Causal DAGs: Inference and Learning

Qing Zhou

UCLA Department of Statistics and Data Science

Stats 201C Advanced Modeling and Inference
Lecture Notes

Outline

1. Causal DAGs and intervention
2. Linear structural equation models
3. Estimation of causal effect
4. Structure learning of DAGs

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 noise variables are jointly independent:

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

- (1) 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 $do(X_i = x_i^*)$:

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

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

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

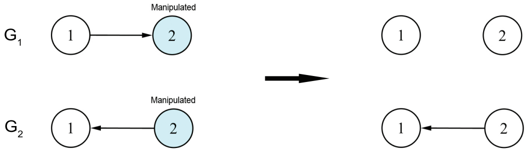
$$P(x_1, \ldots, x_p \mid do(\mathbf{x}^*)) = I(x_S = \mathbf{x}^*) \prod_{j \notin S} P(x_j \mid pa_j). \tag{6}$$

- 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 $x_i = x_i^*$:

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

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). □

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

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

  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

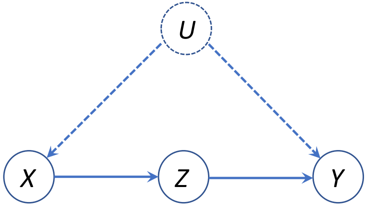
Reference: Pearl (2000) §3.3.

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 $X \rightarrow Z \rightarrow Y$; 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 path 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):

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

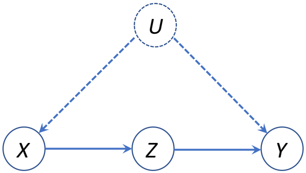
Suppose we have data observed for the three random variables X, Y, Z. Then to estimate the causal effect X on Y:

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,

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

without observing U.

Estimation of causal effect

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

$$\begin{aligned} P(y \mid do(x)) &= P(y \mid x; \mathcal{G}_{\bar{X}}) \\ &= \sum_z P(y \mid x, z; \mathcal{G}_{\bar{X}}) P(z \mid x; \mathcal{G}_{\bar{X}}) \\ &= \sum_z P(y \mid z; \mathcal{G}_{\bar{X}}) P(z \mid do(x)) \\ &= \sum_z P(y \mid do(z)) P(z \mid x). \end{aligned} \tag{12}$$

Linear SEMs:

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

Estimation of causal effect

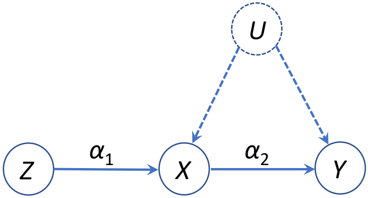
- Eq. (12) is an example of *front-door adjustment* (Theorem 3.3.4, Pearl (2000)):
  1. Z intercepts all directed paths from X to Y;
  2. there is no back-door path from X to Z; and
  3. all back-door paths from Z to Y are blocked by X.

  Then $P(y \mid do(x))$ is identifiable

$$P(y \mid do(x)) = \sum_z P(z \mid x) \sum_{x'} P(y \mid x', z) P(x'). \qquad (13)$$

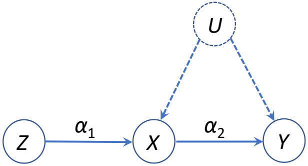
- Rules of do-calculus (Pearl (2000) §3.4): a set of inference rules for transforming intervention and observational probabilities, say to translate causal effect to conditional probabilities.

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?

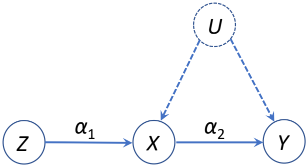
Estimation of causal effect



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*:

$$\alpha_2 = \frac{\beta_Z(Y \sim Z)}{\beta_Z(X \sim Z)} = \frac{\mathrm{Cov}(Y, Z)}{\mathrm{Cov}(X, Z)}. \qquad (14)$$

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 $\widehat{\beta}$.

Structure learning of DAGs

Structure learning: Given data $x_i = (x_{i1}, \ldots, x_{ip}) \sim (\mathcal{G}, \mathbb{P})$ (causal model), $i = 1, \ldots, n$, how to estimate the DAG $\mathcal{G}$?

- Constraint-based methods: Conditional independence tests against $X_i \perp X_j \mid X_S$ for all i, j, S.
- Score-based methods: Optimizing a scoring function over graph space.

See, e.g. Aragam and Zhou (2015) Section 1.2 for recent literature.

Data types:

- Observational data (no intervention)
- Experimental data (intervention available)

Structure learning of DAGs

Assumption: $\mathbb{P}(X_1, \dots, X_p)$ is faithful wrt $\mathcal{G}$:

Definition 1

For a graphical model $(\mathcal{G}, \mathbb{P})$, we say the distribution $\mathbb{P}$ is faithful to the graph $\mathcal{G}$ if for every triple of disjoint sets $A, B, S \subset V$,

$$X_A \perp X_B \mid X_S \Leftrightarrow S \text{ separates } (d\text{-separates}) \; A \text{ and } B.$$

- Conditional independence (CI) in $\mathbb{P} \Leftrightarrow$ d-separation in $\mathcal{G}$, i.e.

$$\mathcal{I}_{\mathbb{P}}(A, B|S) \Leftrightarrow \mathcal{D}_{\mathcal{G}}(A, B|S).$$

- Given $\mathcal{G}$, almost all parameter values in the SEMs will define a faithful $\mathbb{P}$.
- Structure learning: use CI relations learned from data to infer edges in $\mathcal{G}$.

Structure learning of DAGs

Suppose we only have observational data. What can be learned?

Definition 2 (Markov equivalence)

Two DAGs $\mathcal{G}$ and $\mathcal{G}'$ on the same set of nodes V are Markov equivalent if $\mathcal{D}_{\mathcal{G}}(X, Y|\boldsymbol{Z}) \Leftrightarrow \mathcal{D}_{\mathcal{G}'}(X, Y|\boldsymbol{Z})$ for any $X, Y \in V$ and $\mathbf{Z} \subseteq V \setminus \{X, Y\}$.

- Two DAGs are Markov equivalent if and only if they have the same skeletons and the same v-structures.
- A v-structure is a triplet $\{i, j, k\} \subseteq V$ of the form $i \to k \leftarrow j$: i and j are nonadjacent; k is called an *uncovered collider*.
- Equivalent DAGs form an equivalence class.
- DAGs in the same equivalence class cannot be distinguished from observational data. Thus we can only learn the equivalence class of $\mathcal{G}$ from observational data.

Structure learning of DAGs

How to represent an equivalence class? CPDAG (Completed partially DAG).

Two types of edges in a DAG $\mathcal{G}$:

- A directed edge $i \to j$ is *compelled* in $\mathcal{G}$ if for every DAG $\mathcal{G}'$ equivalent to $\mathcal{G}$, the edge $i \to j$ exists in $\mathcal{G}'$.
- If an edge is not compelled in $\mathcal{G}$, then it is *reversible*.

Definition 3 (CPDAG)

The CPDAG of an equivalence class is the PDAG consisting of a directed edge for every compelled edge in the equivalence class, and an undirected edge for every reversible edge in the equivalence class.

Examples:

Structure learning of DAGs

Theorem 3 (Spirtes et al. (1993))

Suppose $(\mathcal{G}, \mathbb{P})$ satisfies the faithfulness assumption. Then there is no edge between a pair of nodes $X, Y \in V$ if and only if there exists a subset $\boldsymbol{Z} \subseteq V \setminus \{X, Y\}$ such that $\mathcal{I}_P(X, Y | \boldsymbol{Z})$.

Constraint-based methods:

1. Find the skeleton of $\mathcal{G}$ by CI tests;
2. Identify v-structures;
3. Orient other edges.

Output: CPDAG (or PDAG)

Structure learning of DAGs

<u>Outline of PC algorithm</u> (Spirtes and Glymour 1991):

1: $E \leftarrow$ edge set of the complete undirected graph on V.
2: **for** $(i, j) \in E$ **do**
3: Search for a subset S_{ij} of either $N_i(E)$ or $N_j(E)$ such that $X_i \perp X_j \mid S_{ij}$. If found, $E \leftarrow E \setminus \{(i, j), (j, i)\}$ and store S_{ij}.
4: **end for**
5: Identify v-structures based on E and $\{S_{ij}\}$.
6: Orient as many edges in E as possible by Meek's rules.

Notes:

1. Line 3: $N_i(E) = \{X_k : (i, k) \in E\}$.
2. For loop: implemented in ascending order of $|S_{ij}| = \ell$ for $\ell = 0, \ldots, \ell_{\max}$.
3. Line 1 to 4: Estimate skeleton $sk(\widehat{\mathcal{G}})$ of $\mathcal{G}$.

Edge orientation steps:

1. Identify v-structures (Line 5) given $sk(\widehat{\mathcal{G}})$:
   For all nonadjacent pair (i, j) with a common neighbor k, orient $i - k - j$ as $i \to k \leftarrow j$ if $k \notin S_{ij}$.
   Because otherwise, $X_i \not\perp X_j \mid S_{ij}$, contradiction.
   After this step, we obtain a PDAG.
2. Meek's rules (Line 6): In the resulting PDAG, orient as many undirected edges as possible by repeated application of four rules (Meek 1995).
   Basic idea: If orienting an undirected edge $i - j$ into $i \to j$ would result in additional v-structures or a directed cycle, then orient it into $i \leftarrow j$.

Structure learning of DAGs

Conditional independence tests ($H_0 : X \perp Y \mid S$):

- Gaussian data: partial correlation $\mathrm{cor}(X, Y \mid S) = 0$.
  1. Sample covariance matrix $\widehat{\Sigma}$ from data columns of (X, Y, S).
  2. $\widehat{\Omega} = (\omega_{ij}) \leftarrow \widehat{\Sigma}^{-1}$ and $\widehat{\rho}_{XY|S} = -\omega_{12}/\sqrt{\omega_{11}\omega_{22}}$.
  3. Fisher z-transformation,

$$z(X, Y|S) = \frac{1}{2} \log \left(\frac{1 + \widehat{\rho}_{XY|S}}{1 - \widehat{\rho}_{XY|S}} \right)$$

  and $\sqrt{n - |S| - 3} \cdot z(X, Y|S) \mid H_0 \sim \mathcal{N}(0, 1)$.

- Discrete data: G^2 or χ^2 test for conditional independence.

$$G^2(X, Y; S = s) = 2 \sum_{x,y} O_{xys} \log(O_{xys}/E_{xys}),$$

$$G^2(X, Y; S) = \sum_{s} G^2(X, Y; S = s) \mid H_0 \sim \chi^2_{(|X|-1)(|Y|-1)|S|},$$

E_{xys}: expected counts under H_0; O_{xys}: observed counts.

Structure learning of DAGs

Correctness and consistency:

Let $\widehat{\mathcal{G}}_n$ be the estimated graph by PC from a sample of size n and $\mathcal{C}$ be the CPDAG of $\mathcal{G}$. Suppose that $\mathbb{P}$ is faithful to $\mathcal{G}$.

1. CI oracles (Spirtes et al. 1993; Meek 1995): If all CI tests are perfect (oracle), then $\widehat{\mathcal{G}}_n = \mathcal{C}$.
2. Large-sample limit: When the sample size $n \to \infty$, all CI tests involved will be perfect (no type I or II error) with high probability. Then the PC algorithm estimates the CPDAG of $\mathcal{G}$ consistently, i.e.

$$\lim_{n\to\infty} \mathbb{P}(\widehat{\mathcal{G}}_n = \mathcal{C}) = 1.$$

Structure learning of DAGs

Score-based methods:

$$\widehat{\mathcal{G}} = \underset{G \in Space}{\text{argmax}}\, S(G, \mathbf{D}). \tag{15}$$

1. $\mathbf{D} = (x_{ij})_{n \times p} = [X_1 \mid \ldots \mid X_p]$ i.i.d. data from $(\mathcal{G}, \mathbb{P})$.
2. $S(G, \mathbf{D})$ is a scoring function: log-likelihood of $\mathbf{D}$ given a graph G with a penalty term on model complexity (number of edges or number of free parameters). For example,

$$S_{\text{BIC}}(G, \mathbf{D}) = \log p(\mathbf{D} \mid \widehat{\theta}, G) - \frac{d}{2} \log n, \tag{16}$$

   $\widehat{\theta}$: MLE of parameters under G, $d =$ dimension of θ.
3. Space of graph: DAG space or equivalence class (CPDAGs).

Structure learning of DAGs

BIC score for Gaussian DAGs:

- Liner SEM for data columns $X_j \in \mathbb{R}^n, j \in [p]$:

$$X_j = \sum_{i \in PA_j} \beta_{ij} X_i + \varepsilon_j, \qquad \varepsilon_j \sim \mathcal{N}_n(0, \omega_j^2 I_n).$$

- Decomposable:

$$\begin{aligned} S_{\mathrm{BIC}}(G, \mathbf{D}) &= \sum_{j=1}^{p} s(X_j, PA_j^G) \qquad (17) \\ &= \sum_j \log p(X_j \mid \widehat{\beta}_j, \widehat{\omega}_j^2, PA_j^G) - \frac{1}{2} |PA_j^G| \log n. \end{aligned}$$

$(\widehat{\beta}_j, \widehat{\omega}_j^2)$: MLEs in Gaussian regression $X_j \sim PA_j^G$.

Bayesian Dirichlet score for discrete DAGs (Heckerman et al. 1995):

- Multinomial distribution: $\theta_{ijk} = \mathbb{P}(X_i = k \mid PA_i = j)$. Parameter for $[X_i \mid PA_i]$ is a $q_i \times r_i$ table:

$$\Theta_i = \left\{ \theta_{ijk} : j \in [q_i], k \in [r_i], \text{such that} \sum_{k=1}^{r_i} \theta_{ijk} = 1 \right\}.$$

- Assume a conjugate prior over Θ_i given G

$$\Theta_i \mid PA_i \sim \text{Product-Dirichlet}((\alpha_{ijk})_{q_i \times r_i}) \Leftrightarrow$$
$$\theta_{ij} = (\theta_{ij1}, \ldots, \theta_{ijr_i}) \mid PA_i \sim_{ind} \text{Dirichlet}(\alpha_{ij1}, \ldots, \alpha_{ijr_i}).$$

  Choose $\alpha_{ijk} = \alpha/(r_i \cdot q_i)$.

- Assume a prior over G: $P(G) \propto \lambda^{d(G)}$, $\lambda \in (0,1)$ and $d(G) = \sum_{i=1}^{p} r_i q_i$ number of parameters.

Structure learning of DAGs

Given $(G, \mathbf{D})$, how to compute the BD score: $(PA_i \equiv PA_i^G)$

- Contingency tables: $N_{ijk} = \#\{PA_i = j \,\&\, X_i = k\}$ in $\mathbf{D}$. For each node, a $q_i \times r_i$ table: $N_i = \{N_{ijk} : j \in [q_i], k \in [r_i]\}$.
- Marginal likelihood of N_{ij} (one row) given PA_i:

$$P(N_{ij} \mid PA_i) = \int P(N_{ij} \mid \theta_{ij}) \pi(\theta_{ij} \mid PA_i) d\theta_{ij}$$
$$= \frac{\Gamma(\alpha/q_i)}{\Gamma(N_{ij\bullet} + \alpha/q_i)} \prod_{k=1}^{r_i} \frac{\Gamma(N_{ijk} + \alpha/(q_i r_i))}{\Gamma(\alpha/(q_i r_i))},$$

  where $N_{ij\bullet} = \sum_k N_{ijk}$ (row sum).
- Marginal likelihood of N_i (the whole table):

$$P(N_i \mid PA_i) = \prod_{j=1}^{q_i} P(N_{ij} \mid PA_i).$$

Structure learning of DAGs

- Marginal likelihood of $\mathbf{D}$ (all p tables, one for each node):

$$P(\mathbf{D} \mid G) = \prod_{i=1}^{p} P(N_i \mid PA_i).$$

  Posterior distribution

$$\begin{aligned} P(G \mid \mathbf{D}) &\propto P(G)P(\mathbf{D} \mid G) \\ &= \prod_{i=1}^{p} \lambda^{q_i r_i} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha/q_i)}{\Gamma(N_{ij\bullet} + \alpha/q_i)} \prod_{k=1}^{r_i} \frac{\Gamma(N_{ijk} + \alpha/(q_i r_i))}{\Gamma(\alpha/(q_i r_i))}. \end{aligned}$$

- BD score is decomposable:

$$S_{BD}(G, \mathbf{D}) := \log P(G) + \log P(\mathbf{D} \mid G) = \sum_{i=1}^{p} s(N_i, PA_i). \tag{18}$$

Structure learning of DAGs

Properties of the scoring functions (17) and (18):

- Score-equivalent: For any two Markov equivalent DAGs G_1 and G_2, we have $S(G_1, \mathbf{D}) = S(G_2, \mathbf{D})$.
- Consistent (Chickering 2002): A scoring function $S(G, \bullet)$ is *consistent* if the following two properties hold for $\mathbf{D}_n \sim_{iid} \mathbb{P}$:
  1. If $\mathbb{P} \in G \setminus H$, then $\lim_n \mathbb{P}\{S(G, \mathbf{D}_n) > S(H, \mathbf{D}_n)\} = 1$.
  2. If $\mathbb{P} \in G \cap H$ and $d(G) < d(H)$, i.e. G has fewer parameters, then $\lim_n \mathbb{P}\{S(G, \mathbf{D}_n) > S(H, \mathbf{D}_n)\} = 1$.

Haughton (1988) established:

1. $S_{\text{BIC}}(G, \bullet)$ (16) is consistent for exponential family.
2. $S_{BD}(G, \mathbf{D}_n) = S_{\text{BIC}}(G, \mathbf{D}_n) + O_p(1) = O_p(n) + O_p(1)$.

Thus, both (17) and (18) are consistent scoring functions.

Structure learning of DAGs

Consistency of score-based learning:

Theorem 4

Suppose $\mathbb{P}$ is faithful to $\mathcal{G}$ and $\mathbf{D}_n \sim_{iid} \mathbb{P}$. If $S(G, \bullet)$ is consistent and score-equivalent, then

$$\lim_{n\to\infty} \mathbb{P}\left\{\underset{G}{\operatorname{argmax}}\, S(G, \mathbf{D}_n) = \mathcal{C}\right\} = 1,$$

where $\mathcal{C} := \{G : G \simeq \mathcal{G}\}$ is the Markov equivalence class of $\mathcal{G}$.

Structure learning of DAGs

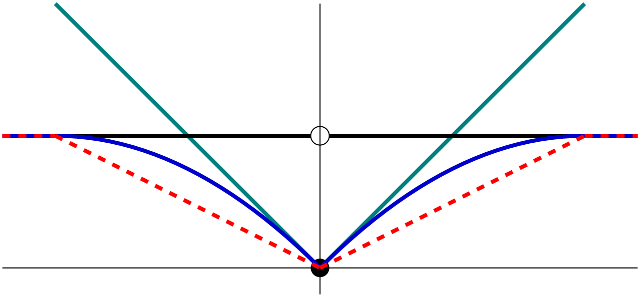
Continuous relaxation of the scoring function:

- Consider Gaussian DAGs for simplicity. The BIC score $S_{BIC}(G, \mathbf{D})$ (17) is over a discrete space and hard to optimize.
- $B = (\beta_{ij}) = [\beta_1 \mid \cdots \mid \beta_p]$ and $\Omega = \text{diag}(\omega_j^2)$.
  Maximum regularized likelihood:

$$(\widehat{B}, \widehat{\Omega}) = \underset{B \in \mathcal{B}, \Omega}{\text{argmax}} \sum_{j=1}^{p} \log p(X_j \mid X\beta_j, \omega_j^2) - \lambda_n \rho(\beta_j). \qquad (19)$$

1. $\mathcal{B}$: weighted adjacency matrices of DAGs, so that $PA_j = \text{supp}(\beta_j)$ and $\text{supp}(B)$ defines a DAG G.
2. $\rho(\beta_j) = \sum_i \rho(|\beta_{ij}|)$: continuous function, e.g. ℓ_1 or concave (Fu and Zhou 2013; Aragam and Zhou 2015).
3. Apply continuous function optimization, such as block-wise coordinate descent.

Compare regularizers: ℓ_1, concave, and ℓ_0.



Black: ℓ_0 penalty; Teal: ℓ_1 penalty; Blue: MCP; Red, dashed: Capped-ℓ_1 penalty.

Structure learning of DAGs

Score-based learning with experimental data:

- If X_i is under intervention, i.e. $do(X_i = x^*)$: delete edges $X_k \to X_i$ for all $k \in PA_i$.
- Let $\mathcal{O}_i$ be the row indices of the data matrix $\mathbf{D}$ for which node X_i is *not* under intervention (i.e. observational). Replace $p(X_i \mid PA_i)$ by $p(X_{\mathcal{O}_i i} \mid PA_{\mathcal{O}_i i})$.
  1. Gaussian data: log-likelihood in (17) and (19) replaced by

$$\ell(B, \Omega; \mathbf{D}) = \sum_{j=1}^{p} \log p(X_{\mathcal{O}_j j} \mid X_{\mathcal{O}_j} \beta_j, \omega_j^2). \tag{20}$$

  2. Multinomial data: Replace N_{ijk} by

$$N_{ijk}(\mathcal{O}_i) = \#\{rows \in \mathcal{O}_i : PA_i = j \,\&\, X_i = k\}.$$

Structure learning of DAGs

Identifiability of causal DAGs:

Assumptions:

(A1) The true parameter Θ^* is faithful to $\mathcal{G}$.

(A2) The parameter for $[X_j \mid PA_j]$ is identifiable.

(A3) Each node X_j is under intervention for $n_j \gg \sqrt{n}$ data points.

Theorem 5 (Gu et al. (2019))

Assume (A1), (A2) and (A3). Denote by $\ell(\Theta; \mathbf{D}_n)$ the log-likelihood of the data $\mathbf{D}_n$. For any $\Theta \neq \Theta^*$,

$$\lim_{n\to\infty} \mathbb{P}\{\ell(\Theta^*; \mathbf{D}_n) > \ell(\Theta; \mathbf{D}_n)\} = 1.$$

1. Gaussian data, $\ell(\Theta; \mathbf{D}_n) = (20)$.
2. Discrete data, $\ell(\Theta; \mathbf{D}_n) = \sum_{i=1}^{p} \sum_{j,k} N_{ijk}(\mathcal{O}_i) \log \theta_{ijk}$.

References I

Bryon Aragam and Qing Zhou. Concave penalized estimation of sparse Gaussian Bayesian networks. *Journal of Machine Learning Research*, 16:2273–2328, 2015.

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

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

Fei Fu and Qing Zhou. Learning sparse causal Gaussian networks with experimental intervention: Regularization and coordinate descent. *Journal of the American Statistical Association*, 108 (501):288–300, 2013.

References II

Jiaying Gu, Fei Fu, and Qing Zhou. Penalized estimation of directed acyclic graphs from discrete data. *Statistics and Computing*, 29:161–176, 2019.

Dominique M.A. Haughton. On the choice of a model to fit data from an exponential family. *Annals of Statistics*, 16:342–355, 1988.

David Heckerman, Dan Geiger, and David M Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243, 1995.

Christopher Meek. Causal inference and causal explanation with background knowledge. *Uncertainty in Artificial Intelligence*, 11: 403–410, 1995.

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

References III

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

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Springer, 1993.

Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1): 62–72, 1991.