TUBAF

Die Ressourcenuniversität.
Seit 1765.

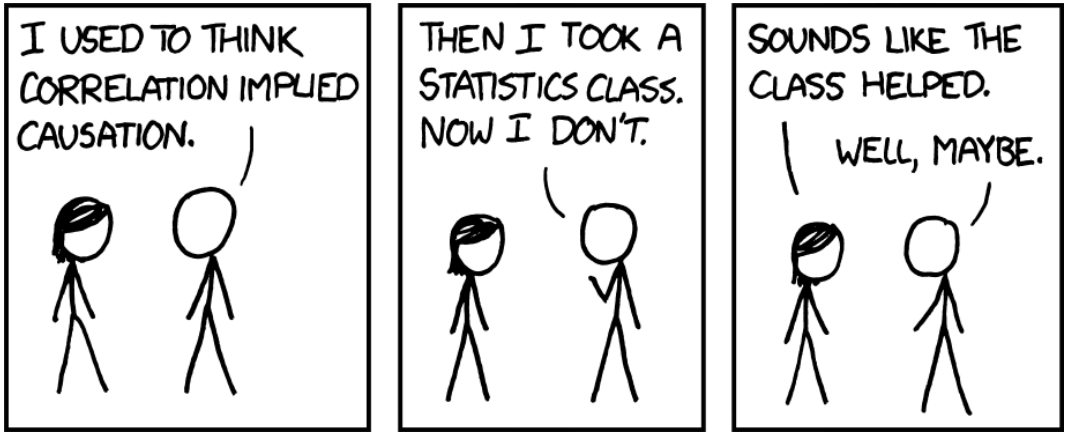
Introduction to Causal Inference

Research seminar on causality

Björn Sprungk

Institute of Stochastics, TU Bergakademie Freiberg

May 08th, 2024



Source: xkcd.com

Content

1. Correlation and Causation

2. Statistical Models and Machine Learning

3. Structural Cause-Effect Models

Correlation and Causation

Correlation does not imply causation

- A correlation between two random variables X_1, X_2, i.e.,

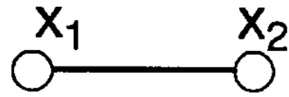
$$\text{Corr}(X_1, X_2) = \text{Corr}(X_2, X_1) \neq 0$$

where

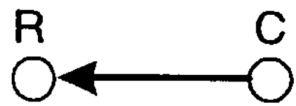
$$\text{Corr}(X_1, X_2) = \mathbb{E}\left[(X_1 - \mathbb{E}[X_1])\,(X_2 - \mathbb{E}[X_2])\right].$$

is a symmetric relation, an association

- Causation is by nature asymmetric: If C is a cause for R than R can not be a cause for C.



1a



1b

1

[1] Figure from: D. R. Cox and N. Wermuth. Some Statistical Aspects of Causality. *European Sociological Review* 17(1):65-74, 2001.

Causation

- When is an event C a cause for another event E which is then the effect ?
- In philosophy: C is a **necessary and sufficient** condition for E to occur
- Obviously too strict for practice (e.g., $C =$ smoking, $E =$ lung cancer)
- One probabilistic approach

Candidate cause[2]

The event C is a candidate cause of E if

$$\mathbb{P}(E \mid C) > \mathbb{P}(E \mid \overline{C}).$$

[2] D. R. Cox. Causality: Some Statistical Aspects. *J. R. Statist. Soc. A* 155(2):291–301, 1992.

Remarks

- By the *total law of probability*

$$\mathbb{P}(E) = \mathbb{P}(C)\,\mathbb{P}(E \mid C) + \mathbb{P}(\overline{C})\,\mathbb{P}(E \mid \overline{C})$$

  we can easily conclude that

$$\mathbb{P}(E \mid C) > \mathbb{P}(E \mid \overline{C}) \quad \Longleftrightarrow \quad \mathbb{P}(E \mid C) > \mathbb{P}(E)$$

  where the latter is also known as positive dependence of E on C

- Positive dependence of events C, E is again a symmetric relation:

$$\mathbb{P}(E \mid C) > \mathbb{P}(E) \quad \Longleftrightarrow \quad \mathbb{P}(E \cap C) > \mathbb{P}(E)\mathbb{P}(C) \quad \Longleftrightarrow \quad \mathrm{Corr}(\mathbf{1}_C, \mathbf{1}_E) > 0$$

- In literature, this is also often called events C and E are correlated

- Thus, if C is a candidate cause for E, also E is a candidate cause for C...

- To express the asymmetry between cause and effect possible restrictions are[3]:
  1. **Temporal ordering:** C has to occur before E in time
  2. **Spatial proximity:** can (alternatively) be used as basis for ordering cause and effect
  3. **Subject-matter knowledge:** "to establish a presumed causal ordering"
- Moreover, candidate causes can still be spurios causes:

  C: high ice cream sales, E: many heatstrokes (the next day)

  satisfies (probably) $\mathbb{P}(E \mid C) > \mathbb{P}(E \mid \overline{C})$.

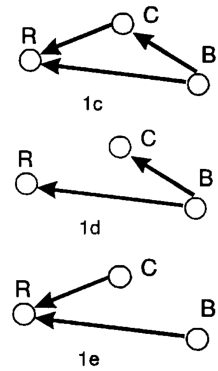
[3] D. R. Cox. Causality: Some Statistical Aspects. *J. R. Statist. Soc. A* 155(2):291–301, 1992.

Spurious causes

If C is a candidate cause of E but there exists a (background) event B which explains the association, i.e.,

$$\mathbb{P}(E \mid C \cap B) = \mathbb{P}(E \mid \overline{C} \cap B)$$

then, we call C a spurious cause of E



- Spurious causes are equivalent to conditional independence

$$\mathbb{P}(E \cap C \mid B) = \mathbb{P}(E \mid B)\mathbb{P}(C \mid B) \quad \Longleftrightarrow \quad E \perp\!\!\!\perp C \mid B$$

- **Example:**

C: high ice cream sales, E: many heatstrokes, B: begin of heat period

Reichenbach's commmon cause principle

If two events A and B are (positively) correlated, i.e. if

$$\mathbb{P}(A \cap B) > \mathbb{P}(A)\,\mathbb{P}(B)$$

then, either

1. A is a cause for B or
2. B is a cause for A or
3. there exists a common cause C for A and B, i.e., an event C such that

$$\mathbb{P}(A \cap B \mid C) = \mathbb{P}(A \mid C)\mathbb{P}(B \mid C).$$

Hans Reichenbach
(1891–1953)

- Rather a requirement/axiom than a theorem
- Applies in many situations in real life but not necessarily in quantum field theory

Further notions of causality [4]

2. **Causality as the Effect of Intervention**
   - If we forbid ice cream sales (C), will this yield a reduction of heatstrokes (E) in practice?
   - Note that this is not captured by simply considering

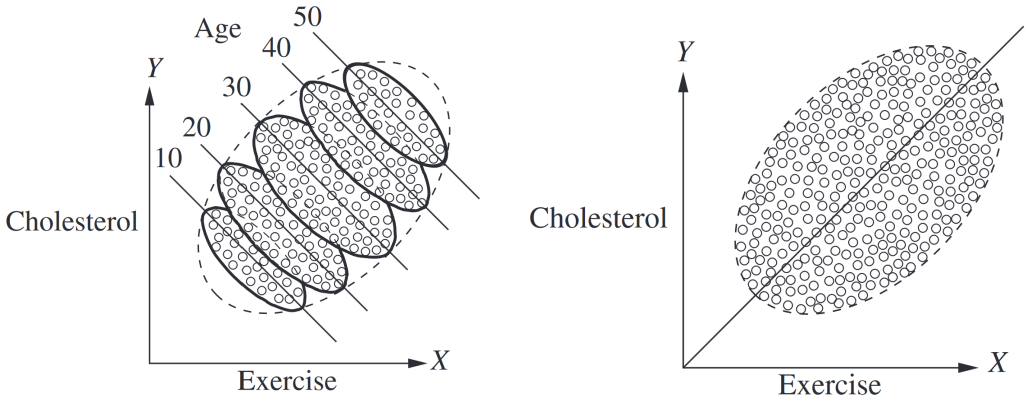
$$\mathbb{P}(E \mid C) - \mathbb{P}(E \mid \overline{C}) > 0$$

   - Of course, we would not detect causality between C and E via an intervention whereas the conditional probabilities would imply it
   - We need a suitable *formalism* to deal with interventions (*Pearl's do-calculus*)
3. **Causality as Explanation of a Process**

[4] D. R. Cox and N. Wermuth. Some Statistical Aspects of Causality. *European Sociological Review* 17(1):65-74, 2001.

Simpson's Paradox

Even more fun with causes, effects, and confounding variables:



Source: J. Pearl, M. Glymour, N. P. Jewell. *Causal Inference in Statistics - A Primer*. Wiley, 2016.

Statistical Models and Machine Learning

Statistical Models

- In general, a statistical model is a pair $(\mathcal{X}, \mathcal{P})$ of a sample/data space $\mathcal{X}$ and a (parametric) family $\mathcal{P}$ of distributions F_θ on $\mathcal{X}$

$$\mathcal{P} = \{F_\theta : \theta \in \Theta\}$$

- **Example:**

$$\mathcal{X} = \mathbb{R}, \qquad \mathcal{P} = \{\mathrm{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \in (0, \infty)\}$$

- We then try to estimate $\theta \in \Theta$ based on data/samples $x_1, \ldots, x_n \in \mathcal{X}$

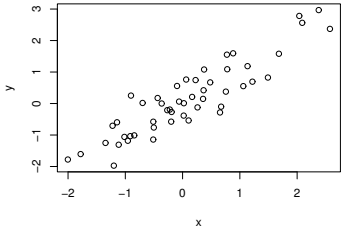
- Think about paired data (x_i, y_i). Can we infer which variable is the cause and which the effect ?

- **Claim:** A (plain) statistical model is not rich enough to express and infer causality, only association.

Linear Regression

- Assume we are given data (x_i, y_i), $i = 1, \ldots, 50$ from an underlying bivariate normal distribution

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sqrt{0.75} \\ \sqrt{0.75} & 1 \end{pmatrix} \right)$$

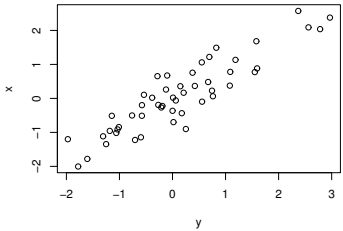


- How can we determine if X or Y is the explanatory variable? I.e., fit

$$Y = b + aX + \varepsilon, \qquad \varepsilon \sim \mathrm{N}(0, \sigma^2),$$

or

$$X = b + aY + \varepsilon, \qquad \varepsilon \sim \mathrm{N}(0, \sigma^2) \ ?$$



Linear Regression cont'd

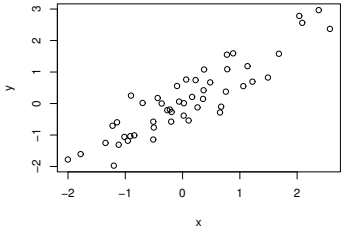
- In this example we can infer the parametric model

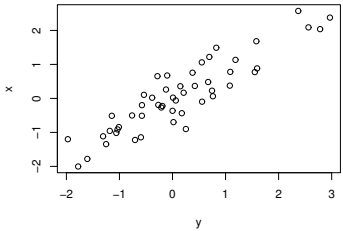
$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho_{XY} \\ \rho_{YX} & \sigma_Y^2 \end{pmatrix} \right)$$

- and, in particular, estimate the correlation

$$\mathrm{Corr}(X, Y) = \mathrm{Corr}(Y, X)$$

- But **without further information** we can not deduce which of the two variables responses to a change in the other.





Prediciton versus causal queries

- Given the bivariate setting we may be interested in properties of a (random) pair (X, Y)
- For regression

$$f(x) = \mathbb{E}[Y \mid X = x]$$

  and for (binary) classification,

$$f(x) = \mathbb{P}[Y = 1 \mid X = x]$$

- Such functions f serve as predictors for unknown Y given a new query point $X = x$ which was generated in a neutral way (without intervention)
- These predictors allow **not** to answer (reasonably) questions like

  *Will there be less heatstrokes in summer if **we forbid** selling ice cream?*

  (because the distribution of (X, Y) is only valid without (non-representative) interventions)

Machine learning and AI

- Machine learning, particularly, supervised (deep) learning is just basically regression for more elaborate (deep) functional models

$$f(x) \approx \mathbb{E}\left[Y \mid X = x\right]$$

- We learn via minimizing an empirical mismatch using a loss function ℓ

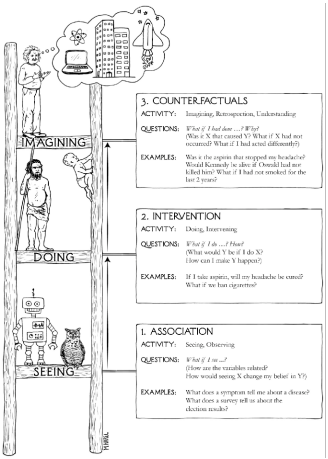
$$\frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i) \;\to\; \min_{f \in \mathcal{F}}$$

over a suitable function class $\mathcal{F}$ of regressors, e.g., neural networks

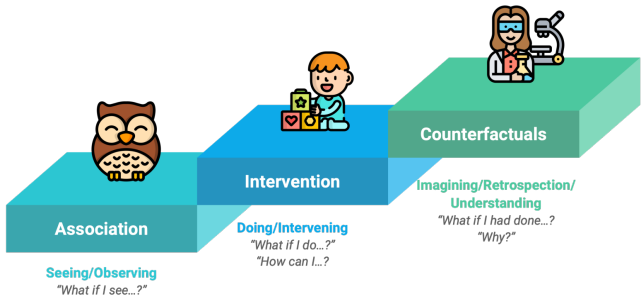
- Statistical learning theory tells us, how difficult this learning is, e.g., how many data points n we need to come close (if at all) to the best regressor $f^\star$ within $\mathcal{F}$

$$f^\star \in \operatorname*{argmin}_{f \in \mathcal{F}} \mathbb{E}\left[\ell(f(X), Y)\right]$$

The ladder of causation



Source: J. Pearl. *The Book of Why: The New Science of Cause and Effect*. Penguin, 2018.



Source: https://www.cajagroup.com

And then there was...

Source: https://openai.com

AMSTATNEWS

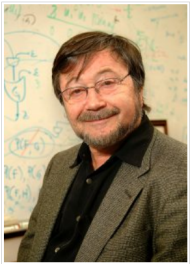
The Membership Magazine of the American Statistical Association

HOME ABOUT PDF ARCHIVES ADVERTISE STATISTICIANS IN HISTORY PODCAST

Home » Artificial Intelligence, Cover Story

Judea Pearl, AI, and Causality: What Role Do Statisticians Play?

1 SEPTEMBER 2023 5,045 VIEWS NO COMMENT



In the first half of 2023, the machine learning programs ChatGPT and GPT-4 changed the landscape of artificial intelligence research seemingly overnight. Judea Pearl's research bridges the subjects of statistics and artificial intelligence and highlights the importance of causality in both settings. Dana Mackenzie, Pearl's co-author for *The Book of Why*, interviews him here to get his take on recent developments. When they wrote their book in 2018, Pearl contended machine learning had not yet moved past the first rung of the "ladder of causation." Computers could not

Mackenzie: Can you tell me your first reactions to ChatGPT and GPT-4? Did you find their capabilities surprising?

Pearl: Aside from being impressed, I have had to reconsider my proof that one cannot get any answer to any causal or counterfactual query from observational studies. What I didn't take into account is the possibility that the text in the training database would itself contain causal information. The programs can simply cite information from the text without experiencing any of the underlying data.

Source: https://magazine.amstat.org/

CAN LARGE LANGUAGE MODELS INFER CAUSATION FROM CORRELATION?

Zhijing Jin[1,2,*,‡] **Jiarui Liu**[3,*] **Zhiheng Lyu**[4] **Spencer Poff**[5]
Mrinmaya Sachan[2] **Rada Mihalcea**[6] **Mona Diab**[3,‡,†] **Bernhard Schölkopf**[1,†]
[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany [2]ETH Zürich
[3]LTI, CMU [4]University of Hong Kong [5]Meta AI [6]University of Michigan
jinzhi@ethz.ch jiarui@cmu.edu zhihenglyu.cs@gmail.com

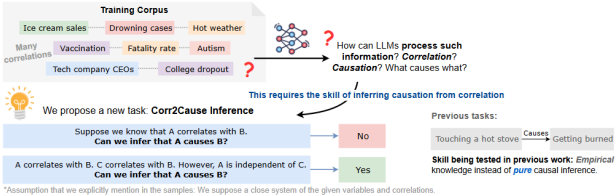


Figure 1: Illustration of the motivation behind our task and dataset.

ABSTRACT

Causal inference is one of the hallmarks of human intelligence. While the field of Causal NLP has attracted much interest in the recent years, existing causal inference datasets in NLP primarily rely on discovering causality from empirical knowledge (e.g., commonsense knowledge). In this work, we propose the first benchmark dataset to test the pure causal inference skills of large language models (LLMs). Specifically, we formulate a novel task CORR2CAUSE, which takes a set of correlational statements and determines the causal relationship between the variables. We curate a large-scale dataset of more than 200K samples, on which we evaluate seventeen existing LLMs. Through our experiments, we identify a key shortcoming of LLMs in terms of their causal inference skills, and show that these models achieve almost close to random performance on the task. This shortcoming is somewhat mitigated when we try to re-purpose LLMs for this skill via finetuning, but we find that these models still fail to generalize – they can only perform causal inference in in-distribution settings when variable names and textual expressions used in the queries are similar to those in the training set, but fail in out-of-distribution settings generated by perturbing these queries. CORR2CAUSE is a challenging task for LLMs, and can be helpful in guiding future research on improving LLMs' pure reasoning skills and generalizability.[1]

Source: Jin et al. Can Large Language Models Infer Causation from Correlation? (arXiv, 2024).

Structural Cause-Effect Models

Structural Causal Models

- Foundation for causal reasoning
- Entail a probability/statistical model but also additional information in form of a structure of dependencies between variables
- Form of structural equation models dating back to Sewall Wright (1889 – 1988) used, e.g., in econometrics and social sciences
- We focus on the simpelst form here with two variables C and E

Assumption underlying causal inference

Principle of independent mechanism

The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other.

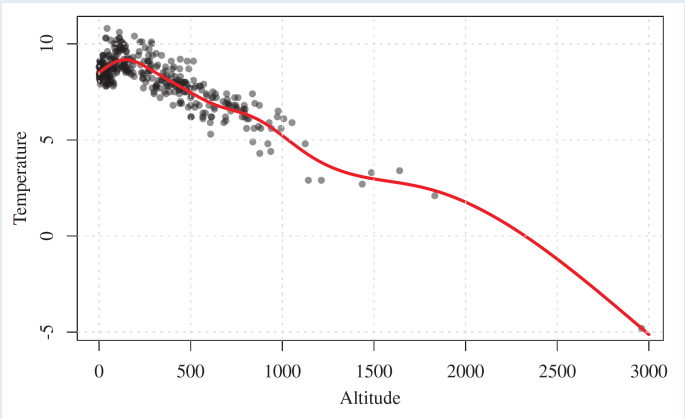
In the probabilistic case, this means that the conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other conditional distributions. In case we have only two variables, this reduces to an independence between the cause distribution and the mechanism producing the effect distribution.

In short:

Independence of cause and mechanism (ICM).

Example

Consider the (cor)relation between altitude and average temperature of cities



Source: J. Peters et al. *Elements of Causal Inference*. MIT Press, 2017.

- Again, we can ask: What's the cause and what's the effect?
- Consider the joint distribution (here: pdf) of altitude A and temperature T

$$\begin{aligned} p(a,t) &= p(a \mid t)\, p(t) \\ &= p(t \mid a)\, p(a) \end{aligned}$$

  which can be decomposed in conditional and marginal density
- The marginal density would correspond to the distribution of the cause
- whereas the conditional density corresponds to the distribution of the mechanism turing cause into effect
- Given the principle of independent mechanism we can now ask:

  Which mechanism ($a \mapsto t$ or $t \mapsto a$), i.e., which conditional distribution $p(t \mid a)$ or $p(a \mid t)$ remains invariant if we change the cause, i.e., marginal $p(a)$ or $p(t)$?

Equilibrium laws

Example: Physical laws for equilibira

- Consider the ideal gas law

$$p \cdot V = n \cdot R \cdot T$$

with pressure p, Volumne V, temperature T, ideal gas constant R, and amount of substance n.

- What's cause and effect here? I.e. changing any of p,V, R will effect the others.

Definition

A structural Cause-Effect model (SCEM) $\mathfrak{C}$ with graph

$$C \to E$$

consists of two assignments

$$C := N_C$$
$$E := f_E(C, N_E)$$

where $N_E \perp\!\!\!\perp N_C$ are independent "noise" random variables on (measurable) spaces $\mathcal{E}$ and $\mathcal{C}$, respectively, and $f_E \colon \mathcal{C} \to \mathcal{E}$ denoting a (measurable) cause-effect mechanism.

We call C a (direct) cause of the effect E.

Given distributions P_{N_C}, P_{N_E} for the "noises" (and f_E), a SCEM yields a joint distribution $P_{C,E}$ for the cause-effect pair (C, E).

Example

Consider the SCEM

$$C := N_C$$
$$E := 4 \cdot C + N_E$$

with $N_E, N_C \sim \mathrm{N}(0, 1)$ iid. Then

$$\begin{pmatrix} C \\ E \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 4 \\ 4 & 17 \end{pmatrix} \right)$$

Interventions

- An intervention is (usually) a change of (one of) the assignments in the SCEM
- which typicall yields a different distribution different from the observational (unintervened) distribution
- **Hard intervention:** Setting one of the two variables to a specific value, e.g.,

$$\mathrm{do}\,(E := 4)$$

- The resulting distribution of the other variables is then denoted by

$$P_C^{\mathrm{do}(E:=4)} = P_C^{\mathfrak{C},\mathrm{do}(E:=4)}$$

  and may differ from the conditional distribution $P_{C|E=4}$

- **Soft intervention:** Keeping a functional dependence, e.g.,

$$\mathrm{do}\left(E := g_E(C) + \tilde{N}_E\right)$$

Example

Consider the SCEM

$$C := N_C$$
$$E := 4 \cdot C + N_E$$

with $N_E, N_C \sim \mathrm{N}(0,1)$ iid. Then for any $x \in \mathbb{R}$

$$P_E \neq P_E^{\mathrm{do}(C:=x)} = \mathrm{N}(4x, 1) = P_{E|C=x}$$

but on the other hand

$$P_C = P_C^{\mathrm{do}(E:=x)} = P_{N_C} = \mathrm{N}(0,1) \neq P_{C|E=x}$$

This reseembles the roles of cause and effect: An intervention on E does not effect C, but on C does effect E.

Counterfactuals

- Modification of a SCEM by changing all of its noise distributions
- Again results in different distributions than the observational distribution

Example

Consider the following setting of an eye disease:

- for 99% of all affected patients the treatment curse the disease ($T = 1, B = 0$) whereas no treatment would yield blindness ($T = 0, B = 1$)
- but for 1% of the patients it is the other way round, i.e., treatment yields blindness whereas by not treatment the recover from the disease

Question: A patient has gone blind after treatment. *What would have happend had the doctor chosen not to treat the patient?*

Corresponding SCEM $\mathfrak{C}$

$$T := N_T$$
$$B := T \cdot N_B + (1 - T) \cdot (1 - N_B)$$

with $N_B \sim \text{Bernoulli}(0.01)$ describing to which of the two groups a patient belongs and N_T the decision on treatment.
To answer the counterfactual question, first condition $\mathfrak{C}$ on observation:

$$\mathfrak{C} \mid B = 1, T = 1 :$$
$$T := 1$$
$$B := T \cdot 1 + (1 - T) \cdot (1 - 1) = 1$$

i.e., we gained knowledge on $N_B = 1$ for the given patient. Then calculate the effect of the intervention

$$\text{do}\,(T = 0)\,.$$

The intervened conditioned SCEM

$$\mathfrak{C} \mid B = 1, T = 1; \mathrm{do}\,(T = 0):$$
$$T := 0$$
$$B := T \cdot 1 + (1 - T) \cdot (1 - 1) = 0$$

yields

$$P_B^{\mathfrak{C}|B=1,T=1;\mathrm{do}(T=0)} = \delta_0$$

i.e., the patient would have not gone blind with certainty (probability 1).

The SCEM provides a computational approach to answer counterfactual questions.

Learning an SCEM from Data

Question: Is the causal structure $\mathfrak{C}$ identifiable from the joint distribution $P^{\mathfrak{C}}_{C,E}$?

Proposition

For every joint distribution $P_{X,Y}$ of a pair (X, Y) of two real-valued random variables, there exists an SCEM

$$Y = f_Y(X, N_Y), \qquad X \perp\!\!\!\perp N_Y,$$

where $f_Y : \mathbb{R} \to \mathbb{R}$ is measurable and N_Y are real-valued noise variable.

Meaning: For $X = C$ and $Y = E$ exists an SCEM and also for $X = E$ and $Y = C$ which yield the same observational distribution $P_{X,Y}$.

Thus: Without additional assumptions the causal structure is not identifiable from data or joint distribution alone.

Summary