Causal Structure Learning in a Nutshell

Short talk focusing on causal discovery from observational data

The Three Layer Causal Hierarchy by Pearl

Level	Typical Activity	Typical Questions
1. Association	Seeing	What is? How does observing X change my belief in Y?
2. Intervention	Doing/Intervening	What if I do X?
3. Counterfactuals	Imagining, Retrospection	Why? Was it X that caused Y?

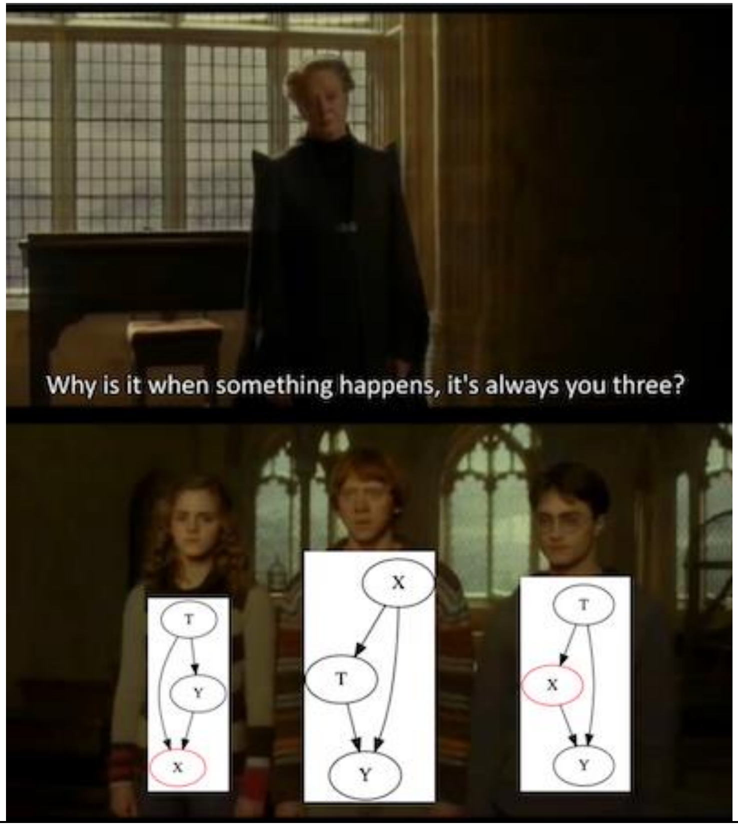
What is Causal Inference?

Inferring the effects of any treatment, policy or intervention

“Effect of some X on some Y”

Examples:

- Effect of treatment on a disease
- Effect of climate change policy on emissions
- Effect of social media on mental health



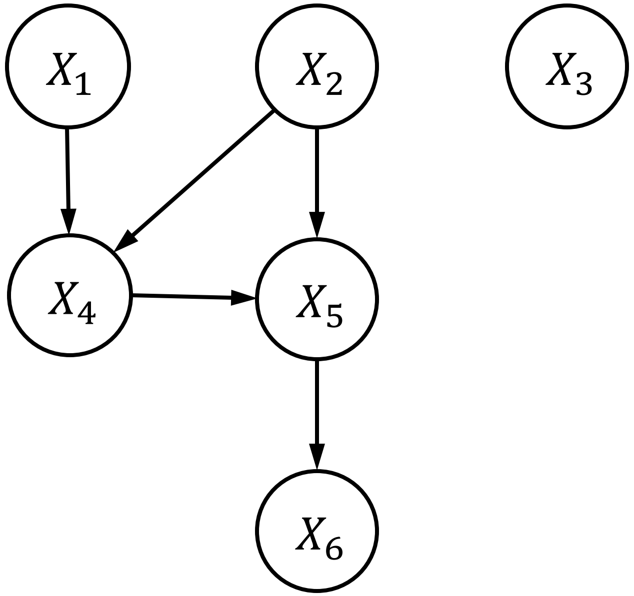
Recap: Bayesian Networks

- Graphical model:
  One-to-one mapping between nodes $i \in V$ of a direct acyclic graph (DAG) $G = (V, E)$ and random variables $X_{i \in V}$

- Local Markov Condition:
  Given the parents $\mathbf{pa}$ of an node $i \in V$ in the DAG G, the corresponding random variable X_i is independent of all its non-descendent random variables $\mathbf{nd}$.

- Bayesian Network Factorization:
  Given a joint probability distribution P_X and a DAG G, P_X factorizes according to G if:

$$P_X \coloneqq P(\{x_i\}_{i \in V}) = \prod_{i \in V} P(x_i | \mathrm{pa}(X_i))$$

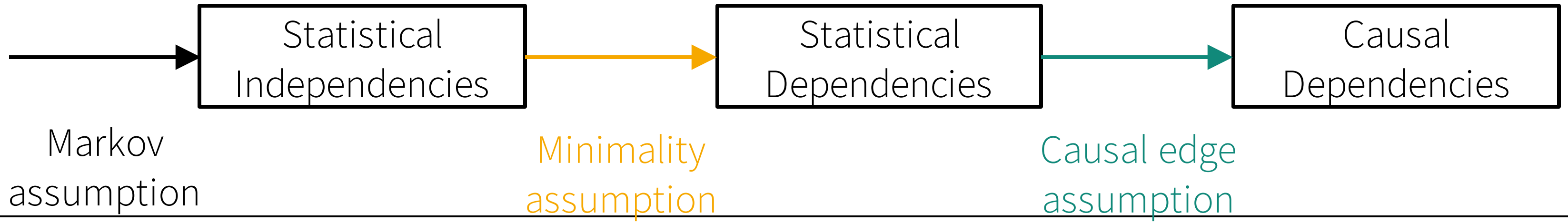


Causal Edges Model Direct and Indirect Influences

Causal Graphs

- <u>Minimality assumption:</u>
    1. *Local Markov condition*
       (implies *d-separation* as *global Markov condition*) $\qquad X_i \perp \mathrm{nd}(X_i) \mid \mathrm{pa}(X_i)$
    2. Adjacent nodes in the DAG $\boldsymbol{G}$ are dependent
       (no additional independences) $\qquad X_i \sim X_j \ in\ G \implies X_i \not\perp X_j$

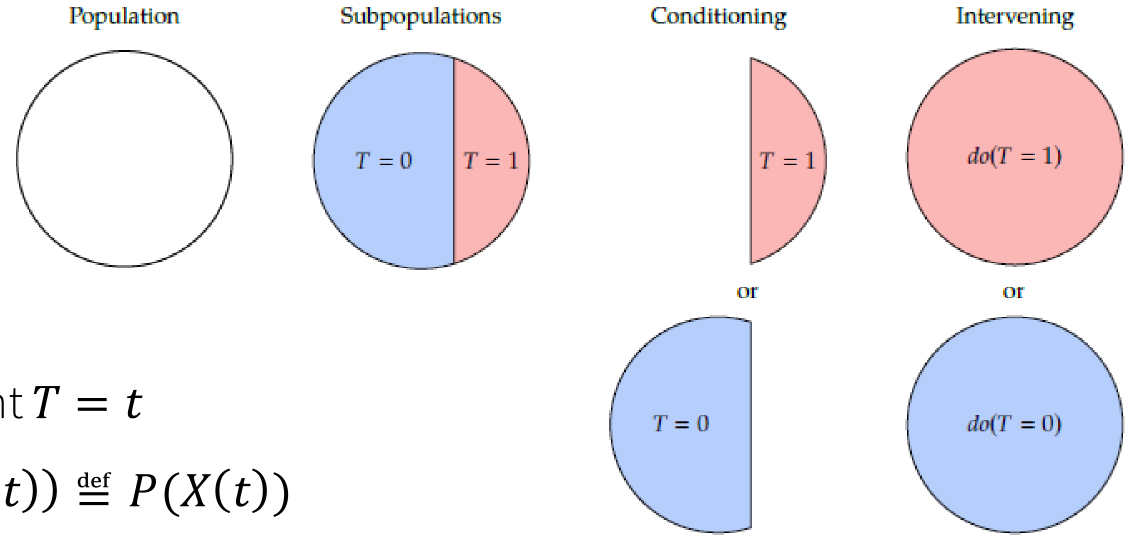
- <u>Strict causal edge assumption</u>:
  Every parent is a direct cause of all its children, i.e. the children can change in response to changes in the parents



Functional Causal Model (FCM)

- **Causal mechanism**: function g_i that generates X_i as the conditional distribution of X_i given all of its direct causes $P(x_i|\mathrm{pa_i})$, s.t. $X_i \coloneqq g_i(\mathrm{pa}_i, \epsilon_i)$

- **SCM**: tuple of:
  - endogenous variables $\boldsymbol{X}$,
  - exogeneous noise variables $\boldsymbol{\epsilon}$ (usually independent)
  - functions $\boldsymbol{g}$, one to generate each endogenous variable as a function of other variables



- (Hard) **intervention**, $\mathrm{do}(T = t)$:
  replacement of the structural function g_i with the assignment $T = t$

$$P(X|T) \coloneqq P(X = x|T = t) \neq P(X = x|do(T = t)) \stackrel{\text{def}}{=} P(X(t))$$

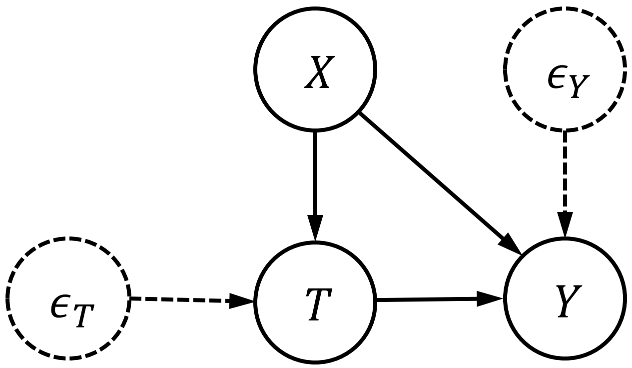
Interventions block any Noncausal-Associations

Causal Structure

- Before any intervention:

$$T \coloneqq f_T(X, \epsilon_T)$$
$$Y \coloneqq f_Y(X, T, \epsilon_Y)$$

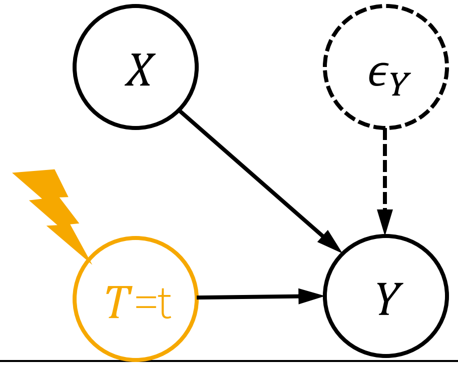
$$P(Y, T, X) = P(Y|T, X)P(T|X)P(X)$$



- After the intervention do(T=t):

$$T \coloneqq t$$
$$Y \coloneqq f_Y(X, T, \epsilon_Y)$$

$$P(X, Y|do(T = t)) = P(Y|T = t, X)P(X)$$
$$\neq$$
$$P(X, Y|T = t) = \frac{P(Y|T = t, X)P(T = t|X)P(X)}{P(T = t)}$$



Interventions Block any Noncausal Associations

Causal Structure

- Before any intervention:

$$T := f_T(X, \epsilon_T)$$
$$Y := f_Y(X, T, \epsilon_Y)$$

$$P(Y, T, X) = P(Y|T, X)P(T|X)P(X)$$

- After the intervention do(T=t):

$$T := t$$
$$Y := f_Y(X, T, \epsilon_Y)$$

$$P(X, Y|do(T = t)) = P(Y|T = t, X)P(X)$$
$$\neq$$
$$P(X, Y|T = t) = \frac{P(Y|T = t, X)P(T = t|X)P(X)}{P(T = t)}$$

universität wien

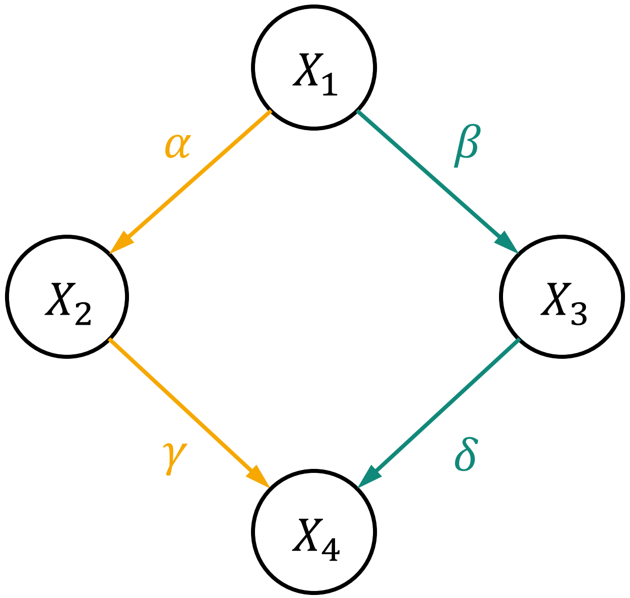
Typical Assumptions for Causal Structure Learning

- Markov assumption:

$$X \perp_G Y \mid Z \quad \Longrightarrow \quad X \perp_P Y \mid Z$$

- Faithfulness:

$$X \perp_G Y \mid Z \quad \Longleftarrow \quad X \perp_P Y \mid Z$$



$$X_4 = \gamma X_2 + \delta X_3 = \underbrace{(\alpha\gamma + \beta\delta)}_{\neq 0} X_1$$

Typical Assumptions for Causal Structure Learning

- Markov assumption:

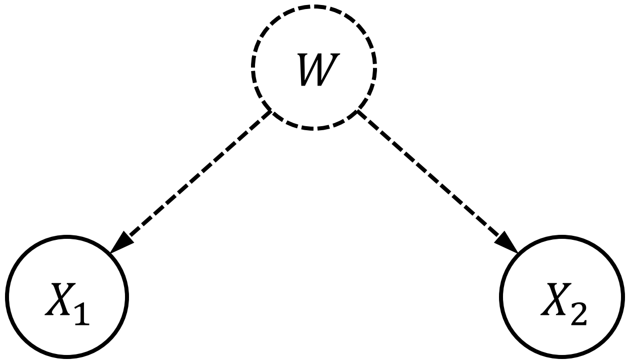
$$X \perp_G Y \mid Z \implies X \perp_P Y \mid Z$$

- Faithfulness:

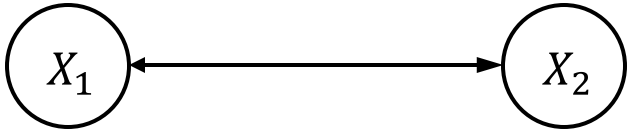
$$X \perp_G Y \mid Z \impliedby X \perp_P Y \mid Z$$

- Causal sufficiency:

No unobserved confounders.



True graph $\quad X_1 \perp X_2$



Projected graph $\quad X_1 \not\perp X_2$

Typical Assumptions for Causal Structure Learning

- Markov assumption:

$$X \perp_G Y \mid Z \quad \Longrightarrow \quad X \perp_P Y \mid Z$$

- Faithfulness:

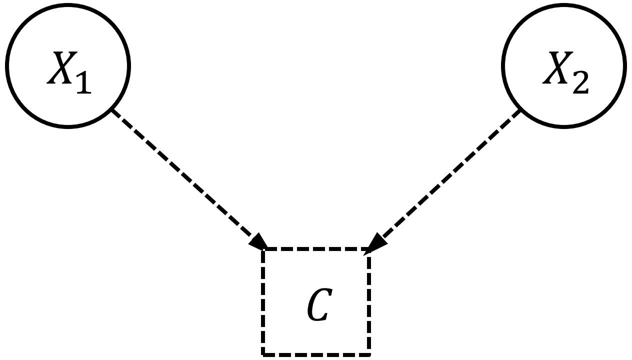
$$X \perp_G Y \mid Z \quad \Longleftarrow \quad X \perp_P Y \mid Z$$

- Causal sufficiency:

  No unobserved confounders.

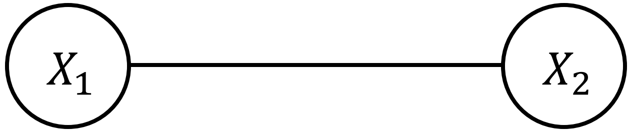
- No selection bias:

  No conditioning on unobserved colliders.



True graph

$X_1 \perp X_2$



Projected graph

$X_1 \not\perp X_2$

Unshielded Colliders can be Detected in Observational Datasets

Review of unshielded 3-node structures

Causal trail $X \rightarrow Z \rightarrow Y$

Evidential trail $X \leftarrow Z \leftarrow Y$

Common cause $X \leftarrow Z \rightarrow Y$

$X \not\perp Y$
$X \perp Y \mid Z$

$X - Z - Y$

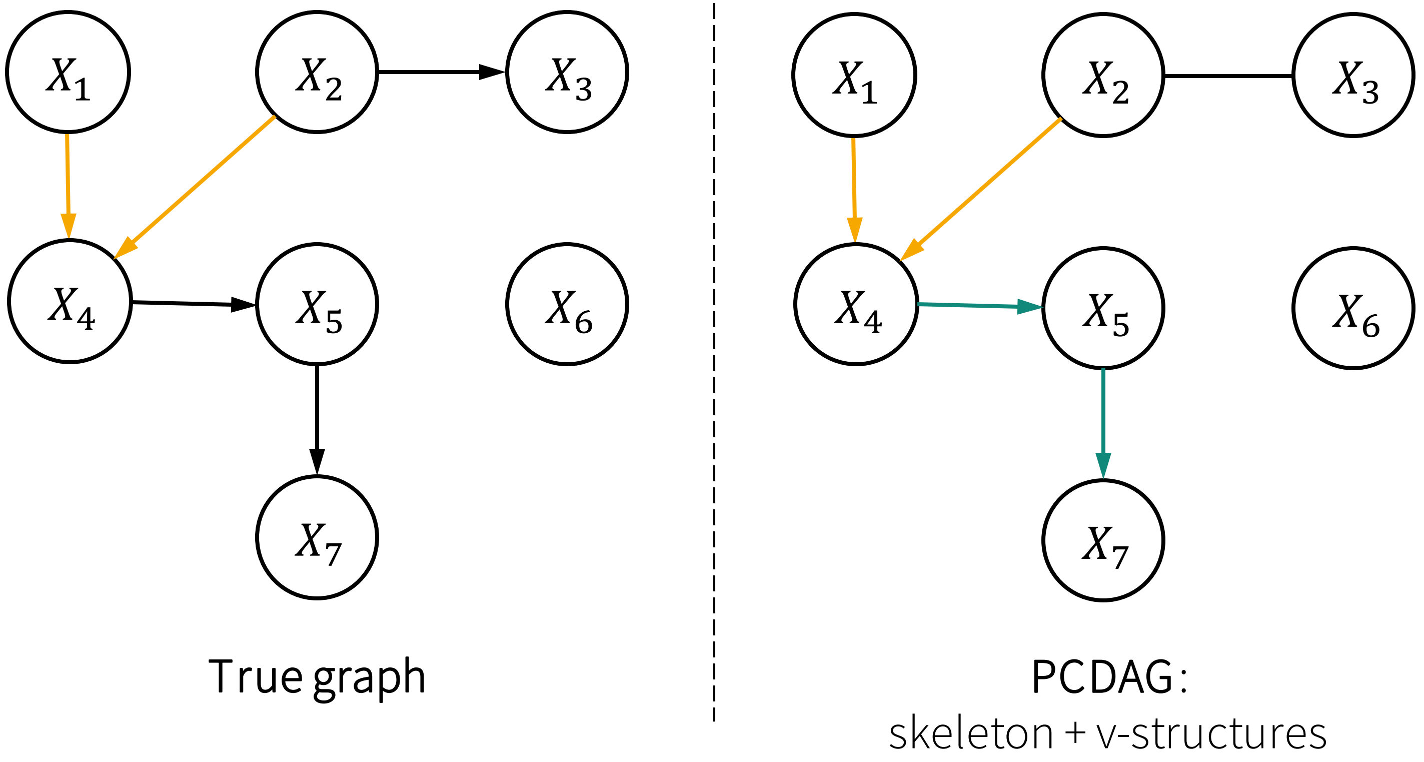
Equivalent undirected graph

Common effect $X \rightarrow Z \leftarrow Y$

$X \perp Y$
$X \not\perp Y \mid Z$

Sketch of the PC-algorithm

1) Start with the complete undirected graph
2) Eliminate edges between variables that are (conditionally) independent
3) Add arrows at colliders in identified v-structure
4) Propagate arrows such that no additional v-structures are formed



Known identifiable FCMs

- Linear Gaussian model with equal or known variance
  (Loh & Bühlmann 2014)

  $$Y := aX + \epsilon \quad \text{with} \quad \epsilon \sim \mathcal{N}(\mu, \sigma)$$

- Linear non-Gaussian model (LiNGAM)
  (Shimizu et al.. 2006)

  $$Y := aX + \epsilon \quad \text{with} \quad \epsilon \nsim \mathcal{N}(\mu, \sigma)$$

- Nonlinear additive noise model (ANL)
  (Hoyer et al., 2008)

  $$Y := g(X) + \epsilon \quad \text{where } g \text{ is nonlinear}$$

- Post-nonlinear causal model (PNL)
  (Zhang & Hyvärinen, 2009)

  $$Y := g(f(X) + \epsilon) \quad \text{where } g \text{ is invertible, } g \text{ nonconstant and both nonlinear}$$

Research Areas in Causal Structure Learning

- Removing assumptions:

No assumed causal sufficiency :	FCI algorithm (Spirtes et al., 2001)
No assumed acyclicity	CCD algorithm (Richardson, 1996)
Neither of both:	SAT-based causal discovery (Hyttinen et. al., 2013)

Research Areas in Causal Structure Learning

- Removing assumptions

- Improving computational scaling

Limiting the number of potential parents:	PNS-algorithm (Bühlmann et al., 2014)
Omitting some CI test:	RFCI-algorithm (Colombo et al, 2012)
Considering only one edge change at a time:	GES-algorithm (Chickering, 2002)
Continuous relaxation of the binary adjacency matrix:	NOTEARS-algorithm (Zheng et al., 2018)

Research Areas in Causal Structure Learning

- Removing assumptions
- Improving computational scaling
- Increasing robustness

Additional CI-tests

Order-independent PC/FCI (Colombo & Maathuis, 2014)

Research Areas in Causal Structure Learning

- Removing assumptions

- Improving computational scaling

- Increasing robustness

- Local structure learning
  - Markov boundary detection
  - Combining Markov boundaries

Research Areas in Causal Structure Learning

- Removing assumptions
- Improving computational scaling
- Increasing robustness
- Focusing only on local structures

Research Areas in Causal Structure Learning

- Removing assumptions
- Improving computational scaling
- Increasing robustness
- Focusing only on local structures
- Modelling uncertainty in the prediction

Research Areas in Causal Structure Learning

- Removing assumptions
- Improving computational scaling
- Increasing robustness
- Focusing only on local structures
- Modelling uncertainty in the prediction
- …

universität wien

END: Causal Structure Learning in a Nutshell

Short talk by Simon Rittel

Data Scientist learned

Causal Structure!

NEXT: Causal Inference for The Brave and the True

Or the Introduction to Causal Inference