Z-INF: Causality

December 28, 2024

Learning Outcomes

- Explain the concept of causality and contrast statistical approaches to causal approaches.
- Design graphical models to encode causal systems and perform statistical, intereventional and counterfactual reasoning on them.
- Perform causal inference from graphs and data.
- Perform causal discovery from data.
- Apply causal methods in machine learning, reinforcement learning and representation learning.
- (Appraise, criticize and code causal algorithms)

Content Outline

1. Introduction to causality. Review of important notions from statistics. Differences between statistics/ML and causality. Reichenbach's principle. Intuitive introduction of the concept of causal structure, interventions.

- Pearl, J., Glymour, M. and Jewell, N.P.. *Causal inference in statistics: A primer.* Chapter 1 (Probability and Statistics, Graphs, Intro to SCMs).
- Peters, J., Janzing, D. and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms.* Chapter 1 (Learning and Causal Modelling, Reichenbach's principle, Principle of Independent Mechanisms); Section 2.1 .
- (Pearl, J., 2015. *Trygve Haavelmo and the emergence of causal calculus.* Econometric Theory, 31(1), pp.152-179.)

▶ Causal models as programs.

2. Graphical Models. Graphical models. BNs, CBNs, SCMs. Graphical structures. d-separation. Markovianity. Faithfulness.

- Pearl, J., Glymour, M. and Jewell, N.P.. *Causal inference in statistics: A primer.* Chapter 2 (Graphical structures: chains, forks and colliders; d-separation).
- Pearl, J.. *Causality.* Section 1.2 (Bayesian Networks), 1.3 (Causal Bayesian Networks), 1.4.1-1.4.2 (Structural Causal Models)
- Peters, J., Janzing, D. and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms.* Section 6.5 (Markov property, Markov equivalence, faithfulness, causal minimality)

▶ `https://pgmpy.org/`

3. Causal Inference: Identifiability (from interventions to do-calculus) Interventions. Identifiability. Randomized experiments. Adjustement formula. Truncated formula. Backdoor criterion. Do-calculus.

- Bareinboim, E., Correa, J.D., Ibeling, D. and Icard, T. *On Pearl's hierarchy and the foundations of causal inference.* Sections 1.1, 1.2, 1.4 (up to 1.4.3.1)
- Pearl, J., Glymour, M. and Jewell, N.P. *Causal inference in statistics: A primer.* Section 3.1-3.3 (Interventions, Adjustment Formula, Backdoor criterion).
- Peters, J., Janzing, D. and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms.* Section 6.7 (Do-calculus).
- (Huszar F., *Causal Inference 2: Illustrating Interventions via a Toy Example*, `https://www.inference.vc/causal-inference-2-illustrating-interventions-in-a-toy-example/`)
- (Heiss A., *Do-calculus adventures!*, `https://www.andrewheiss.com/blog/2021/09/07/do-calculus-backdoors/`)
- (Bareinboim, E., Correa, J.D., Ibeling, D. and Icard, T. *On Pearl's hierarchy and the foundations of causal inference.* Sections 1.5)

▶ `https://causalfusion.net/app`

4. Causal Inference: Identification (from propensity scores to ML methods) Potential Outcomes. ATE. Matching/Stratification. IPW. CEVAE/Dragonnet. (IV. Double regression.)

- Peters, J., Janzing, D. and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms.* Section 6.9 (Potential Outcomes).
- Austin, P.C., 2011. *An introduction to propensity score methods for reducing the effects of confounding in observational studies.*
- Louizos, C., Shalit, U., Mooij, J.M., Sontag, D., Zemel, R. and Welling, M. *Causal effect inference with deep latent-variable models.*
- Shi, C., Blei, D. and Veitch, V. *Adapting neural networks for the estimation of treatment effects.*

▶ `https://github.com/py-why/dowhy`

5. Causal Inference: Counterfactuals Counterfactuals. Computing counterfactuals. Probability of necessity and sufficiency.

- Pearl, J., Glymour, M. and Jewell, N.P.. *Causal inference in statistics: A primer.* Section 4.1-4.2 (Counterfactuals and their computation).
- Peters, J., Janzing, D. and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms.* Section 6.4 (Counterfactuals).
- Pearl, J.. *Causality.* Section 9.2.1, 9.3.1, 9.3.2 (Probability of necessity and sufficiency)
- (Darwiche A. *Causality: Counterfactuals*, `https://www.youtube.com/watch?v=BAQIXS8dvaU`)
- (Pearl, J. *Which Patients are in Greater Need: A counterfactual analysis with reflections on COVID-19*, `https://causality.cs.ucla.edu/blog/index.php/2020/04/02/which-patients-are-in-greater-need-a-counterf`

6. Causal Discovery: Independence-based. Causal discovery assumptions. PC. FCI. Functional assumptions. ANM and LiNGAM.

- Peters, J., Janzing, D. and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms.* Section 7.1 (Identifiability under functional assumptions), 7.2.1 (Independence based-methods).
- Glymour, C., Zhang, K. and Spirtes, P., 2019. *Review of causal discovery methods based on graphical models.*
- (Spirtes, P., Glymour, C. and Scheines, R., 2001. Causation, prediction, and search. Section 5.4.2 (PC algorithms))

- (Shimizu, S., Hoyer, P.O., Hyvärinen, A., Kerminen, A. and Jordan, M., 2006. *A linear non-Gaussian acyclic model for causal discovery.*)

▶ `https://causal-learn.readthedocs.io/en/latest/`

7. Causal Discovery: Functional-based methods and score-based methods Score-based methods. Graph metrics. GES. Unconstrained optimization. NoTears.

- Peters, J., Janzing, D. and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms.* Section 7.2.2 - 7.2.5 (Score-based methods).
- Peters, J. and Bühlmann, P. *Structural intervention distance for evaluating causal graphs.*
- Zheng, X., Aragam, B., Ravikumar, P.K. and Xing, E.P., 2018. *Dags with no tears: Continuous optimization for structure learning.*

8. Causality and Machine Learning Causality for ML. Causal and anti-causal learning. Robust/Invariant learning. Neural causal models.

- Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K. and Mooij, J.. *On causal and anticausal learning.*
- Xia, K., Lee, K.Z., Bengio, Y. and Bareinboim, E., 2021. *The causal-neural connection: Expressiveness, learnability, and inference.*

9. Causal Bandits and Causal Reinforcement Learning Bandits and causality. MABUC. Parallel bandits. RL and causality.

- Bareinboim, E., Forney, A. and Pearl, J., 2015. *Bandits with unobserved confounders: A causal approach.*
- Buesing, L., Weber, T., Zwols, Y., Racaniere, S., Guez, A., Lespiau, J.B. and Heess, N. *Woulda, coulda, shoulda: Counterfactually-guided policy search.*
- (Bareinboim, E. *Causal Reinforcement Learning*, `https://crl.causalai.net/`)

10. Causal Representation Learning and Causal Abstraction CRL. Causal Component Analysis. Causal abstraction. Interventional consistency.

- Wendong, L., Kekić, A., von Kügelgen, J., Buchholz, S., Besserve, M., Gresele, L. and Schölkopf, B. *Causal component analysis.*
- Rubenstein, P.K., Weichwald, S., Bongers, S., Mooij, J.M., Janzing, D., Grosse-Wentrup, M. and Schölkopf, B., 2017. *Causal consistency of structural equation models.*

Released: 17th Jan

Reading Week 1

Z-INF - Causality
Spring 2025

Introduction to causality

Weekly reading

In this week we will introduce informally the core ideas that justify and found the theory of *causality*. We will make our first steps into causality by considering some emblematic scenarios where traditional statistical analysis leads us to paradoxical conclusions, such as the **Simpson's paradox** [Pearl et al., 2016, Sec 1.2]. This paradox will introduce us to the idea of **confounding** (or **common cause**). Confounders are a central concept (and a bane) of causal analysis: when a presumed cause and a presumed effect are both influenced by a common cause, estimating a relation of cause-effect, if one exists, is challenging; this is even more problematic if the confounder can not be observed (*unobserved* or *latent confounder*). Confounders are thus a core theoretical concept in causal theory and a major practical obstacle in causal analysis. Confounder also play an important role in the fact that *causation is not correlation*; **Reichenbach's principle** explains how two things can be correlated without being in a causal relation [Peters et al., 2017, Sec 1.3].

Next, in preparation for delving into causal theory, we will need to review ideas from three fields: *statistics*, *machine learning* and *graph theory*. Statistics provides the notions to deal with data and uncertainty, such as *variables*, *probability distributions*, *expected values* and *variances* [Pearl et al., 2016, Sec 1.3.1-1.3.9]. Machine learning provides concepts for learning and induction, such as *regression* and *empirical risk minimization* [Pearl et al., 2016, Sec 1.3.10-1.3.11] and [Peters et al., 2017, Sec 1.1-1.2]. Graph theory provides tools to express structural knowledge, such as *nodes*, *edges*, *parenthood* and *acyclicity* [Pearl et al., 2016, Sec 1.4].

Finally, relying on these preliminary notions, we will see a first informal definition of a model that will allow us to deal formally with causality: a **structural causal model** (SCM) [Pearl et al., 2016, Sec 1.5]. SCMs are not the only way to formalize causality, but they provide a powerful and versatile language that has found large adoption in machine

learning. A couple of examples in [Peters et al., 2017, Sec 1.4] will illustrate concrete SCMs and discuss the relation of SCMs to other forms of modelling. Importantly, [Peters et al., 2017, Sec 1.4] will also provide an intuitive idea of another central concept for causality: **interventions**. Interventions capture the idea of interacting with a system and performing experiments to understand its inner workings; as such, interventions will be a very important tool to understand relations of cause and effect - indeed the very idea of causality we adopt is sometimes called an *interventionist* description of causality.

Coding

Coding offers a powerful way to study causal models: it allows us to implement models and explore how they behave causally. Indeed, we can think of SCMs as programs [Ibeling and Icard, 2020, Sec "Probabilistic Programs"]. You are invited to implement from scratch Simpson's paradox [Pearl et al., 2016, Sec 1.2] and/or simple causal models [Pearl et al., 2016, SCM 1.5.1].

Optional reading

While SCM has become the main formalism used in machine learning, a similar formalism called **structural equation models** (SEM) has been used in economics for some time. [Pearl, 2015, Sec 1] reviews the historical debt of SCMs towards SEMs, as well as the contribution of a Norwegian pioneer, Trygve Haavelmo. Importantly, this paper underlines the mechanistic aspect of SEMs and the importance of interventions.

References

Duligur Ibeling and Thomas Icard. Probabilistic reasoning across the causal hierarchy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10170–10177, 2020.

Judea Pearl. Trygve Haavelmo and the emergence of causal calculus. *Econometric Theory*, 31(1):152–179, 2015.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: a primer*. John Wiley & Sons, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: Foundations and learning algorithms*. MIT Press, 2017.

Released: 24th Jan

Reading Week 2

Z-INF - Causality
Spring 2025

Graphical models

Weekly reading

In this week we will formalize the idea of *causal models* through *graphical models*. Remember how during the previous week we reviewed ideas from statistics and graph theory. **Graphical models** are models that allow us to join together statistics and graph theory, providing an intuitive and cheap way to represent structured objects with a stochastic (or uncertain) behavior.

First, we will review some additional notions from graph theory; in particular, we will study three basic graphical structures: **chains**, **forks** and **colliders** [Pearl et al., 2016, Sec 2.1-2.3]. We can use these elementary structures to graphically evaluate a key graphical properties of independence between variables: **d-separation** [Pearl et al., 2016, Sec 2.4].

Crucially, since graphical models join the statistical and the graphical worlds, we want agreement on important statistical and graphical properties. For instance, we want the graphical notion of *d-separation* to agree with the statistical notion of *independence*. The property of **Markovianity** capture one part of this requirement: if d-separation holds graphically, then independence holds statistically; by construction, a SCM is Markovian: thus, if we identify graphically a d-separation on the DAG of a SCM, we know the corresponding statistical independence must hold. Markovianity has several formulations clearly summarized in [Peters et al., 2017, Sec 6.5.1-6.5.2]. The property of **faithfulness** captures the other part of the requirement: if independence holds statistically, then d-separation holds graphically; in general, faithfulness does not always hold for SCMs and must be assumed; given faithfulness, if we identify a statistical independence from data from the SCM, then we can identify the compatible graphical structures that satisfy d-separation [Peters et al., 2017, Sec 6.5.3].

Finally, we sum up our work with graphical models by going through a series of progressively more complex graphical models: **Bayesian networks** which encode joint statistical distributions [Pearl, 2009, Sec 1.2.1-1.2.2]; **causal Bayesian networks** which allow for interventions [Pearl, 2009, Sec 1.3]; and **structural causal models** which provide a complete causal model able to deal with observations, interventions and counterfactuals [Pearl, 2009, Sec 1.4-1.4.3]. Critically, [Pearl, 2009, Sec 1.4.2] introduces another slightly different use of the term *Markovianity*: a **Markovian SCM** is an acyclic model with no unobserved confounders (simple setting); a **semi-Markovian SCM** is an acyclic model with unobserved confounders (challenging setting).

Coding

`https://pgmpy.org/` is a standard Python library to encode graphical models like Bayesian networks and causal Bayesian networks. You are invited to use it to implement simple models like [Pearl, 2009, Fig 1.2,1.4] and consider how the distributions change under conditioning and intervention.

References

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: a primer*. John Wiley & Sons, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: Foundations and learning algorithms*. MIT Press, 2017.

Released: 31st Jan

Reading Week 3

Z-INF - Causality
Spring 2025

Causal Inference: Identifiability (from interventions to do-calculus)

Weekly reading

In this week we will start working with the actual theory of causality; specifically we will introduce the problem of **causal inference**: if we are given the graphical structure of a model and data from the model, can we estimate the causal effect of one variable on another? Notice that this question is fundamental whenever we want to successfully control a system.

We start by reviewing the ideas we have explored so far and solidifying them into the **Pearl causal hierarchy** [Bareinboim et al., 2022]. This hierarchy identifies three separate layers: *L1 - observational questions* formalized in Bayesian networks; *L2 - interventional questions* formalized in causal Bayesian networks; and *L3 - counterfactual question* formalized in structural causal models [Bareinboim et al., 2022, Sec 1.1, 1.2, 1.4 - up to 1.4.3.1]. Whereas L1 is the traditional domain of statistics, causal inference will deal with L2.

The first step into causal inference is to formalize the idea of interventions [Pearl et al., 2016, Sec 3.1]. We can conceive of an intervention as an operator that mutilates the graphical structure of an SCM and induces a *new post-intervention model*.

Causal inference on L2 means dealing with *interventional queries*, that is evaluating quantities that involve a *do-operator*. At the heart of this problem lies the idea that a good strategy to answer an interventional query including this new do-operator is to try to reduce it to a traditional statistical quantity in the post-intervention model. If we could perform the required intervention and observe the post-intervention model, answering would be trivial (up to the standard statistical challenges). This is what happens, for instance, in the case of **randomized experiments** [Pearl et al., 2016, Sec 3.1]. However, often we want to answer an interventional query without performing

an intervention but relying only on observational data. This gives rise to two central questions of causal inference:

- **Identifiability:** given the structure of an SCM and observational data, is it possible to answer an interventional query?
- **Identification:** assuming identifiability, how can we estimate the interventional query of interest.

We will consider at first the problem of identifiability. As illustrated by examples in previous weeks, estimating causal effects from observational data means controlling for confounding, that is, finding some *adjustment* of the observational data that would match the interventional quantity we want to estimate. In Markovian SCMs (i.e.: no unobserved confounders) we can *always* identify an interventional query by selecting a proper **adjustment set** [Pearl et al., 2016, Sec 3.2] or relying on the **truncated product formula** (or **g-formula**) [Pearl et al., 2016, Sec 3.2.2]. In semi-Markovian SCMs (i.e.: with unobserved confounders) we cannot always identify an interventional query and we need, instead, to reason about the structure of the graph to decide whether we can find a correct adjustment; the **backdoor adjustment** is an intuitive criterion that defines what variables we should control in order to estimate our interventional query [Pearl et al., 2016, Sec 3.3].

It turns out that the *backdoor adjustment* is just a specific form of adjustment that can be derived using more general rules; these rules, called **do-calculus**, can be always used to decide whether an interventional query can be reduced to observational data [Peters et al., 2017, Sec 6.7].

Coding

Code is a handy tool to develop an intuitive and grounded understanding of interventions. Representing an SCM as a program (as discussed in RW1), how would you implement interventions? How would you collect observational and interventional data from your SCM/program? `https://causalfusion.net/app` is an interactive online platform where you can draw graphical models, define interventional queries, and evaluate whether they are identifiable or not. You are invited to try out the tool and experiment with identifiability.

Optional reading

An intuitive and graphical illustration of interventions and their effect on distributions is available online at [Huszar, 2019]. An intuitive explanations of the do-calculus rules and the derivation of the backdoor criterion is also available online at [Heiss, 2021]. Do-calculus is indeed complete [Shpitser and Pearl, 2008] and thus allow us to always decide when an L2 query can be reduced to an L1 query.

References

Elias Bareinboim, Juan D Correa, Duligur Ibeling, and Thomas Icard. *On Pearl's Hierarchy and the Foundations of Causal Inference*. 2022.

Andrew Heiss. Do-calculus adventures! exploring the three rules of do-calculus in plain language and deriving the backdoor adjustment formula by hand, 2021. URL https://www.andrewheiss.com/blog/2021/09/07/do-calculus-backdoors/.

Ferenc Huszar. Causal inference 2: Illustrating interventions via a toy example, 2019. URL https://www.inference.vc/causal-inference-2-illustrating-interventions-in-a-toy-example/.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: a primer*. John Wiley & Sons, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: Foundations and learning algorithms*. MIT Press, 2017.

Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9(9), 2008.

Released: 7th Feb

Reading Week 4

Z-INF - Causality
Spring 2025

Causal Inference: Identification (from matching to ML methods)

Weekly reading

During the previous week we have discussed the problem of identifiability (*can we answer an interventional query from observational data?*); this week we will move into the problem of identification (*how do we answer an interventional query from observational data?*).

Before looking at the identifiability problem, however, we will take a look at the language of **potential outcomes** (PO) [Peters et al., 2017, Sec 6.9]. PO is an alternative formalism to SCMs which does not rely on graphical models, and a relevant amount of work on identification has been carried out using this language.

We will then use PO to express the scenario in which a treatment T is provided to a patient with certain features X, and an outcome Y is observed. We then want to estimate the **average treatment effect** (ATE). To estimate the proper causal quantity, we need to control for confounders in X; we can either work with the actual features X or compute a synthetic descriptor called *propensity score* $g(X)$ to summarize information in the features X into a low-dimensional representation. [Austin, 2011] discusses the main ways of using propensity scores to estimate ATE: *matching*, *stratification*, *inverse probability*, *covariate adjustment*; all these methods are based on a similar intuition - how can we align individuals with different treatments but similar characteristics, so that by aligning similar characteristics we obtained a balanced data set, as we would in a randomized control trial.

Beyond these standard statistical methods to estimate ATE, *machine learning models* can also be deployed to answer causal queries. In the Markovian setting (no hidden confounder), [Shi et al., 2019] describes a neural network architecture based on TARNET

designed to estimate ATE; the key idea being to use different streams within a neural networks to estimate the quantities necessary to compute the ATE.

Coding

https://github.com/py-why/dowhy is a library providing a set of algorithms for causal effect estimation (and more). You are invited to appreciate the workflow of the library and run the tutorial examples.

Optional reading

The distinction between identifiability and identification may be subtle: identifiability cares about the existence and the uniqueness of the solution, identification cares about actual property of the estimator (e.g., variance of the estimator) [Maclaren and Nicholson, 2019]. SCM and PO rely on different explicit assumptions, but it can be proved that they have the same expressibility [Galles and Pearl, 1998]; however, the difference in language (e.g., reliance of graphs) might make one of the two formalism more or less useful in specific situations. Consider, for instance, the discussion on the selection of covariates in [Austin, 2011, Sec "Variable selection for the propensity score model"]; this discussion makes sense because of the unavailability of a graphical structure; if we had a graphical structure, do-calculus will allow us to decide on the correct set of covariate for adjustments; see [Cinelli et al., 2020] for some examples of good and bad choices. Beyond [Shi et al., 2019], other ML approaches have been used for estimating causal effects in different settings; for instance, [Louizos et al., 2017] consider a semi-Markovian setting (with hidden confounder) and use a VAE-based solution to infer the confounders and then estimate the causal effect.

References

Peter C Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424, 2011.

Carlos Cinelli, Andrew Forney, and Judea Pearl. A crash course in good and bad controls. 2020.

David Galles and Judea Pearl. An axiomatic characterization of causal counterfactuals. *Foundations of Science*, 3:151–182, 1998.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6446–6456, 2017.

Oliver J Maclaren and Ruanui Nicholson. What can be estimated? identifiability, estimability, causal inference and ill-posed inverse problems. *arXiv preprint arXiv:1904.02826*, 2019.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: Foundations and learning algorithms*. MIT Press, 2017.

Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32, 2019.

Released: 14th Feb

Reading Week 5

Z-INF - Causality
Spring 2025

Causal Inference: Counterfactuals

Weekly reading

During this week we move from the interventional layer (L2) to the last layer of Pearl's hierarchy: the **counterfactual layer** (L3).

Counterfactuals are a thorny philosophical and scientific concept. In general, given that something has happened in reality (*factual*), there is no way to know what would have happened if something different had happened (*counterfactual*). From this standpoint, counterfactual has no reality. Look back at causal inference too: when we want to compute the ATE, we want to estimate the difference between providing a treatment or not; however, we have a fundamental *missing data problem*: for each individual, we either give the treatment or we do not give the treatment; thus, we can not compute an **individual ATE**; what we could do was to group individuals who got the treatment and individuals who did not and, controlling for confounding, estimating a **population ATE**.

Now counterfactuals promise the ability of computing causal quantities on an individual level. But to make counterfactuals computable we need the mechanistic assumptions of an SCM. We can then see how counterfactuals are given a precise meaning in an SCM and we can review the procedure to compute a counterfactual based on the following steps: *abduction* (compute the exact state of the world when an event happened by inferring the value of all the exogenous variables), *action* (instantiate the *closest possible alternative world* by performing the only counterfactual action we want to evaluate), *prediction* (compute the outcome of the counterfactual in the alternative world) [Pearl et al., 2016, Sec 4.1.-4.2], [Peters et al., 2017, Sec 6.4]. Notice that, when your model describes an individual and you interpret the value of the exogenous variables as all the unobserved factors that characterize an individual, a counterfactual exactly correspond to evaluating an individualized causal quantity as opposed to a population causal quantity.

Interestingly, counterfactual allows us to compute the probability of causation: we can define quantities such as the *probability of necesssity* (how necessary is a cause for an effect?) or the *probability of sufficiency* (how sufficient is a cause for an effect?) [Pearl, 2009, Sec 9.2.1]. These quantities allow us to apply causal reasoning often more broadly, modelling for instance questions pertaining responsibility and accountability, as illustrated in the example of the coin toss and the firing squad [Pearl, 2009, Sec 9.3.1-9.3.2].

Coding

As we did in RW3, we can take advantage of code as a handy tool to develop now a grounded understanding of counterfactuals. Consider again an SCM as a program, how would you implement counterfactuals? That is, how would you implement the *abduction-intervention-prediction* algorithm for counterfactual computation?

Optional reading

A neat introduction to counterfactuals is given online by [Darwiche, 2022]. For an example of reasoning in terms of counterfactuals, see for instance [Mueller and Pearl, 2020]. Computing counterfactuals is not trivial and it can be performed using *twin networks* [Pearl, 2009, Sec 7.1.4]. Also, in the same way we reasoned about the identifiability and the identification of causal/interventional queries in RW3 and RW4, we can reason about the identifiability and the identification of causal/counterfactual queries; counterfactual queries are in general harder to answer, and often only bounds can be provided; see as an example [Pearl, 2009, Sec 8.3].

References

Adnan Darwiche. Causality: Counterfactuals part a, 2022. URL https://www.youtube.com/watch?v=BAQIXS8dvaU.

Scott Mueller and Judea Pearl. Which patients are in greater need: A counterfactual analysis with reflections on covid-19, 2020. URL https://causality.cs.ucla.edu/blog/index.php/2020/04/02/which-patients-are-in-greater-need-a-counterfactual-analysis-with-reflections-on-c

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: a primer*. John Wiley & Sons, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: Foundations and learning algorithms*. MIT Press, 2017.

Released: 21st Feb

Reading Week 6

Z-INF - Causality
Spring 2025

Causal Discovery: Independence-based and functional-based

Weekly reading

We now leave behind the problem of *causal inference* (can we compute causal quantities given data and causal graph?) and move to the problem of **causal discovery/causal structure learning** (can we learn a causal graph if we only have data?).

This problem defines a new *identifiability problem*: is data sufficient to identify uniquely the causal structure that generated it? It is a legitimate question; after all, we might wonder whether observational data would allow us to learn a causal model that is concerned also with interventional and counterfactual quantities.

There must exist some degree to which causal structure can be identified. Remember our discussion in RW2 about properties connecting statistical and structural aspects of a graphical model; properties like *faithfulness* guarantee that independences in the data must hold in the graphical model as well. This simple observation gives rise to a first family of approaches for causal discovery, that is, **independence-based methods**. These methods are based on discovering independencies in the data and then construct a graphical structure that guarantees these independencies (in terms of d-separation). Unfortunately, though, this approach can rarely pinpoint a single graphical structure: given a set of independencies, there are usually multiple graphs that respect those independencies; this group of graphs is called the **Markov equivalence class**, and it is the theoretical limit of identifiability for independence-based methods. Practically, these methods suffer from other computational limitations such as the sensitivity of the statistical test for independences and the need to run several statistical tests [Peters et al., 2017, Sec 7.2.1].

Part of the difficulty of identifying the causal structure of an SCM follows from the how general is the definition of an SCM - indeed we never make any **assumption on**

the form of the exogenous noise or the endogenous functions. If we could restrict the domains over which SCMs are defined we could gain in identifiability. Interestingly, linear SCMs with Gaussian noise remain, in general, non-identifiable (thanks to the possibility of fitting different forms of Gaussians to the noise), but **linear SCMs with Gaussian noise with equal variances** and **linear SCMs with non-Gaussian noise** (LiNGaM), for instance, are identifiable [Peters et al., 2017, Sec 7.1].

As a summary, Glymour et al. [2019] provides a comprehensive review of algorithms for causal discovery, including *independence-based methods*, *functional-based methods* (which we have just discussed) and *score-based methods* (topic of RW7).

Coding

`https://causal-learn.readthedocs.io/en/latest/` is a library providing a set of algorithms for causal discovery. You are invited to try out different structure learning algorithms. Some datasets for causal discovery can be found online. An interesting one is `https://webdav.tuebingen.mpg.de/cause-effect/` requiring to perform causal discovery over two variables (X, Y); despite the minimal number of variables, this is a very challenging problem.

Optional reading

Purely observational causal discovery is, in general, an underdetermined problem: two models may generate the same observational distributions and still imply different interventional or counterfactual distributions. A natural extension of the problem we have considered is learning also with *interventional data* or in *multiple environments* [Peters et al., 2016], which leads to new *interventional Markov equivalence classes* [Yang et al., 2018]. Prototypical independence-based algorithms for causal discovery are *PC algorithms* discussed in [Spirtes et al., 2001, Sec 5.4.2]. A seminal algorithm for causal discovery based on functional assumptions is [Shimizu et al., 2006].

References

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: Foundations and learning algorithms*. MIT Press, 2017.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2001.

Karren Yang, Abigail Katcoff, and Caroline Uhler. Characterizing and learning equivalence classes of causal dags under interventions. In *International Conference on Machine Learning*, pages 5541–5550. PMLR, 2018.

Released: 28th Feb

Reading Week 7

Z-INF - Causality
Spring 2025

Causal Discovery: Score-based

Weekly reading

During the previous week we have considered two families of methods for graph discovery: *independence-based methods* and *methods based on assumptions on the form of the noise and/or functions*. We now consider a last family of approaches: **score-based methods**.

Score-based methods are based on the definition of a *score function* that can evaluate how well a candidate learned graph fit the available data. Then, the causal discovery problem turns into a *search problem* for the graph achieving the highest score. The search problem in *NP-hard*, but standard heuristics can be applied to solve the problem efficiently (although not exactly) [Peters et al., 2017, Sec 7.2.2].

A critical bottleneck of score-based methods is that they need to move through the discrete space of acyclic graphs. Combinatorial discrete problems are known for being hard to solve, and evaluating acyclicity at every step turns out to be very costly. An alternative approach was introduce by *NoTears*, an algorithm that characterizes acyclicity through a continuous measure [Zheng et al., 2018]. This approach defines a sub-family of **continuous score-based methods** which can be efficiently solved with classical optimization approaches.

Incidentally, notice that evaluating the quality of a learned causal structure is a non-trivial issue. Classical graph-theoretical measures compute the distance between two graphs (e.g.: the ground truth DAG and a learned DAG) via *structural Hamming distance* (SHD), that is the difference in edges between two graphs. However, this measure has no causal meaning, and alternative measures, such as *structural interventional distance* (SID) which account for differences in evaluating interventional quantities, might be more meaningful [Peters and Bühlmann, 2015, Sec 1-2.3].

Coding

A public implementation of NoTears is available at https://github.com/xunzheng/notears. You are invited to compare it against other algorithms from the *causal-learn* library.

Optional reading

A reference algorithm for score-based causal discovery is [Chickering, 2002], while modern approaches often try to improve over continuous score-based methods, see for instance [Massidda et al., 2023]. Our discussion has covered only traditional methods that rely on observational data to estimate a causal graph. As mentioned in RW6, in reality, it might be possible to incorporate interventional data. *Experimental design* considers the problem of determining the best interventions to be taken in order to learn the true causal structure of a system.

References

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Riccardo Massidda, Francesco Landolfi, Martina Cinquini, and Davide Bacciu. Constraint-free structure learning with smooth acyclic orientations. *arXiv preprint arXiv:2309.08406*, 2023.

Jonas Peters and Peter Bühlmann. Structural intervention distance for evaluating causal graphs. *Neural computation*, 27(3):771–799, 2015.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: Foundations and learning algorithms*. MIT Press, 2017.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Released: 14th Mar

Reading Week 8

Z-INF - Causality
Spring 2025

Causality and Machine Learning

Weekly reading

We now move on to consider more closely how *causality* and *machine learning* might interact. There are two main ways in which the two fields may relate:

- **Machine Learning for Causality**: this refers to using ML methods and techniques to solve efficiently causal problems. We have seen examples of this in RW4 with Dragonnet (deploying a neural network to compute ATE) or in RW7 with NoTears (using optimization techniques to learn a causal structure).
- **Causality for Machine Learning**: this refers mainly to relying on causal model and reasoning to improve the learning of models. We will discuss this more in detail now.

Some of the advantages of relying on causality in learning models should already be evident at this point: *unbiasedness* (in estimating effects by controlling for confounders), *robustness* (by accounting for environmental changes through interventions) or *interpretability* (by providing an understandable graphical model).

To see more concrete implications, we will look to the simple case of learning a function between two variables (X, Y). If we can determine the direction of causality (say, $X \to Y$), then we can distinguish two function learning setting: a **causal learning** (that is, learning $f(X) = Y$) and an **anti-causal learning** (that is, learning $f(Y) = X$). Remember that, if we are concerned only with prediction, both tasks might be meaningful. Just distinguishing between causal and anti-causal learning can allow us to decide how our learned function will behave under distribution shifts or in a semi-supervised setting [Schölkopf et al., 2012].

A real challenge in using causal models in machine learning is due to their discrete nature which limits their scalability. Substantial interest is concentrated in how SCMs

may be married with the horsepower of machine learning, neural networks, while preserving the guarantees provided by rigorous causal reasoning. A recent proposal are **neural causal models** (NCM) which defines causal models based on neural networks [Xia et al., 2021].

In the next weeks (RW9, RW10) we will see more specific intersections of ML and causality.

Coding

Some old traditional datasets have been sorted in causal and anticausal tasks here `https://pl.is.tue.mpg.de/p/causal-anticausal/`. You can assess how this form of causal knowledge may be of help.

Optional reading

The use and the benefits of causality in machine learning is clearly too broad of a topic to be dealt in one week (or even in one course). Some more perspectives may be gained by reading position papers such as [Schölkopf, 2019] or by checking more extensive surveys like [Kaddour et al., 2022].

References

Jean Kaddour, Aengus Lynch, Qi Liu, Matt J Kusner, and Ricardo Silva. Causal machine learning: A survey and open problems. *arXiv preprint arXiv:2206.15475*, 2022.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012.

Bernhard Schölkopf. Causality for machine learning. *arXiv preprint arXiv:1911.10500*, 2019.

Kevin Xia, Kai-Zhan Lee, Yoshua Bengio, and Elias Bareinboim. The causal-neural connection: Expressiveness, learnability, and inference. *Advances in Neural Information Processing Systems*, 34:10823–10836, 2021.

Released: 21st Mar

Reading Week 8

Z-INF - Causality
Spring 2025

Causal Representation Learning and Causal Abstraction

Weekly reading

From this week onwards we abandon the core field of causality (theory of structural causal models, causal inference, causal discovery, causal machine learning) and we consider more advanced topics in causality. The choice is somewhat arbitrary and it implies that many other relevant topics will not be covered (e.g.: *causality in time-series*, *experimental design*, *mediation analysis*, *causal fairness*, *causal interpretability*, *causal transportability*...).

Through the course, we have learned that causal graphs are essential for causal analysis; also, we have reviewed several approaches to *causal discovery* aimed at learning such causal graphs. This week we will consider approaches that push the boundaries of causal discovery further.

Recall that the task of causal discovery is aimed at learning a causal graph from data. An implicit assumption is that the variables observed in the data automatically identify the relevant variables/nodes of the causal graph we want to learn. Thus, in a sense, causal discovery learns *only* the edges of a causal graph. **Causal representation learning** (CRL) drops this assumption. In CRL we deal with high-dimensional data (think of the screenshots of a videogame) with a causal dynamics generated at a lower-level (think of the logic of a videogame acting on game variables, not on pixels). The task of CRL is therefore to learn a causal graph over the low-dimensional representation of the data. This implies that in CRL we need to learn both the variables/nodes *and* the edges of a causal model.

In CRL we are confronted with the challenge of learning both *mixing functions* (mapping low-level latent variables to high-level observed variables) and *causal functions* (mapping low-level causal variables to each other). Assumptions, interventional data, and/or parametric constraints are required to make learning feasible. [Wendong et al.,

2023] propose a method based on *independent component analysis* that can provide some guarantees on *identification*.

The ideal outcome of a causal discovery algorithm is a single causal graph that explains the causal dynamics of a system. Yet scientists often rely on multiple models of the same system at different levels of resolution to explain different behaviours (think about how the thermodynamic behavior of a gas might be explained microscopically and macroscopically). **Causal abstraction** (CA) studies how we can relate multiple causal models, how we can assess their degree of approximation, and how we can jointly exploit them; **CA learning** specifically looks at how relations of abstraction between different causal models can be learned from data.

CA learning also presents significant challenges as it might require learning *abstraction functions* (mapping a refined model to a coarse model) and, possibly, *causal functions* (mapping causal variables in the coarse model to each other). Again, assumptions, interventional data, and/or parametric constraints are required to make learning feasible. [Zennaro et al., 2023] propose a method based on the *relaxation of a combinatorial problem* and its solution via *gradient descent* in order to learn a mapping between two given SCMs.

Coding

Code for both the papers in the weekly reading is publicly available online. You can have a look at the code and test it out.

Optional reading

An influential position paper on causal representation learning is [Schölkopf et al., 2021]; for a learning approach based on a (linear) parametric assumption see, for instance, [Squires et al., 2023]. A seminal paper on causal abstraction is [Rubenstein et al., 2017] with its successive refinement in [Beckers and Halpern, 2019]; practical applications of causal abstraction include learning surrogate models [Dyer et al., 2023] and neural networks interpretability [Geiger et al., 2021].

References

Sander Beckers and Joseph Y Halpern. Abstracting causal models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2678–2685, 2019.

Joel Dyer, Nicholas Bishop, Yorgos Felekis, Fabio Massimo Zennaro, Anisoara Calinescu, Theodoros Damoulas, and Michael Wooldridge. Interventionally consistent surrogates for agent-based simulators. *arXiv preprint arXiv:2312.11158*, 2023.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586, 2021.

Paul K Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal consistency of structural equation models. In *33rd Conference on Uncertainty in Artificial Intelligence (UAI 2017)*, pages 808–817. Curran Associates, Inc., 2017.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Chandler Squires, Anna Seigal, Salil S Bhate, and Caroline Uhler. Linear causal disentanglement via interventions. In *International conference on machine learning*, pages 32540–32560. PMLR, 2023.

Liang Wendong, Armin Kekić, Julius von Kügelgen, Simon Buchholz, Michel Besserve, Luigi Gresele, and Bernhard Schölkopf. Causal component analysis. *Advances in Neural Information Processing Systems*, 36:32481–32520, 2023.

Fabio Massimo Zennaro, Máté Drávucz, Geanina Apachitei, W. Dhammika Widanage, and Theodoros Damoulas. Jointly learning consistent causal abstractions over multiple interventional distributions. In *2nd Conference on Causal Learning and Reasoning*, 2023. URL https://openreview.net/forum?id=RNs7aMS6zDq.

Released: 28th Mar

Reading Week 10

Z-INF - Causality
Spring 2025

Causal Bandits and Causal Reinforcement Learning

Weekly reading

In this last week we will engage with another advanced topic in causality, that is, **causal decision-making**. Causal decision-making considers the problem of how knowledge of the causal dynamics of a system can inform us towards better choices and interventions.

We start considering the simplest decision-making problem: the *multi-armed bandit problem* (MAB) [Sutton and Barto, 2018, Sec 2.1-2.4]. While standard MABs assume that our actions and their outcomes are independent, it would be very reasonable to postulate the existence of a causal system that mediates our actions and relates possible outcomes; for instance, in the typical example of providing drugs, it seems reasonable to state that the outcomes of different drugs are not independent but are mediated by an identical causal system (the organism of patients). This leads to **causal MABs** (CMABs), a version of MABs where actions and their outcomes are not independent anymore but related through an underlying causal model.

The relevance of performing causal reasoning and, as always, dealing with confounders, is remarked by [Bareinboim et al., 2015]: if the system you are trying to optimize for is affected by *unobserved confounders*, then standard MAB algorithms are bound to be sub-optimal. Instead, an algorithm relying on causal notions can be used to learn the true optimal solution.

Introducing a time-dimension in the decision-making problems lead from the simple MAB to a *full reinforcement learning problem* (RL) [Sutton and Barto, 2018, Sec 3.1-3.3]. An RL problem has a natural expression in causal terms, since it implies an agent taking action (i.e.: performing interventions) in a given environment (i.e.: on a causal system). Thus, in a **causal RL** setting, a decision-making agent could rely on causal reasoning to improve its outcomes; this might include learning a causal model, inferring the outcomes of actions/interventions from the model, aggregate interventional data

generated by different policies, simulating counterfactual trajectories.

One of the first proposals for using counterfactual reasoning in reinforcement learning is [Buesing et al., 2018], which relied on causal modelling to generate counterfactual trajectories from recorded real data.

Coding

The scenario proposed in [Bareinboim et al., 2015] is very simple: it is recommended to try to implement it in order to get a better understanding of how unobserved confounders do indeed affect even a simple decision-making problem.

Optional reading

A more generic initial treatment of CMABs has been provided by [Lattimore et al., 2016]. A purely graphical approach to reduce the number of actions in a CMABs has been suggested in [Lee and Bareinboim, 2018]. An alternative causal approach to decision-making based on Gaussian processes is discussed in [Aglietti et al., 2020]. `https://crl.causalai.net/` provides a neat introduction to the intersection of causality and reinforcement learning.

References

Virginia Aglietti, Xiaoyu Lu, Andrei Paleyes, and Javier González. Causal bayesian optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3155–3164. PMLR, 2020.

Elias Bareinboim, Andrew Forney, and Judea Pearl. Bandits with unobserved confounders: A causal approach. *Advances in Neural Information Processing Systems*, 28:1342–1350, 2015.

Lars Buesing, Theophane Weber, Yori Zwols, Sebastien Racaniere, Arthur Guez, Jean-Baptiste Lespiau, and Nicolas Heess. Woulda, coulda, shoulda: Counterfactually-guided policy search. *arXiv preprint arXiv:1811.06272*, 2018.

Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: Learning good interventions via causal inference. In *Advances in Neural Information Processing Systems*, pages 1181–1189, 2016.

Sanghack Lee and Elias Bareinboim. Structural causal bandits: where to intervene? In *Advances in Neural Information Processing Systems*, pages 2568–2578, 2018.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT Press, 2018.