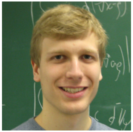
A Measure-Theoretic Axiomatisation of Causality


Junhyung Park[1]


Simon Buchholz[1]


Bernhard Schölkopf[1]


Krikamol Muandet[2]

[1]Max Planck Institute for Intelligent Systems, Tübingen

[2]CISPA Helmholtz Center for Information Security, Saarbrücken

NeurIPS 2023 (New Orleans)

Table of Contents

Table of Contents

Why is Causality Important?

Policy / drug evaluation

Why is Causality Important?

Policy / drug evaluation

Image Classification

Why is Causality Important?

Policy / drug evaluation

Law, blame

Image Classification

Why is Causality Important?

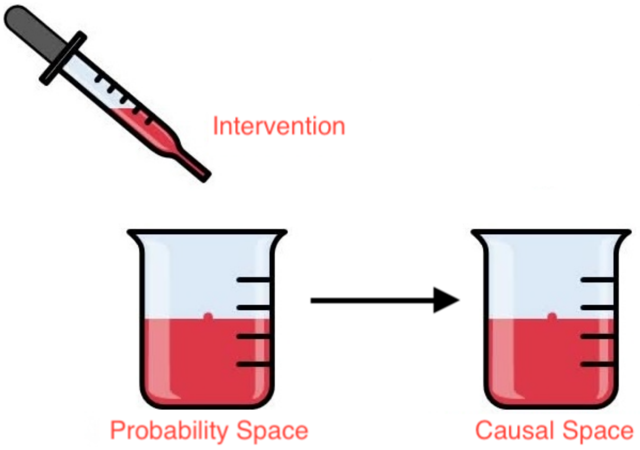
Policy / drug evaluation

Image Classification

Law, blame

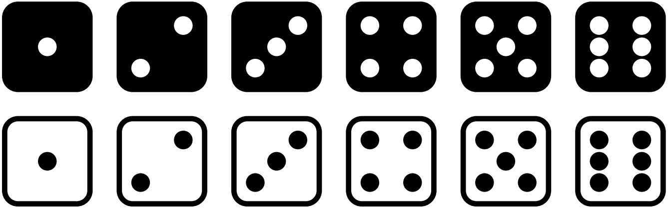
- Natural language processing
- Algorithmic recourse
- transfer learning
- Out of distribution generalisation
- Many more...

Plan



Probability Theory

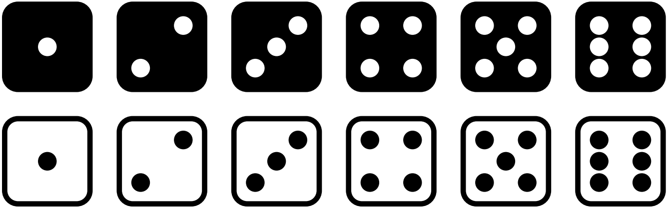
Probability Space



Foundations of the Theory of Probability, Andrei N Kolmogorov, 1933.

Probability Theory

Probability Space



$$(\Omega, \mathcal{H}, \mathbb{P})$$

Set of outcomes — σ-algebra of events — Probability measure

Foundations of the Theory of Probability, Andrei N Kolmogorov, 1933.

Probability vs Statistics

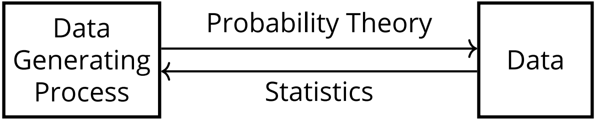


Figure: Statistics is an inverse problem of probability theory.

Probability vs Statistics

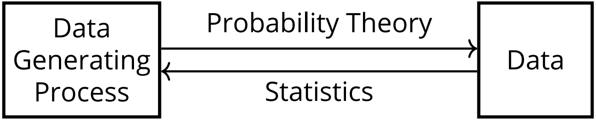


Figure: Statistics is an inverse problem of probability theory.

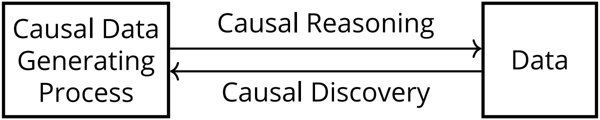
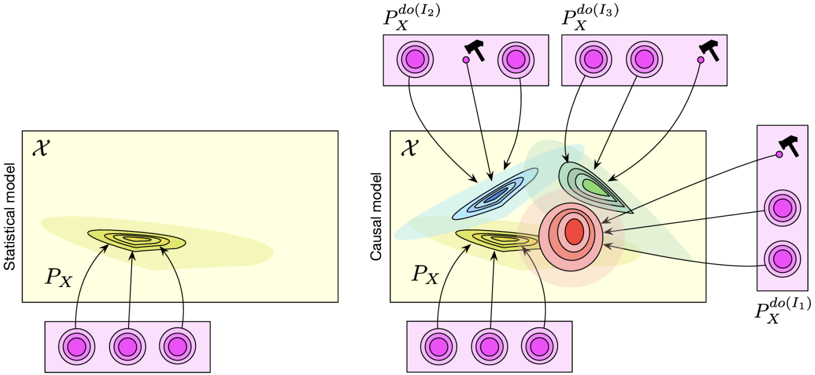


Figure: Causal discovery is an inverse problem of causal reasoning.

Manipulation is at the heart of Causality



We are interested in what happens to a system, when we intervene on a sub-system.

Towards Causal Representation Learning, Schölkopf, Locatello, Bauer, Ke, Kalchbrenner, Goyal, Bengio, Proceedings of the IEEE, 2021.

Table of Contents

Notations

- For a set T, we denote its power set by $\mathcal{P}(T)$.

Notations

- For a set T, we denote its power set by $\mathcal{P}(T)$.
- Product measurable space with index set T:

$$(\Omega, \mathcal{H}) = (\times_{t \in T} E_t, \otimes_{t \in T} \mathcal{E}_t).$$

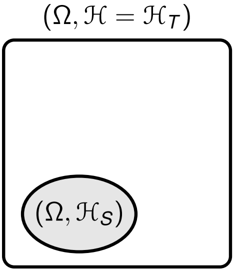
$\mathcal{H}_S$: sub-σ-algebra of $\mathcal{H}$ corresponding to $S \in \mathcal{P}(T)$.

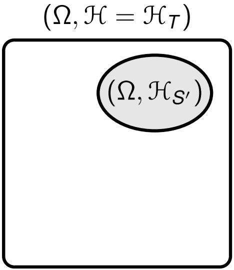
Notations

- For a set T, we denote its power set by $\mathcal{P}(T)$.
- Product measurable space with index set T:

$$(\Omega, \mathcal{H}) = (\times_{t \in T} E_t, \otimes_{t \in T} \mathcal{E}_t).$$

$\mathcal{H}_S$: sub-σ-algebra of $\mathcal{H}$ corresponding to $S \in \mathcal{P}(T)$.
Intuition: $\mathcal{H} = \mathcal{H}_T$ **is the entire space.** $\mathcal{H}_S$ **is a subspace.**





Notations

- For a set T, we denote its power set by $\mathcal{P}(T)$.
- Product measurable space with index set T:

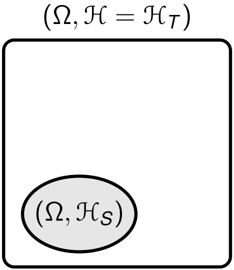
$$(\Omega, \mathcal{H}) = (\times_{t \in T} E_t, \otimes_{t \in T} \mathcal{E}_t).$$

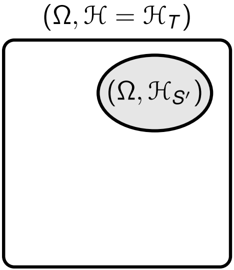
  $\mathcal{H}_S$: sub-σ-algebra of $\mathcal{H}$ corresponding to $S \in \mathcal{P}(T)$.
  Intuition: $\mathcal{H} = \mathcal{H}_T$ **is the entire space.** $\mathcal{H}_S$ **is a subspace.**

- "Transition probability kernel"
  K_S from $(\Omega, \mathcal{H}_S)$ into $(\Omega, \mathcal{H})$:

$$K_S(x, \cdot) \to [0, 1].$$

  For every $x \in (\Omega, \mathcal{H}_S)$, $K_S(x, \cdot)$ is a measure on $(\Omega, \mathcal{H})$.



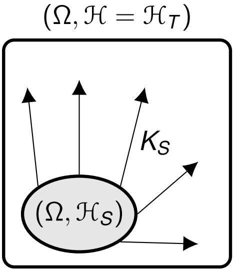


Notations

- For a set T, we denote its power set by $\mathcal{P}(T)$.
- Product measurable space with index set T:

$$(\Omega, \mathcal{H}) = (\times_{t \in T} E_t, \otimes_{t \in T} \mathcal{E}_t).$$

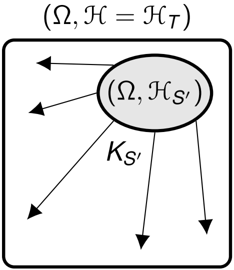
  $\mathcal{H}_S$: sub-σ-algebra of $\mathcal{H}$ corresponding to $S \in \mathcal{P}(T)$.
  Intuition: $\mathcal{H} = \mathcal{H}_T$ **is the entire space.** $\mathcal{H}_S$ **is a subspace.**



- "Transition probability kernel"
  K_S from $(\Omega, \mathcal{H}_S)$ into $(\Omega, \mathcal{H})$:

$$K_S(x, \cdot) \to [0, 1].$$

  For every $x \in (\Omega, \mathcal{H}_S)$, $K_S(x, \cdot)$ is a measure on $(\Omega, \mathcal{H})$.
  Intuition: conditional distribution.



Causal Spaces

A *causal space* is defined as the quadruple $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K})$, where

- $(\Omega, \mathcal{H}, \mathbb{P}) = (\times_{t \in T} E_t, \otimes_{t \in T} \mathcal{E}_t, \mathbb{P})$ is a probability space, and
- $\mathbb{K} = \{K_S : S \in \mathcal{P}(T)\}$ is a collection of transition probability kernels K_S from $(\Omega, \mathcal{H}_S)$ into $(\Omega, \mathcal{H})$, called the *causal kernel on* $\mathcal{H}_S$, such that
  1. for all $A \in \mathcal{H}$ and $x \in \Omega$,
  $$K_\emptyset(x, A) = \mathbb{P}(A);$$
  2. for all $A \in \mathcal{H}_S$ and $x \in \Omega$,
  $$K_S(x, A) = 1_A(x).$$

Causal Spaces

A *causal space* is defined as the quadruple $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K})$, where

- $(\Omega, \mathcal{H}, \mathbb{P}) = (\times_{t \in T} E_t, \otimes_{t \in T} \mathcal{E}_t, \mathbb{P})$ is a probability space, and
- $\mathbb{K} = \{K_S : S \in \mathcal{P}(T)\}$ is a collection of transition probability kernels K_S from $(\Omega, \mathcal{H}_S)$ into $(\Omega, \mathcal{H})$, called the *causal kernel on* $\mathcal{H}_S$, such that
  - (i) for all $A \in \mathcal{H}$ and $x \in \Omega$,
  $$K_\emptyset(x, A) = \mathbb{P}(A);$$
  - (ii) for all $A \in \mathcal{H}_S$ and $x \in \Omega$,
  $$K_S(x, A) = 1_A(x).$$

$\mathbb{P}$ is the “observational distribution”.

Interventions

An intervention is the process of

(a) choosing a sub-σ-algebra $\mathcal{H}_U$, and

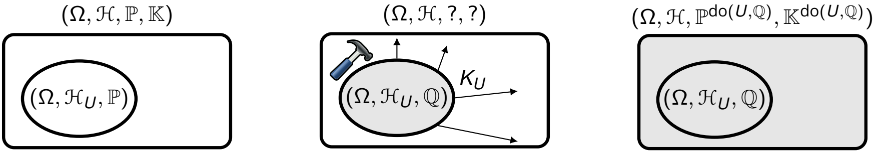
(b) placing any measure $\mathbb{Q}$ on $(\Omega, \mathcal{H}_U)$.

Then we have a new causal space $(\Omega, \mathcal{H}, \mathbb{P}^{\text{do}(U,\mathbb{Q})}, \mathbb{K}^{\text{do}(U,\mathbb{Q})})$, where

$$\mathbb{P}^{\text{do}(U,\mathbb{Q})}(A) = \int \mathbb{Q}(d\omega) K_U(\omega, A) \tag{1}$$

and $\mathbb{K}^{\text{do}(U,\mathbb{Q})} = \{K_S^{\text{do}(U,\mathbb{Q})} : S \in \mathcal{P}(T)\}$ with

$$K_S^{\text{do}(U,\mathbb{Q})}(\omega, A) = \int \mathbb{Q}(d\omega'_{U\setminus S}) K_{S\cup U}((\omega_S, \omega'_{U\setminus S}), A). \tag{2}$$



Causal Spaces

A *causal space* is defined as the quadruple $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K})$, where

- $(\Omega, \mathcal{H}, \mathbb{P}) = (\times_{t \in T} E_t, \otimes_{t \in T} \mathcal{E}_t, \mathbb{P})$ is a probability space, and
- $\mathbb{K} = \{K_S : S \in \mathcal{P}(T)\}$ is a collection of transition probability kernels K_S from $(\Omega, \mathcal{H}_S)$ into $(\Omega, \mathcal{H})$, called the *causal kernel on* $\mathcal{H}_S$, such that
  - (i) for all $A \in \mathcal{H}$ and $x \in \Omega$,
  $$K_\emptyset(x, A) = \mathbb{P}(A);$$
  - (ii) for all $A \in \mathcal{H}_S$ and $x \in \Omega$,
  $$K_S(x, A) = 1_A(x).$$

Intuition on the axioms:

Causal Spaces

A *causal space* is defined as the quadruple $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K})$, where

- $(\Omega, \mathcal{H}, \mathbb{P}) = (\times_{t \in T} E_t, \otimes_{t \in T} \mathcal{E}_t, \mathbb{P})$ is a probability space, and
- $\mathbb{K} = \{K_S : S \in \mathcal{P}(T)\}$ is a collection of transition probability kernels K_S from $(\Omega, \mathcal{H}_S)$ into $(\Omega, \mathcal{H})$, called the *causal kernel on* $\mathcal{H}_S$, such that
  - (i) for all $A \in \mathcal{H}$ and $x \in \Omega$,
  $$K_\emptyset(x, A) = \mathbb{P}(A);$$
  - (ii) for all $A \in \mathcal{H}_S$ and $x \in \Omega$,
  $$K_S(x, A) = 1_A(x).$$

Intuition on the axioms:

(i) $\mathbb{P}^{\text{do}(\emptyset, \mathbb{Q})}(A) = \mathbb{P}(A)$.

Causal Spaces

A *causal space* is defined as the quadruple $(\Omega, \mathcal{H}, \mathbb{P}, \mathbb{K})$, where

- $(\Omega, \mathcal{H}, \mathbb{P}) = (\times_{t \in T} E_t, \otimes_{t \in T} \mathcal{E}_t, \mathbb{P})$ is a probability space, and
- $\mathbb{K} = \{K_S : S \in \mathcal{P}(T)\}$ is a collection of transition probability kernels K_S from $(\Omega, \mathcal{H}_S)$ into $(\Omega, \mathcal{H})$, called the *causal kernel on* $\mathcal{H}_S$, such that
  - (i) for all $A \in \mathcal{H}$ and $x \in \Omega$,
  $$K_\emptyset(x, A) = \mathbb{P}(A);$$
  - (ii) for all $A \in \mathcal{H}_S$ and $x \in \Omega$,
  $$K_S(x, A) = 1_A(x).$$

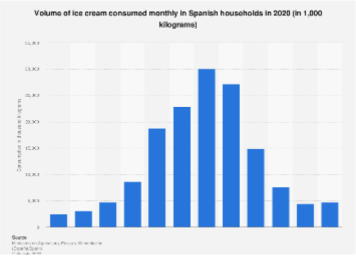
Intuition on the axioms:

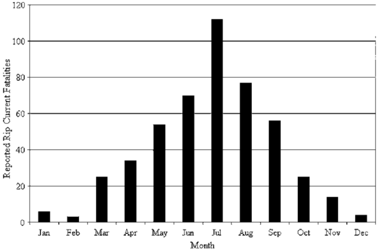
(i) $\mathbb{P}^{\text{do}(\emptyset, \mathbb{Q})}(A) = \mathbb{P}(A)$.

(ii) For $A \in \mathcal{H}_U$, $\mathbb{P}^{\text{do}(U, \mathbb{Q})}(A) = \int \mathbb{Q}(dx) 1_A(x) = \mathbb{Q}(A)$.

Table of Contents

Ice Cream Sales and Fatal Rip Current Accidents

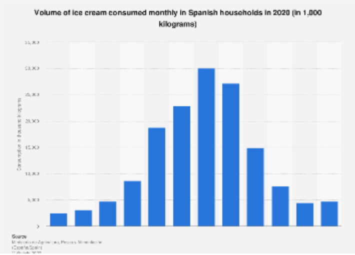


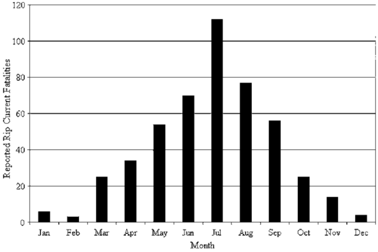


- Causal space: $(E_{\text{ice}} \times E_{\text{acc}}, \mathcal{E}_{\text{ice}} \otimes \mathcal{E}_{\text{acc}}, \mathbb{P}, \mathbb{K})$[1].

[1] $E_{\text{ice}} = E_{\text{acc}} = \mathbb{R}$ and $\mathcal{E}_{\text{ice}} = \mathcal{E}_{\text{acc}}$ is the Lebesgue σ-algebra.

Ice Cream Sales and Fatal Rip Current Accidents

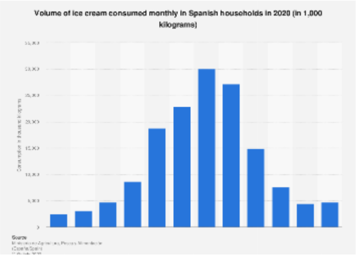


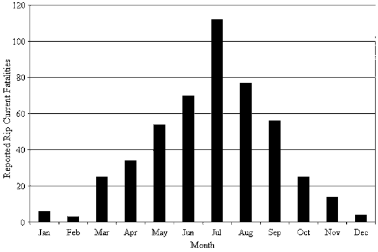


- Causal space: $(E_{\text{ice}} \times E_{\text{acc}}, \mathcal{E}_{\text{ice}} \otimes \mathcal{E}_{\text{acc}}, \mathbb{P}, \mathbb{K})$[1].
- $\mathbb{P}$ has strong correlation.

[1] $E_{\text{ice}} = E_{\text{acc}} = \mathbb{R}$ and $\mathcal{E}_{\text{ice}} = \mathcal{E}_{\text{acc}}$ is the Lebesgue σ-algebra.

Ice Cream Sales and Fatal Rip Current Accidents





- Causal space: $(E_{\text{ice}} \times E_{\text{acc}}, \mathcal{E}_{\text{ice}} \otimes \mathcal{E}_{\text{acc}}, \mathbb{P}, \mathbb{K})$[1].
- $\mathbb{P}$ has strong correlation.
- For causal kernels, let
  - $K_{\text{ice}}(x, A) = \mathbb{P}(A)$ for all $A \in \mathcal{E}_{\text{acc}}$; and
  - $K_{\text{acc}}(y, B) = \mathbb{P}(B)$ for all $B \in \mathcal{E}_{\text{ice}}$.

[1] $E_{\text{ice}} = E_{\text{acc}} = \mathbb{R}$ and $\mathcal{E}_{\text{ice}} = \mathcal{E}_{\text{acc}}$ is the Lebesgue σ-algebra.

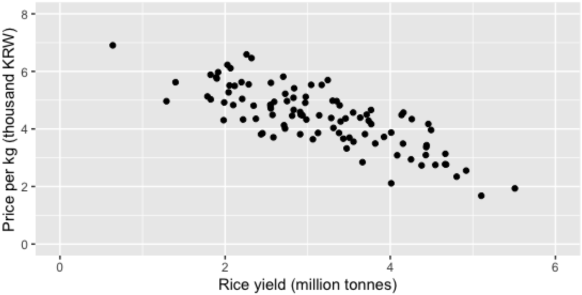
Crop Yield and Price

- Causal space: $(E_{\text{rice}} \times E_{\text{price}}, \mathcal{E}_{\text{rice}} \otimes \mathcal{E}_{\text{price}}, \mathbb{P}, \mathbb{K})$[2].

[2] $E_{\text{rice}} = E_{\text{price}} = \mathbb{R}$ and $\mathcal{E}_{\text{rice}} = \mathcal{E}_{\text{price}}$ is the Lebesgue σ-algebra.

Crop Yield and Price

- Causal space: $(E_{\text{rice}} \times E_{\text{price}}, \mathcal{E}_{\text{rice}} \otimes \mathcal{E}_{\text{price}}, \mathbb{P}, \mathbb{K})^2$.
- Without any intervention, the higher the yield, the more rice there is in the market, and lower the price.



[2] $E_{\text{rice}} = E_{\text{price}} = \mathbb{R}$ and $\mathcal{E}_{\text{rice}} = \mathcal{E}_{\text{price}}$ is the Lebesgue σ-algebra.

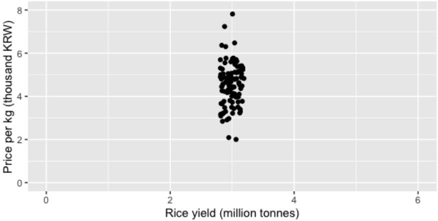
Crop Yield and Price

- If the government intervenes by buying up extra rice or releasing rice into the market from its stock, with the goal of stabilising supply at 3 million tonnes, then the price will stabilise accordingly.

Crop Yield and Price

- If the government intervenes by buying up extra rice or releasing rice into the market from its stock, with the goal of stabilising supply at 3 million tonnes, then the price will stabilise accordingly.
- The corresponding causal kernel at rice $= 3$ for $A \in \mathcal{E}_{\text{price}}$:

$$K_{\text{rice}}(3, A) = \int_A \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-4.5)^2} dx.$$



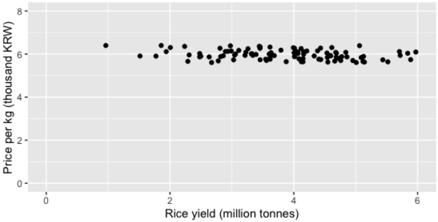
Crop Yield and Price

- If, instead, the government fixes the price of rice at a high price, say 6 thousand Korean Won per kg, then the farmers will be incentivised to produce more.

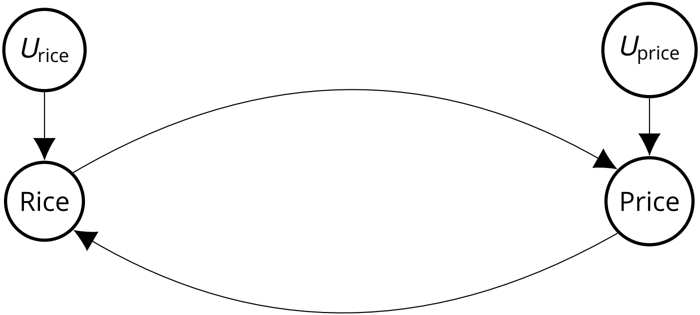
Crop Yield and Price

- If, instead, the government fixes the price of rice at a high price, say 6 thousand Korean Won per kg, then the farmers will be incentivised to produce more.
- The corresponding causal kernel at price $= 6$ for $B \in \mathcal{E}_{\text{rice}}$:

$$K_{\text{price}}(6, B) = \int_B \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-4)^2} dx.$$



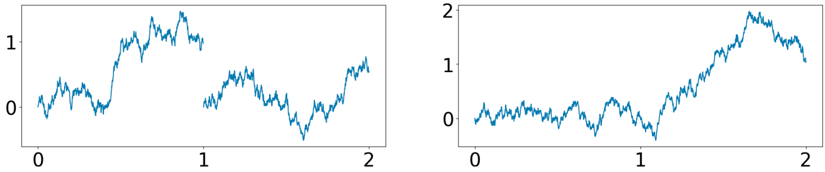
Crop Yield and Price



$$\text{Rice} = f_{\text{rice}}(\text{Price}, U_{\text{rice}}), \qquad \text{Price} = f_{\text{price}}(\text{Rice}, U_{\text{price}})$$

There may not be an observational distribution that is consistent with the structural equations, or there might be many of them.

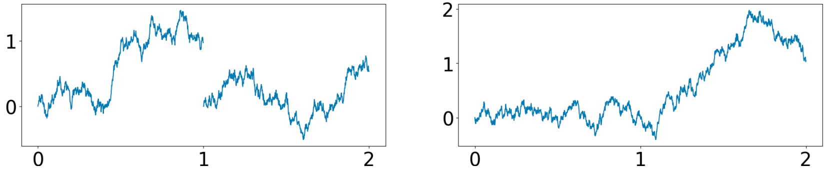
1-dimensional Brownian Motion



- Causal space: $(\times_{t\in\mathbb{R}_+} E_t, \otimes_{t\in\mathbb{R}_+} \mathcal{E}_t, \mathbb{P}, \mathbb{K})$[3].

[3] For each $t \in \mathbb{R}_+$, $E_t = \mathbb{R}$ and $\mathcal{E}_t$ is the Lebesgue σ-algebra.

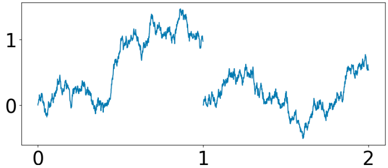
1-dimensional Brownian Motion

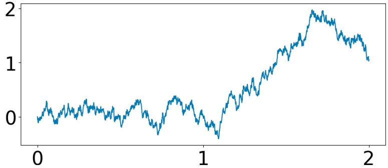


- Causal space: $(\times_{t\in\mathbb{R}_+} E_t, \otimes_{t\in\mathbb{R}_+}\mathcal{E}_t, \mathbb{P}, \mathbb{K})^3$.
- $\mathbb{P}$ is the Wiener measure.

[3] For each $t \in \mathbb{R}_+$, $E_t = \mathbb{R}$ and $\mathcal{E}_t$ is the Lebesgue σ-algebra.

1-dimensional Brownian Motion



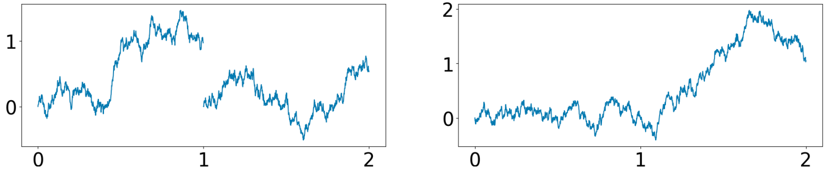


- Causal space: $(\times_{t\in\mathbb{R}_+} E_t, \otimes_{t\in\mathbb{R}_+} \mathcal{E}_t, \mathbb{P}, \mathbb{K})^3$.
- $\mathbb{P}$ is the Wiener measure.
- For any $s < t$, the causal kernels are

$$K_s(x, y) = \frac{1}{\sqrt{2\pi(t-s)}} e^{-\frac{1}{2(t-s)}(y-x)^2}, \qquad K_t(x, y) = \frac{1}{\sqrt{2\pi s}} e^{-\frac{1}{2s}y^2}.$$

[3] For each $t \in \mathbb{R}_+$, $E_t = \mathbb{R}$ and $\mathcal{E}_t$ is the Lebesgue σ-algebra.

1-dimensional Brownian Motion



- Causal space: $(\times_{t\in\mathbb{R}_+} E_t, \otimes_{t\in\mathbb{R}_+}\mathcal{E}_t, \mathbb{P}, \mathbb{K})^3$.
- $\mathbb{P}$ is the Wiener measure.
- For any $s < t$, the causal kernels are

$$K_s(x,y) = \frac{1}{\sqrt{2\pi(t-s)}} e^{-\frac{1}{2(t-s)}(y-x)^2}, \qquad K_t(x,y) = \frac{1}{\sqrt{2\pi s}} e^{-\frac{1}{2s}y^2}.$$

Past values affect the future, but future values do not affect the past.

[3] For each $t \in \mathbb{R}_+$, $E_t = \mathbb{R}$ and $\mathcal{E}_t$ is the Lebesgue σ-algebra.

Table of Contents

Summary

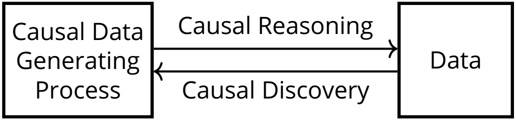
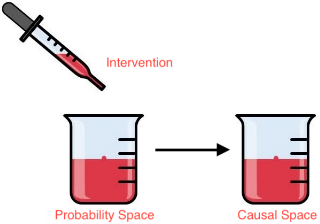


Figure: Causal discovery is an inverse problem of causality reasoning.



- We focused on the forwards direction, and proposed *causal spaces* by endowing probability spaces with causal kernels.

Summary

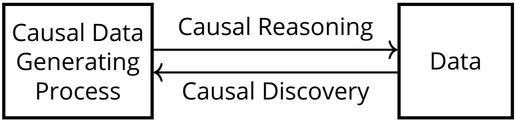
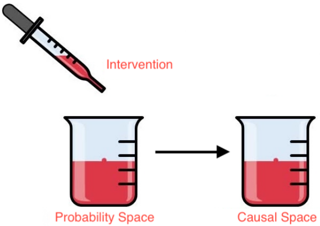


Figure: Causal discovery is an inverse problem of causality reasoning.



- We focused on the forwards direction, and proposed *causal spaces* by endowing probability spaces with causal kernels.
- Causal spaces strictly generalise existing frameworks, while elegantly overcoming some of their drawbacks, such as hidden confounders, cycles and continuous time stochastic processes.

Summary

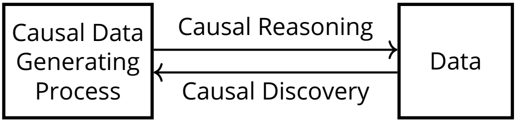
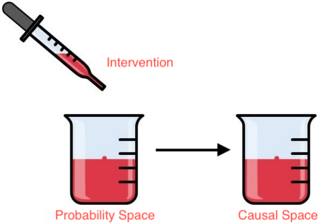


Figure: Causal discovery is an inverse problem of causality reasoning.



- We focused on the forwards direction, and proposed *causal spaces* by endowing probability spaces with causal kernels.
- Causal spaces strictly generalise existing frameworks, while elegantly overcoming some of their drawbacks, such as hidden confounders, cycles and continuous time stochastic processes.
- In the backwards direction, assumptions are unavoidable. The value of a framework is how well and naturally the assumptions can be expressed. For that, existing frameworks excel.