10-708: Probabilistic Graphical Models, Spring 2020

18: Causality II

Lecturer: Kun Zhang   Scribe: Xueying Ding, Naveen Shankar, Koyoshi Shindo, Zeyu Tang, Yilin Yang

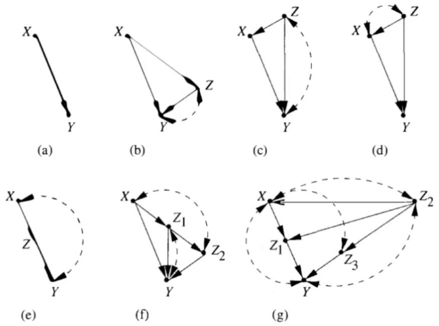
1 Introduction

In this lecture, we will first continue the discussion about causal inference; then we will be talking about causal discovery.

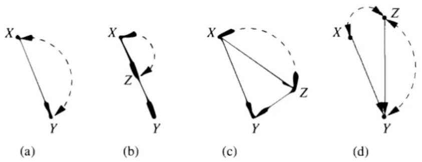
2 Identifying a Causal Effect

A sufficient condition for identifying the causal effect $P(y|do(x))$ from a graph is that there exists no bi-directed path (a path composed entirely of bi-directed arcs) between X and any of its children (Tian and Pearl [2002]).

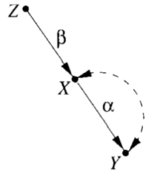
For example, this condition holds in the following graphs:



And does not hold in these graphs:



However, if you assume the system is linear, you can often recover the causal effect even if the condition does not hold. For example, in the following graph:



$\beta = r_{XZ}$ (the regression coefficient of regressing X on Z) and $\alpha\beta = r_{YZ}$. As such, you can recover the true causal effect α by $\alpha = \dfrac{r_{YZ}}{r_{XZ}}$.

3 Propensity Scores

Assuming the back-door criterion (or conditional ignorability) condition holds, the average causal effect (ACE) can be calculated as

$$\text{ACE} = \mathbf{E}[Y|do(x)] - \mathbf{E}[Y|do(x')]$$

Where e.g. x is the treatment condition, x' is the baseline, and $P(Y|do(x)) = \sum_c P(Y|x,c)P(c)$, for some confounding covariate(s) c.

When you have randomized controlled experiments, $P(c)$ is the same across the treatment and baseline conditions, so the ACE is simple to calculate. However, without randomized controlled experiments, there is no such guarantee. Instead, you have to do something like match $P(C|x)$ to $P(C|x')$. Unfortunately, C is usually high-dimensional, which makes such a matching problem difficult. One alternative is to use *propensity scores.*

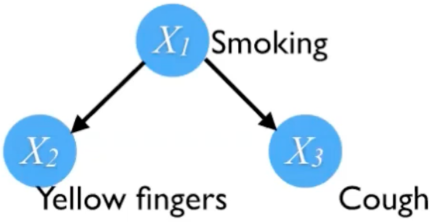
Let the propensity score $h(C) = P(X = 1|C)$, then $X \perp\!\!\!\perp C|h(C)$, and we call $h(C)$ and C confounding-equivalent because:

$$\begin{aligned}\text{ACE} &= \sum_c P(Y|x,c)P(c) = \sum_c \sum_h P(Y|x,c)p(h)p(c|h) = \sum_c \sum_h P(Y|x,c,h)p(h)p(c|h,x) \\ &= \sum_c \sum_h P(Y,c|x,h)P(h) = \sum_h P(Y|x,h)P(h)\end{aligned}$$

In other words, the ACE can be calculated using $P(h)$ instead of $P(c)$. This is extremely useful, since regardless of how high-dimensional C is, $h(C)$ is always a scalar. Thus, the task of matching $P(h)$ across different groups is feasible even when C is high-dimensional, leading to more accurate estimations of the ACE.

4 Counterfactual Reasoning

Consider a simple three variable graph, where each variable is binary:



There are three types of questions we might want to answer about the graph:

- Prediction: Would George cough if we *find* he has yellow fingers? This is $P(X_3|X_2 = 1)$
- Intervention: Would George cough if we *make sure* that he has yellow fingers? This is $P(X_3|do(X_2 = 1))$
- Counterfactual: Would George cough *had* he had yellow fingers, *given that he does not have yellow fingers and coughs*? This is $P(X_{3_{X_2=1}}|X_2 = 0, X_3 = 1)$

In general, statements about counterfactual reasoning are more powerful than statements about intervention, which are more powerful than statements about prediction.

For example, consider the situation where you have two variables X and Y, where $Y = \log(X + E + 3)$; E is some exogenous error, and X causes Y. Furthermore, say we observe for George $x = -1$ and $y = 0.4$, and we are interested in what Y would have been had X been 0 instead. With prediction, the "best" answer we can give is $\mathbf{E}[Y|X = 0]$, or $\log(0 + 3) = 1.10$.

However, with counterfactual inference, we can make claims *specific* to George, rather than general claims about people with $X = 0$. We do this by following three steps of counterfactual inference to calculate $P(Y_{X=0}|X = -1, Y = 0.4, E = e)$:

- Abduction: Find $P(E|\text{evidence}) = P(E|X = x, Y = y)$
- Action: Replace the equation for X by $X = x'$
- Prediction: Use the modified model to predict Y

Here, the process is:

- Abduction: $e = \exp(y) - x - 3 = \exp(0.4) + 1 - 3 = -0.51$
- Action: Set $X = x' = 0$ (instead of $X = x = -1$)
- Prediction: $Y = \log(x' + e + 3) = \log(0 - 0.51 + 3) = 0.91$

As a result, we provide the counterfactual inference value of $Y = 0.91$ instead of the prediction value $Y = 1.10$. Counterfactual inference allows us to make a claim about a specific person, and gives an answer that is both more informative and notably different from the best answer given by prediction. In this case, we can make such a claim by leveraging the fact that once observed, the exogenous error E is specific to George, and would have remained the same in counterfactual scenarios.

5 Causal Discovery (Overview)

Life abounds with examples when causal reasoning could be quite helpful: finding the causal relationship between agriculture, culture, and climate; investigating how certain variables would influence the shape of human skeleton in the long run; distinguishing causes from effects by analysing distribution information; reasoning about the latent variable when we do not have access to such variables; and so on.

Traditional causal discovery methods fall into two categories: constraint-based and score-based methods. Before we look at various causal discovery algorithms, let's first consider the connection between *causal structure* and *statistical properties of data*, and suitable assumptions are needed.

5.1 Modularity

Within a causal system, the whole causal process can be divided into small modules based on the *modularity* property.

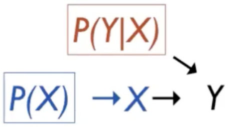


Figure 1: Modularity property

As illustrated in fig. 1, we can see that, if there is no confounder between X and Y, the data generating process of the cause (here is X) and the data generating process of effect from cause (here is generating Y from X) are independent.

5.2 Causal Sufficiency

We say a set of random variable $\mathbf{V}$ is causally sufficient, if $\mathbf{V}$ contains every direct cause (with respect to $\mathbf{V}$) of any pair of variables in $\mathbf{V}$.

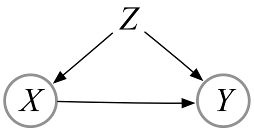


Figure 2: Causal sufficiency illustration

As illustrated in fig. 2, Z is the unobserved common cause (hidden confounder) of X and Y, according to our definition of causal sufficiency, $\mathbf{V}_1 = \{X, Y, Z\}$ is causally sufficient, while $\mathbf{V}_2 = \{X, Y\}$ is not.

5.3 Causal Markov Condition vs. Faithfulness

In the previous lectures, we have learnt the d-separation, namely, "read" the conditional independence relation from the graph. We know that if X and Y are d-separated by Z in graph $\mathcal{G}$, then the conditional independence $X \perp\!\!\!\perp Y \mid Z$ holds true.

Notice that the contrapositive of this proposition, namely, conditional dependence implies d-connection, does not tell us what would happen to the underlying graph if we actually found some conditional independence relations in the data. Therefore, beyond (global) Markov condition, we need something else to connect the conditional independence in data back to the property in the causal graph, that is, faithfulness.

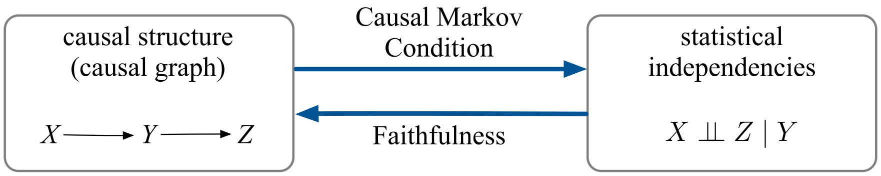


Figure 3: Caption

In the figure above, Causal Markov Condition is saying that, each variable is conditional independent with its non-descendants (non-effect variables) conditional on its parents (direct causes); Faithfulness is saying that, all observed (conditional) independencies are entailed by Markov condition in the graph. Notice that faithfulness is a rather strong assumption, and not every causal discovery algorithm makes this assumption, as will see later when we discuss causal discovery algorithms.

With Causal Markov Condition and faithfulness assumption, we can recover the skeleton (only edges) of the causal graph. After orienting the derived skeleton, we can get the causal graph using the following PC algorithm.

6 PC Algorithm

An example of the constraint-based causal discovery methods is the PC algorithm (Spirtes and Glymour [1991]). This contains two steps:

Step1: X and Y are adjacent iff they are dependent conditional on every subset of the remaining variables.
Step2: Orientation propogation.

More details can be found in Algorithm 1.

Algorithm 1 PC Algorithm

```
A.) Form the complete undirected graph C on the vertex set V.
B.) n=0.
    repeat
        repeat
            select an ordered pair of variables X and Y that are adjacent in C such that Adjacenies(C,X)\{Y}
            has cardinality greater than or equal to n, and a subset S of Adjacenies(C,X)\{Y} of cardinality
            n, and if X and Y are d-separated given S delete edge X-Y from C and record S in Sepset(X,
            Y) and Sepset (Y, X);
        until all ordered pairs of adjacent variables X and Y such that Adjacenies(C,X)\{Y} has cardinality
        greater than or equal to n and all subset S of Adjacenies(C,X)\{Y} of cardinality n have been
        tested or d-separated;
        n = n+1.
    until for each ordered pair of adjacent vertices X, Y, Adjacenies(C,X)\{Y} is of cardinality less than
    n.
C.) For each triple of vertices X, Y, Z such that the pair X, Y and the pair Y, Z are each adjacent in C but
    the pair X, Z are not adjacent in C, orient X-Y-Z as X→Y←Z if and only if Y is not in Sepset(X, Z).
D.) repeat
        If A→B, B and C are adjacent, A and C are not adjacent, and there is no arrowhead at B, then
        orient B-C as B → C.
        If there is a directed path from A to B, and an edge between A and B, then orient A-B as A → B.
    until no more edges can be oriented.
```

7 Equivalent Classes: Patterns

We can define the equivalanet classes: Two DAGs are (independence) equivalent if and only if they have the same skeletons and the same v-structures (Pearl and Verma [1991]). We can only recover such class based on the data using the conditional independent relations. Such (independence) equivalent classes can be represented by pattern, which is defined by: Patterns or CPDAG(Completed Partially Directed Acyclic Graph): graphical representation of (conditional) independence equivalence among models with no latent common causes (i.e., causally sufficient models).

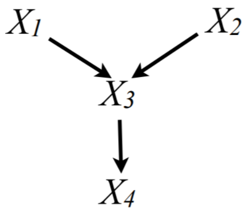
8 FCI

FCI (Fast Causal Inference) can validly infer causal relationships from conditional independence statements even when there are confounders. FCI is quite complex so the detail of the algorithm was not covered in the lecture, but some rules from FCI are covered in the lecture to showcase examples of when we can be certain that there must be no confounders between two variables, and we can be certain that there must be confounders between two variables.

8.1 Example 1. Y-structure implies no confounder

The three conditional independence statements: $X_1 \perp X_2$; $X_1 \perp X_4|X_3$; $X_2 \perp X_4|X_3$ imply that X_1, X_2, X_3, X_4 follow the so-called Y-structure shown in the figure below.

Then it is impossible for X_3 and X_4 to have a confounder. Suppose for contradiction that there is a confounder C that directly causes X_3 and X_4. Then the path $X_1 - X_3 - C - X_4$ will cause X_1 and X_4 to be d-connected given X_3, and the condition $X_1 \perp X_4 | X_3$ no longer holds, hence a contradiction.



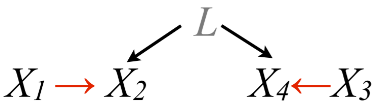
8.2 Example 2. If $A \perp C|B$ and B does not cause A, then there is no confounder between B and C.

Since B does not cause A, then it must hold that either A causes B or A and B have a confounder. In both cases, B causes C must holds and there must be no confounder between B and C. Otherwise, B would be a collider and A and C cannot be conditionally independent given B.

A real life example would be $A = raining$, $B = slippery$, $C = falling\ down$. Since $raining$ and $falling\ down$ are conditionally independent given $slippery$, and $slippery$ does not cause $raining$, we can be certain that $slippery$ causes $falling\ down$ and there is no confounder between $slippery$ and $falling\ down$. Another example would be $A = geographical\ background$, $B = economic\ conditions$, $C = emergence\ of\ science$.

8.3 Example 3. If $X_1 \perp X_3; X_1 \perp X_4; X_2 \perp X_3$, then X_2 and X_4 must have a confounder.

This is illustrated in the figure below. Suppose for contradiction that there are no counfounder between X_2 and X_4. Then applying PC rule, X_1, X_2 and X_4 should follow v-structure which implies X_4 causes X_2. Similarly, X_2, X_3 and X_4 should follow v-structure which implies X_2 causes X_4, then we have a contradction, so there must be a confounder (denoted as L in the figure below) between X_2 and X_4.

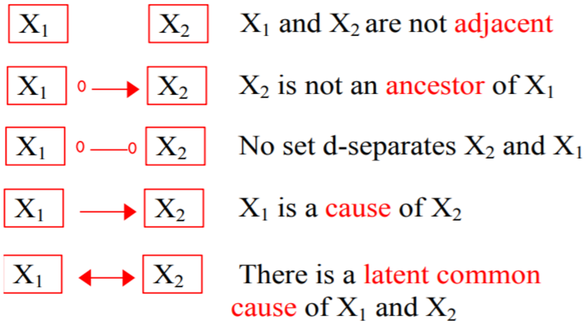


8.4 PAG

The above three examples showcase some rules that are integrated in the FCI algorithm. As can be seen from the example 3, result provided by FCI would make graph over observed variables not necessarily a DAG without the help of latent variables, so FCI returns output called PAG instead of DAG. There are 5 kinds of relationships between two variables in PAG shown in the figure below. A circle denotes that it could be an arrow head or arrow tail.

9 GES

PC algorithm and FCI are constraint-based causal discovery algorithms. Another branch of causal discovery algorithms is score-based. GES (Greedy Equivalence Search) is one kind of score-based causal discovery.



X_1 and X_2 are not adjacent

X_2 is not an ancestor of X_1

No set d-separates X_2 and X_1

X_1 is a cause of X_2

There is a latent common cause of X_1 and X_2

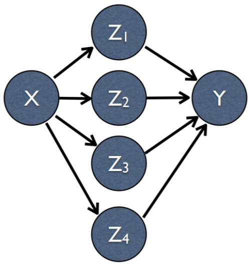
The score function satisfies 3 properties: score equivalent (assigning the same score to equivalent DAGs), locally consistent (score of a DAG increases when adding any edge that eliminates a false independence constraint and vice versa), and decomposable (score of a DAG can be written as summation over a function of node and its parents in the DAG). An example of a score function that satifies above three properties are BIC: $S_B(G, D) = \text{log}p(D|\hat{\theta}, G^h) - \frac{d}{2}\text{log}m$.

The GES algorithm consists of Forward Greedy Search and Backward Greedy Search through the sapce of DAG equivalence classes, and the detailed algorithms are shown below:

Algorithm 2 GES

```
1. Forward Greedy Search (FGS)
       Start from some (sparse) pattern (usually the empty graph)
       Evaluate all possible patterns with one more adjacency that entail strictly fewer CI statements than
       the current pattern
       Move to the one that increases the score most
       Iterate until a local maximum
2. Backward Greedy Search (BGS)  Start from the output of the Forward stage
       Evaluate all possible patterns with one fewer adjacency that entail strictly more CI statements than
       the current pattern
       Move to the one that increases the score most
       Iterate until a local maximum
```

For example, if the data was generated by the DAG shown below, then it is likely that FGS will add an edge between X and Y since they are highly correlated. However, BGS will likely remove this edge between X and Y for a higher score.



10 Linear,Non-Gaussian Models

For classical models, the problem is defined as identifying the casual structures such as $X \longleftarrow Y \longrightarrow Z$, $X \longrightarrow Y \longrightarrow Z$, and $X \longleftarrow Y \longleftarrow Z$, which give rise to the conditional independence between the data as $X \perp\!\!\!\perp Z \mid Y$. However, recovering the causal relations from conditional independences is bounded by the equivalence class. Since the mapping between causal structures and conditional independence is not one-to-one, we cannot reconstruct the exact causal relationship. Furthermore, the classical methods cannot directly characterize and recover the cause-effect relationships for two-variable cases, such as $X \longrightarrow Y$, $X \longleftarrow Y$ and $X \longleftarrow Z \longrightarrow Y$ (where Z is an unobserved cause). Additional weak and reasonable assumptions are needed.

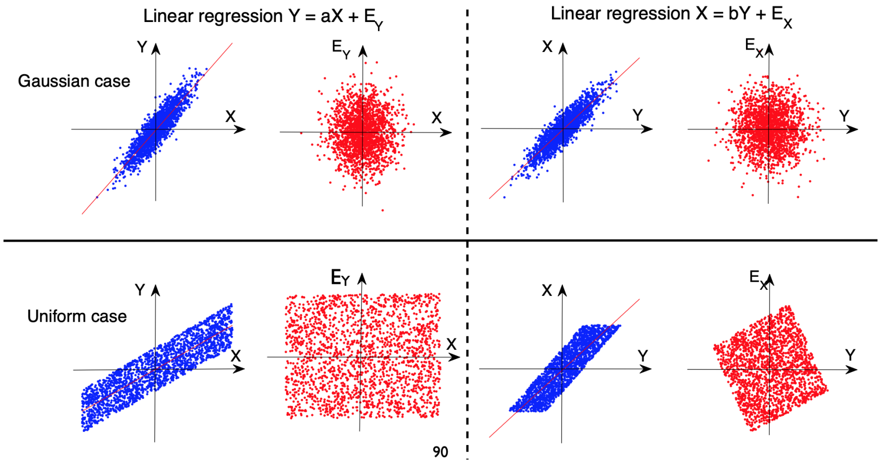
10.1 Functional Causal Model

A functional causal model represents the effect as a function of direct causes and noise: $Y = f(X, E)$, with $X \perp\!\!\!\perp E$.The following are some typical casual models for discovering the causal relations(Shimizu et al. [2006], Hoyer et al. [2008], Zhang and Hyvärinen [2009a], Zhang and Hyvärinen [2009b]):

- Linear non-Gaussian acyclic causal model: $Y = a \cdot X + E$.
- Additive noise model: $Y = f(X) + E$.
- Post-nonlinear causal model: $Y = f_2(f_1(X) + E$.

It is noted that even if we do not give any constraints on the function f, given two random variables X and Y, we can always represent a variable as a function of another variable and an independence noise. However, it is not the case if we have the constraints on the function.

In the linear case, we can easily go from correlation to asymmetry causal relation if we assumed the noise is non-Gaussian. An example is for data generated by $Y = aX + E$ i.e, $X \longrightarrow Y$ The upper case explains regressing Y on X and X on Y when the X and noise E are gaussian. In this case, the error term is uncorrelated from the predictor, which also implies independence. For the bottom cause, X and the noise is generated from uniform distribution. We can observe that the noise term is not independent from Y.



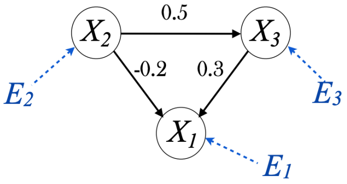
10.2 LiNGAM Model

In the more general case, we have the linear non-Gaussian acyclic model. The Linear non-Gaussian acyclic model is defined as the following:

$$X_i = \sum_{\text{j: parents of i}} b_{ij} X_j + E_i$$
$$\mathbf{X} = \mathbf{BX} + \mathbf{E}$$

with the disturbances(errors) E_i are non-Gaussian(or at most one is Gaussian) and mutually independent. An example is to represent the causal structure of the graph as:

$$X_2 = E_2$$
$$X_3 = 0.5X_2 + E_3$$
$$X_1 = -0.2X_2 + 0.3X_3 + E_1$$



The matrix form for the above LiNGAM model is given as the following. Note that the diagonal entries of the matrix are zero because we assume no self-loops, for simplicity.

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} 0 & -0.2 & 0.3 \\ 0 & 0 & 0 \\ 0 & 0.5 & 0 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} + \begin{bmatrix} E_1 \\ E_2 \\ E_3 \end{bmatrix}$$

Thus, the task is reduced to recover the matrix B given the data. With the above LiNGAM model setting, we can reconstruct a unique solution $\mathbf{B}$ by the model and independent component analysis(ICA).

ICA aims to find a rotation transformation $\mathbf{Y} = \mathbf{WX}$, and to make each Y_i independent, maximum likelihood and mutual information minimization are used.

Thus, $\mathbf{B}$ is achieved from $\mathbf{W}$ by permutation and rescaling (make all diagonal entries non-zero, divide the row by its diagonal entry, and subtract the resulting matrix from the identity matrix).

10.3 Gaussianity or Non-Gaussianity?

The gaussian is widely used in the research area because of its “simplicity” of the form: marginal and conditionals are also gaussian, the mean and covariance matrix are easily computed, central limit theorem, and etc.

However, in practice, the non-gaussianity is actually ubiquitous. By Cramer’s theorem, we observe the linear closure property of Gaussian distribution: if the sum of any finite independent variables is Gaussian, then all summands must be gaussian. But gaussian is only special in the linear case.

Other issues in causal discovery includes nonlinearities, categorical variables or mixed cases, measurement error, selection bias, causality in time series, non-stationary and heterogeneous data, and etc. One important issue is to face the confounding, or find the hidden causes or parents in the graph.

References

P. Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. pages 689–696, 01 2008.

J. Pearl and T. Verma. A theory of inferred causation, 1991.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.*, 7:2003–2030, Dec. 2006. ISSN 1532-4435.

P. Spirtes and C. Glymour. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9 (1):62–72, 1991.

J. Tian and J. Pearl. A general identification condition for causal effects. In *In Eighteenth National Conference on Artificial Intelligence*, pages 567–573, 2002.

K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, page 647–655, Arlington, Virginia, USA, 2009a. AUAI Press. ISBN 9780974903958.

K. Zhang and A. Hyvärinen. Causality discovery with additive disturbances: An information-theoretical perspective. In *Proceedings of the 2009th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part II*, page 570–585, Berlin, Heidelberg, 2009b. Springer-Verlag. ISBN 3642041736.