Introduction to Causal Modeling

Session 2: Causal Discovery

Oisín Ryan

December 7, 2021

This workshop

November: Causal Inference.

- Given our beliefs about the system we are studying (i.e., our causal model), can we estimate a causal effect from data? If so, how?

Today: Causal Discovery/Learning

- Given data, but no strong beliefs about the causal system, can we learn the causal model?

Outline

1. Recap and Q&A
2. Why Causal Discovery?
3. Constraint-Based Methods
   - Hands-on Part 1
4. Restricted SCMs
5. Invariant Causal Prediction
   - Hands-on Part 2
6. Discussion

Quick Recap

November session in a nutshell:

- DAGs represent causal structure: $X \to Y$ stands for "X is a direct cause of Y"

Quick Recap

November session in a nutshell:

- DAGs represent causal structure: $X \rightarrow Y$ stands for "X is a direct cause of Y"
- The DAG tells you *if* and *how* you can estimate different causal effects: Which variables you *should* and *shouldn't* **condition** on
  - A variable can be "conditioned on" in different ways, like using a statistical model (e.g. adding it as a predictor to your regression model) or using a particular design (e.g. including only sick people in your study)
  - E.g. if $X \leftarrow Z \rightarrow Y$ - Condition on Z, estimate $E[Y|X, Z]$
  - E.g. if $X \rightarrow Z \leftarrow Y$ - Don't condition on Z! Estimate $E[Y|X]$

Quick Recap

November session in a nutshell:

- DAGs represent causal structure: $X \rightarrow Y$ stands for "X is a direct cause of Y"
- The DAG tells you *if* and *how* you can estimate different causal effects: Which variables you *should* and *shouldn't* **condition** on
  - A variable can be "conditioned on" in different ways, like using a statistical model (e.g. adding it as a predictor to your regression model) or using a particular design (e.g. including only sick people in your study)
  - E.g. if $X \leftarrow Z \rightarrow Y$ - Condition on Z, estimate $E[Y|X, Z]$
  - E.g. if $X \rightarrow Z \leftarrow Y$ - Don't condition on Z! Estimate $E[Y|X]$
- Upside: **If we know the DAG**, and we have all the relevant variables, then we can (in principle) estimate the effect of an intervention from observational data!

Why do DAGs give us this information?

Why do DAGs give us this information?

DAGs allow us to "read off" all the *paths* by which two variables are **marginally** and **conditionally** (in)dependent

- Formalizes the idea that two variables can be statistically dependent **in a particular way** without one being a *cause* of the other
- d-seperation gives us the explicit rules which govern this: open-paths!

Why do DAGs give us this information?

DAGs allow us to "read off" all the *paths* by which two variables are **marginally** and **conditionally** (in)dependent

- Formalizes the idea that two variables can be statistically dependent **in a particular way** without one being a *cause* of the other
- d-seperation gives us the explicit rules which govern this: open-paths!

If we want a **statistical dependency** to reflect a **causal dependency** we need to condition on variables such that we *block* uninteresting (non-causal) paths

Why do DAGs give us this information?

DAGs allow us to "read off" all the *paths* by which two variables are **marginally** and **conditionally** (in)dependent

- Formalizes the idea that two variables can be statistically dependent **in a particular way** without one being a *cause* of the other
- d-seperation gives us the explicit rules which govern this: open-paths!

If we want a **statistical dependency** to reflect a **causal dependency** we need to condition on variables such that we *block* uninteresting (non-causal) paths

Chain

X: Smoking
Z: Tar
Y: Cancer

$$X \not\perp\!\!\!\perp Y$$

$$X \perp\!\!\!\perp Y \mid Z$$

Fork

X: Storks
Z: Environment
Y: Babies

$$X \not\perp\!\!\!\perp Y$$

$$X \perp\!\!\!\perp Y \mid Z$$

Collider

X: Attractiveness
Z: Being Single
Y: Intelligence

$$X \perp\!\!\!\perp Y$$

$$X \not\perp\!\!\!\perp Y \mid Z$$

Last time: Causal Reasoning

causal model, e.g. DAG G

infer

expectations about
observational data $P(X_1,\ldots,X_5)$
& interventions $P_{do}(X_1,\ldots,X_5)$

X_1	X_2	X_3	X_4	X_5
3.4	−0.3	5.8	−2.1	2.2
1.7	−0.2	7.0	−1.2	0.4
−2.4	−0.1	4.3	−0.7	3.5
2.3	−0.3	5.5	−1.1	−4.4
3.5	−0.2	3.9	−0.9	−3.9
⋮	⋮	⋮	⋮	⋮

Today: Causal Learning (also called: Causal Discovery / Structure Learning)

observational data $P(X_1,\ldots,X_5)$

X_1	X_2	X_3	X_4	X_5
3.4	− 0.3	5.8	− 2.1	2.2
1.7	− 0.2	7.0	− 1.2	0.4
− 2.4	− 0.1	4.3	− 0.7	3.5
2.3	− 0.3	5.5	− 1.1	− 4.4
3.5	− 0.2	3.9	− 0.9	− 3.9
⋮	⋮	⋮	⋮	⋮

infer

?

causal model, e.g. DAG G

Causal Learning

Causal Learning is somewhat different from **Statistical Learning**

Causal Learning

Causal Learning is somewhat different from **Statistical Learning**

Statistical Learning is concerned with finding a model with minimizes out-of-sample prediction error.

- Using data observed under certain conditions to **learn** a model which enables us to make predictions about what we expect to see in a new data point collected under **exactly those conditions**

Causal Learning

Causal Learning is somewhat different from **Statistical Learning**

Statistical Learning is concerned with finding a model with minimizes out-of-sample prediction error.

- Using data observed under certain conditions to **learn** a model which enables us to make predictions about what we expect to see in a new data point collected under **exactly those conditions**

Causal Learning is more general, but harder.

- Using data to **learn** a model which enables predictions about **new** or **different** situations. I.e., using observational data to learn about that says something about the **intervention** setting.

Example

Imagine we are a team of epidemiologists.

We take a blood sample from a random sample of the population and record:

- The level of expression of a particular gene X
- The level of expression of a phenotype Y (e.g. blood insulin levels).

Example

Imagine we are a team of epidemiologists.

We take a blood sample from a random sample of the population and record:

- The level of expression of a particular gene X
- The level of expression of a phenotype Y (e.g. blood insulin levels).

Example

Statistical Learning allows us to obtain a model that allows us to answer questions like:

- If I collect one more individuals data under the same conditions, and I observe a **gene expression score of x**, what is my best guess of what **phenotype level** that person has?

Example

Statistical Learning allows us to obtain a model that allows us to answer **certain types of prediction questions** like:

- If I collect one more individuals data under the same conditions, and I observe a **gene expression score of x**, what is my best guess of what **phenotype level** that person has?

Example Causal vs Statistical Learning

What kinds of questions can we **not answer** with Statistical Learning alone?

Example Causal vs Statistical Learning

What kinds of questions can we **not answer** with Statistical Learning alone?

What if instead of just observing genes and phenotypes, I was to *manipulate*/ intervene on / *change* the expression of that gene.

- E.g., deactivating or suppressing gene expression entirely.
- In SCM terminology, I want to know what happens to Y in the context of $do(X = 0)$

What level of phenotype expression would I expect to see if I did that?

- Predicting phenotype from gene expression in a different setting: The intervention setting instead of the observation setting

Example Causal vs Statistical Learning

Example Causal vs Statistical Learning

Example Causal vs Statistical Learning

Example Causal vs Statistical Learning

Example Causal vs Statistical Learning

The answer to questions about *intervention effects* depends on the **causal graph**

For that reason, **causal learning** is often focused on recovering the **structure** of the causal graph from data

Example Causal vs Statistical Learning

The answer to questions about *intervention effects* depends on the **causal graph**

For that reason, **causal learning** is often focused on recovering the **structure** of the causal graph from data

Causal learning has been applied most notably in fields like **systems biology**, **genetic research**, **neural connectivity** research

- In these fields, predictions about the effects of do-interventions can be easily empirically validated, e.g., gene knockout experiments
- But these methods *can* and *have* been applied in other fields.
- Today we learn about the principles behind these causal discovery methods!

Causal Learning: An overview

In this lecture we will learn about three different strategies for causal discovery.

1. Constraint-based Discovery
2. Restricting the Structural Causal Model
3. Using Different Environments (data from multiple sources)

Different ways to exploit relationship between causal models and data: Different assumptions, able to recover different things

Causal Learning: An overview

In this lecture we will learn about three different strategies for causal discovery.

1. Constraint-based Discovery
2. Restricting the Structural Causal Model
3. Using Different Environments (data from multiple sources)

Different ways to exploit relationship between causal models and data: Different assumptions, able to recover different things

Disclaimer: This is by no means a comprehensive review of causal discovery methods, an area of active research and quick development. Instead, simple introduction to three general strategies that show up in many developments.

Constraint-Based Methods (Using Patterns of Conditional (In)Dependence)

Well-known problem of causal discovery: Correlation does not imply causation

F. H. Messerli: *Chocolate Consumption, Cognitive Function, and Nobel Laureates*, N Engl J Med 2012

Fundamentals of Dependence and Causality

Well-known problem of causal discovery: Correlation does not imply causation

Well-known problem of causal discovery: Correlation does not imply causation

Reichenbach's Common Cause Principle

If $X \not\perp\!\!\!\perp Y$ then either:

- $X \rightarrow ... \rightarrow Y$
- $X \leftarrow ... \leftarrow Y$
- X and Y share a common cause
- A combination of the above

Well-known problem of causal discovery: Correlation does not imply causation

Reichenbach's Common Cause Principle

If $X \not\perp\!\!\!\perp Y$ then either:

- $X \rightarrow ... \rightarrow Y$
- $X \leftarrow ... \leftarrow Y$
- X and Y share a common cause
- A combination of the above

Well-known problem of causal discovery: Correlation does not imply causation

Reichenbach's Common Cause Principle

If $X \not\perp\!\!\!\perp Y$ then either:

- $X \rightarrow ... \rightarrow Y$
- $X \leftarrow ... \leftarrow Y$
- X and Y share a common cause
- A combination of the above

Or: we have uknowingly conditioned on a collider Z

- i.e. $X \not\perp\!\!\!\perp Y$ is actually $X \not\perp\!\!\!\perp Y|Z$
- In practice - selection bias

Causal Discovery using Conditional Independence

Basic Idea:

1. Find all conditional independence relations present in the data
2. Draw the DAG in which all (and only) those independences follow from d-seperation rules

Assume:

- **Sufficiency**: No unobserved common causes
- No selection bias (no conditioning on unobserved colliders)
- **Faithfulness**: We'll explain that in a few slides time!

Chain

X: Smoking
Z: Tar
Y: Cancer

$$X \not\perp\!\!\!\perp Y$$

$$X \perp\!\!\!\perp Y \mid Z$$

Fork

X: Storks
Z: Environment
Y: Babies

$$X \not\perp\!\!\!\perp Y$$

$$X \perp\!\!\!\perp Y \mid Z$$

Collider

X: Attractiveness
Z: Being Single
Y: Intelligence

$$X \perp\!\!\!\perp Y$$

$$X \not\perp\!\!\!\perp Y \mid Z$$

Example 1: CI-based discovery

We gather a very large dataset containing three-variables: A, B and C

We choose an appropriate test for independence and find:

$$A \perp\!\!\!\perp C$$
$$A \not\perp\!\!\!\perp C \mid B$$

All other combinations of variables are dependent (e.g. $A \not\perp\!\!\!\perp B$, and $B \not\perp\!\!\!\perp C \mid A$)

What is the data-generating DAG?
menti.com code **1008 3599**

Example 1: CI-based discovery

We gather a very large dataset containing three-variables: A, B and C

We choose an appropriate test for independence and find:

$$A \perp\!\!\!\perp C$$
$$A \not\perp\!\!\!\perp C \mid B$$

All other combinations of variables are dependent (e.g. $A \not\perp\!\!\!\perp B$, and $B \not\perp\!\!\!\perp C \mid A$)

What is the data-generating DAG?
menti.com code **1008 3599**

Chain

X: Smoking
Z: Tar
Y: Cancer

$$X \not\perp\!\!\!\perp Y$$

$$X \perp\!\!\!\perp Y \mid Z$$

Fork

X: Storks
Z: Environment
Y: Babies

$$X \not\perp\!\!\!\perp Y$$

$$X \perp\!\!\!\perp Y \mid Z$$

Collider

X: Attractiveness
Z: Being Single
Y: Intelligence

$$X \perp\!\!\!\perp Y$$

$$X \not\perp\!\!\!\perp Y \mid Z$$

Example 2: CI-based discovery

We gather a very large dataset containing three-variables: A, B and C

We choose an appropriate test for independence and find:

$$A \perp\!\!\!\perp C \mid B$$

All other combinations of variables are dependent

What is the data-generating DAG?
menti.com code **1008 3599**

Example 2: CI-based discovery

We gather a very large dataset containing three-variables: A, B and C

We choose an appropriate test for independence and find:

$$A \perp\!\!\!\perp C \mid B$$

All other combinations of variables are dependent

What is the data-generating DAG?
menti.com code **1008 3599**

Example 1: CI-based discovery

We gather a very large dataset containing three-variables: A, B and C

We choose an appropriate test for independence and find:

$$A \perp\!\!\!\perp C$$
$$A \not\perp\!\!\!\perp C \mid B$$

All other combinations of variables are dependent (e.g. $A \not\perp\!\!\!\perp B$ and $B \not\perp\!\!\!\perp C \mid A$)

What is the data-generating DAG?
menti.com code **80 76 34 5**

Example 2: CI-based discovery

We gather a very large dataset containing three-variables: A, B and C

We choose an appropriate test for independence and find:

$$A \perp\!\!\!\perp C \mid B$$

All other combinations of variables are dependent

What is the data-generating DAG?
menti.com code **80 76 34 5**

Basics of CI-based discovery

In the general case, we usually cannot **uniquely** identify the DAG based only on patterns of statistical independence and dependence.

We already know this from d-seperation rules, since they tell us that different DAGs can yield the same statistical dependencies

Markov Equivalence:

Two DAGs are *Markov Equivalent* if they satisfy the same d-seperation statements, and so the same set of (conditional) (in)dependence relations

Basics of CI-based discovery

But we can learn **something** about the underlying DAG!

Implications of d-seperation:

1. If two variables share a *direct causal path* $X \to Y$ OR $X \leftarrow Y$, then you can **never** block that path by conditioning on other variables!
   - If two variables are dependent **no matter what we condition on**, they share a direct causal link **of some direction**

Dehydration
Sex a.b
Aspirin
Recovery
Blood Pressure

Dehydration
Sex a.b
Aspirin
Recovery
Blood Pressure

Basics of CI-based discovery

But we can learn **something** about the underlying DAG!

Implications of d-seperation:

1. If two variables share a *direct causal path* $X \rightarrow Y$ OR $X \leftarrow Y$, then you can **never** block that path by conditioning on other variables!
   - If two variables are dependent **no matter what we condition on**, they share a direct causal link **of some direction**

2. Colliders imply a different pattern of dependencies than chains and forks

Chain

X: Smoking
Z: Tar
Y: Cancer

$$X \not\perp\!\!\!\perp Y$$

$$X \perp\!\!\!\perp Y \mid Z$$

Fork

X: Storks
Z: Environment
Y: Babies

$$X \not\perp\!\!\!\perp Y$$

$$X \perp\!\!\!\perp Y \mid Z$$

Collider

X: Attractiveness
Z: Being Single
Y: Intelligence

$$X \perp\!\!\!\perp Y$$

$$X \not\perp\!\!\!\perp Y \mid Z$$

$A \perp\!\!\!\perp C$

$A \perp\!\!\!\perp C \mid B$

The basics of CI-based discovery

CI-based methods can discover:

A Which variables share a direct causal link and which variables don't
- This is known as the *skeleton* of a DAG
- If $X \rightarrow Y$ or $X \leftarrow Y$, then X and Y *will never be statistically independent* (marginally or conditionally).

B If there's a collider, we can detect the direction of causal arrows, **as long as** the two "cause" variables don't also cause each other
- Also known as an "immorality" - two parent nodes who share a child but are not "married" (!!!)
- **If everything is dependent on everything - very little can be learned!**

In general, outside of (immoral) colliders, we can't determine the direction of the causal arrow

Assumptions for CI-based discovery

Global Markov Condition:

X and Y are d-seperated by $S \implies X \perp\!\!\!\perp Y \mid S$

Assumptions for CI-based discovery

Global Markov Condition:

X and Y are d-seperated by $S \implies \mathsf{X} \perp\!\!\!\perp Y \mid S$

Faithfulness:

$X \perp\!\!\!\perp Y \mid S \implies X$ and Y are d-seperated by S

Violations of Faithfulness

$$A := \epsilon_A$$
$$B := .5A + \epsilon_B$$
$$C := -.25A + .5B + \epsilon_C$$

where

- $\epsilon_A, \epsilon_B, \epsilon_C$ are iid, $\sim \mathcal{N}(0,1)$

$$\begin{pmatrix} A \\ B \\ C \end{pmatrix} = \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & .5 & 0 \\ .5 & 1.25 & .5 \\ 0 & .5 & 1.25 \end{pmatrix}\right]$$

Violations of Faithfulness

$$A := \epsilon_A$$
$$B := .5A + \epsilon_B$$
$$C := -.25A + .5B + \epsilon_C$$

where

- $\epsilon_A, \epsilon_B, \epsilon_C$ are iid, $\sim \mathcal{N}(0, 1)$

$$A \perp\!\!\!\perp C$$
$$A \not\perp\!\!\!\perp C \mid B$$

Violations of Faithfulness

$$A \perp\!\!\!\perp C$$

$$A \not\perp\!\!\!\perp C \mid B$$

Assume Faithfulness

Violations of Faithfulness

$$A \perp\!\!\!\perp C$$

$$A \not\perp\!\!\!\perp C \mid B$$

Assume Faithfulness

Assumptions for CI-based discovery

Global Markov Condition:

P is Markov w.r.t G if

$$X \text{ and } Y \text{ are d-seperated by } S \implies \mathrm{X} \perp\!\!\!\perp Y \mid S$$

Faithfulness:

P satisfies faithfulness w.r.t G if

$$X \perp\!\!\!\perp Y \mid S \implies X \text{ and } Y \text{ are d-seperated by } S$$

Essentially: Paths never "perfectly cancel out"

Statistical (conditional) Independence $\implies$ causal independence (d-seperation)

The basics of CI-based discovery

Typically CI-based methods cannot *uniquely* identify the underlying DAG from observational data. Instead they can identify a *Markov Equivalence set*

Markov Equivalence:

Two DAGs are *Markov Equivalent* if they satisfy the same d-seperation statements, and so the same set of (conditional) (in)dependence relations

The basics of CI-based discovery

Typically CI-based methods cannot *uniquely* identify the underlying DAG from observational data. Instead they can identify a *Markov Equivalence set*

Markov Equivalence:

Two DAGs are *Markov Equivalent* if they satisfy the same d-seperation statements, and so the same set of (conditional) (in)dependence relations

You can determine whether two DAGs are markov-equivalent by checking:

A They have the same skeleton

B They have the same immoralities

The basics of CI-based discovery

Typically CI-based methods cannot *uniquely* identify the underlying DAG from observational data. Instead they can identify a *Markov Equivalence set*

Markov Equivalence:

Two DAGs are *Markov Equivalent* if they satisfy the same d-seperation statements, and so the same set of (conditional) (in)dependence relations

You can determine whether two DAGs are markov-equivalent by checking:

A They have the same skeleton

B They have the same immoralities

Represented by a Complete Partially-Oriented Directed Acyclic Graph **(CPDAG)**

CPDAG

Marov-Equivalence Set

CPDAG

Marov-Equivalence Set

Constraint-based vs Score-based methods

These types of strategies are divided into "**constraint-based**" and "**score-based**" (Spirtes et al. 2000)

Constraint-based algorithms:

- Test a large number of possible conditional independence statements
- In practice, don't need to test **all** independenceis, but can perform a quicker search
- PC algorithm / FCI: Start with a full graph. If marg. independent remove edge: no need to test any more independence relations for that pair. Repeat for cond. independence

Score-based algorithms:

- GES: Start with an empty graph. Add the edge that improves model fit according to BIC. Repeat. Note equivalences
- Generally speaking able to identify the same things as constraint-based methods, but less well able to deal with violations of sufficiency

In practice

Constraint-based methods

- Extensions exist that deal with violations of *sufficiency*
- Fundamental assumption: faithfulness

In practice

Constraint-based methods

- Extensions exist that deal with violations of *sufficiency*
- Fundamental assumption: faithfulness

Disadvantages:

- Population CI's aren't known: We need statistical tests + sample data
  - All of statistics is relevant here, e.g., sample size considerations
  - In a given sample : Type I and II errors
  - Faithfulness $\neq$ No false negatives!

In practice

Constraint-based methods

- Extensions exist that deal with violations of *sufficiency*
- Fundamental assumption: faithfulness

Disadvantages:

- Population CI's aren't known: We need statistical tests + sample data
  - All of statistics is relevant here, e.g., sample size considerations
  - In a given sample : Type I and II errors
  - Faithfulness $\neq$ No false negatives!
- CI testing easy if linear + Gaussian (partial correlation / regression) or discrete (cross-tables)
  - Can be difficult/time-consuming in other cases (Shah & Peters, 2020)
  - Non-parametric methods exist but require large sample size and computation time (e.g., the kernel PC package)

Understanding human functioning using graphical models

Markus Kalisch, Bernd AG Fellinghauer, Eva Grill, Marloes H Maathuis, Ulrich Mansmann, Peter Bühlmann & Gerold Stucki

BMC Medical Research Methodology **10**, Article number: 14 (2010) | Cite this article

Figure 1

Five largest connected components. All five connected components containing at least three nodes in the skeleton computed on the SCI data.

Spinal Cord Injury (SCI) data

The multi center, cross-sectional study of patients with SCI was conducted in 14 countries. Individuals were included if they had sustained a SCI with an acute onset, or if they were receiving rehabilitation in the early post-acute situation, or if they were in a long-term context. Individuals had to be at least 18 years old and had to be able to understand purpose and

Challenges and Opportunities with Causal Discovery Algorithms: Application to Alzheimer's Pathophysiology

Xinpeng Shen, Sisi Ma, Prashanthi Vemuri, Gyorgy Simon & the Alzheimer's Disease Neuroimaging Initiative

Scientific Reports **10**, Article number: 2975 (2020) | Cite this article

Causal Discovery with Conditional Independence: In a Nutshell

observational data $P(X_1,\ldots,X_5)$

X_1	X_2	X_3	X_4	X_5
3.4	− 0.3	5.8	− 2.1	2.2
1.7	− 0.2	7.0	− 1.2	0.4
− 2.4	− 0.1	4.3	− 0.7	3.5
2.3	− 0.3	5.5	− 1.1	− 4.4
3.5	− 0.2	3.9	− 0.9	− 3.9
⋮	⋮	⋮	⋮	⋮

infer

?

causal model, e.g. DAG G

Causal Discovery with Conditional Independence: In a Nutshell

observational data $P(X_1,\ldots,X_5)$

X_1	X_2	X_3	X_4	X_5
3.4	− 0.3	5.8	− 2.1	2.2
1.7	− 0.2	7.0	− 1.2	0.4
− 2.4	− 0.1	4.3	− 0.7	3.5
2.3	− 0.3	5.5	− 1.1	− 4.4
3.5	− 0.2	3.9	− 0.9	− 3.9
⋮	⋮	⋮	⋮	⋮

independence tests →

$$X_5 \perp\!\!\!\perp X_2 \mid X_1 X_3 X_4$$
$$X_3 \perp\!\!\!\perp X_5 \mid X_1 X_2 X_4$$
$$\vdots$$
$$X_1 \not\perp\!\!\!\perp X_2 \mid X_3 X_4$$

Faithfulness →
Markov Cond.

G

Causal Discovery with Conditional Independence: In a Nutshell

observational data $P(X_1,\ldots,X_5)$

X_1	X_2	X_3	X_4	X_5
3.4	− 0.3	5.8	− 2.1	2.2
1.7	− 0.2	7.0	− 1.2	0.4
− 2.4	− 0.1	4.3	− 0.7	3.5
2.3	− 0.3	5.5	− 1.1	− 4.4
3.5	− 0.2	3.9	− 0.9	− 3.9
⋮	⋮	⋮	⋮	⋮

independence tests

$$X_5 \perp\!\!\!\perp X_2 \mid X_1 X_3 X_4$$
$$X_3 \perp\!\!\!\perp X_5 \mid X_1 X_2 X_4$$
$$\vdots$$
$$X_1 \not\perp\!\!\!\perp X_2 \mid X_3 X_4$$

Faithfulness

Markov Cond.

Outline

1. Recap and Q&A
2. Why Causal Discovery?
3. Constraint-Based Methods
   - **Hands-on Part 1**
4. Restricted SCMs
5. Invariant Causal Prediction
   - Hands-on Part 2
6. Discussion

Restricted Causal Models

Restricted SCMs

Recall: SCMs are in principle non-parametric - relationships can take any form, and the errors can have any distribution

In some special cases we can "buy" an ability to learn the causal structure if we are willing to make more assumptions about the types of relationships and/or distribution of the noise term in the SCM

Restricted SCMs

Recall: SCMs are in principle non-parametric - relationships can take any form, and the errors can have any distribution

In some special cases we can "buy" an ability to learn the causal structure if we are willing to make more assumptions about the types of relationships and/or distribution of the noise term in the SCM

Examples so far: $Y := BX + \epsilon$, where ϵ is Gaussian, $\epsilon \sim N(0, \sigma^2)$

- This is a special type of model called a linear model with additive Gaussian noise
- Recall: In an SCM, the *noise variable* is independent of the cause variable. $\epsilon \perp\!\!\!\perp X$

Restricted SCMs

Recall: SCMs are in principle non-parametric - relationships can take any form, and the errors can have any distribution

In some special cases we can "buy" an ability to learn the causal structure if we are willing to make more assumptions about the types of relationships and/or distribution of the noise term in the SCM

Examples so far: $Y := BX + \epsilon$, where ϵ is Gaussian, $\epsilon \sim N(0, \sigma^2)$

- This is a special type of model called a linear model with additive Gaussian noise
- Recall: In an SCM, the *noise variable* is independent of the cause variable. $\epsilon \perp\!\!\!\perp X$
- As we have seen: In general, many different causal models of this form that are **statistically equivalent**

Bivariate SCMs

Bivariate SCMs

$$Y := B_X X + \epsilon_Y \quad \epsilon_Y \sim \mathcal{N}(0, \sigma_Y^2)$$

$$X := B_Y Y + \epsilon_X \quad \epsilon_X \sim \mathcal{N}(0, \sigma_X^2)$$

Bivariate SCMs: Linear with Gaussian Additive Noise

$$Y := BX + \epsilon_Y \qquad \epsilon_Y \sim \mathcal{N}(0, {\sigma_Y}^2)$$

Bivariate SCMs: Linear with Gaussian Additive Noise

$$Y := BX + \epsilon_Y \qquad \epsilon_Y \sim \mathcal{N}(0, {\sigma_Y}^2)$$

Restricted SCMs

If we are dealing with a **bivariate system** with linear relationships and Gaussian noise, we can never distinguish if $X \to Y$ or $Y \to X$.

- In the true model we know that the cause variable is independent of the error, $\epsilon_Y \perp\!\!\!\perp X$
- But in this situation, when we fit a statistical model in the **wrong** causal direction, we still get errors which are independent of the predictor, $\epsilon_X \perp\!\!\!\perp Y$

Restricted SCMs

If we are dealing with a **bivariate system** with linear relationships and Gaussian noise, we can never distinguish if $X \to Y$ or $Y \to X$.

- In the true model we know that the cause variable is independent of the error, $\epsilon_Y \perp\!\!\!\perp X$
- But in this situation, when we fit a statistical model in the **wrong** causal direction, we still get errors which are independent of the predictor, $\epsilon_X \perp\!\!\!\perp Y$

But causal Models **are** statistically distinguishable if we are willing to assume the SCM is **either**:

A Linear Model with **Non-Gaussian** Additive Noise

B **Non-Linear** Model with Gaussian Additive Noise

Bivariate SCMs: Linear with non-Gaussian Additive Noise

$$Y := BX + \epsilon_Y \qquad \epsilon_Y \sim U(lb, ub)$$

Bivariate SCMs: Linear with non-Gaussian Additive Noise

$$Y := BX + \epsilon_Y \qquad \epsilon_Y \sim U(lb, ub)$$

Bivariate SCMS: Linear with non-Gaussian Additive Noise

We assume an underlying SCM where the *noise variable* is independent of the cause variables.

$$Y := BX + \epsilon_Y$$
$$\epsilon_Y \perp\!\!\!\perp X$$

If we fit a model with the "wrong" causal direction, the *noise* of that regression model will not be independent of the predictors

$$X = B_Y Y + \epsilon_X$$
$$\epsilon_X \not\perp\!\!\!\perp Y$$

This only works for non-Gaussian noise, but there it allows us to identify the direction of the causal arrow!

Bivariate SCMS: Linear with non-Gaussian Additive Noise

$$Y := BX + \epsilon_Y \qquad \epsilon_Y \sim U(lb, ub)$$

Bivariate SCMS: Non-Linear with Gaussian Additive Noise

$$Y := sin(X) + \epsilon_Y \qquad \epsilon_Y \sim \mathcal{N}(0, \sigma^2)$$

Bivariate SCMS: Non-Linear with Gaussian Additive Noise

$$Y := sin(X) + \epsilon_Y \qquad \epsilon_Y \sim \mathcal{N}(0, \sigma^2)$$

Bivariate SCMS: Non-Linear with Gaussian Additive Noise

We assume an underlying SCM

$$Y := f(X) + \epsilon_Y$$
$$\epsilon_Y \perp\!\!\!\perp X$$

If $f(X)$ is *non-linear* (e.g. $sin(X), X^p, e^X$), there is **no equivalent statistical model** that satisfies

$$X := g(Y) + \epsilon_X$$
$$\epsilon_X \perp\!\!\!\perp Y$$

If we fit a model in the wrong causal direction, we will find $\epsilon_X \not\perp\!\!\!\perp Y$ We can test this by fitting non-linear models in both 'directions' and testing for independence of errors!

Summary

Linear Non-Gaussian:

- LiNGAM algorithm (Shimizu et al. 2006)
- Extends this idea to the multivariate case
- Can uniquely identify multivariate DAGs!

Disadvantages:

- Need sufficiency (some extensions)
- How common are these systems?
- Degree of non-normality?

Summary

Linear Non-Gaussian:

- LiNGAM algorithm (Shimizu et al. 2006)
- Extends this idea to the multivariate case
- Can uniquely identify multivariate DAGs!

Disadvantages:

- Need sufficiency (some extensions)
- How common are these systems?
- Degree of non-normality?

Non-Linear Gaussian:

- Mooij et al (2016)
- Use non-parametric/smoothing methods
- If we know the noise distribution, can choose best fitting model
- Works well in bivariate case with no confounding

Disadvantages:

- Again, sufficiency needed
- Difficult to scale to multivariate case

Network modelling methods for FMRI

Stephen M. Smith [a,*], Karla L. Miller [a], Gholamreza Salimi-Khorshidi [a], Matthew Webster [a], Christian F. Beckmann [a,b], Thomas E. Nichols [a,c], Joseph D. Ramsey [d], Mark W. Woolrich [a,e]

Nonlinear causal discovery with additive noise models

Patrik O. Hoyer
University of Helsinki
Finland

Dominik Janzing
MPI for Biological Cybernetics
Tübingen, Germany

Joris Mooij
MPI for Biological Cybernetics
Tübingen, Germany

Jonas Peters
MPI for Biological Cybernetics
Tübingen, Germany

Bernhard Schölkopf
MPI for Biological Cybernetics
Tübingen, Germany

Figure 26: Scatter plots of pairs from D.9. pair0033: alcohol consumption → mean corpuscular volume, pair0034: alcohol consumption → alkaline phosphotase, pair0035: alcohol consumption → alanine aminotransferase, pair0036: alcohol consumption → aspartate aminotransferase, pair0037: alcohol consumption → gamma-glutamyl transpeptidase.

atter plots of pairs from D.7. pair0022: age → height, pair0023: age → 024: age → heart rate.

An estimation of causal structure based on Latent LiNGAM for mixed data

Mako Yamayoshi, Jun Tsuchida & Hiroshi Yadohisa

Behaviormetrika **47**, 105–121 (2020) | Cite this article

Variables	Details
low	Indicator of birth weight less than 2.5 kg (binary)
age	Mother's age in years
lwt	Mother's weight in pounds at last menstrual period (*lb*)
smoke	Smoking status during pregnancy (binary)
ptl	Number of previous premature labors
ht	History of hypertension (binary)
ui	Presence of uterine irritability (binary)
ftv	Number of physician visits during the first trimester

(b)

Restricted Causal Models: In a Nutshell

observational data $P(X_1,\ldots,X_5)$

X_1	X_2	X_3	X_4	X_5
3.4	− 0.3	5.8	− 2.1	2.2
1.7	− 0.2	7.0	− 1.2	0.4
− 2.4	− 0.1	4.3	− 0.7	3.5
2.3	− 0.3	5.5	− 1.1	− 4.4
3.5	− 0.2	3.9	− 0.9	− 3.9
⋮	⋮	⋮	⋮	⋮

infer

?

causal model, e.g. DAG G

Restricted Causal Models: In a Nutshell

observational data $P(X_1,\ldots,X_5)$

X_1	X_2	X_3	X_4	X_5
3.4	− 0.3	5.8	− 2.1	2.2
1.7	− 0.2	7.0	− 1.2	0.4
− 2.4	− 0.1	4.3	− 0.7	3.5
2.3	− 0.3	5.5	− 1.1	− 4.4
3.5	− 0.2	3.9	− 0.9	− 3.9
⋮	⋮	⋮	⋮	⋮

Assumptions about SCM

Non-Linear / Non-Gaussian

Search for stat. model in which

Noise(X)

$\perp\!\!\!\perp$

Parents (X)

Invariant Causal Prediction

So far we looked at methods for causal discovery that work with a **single observational dataset**

What if we have a mix of **observational** and **experimental** data? Can we do something different?

This is the idea behind causal discovery using *Invariant Causal Prediction*

Invariant Causal Prediction

Basic Idea:

1. There is some unknown causal graph, but we only care about learning **the direct causes** of one *target variable* Y
2. We have data drawn from different *environments*
   - Here: A mix of observational and intervention data
   - Interventions act on variables other than Y
3. We identify the direct causes of Y by looking for those conditional **dependencies** that are *invariant* (remain the same) across environments

Invariant Causal Prediction (ICP)

Basic Idea:

1. There is some unknown causal graph, but we only care about learning **the direct causes** of one *target variable* Y
2. We have data drawn from different *environments*
   - Here: A mix of observational and intervention data
   - Interventions act on variables other than Y
3. We identify the direct causes of Y by looking for those conditional **dependencies** that are *invariant* (remain the same) across environments

ICP: Key Assumption

Modularity and Localized Interventions:

We assume that it is possible to intervene on a variable without fundamentally changing how it relates to other variables

- We can change $p(X)$ without changing $p(Y \mid X)$
- We can change one cause-effect mechanism without changing the others

ICP Rationale

Let $X \to Y$ and let $pa(Y)$ denote all other **direct causes** of Y. **Modularity** implies that: $p(Y \mid X, pa(Y))$ remains the same no matter whether we have:

- Observational data
- Data where we intervene on X
- Data where we intervene on a common cause of X and Y

ICP Rationale

Let $X \to Y$ and let $pa(Y)$ denote all other **direct causes** of Y. **Modularity** implies that: $p(Y \mid X, pa(Y))$ remains the same no matter whether we have:

- Observational data
- Data where we intervene on X
- Data where we intervene on a common cause of X and Y

However, $p(Y \mid X, pa(Y))$ will be different if:

- X is not a direct cause of Y (e.g., $X \leftarrow Y$)

ICP Rationale

Let $X \to Y$ and let $pa(Y)$ denote all other **direct causes** of Y. **Modularity** implies that: $p(Y \mid X, pa(Y))$ remains the same no matter whether we have:

- Observational data
- Data where we intervene on X
- Data where we intervene on a common cause of X and Y

However, $p(Y \mid X, pa(Y))$ will be different if:

- X is not a direct cause of Y (e.g., $X \leftarrow Y$)
- $X \to Y$ but we have intervened on Y

ICP Rationale

Let $X \to Y$ and let $pa(Y)$ denote all other **direct causes** of Y. **Modularity** implies that: $p(Y \mid X, pa(Y))$ remains the same no matter whether we have:

- Observational data
- Data where we intervene on X
- Data where we intervene on a common cause of X and Y

However, $p(Y \mid X, pa(Y))$ will be different if:

- X is not a direct cause of Y (e.g., $X \leftarrow Y$)
- $X \to Y$ but we have intervened on Y

Invariant Causal Prediction: Only conditional dependence relationships (i.e. *predictions*) that reflect direct cause-effect relationships (i.e. *causal*) will remain the same (i.e. *invariant*) across different **environments**

Example

Observational Data

Example

Observational Data

What **open paths** make up the statistical dependency $cor(Y, X_1|X_4)$?

Example

Observational Data

What **open paths** make up the statistical dependency $cor(Y, X_1|X_4)$?

- $X1 \rightarrow Y$

Example

Observational Data

What **open paths** make up the statistical dependency $cor(Y, X_1|X_4)$?

- $X1 \rightarrow Y$

What **open paths** make up the statistical dependency $cor(Y, X_3)$?

- $X_3 \rightarrow X_4 \rightarrow Y$
- $X_3 \rightarrow X_1 \rightarrow Y$
- $X_3 \rightarrow X_4 \rightarrow X_1 \rightarrow Y$

Example

Intervention Data: e.g. $do(X_1)$

What **open paths** make up the statistical dependency $cor(Y, X_1|X_4)$?

- $X1 \rightarrow Y$

Example

Intervention Data: e.g. $do(X_1)$

What **open paths** make up the statistical dependency $cor(Y, X_1|X_4)$?

- $X1 \rightarrow Y$

What **open paths** make up the statistical dependency $cor(Y, X_3)$?

- $X_3 \rightarrow X_4 \rightarrow Y$

Fundamental assumption: $X_1, X_4 \to Y$ is invariant under interventions.

cf. modularity, autonomy, Haavelmo 1944, Aldrich 1989, Pearl 2009, ...

Fundamental assumption: $X_1, X_4 \to Y$ is invariant under interventions.

cf. modularity, autonomy, Haavelmo 1944, Aldrich 1989, Pearl 2009, ...

Key idea: Use and data and search for invariant models.

set	$\emptyset$	$\{1\}$	$\{2\}$	$\{3\}$	$\cdots$	$\{1,4\}$	$\{2,4\}$	$\cdots$	$\{1,3,4\}$
invariance	✗	✗	✗	✗	$\cdots$	✓	✗	$\cdots$	✓

$$\hat{S} := \bigcap_{S \text{ invariant}} S = \{1,4\}$$

ICP in Practice

Invariant Causal Prediction (Peters, Buhlmann, Meinshausen, 2016)

- Can be used even when it is not known what variables are intervened on or what interventions are applied ("fat hand interventions")
- Basic idea can be extended to other "environments" - non-descendants and time
- Can also be extended to learning the full graph, with some caveats
- Extended/Generalized to non-linear case

Disadvantages:

- Sufficiency (again, some extensions)
- Need knowledge that interventions are not directly on Y
- **Need different environments**: The more environments the better!

Methods for causal inference from gene perturbation experiments and validation

Nicolai Meinshausen[a], Alain Hauser[b], Joris M. Mooij[c], Jonas Peters[d], Philip Versteeg[c], and Peter Bühlmann[a,1]

Fig. 3. The graph of estimated causal relations between the biochemical agents in the ref. 44 data. Blue edges are found by an invariant prediction approach, whereas red edges are found if allowing hidden variables and feedback with invariant prediction. Purple edges are found with invariant prediction whether allowing for hidden variables or not. The solid edges

Fig. 1. The activities of two pairs of genes. Observational data are shown as red crosses and interventional data as blue circles. A blue arrow marks the experiment where an intervention on gene YOL067C occurs (activity shown on the *x* axis in the left panel) and analogously in the right panel. The in-

Christina Heinze-Deml*, Jonas Peters, and Nicolai Meinshausen

Invariant Causal Prediction for Nonlinear Models

https://doi.org/10.1515/jci-2017-0016
Received January 7, 2017; revised May 9, 2018; accepted August 24, 2018

Figure 1

Three candidates for a causal DAG with target total fertility rate (TFR) and four potential causal predictor variables. We would like to infer the parents of TFR in the true causal graph. We use the continent as the environment variable E. If the true DAG was one of the two graphs on the left, the environmental variable would have no direct influence on the target variable TFR and 'Continent' would be a valid environmental variable, see Definition 1.

observational data $P(X_1,\ldots,X_5)$

X_1	X_2	X_3	X_4	X_5
3.4	− 0.3	5.8	− 2.1	2.2
1.7	− 0.2	7.0	− 1.2	0.4
− 2.4	− 0.1	4.3	− 0.7	3.5
2.3	− 0.3	5.5	− 1.1	− 4.4
3.5	− 0.2	3.9	− 0.9	− 3.9
⋮	⋮	⋮	⋮	⋮

infer

?

causal model, e.g. DAG G

In a Nutshell: ICP

observational data $P(X_1,\ldots,X_5)$
&/ intervention data $P_{do}(X_1,\ldots,X_5)$

X_1	X_2	X_3	X_4	X_5
3.4	− 0.3	5.8	− 2.1	2.2
1.7	− 0.2	7.0	− 1.2	0.4
− 2.4	− 0.1	4.3	− 0.7	3.5
2.3	− 0.3	5.5	− 1.1	− 4.4
3.5	− 0.2	3.9	− 0.9	− 3.9
⋮	⋮	⋮	⋮	⋮

infer

?

causal model, e.g. DAG G

In a Nutshell: ICP

observational data $P(X_1,\ldots,X_5)$
&/ intervention data $P_{do}(X_1,\ldots,X_5)$

X_1	X_2	X_3	X_4	X_5
3.4	− 0.3	5.8	− 2.1	2.2
1.7	− 0.2	7.0	− 1.2	0.4
− 2.4	− 0.1	4.3	− 0.7	3.5
2.3	− 0.3	5.5	− 1.1	− 4.4
3.5	− 0.2	3.9	− 0.9	− 3.9
⋮	⋮	⋮	⋮	⋮

Modularity

Interventions
Not on Y

causal model, e.g. DAG G

Outline

1. Recap and Q&A
2. Why Causal Discovery?
3. Constraint-Based Methods
   - **Hands-on Part 1**
4. Restricted SCMs
5. Invariant Causal Prediction
   - **Hands-on Part 2**
6. Discussion

Can we infer causal structure from data?

Short answer:

- In general, without assumptions, **no**
- There is usually more than one SCM that generates the same observational data

Long answer:

- Yes, or at least, we can learn *something* about the causal structure
- But only under certain conditions - if we are willing to make certain assumptions about the causal system
- Note: This is a limitation of all science - no free lunch!

Summary

In this lecture we learned about three different strategies for causal discovery.

1. Constraint-based and Score-based methods
   - Going from estimated conditional and marginal (in)dependence to the DAG structure
   - Generally: Not able to *uniquely identify* the DAG, but instead a set of possible DAGs (Markov Equivalence Set)

Summary

In this lecture we learned about three different strategies for causal discovery.

1. Constraint-based and Score-based methods
   - Going from estimated conditional and marginal (in)dependence to the DAG structure
   - Generally: Not able to *uniquely identify* the DAG, but instead a set of possible DAGs (Markov Equivalence Set)
2. Restricting the Structural Causal Model
   - If we know something about *how exactly* variables look like and/or relate to one another, we can use that to identify the causal model!
   - Specifically: Linear but Non-Gaussian or Non-linear models - succeed in getting a single underlying DAG

Summary

In this lecture we learned about three different strategies for causal discovery.

1. Constraint-based and Score-based methods
   - Going from estimated conditional and marginal (in)dependence to the DAG structure
   - Generally: Not able to *uniquely identify* the DAG, but instead a set of possible DAGs (Markov Equivalence Set)
2. Restricting the Structural Causal Model
   - If we know something about *how exactly* variables look like and/or relate to one another, we can use that to identify the causal model!
   - Specifically: Linear but Non-Gaussian or Non-linear models - succeed in getting a single underlying DAG
3. Using Invariance and Data from Different Environments
   - ICP: Look for predictive relationships that stay the same in different settings (observational/interventional)
   - Works if we care about causes of a single variable, and we don't intervene on that variable

Some Topics Not Covered Today

Extensions and combinations of these ideas

- PC, ICP, LiNGAM extensions to deal with hidden/latent variables
- ICP and PC algorithm have both been extended to the generalized setting: Testing conditional independence without parametric assumptions (kernel-based independence testing)
- Tetrad project (link in labs) and JCI (Joint Causal Inference; Mooij et al., 2020)
- Cyclic Causal Models

Longitudinal/Time-series Data

- Granger and sims causality
- Sequential ICP
- Interrupted Time-series
- Synthetic Controls

How to incorporate causal learning into machine learning?

- The idea of **invariant** statistical models (ICP) looks like being key there

Summary

There are different ways to exploit relationship between causal models and data to recover the causal model itself: Different assumptions, able to recover different things

- Many extensions and combinations of these methods also exist, and different strategies are possible

Summary

There are different ways to exploit relationship between causal models and data to recover the causal model itself: Different assumptions, able to recover different things

- Many extensions and combinations of these methods also exist, and different strategies are possible

Learning predictive models $\neq$ learning causal models

Summary

There are different ways to exploit relationship between causal models and data to recover the causal model itself: Different assumptions, able to recover different things

- Many extensions and combinations of these methods also exist, and different strategies are possible

Learning predictive models $\neq$ learning causal models

All causal learning methods need statistical techniques

- So problems of stat. learning also apply, e.g. sample size and data quality
- "Garbage in $\rightarrow$ garbage out"

Summary

There are different ways to exploit relationship between causal models and data to recover the causal model itself: Different assumptions, able to recover different things

- Many extensions and combinations of these methods also exist, and different strategies are possible

Learning predictive models $\neq$ learning causal models

All causal learning methods need statistical techniques

- So problems of stat. learning also apply, e.g. sample size and data quality
- "Garbage in $\rightarrow$ garbage out"

No free lunch principle

- All discovery algorithms rely on assumptions about the underlying causal process
- Assumptions/Knowledge in $\rightarrow$ Knowledge Out

Thanks!

o.ryan@uu.nl

Extra References

- Spirtes, Peter, Clark N. Glymour, Richard Scheines, and David Heckerman. Causation, prediction, and search. MIT press, 2000.
- Shimizu, S., Hoyer, P. O., Hyvärinen, A., Kerminen, A., & Jordan, M. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research,* 7(10).
- Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., & Schölkopf, B. (2016). Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research,* 17(1), 1103-1204.
- Peters, J., Bühlmann, P., & Meinshausen, N. (2016). Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 947-1012.
- Peters, J., Janzing, D., & Schölkopf, B. (2017). Elements of causal inference: foundations and learning algorithms. The MIT Press.

- Credit some slides: Jonas Peeters Causal Discovery workshop