Summary for A Theory of Inferred Causation[Causality Chap.2 (Judea Pearl)]

Shirley Wu
2021.2.1

1 The Causal Discovery Intuition

The clues that we explore in this chapter come from certain patterns of statistical associations that are characteristic of causal organizations – patterns that, in fact, can be given meaningful interpretation only in terms of causal directionality. Consider, for example, the following intransitive pattern of dependencies among three events: A and B are dependent, B and C are dependent, yet A and C are independent. An example satisfying such events will be $A \rightarrow B \leftarrow C$ (In Pearl's favorite example, A and C are the outcomes of two fair coins, and B represents a bell that rings whenever either coin comes up heads.)

2 The Causal Discovery Framework

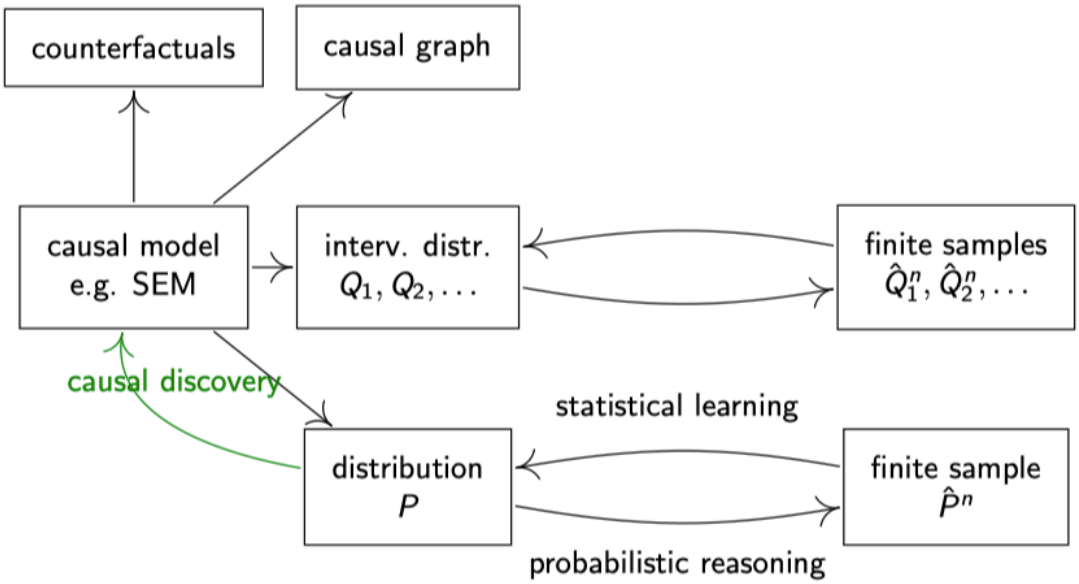

Figure 1: Causal discovery structure, picture from Wenjie Wang

Causal Discovery:

(i) An induction game that scientists play against Nature.

(ii) Nature possesses stable causal mechanisms that are deterministic functional relationships between variables, some of which are unobservable.

(iii) These mechanisms can be organized in the form of an acyclic structure, which the scientist attempts to identify from the available observations.

(iv) One-sentence definition: learning causal relationships from raw data.

Causal Structure:
A causal structure of a set of variables V is a directed acyclic graph (DAG) in which each node corresponds to a distinct element of V, and each link represents a direct functional relationship among the corresponding variables.

Causal Model:
A causal model is a pair $M = \langle D, \Theta_D \rangle$ consisting of a causal structure D and a set of parameters Θ_D compatible with D. The parameters Θ_D assign a function $x_i = f_i\left(pa_i, u_i\right)$ to each $X_i \in V$ and a probability measure $P\left(u_i\right)$ to each u_i, where PA_i are the parents of X_i in D and where each U_i is a random disturbance distributed according to $P\left(u_i\right)$ independently of all other u.

The definitions of causal model reveal that a causal structure serves as a blueprint(backbone) for forming a "causal model" – a precise specification of how each variable is influenced by its parents in the DAG, while that Nature is at liberty to introduce arbitrary(yet mutually independent) disturbances.
The assumption of independent disturbances renders the model Markovian. The Markov condition guides us in deciding when a set of parents PA_i is considered to include all the relevant causes of variable X_i. If a set PA_i in a model is too narrow, there will be disturbance terms that influence several variables simultaneously, and the Markov property will be lost. Such disturbances will be treated explicitly as "latent" variables. The Markov property is restored once we acknowledge the existence of latent variables and represent their existence explicitly as nodes in a graph.

3 Model Preference (Occam's Razor)

Latent Structure:
A latent structure is a pair $L = \langle D, O \rangle$, where D is a causal structure over V and where $O \subseteq V$ is a set of observed variables(and some unobserved variables exist).

Structure Preference:
One latent structure $L = \langle D, O \rangle$ is preferred to another $L' = \langle D', O \rangle$ (written $L \preceq L'$) if and only if D' can mimic D over $O-$ that is, if and only if for every Θ_D there exists a $\Theta'_{D'}$ such that $P_{[O]}\left(\langle D', \Theta'_{D'} \rangle\right) = P_{[O]}\left(\langle D, \Theta_D \rangle\right)$. Two latent structures are equivalent, written $L' \equiv L$, if and only if $L \preceq L'$ and $L \succeq L'$.

Minimality:
A latent structure L is minimal with respect to a class $\mathcal{L}$ of latent structures if and only if there is no member of $\mathcal{L}$ that is strictly preferred to L- that is, if and only if for every $L' \in \mathcal{L}$ we have $L \equiv L'$ whenever $L' \preceq L$.
A set of latent structures are minimal, for they entail the observed independencies and none other.

Consistency:

A latent structure $L = |D, O|$ is consistent with a distribution $\hat{P}$ over O if D can accommodate some model that generates $\hat{P}-$ that is, if there exists a parameterization Θ_D such that $P_{[O]}(\langle D, \Theta_D \rangle) = \hat{P}$

Inferred Causation:
Given $\hat{P}$, a variable C has a causal influence on variable E if and only if there exists a directed path from C to E in every minimal latent structure consistent with $\hat{P}$.

The only assumption invoked in this implication is minimality– models that overfit the data are ruled out. One reason scientists prefer minimality is that such theories are more constraining and thus more falsifiable.

Example:
(a) and (b) in Figure 2 are minimal, for they entail the observed independencies and none other.

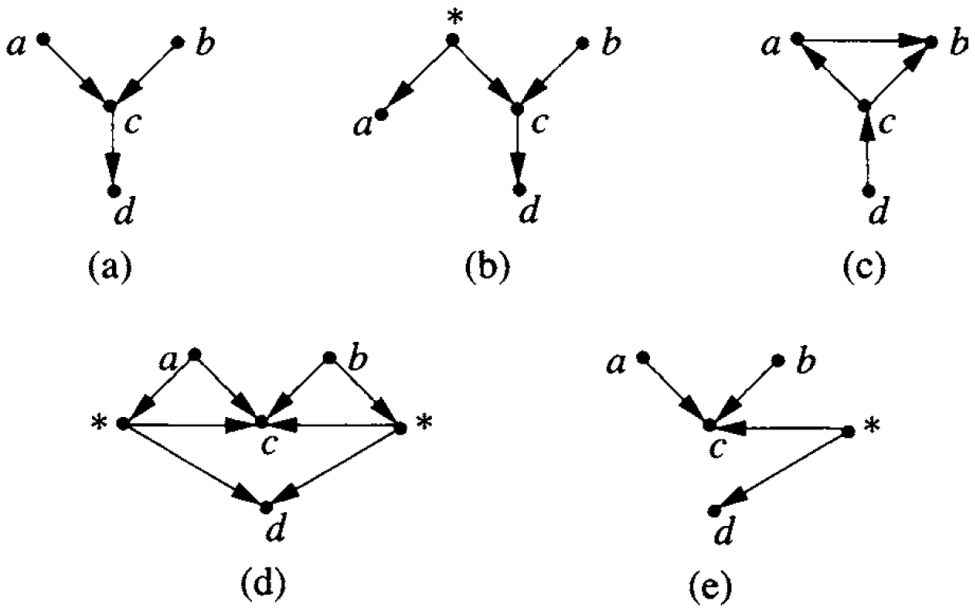

Figure 2: Observations taken over four variables {a, b, c, d} reveal two independencies: "a is independent of b and "d is independent of {a, b} given c. (a = having a cold, b = having hay fever, c = having to sneeze, d = having to wipe one's nose.)

(c) is not minimal for it accommodates distributions with arbitrary relations between a and b. Similarly, (d) is not minimal because it fails to impose the conditional independence between d and {a, b} given c and will therefore accommodate distributions in which d and {a, b} are dependent given c. In contrast, (e) is not consistent with the data, since it imposes an unobserved marginal independence between {a, b} and d.

4 Stable Distributions

Stability:
Let $I(P)$ denote the set of all conditional independence relationships embodied in P. A causal

model $M = \langle D, \Theta_D \rangle$ generates a stable distribution if and only if $P(\langle D, \Theta_D \rangle)$ contains no extraneous independences - that is, if and only if $I(P(\langle D, \Theta_D \rangle)) \subseteq I(P(\langle D, \Theta'_D \rangle))$ for any set of parameters Θ'_D. The stability condition states that, as we vary the parameters from Θ to Θ' no independence in P can be destroyed.

Figure 3: One chair or two chiars?

T1: The object in the picture is a chair.

T2: The object in the picture is either a chair or two chairs positioned such that one hides the other.

Our preference for T_1 over T_2 can be justified on two principles, one based on minimality and the other on stability. The minimality principle argues that T_1 is preferred to T_2 because the set of scenes composed of single objects is a proper subset of scenes composed of two or fewer objects. The stability principle rules out T_2 a priori, arguing that it would be rather unlikely for two objects to align themselves so as to have one perfectly hide the other. Such an alignment would be unstable relative to slight changes in environmental conditions or viewing angle.

For another example, in the structure $Z \leftarrow X \rightarrow Y$, which stands for the relations

$$z = f_1(x, u_1), \quad y = f_2(x, u_2)$$

the variables Z and Y will be independent, conditional on X, for all functions f_1 and f_2. In contrast, if we add an arrow $Z \rightarrow Y$ to the structure and use a linear model

$$z = \gamma x + u_1, \quad y = \alpha x + \beta z + u_2$$

with $\alpha = -\beta\gamma$, then Y and X will be independent. However, the independence between Y and X is unstable because it disappears as soon as the equality $\alpha = -\beta\gamma$ is violated. The stability

assumption presumes that this type of independence is unlikely to occur in the data, that all independencies are structural.

Similarly, consider the chain model $Y \rightarrow X \rightarrow Z$. The reason for us to consider the equality $\rho_{YZ} = \rho_{XZ} \cdot \rho_{YX}$ "stable" and the equality $\alpha = -\beta\gamma$ "accidental" is that—each variable in causal models is determined by a set of other variables through a mechanism that remains invariant when other mechanisms are subjected to external influences. This invariance means that mechanisms can vary independently of one another, which in turns implies that the set of structural coefficients (e.g., α, β, γ) – rather than other types of parameters (e.g., $\rho_{YZ}, \rho_{XZ}, \rho_{YX}$) – can and will vary independently when experimental conditions change.

5 Recovering DAG Structures

Algorithm 1 ***IC Algorithm (Inductive Causation)***

Input: $\hat{P}$, a stable distribution on a set V of variables.
Output: a pattern $H(\hat{P})$ compatible with $\hat{P}$.

(1) For each pair of variables a and b in V, search for a set S_{ab} such that $(a \perp\!\!\!\perp b \mid S_{ab})$ holds in $\hat{P}-$ in other words, a and b should be independent in $\hat{P}$, conditioned on S_{ab}. Construct an undirected graph G such that vertices a and b are connected with an edge if and only if no set S_{ab} can be found.

(2) For each pair of nonadjacent variables a and b with a common neighbor c, check if $c \in S_{ab}$. If it is, then continue. If it is not, then add arrowheads pointing at c ($a \rightarrow c \leftarrow b$).

(3) In the partially directed graph that results, orient as many of the undirected edges as possible subject to two conditions:

(i) Any alternative orientation would yield a new v -structure; or

(ii) Any alternative orientation would yield a directed cycle.

To be more specific, step 3 of the IC algorithm can be systematized into these four rules:

- R_1 : Orient $b - c$ into $b \rightarrow c$ whenever there is an arrow $a \rightarrow b$ such that a and c are nonadjacent.
- R_2 : Orient $a - b$ into $a \rightarrow b$ whenever there is chain $a \rightarrow c \rightarrow b$.
- R_3 : Orient $a - b$ into $a \rightarrow b$ whenever there are two chains $a - c \rightarrow b$ and $a - d \rightarrow b$ such that c and d are nonadjacent.
- R_4 : Orient $a - b$ into $a \rightarrow b$ whenever there are two chains $a - c \rightarrow d$ and $c \rightarrow d \rightarrow b$ such that c and b are nonadjacent and a and d are adjacent.

This algorithum can generates the pattern(graph) $H(\hat{P})$ with variables all observable. Latent structures, however, require special treatment, because the constraints that a latent structure imposes upon the distribution cannot be completely characterized by any set of conditional independence statements.

6 Recovering Latent Structures

Algorithm 2 ***IC Algorithm (Inductive Causation with Latent Variables)***

Input: $\hat{P}$, a stable distribution (with respect to some latent structure).
Output: core $(\hat{P})$, a marked pattern.

(1) For each pair of variables a and b, search for a set S_{ab} such that a and b are independent in $\hat{P}$, conditioned on S_{ab}. If there is no such S_{ab}, place an undirected link between the two variables, $a - b$.

(2) For each pair of nonadjacent variables a and b with a common neighbor c, check if $c \in S_{ab}$. If it is, then continue. If it is not, then add arrowheads pointing at c (i.e., $a \rightarrow c \leftarrow b$).

(3) In the partially directed graph that results, add (recursively) as many arrowheads as possible, and mark as many edges as possible, according to the following two rules:

(i) R_1 : For each pair of nonadjacent nodes a and b with a common neighbor c, if the link between a and c has an arrowhead into c and if the link between c and b has no arrowhead into c, then add an arrowhead on the link between c and b pointing at b and mark that link to obtain $c \xrightarrow{*} b$

(ii) R_2 : If a and b are adjacent and there is a directed path (composed strictly of marked links) from a to b (as in Figure 2.2), then add an arrowhead pointing toward b on the link between a and b.

Example:

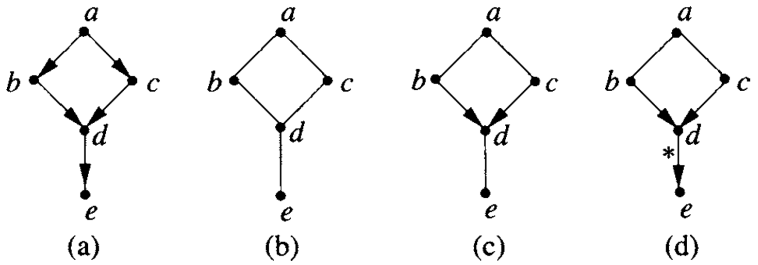

Figure 4: Graphs constructed by the IC* algorithm. (a) Underlying structure. (b) After step 1. (c) After step 2. (d) Output of IC*.

- Step1: $S_{ad} = \{b, c\}, S_{ae} = \{d\}, S_{bc} = \{a\}, S_{be} = \{d\}, S_{ce} = \{d\}$
- Step2: d is not in S_{bc}, thus we add $(b \rightarrow d \leftarrow c)$.
- Step3: b and e has one commmon neighbor d, and the link between b and d has an arrow into d and the link between d and e has no arrowhead into d, then we can add $d \xrightarrow{*} e$. Same does (c, e, d).

7 Local Criteria for Inferring Causal Relations

Potential Cause:
A variable X has a potential causal influence on another variable Y (that is inferable from $\hat{P}$) if

the following conditions hold.

1. X and Y are dependent in every context.
2. There exists a variable Z and a context S such that

(i) X and Z are independent given S(i.e., $X \perp\!\!\!\perp Z \mid S$) and

(ii) Z and Y are dependent given S(i.e., $Z \not\perp\!\!\!\perp Y \mid S$).

By "context" we mean a set of variables tied to specific values. Intuitively, the asymmetrical relationship between Z and X and Y implies the direction of information flow.

Genuine Cause:
A variable X has a genuine causal influence on another variable Y if there exists a variable Z such that either:

1. X and Y are dependent in any context and there exists a context S satisfying

(i) Z is a potential cause of X,

(ii) Z and Y are dependent given S(i.e., $Z \not\perp\!\!\!\perp Y \mid S$), and

(iii) Z and Y are independent given $S \cup X$(i.e., $Z \perp\!\!\!\perp Y \mid S \cup X$); Or

2. X and Y are in the transitive closure of the relation defined in criterion 1 .

Conditions (i)-(iii) are illustrated in Figure4 with $X = d, Y = e, Z = b$, and $S = \varnothing$. The destruction of the dependence between b and e through conditioning on d cannot be attributed to spurious association between d and e; genuine causal influence is the only explanation.

Spurious Association:
Two variables X and Y are spuriously associated if they are dependent in some context and there exist two other variables (Z_1 and Z_2) and two contexts (S_1 and S_2) such that:

1. Z_1 and X are dependent given S_1 ($Z_1 \not\perp\!\!\!\perp X \mid S_1$);
2. Z_1 and Y are independent given S_1 ($Z_1 \perp\!\!\!\perp Y \mid S_1$);
3. Z_2 and Y are dependent given S_2 ($Z_2 \not\perp\!\!\!\perp Y \mid S_2$)
4. Z_2 and X are independent given S_2 ($Z_2 \perp\!\!\!\perp X \mid S_2$).

Conditions 1 and 2 use Z_1 and S_1 to disqualify Y as a cause of X, ; conditions 3 and 4 use Z_2 and S_2 to disqualify X as a cause of Y. This leaves the existence of a latent common cause as the only explanation for the observed dependence between X and Y, as exemplified in the structure $Z_1 \rightarrow X \rightarrow Y \leftarrow Z_2$

These theories become simpler when temporal information is available since every variable preceding and adjacent to X now qualifies as a "potential cause of X.

Genuine Causation with Temporal Information:
A variable X has a causal influence on Y if there is a third variable Z and a context S, both occurring before X, such that:

1. $(Z \not\perp\!\!\!\perp Y \mid S)$;
2. $(Z \perp\!\!\!\perp Y \mid S \cup X)$

Spurious Association with Temporal Information:
Two variables X and Y are spuriously associated if they are dependent in some context S, if X precedes Y, and if there exists a variable Z satisfying:

1. $(Z \perp\!\!\!\perp Y \mid S)$;
2. $(Z \not\perp\!\!\!\perp X \mid S)$.

8 Nontemporal Causation and Statistical Time

Statistical Time:
Given an empirical distribution P, a statistical time of P is any ordering of the variables that agrees with at least one minimal causal structure consistent with P.

(...)

9 Conclusion

How powerful is causal discovery?:
Although statistical analysis cannot distinguish genuine causation from spurious covariation in every conceivable case, in many cases it can. Under the assumptions of model minimality, there are patterns of dependencies that should be sufficient to uncover genuine causal relationships. These relationships cannot be attributed to hidden causes lest we violate one of the basic maxims of scientific methodology: the semantical version of Occam's razor.

How safe are the causal relationships inferred by the IC algorithm?:
Not absolutely safe, but good enough to tell a tree from a house and good enough to make useful inferences without having to touch every physical object that we see. For causal inference, our question amounts to assessing whether there are enough discriminating clues in a typical learning environment to allow us to make reliable discriminations between cause and effect. Rephrased as a logical guarantee, we can categorically assert that the IC* algorithm will never label an arrow $a \rightarrow b$ as genuine if in fact a has no causal influence on b and if the observed distribution is stable relative to its underlying causal model.

How causal discovery could benefit current research in machine learning?:
These method resembles a standard, machine-learning search through a space of hypotheses where each hypothesis stands for a causal model. Unfortunately, even if the training sample exhausts the hypothesis subset, we are still left with a vast number of equivalent causal theories, each stipulating a drastically different set of causal claims. Therefore, fitness to data is an insufficient criterion for validating causal theories. Whereas in traditional learning tasks we attempt to generalize from one set of instances to another, the causal modeling task is to generalize from behavior under one set of conditions to behavior under another set. Causal models should therefore be chosen by a criterion that challenges their stability against changing conditions, and this is indeed what scientists attempt to accomplish through controlled experimentation. Absent such experimentation, the best one can do is to rely on virtual control variables.