Causality

Ricardo Silva
Centre for Computational Statistics and Machine Learning,
University College London

October 27, 2014

Definition

The main task in causal inference is the prediction of the outcome of an intervention. For example, a treatment assigned by a doctor that will change the patient's heart condition is an intervention. Predicting the change in patient condition is a causal inference task. In general, an intervention is an action taken by an external agent that changes the original values, or the probability distributions, of some of the variables in the system. Besides predicting outcomes of actions, causal inference is also concerned with explanation: identifying which were the causes of a particular event that happened in the past.

Motivation and Background

Many problems in machine learning are prediction problems. Given a feature vector $\mathbf{X}$, the task is to provide an estimate of some output vector $\mathbf{Y}$, or its conditional probability distribution $P(\mathbf{Y}|\mathbf{X})$. This typically assumes that the distribution of $\mathbf{Y}$ given $\mathbf{X}$ during learning is the same distribution at prediction time. There are many scenarios where this is not the case.

Epidemiology and several medical sciences provide counterexamples. Consider two seemingly straightforward learning problems. In the first example, one is given epidemiological data where smokers are clearly more inclined than non-smokers to develop lung cancer. Can I use this data to learn that smoking causes cancer? In the second example, consider a group of patients suffering from a type of artery disease. In this group, those that receive a bypass surgery are likely to survive longer than those that receive a particular set of drugs with no surgery.

There is no fundamental problem on using such datasets to predict the probability of a smoker developing lung cancer, or the life expectancy of someone who went through surgery. Yet, the data does not necessarily tell you if smoking is a cause of lung cancer, or that nationwide the government should promote surgery as the treatment of choice for that particular heart disease. What is going on?

There are reasons to be initially suspicious of such claims. This is well-known in statistics as the expression "association is not causation" (Wasserman, 2004, p. 253). The data generating mechanism for our outcome $\mathbf{Y}$ ("developing lung cancer," "getting cured from artery disease") given the relevant inputs $\mathbf{X}$ ("smoking habit," "having a surgery") might change under an *intervention* for reasons such as follows.

In the smoking example, the reality might be that there are several *hidden common causes* that are responsible for the observed association. A genetic factor, for instance: the possibility that there is a class of genotypes on which people are more likely to pick up smoking *and* develop lung cancer, without any direct causal connection between these two variables. In the artery disease example, surgery might not be the best choice to be made by a doctor. It might have been the case that so far patients in better shape were more daring in choosing, by themselves, the surgery treatment. This *selection bias* will favor surgery over drug treatment, since from the outset the patients that are most likely to improve take that treatment.

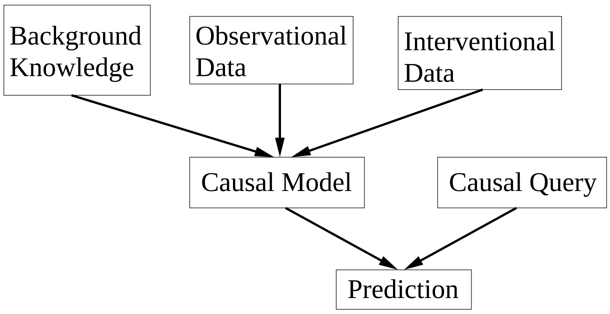
When treatment is enforced by an *external agent* (the doctor), such selection bias disappears, and the resulting $P(\mathbf{Y}|\mathbf{X})$ will not be the same. One way of learning this relationship is through *randomized trials* (Rosenbaum, 2002). The simplest case consists on flipping a coin for each patient on the training set. Each face of the coin corresponds to a possible treatment, and assignment is done accordingly. Since assignment does not depend on any hidden common cause or selection bias, this provides a basis for learning causal effects. Machine learning and statistical techniques can be applied directly in this case (e.g., logistic regression). Data analysis performed with a randomized trial is sometimes called an *interventional study*.

The smoking case is more complicated: a direct intervention is not possible, since it is not acceptable to force someone to smoke or not to smoke.

The inquiry asks only for a *hypothetical intervention*, i.e., if someone is forced to smoke, will his or her chances of developing lung cancer increase? Such an intervention will not take place, but this still has obvious implications in public policy. This is the heart of the matter in issues such as deciding on raising tobacco taxes, or forbidding smoking in public places. However, data that measures this interventional data generation mechanism will never be available for ethical reasons. The question has to be addressed through an *observational study*, i.e., a study for causal predictions without interventional data.

Observational studies arise not only under the impossibility of performing interventions, but also in the case where performing interventions is too expensive or time consuming. In this case, observational studies, or a combination of observational and interventional studies, can provide extra information to guide an experimental analysis (Hyttinen et al., 2013; Sachs et al., 2005; Cooper and Yoo, 1999; Eaton and Murphy, 2007). The use of observational data, or the combination of several interventional datasets, is where the greatest contributions of machine learning to causal inference rest.

Structure of the Learning System



Structure of Causal Inference

In order to use observational data, a causal inference system needs a way of linking the state of the world under an intervention to the *natural* state of the world. The natural state is defined as the one to which no external intervention is applied. In the most general formulation, this link between the natural state and the manipulated world is defined for interventions in any subset of variables in the system.

A common language for expressing the relationship between the different states of the world is a *causal graph*, as explained in more detail in the next section. A causal model is composed of the graph and a probability distribution that factorizes according to the graph, as in a standard graphical model. The only difference between a standard graphical model and a causal graphical model is that in the latter extra assumptions are made. The graphical model can be seen as a way of encoding such assumptions.

The combination of assumptions, observational and interventional data generates such a graphical causal model. In the related problem of reinforcement learning, the agent has to maximize a specific utility function and typically has full control on which interventions (actions) can be performed. Here we will focus on the unsupervised problem of learning a causal model for a fixed input of observational and interventional data.

Because only some (or no) interventional data might be available, the learning system might not be able to answer some causal queries. That is, the system will not provide an answer for some prediction tasks.

Languages and assumptions for causal inference

Directed acyclic graphs (DAGs) are a popular language in machine learning to encode qualitative statements about causal relationships. A DAG is composed of a set of vertices and a set of directed edges. The notion of parents, children, ancestors and descendants are the usual ones found in graphical modeling literature.

In terms of causal statements, a directed edge $A \rightarrow B$ states that A is a *direct* cause of B: that is, different interventions on A will result on different distributions for B, even if we intervene on all other variables. The assumption that A is a cause of B is not used in non-causal graphical models.

A causal DAG G satisfies the *causal Markov condition* if and only if a vertex is independent of all of its non-descendants given its direct causes (parents). In Figure 1(a), A is independent of D, E and F given its parents, B and C. It may or may not be independent of G given B and C.

The causal Markov condition implies several other conditional independence statements. For instance, in Figure 1(a) we have that H is independent of F given A. Yet, this is not a statement about the parents of any vertex.

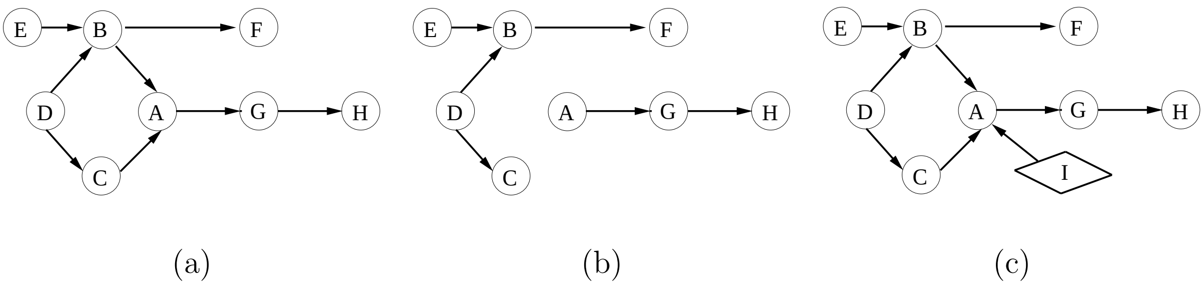


Figure 1: (a) A causal DAG. (b) Structure of the causal graph under some intervention that sets the value of A to a constant. (c) Structure of the causal graph under some intervention that changes the distribution of A.

Pearl's d-separation criterion (Pearl, 2000) is a sound and complete way of reading off independencies, out of a DAG, which are entailed by the causal Markov condition. We assume that the joint probability distribution over the vertice variables is *Markov* with respect to the graph, that is, any independence statement that is encoded by the graph should imply the corresponding independence in the distribution.

Representing Interventions

The local modularity given by the causal Markov condition leads to a natural notion of intervention. Random variable V, represented by a particular vertex in the graph, is following a *local mechanism*: its direct causes determine the distribution of V before its direct effects are generated. The role of an intervention is to *override* the natural local mechanism. An external agent substitutes the natural $P(V|Parents(V))$ by a new distribution $P_{Man}(V|Parents(V))$ while keeping the rest of the model unchanged ("Man" here stands for a particular manipulation). The notion of intervening by changing a single local mechanism is sometimes known as an *ideal intervention*. Other general types of interventions can be defined (Eaton and Murphy, 2007), but the most common frameworks for calculating causal effects rely on this notion.

A common type of intervention is the point mass intervention, which happens when V is set to some constant v. This can be represented graphically by "wiping out" all edges into V. Figure 1(b) represents the resulting graph

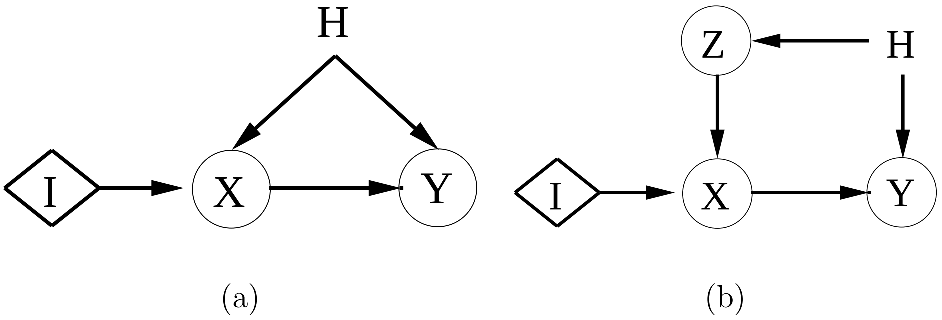


Figure 2: (a) X and Y have a hidden common cause H. (b) Y is dependent on the intervention node I given X, but conditioning on Z and marginalizing it out will allow us to eliminate the "back-door" path that links X and Y through the hidden common cause H.

in (a) under a point manipulation of A. Notice that A is now d-separated from its direct causes under this regime. It is also probabilistically independent, since A is now constant. This allows for a graphical machinery that can read off independencies out of a *manipulated* graph (i.e., the one with removed edges). It is the idea of representing the natural state of the world with a single causal graph, and allowing for modifications in this graph according to the intervention of choice, that links the different regimes obtained under different interventions.

For the general case where a particular variable V is set to a new distribution, a *manipulation node* is added as an extra parent of V: this represents that an external agent is acting over that particular variable (Spirtes et al., 2000; Pearl, 2000; Dawid, 2003), as illustrated in Figure 1(c). $P(V|Parents(V))$ under intervention I is some new distribution $P_{Man}(V|Parents(V), I)$.

Calculating Distributions under Interventions

The notion of independence is a key aspect of probabilistic graphical models, where it allows for efficient computation of marginal probabilities. In causal graphical models, it also fulfils another important role: independencies indicate that the effect of some interventions can be estimated using observational data.

We will illustrate this concept with a simple example. One of the key difficulties in calculating a causal effect is *unmeasured confounding*, i.e., hidden common causes. Consider Figure 2(a), where X is a direct cause of Y, and H is a hidden common cause of both. I is an intervention vertex. Without extra assumptions, there is no way of estimating the effect of X on Y using a training set that is sampled from the observed marginal $P(X, Y)$. This is more easily seen in the case where the model is multivariate Gaussian with zero mean. In this case, each variable is a linear combination of its parents with standard Gaussian additive noise

$$\begin{array}{rcl} X & = & aH + \epsilon_X \\ Y & = & bX + cH + \epsilon_Y \end{array}$$

where H is also a standard normal random variable. The manipulated distribution $P_{Man}(Y|X, I)$, where I is a point intervention setting $X = x$, is a Gaussian distribution with mean $b \cdot x$. Value x is given by construction, but one needs to learn the unknown value b.

One can verify that the covariance of X and Y in the natural state is given by $a + bc$. Observational data, i.e., data sampled from $P(X, Y)$, can be used to estimate the covariance of X and Y in the natural state, but from that it is not possible to infer the value of b: there are too many degrees of freedom.

However, there *are* several cases where the probability of $\mathbf{Y}$ given some intervention on $\mathbf{X}$ can be estimated with observational data and a given causal graph. Consider the graph in Figure 2(b). The problem again is to learn the distribution of Y given X under regime I, i.e., $P(Y|X, I)$. It can be read off from the graph that I and Y are not independent given X, which means $P(Y|X, I) \neq P(Y|X)$. How can someone then estimate $P(Y|X, I)$ if no data for this process has been collected? The answer lies on *reducing the "causal query" to a "probabilistic query"* where the dependence on I disappears (and, hence, the necessity of having data measured under the I intervention). This is done by relying on the assumptions encoded by the graph:

$$\begin{array}{rcll} P(Y|X, I) & = & \sum_z P(Y|X, I, z)P(Z = z|X, I) & \text{(Z is discrete in this example)} \\ & = & \sum_z P(Y|X, z)P(Z = z|X, I) & \text{(Y and I are independent given Z)} \\ & \propto & \sum_z P(Y|X, z)P(X|z, I)P(Z = z|I) & \text{(By Bayes' rule)} \\ & = & \sum_z P(Y|X, z)P(X|z, I)P(Z = z) & \text{(Z and I are marginally independent)} \end{array}$$

In the last line, we have $P(Y|X, Z)$ and $P(Z)$, which can be estimated with observational data, since no intervention variable I appears on the expression. $P(X|Z, I)$ is set by the external agent: its value is known by construction. This means that the causal distribution $P(Y|X, I)$ can be learned even in this case where X and Y share a hidden common cause H.

There are several notations for denoting an interventional distribution such as $P(Y|X, I)$. One of the earliest was due to Spirtes et al. (2000), which used the notation $P(Y|set\ X = x)$ to represent the distribution under an intervention I that fixed the value of X to some constant x. Pearl (2000) defines the operator *do* with an analogous purpose:

$$P(Y|do(X = x)) \tag{1}$$

Pearl's *do*-calculus is essentially a set of operations for reducing a probability distribution that is a function of some intervention to a probability distribution that does not refer to any intervention. All reductions are conditioned on the independencies encoded in a given causal graph. This is in the same spirit of the example presented above.

The advantage of such notations is that, for point interventions, they lead to simple yet effective transformations (or to a report that no transformation is possible). Spirtes et al. (2000) and Pearl (2000) provide a detailed account of such prediction tools. By making a clear distinction between $P(Y|X)$ (X under the natural state) and $P(Y|do(X))$ (X under some intervention), much of the confusion that conflates causal and non-causal predictions disappears.

It is important to stress that methods such as the *do*-calculus are nonparametric, in the sense that they rely on conditional independence constraints only. More informative reductions are possible if one is willing to provide extra information, such as assuming linearity of causal effects. For such cases, other parametric constraints can be exploited (Spirtes et al., 2000; Pearl, 2000).

Learning Causal Structure

In all of the previous section, we assumed that a causal graph was available. Since background knowledge is often limited, learning such graph structures becomes an important task. However, this cannot be accomplished without extra assumptions. To see why, consider again the example in Figure 2(a): if $a + bc = 0$, it follows that the X and Y are independent in the natural state. However, Y is *not* causally independent of X (if $b \neq 0$): $P(Y|do(X = x_1))$ and $P(Y|do(X = x_2))$ will be two different Gaussians with means $b \cdot x_1$ and $b \cdot x_2$, respectively.

This example demonstrates that an independence constraint that is testable by observational data does not warrant causal independence, at least based on the causal Markov condition only. However, an independence constraint that arises from particular identities such as $a + bc = 0$ is not *stable*, in the sense that it does not follow from the qualitative causal relations entailed by the Markov condition: a change in any of the parameter values will destroy such a constraint.

The artificiality of unstable independencies motivates an extra assumption: the *faithfulness* condition (Spirtes et al., 2000), also known as the *stability* condition (Pearl, 2000). We say that a distribution P is faithful to a causal graph G if P is Markov with respect to G, *and* if each conditional independence in P corresponds to some d-separation in G. That is, on top of the causal Markov condition we assume that all independencies in P are entailed by the causal graph G.

The faithfulness condition allows us to reconstruct classes of causal graphs from observational data. In the simplest case, observing that X and Y are independent entails that there is no causal connection between X and Y. Consequently, $P(Y|do(X)) = P(Y|X) = P(Y)$. No interventional data was necessary to arrive at this conclusion, given the faithfulness condition.

In general, the solution is undetermined: more than one causal graph will be compatible with a set of observable independence constraints. Consider a simple example, where data is generated by a causal model with a causal graph given as in Figure 3(a). This graph entails some independencies. For instance, that X and Z are independent given W, or that X and Y are not independent given any subset of $\{W, Z\}$. However, several other graphs entail the same conditional independences. The graph in Figure 3(b) is one example. The learning task is then discovering an *equivalence class* of graphs, not necessarily a particular graph. This is in contrast with the problem of learning the structure of non-causal graphical models: the fact that there are other structures compatible with the data is not important in this case, since we will not use such graphical models to predict the effect of some hypothetical intervention. An equivalence class might not be enough information to reduce a desired causal query to a probabilistic query, but it might require much less prior knowledge than specifying a full causal graph.

Assume for now that no hidden common causes exist in this domain. In particular, the graphical object in Figure 3(c) is a representation of the equivalence class of graphs that are compatible with the independencies encoded in Figure 3(a) (Pearl, 2000; Spirtes et al., 2000). All members of the

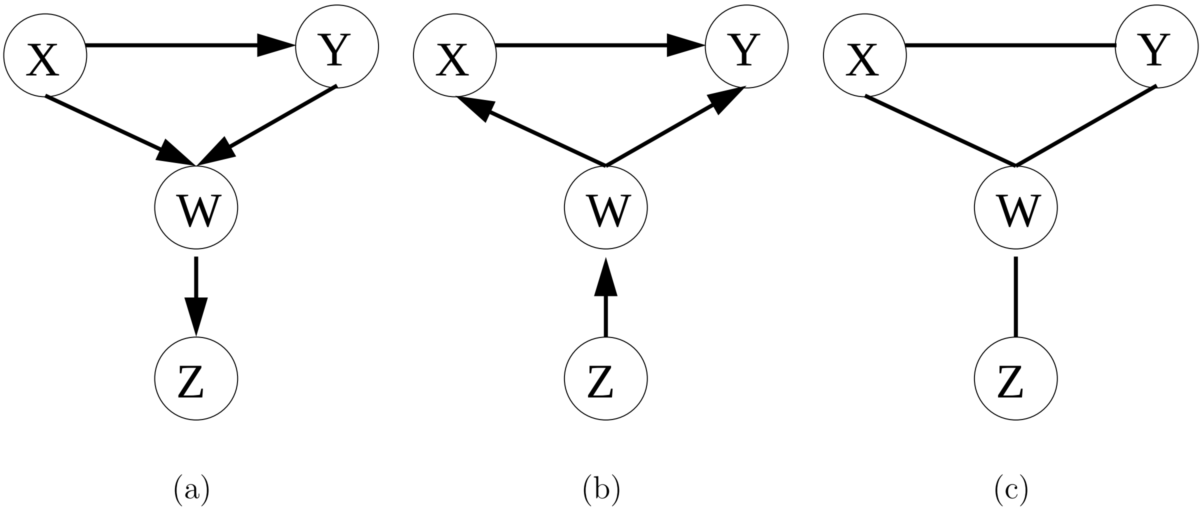


Figure 3: (a) A particular causal graph which entails a few independence constraints, such as X and Z being independent given W. (b) A different causal graph that entails exactly the same independence constraints as in (a). (c) A representation for all graphs that entail the same conditional independencies as (a) and (b).

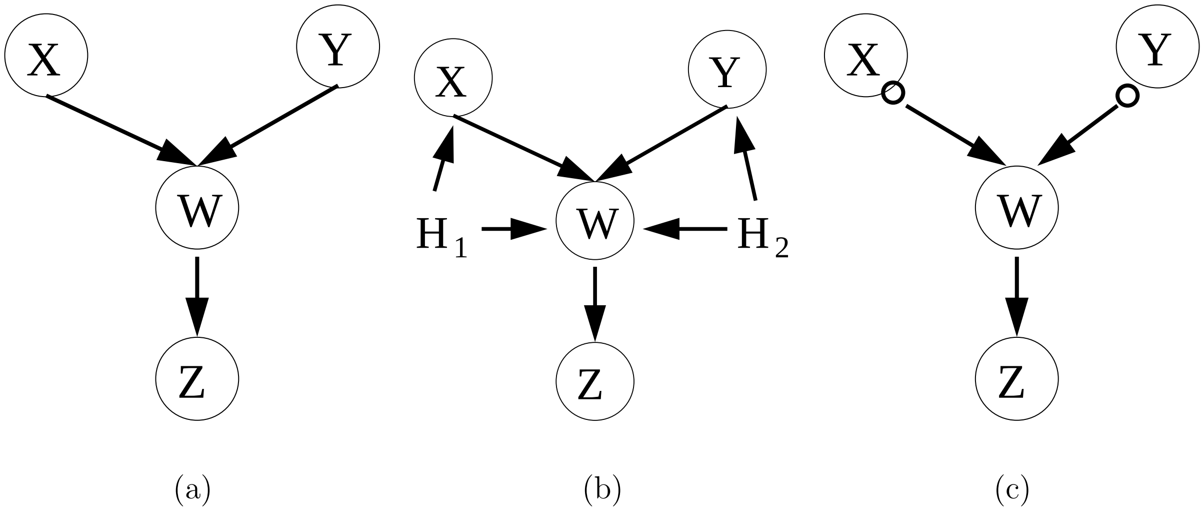


Figure 4: (a) A particular causal graph with no other member on its equivalence class (assuming there are no hidden common causes). (b) Graph under the presence of two hidden common causes H_1 and H_2. (c) A representation for all graphs that entail the same conditional independencies as (a), without assuming the non-existence of hidden common causes.

equivalence class will have the same *skeleton* of this representation, i.e., the same adjacencies. An undirected edge indicates that there are two members in the equivalence class where directionality of this particular edge goes in opposite directions. Some different directions are illustrated in Figure 3(b). One can verify from the properties of d-separation that, if an expert or an experiment indicates that $X - W$ should be directed as $X \rightarrow W$, then the edge $W - Z$ is *compelled* to be directed as $W \rightarrow Z$: the direction $W \leftarrow Z$ is incompatible with the simultaneous findings that X and Z are independent given W, and that X causes W .

More can be discovered if more independence constraints exist. In Figure 4(a), X is not a cause of Y. If we assume no hidden common causes exist in this domain, then no other causal graph is compatible with the independence constraints of Figure 4(a): the equivalence class is this graph only. However, the assumption of no hidden common causes is strong and undesirable. For instance, the graph in Figure 4(b), where H_1 and H_2 are hidden, is in the same equivalence class of (a). Yet, the graph in (a) indicates that $P(W|do(X)) = P(W|X)$, which can be arbitrarily different from the real $P(W|do(X))$ if Figure 4(b) is the real graph. Some equivalence class representations, such as the Partial Ancestral Graph representation (Spirtes et al., 2000), are robust to hidden common causes: in Figure 4(c), and edge that has a circle as endpoint indicates that is not known if there is a causal path into both, e.g., X and W (which would be the case for a hidden common cause of X and W). The arrow into W does indicate that W cannot be a cause of X. A fully directed edge such as $W \to Z$ indicates total information: W is a cause of Z, Z is not a cause of W, and W and Z have no hidden common causes.

Given equivalence class representations and background knowledge, different types of algorithms explore independence constraints to learn an equivalence class. It is typically assumed that the true graph is acyclic. The basic structure is to evaluate how well a set of conditional independence hypotheses is supported by the data. Depending on which constraints are judged to hold in the population, we keep, delete or orient edges accordingly. Some algorithms, such as the PC algorithm (Spirtes et al., 2000), test a single independence hypothesis at a time, and assemble the individual outcomes in the end into an equivalence class representation. Other algorithms such as the GES algorithm (Meek, 1997; Chickering, 2002) start from a prior distribution for graphs and parameters, and proceed to compare the marginal likelihood of members of different equivalence classes (which can be seen as a Bayesian joint test of independence hypotheses). In the end, this reduces to a search for the maximum a posteriori equivalence class estimator. Both PC and GES have consistency properties: in the limit of infinite data, they return the right equivalence class under the faithfulness assumption. However, both PC and GES, and most causal discovery algorithms, assume that there are no hidden common causes in the domain. The Fast Causal Inference (FCI) algorithm of Spirtes et al. (2000) is able to generate equivalence class representations as in Figure 4(c). As in the PC algorithm, this is done by testing a single independence hypothesis at a time, and therefore is a high variance estimator given small samples. A GES-like algorithm with the consistency properties of FCI is not currently known. An algorithm that allows for cyclic networks is discussed by Richardson (1996).

Semiparametric models

Our examples relied on conditional independence constraints. In this case, the equivalence class is known as the *Markov equivalence class.* Markov equivalence classes are "nonparametric", in the sense that they do not refer to any particular probability family. In practice, this advantage is limited by our ability to test independence hypotheses within flexible probability families. Another shortcoming of Markov equivalence classes is that they might be poorly informative if few independence constraints exist in the population. This will happen, for instance, if a single hidden variable is a common cause of all observed variables. If one is willing to incorporate further assumptions, such as linearity of causal relationships, semiparametric constraints can be used to define other types of equivalence classes that are more discriminative than the Markov equivalence class. Silva et al. (2006) describe how some rank constraints in the covariance matrix of the observed variables can be used to learn the structure of linear models, even if no independence constraints are observable. Shimizu et al. (2006) provide a solution to find the causal ordering of a linear DAG model without latent variables, by exploiting information beyond the second moments of a distribution in the non-Gaussian case. Entner et al. (2012) introduce an approach to estimate causal effects in non-Gaussian linear systems under some assumptions of directionality but allowing for unmeasured confounding. Peters et al. (2014) develop a general method for learning directionality in non-linear models with additive noise.

Confidence intervals

Several causal learning algorithms such as the PC and FCI algorithms (Spirtes et al., 2000) are consistent, in the sense that they can recover the correct equivalence class given the faithfulness assumption and an infinite amount of data. Although point estimates of causal effects are important, it is also important to provide confidence intervals. From a frequentist perspective, it has been shown that this is not possible given the faithfulness assumption only (Robins et al., 2003). An intuitive explanation is as follows: consider the model such as the one in Figure 2(a). For any given sample size, there is at least one model such that the association due to the paths $X \leftarrow H \rightarrow Y$ and $X \rightarrow Y$ nearly cancel each other (faithfulness is still preserved), making the covariance of X and Y statistically indistinguishable from zero. In order to achieve uniform consistency, causal inference algorithms need assumptions stronger than faithfulness. Zhang and Spirtes (2003) provide some directions.

Other Languages and Tasks in Causal Learning

A closely related language for representing causal models is the *potential outcomes* framework popularized by Donald Rubin (Rubin, 2005). In this case, random variables for a same variable Y are defined for each possible state of the intervened variable X. Notice that, by definition, only one of the possible Y outcomes can be observed for any specific data point. This framework is popular in the statistics literature as a type of missing data model. The relation between potential outcomes and several other representations of causality is discussed by Richardson and Robins (2013).

A case where potential outcomes become particularly motivated is in *causal explanation*. In this setup, the model is asked for the probability that a particular event in time was the cause of a particular outcome. This is often cast as a *counterfactual question*: had A been false, would B still have happened? Questions in History and law are of this type: the legal responsibility of an airplane manufacturer in an accident depends on technical malfunction being an *actual cause* of the accident. Ultimately, such issues of causal explanation, actual causation and other counterfactual answers, are untestable. Although machine learning can be a useful tool to derive the consequences of assumptions combined with data about other events of the same type, in general the answers will not be robust to changes in the assumptions, and the proper assumptions ultimately cannot be selected with the available data. Some advances in generating explanations with causal models are described by Halpern and Pearl (2005).

References

D. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

G. Cooper and C. Yoo. Causal discovery from a mixture of experimental and observational data. *Proceedings of the 15th conference on Uncertainty in Artificial Intelligencem (UAI-1999)*, pages 116–125, 1999.

A.P. Dawid. Causal inference using influence diagrams: the problem of partial compliance. In P.J. Green, N.L. Hjort, and S. Richardson, editors, *Highly Structured Stochastic Systems*, pages 45–65. Oxford University Press, 2003.

D. Eaton and K. Murphy. Exact Bayesian structure learning from uncertain interventions. *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (AISTATS-2007)*, pages 107–114, 2007.

D. Entner, P.O. Hoyer, and P. Spirtes. Statistical test for consistent estimation of causal effects in linear non-gaussian models. *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS-2012)*, pages 364–372, 2012.

J. Halpern and J. Pearl. Causes and explanations: a structural-model approach. Part II: Explanations. *British Journal for the Philosophy of Science*, 56:889–911, 2005.

A. Hyttinen, F. Eberhardt, and P. O. Hoyer. Experiment selection for causal discovery. *Journal of Machine Learning Research*, 14:3041–3071, 2013.

C. Meek. *Graphical Models: Selecting Causal and Statistical Models.* PhD Thesis, Carnegie Mellon University, 1997.

J. Pearl. *Causality: Models, Reasoning and Inference.* Cambridge University Press, 2000.

J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

T. S. Richardson. A discovery algorithm for directed cyclic graphs. *Proceedings of 12th Conference on Uncertainty in Artificial Intelligence*, 1996.

T.S. Richardson and J. Robins. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. *Working Paper Number 128, Center for Statistics and the Social Sciences, University of Washington*, 2013.

J. Robins, R. Scheines, P. Spirtes, and L. Wasserman. Uniform consistency in causal inference. *Biometrika*, 90:491–515, 2003.

P. Rosenbaum. *Observational Studies.* Springer-Verlag, 2002.

D. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, pages 322–331, 2005.

K. Sachs, O. Perez, D. Pe'er, D. Lauffenburger, and G. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308, 2005.

S. Shimizu, P. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

R. Silva, R. Scheines, C. Glymour, and P. Spirtes. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7: 191–246, 2006.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search.* Cambridge University Press, 2000.

L. Wasserman. *All of Statistics.* Springer-Verlag, 2004.

J. Zhang and P. Spirtes. Strong faithfulness and uniform consistency in causal inference. *Proceedings of the 19th Conference in Uncertainty in Artificial Intelligence (UAI-2013)*, pages 632–639, 2003.