

Lecture 5: Causal inference with endogeneous treatments

Lecturer: Yunyi Zhang

Suggested Reading:

1 Model & Distribution

There are two languages that describe the behaviors of random variables: the model (such as linear model) and the distribution (like density, pmf, cdf, etc). Statisticians always use the former, but we also need the latter in today's topic. So before today's discussion, **we make a connection between statistical model and the distribution**. We will not explicitly separate those descriptions in the following study.

Example 1 Suppose a linear model

$$Y = \sum_{j=1}^p X_j \beta_j + \epsilon,$$

where ϵ has cdf $F_\epsilon(\cdot)$, and ϵ is independent of X_j . When we establish this model, we assume the conditional distribution of Y conditional on X_1, \dots, X_p . Thus,

$$\begin{aligned} F(Y \leq y | X_1 = x_1, \dots, X_p = x_p) &= \text{Prob}(Y \leq y | X_1 = x_1, \dots, X_p = x_p) \\ &= \text{Prob}(\epsilon \leq y - \sum_{j=1}^p x_j \beta_j) = F_\epsilon(y - \sum_{j=1}^p x_j \beta_j). \end{aligned}$$

Correspondingly, the conditional density is

$$f_\epsilon(y - \sum_{j=1}^p x_j \beta_j).$$

Moreover, if we want to calculate the probability $P(Y \leq y)$, then it becomes

$$\mathbf{E} F_\epsilon(y - \sum_{j=1}^p X_j \beta_j).$$

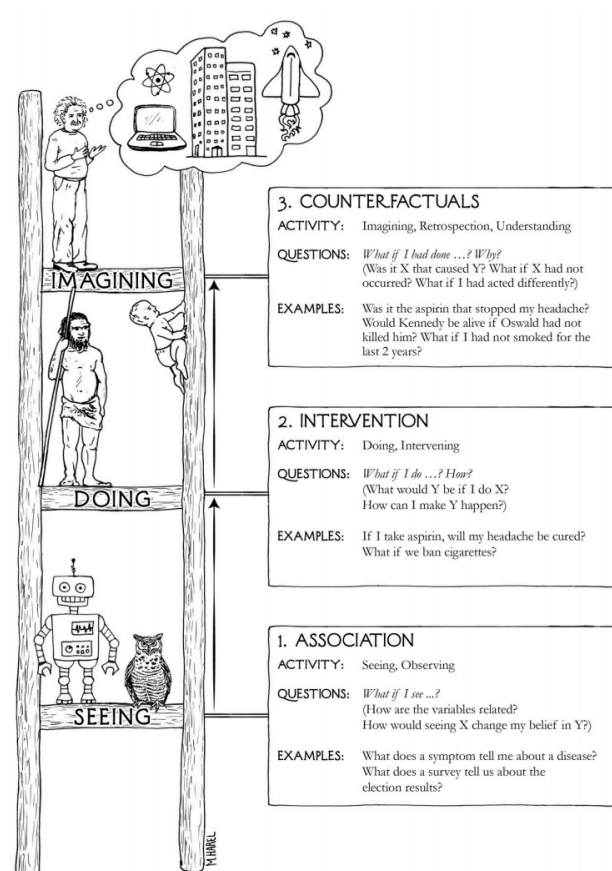
Example 2 More specifically, let us suppose $Y = \sum_{j=1}^p X_j \beta_j + \epsilon$ with $\epsilon \sim N(0, \sigma^2)$. If that happens, the conditional distribution of $Y | X_j, j = 1, \dots, p$ is $N(\sum_{j=1}^p X_j \beta_j, \sigma^2)$, with density

$$f(y | X_1, \dots, X_p) = \frac{1}{\sqrt{2\pi}\sigma^2} \exp \left(-\frac{1}{2\sigma^2} (y - \sum_{j=1}^p X_j \beta_j)^2 \right).$$

2 Judea Pearl's point of view of causal inference

Judea Pearl believes that there are **three levels of human cognitive ability**, i.e., seeing, doing, and imagining, as summarized in the ladder of causation. The machine learning scientists then aim to emulate those abilities using a computer.

- **Association (observing)**: What happens to Y if I see X , like regression model.
- **Intervention**: How can I make Y happens, not only imitations.
- **Counterfactuals**: image the world that does not exist, like today's generative model.



3 Structural causal model and DAG

When discussing methods for treatment effect estimation under unconfoundedness, we have effectively assumed that potentially after conditioning on observed covariates the treatment assignment is determined by as-good-as-random factors that are irrelevant to the causal inference question at hand. In other words, we have effectively assumed treatment assignment is exogenous to the system we are studying.

In some applications, however, such exogeneity assumptions are simply not plausible. For example, when studying the effect of prices on demand, it is unrealistic to assume that potential outcomes of

$(Y(1), Y(0))$

W 与 $(Y(0), Y(1))$ 相互决定 \Rightarrow 内生性

W

demand (i.e., what demand would have been at given prices) are independent of what prices actually were. Instead, its much more plausible to assume that prices and demand both respond to each other until a supply-demand equilibrium is reached. 考虑内生性 (无法通过控制 X 使得 $W \perp (Y(0), Y(1)) | X$)

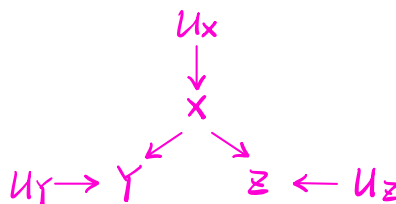
This section present basic methods and concepts for causal inference in settings where unconfoundness does not hold and treatment assignment is instead endogenous, i.e., treatments are assigned in a way that depends on the interplay of other variables within the system.

1. DAG

Definition 3.1 (Directed acyclic graph (DAG)). A directed graph with nodes indexed $j = 1, \dots, p$ is characterized by a set of edges E_{ij} where $E_{ij} = 1$ denotes the presence of an edge from node i to node j and $E_{ij} = 0$ denotes lack of such an edge. Within a directed graph, a directed path is an ordered set of at least two nodes $i_1, i_2, \dots, i_k \in \{1, \dots, p\}$ such that $E_{i_1 i_2} = E_{i_2 i_3} = \dots = E_{i_{k-1} i_k} = 1$. A directed graph is acyclic (i.e., a DAG) if it contains no directed cycles, i.e., directed paths with $i_1 = i_k$. Within a DAG, we say that a node i is upstream of j (and that j is downstream of i) if there exists a directed path starting at i and ending at j . We define the set of parents of node j as the set of nodes i with $E_{ij} = 1$.

Example 3 Suppose the system contains 3 random variables X, Y, Z and 3 mutually independent exogenous random variables U_X, U_Y, U_Z satisfying the following system

$$\begin{aligned} X &= U_X \\ Y &= \frac{1}{3}X + U_Y \\ Z &= \frac{1}{4}X + U_Z. \end{aligned}$$



Structural causal model (SCM) constitutes an important tool to link causal description and probability statements.

2. SCM

Definition 3.2. An SCM consists of two assignments

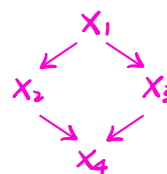
$$C = N_C, \quad E = f_E(C, N_E), \quad \text{where } N_E \text{ is independent of } N_C.$$

effect *cause error*

We call C the cause and E the effect. Furthermore, C is the parents (direct causes) of E , and referring to $C \rightarrow E$ in the causal graph.

Example 4 Suppose the random vector $(X_1, X_2, X_3, X_4) \in \mathbf{R}^p$ satisfies

$$\begin{aligned} X_1 &= U_1, \\ X_2 &= 0.7X_1 + U_2 \\ X_3 &= -0.5X_1 + U_3 \\ X_4 &= 0.3X_2 + 0.4X_3 + U_4. \end{aligned}$$



Then X_1 is the direct cause for X_2 and X_3 . X_2, X_3 are the cause for X_4 .

Example 5 Basketball performance based on height & gender Suppose the following relations

$$Gender = U_1,$$

$$Height = f_h(Gender, U_2),$$

$$Performance = f_p(Gender, Height, U_3).$$

After introducing the following structural model, we can fit the f_h and f_p , then predict the performance of the players.

Example 6 Meinshausen et al. [2016] use structural equation models to study the relationship between the expression of different genes in the yeast *saccharomyces cerevisiae*. The authors have access to expression levels for 6,170 genes and are interested in questions of the type: How will the expression of gene i in the yeast be affected by inactivating gene j ? To formalize this question, they posit that gene expressions can be modeled using a DAG, and posit a linear SEM

$$Z_i = \sum_{j \in pa_i} \beta_{ij} Z_j + \epsilon_i,$$

where Z_i measures the expression level of the i th gene; the statistical task then reduces to estimating β_{ij} in this model. They estimate these quantities using the method of Peters, Blümlmann, and Meinshausen [2016] which assumes cross-environment invariance of the SEM coefficients to identify causal effects.

4 Use causal graph to simplify model

Suppose the causal graph is acyclic, then the joint distribution(density or pmf) is given by

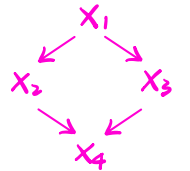


$$f(x_1, \dots, x_n) = \prod_{i=1}^n f(x_i | pa_i),$$

where pa_i represents the parents of x_i .

Example 7 Consider example 4. Suppose the density of U_i is f_U , then

$$\begin{aligned} f(y_1, y_2, y_3, y_4) &= f(X_4 = y_4 | X_2 = y_2, X_3 = y_3) f(X_3 = y_3 | X_1 = y_1) \\ &\quad \times f(X_2 = y_2 | X_1 = y_1) f(X_1 = y_1) \\ &= f_U(y_4 - 0.3y_2 - 0.4y_3) f_U(y_3 + 0.5y_1) \times f_U(y_2 - 0.7y_1) \times f_U(y_1). \end{aligned}$$



Example 8 Suppose a graph $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$ of binary random variables satisfy the following equations:

$$P(X_i = 1|X_{i-1} = 1) = p, \quad P(X_i = 1|X_{i-1} = 0) = q, \quad P(X_1 = 1) = p_0.$$

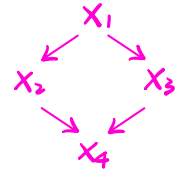
Then the joint probability

$$\begin{aligned} P(X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 0) &= P(X_4 = 0|X_3 = 1)P(X_3 = 1|X_2 = 0)P(X_2 = 0|X_1 = 1)P(X_1 = 1) \\ &= (1 - p) \times q \times (1 - p) \times p_0. \end{aligned}$$

With the help of the causal graph, statisticians can simplify the models, from modelling $f(x_1, \dots, x_n)$ to modelling $f(x_i|pa_i)$, which significantly decreases the complexity.

Example 9 Adopting causal graph in fitting linear models Suppose the causal model

$$\begin{aligned} X_1 &= U_1, \\ X_2 &= \alpha_{21}X_1 + U_2 \\ X_3 &= \alpha_{31}X_1 + U_3 \\ X_4 &= \alpha_{42}X_2 + \alpha_{43}X_3 + U_4. \end{aligned}$$



After collecting data $(X_{i1}, X_{i2}, X_{i3}, X_{i4}), i = 1, \dots, n$, we may estimate the coefficients

$$\hat{\alpha}_{21} = \sum_{i=1}^n X_{i2}X_{i1} / \sum_{i=1}^n X_{i1}^2, \quad \hat{\alpha}_{31} = \sum_{i=1}^n X_{i3}X_{i1} / \sum_{i=1}^n X_{i1}^2, \quad (\hat{\alpha}_{42}, \hat{\alpha}_{43})^\top = (Z^\top Z)^{-1} Z^\top W,$$

where

$$Z = \begin{bmatrix} X_{12} & X_{13} \\ X_{22} & X_{23} \\ \vdots & \vdots \\ X_{n2} & X_{n3} \end{bmatrix}, \quad W = (X_{14}, \dots, X_{n4})^\top.$$

5 Special structures in causal graph

1. chains

There is some typical structures in causal graph. The first structure is called “chains”: $X \rightarrow Y \rightarrow Z$.

Example 10 High school funding, SAT score, Acceptance

$$X \perp Z \mid Y = y$$

$$X = U_X, \quad Y = \alpha X + U_Y, \quad Z = \beta Y + U_Z.$$

- ① $Z \rightarrow W \leftarrow X \rightarrow Y$: d-separated by $\{\emptyset\}, \{X\}$
 ② $Z \rightarrow W \rightarrow U$: d-separated by $\{W\}$
 ③ $U \leftarrow W \leftarrow X \rightarrow Y$: d-separated by $\{W\}, \{X\}$

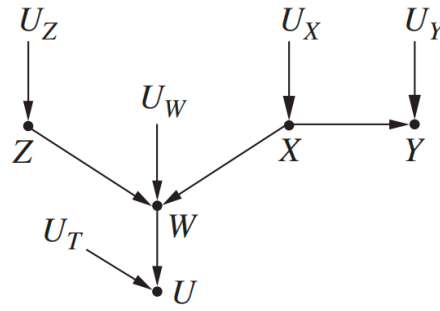


Figure 1: Figure 1

The special character of a chain is that X, Z are independent conditioning on $Y = y$.

2. forks The second structure is called “forks”, which is of the form. $Y \leftarrow X \rightarrow Z$: $Y \perp Z | X = x$

Example 11 Temperature, Ice cream sales, forest fires

$$X = U_X, Y = \alpha X + U_Y, Z = \beta X + U_Z.$$

Conditioning on $X = x$, Y, Z are independent.

3. collider The third structure is called collision, whose structure is as follows: $X \rightarrow Z \leftarrow Y$. The special stuff of collision is that X, Y are independent, but they are dependent after conditioning on Z . $Y \perp X$ & $Y \not\perp X | Z$

Example 12 Suppose we independently toss two dices. Recall the result $Z = X + Y$. In this case, X and Y are independent. However, once we know that $Z = 10$, then X and Y can be determined from each other.

With the special structures chains, forks, and collider, we introduce a notion called “d(directional)–separation.”

4. d-separation Definition 5.1 (d-separation). A path in the causal graph p (a linked sequence a_1, a_2, \dots, a_m of nodes you want to research) is blocked by a set of nodes $Z = \{z_1, \dots, z_m\}$ if and only if

- p contains a chain of nodes $A \rightarrow B \rightarrow C$ or a fork $A \leftarrow B \rightarrow C$ such that B is in Z , or
- p contains a collider $A \rightarrow B \leftarrow C$ such that B , and all descendants of B , are not in Z .

Remark Despite being abstract, the idea of d-separation is clear, i.e., after conditional on Z , the endpoints a_1 and a_m should be independent.

Example 13 Consider Figure 1. Suppose we care about the path Z, W, X, Y , then this path is d-separated by $\{\emptyset\}, \{X\}$, but is not d-separated by $\{U\}$. The path ZWU is d-separated by $\{W\}$. The path $UWXY$ is d-separated by $\{X\}$, or $\{W\}$.

Example 14 Consider Figure 1. Suppose we care about the path $TZWU$, then $\{Z\}$ or $\{W\}$ d-separates the path. Suppose we care about the path $UWXYT$, then $\{\emptyset\}, \{X\}, \{W\}$ d-separate the path, but $\{Y\}$ does not d-separate the path.

一句话总结

do-operator = 上帝之手，表示“人为强制改变某个变量”，而不是被动观察它。

目的：打破被动观察的局限，估计「如果我对 X 做某个操作（比如设 $X=1$ ），Y 会如何变化？」——这就是因果效应！

为什么需要 do-operator?

传统统计的局限：

- 我们熟悉的条件概率 $P(Y|X=1)$ 是“观察到的关联”（比如看到 $X=1$ 时 Y 的分布），但可能混杂了其他因素。
- 例子：
 - 假设 X 是“吃药”，Y 是“康复”。
 - 如果病人自己选择吃药 ($X=1$)，可能病情更重（存在混杂因子 Z =病情），导致 $P(Y=1|X=1)$ 很低（看似吃药没用）。
 - 但真实因果效应可能是正的（吃药有效），因为混杂因子 Z 扭曲了观察结果。

do-operator 的作用：

- $P(Y | \text{do}(X=1))$ 表示“强制让所有人吃药后，康复的概率”。
- 它消除了混杂因子的影响，直接反映 X 对 Y 的因果效应。

do-operator 的直观理解

假设因果图如下：

$Z \rightarrow X \rightarrow Y$
 $U \nearrow$
U (未观测的混杂因子)

- 观察 X 和 Y 的关联 ($P(Y|X)$)：混杂因子 Z 和 U 会影响结果，导致相关性 \neq 因果性。
- 干预 $\text{do}(X=1)$ ：直接切断 X 的所有 incoming 箭头（比如 $Z \rightarrow X$ 和 $U \rightarrow X$ ），让 X 独立于其他因素。此时 X 的变化只影响 Y，而不是被其他：明。

和普通条件概率的区别

- $P(Y | X=1)$ ：被动观察 $X=1$ 的人群中 Y 的分布（可能混杂）。
- $P(Y | \text{do}(X=1))$ ：主动干预让 $X=1$ 后 Y 的分布（因果效应）。

例子：

- 冰激凌销量 (X) 和溺水数 (Y) 在夏天 (Z) 时高度相关。
 - $P(Y | X=\text{高})$ ：夏天时两者都高，但这是伪相关。
 - $P(Y | \text{do}(X=\text{高}))$ ：如果强制让冰激凌销量高（比如冬天促销），溺水数不会增加，真实因果效应为 0。

① $T \rightarrow Z \rightarrow W \rightarrow U$: d-separated by $\{Z\}, \{W\}$

② $T \rightarrow Y \leftarrow X \rightarrow W \rightarrow U$: d-separated by $\{\emptyset\}, \{X\}, \{W\}$

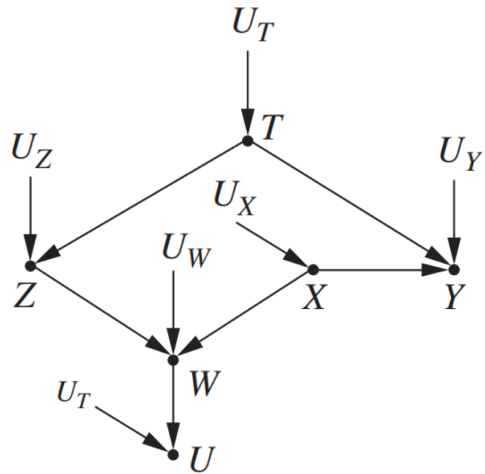


Figure 2: Figure 14

6 Causal query and “do” operator

Example 15 Motivation

In the beginning of the class, we say that the sales of ice – cream and the wild fire has strong dependency, due to the weather. In the causal graph sense, the graph is written by the following

$$ice\ cream \leftarrow weather \rightarrow fire.$$

Suppose we want to test the causal relation between ice cream and fire (such as $fire = f(ice\ cream, weather, U_f)$), a simple strategy is fixing the sales of ice cream to different levels, and test whether the fire has significant difference. This motivates the use of the notion “do” operator / causal query.

Given a SEM, a causal query involves exogenously setting the values of some nodes of the graph G , and seeing how this affects the distribution of other nodes. Given two disjoint sets of nodes $W, Y \subset Z$, the causal effect of setting W to w is written $P(Y|do(W = w))$, and corresponds to deleting all equations used to generate W and plugging in w for W in the rest. Formal definition is as follows:

1. do operator

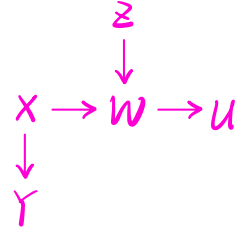
Definition 6.1 (do operator). Suppose a system with n features X_1, \dots, X_n with corresponding structural causal model

$$X_j = f_j(pa_j, U_j), j = 1, \dots, n \text{ and corresponding causal graph } G,$$

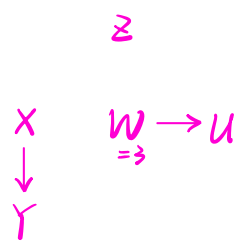
where pa_j refers to the parents of X_j . The do operator $do(X_k = q)$ generates a new structural causal model as follows:

- Replace X_k by q in all of the structural causal model equations.
- In the causal graph G , delete all edges that point X_k .

Example 16 Suppose a SEM

$$\begin{aligned}
 X &= U_X, \\
 Y &= \alpha_{YX}X + U_Y, \\
 Z &= U_Z, \\
 W &= \alpha_{WX}X + \alpha_{WZ}Z + U_W, \\
 U &= \alpha_{UW}W + U_U.
 \end{aligned}$$


After “ $do(W = 3)$ ”, the system becomes

$$\begin{aligned}
 X &= U_X, \\
 Y &= \alpha_{YX}X + U_Y, \\
 Z &= U_Z, \\
 W &= 3, \\
 U &= 3\alpha_{UW} + U_U.
 \end{aligned}$$


Clearly, if do the points that separate a path, then we can construct independent random variables.

Example 17 Testing direct causal effect Suppose binary random variables X, Y, Z satisfy the following causal graph. Suppose we want to test that there is no causal effects between Z and Y . To achieve

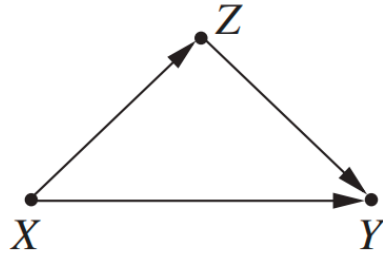


Figure 3: Example 3

this goal, we estimate the difference

$$\tau = \text{Prob}(Y = 1 | do(Z = 1)) - \text{Prob}(Y = 1 | do(Z = 0))$$

and test the hypothesis

$$H_0 : \tau = 0 \text{ versus } H_1 : \tau \neq 0.$$

That is, we first $do(Z = 1)$, collect the corresponding data $Y_{i1}, i = 1, \dots, n_1$. After that, we $do(Z = 0)$, collect the corresponding data $Y_{i2}, i = 1, \dots, n_2$. Then derive the estimator and the test statistics

$$\hat{\mu}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} Y_{i1}, \quad \hat{\mu}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_{i2}, \quad \hat{\tau} = \hat{\mu}_1 - \hat{\mu}_2, \quad \hat{T} = \frac{\hat{\tau}}{\sqrt{\frac{\hat{\mu}_1(1-\hat{\mu}_1)}{n_1} + \frac{\hat{\mu}_2(1-\hat{\mu}_2)}{n_2}}}.$$

If we cannot reject H_0 , then we may believe that there is no causal relation between Z, Y , and so we can delete the edge $Z \rightarrow Y$.

7 The difference between “do” operator and conditional

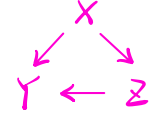
Let us consider the following example

主动控制

被动观察

Example 18 Consider the following SCM:

$$\begin{aligned} X &= U_X, \\ Y &= U_Y \times X + (1 - U_Y) \times Z, \\ Z &= X \times U_Z, \end{aligned}$$

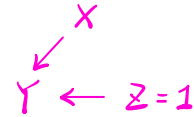


where U_X, U_Y, U_Z, U_W satisfy $Prob(U_i = 1) = 1/2$ and $Prob(U_i = 0) = 1/2$. Let us first calculate the probability

$$Prob(Y = 1|Z = 1) = \frac{Prob(Y = 1, Z = 1)}{Prob(Z = 1)} = \frac{Prob(U_X = 1, U_Z = 1)}{Prob(U_X = 1, U_Z = 1)} = 1.$$

On the other hand, let us consider $do(Z = 1)$. After this operation, the system becomes

$$\begin{aligned} X &= U_X, \\ Y &= (1 - U_Y) + XU_Y, \\ Z &= 1. \end{aligned}$$



Therefore,

$$Prob(Y = 1|do(Z = 1)) = Prob(U_Y = 0) + Prob(U_Y = 1, U_X = 1) = \frac{3}{4}.$$

Therefore, generally speaking, the do operation does not coincide with the conditional. If we consider the graph after operation, whose probability is denoted by $Prob_m(\cdot)$, then we have

$$\begin{aligned} Prob_m(Y = 1|Z = 1) &= Prob_m(Y = 1, X = 0|Z = 1) + Prob_m(Y = 1, X = 1|Z = 1) \\ &= Prob_m(Y = 1|Z = 1, X = 0)Prob_m(X = 0|Z = 1) + Prob_m(Y = 1|Z = 1, X = 1)Prob_m(X = 1|Z = 1) \\ &= Prob_m(U_Y = 0) \times \frac{1}{2} + \frac{1}{2} = \frac{3}{4}, \end{aligned}$$

which equals $Prob(Y = 1|do(Z = 1))$. This kind of formula is called the “adjustment formula” that connects the “do ” operation with the conditional probability.

Example 19 Consider the following causal graph and the corresponding SCM

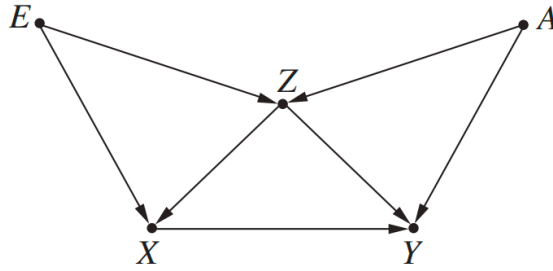


Figure 4: Caption

$$\begin{aligned} E &= U_E, \\ A &= U_A, \\ Z &= \alpha_{ZE}E + \alpha_{ZA}A + U_Z, \\ X &= \alpha_{XZ}Z + \alpha_{XE}E + U_X, \\ Y &= \alpha_{YZ}Z + \alpha_{YX}X + \alpha_{YA}A + U_Y, \end{aligned} \tag{1}$$

where U_E, U_A, U_Z, U_X, U_Y are i.i.d. random variables with mean 0 and variance σ^2 . Let us consider

$$\begin{aligned} \mathbf{E}(Y|Z = z) &= \alpha_{YZ}z + \alpha_{YX}\mathbf{E}X|Z + \alpha_{YA}\mathbf{E}A|Z \\ &= \alpha_{YZ}z + \alpha_{YX}(\alpha_{XZ}z + \alpha_{XE}\mathbf{E}E|Z) + \alpha_{YA}\mathbf{E}A|Z \end{aligned}$$

but due to the dependency between E and Z , calculating this is not very simple. On the other hand, after $do(Z = z)$, the system becomes

$$\begin{aligned} E &= U_E, \\ A &= U_A, \\ Z &= z, \\ X &= \alpha_{XZ}z + \alpha_{XE}E + U_X, \\ Y &= \alpha_{YZ}z + \alpha_{YX}X + \alpha_{YA}A + U_Y. \end{aligned}$$

So

$$\mathbf{E}Y|do(Z = z) = (\alpha_{YZ} + \alpha_{YX}\alpha_{XZ})z,$$

making the calculation accessible.

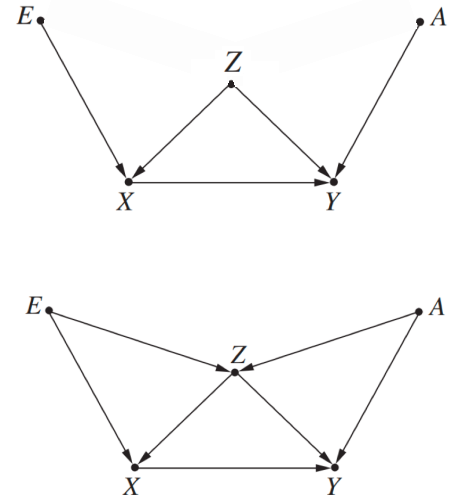


Figure 4: Caption

8 Frontdoor and Backdoor criterion

To illustrate the idea of backdoor and frontdoor criterion, we consider the following illustrative example:

Example 20 Suppose random variables (X, Y, Z, E, A) satisfy the causal relation demonstrated in Figure 4. We want to test the causality between X and Y with knowing the presence of confounders. Despite knowing the presence of confounders, sometimes it is impossible to acquire all desired data. For example, suppose X is the use of a medicine, Y is the performance of the medicine, and E, Z, A are confounders. For example, Z can be the blood pressure of the patient E can be the diet, and A can be the income of the patient. In order to test the causal effect $X \rightarrow Y$, we $do(X = 1)$, collect some data, then $do(X = 0)$, collect other data, and calculate the ATE. The issue here is, we may not be able to collect E and A . So we wonder only through collecting Z can we describe the causal relation between X and Y . The backdoor criterion provides us with a solution for this problem.

1. backdoor
criteria

Definition 8.1. Given an ordered pair of variables (X, Y) ($\neq (Y, X)$), in a directed acyclic graph G , a set of variables Z satisfies the “back – door” criterion relative to (X, Y) if no node in Z is a descendant of X , and Z blocks every path between X and Y that contains an arrow into X . (避免 collider)

逆因果路径

Generally, the “backdoor” path makes X and Y dependent, but not transmitting causal influence from X (i.e., the dependence comes from other nodes). Moreover, we cannot condition on descendent of X that introduces new dependence.

Example 21 Let us consider the following graph with pair (X, Y) . This graph has two path $Z \rightarrow X$ and $Z \rightarrow W \rightarrow Y \leftarrow X$. Consider the set $\{W\}$. Then the path $Z \rightarrow W \rightarrow Y \leftarrow X$ satisfies backdoor

寻找满足 backdoor criteria 的 sets 的方法:

① 找出 X, Y 间所有指向 X 的 path $\{P\}$.

② 找到一组 nodes $\{Z\}$, 确保

(1) $\{Z\}$ 中不含有 X 的 descendants

(2) condition on $\{Z\}$ 时, X 和 Y 不通过 $\{P\}$ 构成 dependence

criterion since W blocks the path.

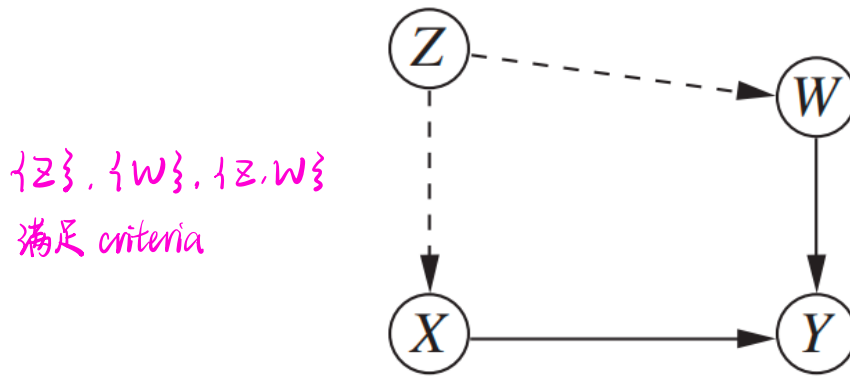


Figure 5: Caption

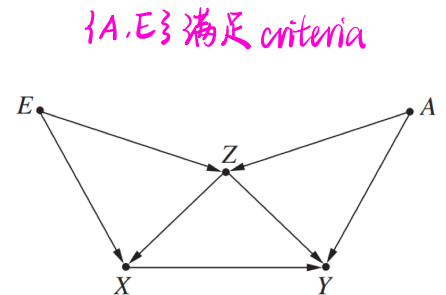


Figure 4: Caption

Example 22 Let us continue the discussion of figure 4. There are 4 paths from X to Y that contain an arrow pointing to X : $XEZY$, $XEZAY$, XZY , $XZAY$. However, in the path $XEZAY$, Z is in a collision EZA , so Z does not block all paths and so Z does not satisfy backdoor criterion. Indeed, when we fix X , i.e., $do(X = x)$, we have

$$Prob(Y = y | do(X = x)) = \sum_{(Z,A)=(z,a)} Prob_m(Y = y | Z = z, A = a, X = x) Prob_m(Z = z, A = a),$$

and collecting Z is not sufficient to describe the behavior of Y .

What happens if we can collect another set of data A . After collecting A , all paths are blocked. So that we satisfy the backdoor criterion.

Generally speaking, the reason for considering backdoor criterion is that, once you find this set, you have



$$P(Y = y | do(X = x)) = \sum_w Prob(Y = y | X = x, W = w) Prob(W = w).$$

In other words, the data you collect can describe the behavior of Y after $do(X = x)$. Especially, if there is no backdoor path from X to Y , then we have $Prob(Y = y | do(X = x)) = Prob(Y = y | X = x)$.

2. frontdoor criteria

The frontdoor operator criterion is defined as follows:

Definition 8.2. A set of variables Z is said to satisfy the frontdoor criterion relative to an ordered pair of variables (X, Y) if

- i. Z intercepts all directed paths from X to Y ,
- ii. There is no (unblocked) back-door path from X to Z ,
- iii. All backdoor paths from Z to Y are blocked by X .

Example 23 Consider the following graph Let us consider the variable set $\{Z\}$ and the causal relation

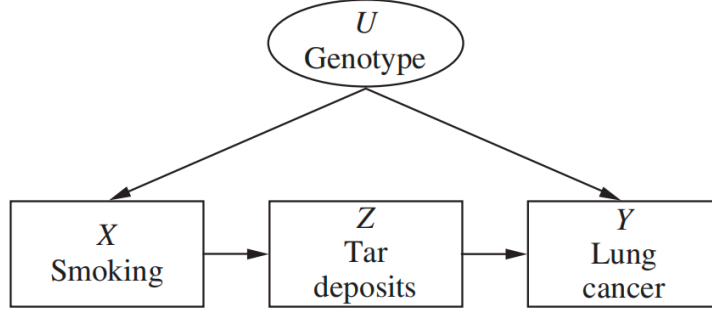


Figure 6: Caption

$X \rightarrow Y$. First Z intercepts the direct path X to Y , and there is no unblocked path from X to Z . Finally, the backdoor path from Z to Y is $ZXUY$, which is blocked by X . So Z satisfy the frontdoor criterion.

Example 24 Consider the other graph Suppose we want to evaluate the causal relation $Z \rightarrow Y$, then

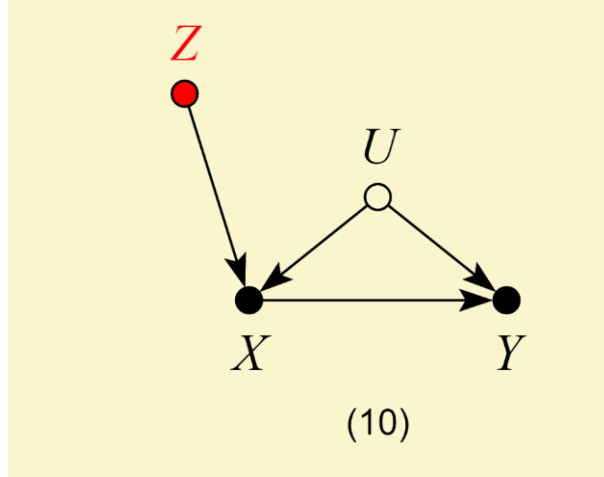


Figure 7: Caption

X does not form the frontdoor criterion since the backdoor path XUY does not contain Z .

Similar to the backdoor criterion, we have the frontdoor adjustment

$$Prob(Y = y|do(X = x)) = \sum_z Prob(Z = z|X = x) \sum_{x'} Prob(Y = y|X = x', Z = z) Prob(X = x').$$

9 Meditation effect and meditation analysis

So far we have discussed how to find the causality with the presence of the confounder W , i.e., $X \leftarrow W \rightarrow Y$. This section consider another situation. When one variable X causes another variable Y , it may both affects directly, i.e., $X \rightarrow Y$, or indirectly through mediating variables, like $X \rightarrow Z \rightarrow Y$. For example, consider the variables ‘Gender’, ‘Qualification’, and ‘Hiring’ in the following causal graph. Suppose we

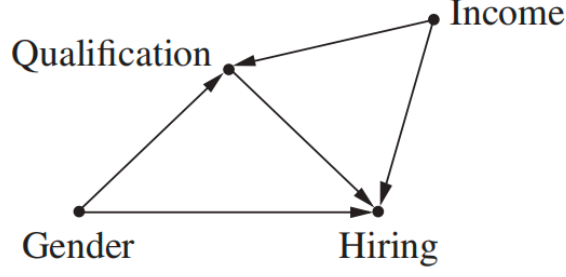


Figure 8: Caption

want to examine the gender discrimination in hiring, we want to evaluate the direct effect $G \rightarrow H$. Normally, in order to achieve the goal, we want to calculate $Prob(H|G = 1) - Prob(H|G = 0)$. However, gender may also affects job qualification, like some jobs that are more suitable for women to do while some jobs are more suitable for man to do. In this figure, $do(G = 1)$ is meaningless, because it still affects qualification, and does not reveal the direct effect $G \rightarrow H$ in hiring process. To illustrate, even if there is no direct causation $G \rightarrow H$, due to the qualification effect, the difference $Prob(H|G = 1) - Prob(H|G = 0)$ still may not be 0. We call ‘Qualification’ here the mediation variable.

Another idea is to conditional on the mediation variables. But it still has problems. To see why, if we conditional on qualification, then the gender can affect hiring through the path $GQIH$, which is still not good.

The third idea to solve this issue is through do two stuffs, i.e., consider the so-called “controlled direct effect ”

$$CDE = Prob(Y = y|do(X = x), do(Z = z)) - Prob(Y = y|do(X = x'), do(Z = z)),$$

and evaluate this for all z .

Example 25 Suppose the SCM for the random variables are

$$\begin{aligned}
 G &= U_G, \\
 I &= U_I, \\
 Q &= \alpha_{QG}G + \alpha_{QI}I + U_Q, \\
 H &= \alpha_{HG}G + \alpha_{HQ}Q + \alpha_{HI}I + U_H,
 \end{aligned} \tag{2}$$

After the do operation $do(G = g), do(Q = q)$, the new system becomes

$$G = g,$$

$$I = U_I,$$

$$Q = q,$$

$$H = \alpha_{HG}g + \alpha_{HQ}q + \alpha_{HI}I + U_H,$$

$$\text{and } (\mathbf{E}H|do(G = 1), do(Q = q) - \mathbf{E}H|do(G = 0), do(Q = q)) = \alpha_{HG}.$$

Quantifying the presence of mediation effect is also an interesting and important topic in causal inference, so we spend some time to introduce this.

Suppose the data we collect is (T_i, M_i, Y_i) , where T_i is the binary treatment variable, $T_i = 1$ means the patient receives treatment. The mediator variable is M_i , and the outcome is Y_i . Following Imai et al. [2010](#), we fit the following three models:

$$i.Y_i = \alpha_1 + \beta_1 T_i + U_i,$$

$$ii.M_i = \alpha_2 + \beta_2 T_i + V_i,$$

$$iii.Y_i = \alpha_3 + \beta_3 T_i + \gamma M_i + W_i.$$

After that, we test the 3 hypothesis $\beta_1 = 0$, $\beta_2 = 0$, $\gamma = 0$. If all these 3 hypothesis are rejected, then we plug-in i, ii to iii, and have

$$Y_i = \alpha_3 + \beta_3 T_i + \gamma(\alpha_2 + \beta_2 T_i + V_i) + W_i = \alpha_3 + \gamma\alpha_2 + (\beta_3 + \gamma\beta_2)T_i + \gamma V_i + W_i.$$

Therefore, we can consider $\gamma\beta_2$ to be the mediation effect of T on Y .

10 Causal inference in linear systems & Causal discovery

The linear SCM assumes that all relations are linear, that is,

$$\mathbf{E}X_i|X_j, j \in pa_i = \sum_{j \in pa_i} r_j X_j.$$

By adopting the linearity assumption, the path coefficient r_j along each edge in the causal graph fully quantifies the contribution of X_j on X_i . Moreover, in the linear SCM, statisticians can iterate the SCM to derive the (direct/total) causal effect.

Example 26 Suppose eq.(2) and we want to evaluate the causal effect between G and H . Notice that

$$\begin{aligned} H &= \alpha_{HG}G + \alpha_{HQ}(\alpha_{QG}G + \alpha_{QI}U_I + U_Q) + \alpha_{HI}U_I + U_H \\ &= (\alpha_{HG} + \alpha_{HQ}\alpha_{QG})G + (\alpha_{HQ}\alpha_{QI} + \alpha_{HI})U_I + \alpha_{HQ}U_Q + U_H. \end{aligned}$$

So the total causal effect is given by $(\alpha_{HG} + \alpha_{HQ}\alpha_{QG})$. In other words, if we make a regression of H on G , then the what we estimate is $\alpha_{HG} + \alpha_{HQ}\alpha_{QG}$ rather than α_{HG} .

Example 27 Suppose eq.(3). If we want to evaluate the causal effect from Z to Y , then we have

$$\begin{aligned} Y &= \alpha_{YZ}Z + \alpha_{YX}(\alpha_{XZ}Z + \alpha_{XE}U_E + U_X) + \alpha_{YA}U_A + U_Y \\ &= (\alpha_{YZ} + \alpha_{YX}\alpha_{XZ})Z + \alpha_{YX}\alpha_{XE}U_E + \alpha_{YX}U_X + \alpha_{YA}U_A + U_Y. \end{aligned}$$

Therefore, if we want to estimate the total effect, we may collect data, and do a regression of Y on Z . However, we should very carefully because the coefficient we estimate here is not the structural coefficient α_{YZ} .

In order to identify the system, we further assume that the errors U_S are non-Gaussian, resulting in the so-called “Linear, Non-Gaussian, Acyclic Model(LiNGAM)”.

Definition 10.1 (LiNGAM). Suppose the data $\mathbf{x}^{(i)} = (\mathbf{x}_1^{(i)}, \dots, \mathbf{x}_j^{(i)})^\top \in \mathbf{R}^p$ are generated from a process with the following properties:

- i. The observed dimensions $\mathbf{x}_j^{(i)}$ can be arranged in a causal order $\{k(1), k(2), \dots, k(p)\}$, where $k(i)$ is a permutation of $1, 2, \dots, p$, such that no later variable causes earlier variable.
- ii. The value assigned to each dimension $\mathbf{x}_j^{(i)}$ is a linear function of the values that already assigned to the earlier variables and a noise term, i.e.,

$$\mathbf{x}_j^{(i)} = \sum_{k(s) < k(j)} b_{js}\mathbf{x}_s^{(i)} + e_j^{(i)} + c_s.$$

- iii. The noises $e_j^{(i)}$ are **non-Gaussian** i.i.d. random variables.

Example 28 Suppose the data $(\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, \mathbf{x}_3^{(i)})^\top$ are generated from the model

$$\begin{aligned} \mathbf{x}_2^{(i)} &= 1.0 + e_2^{(i)}, \\ \mathbf{x}_3^{(i)} &= -0.7 + e_3^{(i)}, \\ \mathbf{x}_1^{(i)} &= 0.7 + 0.5\mathbf{x}_2^{(i)} - 0.4\mathbf{x}_3^{(i)} + e_3^{(i)}, \\ \mathbf{x}_4^{(i)} &= 0.2 - 0.9\mathbf{x}_1^{(i)} + 0.7\mathbf{x}_2^{(i)} + e_4^{(i)}, \end{aligned}$$

then the data are from a LiNGAM with order 2314 or 3214. Clearly, the order is not unique. If we

write the matrix $\mathbf{B} = \{b_{js}\}_{j,s=1,\dots,p}$, then there is an interesting relation

$$\mathbf{x}^{(i)} = \mathbf{B}\mathbf{x}^{(i)} + \mathbf{e}^{(i)} + \mathbf{c} \Rightarrow \mathbf{E}\mathbf{x}^{(i)} = \mathbf{B}\mathbf{E}\mathbf{x}^{(i)} + \mathbf{c}.$$

Define $\mathbf{z}^{(i)} = \mathbf{x}^{(i)} - \mathbf{E}\mathbf{x}^{(i)}$, then

$$\begin{aligned} \mathbf{z}^{(i)} - \mathbf{B}\mathbf{z}^{(i)} &= \mathbf{x}^{(i)} - \mathbf{E}\mathbf{x}^{(i)} - \mathbf{B}\mathbf{x}^{(i)} + \mathbf{B}\mathbf{E}\mathbf{x}^{(i)} \\ &= \mathbf{B}\mathbf{x}^{(i)} + \mathbf{e}^{(i)} + \mathbf{c} - \mathbf{E}\mathbf{x}^{(i)} - \mathbf{B}\mathbf{x}^{(i)} + \mathbf{B}\mathbf{E}\mathbf{x}^{(i)} = \mathbf{e}^{(i)} \\ &\Rightarrow \mathbf{z}^{(i)} = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{e}^{(i)}. \end{aligned}$$

Besides, define $\mathbf{w}^{(i)} = \mathbf{P}\mathbf{z}^{(i)}$ for some permutation matrix, then

$$\mathbf{z}^{(i)} = \mathbf{B}\mathbf{z}^{(i)} + \mathbf{e}^{(i)} \Rightarrow \mathbf{w}^{(i)} = \mathbf{P}\mathbf{z}^{(i)} = \mathbf{P}\mathbf{B}\mathbf{P}^\top \mathbf{w}^{(i)} + \mathbf{P}\mathbf{e}^{(i)},$$

so the representation is invariant under permutation.

Remark [permutation matrix] A permutation matrix is a square binary matrix that has exactly one entry of 1 in each row and each column and 0s elsewhere. When used to multiply another matrix, say \mathbf{A} , results in permuting the rows (when pre-multiplying, to form $\mathbf{P}\mathbf{A}$) or columns (when post-multiplying, to form $\mathbf{A}\mathbf{P}$) of the matrix \mathbf{A} .

Moreover, the matrix \mathbf{B} has special structures. To illustrate, consider example [28](#). We have

$$\begin{bmatrix} \mathbf{z}_1^{(i)} \\ \mathbf{z}_2^{(i)} \\ \mathbf{z}_3^{(i)} \\ \mathbf{z}_4^{(i)} \end{bmatrix} = \begin{bmatrix} 0 & 0.5 & -0.4 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -0.9 & 0.7 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{z}_1^{(i)} \\ \mathbf{z}_2^{(i)} \\ \mathbf{z}_3^{(i)} \\ \mathbf{z}_4^{(i)} \end{bmatrix} + \mathbf{e}^{(i)}. \quad (3)$$

That is, with suitable order, the matrix \mathbf{B} will form a lower triangular matrix with first row 0. Therefore, if we are able to find an estimator of the matrix $\mathbf{S} = (\mathbf{I} - \mathbf{B})^{-1}$ (thus \mathbf{B}), then we can use this relation to establish the causal relation. Definite, this kind of definition raises identification issues. That is, maybe different kinds of systems share the same matrix \mathbf{B} . However, the following theorem ensures that the system is identifiable when the noises are not normal.

Theorem 10.1 (Darmois-Skitovič). *Let X_1, \dots, X_d be independent, non-degenerated random variables (variance not 0). If there exists non-vanishing coefficients $a_1, \dots, a_d, b_1, \dots, b_d$ (i.e., for all i ,*

$a_i \neq 0 \neq b_i$) such that two linear combinations

$$l_1 = \sum_{i=1}^d a_i X_i, \quad l_2 = \sum_{i=1}^d b_i X_i$$

are independent, then each X_i is normally distributed.

Example 29 Counter-example Consider the system

$$Y = 2X + e_Y, \quad X = e_X, \quad e_X, e_Y \sim N(0, 1).$$

Consider another representation

$$X = \frac{1}{2}Y - \frac{1}{2}e_Y$$

To achieve our goal, we introduce an important tool named “Independent Component Analysis(ICM)” that finds the matrix \mathbf{B} .

11 Independent Component Analysis

Suppose for each $i = 1, 2, \dots, n$ and each $j = 1, \dots, p$, we have a vector $\mathbf{s}^{(i)} = (\mathbf{s}_1^{(i)}, \dots, \mathbf{s}_p^{(i)})^\top$ consisting of i.i.d. elements. Assume that

$$\mathbf{x}^{(i)} = \mathbf{M}\mathbf{s}^{(i)} = \sum_{j=1}^p \mathbf{s}_j^{(i)} \mathbf{M}_{.j},$$

we want to estimate the matrix \mathbf{M} . First we have some clarifications as follow:

- i. It is impossible to estimate the variance of $\mathbf{s}_j^{(i)}$, since

$$\mathbf{x}^{(i)} = \sum_{j=1}^p \frac{\mathbf{s}_j^{(i)}}{\zeta_j} \times \zeta_j \mathbf{M}_{.j}.$$

- ii. It is impossible to estimate the order of $\mathbf{s}^{(i)}$ (in other words, the j here is only dummy variable) since

$$\mathbf{x}^{(i)} = \mathbf{M}\mathbf{s}^{(i)} = \mathbf{M}\mathbf{P}\mathbf{P}^\top \mathbf{s}^{(i)}$$

for any permutation matrix. Since

$$\mathbf{s}^{(i)} = \mathbf{M}^{-1}\mathbf{x}^{(i)} \Rightarrow \mathbf{E}\mathbf{s}^{(i)} = \mathbf{M}^{-1}\mathbf{E}\mathbf{x}^{(i)},$$

so we can consider demean before doing ICM.

Let us consider finding an independent component. Consider a linear combination vector $\mathbf{p} \in \mathbf{R}^p$,

Input: data $\mathbf{x}^{(i)}, i = 1, 2, \dots, n$.

1. Calculate the sample mean $\bar{\mathbf{x}}_j = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_j^{(i)}$, then calculate $\mathbf{z}_j^{(i)} = \mathbf{x}_j^{(i)} - \bar{\mathbf{x}}_j$.
2. Define the function $H(\mathbf{u}), \mathbf{u} \in \mathbf{R}^p$ as follows:

$$H(\mathbf{u}) = \frac{1}{n} \sum_{i=1}^n (\mathbf{u}^\top \mathbf{z}^{(i)})^4 - \frac{3}{n} \left(\frac{1}{n} \sum_{i=1}^n (\mathbf{u}^\top \mathbf{z}^{(i)})^2 \right)^2$$

3. Find the maximizers $\hat{\mathbf{u}}$ of $H^2(\cdot)$, and the independent components in $\mathbf{x}^{(i)}$ is given by $\hat{\mathbf{u}}^\top \mathbf{x}^{(i)}$.
-

we have

$$\lambda_i = \mathbf{p}^\top \mathbf{x}^{(i)} = \mathbf{p}^\top \mathbf{M} \mathbf{s}^{(i)} = \mathbf{q}^\top \mathbf{s}^{(i)} = \sum_{j=1}^p \mathbf{q}_j \mathbf{s}_j^{(i)}, \text{ where } \mathbf{q} = \mathbf{p}^\top \mathbf{M}.$$

For $\mathbf{s}_j^{(i)}$ are i.i.d., what can expect is, if all $\frac{\mathbf{q}_j}{\|\mathbf{q}\|_2}$ are not large, then due to the central limit theorem, the distribution of λ_i should be similar to the normal distribution. On the other hand, if occasionally only one element in \mathbf{q} is not 0, from the assumption the distribution of λ_i should not be normal. This idea forms the motivation of our algorithm, that is, to maximize non-Gaussianity.

To achieve the goal, we need some index that illustrates the difference between normal distribution and others. Kurtosis, defined by

$$Kurt(y) = \mathbf{E}y^4 - 3(\mathbf{E}y^2)^2,$$

is differently one of the useful criteria for measuring non-Gaussianity.

Remark For standard normal random variable, we have $Kurt(y) = 0$.

Now we illustrate how to estimate the independent component.

However, this algorithm is not very helpful in the sense that the function H is highly non-convex (actually it should contain $2p$ maximizers). Therefore, machine learning scientists have developed various more practical methods for ICA, like those in Sklearn.

With the help of the ICA, we are now able to develop the causal order, i.e., in which order the matrix \mathbf{B} will form a lower triangular matrix. We introduce the method proposed in Shimizu et al. [2006](#) to find the desired order.

We explain the validity of this algorithm using eq.([3](#)). The algorithm actually has two steps, i.e., reconstructing the matrix \mathbf{B} and find the correct order. From eq.([3](#)), the ICA may recover the matrix $\mathbf{M} = (\mathbf{I} - \mathbf{B})^{-1} \mathbf{V}$ with some permutation matrix \mathbf{V} . Therefore,

$$\mathbf{M}^{-1} = \mathbf{V}^\top (\mathbf{I} - \mathbf{B}) = \mathbf{V}^\top \begin{bmatrix} 1 & -0.5 & 0.4 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0.9 & -0.7 & 0 & 1 \end{bmatrix}.$$

Thus, step 3 tries to eliminate the effect of the permutation matrix \mathbf{V} . After recovering \mathbf{B} , since we

know that the structural model is equivalent under permutation, we want to find a form that makes \mathbf{B} more close to lower triangular matrix. The matrix we use can be

$$\mathbf{P}^\dagger = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Since $K(1) = 2, K(2) = 3, K(3) = 1, K(4) = 4$, the order can be 2314, and we find the order.

Remark [Practical issue in finding permutations] When the dimension is relatively large, it can be hard to find the suitable permutation with brute force. Therefore, we want to seek an efficient method to calculate the permutation. For step 3, suppose the permutation is $k(i)$, then

$$\tilde{\mathbf{L}}_{ii} = \hat{\mathbf{L}}_{k(i)i}.$$

So essentially we want to minimize the term

$$\sum_{i=1}^p \frac{1}{|\hat{\mathbf{L}}_{k(i)i}|},$$

and there are existing polynomial-order algorithm that solves this in the literature.

The algorithm for finding the lower-triangular matrix is as follows:

12 Pruning edges after finding the order

After finding the orders, we have

$$\mathbf{x}_j^{(i)} = \sum_{k(s) < k(j)} b_{js} \mathbf{x}_s^{(i)} + e_j^{(i)} + c_s.$$

In other words, we can ensure that the errors $e_j^{(i)}$ are exogenous with the $\mathbf{x}_s^{(i)}$, i.e., the $e_j^{(i)}$ is independent of $\mathbf{x}_s^{(i)}$. That said, we can use the hypothesis testing for testing the existence of certain edges.

References

Imai, Kosuke, Luke Keele, and Teppei Yamamoto (2010). “Identification, Inference and Sensitivity Analysis for Causal Mediation Effects”. *Statistical Science* 25.1, pp. 51–71.

Input: the data $\mathbf{x}^{(i)} \in \mathbf{R}^p$ with $i = 1, 2, \dots, n$.

1. Calculate the sample mean $\bar{\mathbf{x}}_j = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_j^{(i)}$, then derive $\mathbf{z}_j^{(i)} = \mathbf{x}_j^{(i)} - \bar{\mathbf{x}}_j$.
2. Perform ICA on the data $\mathbf{z}^{(i)}$, i.e., find the matrix $\hat{\mathbf{B}}$ such that approximately $\mathbf{z}^{(i)} = \hat{\mathbf{M}}\mathbf{s}^{(i)}$ with $\mathbf{s}^{(i)}$ consisting of independent components. Define $\hat{\mathbf{L}} = \hat{\mathbf{M}}^{-1}$.
3. Find the permutation of rows of $\hat{\mathbf{L}}$, i.e., $\tilde{\mathbf{L}}$, that minimizes

$$H = \sum_{i=1}^p \frac{1}{|\tilde{\mathbf{L}}_{ii}|}.$$

After that, divide each row of $\tilde{\mathbf{L}}_{ii}$ by its corresponding diagonal element, yielding the matrix $\tilde{\mathbf{L}}'$ such that

$$\tilde{\mathbf{L}}'_{ii} = \frac{\tilde{\mathbf{L}}_{ij}}{\tilde{\mathbf{L}}_{ii}}.$$

4. Calculate the matrix $\hat{\mathbf{B}} = \mathbf{I} - \tilde{\mathbf{L}}'$. Then we find a permutation matrix \mathbf{P} making the matrix $\tilde{\mathbf{B}} = \mathbf{P}\hat{\mathbf{B}}\mathbf{P}^\top$ be “as close as possible to lower triangular matrix”. In practice, we find the \mathbf{P} that minimizes

$$\sum_{i=1}^p \sum_{j>i} \tilde{\mathbf{B}}_{ij}^2.$$

5. The permutation is given by $k(i) = j$ where $\mathbf{P}_{ij} = 1$.
-

-
1. Initialize the permutation p to be an empty list.
 2. Repeat until $\tilde{\mathbf{B}}$ has no elements:
 - 2.i Find a row i of $\tilde{\mathbf{B}}$ containing all zeros.
 - 2.ii Append i to the list p .
 - 2.iii Remove the i th row and column from $\tilde{\mathbf{B}}$.
-

Input: the observations $\mathbf{x}^{(i)}$.

1. Fit the linear model of $\mathbf{x}_j^{(i)}$ on \mathbf{x}_s with $k(s) < k(j)$, derive the coefficients \hat{b}_{js} .
2. Perform T-test on the hypothesis

$$H_0 : b_{js} = 0 \text{ versus } H_1 : b_{js} \neq 0,$$

i.e., calculate the estimator and the p-value

$$\hat{T} = \frac{|\hat{b}_{js}|}{\sqrt{\hat{\sigma}^2 \mathbf{e}_{k(s)}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{e}_{k(s)}}}, \quad \hat{p} = 2 - 2\Phi(\hat{T}).$$

Reject the null hypothesis if \hat{p} is small.

Shimizu, Shohei, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen (2006). “A Linear Non-Gaussian Acyclic Model for Causal Discovery”. *Journal of Machine Learning Research* 7.72, pp. 2003–2030.