Causal Inference
Theory and Applications in Enterprise Computing

Christopher Hagedorn, Johannes Huegle, Dr. Michael Perscheid

May 26, 2020

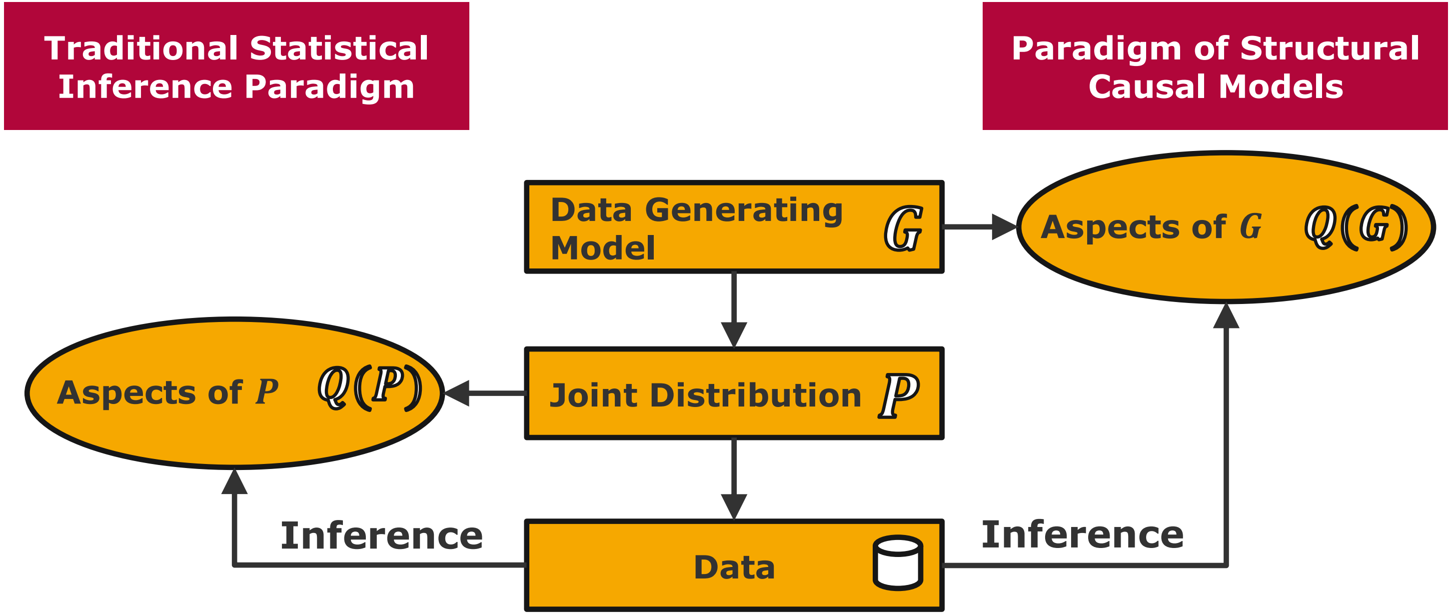
Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Agenda
May 26, 2020

- **Embedding: Causal Inference in a Nutshell**
- **Introduction to Causal Calculus**

Embedding: Causal Inference in a Nutshell

Embedding: Causal Inference in a Nutshell
Concept



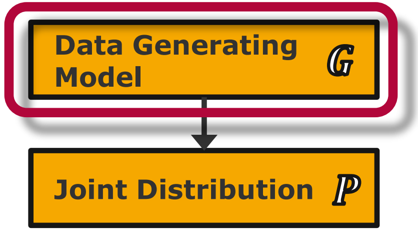
E.g., what is the sailors' probability of recovery when **we see** a treatment with lemons?

$$Q(P) = P(recovery|lemons)$$

E.g., what is the sailors' probability of recovery if **we do** treat them with lemons?

$$Q(G) = P(recovery|do(lemons))$$

Recap: Causal Inference in a Nutshell
Causal Graphical Models



Causal Graphical Model

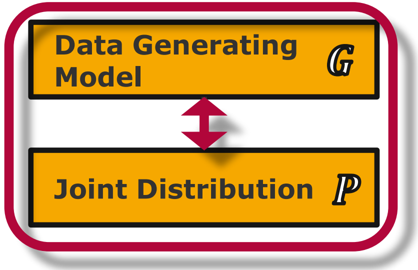
- *Directed Acyclic Graph (DAG)* $G = (V, E)$
  - *Vertices* $V_1, \ldots, V_n$
  - *Directed edges* $E = (V_i, V_j)$, i.e., $V_i \to V_j$
  - *No cycles*
- *Directed Edges* encode direct causes via
  - $V_j = f_j\big(\mathrm{Pa}(V_j), N_j\big)$ with independent noise $N_1, \ldots, N_n$

Causal Sufficiency

- All relevant variables are included in the DAG G

Recap: Causal Inference in a Nutshell

Connecting G and P



$$(X \perp\!\!\!\perp Y|Z)_G \Rightarrow (X \perp\!\!\!\perp Y|Z)_P$$

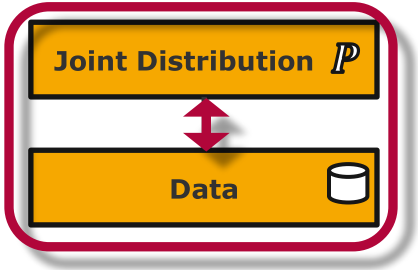
- Key Postulate: *(Local) Markov Condition*
- Essential mathematical concept: *d-Separation*
  - Idea: *Blocking* of paths
  - Implication: *Global Markov Condition*

$$(X \perp\!\!\!\perp Y|Z)_G \Leftarrow (X \perp\!\!\!\perp Y|Z)_P$$

- Key Postulate: *Causal Faithfulness*

Recap: Causal Inference in a Nutshell

Connecting P and



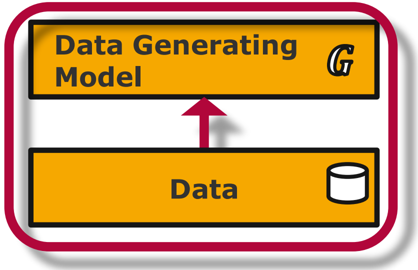
Statistical Inference

- Essential concept: *Point estimator* $\hat{\Theta}$
  - *Statistic* $g(X_1, \dots, X_n)$ of *random samples* $X_1, \dots, X_n$ to estimate *population parameter* Θ
- Inference: *Statistical Hypothesis Test*
  - *Null Hypothesis* H_0, claim on a population's property initially assumed to be true
  - *Alternative Hypothesis* H_1, a claim that contradicts H_0
  - Rejection criteria for H_0: c-value $T(x) > c$ or equivalently p-value $P_{H_0}\big(T(X) > T(x)\big) < \alpha$

$$(\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} | \boldsymbol{Z})_{\boldsymbol{P}} \Leftarrow$$

- Key idea: *Conditional Independence Test*
  - Distribution of $\boldsymbol{V} = \{V_1, \dots, V_N\} \Rightarrow$ dependence measure $T(V_i, V_j, \boldsymbol{S}) \Rightarrow$ hypothesis $H_0: t = 0$

Recap: Causal Inference in a Nutshell
Causal Structure Learning



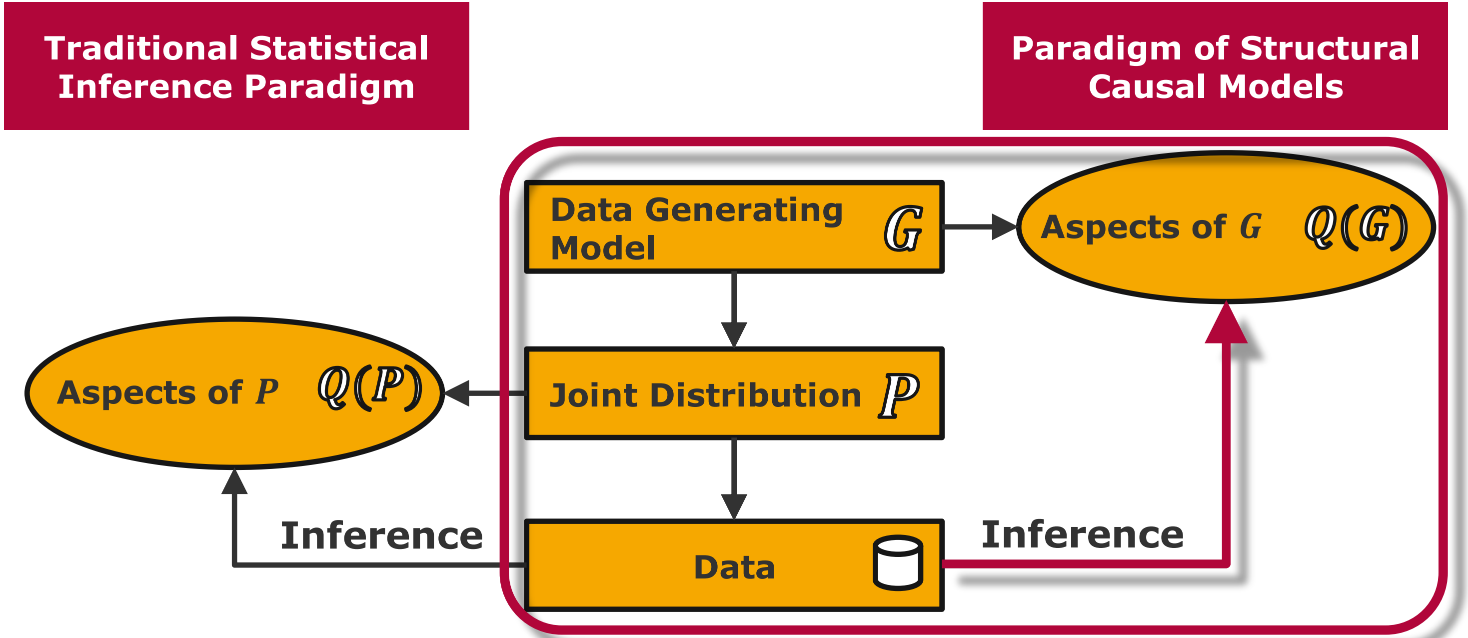
Causal Structure Learning

- Assumptions: *Causal Sufficiency, Markov Condition, Causal Faithfulness*
- Idea: Accept only those DAG's G for which $(X \perp\!\!\!\perp Y|\, Z)_G \Leftrightarrow (X \perp\!\!\!\perp Y|\, Z)_P$
  - Identifies DAG up to *Markov equivalence class* (i.e., same *skeleton C* and *v-structures*)
  - Markov equivalence class uniquely described by *completed partially directed acyclic graph (CPDAG)*
- Basis: V_i and V_j are linked if and only if there is no $S(V_i, V_j)$ s.t. $\left(V_i \perp\!\!\!\perp V_j|\, S(V_i, V_j)\right)_P$
- Methods:
  - *Constraint-based*: CI testing to derive skeleton together with edge orientation rules
  - *Score-based*: "search-and-score approach"
  - *Hybrid*: Constraint-based skeleton derivation and score-based edge orientation

HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Introduction to Causal Calculus

1. Introduction
The Concept



E.g., what is the sailors' probability of recovery when **we see** a treatment with lemons?

$$Q(P) = P(recovery|lemons)$$

E.g., what is the sailors' probability of recovery if **we do** treat them with lemons?

$$Q(G) = P(recovery|do(lemons))$$

1. Introduction
Simpson's Paradox

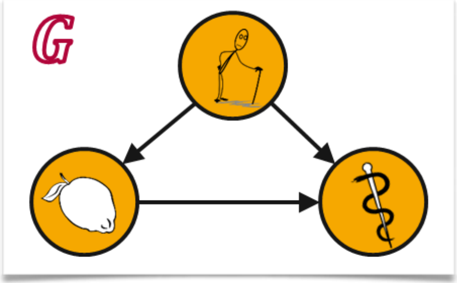
Recap the scurvy experiment:

- We observed

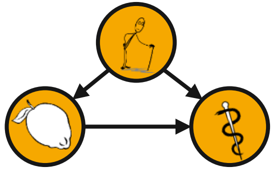
$$P(recovery|lemons, old) > P(recovery|no\ lemons, old)$$
$$P(recovery|lemons, young) > P(recovery|no\ lemons, young)$$

But: $$P(recovery|lemons) < P(recovery|no\ lemons)$$

- This reversal of the association between two variables after considering the third variable is called **Simpson's paradox**.

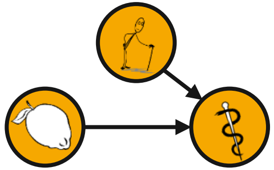


➡ **Observational Regime**



vs.

Interventional Regime



- Pearl extends probability calculus by introducing a new operator for describing interventions, the **do-operator**.

The do-operator

- $do(\ldots)$ marks intervention in the model
  - *In an algebraic model:* we replace certain functions with a constant $X = x$
  - *In a graph:* we remove edges going into the target of intervention, but preserve edges going out of the target.
- The causal calculus uses
  - *Bayesian conditioning*, $p(y|x)$, where x is observed variable
  - *Causal conditioning*, $p(y|do(x))$, where we force a specific value x

➡ **Goal:** Generate probabilistic formulas for the effect of interventions in terms of the observed probabilities.

Resolution of Simpson's paradox

- Simpson's paradox is only paradoxical if we misinterpret

  $P(recovery|lemons)$ as $P(recovery|do(lemons))$
- We should treat scurvy with lemons if

$$P(recovery|do(lemons)) > P(recovery|do(no\ lemons))$$

1. Introduction
Resolution of Simpson's Paradox

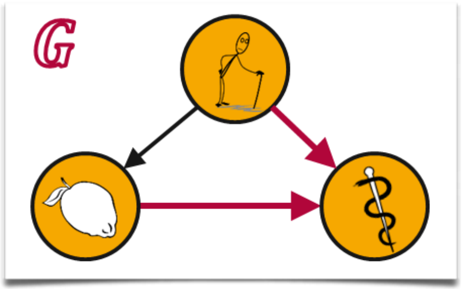
- The treatment does not affect the distribution of the subpopulations, i.e.,

$$P(old|do(lemons) = P(old|do(no\ lemons)) = P(old)$$

- Then, it is impossible that we have, simultaneously,

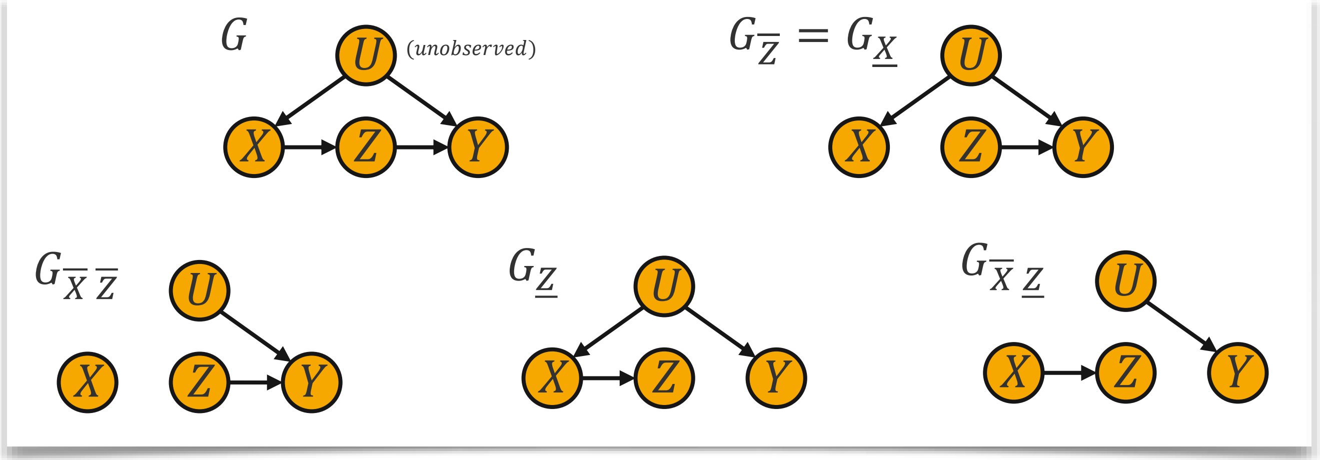
$$P(recovery|lemons, old) > P(recovery|no\ lemons, old)$$
$$P(recovery|lemons, young) > P(recovery|no\ lemons, young)$$

But: $P(recovery|lemons) < P(recovery|no\ lemons)$



- **Proof:**
  - $P(recovery|do(lemons)) = P(recovery|do(lemons), old)\ P(old|do(lemons)) + P(recovery|do(lemons), young)\ P(young|do(lemons)) = P(recovery|do(lemons), old)\ P(old) + P(recovery|do(lemons), young)\ P(young)$
  - $P(recovery|do(no\ lemons)) = P(recovery|do(no\ lemons), old)\ P(old) + P(recovery|do(no\ lemons), young)\ P(young)$
  - *Hence:* $P(recovery|do(lemons)) > P(recovery|do(no\ lemons))$

2. The Calculus of Intervention
Perturbed Graphs



- G Graph
- W, X, Y, Z, U disjoint subsets of the variables
- $G_{\overline{X}}$ perturbed graph in which all edges *pointing to* X have been deleted
- $G_{\underline{X}}$ perturbed graph in which all edges *pointing from* X have been deleted
- $Z(W)$ set of nodes in Z which are *not ancestors of* W

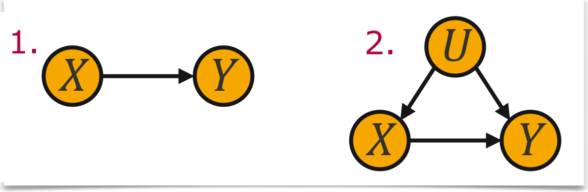
Causal Inference
Theory and Applications in Enterprise Computing

Hagedorn, Huegle, Perscheid

Definition:

Let $Q(M)$ be any computable quantity of a model M. We say that Q *is identifiable* in a class M of models if, for any pairs of models M_1 and M_2 from M, $Q(M_1) = Q(M_2)$ whenever $P_{M_1}(v) = P_{M_2}(v)$.

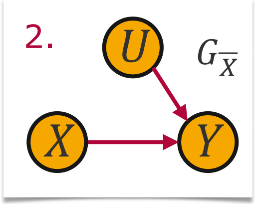
- I.e., $P(y|do(x))$ is identifiable
  if it can be consistently estimated from data involving only observed variables.
- **Examples:**



- Can you estimate $P(y \mid do(x))$, given $P(x, y)$?
  1. Yes, since $P(y|do(x)) = P(y|x)$, i.e., $P(y|do(x))$ is identifiable
  2. No (observational regime), since $P(x, y) = \sum_u P(x, y, u) = \sum_u P(y|x, u)P(x|u)P(u)$
     $$P(y|do(x)) = \sum_u P(y|x, u)P(u)$$

2. The Calculus of Intervention
Back-Door Criterion

- **But:** after adjustment for direct causes (intervention)
  - $P(x, y) = \sum_u P(x, y, u) = \sum_u P(y|x, u)\cancel{P(x|u)}P(u) = P(y|do(x))$
  - Hence, $P(y|do(x))$ is identifiable
- Any common ancestor of X and Y is a *confounder*
- Confounders originate "back-door" paths that need to be blocked by conditioning
- This defines a basic criterion for identifiability:



Back-Door Criterion (Pearl 1993):
A set of variables Z satisfies the *back-door criterion* relative to an ordered pair of variables (V_i, V_j) in a DAG G if:
1. no node in Z is a descendant of V_i; and
2. Z blocks every path between V_i and V_j that contains an arrow to V_i.

➡ Back-door adjustment: $P(v_j|do(v_i)) = \sum_z P(v_j|v_i, z)P(z)$

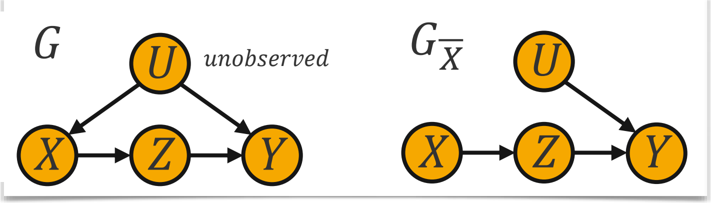
2. The Calculus of Intervention
Front-Door Criterion

- **But:** If U is hidden (unobserved), then there is no data for conditioning
- Then, $P(y|do(x))$ is also identifiable!

  $P(y|do(x)) = \sum_z P(y|do(z))P(z|do(x))$

  $= \sum_z P(y|do(z))P(z|x)$ (direct effect)

  $= \sum_{x'} P(y|x', z)P(x')P(z|x)$ (back-door)



- This defines a basic criterion for identifiability with unobserved variables:

Front-Door Criterion (Pearl 1993):
A set of variables Z satisfies the *front-door criterion* relative to an ordered pair of variables (V_i, V_j) in a DAG G if:
1. Z intercepts all directed paths from V_i to V_j ; and
2. there is no unblocked back-door path from V_i to Z; and
3. all back-door paths from Z to V_j are blocked by V_i

➡ Front-door adjustment: $P\left(v_j \middle| do(v_i)\right) = \sum_z P(z|v_i) \sum_{v_i'} P(v_j | v_i', z)P(v_i')$

2. The Calculus of Intervention

The do-Calculus (Pearl 1995)

The do-Calculus:

Let X, Y, Z, and W be arbitrary disjoint sets of nodes in a causal DAG G.

Rule 1: Ignoring observations

$$p(y|do(x), z, w) = p(y|do(x), w) \quad if\ (Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}}}$$

Rule 2: Action/Observation exchange (Back-Door)

$$p(y|do(x), do(z), w) = p(y|do(x), z, w) \quad if\ (Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}, \underline{Z}}}$$

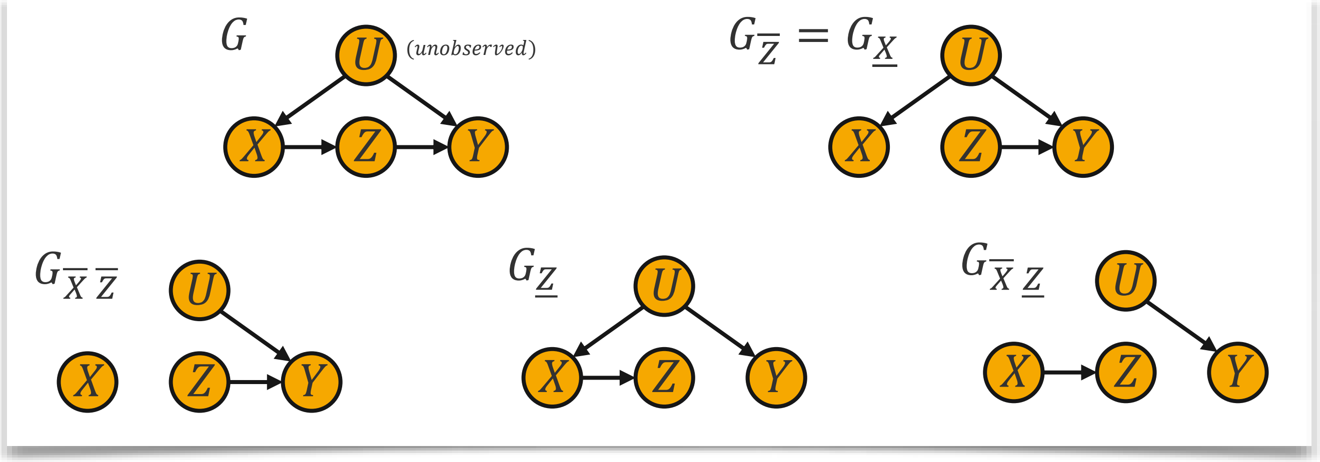
Rule 3: Ignoring actions/interventions

$$p(y|do(x), do(z), w) = p(y|do(x), w) \quad if\ (Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}, \overline{Z(W)}}}$$

Notes:

- *Allows a syntactical derivation of claims about interventions*
- *The calculus is sound and complete*
  - *Sound*: If the do-operations can be removed by repeated application of these three rules, the causal effect is identifiable. (Galles et al. 1995)
  - *Complete:* If identifiable, the do-operations can be removed by repeated application of these three rules. (Huang et al. 2012)
  - I.e., "it works on all inputs and always gets the right result"
- *Also allows for identifiability of causal effects in MAGs*

3. Estimating Causal Effects
Deriving Causal Effects using the do-Calculus



- **Example: Compute $P(y|do(z))$**

  We have $P(y|do(z)) = \sum_x P(y|x, do(z))\, P(x|do(z))$

  $= \sum_x P(y|x, do(z))\ P(x)$ (Rule 1: $(Z \perp\!\!\!\perp X)_{G_{\overline{Z}}}$)

  $= \sum_x P(y|x, z)\ P(x)$ (Rule 2: $(Z \perp\!\!\!\perp Y)_{G_{\underline{Z}}}$)

3. Estimating Causal Effects
Quantifying Causal Strength

- The *Causal Effect of $V_i = v_i$ on V_j* is given by $P\big(\mathrm{V_j}\big|do(V_i = v_i)\big)$
  - I.e., the distribution of V_j given that we force V_i to be v_i
  - This defines the basis of the examination of causal effects
- But: Quantifying the causal influence of V_i on V_j is a nontrivial question!
- Many *measures of causal strength* depending on the causal structures have been proposed, e.g.,
  - *Average Treatment Effect (ATE):*

    $E[V_j|do(V_i = 1)] - E[V_j|do(V_i = 0)]$ for binary V_i
  - *Average Causal Effect (ACE)*:

    $\frac{\partial}{\partial v_i} E\big[V_j\big|do(V_i = v_i)\big]$ for continuous V_i, V_j
  - *Conditional Mutual Information (CI):*

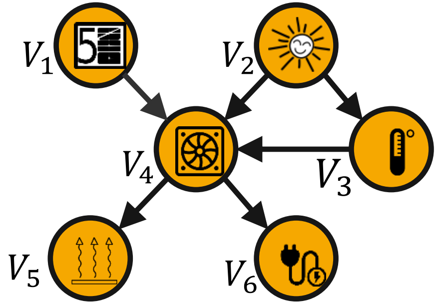
    $\sum_{v_i,v_j} P(v_i)P\big(v_j|do(V_i = v_i)\big) \log \frac{P(v_j|do(V_i=v_i))}{\sum_{v_i'} P(V_i = v_i')P(v_j|do(V_i = v_i'))}$ for categorical V_i, V_j
  - *Relative Entropy, etc.*

Hasso Plattner Institut

Recap the cooling house example

- We are in the multivariate normal case
- Hence, average causal effects are given by
  - $\mathrm{ACE}(V_4, V_1, v_1) = \frac{\partial}{\partial v_1} E[V_4 | do(V_1 = v_1)]$
    $= E[V_4 | do(V_1 = v_1 + 1)] - E[V_4 | do(V_1 = v_1)]$ (linear f)
    $= \beta_{V_1 \to V_4} \;= 4$
  - $\mathrm{ACE}(V_6, V_1, v_1) = \frac{\partial}{\partial v_1} E[V_6 | do(V_1 = v_1)]$
    $= E[V_6 | do(V_1 = v_1 + 1)] - E[V_6 | do(V_1 = v_1)]$
    $= \beta_{V_1 \to V_4} \cdot \beta_{V_4 \to V_6} = 4 \cdot 1.2 \;= 4.8$
  - $\mathrm{ACE}(V_4, V_2, v_2) = \frac{\partial}{\partial v_2} E[V_4 | do(V_2 = v_2)]$
    $= E[V_4 | do(V_2 = v_2 + 1)] - E[V_4 | do(V_2 = v_2)]$
    $= \beta_{V_2 \to V_4} + \beta_{V_2 \to V_3} \cdot \beta_{V_3 \to V_4} \;= 5 + 3 \cdot 0.7 \;= 7.1$
  - $\mathrm{ACE}(V_6, V_5, v_5) = 0$

Cooling House Example:



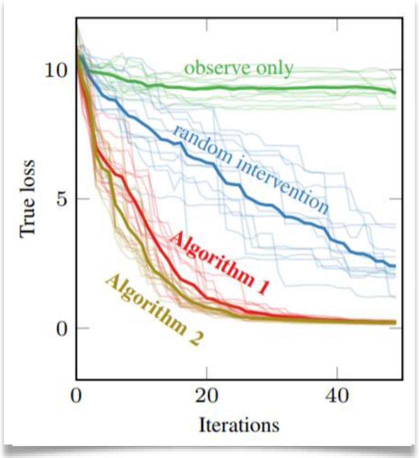
- $V_1 = N(0,1)$
- $V_2 = N(0,1)$
- $V_3 = 3\,V_2 + N(0,1)$
- $V_4 = 4\,V_1 + 5\,V_2 + 0.7\,V_3 + N(0,1)$
- $V_5 = V_4 + N(0,1)$
- $V_6 = 1.2\,V_4 + N(0,1)$

4. Excursion
Causal Functional System (e.g., Rubenstein 2017)

Idea:

The identification of the underlying causal graph G allows to learn the functions computing children from parents in the structural causal model.

- I.e., the logical second step after the causal discovery
- The do-operator builds a natural basis of probabilistic learning algorithms for estimating the functional system:
  - Active Bayesian learning allows for identification of interventions that are optimally informative about all of the unknown functions (Algorithm 1)
  - Exploiting factorization properties allows for vectorization and simultaneous calculations in a dynamic programming approach (Algorithm 2)
- *Probabilistic active learning of functions* significantly improves the estimation compared to unstructured base-lines (Observe only, random intervention).



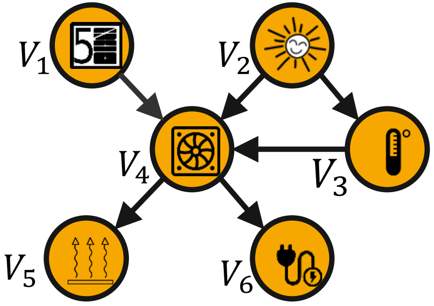
4. Excursion
Causal Functional System (A Naive Example!)

- **Goal:** Estimate $\beta_{V_1 \to V_4}$
- **Recall**: True $\beta_{V_1 \to V_4} = 4$

- **Linear Regression Model Approach:**
  - Fit linear model $V_4 = lm(V_1, V_2, V_3, V_5, V_6)$
  - Then $\hat{\beta}_{V_1 \to V_4} = 1.14$
  - ⇨ Underestimated $\beta_{V_1 \to V_4}$

- **Causal Structural Approach:**
  - From estimated CPDAG $\hat{G}$ we know $V_1 = Pa(V_4)$
  - Hence, $\hat{\beta}_{V_1 \to V_4} = \widehat{ACE}(V_4, V_1, v_1) \in \{4.09, 4.09\}$
  - ⇨ Estimated $\beta_{V_1 \to V_4}$ (up to the equivalence class)

Cooling House Example:



- $V_1 = N(0,1)$
- $V_2 = N(0,1)$
- $V_3 = 3\, V_2 + N(0,1)$
- $V_4 = 4\, V_1 + 5\, V_2 + 0.7\, V_3 + N(0,1)$
- $V_5 = V_4 + N(0,1)$
- $V_6 = 1.2\, V_4 + N(0,1)$

References
Literature

- Pearl, J. (2009). *Causal inference in statistics: An overview*. Statistics Surveys.
- Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.
- Spirtes et al. (2000). *Causation, Prediction, and Search*. The MIT Press.
- Pearl, J. (1995). *Causal diagrams for empirical research*. Biometrika.
- Maathuis et al. (2013). *A generalized backdoor criterion*. arXiv.
- Galles et al. (1995). *Testing identifiability of causal effects*. In Proceedings of UAI-95.
- Huang et al. (2012). *Pearl's Calculus of Intervention Is Complete*. arXiv.
- Pearl, J (2012). *The Do-Calculus Revisited*. arXiv.
- Janzing et al. (2013) *Quantifying causal influences*. The Annals of Statistics.
- Rubenstein et al. (2017). *Probabilistic Active Learning of Functions in Structural Causal Models*. arXiv.

Thank you
for your attention!