DAG-Based Causal Inference

Qing Zhou

UCLA Department of Statistics

Stats 212 Graphical Models
Lecture Notes

Outline

1. Causal DAGs and intervention
2. Linear structural equation models
3. Estimation of causal effect
4. Semi-Markov causal models
5. Potential outcome approach

Causal DAGs and intervention

(Reference: Pearl (2000) §3.1 and §3.2; Pearl (1995))

Definition: A causal model among $X_1, \ldots, X_p$ is defined by a DAG $\mathcal{G}$ and a distribution $\mathbb{P}(\varepsilon) = \mathbb{P}(\varepsilon_1, \ldots, \varepsilon_p)$.

- Each child-parent relationship in $\mathcal{G}$, (X_j, PA_j), represents a functional relationship (structural equation model, SEM):

$$X_j = f_j(PA_j, \varepsilon_j), \qquad j = 1, \ldots, p. \tag{1}$$

- The background (error) variables are jointly independent:

$$\mathbb{P}(\varepsilon_1, \ldots, \varepsilon_p) = \prod_j \mathbb{P}(\varepsilon_j). \tag{2}$$

- (18) and (2) imply that $\mathbb{P}(X_1, \ldots, X_p)$ is Markovian with respect to the DAG $\mathcal{G}$:

$$\mathbb{P}(X_1, \ldots, X_p) = \prod_{j=1}^{p} \mathbb{P}(X_j \mid PA_j). \tag{3}$$

Causal DAGs and intervention

Causal effect defined via external intervention:

- Consider an atomic intervention that forces X_i to some fixed value x_i, which we denote by $do(X_i = x_i)$ or $do(x_i)$ for short.
- Effect of $do(x_i)$: to replace the SEM for X_i by $X_i = x_i$ and substitute $X_i = x_i$ in the other SEMs.
- For two distinct sets of variables X and Y, the causal effect of X on Y is determined by the mapping

$$x \mapsto \mathbb{P}[Y \mid do(X = x)] \equiv \mathbb{P}(Y \mid do(x)).$$

Examples of causal effects.

1. linear SEM: Causal effect $\frac{\partial \mathbb{E}(Y|do(x))}{\partial x}$.
2. Treatment ($X = 1$) vs control ($X = 0$): Causal effect $\mathbb{E}(Y \mid do(X = 1)) - \mathbb{E}(Y \mid do(X = 0))$.

Causal DAGs and intervention

Model interventions as variables:

- Treat intervention as additional variable in the DAG: F_j for intervention on X_j.
- SEM for X_j change to

$$X_j = h_j(PA_j, F_j, \varepsilon_j) = \begin{cases} f_j(PA_j, \varepsilon_j), & \text{if } F_j = idle \\ x, & \text{if } F_j = do(x). \end{cases} \tag{4}$$

- Augment the parents of X_j to $PA_j \cup \{F_j\}$:

$$\mathbb{P}(X_j = x_j \mid PA_j, F_j) = \begin{cases} \mathbb{P}(X_j = x_j \mid PA_j), & \text{if } F_j = idle \\ I(x_j = x), & \text{if } F_j = do(x), \end{cases}$$

  assuming all X_j are *discrete* for convenience.

Causal DAGs and intervention

Computing causal effect (of interventions): To simplify notation, consider discrete X_j and write $\mathbb{P}(X = x) = P(x)$.

- *Truncated factorization* of $P(x_1, \ldots, x_p)$ given $do(X_i = x_i^*)$:

$$P(x_1, \ldots, x_p \mid do(x_i^*)) = I(x_i = x_i^*) \prod_{j \neq i} P(x_j \mid pa_j), \qquad (5)$$

  where $pa_j = (x_k : k \in PA_j)$.

- Multiple interventions $do(X_S = \mathbf{x}^*)$, $S \subset \{1, \ldots, p\}$:

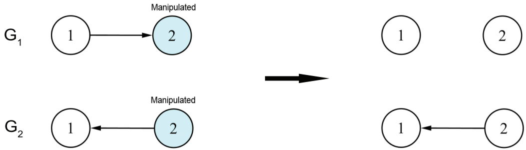
$$P(x_1, \ldots, x_p \mid do(\mathbf{x}^*)) = I(x_S = \mathbf{x}^*) \prod_{j \notin S} P(x_j \mid pa_j). \qquad (6)$$

- Graph structure change when $do(X_i = x_i^*)$: delete edges $X_j \to X_i$ for all $j \in PA_i$, i.e. change $\mathcal{G}$ to $\mathcal{G}_{\bar{X}_i}$.

Causal DAGs and intervention

Difference between $P(y \mid do(x))$ and $P(y \mid x)$.

- Two DAGs G_1 and G_2 on X_1, X_2:



- Find $P(x_1 \mid do(x_2))$ with respect to G_1 and G_2.

$$
\begin{aligned}
G_1: &\quad P(x_1 \mid do(x_2)) = P(x_1), \\
G_2: &\quad P(x_1 \mid do(x_2)) = P(x_1 \mid x_2).
\end{aligned}
$$

Causal DAGs and intervention

From (5), putting $x_i = x_i^*$:

$$\begin{aligned} P(x_{-i} \mid do(x_i^*)) &= \prod_{j \neq i} P(x_j \mid pa_j) \cdot \frac{P(x_i^* \mid pa_i)}{P(x_i^* \mid pa_i)} \\ &= \frac{P(x_1, \ldots, x_p)}{P(x_i^* \mid pa_i)} \\ &= P(x_j, j \in B \mid x_i^*, pa_i) P(pa_i), \end{aligned} \tag{7}$$

where $B = [p] \setminus \{i, PA_i\}$ and $[p] := \{1, \ldots, p\}$.

- Intervention event (*do*-operator) *not* on the right-hand side.
- Compute causal effect (intervention probability) by conditional probabilities (pre-intervention probabilities) that can be estimated from observational data.

Causal DAGs and intervention

Theorem 1 (Adjustment for direct causes)

Let PA_i be the parents of X_i and Y be any set of other variables in a causal DAG $\mathcal{G}$. Then the causal effect of $do(X_i = x_i)$ on Y is given by

$$P(y \mid do(x_i)) = \sum_{pa_i} P(y \mid x_i, pa_i) P(pa_i), \tag{8}$$

where $P(y \mid x_i, pa_i)$ and $P(pa_i)$ are pre-intervention probabilities.

Proof.

Marginalize out $X_j \notin Y \cup \{X_i\}$ on both sides of (7). $\square$

Causal DAGs and intervention

A simple implication of Theorem 1:

If Y is a set of non-descendants of X_i, then

$$Y \perp X_i \mid PA_i.$$

By Theorem 1

$$\begin{aligned} P(y \mid do(x_i)) &= \sum_{pa_i} P(y \mid x_i, pa_i) P(pa_i) \\ &= \sum_{pa_i} P(y \mid pa_i) P(pa_i) = P(y), \end{aligned}$$

which is independent of the intervention on X_i. Thus, X_i has no causal effect on Y.

Linear structural equation models

A causal model $(\mathcal{G}, \mathbb{P}_\varepsilon)$ with linear SEMs:

- A linear model for each child-parent relationship:

$$X_j = \sum_{i \in PA_j} \beta_{ij} X_i + \varepsilon_j, \qquad j = 1, \ldots, p. \tag{9}$$

- ε_j's are independent and $\mathbb{E}(\varepsilon_j) = 0$;
- Usually assume $\varepsilon_j \sim \mathcal{N}(0, \omega_j^2)$. In this case, the DAG is called a Gaussian DAG and the graphical model is called a Gaussian Bayesian network.

Linear structural equation models

Causal effect:

- The causal effect of X_k on X_j

$$\begin{aligned}\gamma_{kj} &:= \frac{\partial \mathbb{E}(X_j \mid do(X_k = x))}{\partial x} \\ &= \mathbb{E}(X_j \mid do(X_k = c+1)) - \mathbb{E}(X_j \mid do(X_k = c)), \end{aligned} \tag{10}$$

  for any $c \in \mathbb{R}$, due to the linear model assumption.

- Using modified DAG $\mathcal{G}_{\bar{X}_k}$ after intervention,

$$\mathbb{E}(X_j \mid X_k = x; \mathcal{G}_{\bar{X}_k}) = \gamma_{kj} x,$$

  where $\mathbb{E}(\bullet; \mathcal{G}_{\bar{X}_k})$ takes expectation with respect to $\mathcal{G}_{\bar{X}_k}$.

Linear structural equation models

Apply Theorem 1 to find γ_{kj}:

- Let $Z = PA_k$ and z denote the value of PA_k,

$$p(x_j \mid do(X_k = x_k)) = \int_z p(x_j \mid x_k, z) p(z) dz,$$

  where the p on the right side is given by the pre-intervention distribution (that of $\mathcal{G}$).

- Let (β, α) be the regression coefficient of X_j on (X_k, PA_k), that is, $\mathbb{E}(X_j \mid X_k, Z) = \beta X_k + \alpha^\mathsf{T} Z$, which can be estimated from observational data.

- Then the causal effect

$$\begin{aligned}\gamma_{kj} &= \frac{\partial}{\partial x_k} \mathbb{E}(X_j \mid do(X_k = x_k)) \\ &= \frac{\partial}{\partial x_k} \int_z \left\{ \beta x_k + \alpha^\mathsf{T} z \right\} p(z) dz = \beta.\end{aligned}$$

Estimation of causal effect

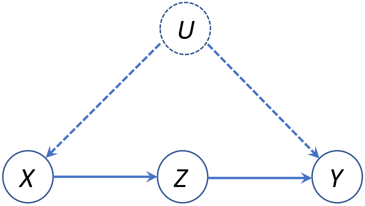
Reference: Pearl (2000) §3.3.

Problem setup:

- Given a causal DAG $\mathcal{G}$, if $P(y \mid do(x))$ can be uniquely computed from the (pre-intervention) distributions of observed variables in $\mathcal{G}$, then we say the causal effect of X on Y is identifiable.
- Note that we allow unobserved nodes in $\mathcal{G}$.
- Only observational data are collected.

Estimation of causal effect

Example: Observed nodes $X \to Z \to Y$; hidden node U, a common parent of X and Y (sometimes called a confounder).



Can we estimate the causal effect of X on Y or of Z on Y from observational data collected for (X, Y, Z)?

Estimation of causal effect

Back-door adjustment:

- Theorem 1 implies: If X, PA_X, Y are observed, then $P(y \mid do(x))$ is identifiable by (8).
- Theorem 1 is a special case of back-door adjustment: PA_X satisfies the back-door criterion relative to X and Y.
- *Back-door criterion*: A set of variables Z satisfies the back-door criterion relative to an ordered pair of variables (X, Y) in a DAG $\mathcal{G}$ if
  1. no nodes in Z is a descendant of X;
  2. Z blocks every chain between X and Y that contains an arrow into X (backdoor path).

Estimation of causal effect

Theorem 2 (Back-door adjustment)

If Z satisfies the back-door criterion relative to (X, Y). Then the causal effect of X on Y is given by

$$P(y \mid do(x)) = \sum_z P(y \mid x, z)P(z). \tag{11}$$

Proof.

Add intervention variable $F_X \to X$ to $\mathcal{G}$:

$$\begin{aligned} P(y \mid do(x)) &= \sum_z P(y \mid do(x), z)P(z \mid do(x)) \\ &= \sum_z P(y \mid F_X = do(x), x, z)P(z). \end{aligned}$$

Invoke that (X, Z) d-separates F_X and Y. $\square$

Estimation of causal effect

Linear SEM: By (11), the causal effect can be identified by regressing Y on (X, Z):

$$\gamma_{X \to Y} := \frac{\partial}{\partial x} \mathbb{E}(Y \mid do(x)) = \beta_X(Y \sim X + Z).$$

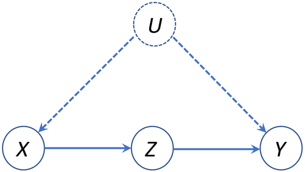
Suppose we have data observed for the three random variables X, Y, Z. Then to estimate the causal effect X on Y:

1. Discrete data: estimate $P(y \mid x, z)$ and $P(z)$ from data. Then plug into (11).
2. Linear SEM: least-squares regression Y on (X, Z), then

$$\widehat{\gamma}_{X \to Y} = \widehat{\beta}_X(Y \sim X + Z).$$

Estimation of causal effect

Example:



By Theorem 2,

$$P(y \mid do(z)) = \sum_x P(y \mid x, z)P(x), \quad P(z \mid do(x)) = P(z \mid x), \tag{12}$$

without observing U.

Estimation of causal effect

Is $P(y \mid do(x))$ identifiable? Yes, because:

$$\begin{aligned} P(y \mid do(x)) &= \sum_z P(y, z \mid do(x)) \\ &= \sum_z P(z \mid do(x)) P(y \mid z, do(x)) \\ &= \sum_z P(z \mid do(x)) P(y \mid do(z)). \end{aligned} \tag{13}$$

Last step uses $Y \perp F_Z \mid \{Z, do(x)\}$:

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

Then, plug (12) into (13) to get

$$P(y \mid do(x)) = \sum_z P(z \mid x) \sum_{x'} P(y \mid x', z) P(x'). \tag{14}$$

Estimation of causal effect

Eq. (14) is an example of *front-door adjustment*.

- Front-door criterion: Z satisfies the front-door criterion relative to (X, Y) if
  1. Z intercepts all directed paths from X to Y;
  2. there is no back-door path from X to Z; and
  3. all back-door paths from Z to Y are blocked by X.

Theorem 3 (Front-door adjustment)

If Z satisfies the front-door criterion relative to (X, Y), then

$$P(y \mid do(x)) = \sum_z P(z \mid x) \sum_{x'} P(y \mid x', z) P(x'). \tag{15}$$

- Linear SEMs:

$$\gamma_{X \to Y} = \gamma_{X \to Z} \times \gamma_{Z \to Y} = \beta_X(Z \sim X) \times \beta_Z(Y \sim Z + X).$$

Estimation of causal effect

Proof of Theorem 3.

(i) Condition 1 implies

$$P(y \mid do(x)) = \textstyle\sum_z P(z \mid do(x))P(y \mid do(z)).$$

(ii) Backdoor adjustment with Condition 2 shows that

$$P(z \mid do(x)) = P(z \mid x).$$

(iii) Backdoor adjustment with Condition 3 shows that

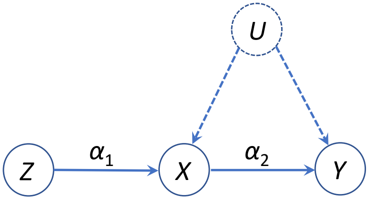
$$P(y \mid do(z)) = \textstyle\sum_{x'} P(y \mid x', z)P(x').$$

$\square$

Rules of do-calculus (Pearl (2000) §3.4): a set of inference rules for transforming intervention and observational probabilities, say to translate causal effect to conditional probabilities.

Estimation of causal effect

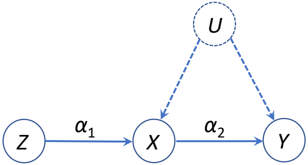
Instrumental variable formula (Bowden and Day 1984) (assume linear SEMs)



Observed nodes $Z \to X \to Y$, and U is hidden common parent of X and Y. Is $\gamma_{X \to Y} = \alpha_2$ identifiable?

Example. $X =$ college education, $Y =$ job after college, $U =$ family social/educational background, $Z =$ randomly assigned high-school fellowship for college application.

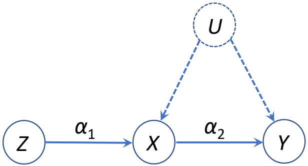
Estimation of causal effect



1. Z has no parents, thus α_1 is identifiable by regressing X on Z: $\alpha_1 = \beta_Z(X \sim Z)$.
2. Similarly, the causal effect of Z on Y, $\alpha_1\alpha_2$, is also identifiable: $\alpha_1\alpha_2 = \beta_Z(Y \sim Z)$.
3. Combined we have the *instrumental variable formula*:

$$\alpha_2 = \frac{\beta_Z(Y \sim Z)}{\beta_Z(X \sim Z)} = \frac{\mathrm{Cov}(Y, Z)}{\mathrm{Cov}(X, Z)}. \tag{16}$$

Estimation of causal effect



Two-stage least-squares:

1. Regress X on Z so $\alpha_1 = \beta_Z(X \sim Z)$ and let $\widehat{X} = \alpha_1 Z$.
2. Regress Y on $\widehat{X}$ and then $\alpha_2 = \beta_{\widehat{X}}(Y \sim \widehat{X})$:

$$\beta_{\widehat{X}}(Y \sim \widehat{X}) = \frac{\mathrm{Cov}(Y, \alpha_1 Z)}{\mathrm{Var}(\alpha_1 Z)} = \frac{\mathrm{Cov}(Y, Z)}{\alpha_1 \mathrm{Var}(Z)} = \alpha_2.$$

Note: To estimate α_2 from samples of (X, Y, Z), $\beta \to$ LSE $\widehat{\beta}$.

Estimation of causal effect

Conditional instrumental variable (Brito and Pearl 2002):
Z is said to be a *conditional instrumental variable* given S relative to (X, Y) if

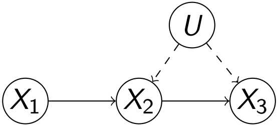
1. S contains no descendants of X or Y;
2. S d-separates Z from Y but not from X in the graph obtained after deleting all edges emerging from X.

Then, the causal effect of X on Y

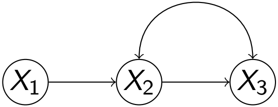
$$\gamma_{X \to Y} = \frac{\mathrm{Cov}(Y, Z \mid S)}{\mathrm{Cov}(X, Z \mid S)} = \frac{\beta_Z(Y \sim Z + S)}{\beta_Z(X \sim Z + S)}. \tag{17}$$

Semi-Markov causal models

DAG with hidden variables



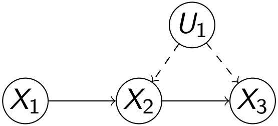
ADMG



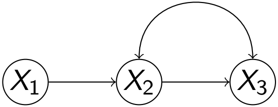
- If two nodes X_i and X_j share a common hidden parent U, remove U from the DAG and add a bidirected edge $X_i \leftrightarrow X_j$: acyclic directed mixed graph (ADMG).
- $X_i \leftrightarrow X_j$: their background variables ε_i and ε_j are dependent.
- A causal model with dependent background variables is called a semi-Markov causal model (SMCM).

Semi-Markov causal models

DAG with hidden variables



ADMG



- SEM for SMCM over $X = \{X_1, \ldots, X_p\}$:

$$X_j = f_j(PA_j, \varepsilon_j), \qquad j = 1, \ldots, p. \tag{18}$$

  $\varepsilon_i \perp \varepsilon_j$ if no bidirected edge between i and j.

- The joint distribution $\mathbb{P}(X)$ is obtained by marginalization of $\mathbb{P}(X, U)$ defined by a DAG on $X \cup U$:

$$P(x_1, \ldots, x_p) = \sum_{u_1, \ldots, u_d} P(x_1, \ldots, x_p \mid u_1, \ldots, u_d) \prod_i P(u_i).$$

Semi-Markov causal models

Let $Y(x) \equiv [Y \mid do(X = x)]$. Restrictions encoded by SMCM:

1. *Exclusion:* For any $S \subset V \setminus (PA_Y \cup \{Y\})$ (no directed edge from S to Y),

$$Y(pa_Y) = Y(pa_Y, s). \tag{19}$$

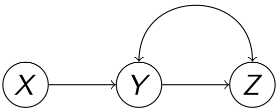
$\because$ both $= f_Y(pa_Y, \varepsilon_Y)$.

2. *Independence:* For any $Z \in V$ not connected to Y via bidirected edges,

$$Y(pa_Y) \perp Z(pa_Z). \tag{20}$$

$\because Y(pa_Y) = f_Y(pa_Y, \varepsilon_Y)$, $Z(pa_Z) = f_Z(pa_Z, \varepsilon_Z)$ and $\varepsilon_Y \perp \varepsilon_Z$.

Semi-Markov causal models



- Exclusion restrictions: $Y(x) = Y(x, z)$ and $X = X(y, z)$.
- Independence restrictions: $X \perp \{Y(x), Z(y)\}$, but $Y(x) \not\perp Z(y)$.

Potential outcome approach

Under potential outcome framework (Rubin 1990):

- $Y(x)$ is a counterfactual entity representing the potential outcome of Y had X been x.
- Suppose $X \in \{0, 1\}$ (treatment vs control). Want to estimate causal effect $\mathbb{E}[Y(1) - Y(0)]$ or $\mathbb{E}[Y(1) - Y(0) \mid X = 1]$.
- $\mathbb{P}^*[Y(x)]$ corresponds to $\mathbb{P}(Y \mid do(x))$. Making assumptions to calculate when $Y(x)$ is missing.

X	$Y(1)$	$Y(0)$	Z (covariates)
1	✓	?	✓,···,✓
1	✓	?	✓,···,✓
1	✓	?	✓,···,✓
0	?	✓	✓,···,✓
0	?	✓	✓,···,✓

Potential outcome approach

Assume conditional ignorability (Rosenbaum and Rubin 1983):
$Y(x) \perp X \mid Z$.

$$
\begin{aligned}
\mathbb{P}^*(Y(x) = y) &= \sum_z \mathbb{P}^*(Y(x) = y \mid z) P(z) \\
&= \sum_z \mathbb{P}^*(Y(x) = y \mid x, z) P(z) \\
&= \sum_z \mathbb{P}(Y = y \mid x, z) P(z) \quad \text{backdoor adjustment.}
\end{aligned}
$$

$$
\begin{tikzcd}
 & Z \arrow[dl] \arrow[dr] & \\
X \arrow[rr] & & Y
\end{tikzcd}
$$

$$
\begin{aligned}
&Y(x) = f_Y(x, Z, \varepsilon_Y) = h(Z, \varepsilon_Y) \\
&X = f_X(Z, \varepsilon_X) \\
&Y(x) \perp X \mid Z \Leftrightarrow \varepsilon_Y \perp \varepsilon_X \mid Z.
\end{aligned}
$$

References I

R.J. Bowden and N.E. Day. *Instrumental Variables*. Cambridge University Press, 1984.

C. Brito and Judea Pearl. Generalized instrumental variables. *Proceedings of the 18th Conferences on Uncertainty in Artificial Intelligence*, pages 85–93, 2002.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82:669–710, 1995.

Judea Pearl. *Causality: Models, reasoning and inference*. Cambridge Univ Press, 2000.

P. Rosenbaum and D.B. Rubin. The central role of propensity score in observational studies for causal effects. *Biometrika*, 70: 41–55, 1983.

D.B. Rubin. Formal models of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, 25: 279–292, 1990.