Causal Inference
Theory and Applications in Enterprise Computing

Christopher Hagedorn, Johannes Huegle, Dr. Michael Perscheid

May 13, 2020

Hasso Plattner Institut
Digital Engineering • Universität Potsdam

Agenda
May 13, 2020

- **Group Topic Assignment**
  1. Causal Structure Learning II
  2. Application Scenario

- **Recap: Causal Inference in a Nutshell**
  - Conditional Independence Tests

- **Jupyter Lab**
  1. Causal Inference in a Nutshell - Cooling House Scenario
  2. CI Tests - Exercises

Topic Assignment

Heterogeneous Computing for Constraint-Based Causal Structure Learning

GPUs provide massive parallel compute power to conduct independence tests in parallel, which have been adopted in the PC-Algorithm. Yet, during learning CPU resources remain widely unused.

Hagedorn, Huegle, Perscheid

Introduction to Group Topics
Causal Structure Learning II

Research Question/Task:

How can CPU resources be efficiently utilized during the GPU-accelerated algorithm‘s execution to reduce computation time? Further, is there a sweet spot between GPU/CPU execution, in which switching between the mode‘s is preferable?

Reading Material:

- cuPC: CUDA-based Parallel PC Algorithm for Causal Structure Learning on GPU
  https://arxiv.org/abs/1812.08491
  Implementation: https://github.com/LIS-Laboratory/cupc
- PC-Algorithm R-Implementation
  https://cran.r-project.org/web/packages/pcalg/index.html
  Framework for Parallel Constraint-Based Learning
  https://www.jstatsoft.org/article/view/v077i02

From Data to Causal Inference:
A Printing Press Example

Learning causal models has a large potential in real world settings. Despite the limitations, given that data is usually not captured with the application of causal inference in mind, it is important to understand the value of learnt causal graphical models.

Introduction to Group Topics
Application Scenario

Research Question/Task:

Understand the provided real world use case and its data. What needs to be considered during data preprocessing and handling in order to apply causal structure learning and further apply causal inference on the learnt model?

Reading Material:

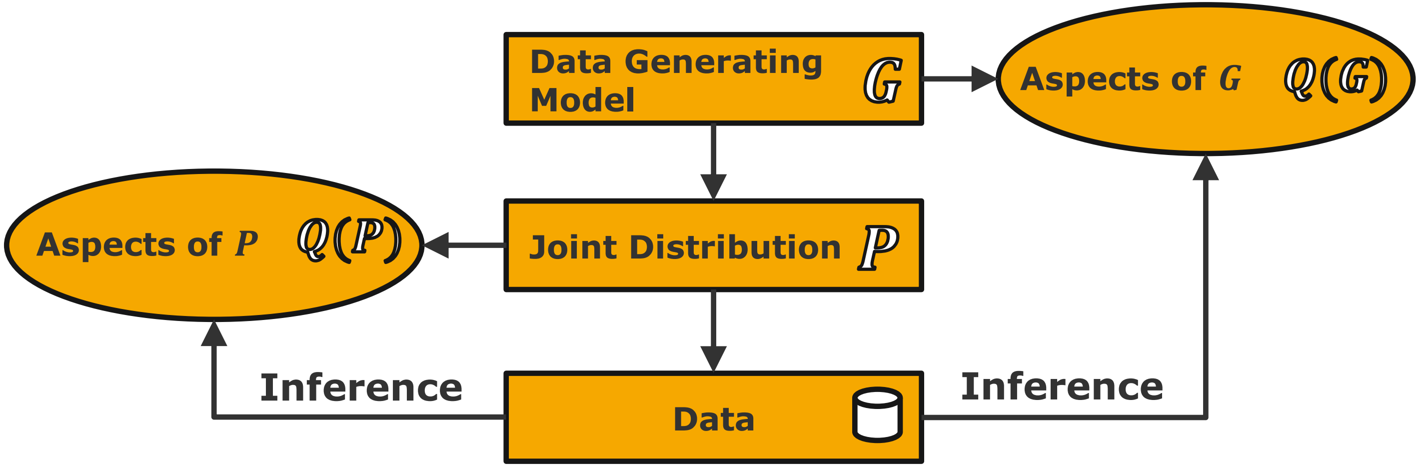
- Master Thesis by Daniel Thevessen
  excerpts provided via e-mail
- Slides from use case
  provided via e-mail

Recap: Causal Inference in a Nutshell

Recap: Causal Inference in a Nutshell
Concept

Traditional Statistical Inference Paradigm

Paradigm of Structural Causal Models



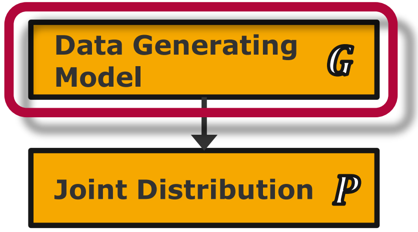
E.g., what is the sailors' probability of recovery when **we see** a treatment with lemons?

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

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



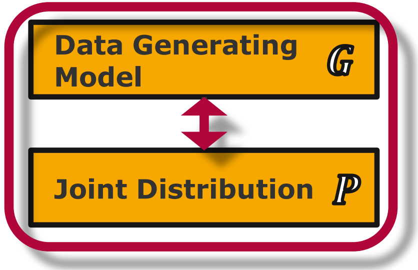
Causal Graphical Model

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

Causal Sufficiency

- All relevant variables are included in the DAG G

Recap: Causal Inference in a Nutshell
Connecting G and P



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

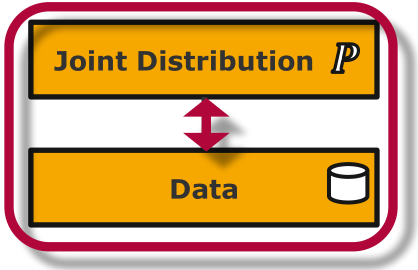
- Key Postulate: *(Local) Markov Condition*
- Essential mathematical concept: *d-Separation*
  - Idea: *Blocking* of paths
  - Implication: *Global Markov Condition*

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

- Key Postulate: *Causal Faithfulness*

Recap: Causal Inference in a Nutshell

Connecting P and



Statistical Inference:

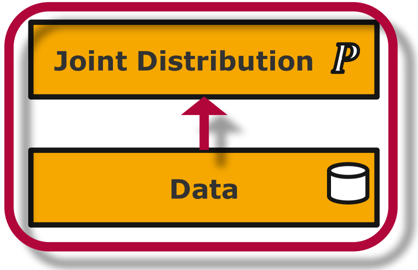
Deduce properties of a population's distribution P on the basis of random sampling.

Random samples $X_1, \dots, X_n$

- *independent and identically distributed* (*i.i.d.*) random variables $X_1, \dots, X_n$

Point estimator $\widehat{\Theta}$

- A *statistic* $g(X_1, \dots, X_n)$ of the random samples $X_1, \dots, X_n$ to estimate *a population parameter* θ
- Follows a probability distribution (*sampling distribution*)



Statistical Hypothesis Test

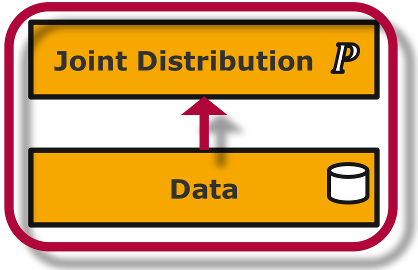
- Decision rule to examine a hypothesis on a population's property via a test statistic T
  - *Null Hypothesis* H_0 is the claim that is initially assumed to be true
  - *Alternative Hypothesis* H_1 is a claim that contradicts H_0

Concept

- Approximate T under H_0 by a known sampling distribution P_{H_0}
- Derive rejection criteria for H_0
  - c-value: reject H_0 if $T(x) > \text{c}$ for a $c \in \mathbb{R}$
  - p-value: reject H_0 if $P_{H_0}\big(T(X) > T(x)\big) < \alpha$

  } are equivalent

Recap: Causal Inference in a Nutshell
Conditional Independence Tests



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

Conditional Independence Test

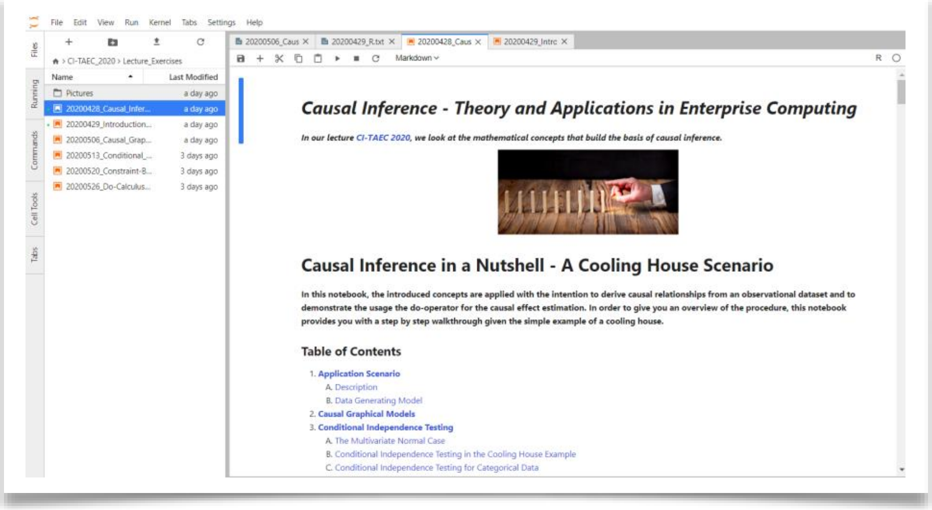
- Distribution of $\boldsymbol{V} = \{V_1, \dots, V_N\} \Rightarrow$ dependence measure $T(V_i, V_j, \boldsymbol{S})$
  $\Rightarrow$ hypothesis test $H_0: t = 0$

Multivariate Normal Distributed $\boldsymbol{V}$

- V_i and V_j are conditionally independent given the separation set $\boldsymbol{S} \subset V / \{V_i, V_j\}$ if and only if the corresponding partial correlation $\rho(V_i, V_j \mid \boldsymbol{S})$ to zero.
- Given significance level α, we reject the null-hypothesis $H_0: \rho(V_i, V_j | \boldsymbol{S}) = 0$ if for the corresponding estimated p-value it holds that $\hat{p}(V_i, V_j | \boldsymbol{S}) \leq \alpha$

Jupyter Lab

Jupyter Lab
Causal Inference in a Nutshell - Cooling House



Topics

- The Multivariate Normal Case
- CI Testing in the Cooling House Example
- CI Testing for Categorical Data

Jupyter Lab
CI Tests - Exercises

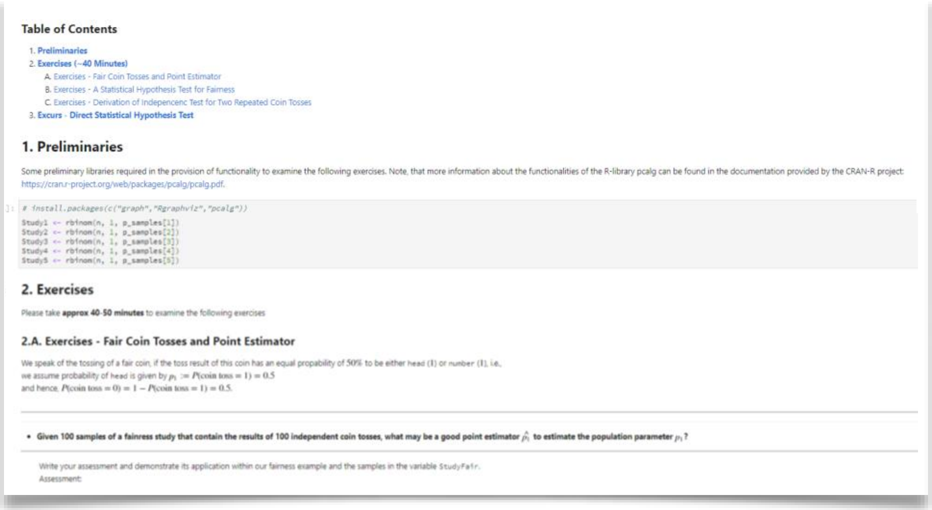


Table of Contents

1. Preliminaries

Some preliminary libraries required in the provision of functionality to examine the following exercises. Note, that more information about the functionalities of the R-library pcalg can be found in the documentation provided by the CRAN-R project: https://cran.r-project.org/web/packages/pcalg/pcalg.pdf.

```
# install.packages(c("graph","Rgraphviz","pcalg"))
Study1 <- rbinom(n, 1, p_samples[1])
Study2 <- rbinom(n, 1, p_samples[2])
Study3 <- rbinom(n, 1, p_samples[3])
Study4 <- rbinom(n, 1, p_samples[4])
Study5 <- rbinom(n, 1, p_samples[5])
```

2. Exercises

Please take **approx 40-50 minutes** to examine the following exercises

2.A. Exercises - Fair Coin Tosses and Point Estimator

We speak of the tossing of a fair coin, if the toss result of this coin has an equal propability of 50% to be either head (1) or number (1), i.e.,
we assume probability of head is given by $p_1 := P(\text{coin toss} = 1) = 0.5$
and hence, $P(\text{coin toss} = 0) = 1 - P(\text{coin toss} = 1) = 0.5$.

- **Given 100 samples of a fairness study that contain the results of 100 independent coin tosses, what may be a good point estimator $\hat{p}_1$ to estimate the population parameter p_1?**

Write your assessment and demonstrate its application within our fairness example and the samples in the variable StudyFair.
Assessment:

Exercises

- Fair Coin Tosses and Point Estimator
- A Statistical Hypothesis Test for Fairness
- Derivation of Independence Test for Two Repeated Coin Tosses

Thank you
for your attention!