Detecting hidden confounding in observational data by combining heterogeneous environments

4 Oct 2022, Eindhoven

Rickard Karlsson
PhD student

Jesse Krijthe
Assistant professor

Pattern Recognition and Bioinformatics group, TU Delft

TU Delft

Causal inference from observational data

Goal Estimate the causal effect of doing action T on outcome Y

Example

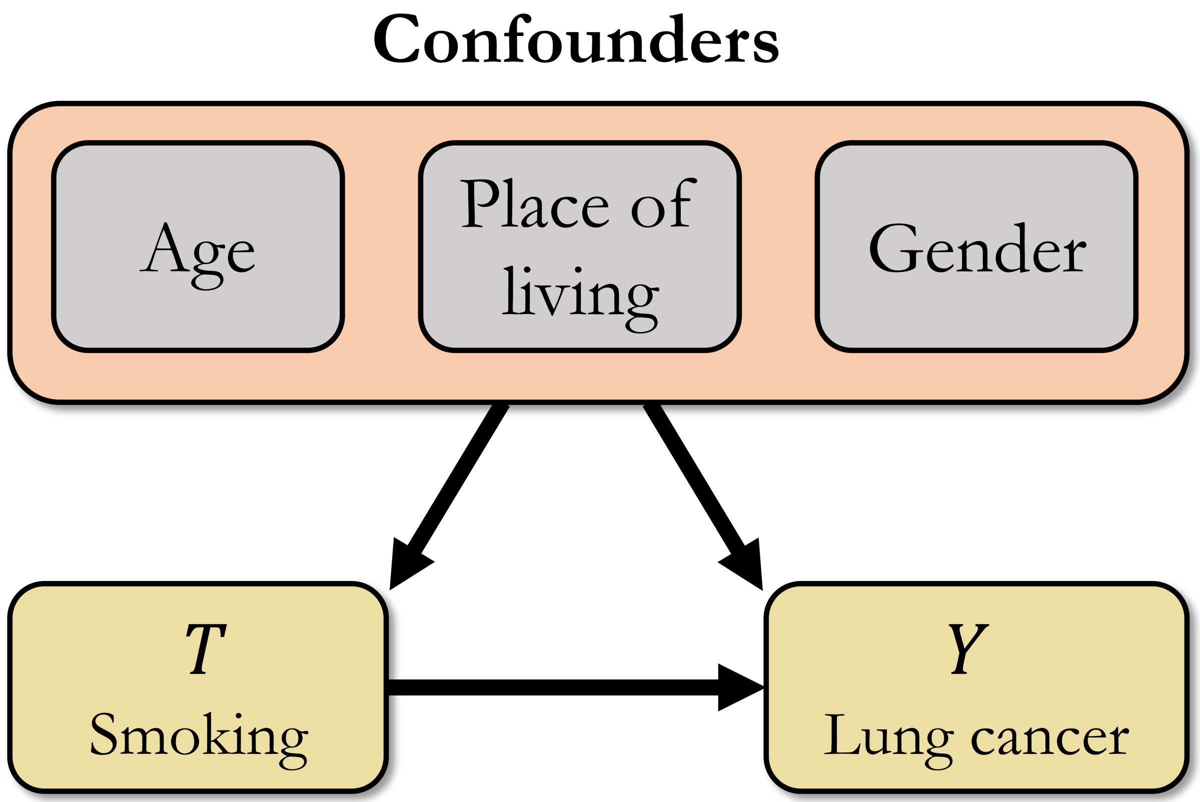
- T : Smoking tobacco for five years
- Y : Risk of having lung cancer

T and Y can be correlated, but *association does not imply causation*.

Causal inference from observational data

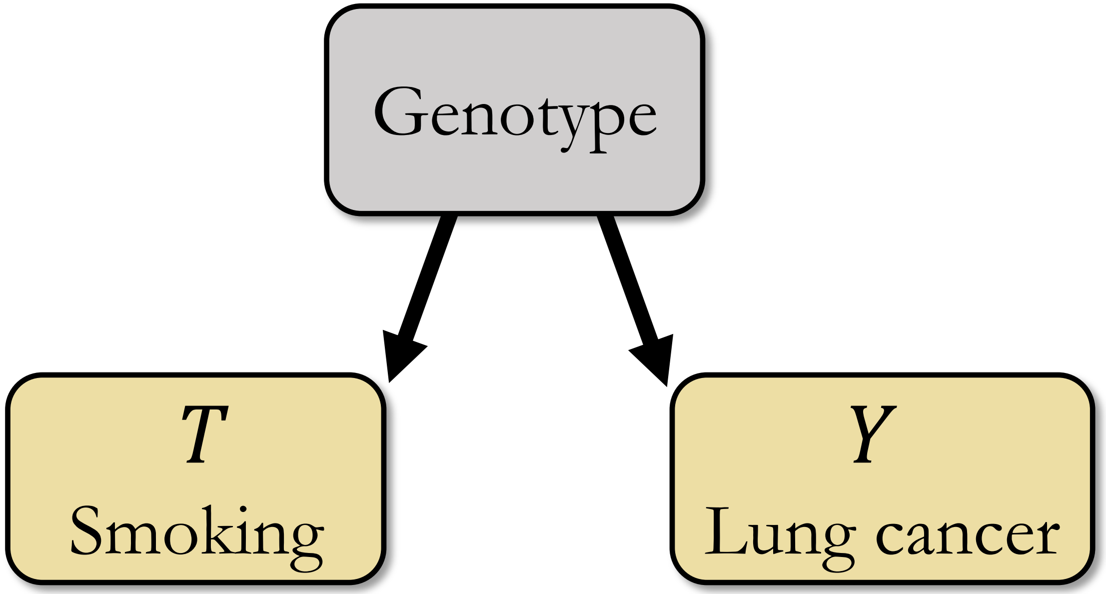
Main assumption

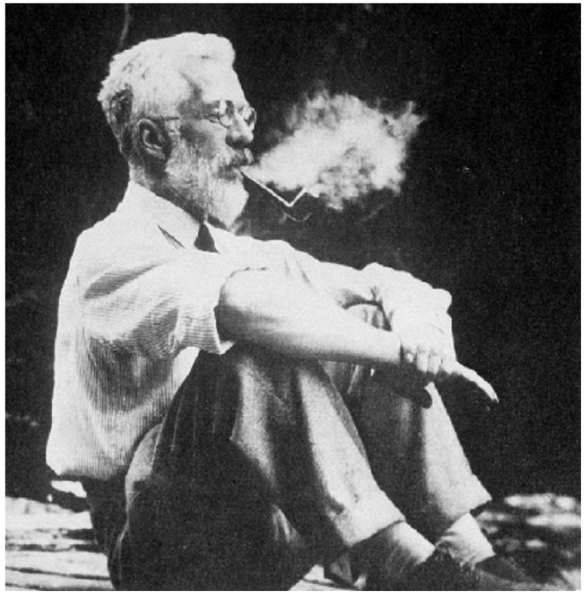
We have observed all relevant confounders in the study population.



Cancer and Smoking (Fisher, 1958)

Fisher argued there exists a hidden confounder between smoking and lung cancer





R.A. Fisher smoking a pipe, 1956. (Source: M. Parascandola)

Confounding is a main reason for why association $\neq$ causation

This talk

In general, we can not know if we have observed all confounders.

But if we have data from multiple environments, **we show ways to statistically test the presence of unobserved confounders**.

An environment can be e.g. data from different hospitals or time periods.

Preliminaries

Causal graphical models

A causal graphical model M for variables $\boldsymbol{X} = (X_1, X_2, \dots, X_d)$ consists of

1. a directed acyclic graph G with vertices $\boldsymbol{X}$ and
   $X_i \to X_j$ iff X_i is a *direct cause* to X_j
2. a joint distribution $P_{\boldsymbol{X}}$ over the variables

For the given graph, we have the *causal factorization*

$$P_{\boldsymbol{X}}(\boldsymbol{X}) = \prod_{i=1}^{d} \underbrace{P_{\boldsymbol{X}}(\, X_i \mid Pa(X_i)\,)}_{\text{causal mechanism}}$$

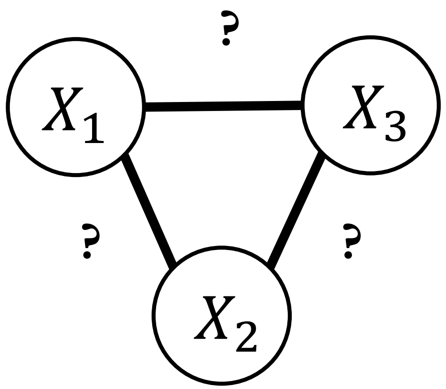
Learning causal structure from data

The structure of the graph G implies certain conditional independencies in $P_{\boldsymbol{X}}$ [1].

Example

We have (X_1, X_2, X_3):

$$X_1 \perp_{P_{\boldsymbol{X}}} X_3 \qquad \Leftrightarrow$$



[1] Assuming G and $P_{\boldsymbol{X}}$ fulfill the faithfulness and causal Markov properties.

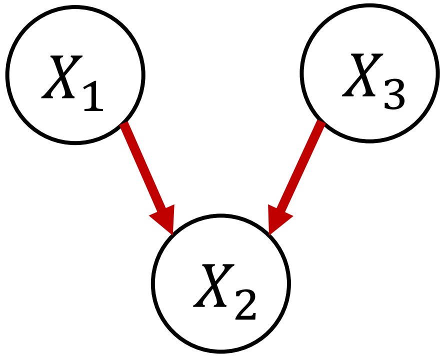
Learning causal structure from data

The structure of the graph G implies certain conditional independencies in $P_{\boldsymbol{X}}$ [1].

Example

We have (X_1, X_2, X_3):

$$X_1 \perp_{P_{\boldsymbol{X}}} X_3 \qquad \Longleftrightarrow$$



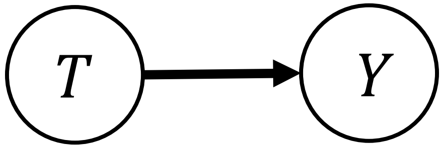
[1] Assuming G and $P_{\boldsymbol{X}}$ fulfill the faithfulness and causal Markov properties.

Reichenbach’s common cause principle

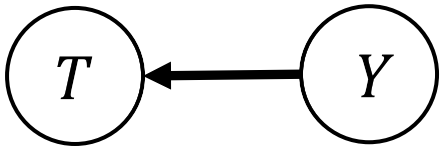
But often we can not learn the exact structure, even for two variables.

Let variables T, Y be correlated, then either of the following can be true:

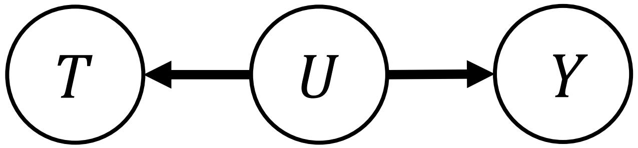
i)



ii)

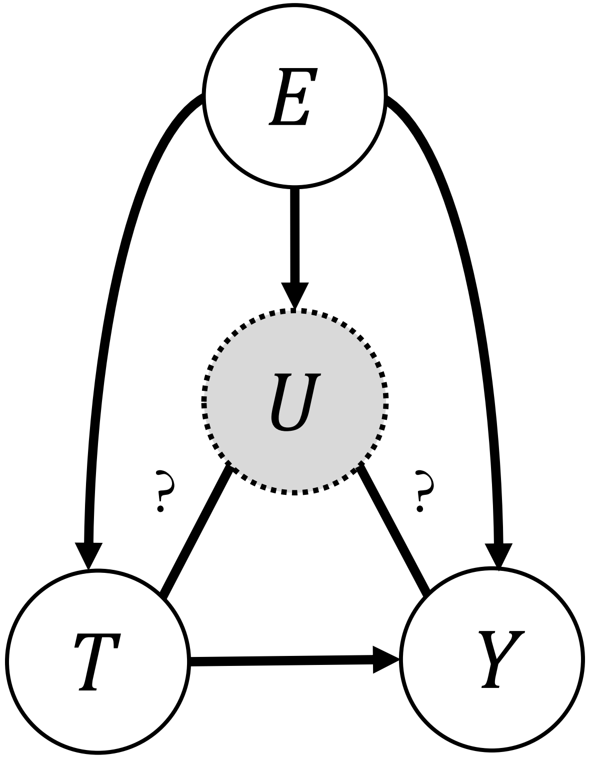


iii) There exists a latent U s.t.



Hans Reichenbach. *The Direction of Time*. Dover Publications, 1956.

Problem statement

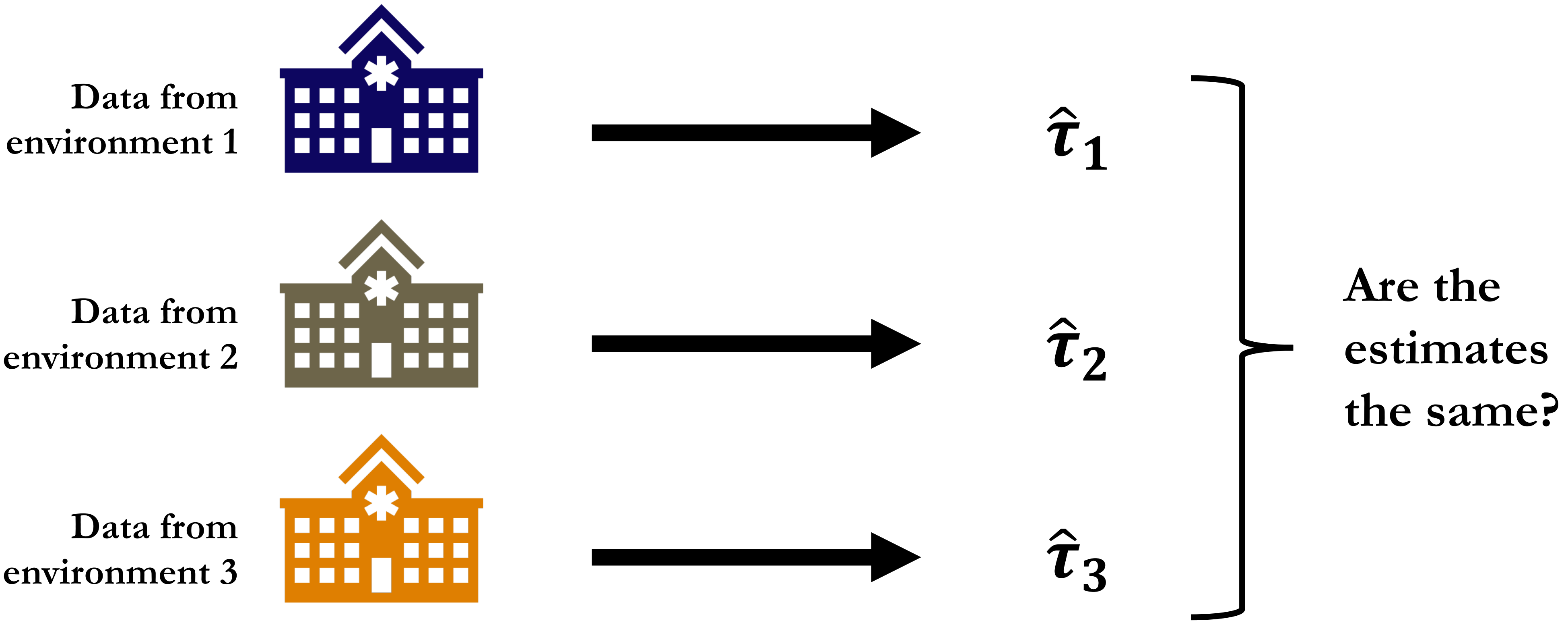


We observe treatment T and outcome Y in different environments E.

All environments share the same unknown causal structure, but $P(T, Y \mid E)$ may change for different enviroments.

Under what conditions can we detect the presence of a hidden confounder $\boldsymbol{U}$?

A first "naïve" approach to check for hidden confounding

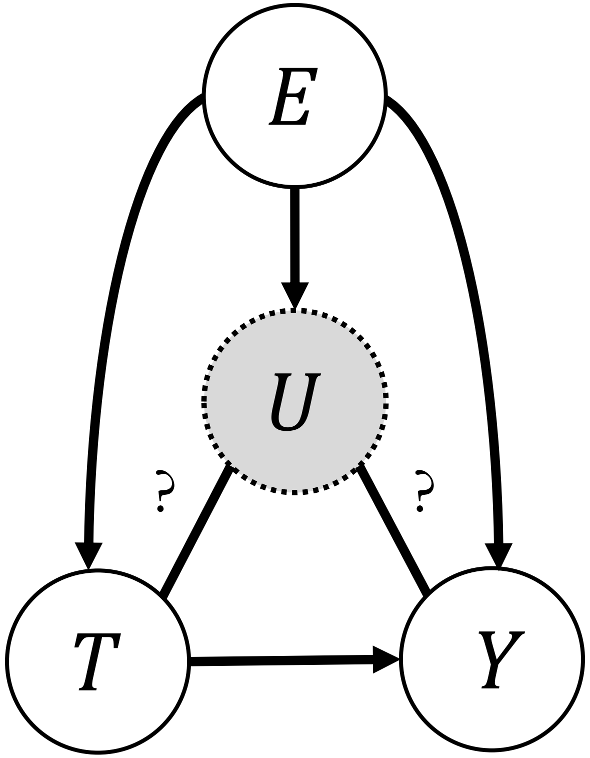


Main assumption: independent causal mechanisms

Let $P(\cdot \mid E) = P_E(\cdot)$

We have the causal factorization

$$P_E(T, Y, U) = \underbrace{P_E(Y \mid Pa(Y))P_E(T \mid Pa(T))P_E(U \mid Pa(U))}_{\text{conditional probabilities (causal mechanisms) vary independently across environments}}$$



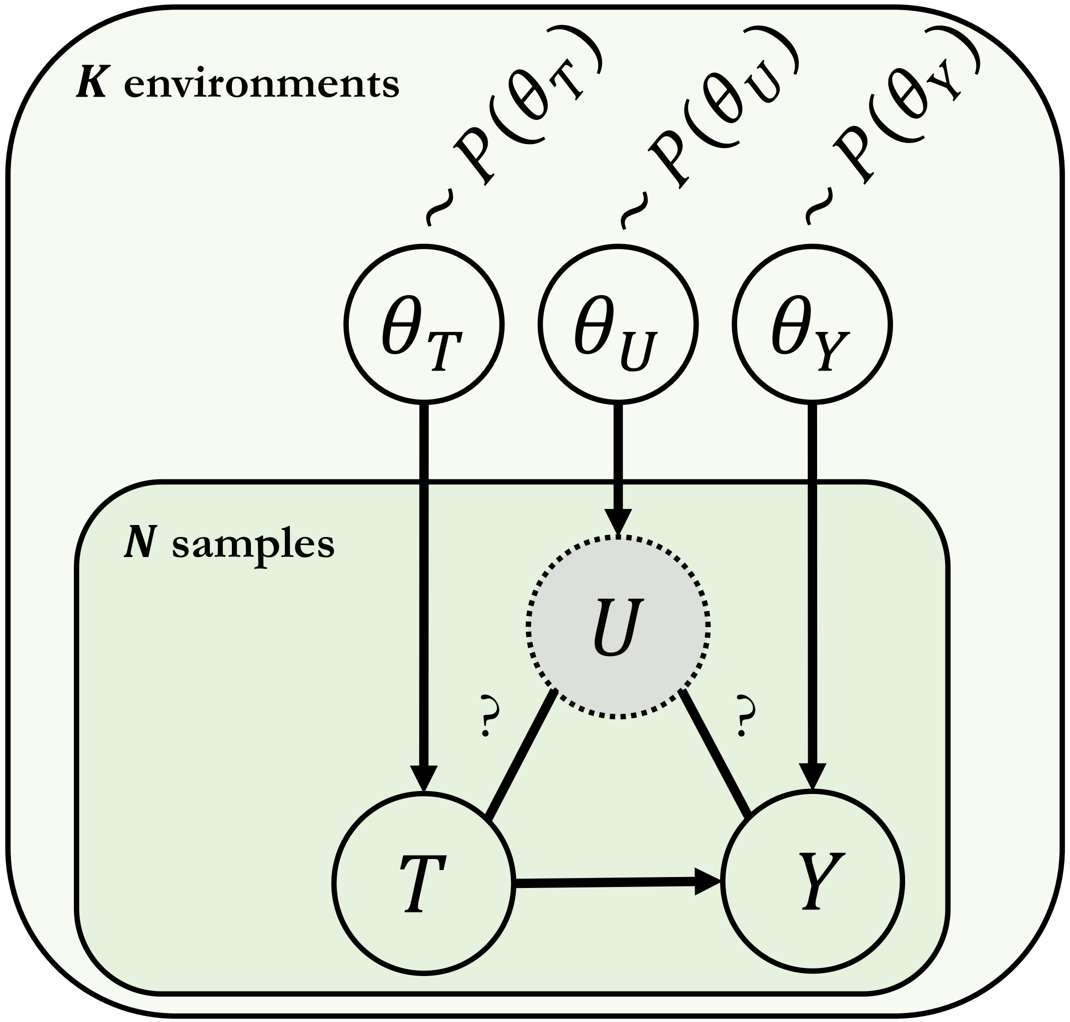
Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 1st edition, 2017.

Main assumption: independent causal mechanisms

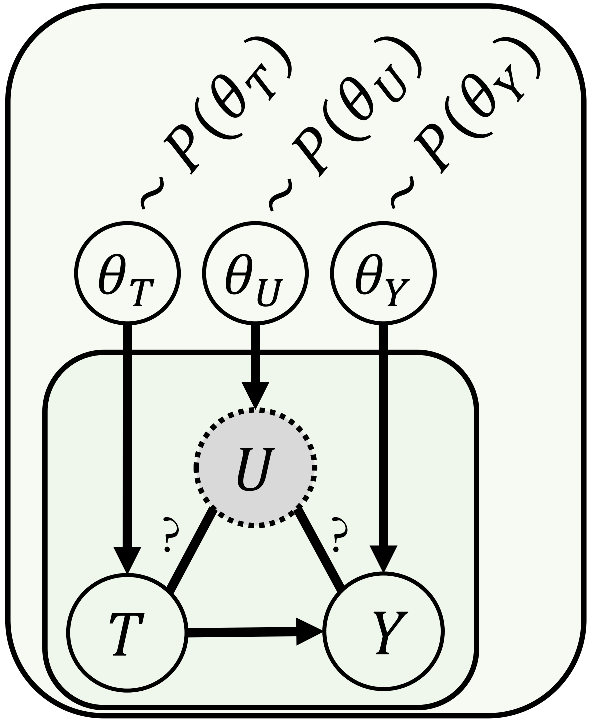
$\theta_T, \theta_Y, \theta_U$: causal mechanisms

Data-generating process

1. An environment is sampled from $P(\theta_T)$, $P(\theta_U)$ and $P(\theta_Y)$
2. In each environment, sample data from $P_{\theta_T,\theta_U,\theta_Y}(T, Y, U)$



Testable implications of hidden confounding



Consider the random sample variables

$$T_i, Y_i \sim P(T_i, Y_i) = \underbrace{\int P_{\theta_T,\theta_Y,\theta_U}(T_i, Y_i) dP(\theta_T) dP(\theta_Y)\, dP(\theta_U)}_{\text{the distribution of "sample } i\text{" marginalizing out the environments}}$$

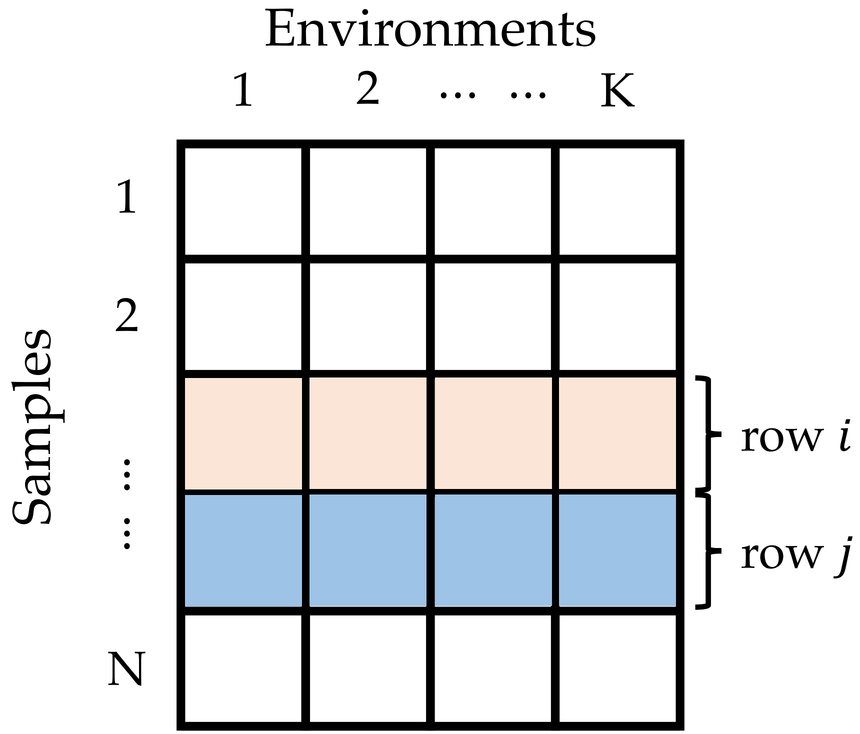
Theorem (informal)
Assuming our data-generating process with independent causal mechanisms, we have:

$T_j \perp Y_i \mid T_i$ for $i \neq j$ $\iff$ There can not exist a confounder U between T and Y

Testing confounding from data

How do we test $\boldsymbol{T_j \perp Y_i \mid T_i}$?

We sample from $P(T_i, Y_i)$ by selecting data from row i, and same for $P(T_j)$ with a different row j.

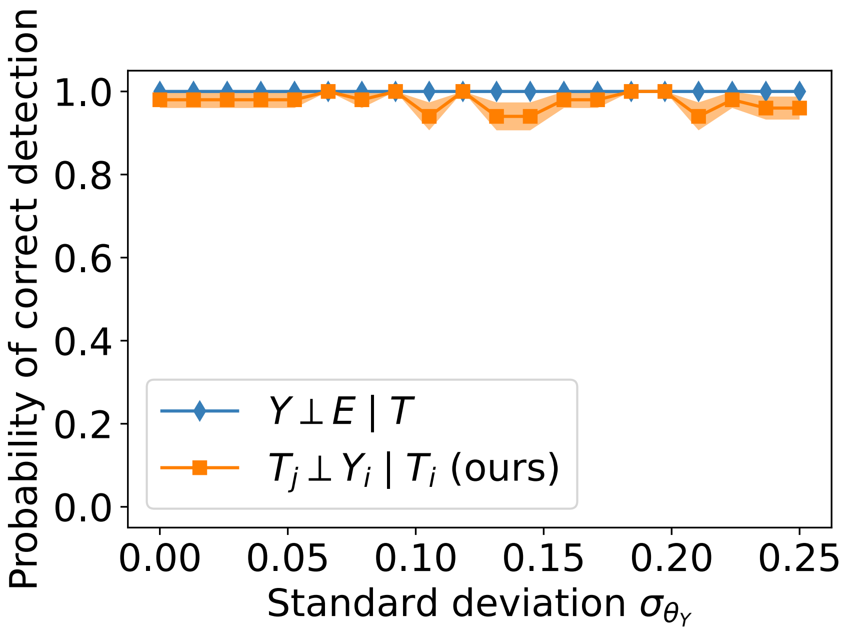


Challenges

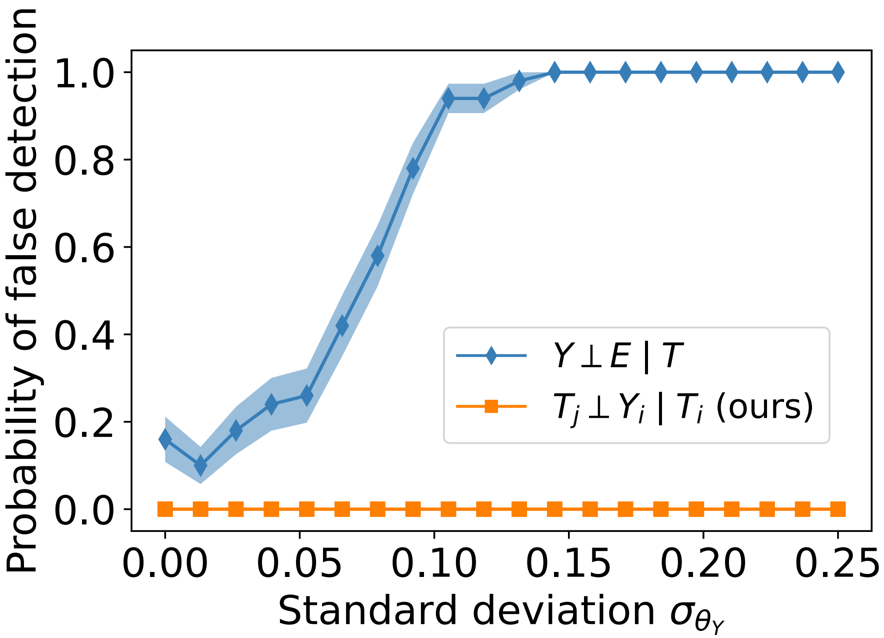
- The "sample size" of the test is the number of environments.
- We need to perform multiple tests for different pairs of (i, j).

Comparison to the naïve approach

Confounding is present



Confounding is *not* present



Variation of $\boldsymbol{\theta_Y}$
between environments

Take-aways

- We can detect hidden confounders when we have data from multiple environments
- It remains a challenge on how to efficiently test the conditional independencies in our theory
- There could be other interesting implications from assuming independent causal mechanism

arXiv pre-print

Rickard K.A. Karlsson and Jesse H. Krijthe. *Combining observational datasets from multiple environments to detect hidden confounding*, 2022.

Email

r.k.a.karlsson@tudelft.nl