Towards a Measure-Theoretic Axiomatisation of Causality

Junhyung Park Simon Buchholz Bernhard Schölkopf Krikamol Muandet

Max Planck Institute for Intelligent Systems, Tübingen

CISPA

Colloquium “When Causal Inference Meets Statistical Analysis”
Paris, France
20th April, 2023

Table of Contents

Table of Contents

Probability Theory

- Probability theory is a *mathematisation* of the concept of *randomness* (or *stochasticity*).

[1] *Foundations of the Theory of Probability*, Andrei N Kolmogorov, 1933

Probability Theory

- Probability theory is a *mathematisation* of the concept of *randomness* (or *stochasticity*).
- There is a *universally accepted* axiomatisation[1] of probability theory based on measure theory.

[1] *Foundations of the Theory of Probability*, Andrei N Kolmogorov, 1933

Probability Theory

- Probability theory is a *mathematisation* of the concept of *randomness* (or *stochasticity*).
- There is a *universally accepted* axiomatisation[1] of probability theory based on measure theory.
- A probability space is a triple $(\Omega, \mathcal{H}, \mathbb{P})$, where:
  1. Ω is a set of *outcomes*;
  2. $\mathcal{H}$ is a collection of *events* forming a *σ-algebra*, i.e. a non-empty collection of subsets of Ω such that
     - $\Omega \in \mathcal{H}$;
     - if $A \in \mathcal{H}$, then $\Omega \setminus A \in \mathcal{H}$;
     - if $A_1, A_2, \ldots \in \mathcal{H}$, then $\cup_n A_n \in \mathcal{H}$;
  3. $\mathbb{P}$ is a *probability measure* on $(\Omega, \mathcal{H})$, i.e. a function $\mathbb{P} : \mathcal{H} \to [0, 1]$ satisfying
     - $\mathbb{P}(\emptyset) = 0$;
     - $\mathbb{P}(\cup_n A_n) = \sum_n \mathbb{P}(A_n)$ for any disjoint sequence (A_n) in $\mathcal{H}$;
     - $\mathbb{P}(\Omega) = 1$.

[1] *Foundations of the Theory of Probability*, Andrei N Kolmogorov, 1933

Probability Theory

Dice Roll

Example

- Outcomes: $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Probability Theory

Dice Roll

Example

- Outcomes: $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- "Probability of rolling a one":

$$\mathbb{P}(\{1\}) = \frac{1}{6}.$$

Probability Theory

Dice Roll

Example

- Outcomes: $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- "Probability of rolling a one":

$$\mathbb{P}(\{1\}) = \frac{1}{6}.$$

- "Probability of rolling an even number":

$$\mathbb{P}(\{2, 4, 6\}) = \frac{1}{2}.$$

Probability vs Statistics

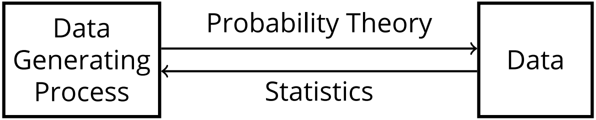


Figure: Statistics is an inverse problem of probability theory.

Probability vs Statistics

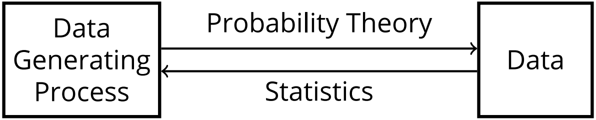


Figure: Statistics is an inverse problem of probability theory.

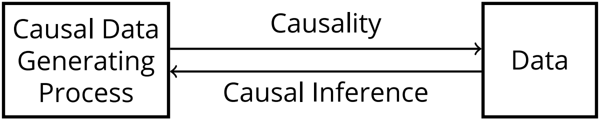


Figure: Causal inference is an inverse problem of causality.

Probability vs Statistics

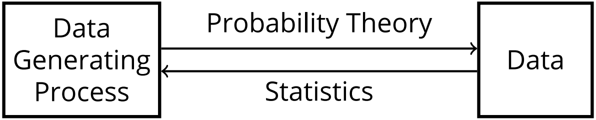


Figure: Statistics is an inverse problem of probability theory.

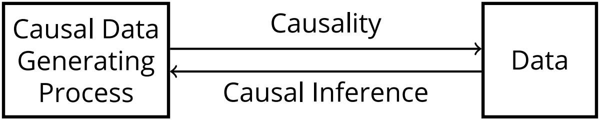
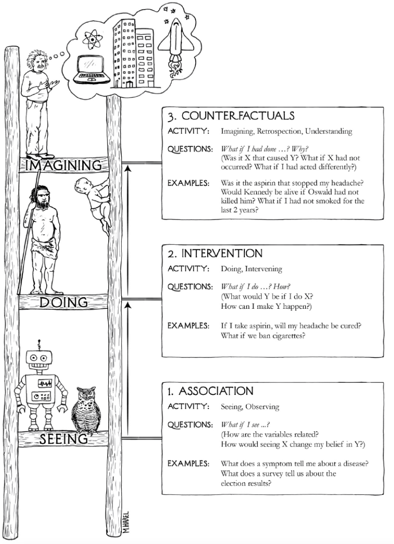


Figure: Causal inference is an inverse problem of causality.

Probability theory is not rich enough to capture causal concepts. But it may give us hints about how to axiomatise the concept of “causality”, which is our goal.

Pearl's Ladder of Causation



[1] *The Book of Why: The New Science of Cause and Effect*, Pearl and MacKenzie, 2018

Pearl's Ladder of Causation

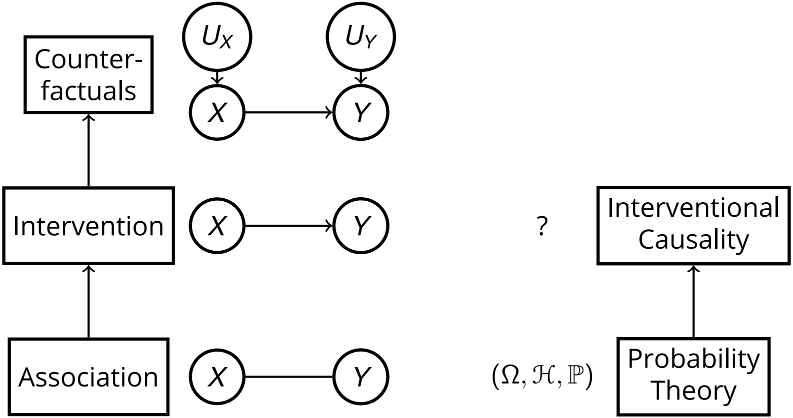


Figure: Primitive objects in each rung of the ladder.

Manipulation is at the heart of Causality

"Descriptive knowledge, by contrast, is knowledge that, although it may provide a basis for prediction, classification, or more or less unified representation or systemisation, does not provide information potentially relevant to manipulation. It is in this that the fundamental contrast between causal explanation and description consists. On this way of looking at matters, our interest in causal relationships and explanation initially grows out of a highly practical interest human beings have in manipulation and control; it is then extended to contexts in which manipulation is no longer a practical possibility[2]."

[2] Making Things Happen: A Theory of Causal Explanation, Woodward, 2005

Manipulation is at the heart of Causality

"Descriptive knowledge, by contrast, is knowledge that, although it may provide a basis for prediction, classification, or more or less unified representation or systemisation, does not provide information potentially relevant to manipulation. It is in this that the fundamental contrast between causal explanation and description consists. On this way of looking at matters, our interest in causal relationships and explanation initially grows out of a highly practical interest human beings have in manipulation and control; it is then extended to contexts in which manipulation is no longer a practical possibility[2]."

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

[2] Making Things Happen: A Theory of Causal Explanation, Woodward, 2005

Table of Contents

Causal Spaces

Definition

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 (product) probability space and $\mathbb{K} = \{K_S : S \in \mathcal{P}(T)\}$, called the *causal mechanism*, 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 $\omega \in \Omega$,

$$K_\emptyset(\omega, A) = \mathbb{P}(A);$$

(ii) for all $A \in \mathcal{H}_S$ and $B \in \mathcal{H}$,

$$K_S(\omega, A \cap B) = 1_A(\omega) K_S(\omega, B).$$

Causal Spaces

Definition

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 (product) probability space and $\mathbb{K} = \{K_S : S \in \mathcal{P}(T)\}$, called the *causal mechanism*, 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 $\omega \in \Omega$,

$$K_\emptyset(\omega, A) = \mathbb{P}(A);$$

(ii) for all $A \in \mathcal{H}_S$ and $B \in \mathcal{H}$,

$$K_S(\omega, A \cap B) = 1_A(\omega) K_S(\omega, B).$$

- $\mathbb{P}$ is the "observational distribution".

Causal Spaces

Definition

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 (product) probability space and $\mathbb{K} = \{K_S : S \in \mathcal{P}(T)\}$, called the *causal mechanism*, 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 $\omega \in \Omega$,

$$K_\emptyset(\omega, A) = \mathbb{P}(A);$$

(ii) for all $A \in \mathcal{H}_S$ and $B \in \mathcal{H}$,

$$K_S(\omega, A \cap B) = 1_A(\omega) K_S(\omega, B).$$

- $\mathbb{P}$ is the "observational distribution".
- Causal kernels K_S encode the causal information.

Causal Spaces

Definition

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 (product) probability space and $\mathbb{K} = \{K_S : S \in \mathcal{P}(T)\}$, called the *causal mechanism*, 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 $\omega \in \Omega$,

$$K_\emptyset(\omega, A) = \mathbb{P}(A);$$

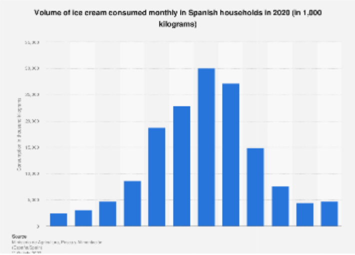
(ii) for all $A \in \mathcal{H}_S$ and $B \in \mathcal{H}$,

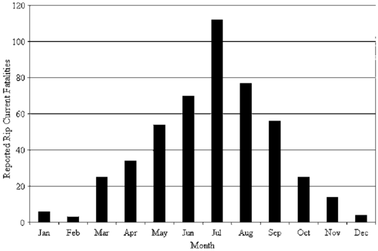
$$K_S(\omega, A \cap B) = 1_A(\omega) K_S(\omega, B).$$

- $\mathbb{P}$ is the "observational distribution".
- Causal kernels K_S encode the causal information.
- For each $\omega \in \Omega$, $K_S(\omega, \cdot)$ is a probability measure on $(\Omega, \mathcal{H})$.

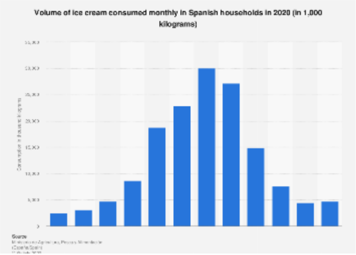
Table of Contents

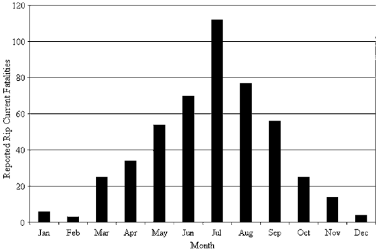
Ice Cream Sales and Fatal Rip Current Accidents





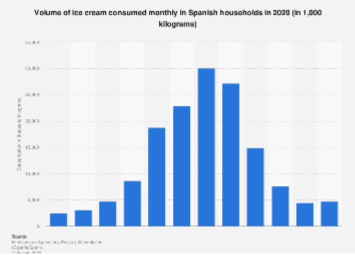
Ice Cream Sales and Fatal Rip Current Accidents

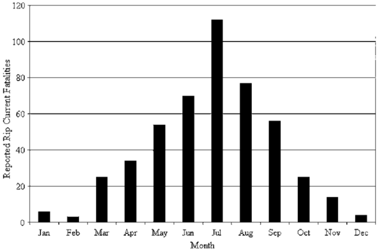




- We take $(E_{\text{ice}} \times E_{\text{acc}}, \mathcal{E}_{\text{ice}} \otimes \mathcal{E}_{\text{acc}}, \mathbb{P}, \mathbb{K})$, where $E_{\text{ice}} = E_{\text{acc}} = \mathbb{R}$ is the set of real numbers, $\mathcal{E}_{\text{ice}} = \mathcal{E}_{\text{acc}}$ is the Lebesgue σ-algebra and $\mathbb{P}$ is a probability measure with strong correlation.

Ice Cream Sales and Fatal Rip Current Accidents





- We take $(E_{\text{ice}} \times E_{\text{acc}}, \mathcal{E}_{\text{ice}} \otimes \mathcal{E}_{\text{acc}}, \mathbb{P}, \mathbb{K})$, where $E_{\text{ice}} = E_{\text{acc}} = \mathbb{R}$ is the set of real numbers, $\mathcal{E}_{\text{ice}} = \mathcal{E}_{\text{acc}}$ is the Lebesgue σ-algebra and $\mathbb{P}$ is a probability measure with strong correlation.
- For causal kernels, let $K_{\text{ice}}(x, A) = \mathbb{P}(A)$ for any $A \in \mathcal{E}_{\text{acc}}$ and $K_{\text{acc}}(x, B) = \mathbb{P}(B)$ for any $B \in \mathcal{E}_{\text{ice}}$.

Ice Cream Sales and Fatal Rip Current Accidents

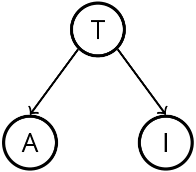
Figure: Correlation but no causation between ice-cream sales and rip current accidents. A stands for the number of fatal rip current accidents, I for ice cream sales, T for temperature, E for economy and W for world.

(a)

Ice Cream Sales and Fatal Rip Current Accidents

Figure: Correlation but no causation between ice-cream sales and rip current accidents. A stands for the number of fatal rip current accidents, I for ice cream sales, T for temperature, E for economy and W for world.

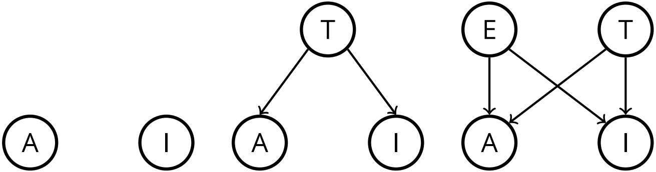
(a)



(b)

Ice Cream Sales and Fatal Rip Current Accidents

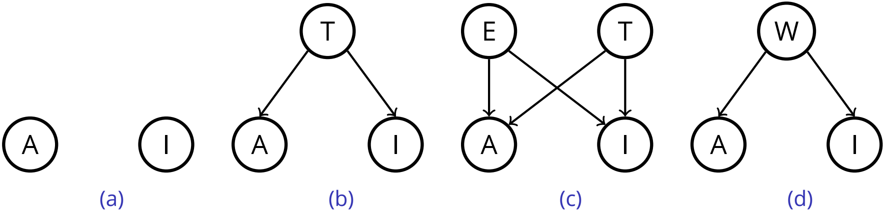
Figure: Correlation but no causation between ice-cream sales and rip current accidents. A stands for the number of fatal rip current accidents, I for ice cream sales, T for temperature, E for economy and W for world.



(a) (b) (c)

Ice Cream Sales and Fatal Rip Current Accidents

Figure: Correlation but no causation between ice-cream sales and rip current accidents. A stands for the number of fatal rip current accidents, I for ice cream sales, T for temperature, E for economy and W for world.



Crop Yield and Price

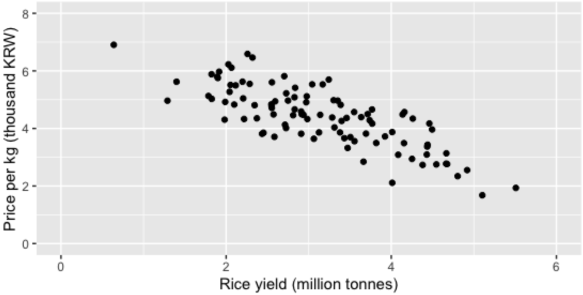
Example

- We take $(E_{\text{rice}} \times E_{\text{price}}, \mathcal{E}_{\text{rice}} \otimes \mathcal{E}_{\text{price}}, \mathbb{P}, \mathbb{K})$, where $E_{\text{rice}} = E_{\text{price}} = \mathbb{R}$, $\mathcal{E}_{\text{rice}} = \mathcal{E}_{\text{price}}$ is the Lebesgue σ-algebra and $\mathbb{P}$ is the observational distribution, for simplicity assumed to be jointly Gaussian.

Crop Yield and Price

Example

- We take $(E_{\text{rice}} \times E_{\text{price}}, \mathcal{E}_{\text{rice}} \otimes \mathcal{E}_{\text{price}}, \mathbb{P}, \mathbb{K})$, where $E_{\text{rice}} = E_{\text{price}} = \mathbb{R}$, $\mathcal{E}_{\text{rice}} = \mathcal{E}_{\text{price}}$ is the Lebesgue σ-algebra and $\mathbb{P}$ is the observational distribution, for simplicity assumed to be jointly Gaussian.
- Without any intervention, the higher the yield, the more rice there is in the market, and lower the price.

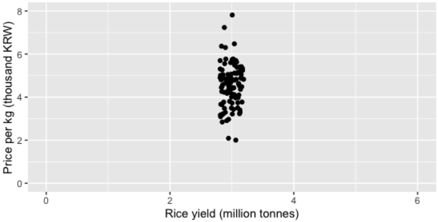


Crop Yield and Price

Example

- 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 $= 3$ will again be Gaussian, say with mean 4.5 and variance 1:

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

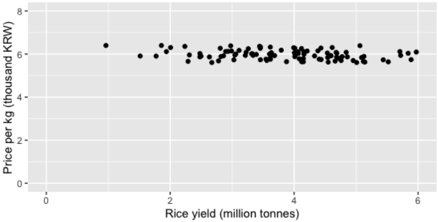


Crop Yield and Price

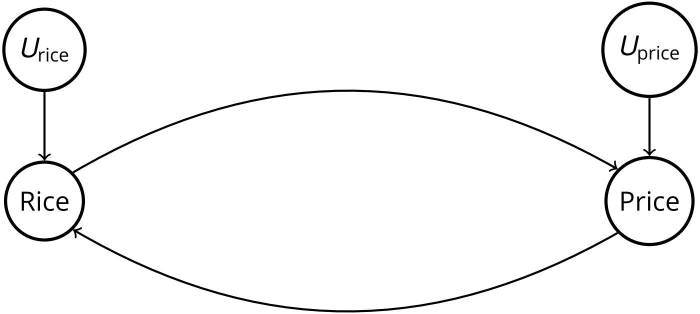
Example

- 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 $=6$ will again be Gaussian, say with mean 4 and variance 1:

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



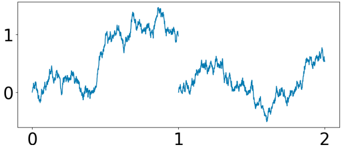
Crop Yield and Price

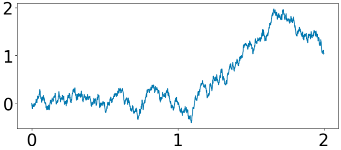


$$\text{Rice} = f_{\text{rice}}(\text{Price}, U_{\text{rice}}), \qquad \text{Price} = f_{\text{price}}(\text{Rice}, U_{\text{price}})$$

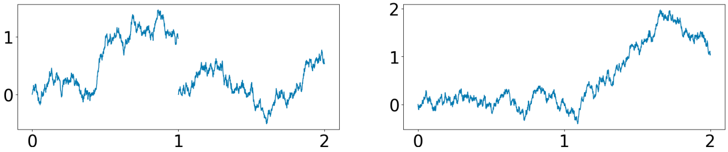
There may not be an observational distribution that is consistent with the structural equations, or there might be many of them.

1-dimensional Brownian Motion



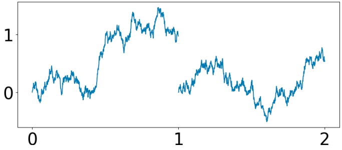


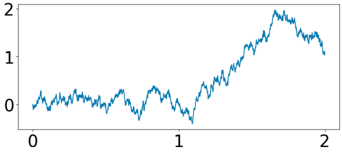
1-dimensional Brownian Motion



- Take $(\times_{t\in\mathbb{R}_+} E_t, \otimes_{t\in\mathbb{R}_+}\mathcal{E}_t, \mathbb{P}, \mathbb{K})$, where, for each $t \in \mathbb{R}_+$, $E_t = \mathbb{R}$ and $\mathcal{E}_t$ is the Lebesgue σ-algebra, and $\mathbb{P}$ is the Wiener measure.

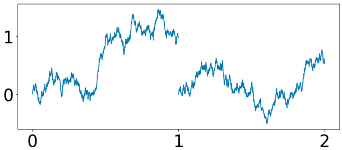
1-dimensional Brownian Motion

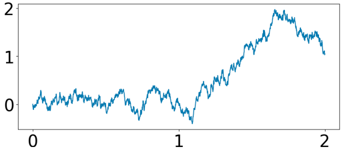




- Take $(\times_{t\in\mathbb{R}_+} E_t, \otimes_{t\in\mathbb{R}_+} \mathcal{E}_t, \mathbb{P}, \mathbb{K})$, where, for each $t \in \mathbb{R}_+$, $E_t = \mathbb{R}$ and $\mathcal{E}_t$ is the Lebesgue σ-algebra, and $\mathbb{P}$ is the Wiener measure.
- For any $s < t$, we have causal kernels $K_s(x, y) = \frac{1}{\sqrt{2\pi(t-s)}} e^{-\frac{1}{2(t-s)}(y-x)^2}$ and $K_t(x, y) = \frac{1}{\sqrt{2\pi s}} e^{-\frac{1}{2s}y^2}$.
  "Past values affect the future, but future values do not affect the past."

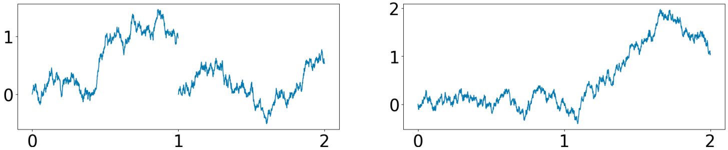
1-dimensional Brownian Motion





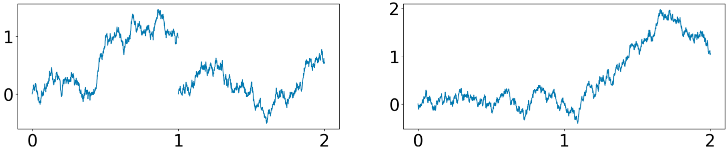
- In Markov continuous-time stochastic processes, no time point has a “parent”, since, between any two time points, there are infinitely many time points.

1-dimensional Brownian Motion



- In Markov continuous-time stochastic processes, no time point has a "parent", since, between any two time points, there are infinitely many time points.
- Since SCMs are explicitly dependent on parents, continuous time stochastic processes cannot be expressed via SCMs, or DAGs.

1-dimensional Brownian Motion



- In Markov continuous-time stochastic processes, no time point has a "parent", since, between any two time points, there are infinitely many time points.
- Since SCMs are explicitly dependent on parents, continuous time stochastic processes cannot be expressed via SCMs, or DAGs.
- Brownian motion is not differentiable, so no approach based on dynamical systems is applicable.

Further Comparisons with SCMs

How is the causal information encoded?

- In SCMs, causal information is encoded in the structural equations, $X_j := f_j(\mathbf{PA}_j, N_j)$, $j = 1, ..., d$.
- What is encoded here is the causal effect on the *subsystem* X_j of the *whole system*, i.e. the rest of the variables $X_1, ..., X_{j-1}, X_{j+1}, ..., X_d$.

Further Comparisons with SCMs

How is the causal information encoded?

- In SCMs, causal information is encoded in the structural equations, $X_j := f_j(\mathbf{PA}_j, N_j)$, $j = 1, ..., d$.
- What is encoded here is the causal effect on the *subsystem* X_j of the *whole system*, i.e. the rest of the variables $X_1, ..., X_{j-1}, X_{j+1}, ..., X_d$.
- In causal spaces, causal information is encoded in the causal kernels, $K_S : S \in \mathcal{P}(T)$.
- What is encoded here is the causal effect on the *whole system* $(\Omega, \mathcal{H})$ of the *subsystem*, $(\Omega, \mathcal{H}_S)$.

What are the primitive objects?

SCMs

1. Variables (or nodes) $X_1, \dots, X_d$
2. Structural equations $f_j(\mathbf{PA}_j, N_j), j = 1, \dots, d$
3. Noise distribution $P_{\mathbf{N}}$

What are the primitive objects?

SCMs

1. Variables (or nodes) $X_1, ..., X_d$
2. Structural equations $f_j(\mathbf{PA}_j, N_j), j = 1, ..., d$
3. Noise distribution $P_{\mathbf{N}}$

- Good interpretability.
- Only finite number of variables considered, and latent variables and cycles are not allowed.
- Existence and uniqueness of distribution not guaranteed without acyclicity.

What are the primitive objects?

SCMs

1. Variables (or nodes) $X_1, ..., X_d$
2. Structural equations $f_j(\mathbf{PA}_j, N_j), j = 1, ..., d$
3. Noise distribution $P_{\mathbf{N}}$

- Good interpretability.
- Only finite number of variables considered, and latent variables and cycles are not allowed.
- Existence and uniqueness of distribution not guaranteed without acyclicity.

Causal Spaces

1. Probability space $(\Omega, \mathcal{H}, \mathbb{P})$
2. Causal kernels K_S, $S \in \mathcal{P}(T)$

What are the primitive objects?

SCMs

1. Variables (or nodes) $X_1, ..., X_d$
2. Structural equations $f_j(\mathbf{PA}_j, N_j), j = 1, ..., d$
3. Noise distribution $P_{\mathbf{N}}$

- Good interpretability.
- Only finite number of variables considered, and latent variables and cycles are not allowed.
- Existence and uniqueness of distribution not guaranteed without acyclicity.

Causal Spaces

1. Probability space $(\Omega, \mathcal{H}, \mathbb{P})$
2. Causal kernels K_S, $S \in \mathcal{P}(T)$

- No interpretability.
- No restrictions on distribution and causal interactions between variables.
- Existence and uniqueness of observational and interventional distributions always guaranteed.

Summary

- Causality is an important concept in many research domains, but while many competing frameworks exist, there is no universally agreed axiomatisation of it, and existing frameworks are not general enough to express all possible distributions and causal interactions.

Summary

- Causality is an important concept in many research domains, but while many competing frameworks exist, there is no universally agreed axiomatisation of it, and existing frameworks are not general enough to express all possible distributions and causal interactions.
- Viewing causality as an extension of probability theory, and taking interventions as a central idea, we proposed an axiomatisation of causality based on measure theory.

Summary

- Causality is an important concept in many research domains, but while many competing frameworks exist, there is no universally agreed axiomatisation of it, and existing frameworks are not general enough to express all possible distributions and causal interactions.
- Viewing causality as an extension of probability theory, and taking interventions as a central idea, we proposed an axiomatisation of causality based on measure theory.
- It is important to stress that existing frameworks such as SCMs or potential outcomes are great for what they are set out to do, namely identification from observational data, for which assumptions are unavoidable. Our goal is not to replace existing frameworks.