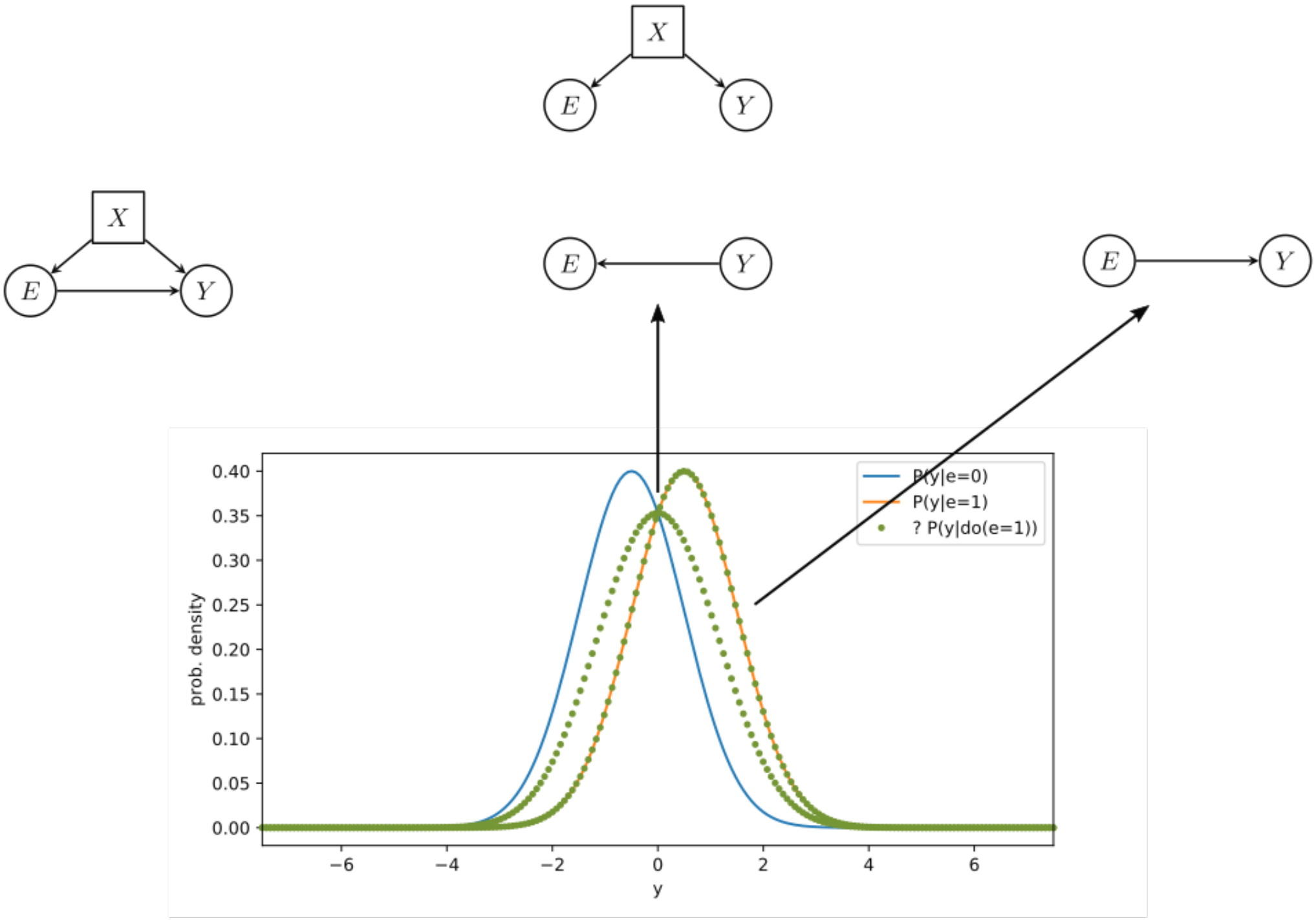
Introduction to causal discovery and inference

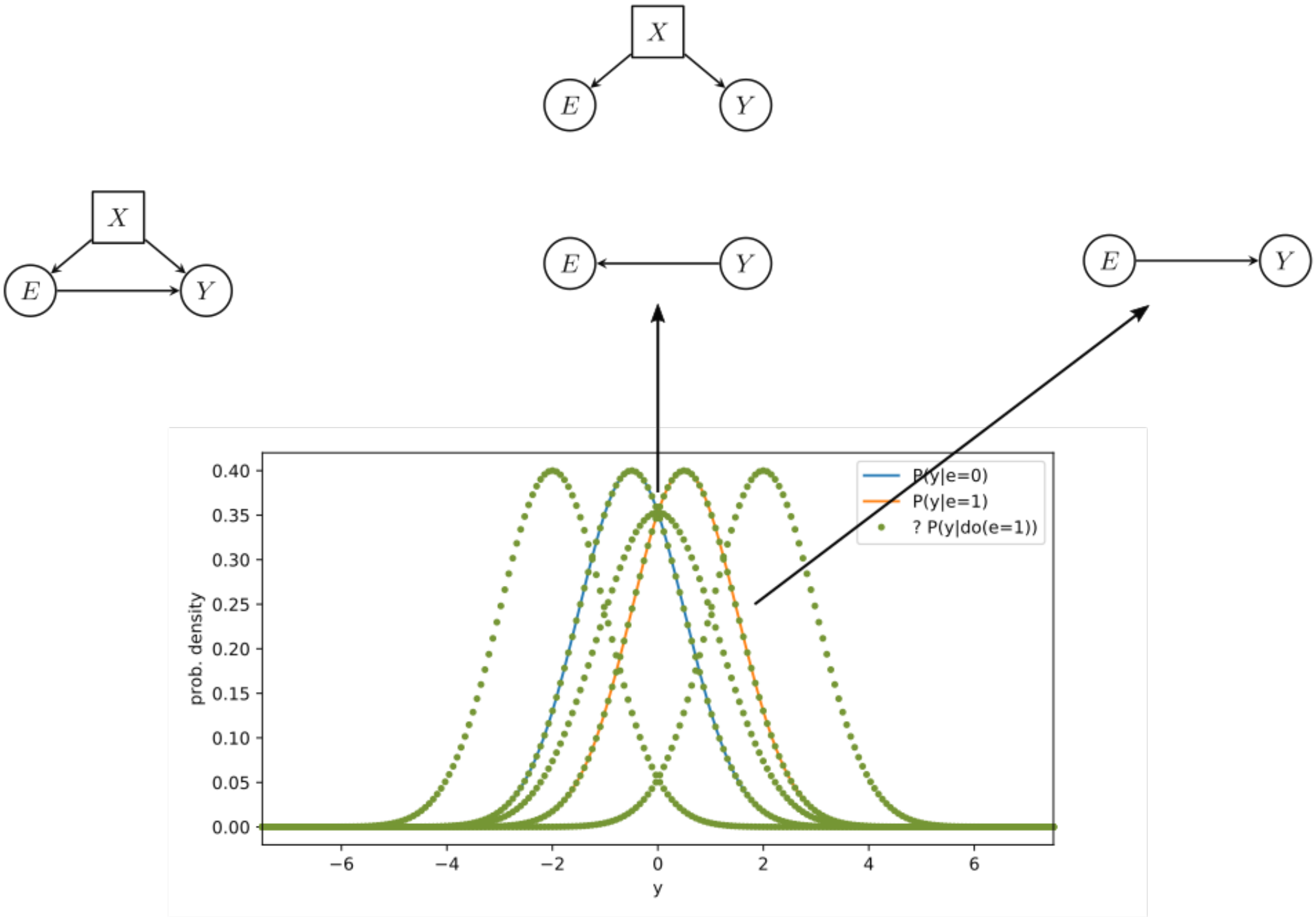
Robert Różański


P(y|e=0)
P(y|e=1)
prob. density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−6
−4
−2
0
2
4
6
y

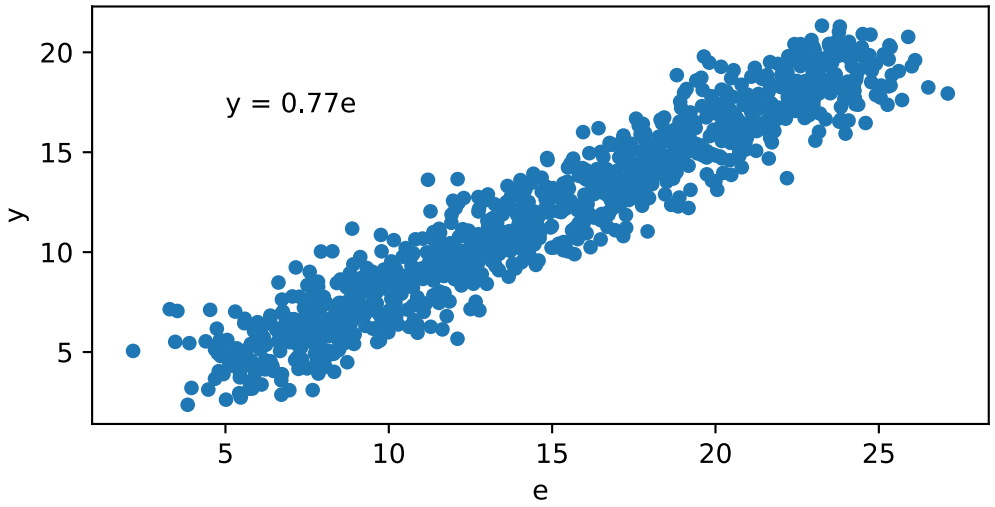

E
Y
P(y|e=0)
P(y|e=1)
? P(y|do(e=1))
prob. density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−6
−4
−2
0
2
4
6
y


X
E
Y
E
Y
E
Y
P(y|e=0)
P(y|e=1)
? P(y|do(e=1))
prob. density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−6
−4
−2
0
2
4
6
y

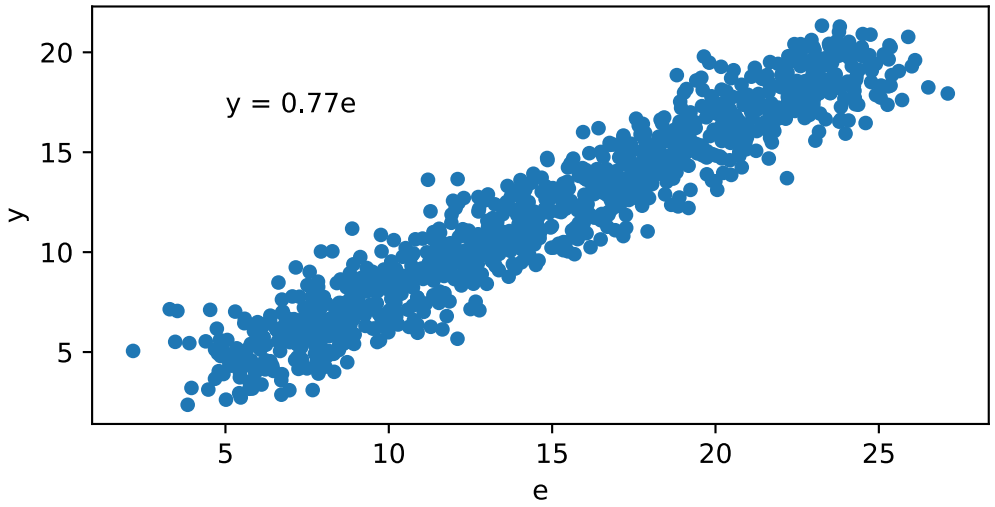

X
E
Y
X
E
Y
E
Y
E
Y
P(y|e=0)
P(y|e=1)
? P(y|do(e=1))
prob. density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−6
−4
−2
0
2
4
6
y

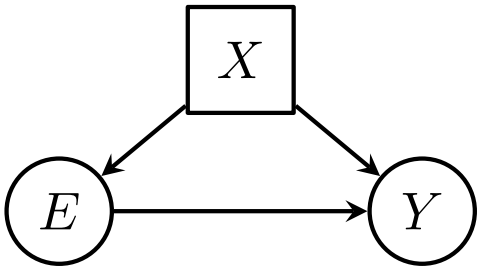

X
E
Y
X
E
Y
E
Y
E
Y
P(y|e=0)
P(y|e=1)
? P(y|do(e=1))
prob. density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−6
−4
−2
0
2
4
6
y

Simpson's paradox

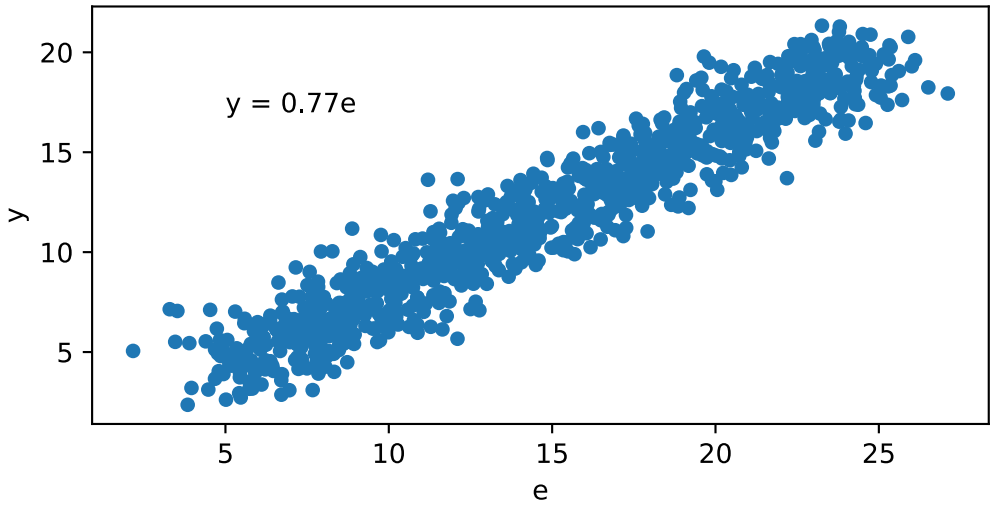


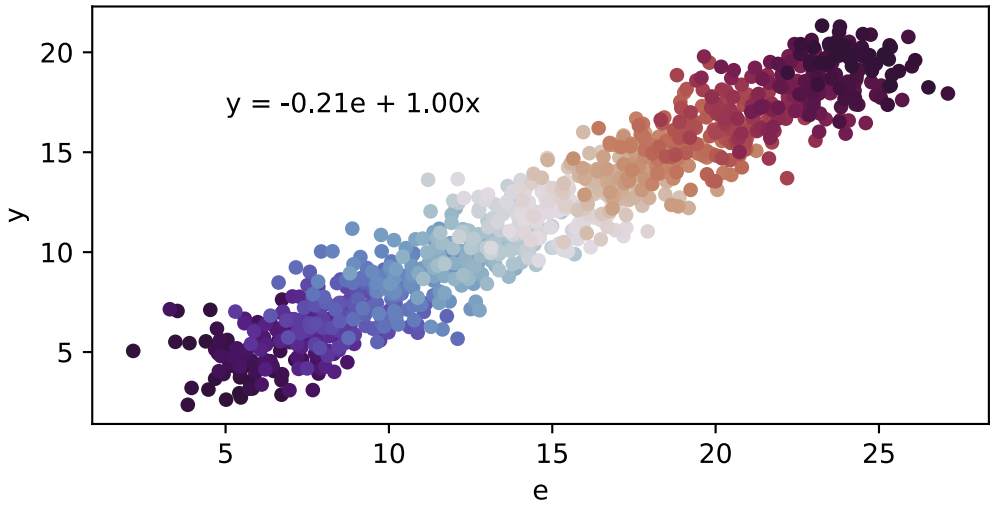
Simpson's paradox

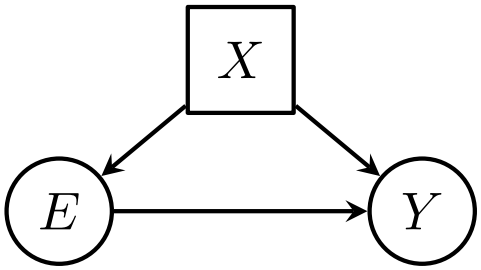




Simpson's paradox





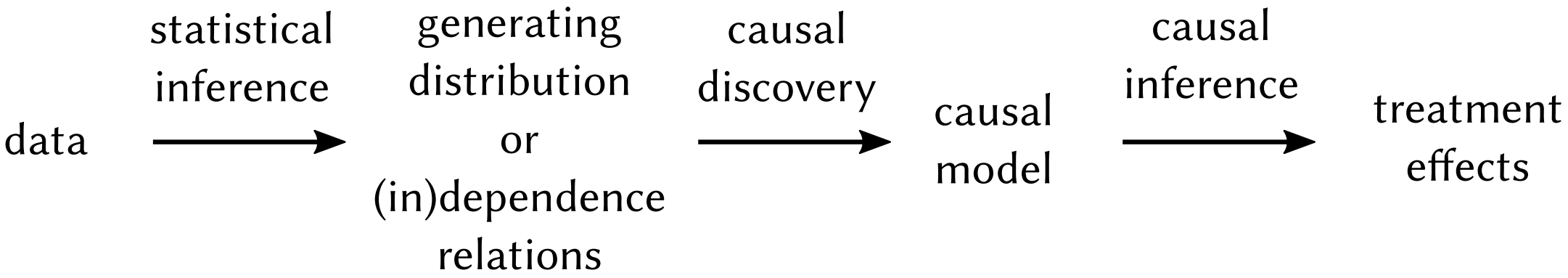


$$e = x + N(\mu_e, \sigma_e)$$
$$y = x - 0.2e + N(\mu_y, \sigma_y)$$

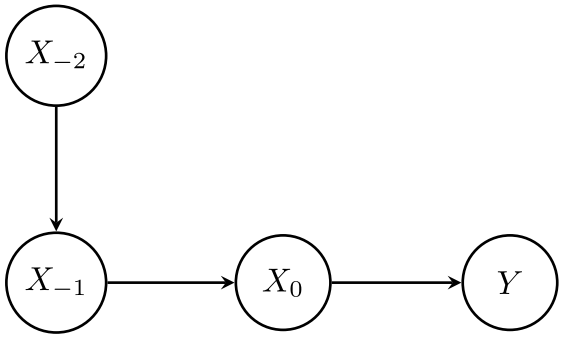
Causality: two tasks

Causal Discovery: identifying causal structure from observational data

Causal Inference: estimating effects of interventions from observational data and causal structure

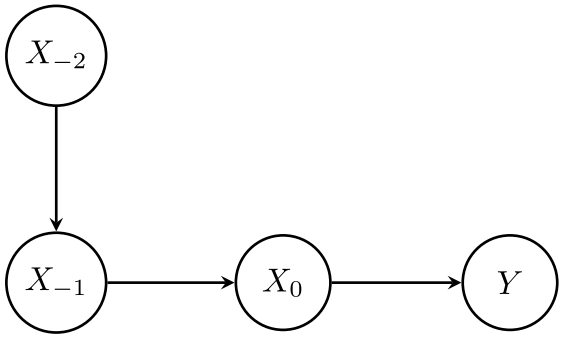


Causal Graphical Models



represent direct causal relations
(relative to variables in graph)

Causal Graphical Models



represent direct causal relations (relative to variables in graph)

Markov property: given X_0, Y is independent of X_{-1} X_{-2}

Causal Discovery: the basic idea

Markov assumption: every node in the graph is probabilistically independent of its non-descendents given its parents

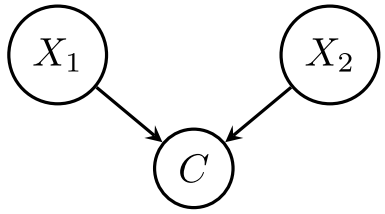
Causal faithfullness assumption: if a variable X is independent of Y given a conditioning set C in the probability distribution P(V), then X is d-separated from Y given C

structural properties $X \perp Y|C \Leftrightarrow X \perp\!\!\!\perp Y|C$ statistical properties

d-separation

d-connection: a path p between X and Y d-connects X and Y given a conditioning set S iff:

d-separation

collider: 

d-connection: a path p between X and Y d-connects X and Y given a conditioning set S iff:

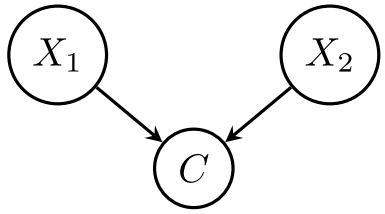
(i) all colliders on p are in S or have a descendent in S and

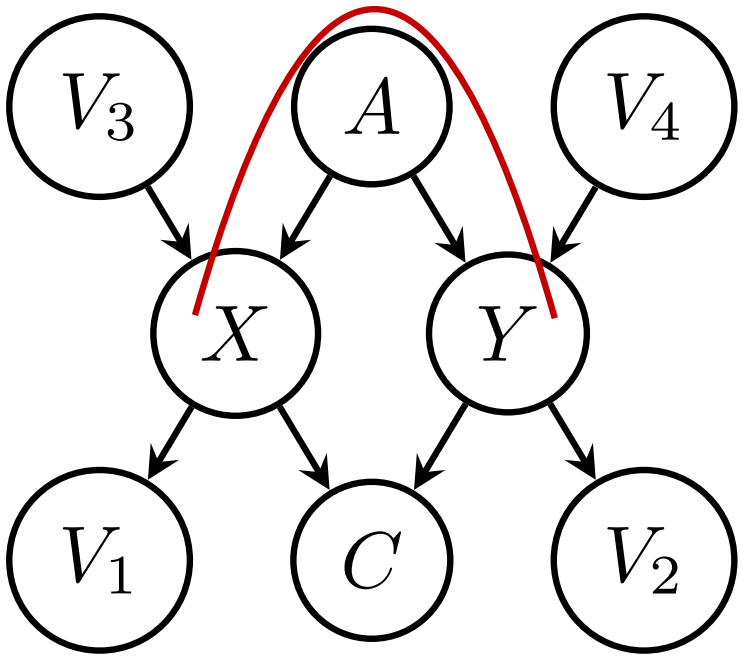
(ii) no non-colliders of p are in S

X and Y are d-separated iff there are no d-connecting paths between them

d-separation

collider:





d-connection: a path p between X and Y d-connects X and Y given a conditioning set S iff:

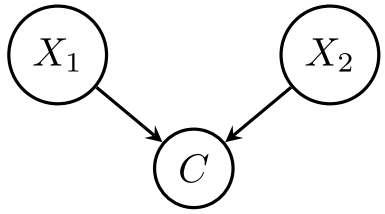
(i) all colliders on p are in S or have a descendent in S and

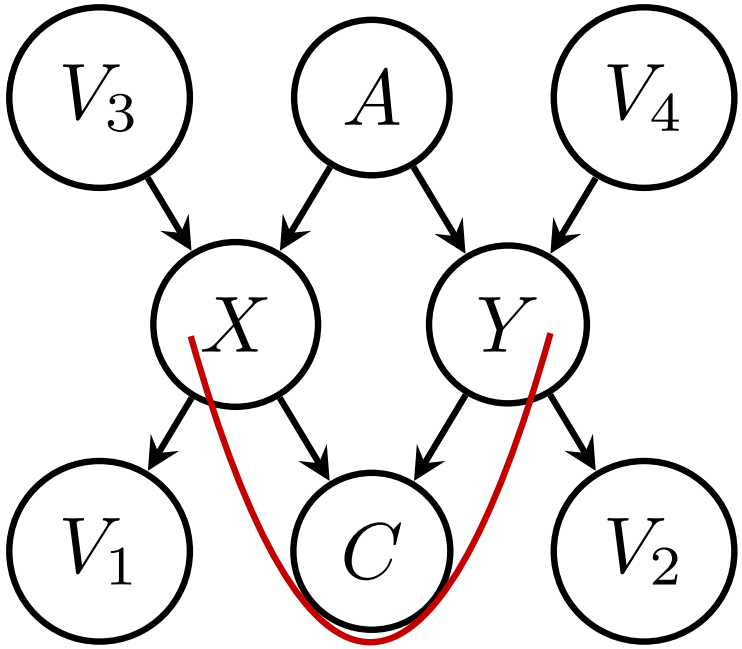
(ii) no non-colliders of p are in S

X and Y are d-separated iff there are no d-connecting paths between them

d-separation

collider:





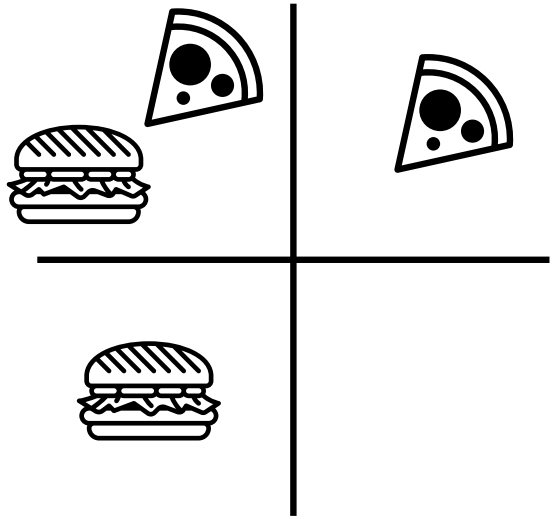
d-connection: a path p between X and Y d-connects X and Y given a conditioning set S iff:

(i) all colliders on p are in S or have a descendent in S and

(ii) no non-colliders of p are in S

X and Y are d-separated iff there are no d-connecting paths between them

Berkson's paradox



3	27
7	63

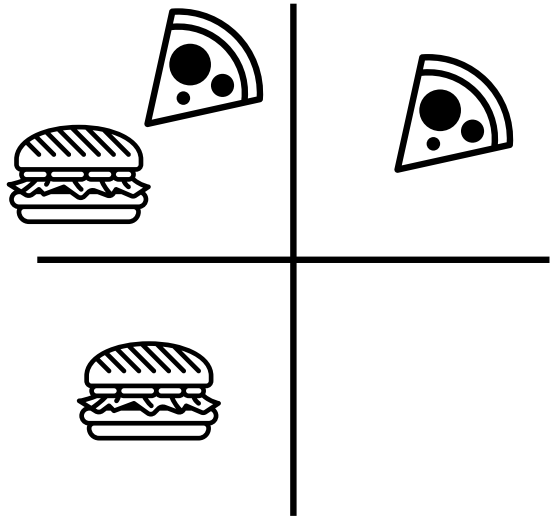
- 10% all places serve burgers
- 10% of pizza places serve burgers too

but if we only go where burgers or pizza are served?

- 27% of places serve burgers
- still only 10% of pizza places serve burgers too

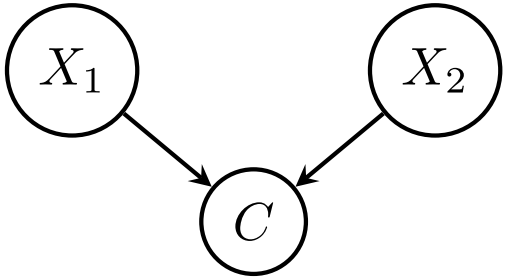
spurious negative correlation

Berkson's paradox



3	27
7	63

conditioning on a common effect of two independent causes makes them dependent



Causal Discovery: the basic idea

the basic assumptions provide some structural constraints given observational data, i.e. they rule out some structures

depending on additional assumptions and particular circumstances constraints can be stronger, up to leaving only one structure

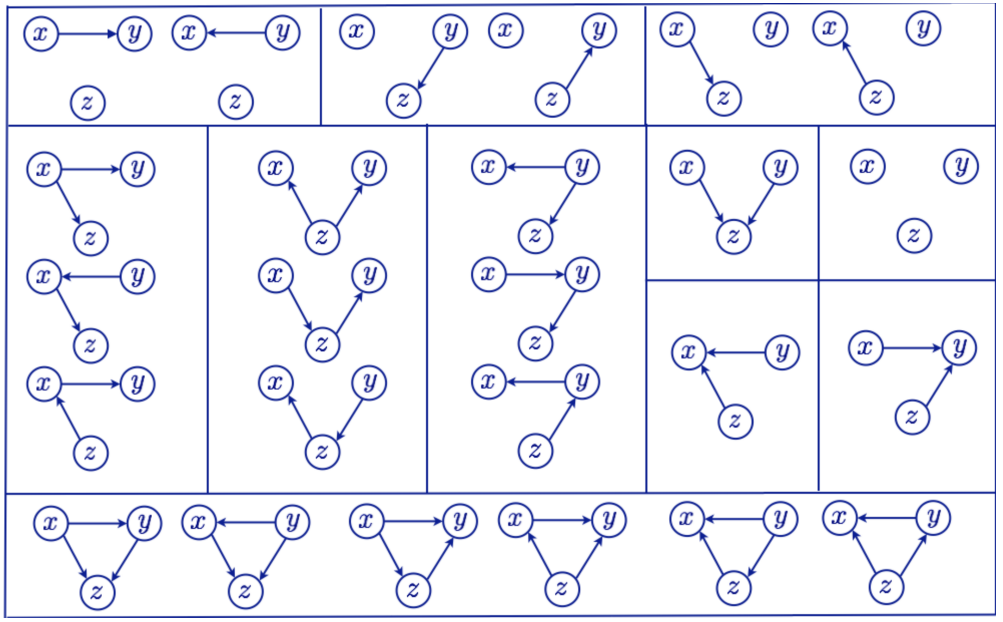
acyclicity + causal sufficiency

causal sufficiency: no unmeasured common causes of any pair of variables (no hidden confounders)

if two (Markov, faithful, acyclic, sufficient) structures have the same (conditional) (in)dependencies, then they have the same adjacencies and unshielded colliders

acyclicity + causal sufficiency

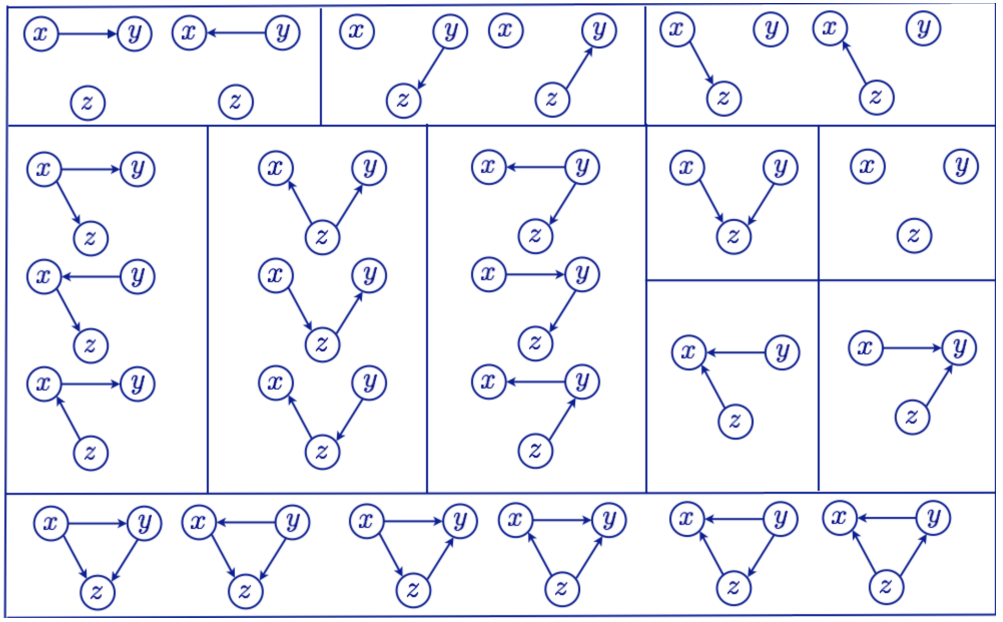
Markov
equivalence
classes



"Introduction to the Foundations of Causal Discovery" F. Eberhardt

acyclicity + causal sufficiency

Markov
equivalence
classes



For linear Gaussian and for multinomial causal relations, an algorithm that identifies the Markov equivalence class is complete
(i.e. it extracts all available information about the underlying causal structure)

"Introduction to the Foundations of Causal Discovery" F. Eberhardt

non-linear / non-gaussian

causal rel.	noise	identifiability
linear	gaussian	Markov-equiv.
	non-gaussian	uniquely*
non-linear	gaussian	uniquely

other topics

- weakening assumptions (allowing cycles, latent confounders, etc.)
- discrete models
- use of background & experimental knowledge
- dynamics and time series
- variable construction
- relational data

Causal Inference

Causal Inference: estimating **effects** of **interventions** from observational data and a causal model

Interventions

usually understood as setting value of a chosen variable

more subtle approches: adding an icoming node (influencing value, but not determining it); modification of outcoming causal functions

important to consider if the actual intervention will match the modelling assumptions

Effect

usually, the population Average Treatment Effect (ATE)

for particular group: Conditional ATE (CATE)

sometimes its is only possible to estimate effect for particular group, not population (sample, treated or controls)

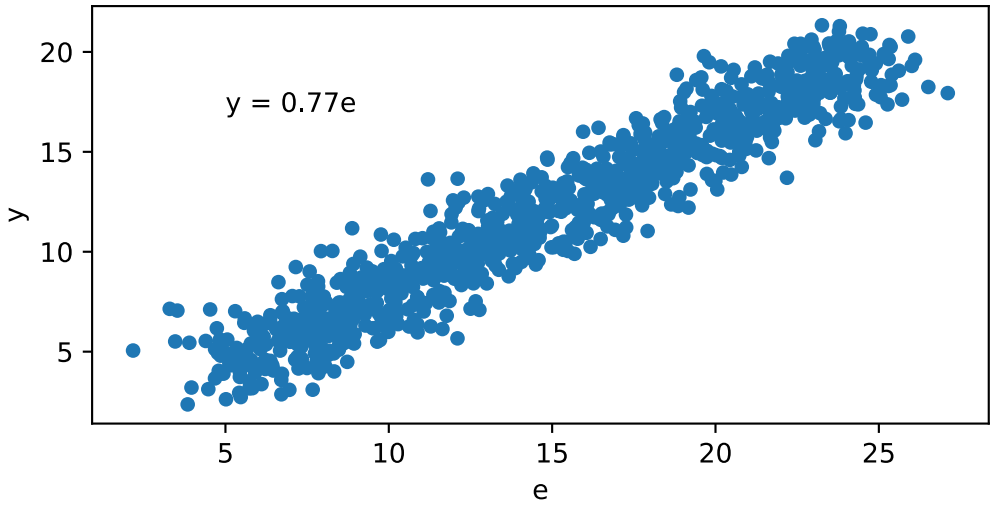
How to estimate?

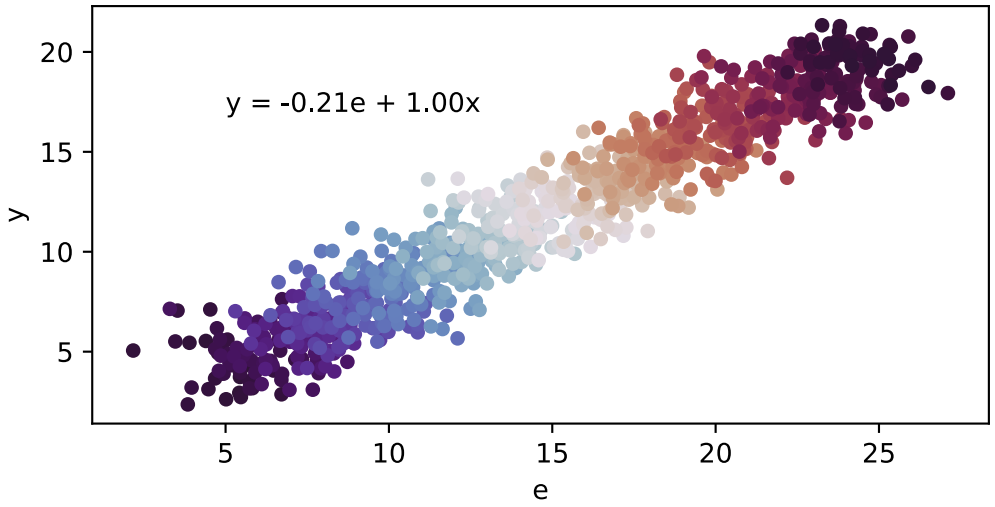
Backdoor

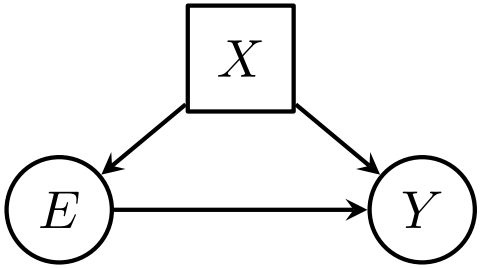
Frontdoor

Instrumental Variables

Backdoor method

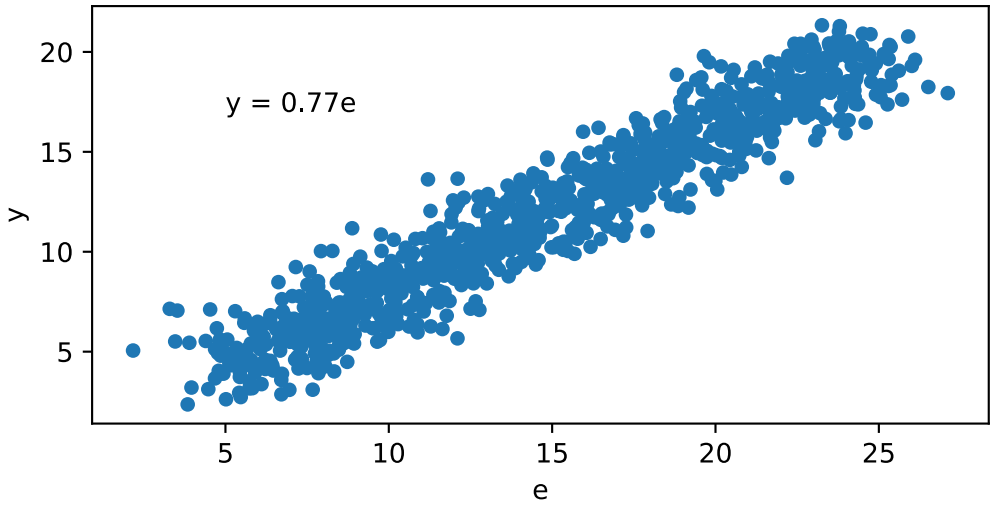


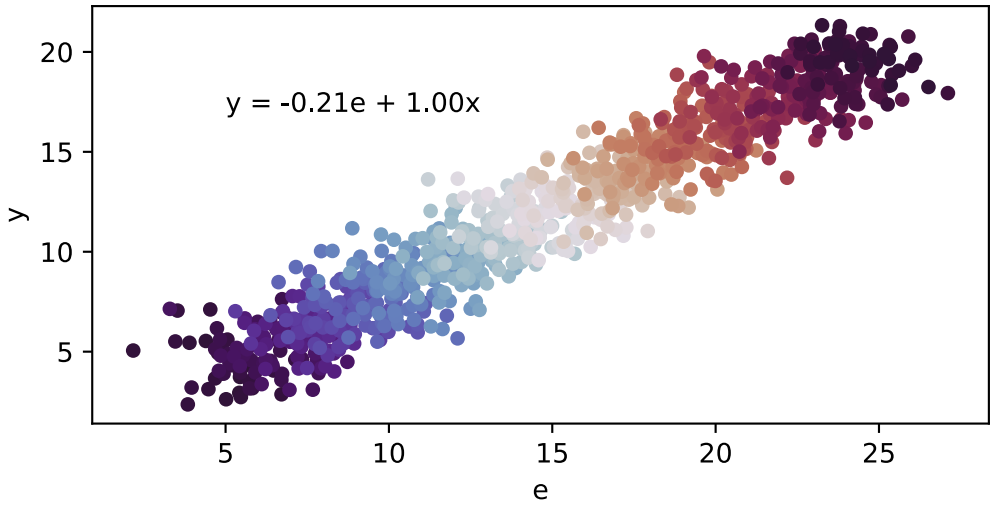


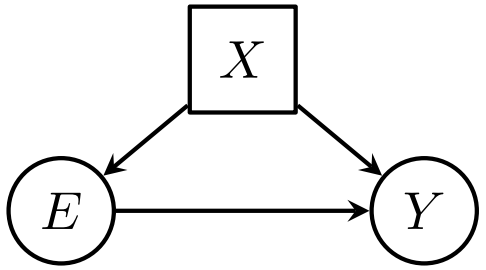


$$e = x + N(\mu_e, \sigma_e)$$
$$y = x - 0.2e + N(\mu_y, \sigma_y)$$

Backdoor method







$$e = x + N(\mu_e, \sigma_e)$$
$$y = x - 0.2e + N(\mu_y, \sigma_y)$$

But what variables to adjust for?

Backdoor method

Why not just adjust for everything?

Backdoor method

Why not just adjust for everything?

- spurious adjustment can lead to bias (e.g. collider bias)

Backdoor method

Why not just adjust for everything?

- spurious adjustment can lead to bias (e.g. collider bias)

- some adjustments are unnecessary (it is good to reduce dimensionality)

Backdoor method

Why not just adjust for everything?

- spurious adjustment can lead to bias (e.g. collider bias)

- some adjustments are unnecessary (it is good to reduce dimensionality)

- how to determine the adjustment set for complex graphs?

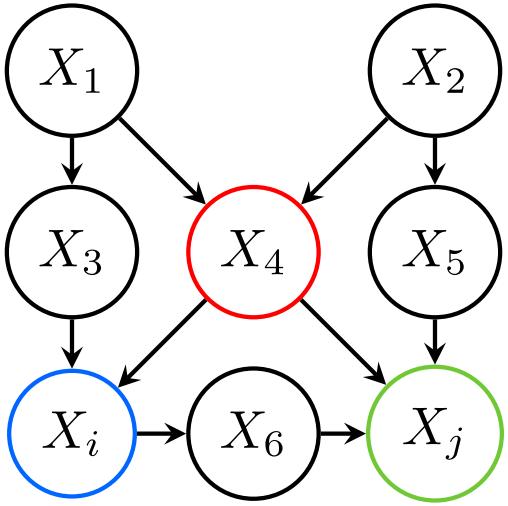
Backdoor method

sets that satisfy the backdoor criterion are sufficient adjustment sets

backdoor path starts with an arrow into X_i

a set S satisfies the backdoor criterion if:
(i) S blocks all backdoor paths
- there is a non-collider on the path that is in S
- there is a collider on the path such that neither this collider nor any of its descentants are in S

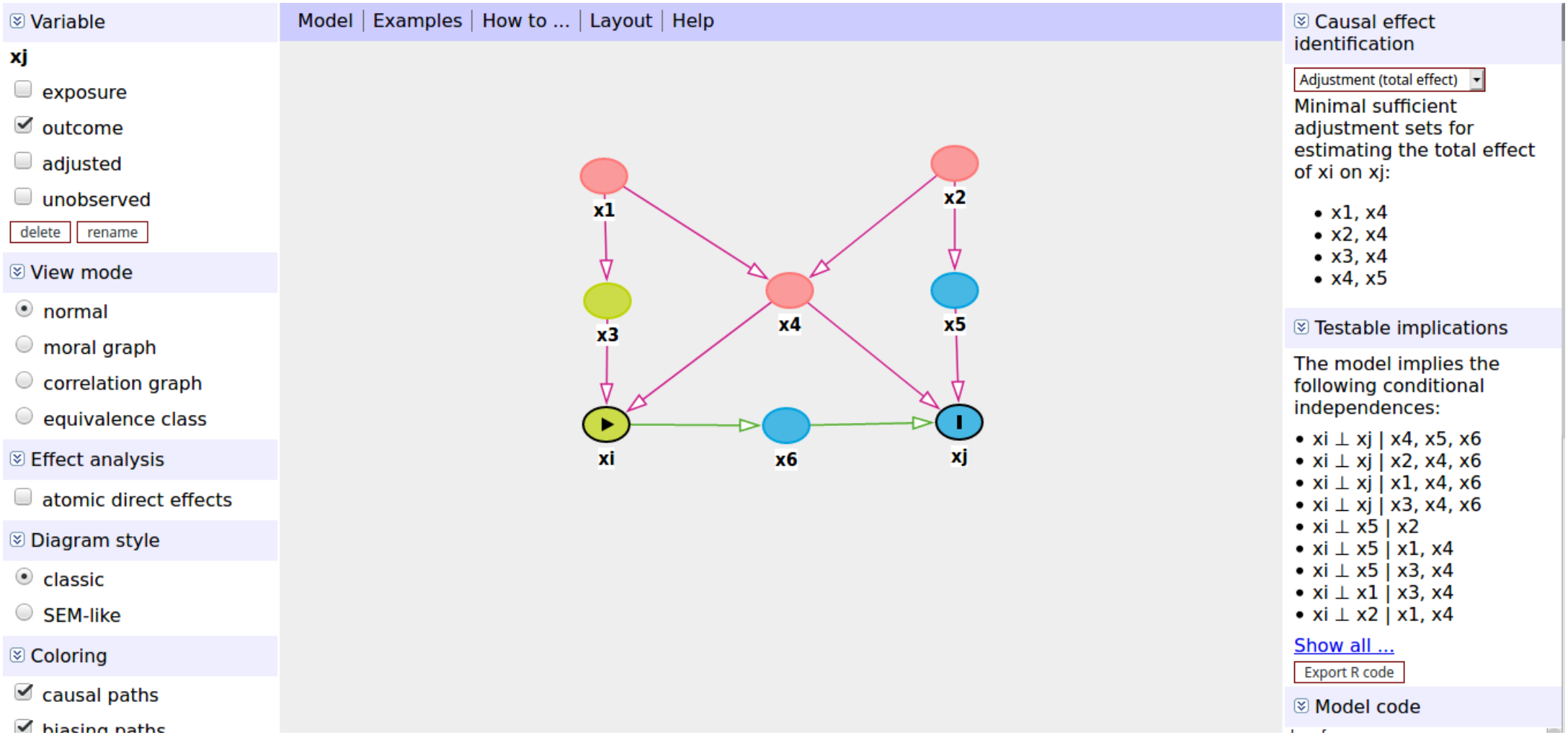
(ii) S does not contain any descendants of X_i



$$S_1 = \{X_1, X_4\}$$
$$S_2 = \{X_2, X_4\}$$
$$S_3 = \{X_3, X_4\}$$
$$S_4 = \{X_4, X_5\}$$

Backdoor method

http://dagitty.net/



Backdoor method

adjustment methods:

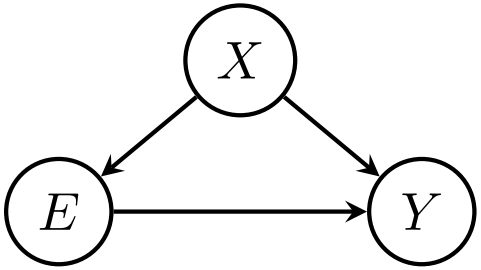
- plain "preprocessing"
  - stratification
  - matching

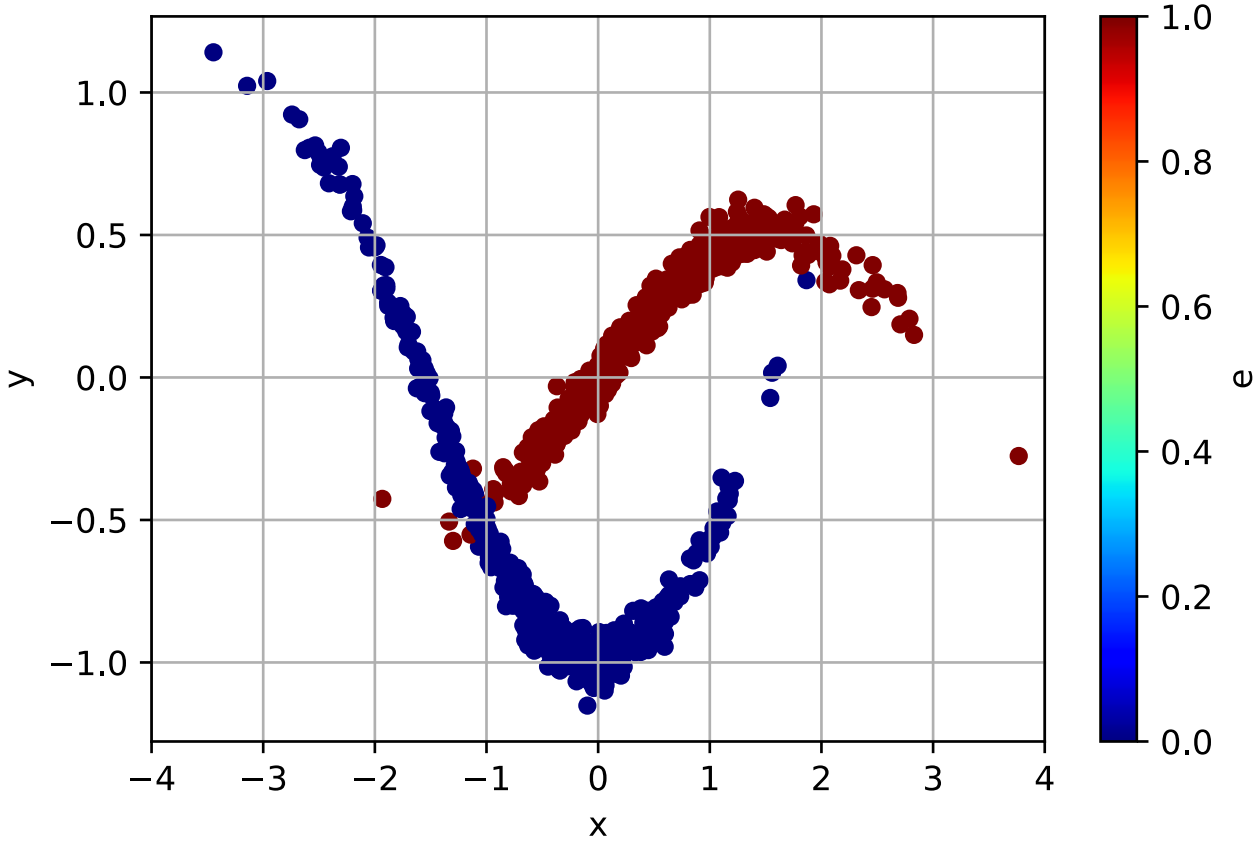
- modelling assignment mechanism

- modelling response surface

Backdoor method

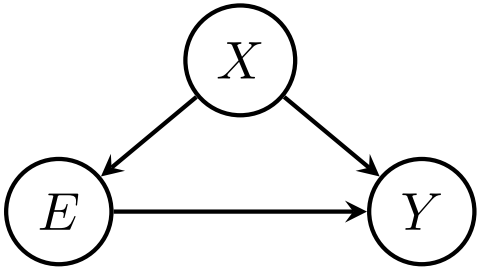
stratification & matching

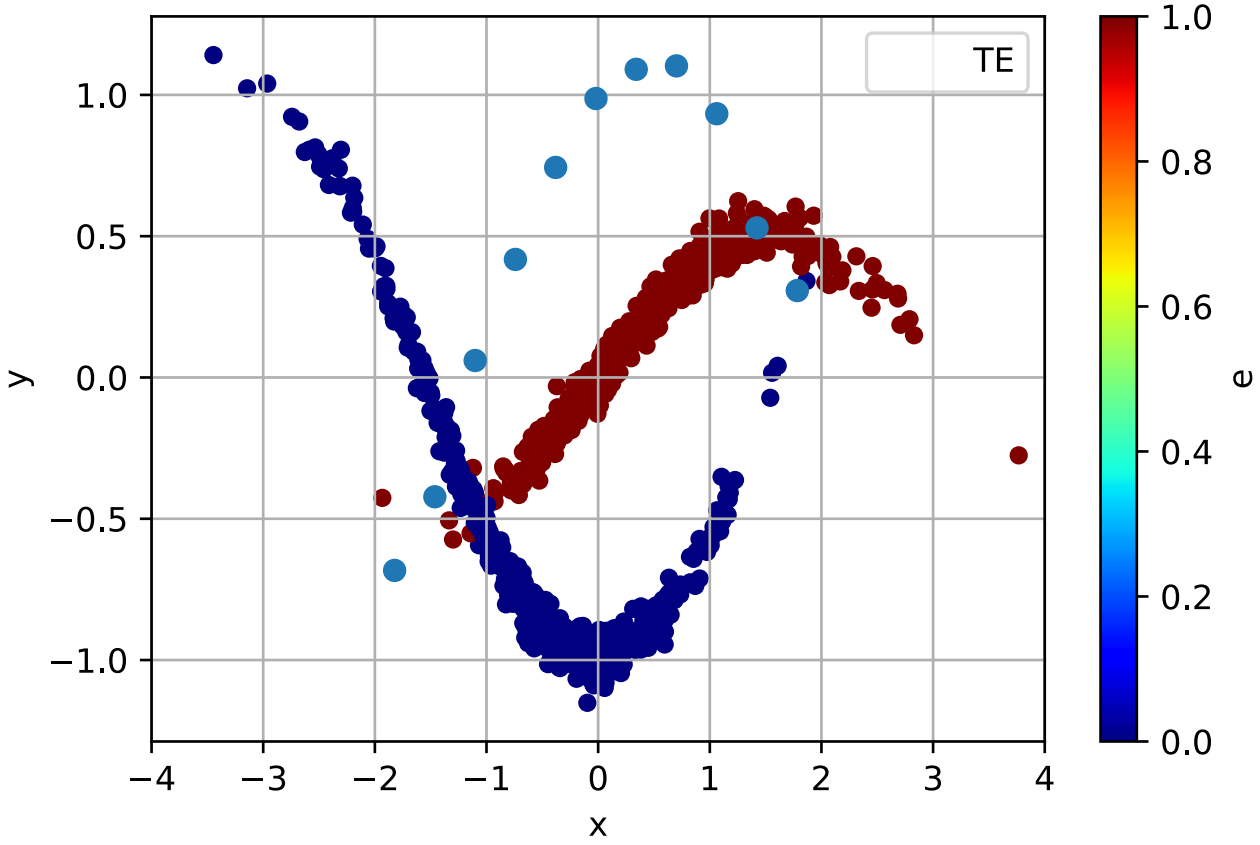




Backdoor method

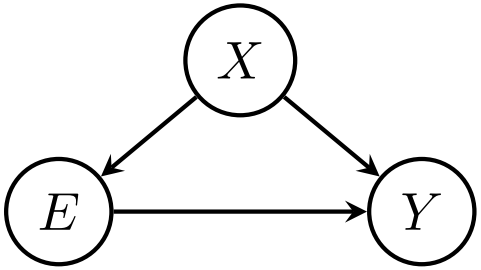
stratification & matching

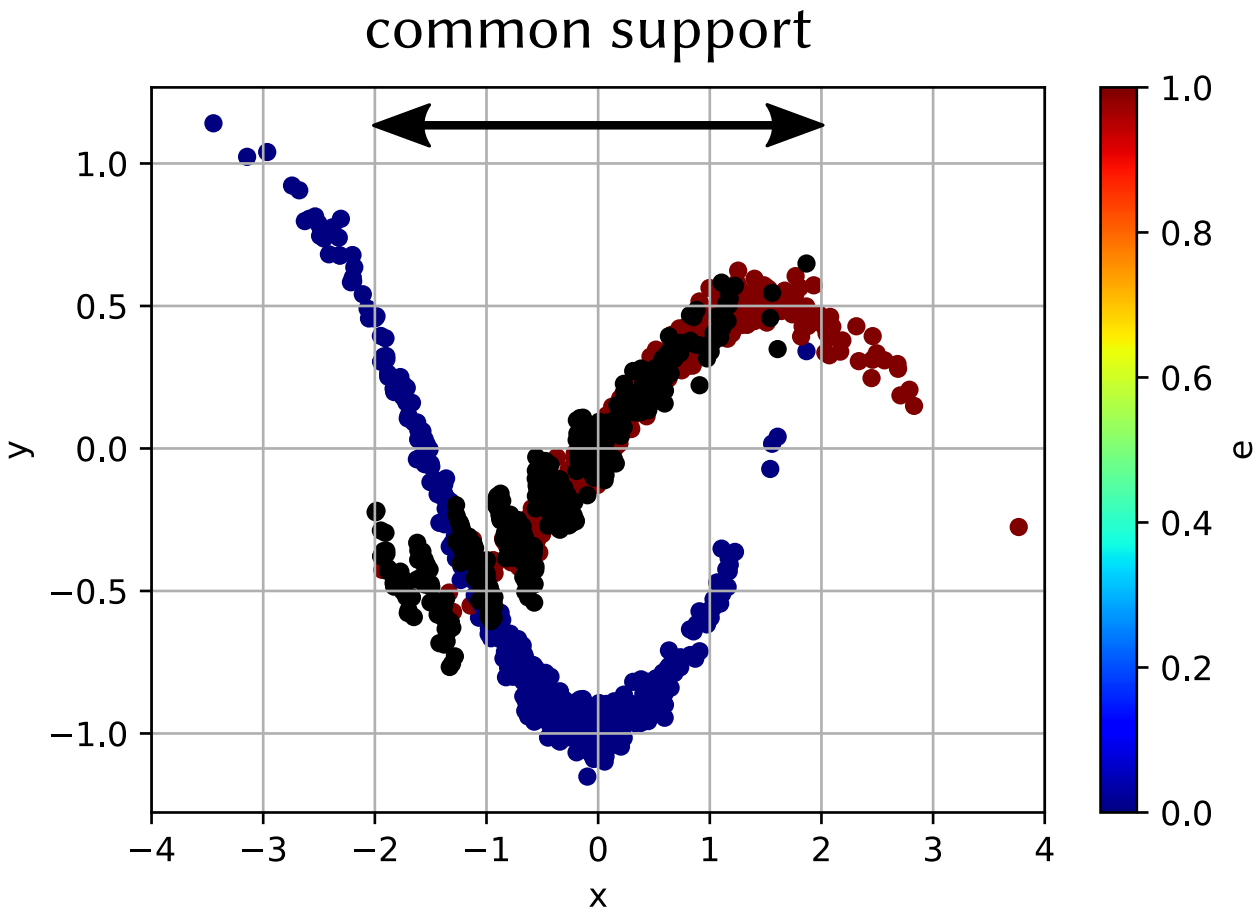




Backdoor method

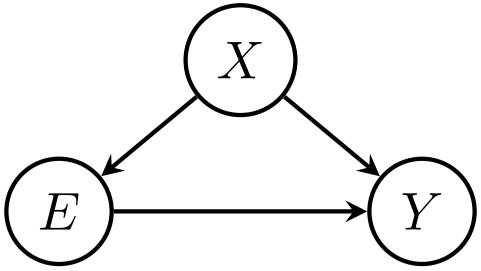
stratification & matching

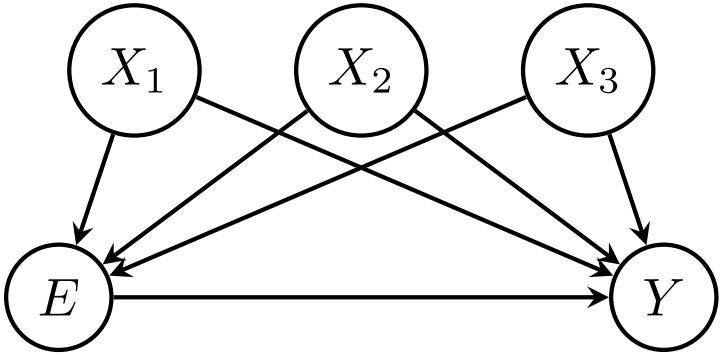




Backdoor method

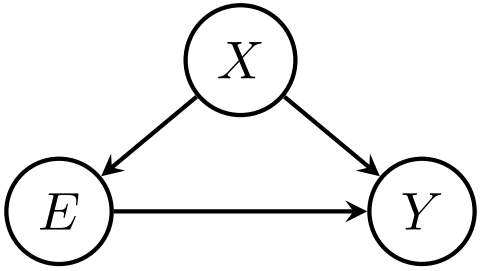
modelling the assignment mechanism

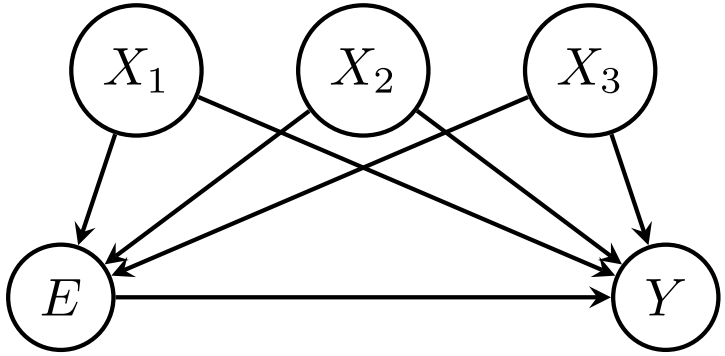




Backdoor method

modelling the assignment mechanism

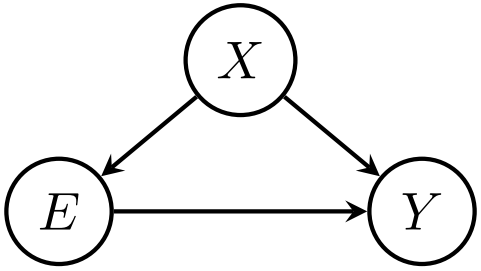


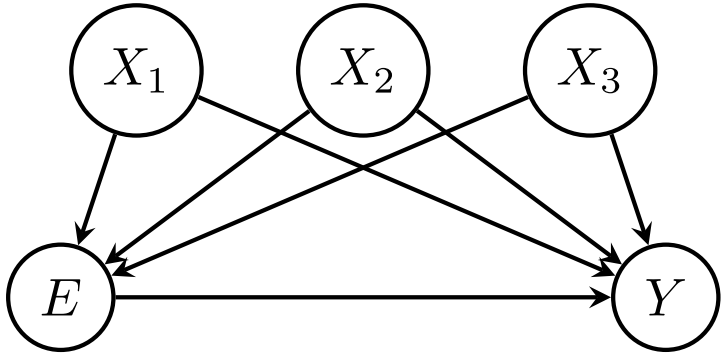


- high number of confounders makes stratification and matching difficult (high dimensionality)
- solution: replace confounders with a probability of treatment (Propensity Score)

Backdoor method

modelling the assignment mechanism





- high number of confounders makes stratification and matching difficult (high dimensionality)
- solution: replace confounders with a probability of treatment (Propensity Score)
- PS used then in stratification or matching

Backdoor method

modelling the response surface

Backdoor method

modelling the response surface

- i.e. just train some estimator $y = f(e, x_0, x_1, ..., x_n)$

Backdoor method
modelling the response surface

- i.e. just train some estimator $y = f(e, x_0, x_1, ..., x_n)$

- or using propensity score $y = f(e, PS)$

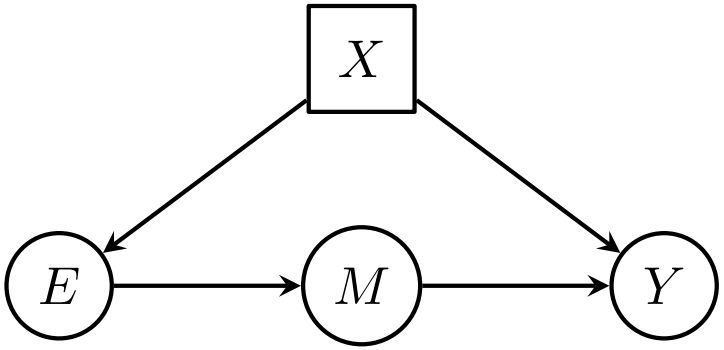
Backdoor method
modelling the response surface

- i.e. just train some estimator $y = f(e, x_0, x_1, ..., x_n)$

- or using propensity score $y = f(e, PS)$

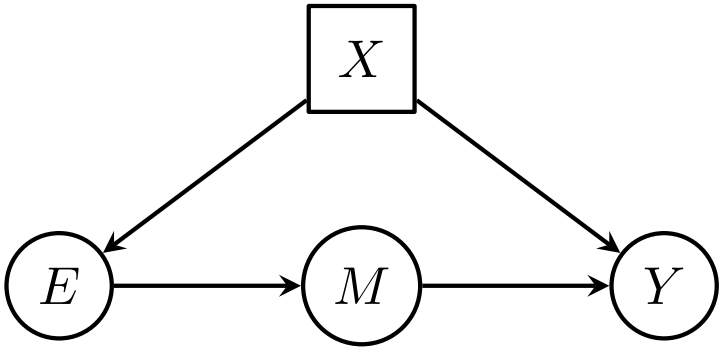
- two of the most popular tools are Bayesian Additive Regression Trees (BART) and TMLE (Targeted Maximum Likelihood Estimation)

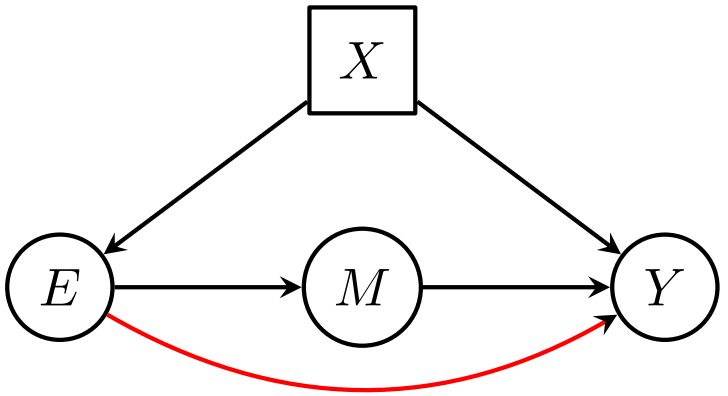
Frontdoor method



$$y = f_{m \to y}(f_{e \to m}(e))$$

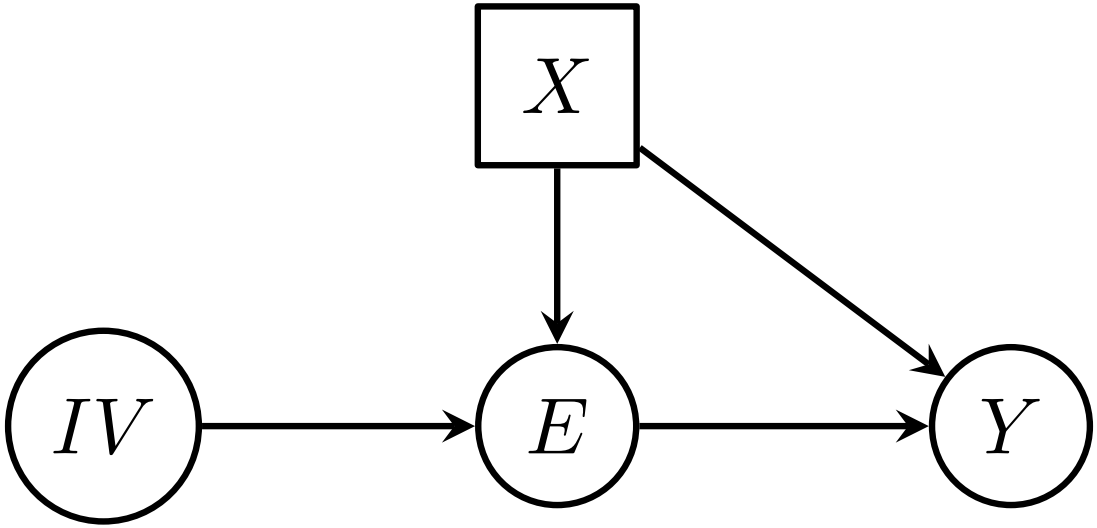
Frontdoor method





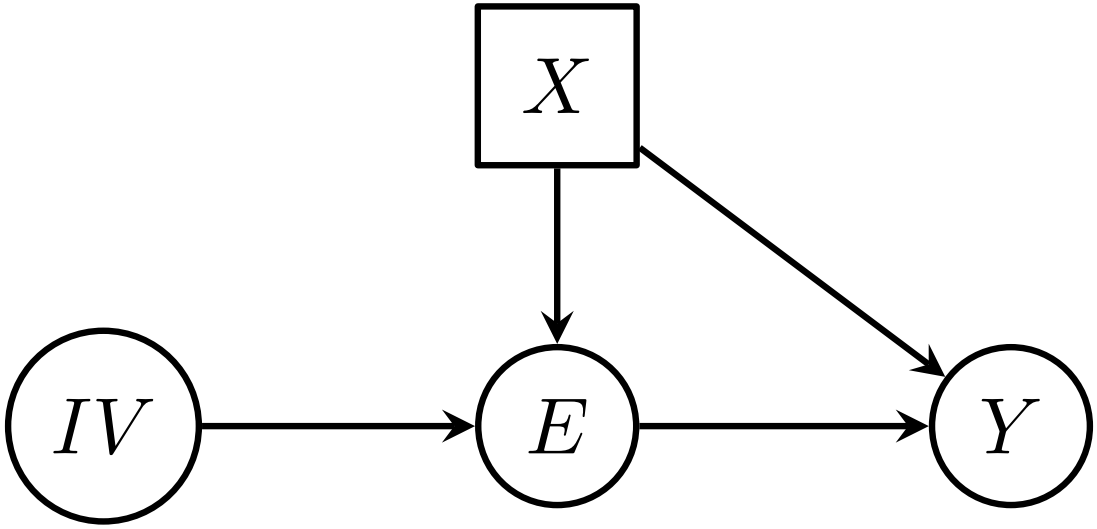
$$y = f_{m \to y}(f_{e \to m}(e))$$

Instrumental Variable method



$$y = f_{e \to y}(f_{iv \to e}(iv))$$

Instrumental Variable method



$$y = f_{e \to y}(f_{iv \to e}(iv))$$

the relative strength of $IV \to E$ is crucial

Other topics

- ATT, ACT, ITE
- do-calculus
- Rubin causal model (potential outcomes)
- Inverse Probability of Treatment Weighting
- double robustness
- time series data
- multiple treatments
- weak treatments
- continous treatment
- datasets
- ...

Bibliography

intro:
- http://www.degeneratestate.org/posts/2018/Mar/24/causal-inference-with-python-part-1-potential-outcomes/
- https://www.inference.vc/untitled/
- "Introduction to the Foundations of Causal Discovery" F. Eberhardt

overview:
- "Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition" Dorie et al.
- "Elements of Causal Inference" Peters et al.

TMLE:
- "Targeted learning: causal inference for observational and experimental data" Van der Laan & Rose