A Brief Introduction to Causal Inference

Son Phuc Nguyen^{1,*}

¹Department of Mathematics, University of Economics and Law, Vietnam National University - Ho Chi Minh city

ABSTRACT Causal inference has been of interest in economics for many decades with a great deal of notable work like the Granger's causality which directly lead to a Nobel Prize in Economics. The question of cause and effect is of paramount importance in making high-stake decisions such as economic policies. Besides, in the last ten years, causal inference in artificial intelligence has gradually become a mainstream with remarkable work such as the do-calculus by Judea Pearl. In this paper, we would like to discuss some fundamental ideas in causal inference.

KEYWORDS

Causal inference
causality
machine learning
economics
econometrics

DOI: <https://doi.org/10.53901/ezfy6oj557r>

ARTICLE HISTORY

Received: 01.06.2020
Accepted: 30.06.2020
Online: 10.08.2020

INTRODUCTION

A commonsensical idea about causation is that causal relationships are relationships that are potentially exploitable for purposes of manipulation and control: very roughly, if C is genuinely a cause of E , then if I can manipulate C in the right way, this should be a way of manipulating or changing E . Therefore, causal inference, to certain extent, answer the question "What if?". What would happen in the future if we change our course of action? In retrospect, what would have happened in the the past if we had made a different choice in the distant past? Human usually face situations with various options. There are always pros and cons in any choice of action, therefore, there is a demand for inferring possible scenarios to assist the decision making process.

In contrast to a common belief that causality is a recent innovation, the ideas dates back to founders of modern statistics like Neyman and Fischer. Neyman and Fischer asserted the importance of randomized experiments in assessing the effect of any treatment. In addition, standard assumption like SUTVA (Stable-Unit-Treatment Value Assumption) was used by Neyman although he did not call SUTVA in his work.

A deeper question is the question of cause and effect, and naturally it is a hard question in most situations. Physics has studied this type of questions for many years. In fact, physicists have tried to describe the dynamics of nature via mathematical tools like differential equations. Causal inference is in general less rigid than physics in the sense that it provides a systematic understanding of certain phenomena but do not seek to find laws that govern

the dynamics of the phenomena. On the other hand, compared to statistics, causal inference place more emphasis on understanding the relations between quantities of interest, more emphasis of the "effect" of a variable on another one.

In recent years, causation has been drawing increasing attentions in both machine learning and economics communities. One of the major incentives comes from recent applications which involve more and more areas with high-stakes decisions such as health care, financial technology and law. In high-stakes decision making, understanding risk is of priority, therefore, there is an ongoing shift towards more stable solutions (with risk evaluations) from the traditional performance-driven solutions (predictive ability is the only criterion).

In this short paper, we would like to give a somewhat bird's eye view of the field to the best of our knowledge. Nowadays, there are two major communities that contribute greatly to the recent advancements of causal inference. In the next section, we will introduce main ideas of the two approaches

PRELIMINARIES

Let us recall some fundamentals in probability that form the fundamental tool in causal inference

Independence

Two events A and B are called *independent* if

$$P(A \cap B) = P(A) \cdot P(B)$$

More generally, $A_1, \dots, A_n$ are called independent if for every subset $S \subseteq \{1, \dots, n\}$

$$P\left(\bigcap_{i \in S} A_i\right) = \prod_{i \in S} P(A_i)$$

Note that pairwise independence does not imply independence.

Two real-valued random variables X and Y are called independent

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

if for every $a, b \in \mathbb{R}$, the events $\{X \leq a\}$ and $\{Y \leq b\}$ are independent.

Equivalently, in terms of densities

$$f(x, y) = f(x)f(y)$$

Note that if $X \perp\!\!\!\perp Y$ then

$$E[XY] = E[X]E[Y] \quad (1)$$

$$\text{Cov}(X, Y) = 0 \quad (2)$$

But, the converse is not true.

The key concept that plays vital role in causal inference is *conditional independence*. Two random variables X and Y are conditionally independent given a third random variable Z

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

if

$$f(x, y \mid z) = f(x \mid z)f(y \mid z)$$

for all x, y and for all z such that $f(z) > 0$

Note that there are pairs X, Y which are conditionally independent given Z but unconditionally dependent, and vice versa.

Intervention

Informally, intervention means we can set the cause C to a specific value c_0 and via that action, manipulate the effect E as we wish. The first question is, of course, to identify which variable is the cause and which variable is the effect. Let us consider an example from [1]: Let A be the altitude and T be the temperature of a city. Suppose we have collected a sample of cities in a certain country which allows us to estimate the joint density $p(a, t)$. There are two ways to express $p(a, t)$

$$p(a, t) = p(a \mid t) \cdot p(t) \quad (3)$$

$$= p(t \mid a) \cdot p(a) \quad (4)$$

In statistics, the uses of the two equations are somewhat indistinguishable, in the sense that, which formula being used only depends on which variable is observed.

In causation, the two formulas are quite different since the former describes T as the cause and A as the effect while the latter does the opposite, namely, A being the cause and T being the effect. In other word, the causal graph which illustrates the first equation is $T \rightarrow A$ while the causal graph which illustrates the second equation is $A \rightarrow T$.

Therefore, considering intervention, with the graph $T \rightarrow A$, we would change the temperature of a certain city with the hope that its altitude will change accordingly. On the other hand, with the graph $A \rightarrow T$, we would change the altitude of a city with the hope that the temperature would change. In practice, we cannot change either the altitude or the temperature of a fixed city but we can certainly find another the city in the same country (a proxy) with comparable causal values and observe the difference in the

effect variable. For this particular example, by common sense, we are inclined to assume the graph $A \rightarrow T$ to be the correct causal direction. Unfortunately, not all situations has a clear-cut choice of causal direction, with the chicken-and-egg paradox being an example.

Finding a reasonable causal graph as the example above is the goal of causal discovery. However, it is necessary to rephrase and postulate some aspects of our intuition, namely

1. It is possible to perform a *localized intervention* on A , in other word, to change $p(a)$ without changing $p(t \mid a)$
2. $p(a)$ and $p(t \mid a)$ are *autonomous* or *invariant* mechanisms in the world.

To be more specific, Assumption 1 states the *independence of cause and mechanism*, namely, in the causal factorization $p(a, t) = p(t \mid a)p(a)$, we would expect that the conditional density $p(t \mid a)$ (viewed as a function of t and a) provides no information about the marginal density function $p(a)$. This holds true if $p(t \mid a)$ is a model of a physical mechanism that does not care about what distribution $p(a)$ we feed into it. In other words, the mechanism is not influenced by the ensemble of cities to which we apply it.

Assumption 2 is the so-called *Independent mechanisms* principle which states (see [1]), the conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other conditional distributions. In case we have only two variables, this reduces to an independence between the cause distribution and the mechanism producing the effect distribution.

CAUSAL INFERENCE AND DIRECTED GRAPH

A more general question than the causal direction above is *identification*. Suppose there are a set of variables of interest, some are observed and some are not, what are the causal relationship among them? In [2], [1], a directed acyclic graph (DAG) is used to express the relationship among variables. For example, in a simple linear regression y is expressed as a linear function of x

$$y = ax + b$$

In this form, we implicitly assume that x causes y , in other words, because x changes, y changes. In DAG, the graph looks like

$$x \rightarrow y$$

Nevertheless, x can also be expressed as a linear function of y

$$x = my + n$$

So, is there a possibility that, in fact, the DAG should look like

$$y \rightarrow x \quad ?$$

In practice, domain knowledge provides the direction of the arrow. However, there are chicken-and-egg situation where it is not clear which way the arrow should go. In [3], Reichenbach introduced the *Common Cause Principle*; if two variables x and y are statistically dependent, then there exists a variable z that causally influences both and explains all the dependence between x and y . In other words, conditioned on z , x and y are independent.

$$P(x, y \mid z) = P(x \mid z)P(y \mid z)$$

Finding common cause is a problem in *causal discovery*, see [4]. However, without any further assumption, it is not possible to distinguish the three DAG's below

$$x \longrightarrow y \quad (5)$$

$$y \longrightarrow x \quad (6)$$

$$y \longleftarrow z \longrightarrow y \quad (7)$$

To be more precise, relying on only observational data, the distributions on x and y that can be realized by these models are the same in all three DAGs, see [5].

Fortunately, it is not hopeless to identify the graphical structure if the causal structure implies nontrivial conditional independence in a more-variable situation. Therefore, the authors in [6], [7] propose direct functional *parent-child* relationships rather than conditionals. The key insight is a merger of probabilistic graphical models and the notions of interventions. The quote below is from [6], page 104

We played around with the possibility of replacing the parents-child relationship $P(X_i|PA_i)$ with its functional counterpart $X_i = f_i(PA_i, U_i)$ and, suddenly, everything began to fall into place: We finally had a mathematical object to which we could attribute familiar properties of physical mechanisms instead of those slippery epistemic probabilities $P(X_i|PA_i)$ with which we had been working so long in the study of Bayesian networks.

To be clear, given a set of observable quantities $x_1, \dots, x_n$ (modeled as random variables) associated with vertices of a directed acyclic graph G , the structural causal model (SCM) assumes that each x_i is the result of a functional form

$$x_i = f_i(PA_i, u_i)$$

where PA_i is the set of parents of x_i , f_i is a deterministic function, and u_i is a stochastic unexplained variable. In this model, directed edges represent direct causations and the noises u_i 's are jointly independent which allows the identification of the formula above via conditional probability $P(x_i|PA_i)$. These adjustment formulas ensure causal sufficiency (in brief, the property that causation can be expressed via conditional probability distributions) which let us compute the joint distribution $p(x_1, \dots, x_n)$. This distribution has structural properties inherited from the graph [6]; in particular, it satisfies the *causal Markov condition* stating that conditioned on its parents, each x_i is independent of its non-descendants. It should be noted that the structural causal model makes it straightforward to formalize *interventions* by setting x_i to a constant or by changing u_i . Besides, there is a causal factorization as follows

$$p(x_1, \dots, x_n) = \prod_{i=1}^n p(x_i | x_{i+1}, \dots, x_n)$$

CAUSAL INFERENCE AND POTENTIAL OUTCOMES

Causation has been studied in economics and econometrics for many years. However, the emphases are somewhat different in this setting. In section , we are very much concerned about the directions of the edges in the directed acyclic graph. In economics, nevertheless, qualitative studies provide the direction of causality or the direction of edges in the graph. In other words, economists know the cause and the outcomes, the primary questions are about the magnitude of the causal effects and the possible presence of unmeasured confounders.

Figure 1 Potential outcome example

One very popular approach in economics is the potential outcome (PO) framework which is also known as the Rubin causal model [8], [9]. There are three components of the PO approach:

1. There is a treatment/cause that can take on different values for each unit. Let x denote a value of the treatment. Then, each unit in the population has a collection of potential outcomes $Y(x)$ for any x . Only one of these $y(x)$ is observed. Think about a drug X , a patient can either take it or not take it. The unobserved $Y(x)$ are counterfactuals. Mathematically, the situation can be expressed as

$$Y^{\text{obs}} = Y(X) = \sum_x Y(x) \mathbf{1}_{X=x}$$

The effects correspond to comparisons of the potential outcomes, in particular, between the observed one and other missing ones.

2. To accommodate the first component, we need multiple units which receive the possible values of treatments. Think about volunteers for clinical trials of a new drug. They can either take or not take the drug.
3. The last component is an assignment mechanism that determines which units receive which treatments. Think about a method to assign volunteers to the control group or treatment group.

To keep things simple, usually a binary treatment is considered and the main question is to estimate the average treatment effect for the entire population or some subpopulation.

$$\tau = E[Y_i(1) - Y_i(0)]$$

In machine learning, in particular the *do*-calculus by Pearl et al., the formula above is written as $\tau = E[Y(\text{do}(1)) - Y(\text{do}(0))]$. Up to this point the economic approach and the machine learning approach are by and large similar.

Let us consider a simple example as in Figure 2 where the confounder X is known a priori and the treatment variable is binary, i.e. $T = 0$ or $T = 1$. In other words, we divide the samples into the control group ($T = 0$) and the treated group ($T = 1$). The goal is to estimate the *average causal effect of treatment (ATE)*

$$ATE = E[Y(T = 1) - Y(T = 0)]$$

Since, for each individual, only $Y(T = 0)$ or $Y(T = 1)$ can be observed (but not both), we have to reason about the world that doesn't exist (called the counterfactual world). To make it possible, we need to assume that everything in the counterfactual world is the same as in the real world, except the treatment. One mechanism which accommodates that assumption is randomized experiments.

There are various methods to ensure the same confounder X such as *matching* and *propensity score*:

1. *Matching* ([10]): The goal is to identify pairs of treated ($T = 1$) and control ($T = 0$) units whose confounders X are similar or even identical to each other

$$Distance(X_i, X_j) \leq \epsilon$$

Paired units guarantee that everything else are approximately constant. A small ϵ produces less bias but more variance. This method is appropriate for low-dimensional settings but not for high-dimensional settings since it is a lot harder to find matches.

2. *Propensity score* ([11]): The propensity score $e(X)$ is the probability of a unit to get treated

$$e(x_i) = P(T = 1 | X = x_i)$$

In [11], it can be proved that

$$T \perp\!\!\!\perp X | e(X)$$

Hence,

$$T \perp\!\!\!\perp (Y(1), Y(0)) | e(X)$$

Propensity score is not observed; it has to be estimated. Supervised learning can be used for this task, namely, predicting a label T based on observed covariates X . Then, pairs can be matched by distance between propensity score

$$Distance(X_i, X_j) = |\hat{e}(X_i) - \hat{e}(X_j)|$$

CAUSAL INFERENCE AND POTENTIAL OUTCOMES

Causation has been studied in economics and econometrics for many years. However, the emphases are somewhat different in this setting. In section , we are very much concerned about the directions of the edges in the directed acyclic graph. In economics, nevertheless, qualitative studies provide the direction of causality or the direction of edges in the graph. In other words, economists know the cause and the outcomes, the primary questions are about the magnitude of the causal effects and the possible presence of unmeasured confounders.

One very popular approach in economics is the potential outcome (PO) framework which is also known as the Rubin causal model [8], [9]. There are three components of the PO approach:

1. There is a treatment/cause that can take on different values for each unit. Let x denote a value of the treatment. Then, each unit in the population has a collection of potential outcomes $Y(x)$ for any x . Only one of these $y(x)$ is observed. Think about a drug X , a patient can either take it or not take it. The unobserved $Y(x)$ are counterfactuals. Mathematically, the situation can be expressed as

$$Y^{\text{obs}} = Y(X) = \sum_x Y(x) \mathbf{1}_{X=x}$$

The effects correspond to comparisons of the potential outcomes, in particular, between the observed one and other missing ones.

2. To accommodate the first component, we need multiple units which receive the possible values of treatments. Think about volunteers for clinical trials of a new drug. They can either take or not take the drug.

Figure 2 Potential outcome example

3. The last component is an assignment mechanism that determines which units receive which treatments. Think about a method to assign volunteers to the control group or treatment group.

To keep things simple, usually a binary treatment is considered and the main question is to estimate the average treatment effect for the entire population or some subpopulation.

$$\tau = E[Y_i(1) - Y_i(0)]$$

In machine learning, in particular the *do*-calculus by Pearl et al., the formula above is written as $\tau = E[Y(do(1)) - Y(do(0))]$. Up to this point the economic approach and the machine learning approach are by and large similar.

Let us consider a simple example as in Figure 2 where the confounder X is known a priori and the treatment variable is binary, i.e. $T = 0$ or $T = 1$. In other words, we divide the samples into the control group ($T = 0$) and the treated group ($T = 1$). The goal is to estimate the *average causal effect of treatment* (ATE)

$$ATE = E[Y(T = 1) - Y(T = 0)]$$

Since, for each individual, only $Y(T = 0)$ or $Y(T = 1)$ can be observed (but not both), we have to reason about the world that doesn't exist (called the counterfactual world). To make it possible, we need to assume that everything in the counterfactual world is the same as in the real world, except the treatment. One mechanism which accommodates that assumption is randomized experiments.

There are various methods to ensure the same confounder X such as *matching* and *propensity score*:

1. *Matching* ([10]): The goal is to identify pairs of treated ($T = 1$) and control ($T = 0$) units whose confounders X are similar or even identical to each other

$$Distance(X_i, X_j) \leq \epsilon$$

Paired units guarantee that everything else are approximately constant. A small ϵ produces less bias but more variance. This method is appropriate for low-dimensional settings but not for high-dimensional settings since it is a lot harder to find matches.

2. *Propensity score* ([11]): The propensity score $e(X)$ is the probability of a unit to get treated

$$e(x_i) = P(T = 1 | X = x_i)$$

In [11], it can be proved that

$$T \perp\!\!\!\perp X | e(X)$$

Hence,

$$T \perp\!\!\!\perp (Y(1), Y(0)) \mid e(X)$$

Propensity score is not observed; it has to be estimated. Supervised learning can be used for this task, namely, predicting a label T based on observed covariates X . Then, pairs can be matched by distance between propensity score

$$Distance(X_i, X_j) = |\hat{e}(X_i) - \hat{e}(X_j)|$$

CONCLUSION

In this short paper, we have gone over a few key ideas in causal inference including some common methods in practice. Eventually, we rely on statistical and machine learning methods to do all the estimation. However, causation usually goes beyond that, requiring domain knowledge of the situation, especially in the identification problem. For applications that emphasize stability and generalization of solutions, a combination of data-driven methods and domain expertise is necessary to guarantee desired results.

REFERENCES

- [1] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference*. The MIT Press, 2017.
- [2] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, 2018.
- [3] Hans Reichenbach. *The direction of time*, volume 65. Univ of California Press, 1956.
- [4] Ruocheng Guo, Lu Cheng, Jundong Li, P Richard Hahn, and Huan Liu. A survey of learning causality with data: Problems and methods. *arXiv preprint arXiv:1809.09337*, 2018.
- [5] Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696, 2009.
- [6] Judea Pearl et al. Models, reasoning and inference. Cambridge, UK: CambridgeUniversityPress, 2000.
- [7] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.
- [8] Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.
- [9] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.
- [10] Paul R Rosenbaum and Donald B Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, 79(387):516–524, 1984.
- [11] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.