Directed Acyclic Graphs

Statistics · Module 11 — Causality

Node 11.2

Position in Knowledge Graph

System: Statistics. Structural Domain: Causality (Module 11). Node 11.2.

Directed acyclic graphs (DAGs) provide the graphical language for encoding causal assumptions, deriving conditional independence relations, and defining interventions. Where the potential outcomes framework (Node 11.1) specifies the estimands, DAGs encode the structural relationships that make identification possible. The d-separation algorithm reads conditional independences from the graph; the do-operator formalizes interventions via graph surgery. This node enables the identifiability and backdoor criterion (Node 11.3).

Formal Definition

A **directed acyclic graph** $\mathcal{G} = (V, E)$ consists of: - A vertex set $V = \{X_1, \ldots, X_p\}$ representing random variables. - A set E of directed edges $X_i \to X_j$, with no directed cycles.

Parents: $\mathrm{pa}(X_j) = \{X_i : X_i \to X_j \in E\}$. **Children:** $\mathrm{ch}(X_j) = \{X_k : X_j \to X_k \in E\}$. **Descendants:** $\mathrm{desc}(X_j)$ = all variables reachable from X_j via directed paths. **Ancestors:** $\mathrm{an}(X_j)$ = all variables from which X_j is reachable via directed paths. **Non-descendants:** $\mathrm{nd}(X_j) = V \setminus (\{X_j\} \cup \mathrm{desc}(X_j))$.

Intuition

A DAG is a map of causal relationships: an arrow $X \to Y$ means X is a direct cause of Y. The absence of an arrow means no direct causal relationship. The graph structure determines which statistical associations are causal, which are confounded, and which are spurious. The power of the DAG framework is that complex causal reasoning reduces to graphical algorithms (d-separation), and interventions reduce to graph surgery (removing arrows).

Structural Role

Dependencies: Builds on conditional independence (Node 1.7) and joint distributions (Node 2.1). **Enables:** Identifiability and the backdoor criterion (Node 11.3), which uses d-separation to determine when causal effects can be estimated from observational data.

DAGs separate the statistical model (the factorization of p) from the causal model (the structural equations and interventions). Statistical dependence is readable from d-separation; causal effects require the interventional distribution, which depends on the graph structure.

Mathematical Development

Markov Factorization

A distribution P is **Markov** with respect to $\mathcal{G}$ if

$$p(x_1, \dots, x_p) = \prod_{j=1}^{p} p(x_j \mid \mathrm{pa}(x_j)).$$

Local Markov Property. Each variable is conditionally independent of its non-descendants given its parents:

$$X_j \perp\!\!\!\perp \mathrm{nd}(X_j) \mid \mathrm{pa}(X_j).$$

The local Markov property is equivalent to the factorization for positive distributions.

d-Separation: The Algorithm

For disjoint sets $A, B, C \subseteq V$, determine whether A and B are d-separated by C:

1. Enumerate all paths between any node in A and any node in B (paths may traverse edges in either direction).
2. A path is **blocked** by C if it contains at least one of:
   - A **chain** $X_i \to X_k \to X_j$ with $X_k \in C$.
   - A **fork** $X_i \leftarrow X_k \to X_j$ with $X_k \in C$.
   - A **collider** $X_i \to X_k \leftarrow X_j$ with $X_k \notin C$ and no descendant of X_k in C.
3. A and B are d-separated by C if **every** path between A and B is blocked.

Write $A \perp_{\mathcal{G}} B \mid C$ for d-separation.

Key subtlety: Conditioning on a collider (or its descendant) **opens** the path, creating a spurious association. This is the graphical basis of "collider bias" or "Berkson's paradox."

Global Markov Property

Theorem. If P is Markov with respect to $\mathcal{G}$, then

$$A \perp_{\mathcal{G}} B \mid C \implies A \perp\!\!\!\perp B \mid C \quad \text{under } P.$$

d-separation in the graph implies conditional independence in the distribution.

Faithfulness

A distribution P is **faithful** to $\mathcal{G}$ if the converse holds:

$$A \perp\!\!\!\perp B \mid C \quad \text{under } P \implies A \perp_{\mathcal{G}} B \mid C.$$

Faithfulness means the distribution has no "accidental" conditional independences beyond those implied by the graph. It fails on a Lebesgue-measure-zero set of parameters, so it is generically true but can fail in special cases (e.g., exact parameter cancellations).

Under both the Markov property and faithfulness, the conditional independence structure of P is exactly the set of d-separations in $\mathcal{G}$.

Causal DAGs and Structural Equations

A **causal DAG** augments the graphical structure with **structural equations** (also called structural causal models, SCMs):

$$X_j = f_j(\mathrm{pa}(X_j), U_j), \quad j = 1, \dots, p,$$

where $U_1, \dots, U_p$ are mutually independent exogenous noise variables. The structural equations define a **data-generating mechanism**: they specify not just the conditional distribution of X_j given its parents, but the functional relationship that would persist under interventions.

The joint distribution implied by the SCM is Markov with respect to $\mathcal{G}$ (the exogenous noise independence implies the factorization).

The do-Operator

The **intervention** $do(X_j = t)$ is defined by:

1. Replace the structural equation for X_j with $X_j = t$ (deterministic).
2. Leave all other structural equations unchanged.

Graphically: remove all edges into X_j (the **mutilated graph** $\mathcal{G}_{\overline{X_j}}$) and set $X_j = t$.

The **interventional distribution** is

$$p(x_1, \dots, x_p \mid do(X_j = t)) = \prod_{k \neq j} p(x_k \mid \mathrm{pa}(x_k))\Big|_{x_j = t}.$$

This generally differs from the observational conditional $p(x_1, \dots, x_p \mid X_j = t)$, because conditioning does not remove the causal influence of X_j's parents.

Observational vs. Interventional Distributions

Example. Consider $U \to T \to Y$ and $U \to Y$ (confounding).

Observational: $p(Y \mid T = t) = \int p(Y \mid T = t, U = u)\, p(U \mid T = t)\, du$. The confounder U has a different distribution for different values of T.

Interventional: $p(Y \mid do(T = t)) = \int p(Y \mid T = t, U = u)\, p(U)\, du$. Under intervention, U retains its marginal distribution because the arrow $U \to T$ is severed.

The difference between these two expressions is the essence of confounding.

Markov Equivalence

Two DAGs are **Markov equivalent** if they imply the same set of conditional independences (same d-separations). Markov equivalent DAGs have: - The same skeleton (undirected edges). - The same **v-structures** (immoralities): patterns $X_i \to X_k \leftarrow X_j$ where X_i and X_j are not adjacent.

The set of all DAGs Markov equivalent to $\mathcal{G}$ is the **Markov equivalence class**, represented by a **completed partially directed acyclic graph** (CPDAG). Observational data alone can identify the Markov equivalence class but not the DAG itself (without additional assumptions or interventional data).

Worked Examples

Example 1: d-Separation in a Confounding Structure

DAG: $X \leftarrow U \rightarrow Y$, $X \rightarrow Y$. Variables: confounder U, treatment X, outcome Y.

Is $X \perp_{\mathcal{G}} Y \mid \emptyset$**?** Path 1: $X \rightarrow Y$ (direct, no collider, not blocked). Answer: No.

Is $X \perp_{\mathcal{G}} Y \mid U$**?** Path 1: $X \rightarrow Y$ (chain with no middle node in $C = \{U\}$, so not blocked). Answer: No. Conditioning on U blocks the backdoor path $X \leftarrow U \rightarrow Y$ but does not block the direct path $X \rightarrow Y$.

Example 2: Collider Bias

DAG: $X \rightarrow C \leftarrow Y$. No direct path between X and Y.

Is $X \perp_{\mathcal{G}} Y \mid \emptyset$**?** The only path is $X \rightarrow C \leftarrow Y$, which contains a collider at C. Since $C \notin \emptyset$, the path is blocked. Answer: Yes, $X \perp_{\mathcal{G}} Y$.

Is $X \perp_{\mathcal{G}} Y \mid C$**?** Now $C \in \{C\}$, so the collider is conditioned on, which **opens** the path. Answer: No, $X \not\perp_{\mathcal{G}} Y \mid C$.

This is Berkson's paradox: conditioning on a common effect creates a spurious association between its causes. For instance, if X = talent and Y = luck, and C = success (which requires one or both), then among successful people, talent and luck appear negatively correlated.

Cross-Links

AI. DAGs are the foundation of Bayesian networks. The d-separation algorithm underlies message-passing (belief propagation) for efficient inference in graphical models. Causal discovery algorithms (PC, GES, FCI) learn DAG structure from data using conditional independence tests.

Economics. Structural equation models in econometrics are causal DAGs with specific functional forms (typically linear). The distinction between structural and reduced-form equations corresponds to the distinction between causal and observational distributions.

Linear Algebra. The adjacency matrix A of a DAG (with $A_{ij} = 1$ if $X_i \rightarrow X_j$) is strictly upper-triangular (after topological sorting), hence nilpotent. The reachability matrix $(I - A)^{-1} = \sum_{k=0}^{p-1} A^k$ encodes all ancestor-descendant relationships.

Quantum Mechanics. Quantum causal models extend DAGs to quantum channels, where the edges represent quantum operations rather than classical conditional distributions. The no-signaling principle in quantum mechanics is the analogue of d-separation: spacelike-separated parties are d-separated by their common past.

Structural Compression

Invariant

d-separation in $\mathcal{G}$ implies conditional independence in any distribution Markov with respect to $\mathcal{G}$; under faithfulness, the implication is an equivalence. The interventional distribution is the observational distribution of the mutilated graph.

Mechanism

The Markov factorization decomposes the joint distribution into local conditional distributions, one per variable. Interventions sever parent-child links by replacing a structural equation with a constant, altering the factorization and producing the interventional distribution. d-separation is an algorithmic procedure for reading off conditional independences from the factorization structure.

Cross-System Echo

DAGs are the graphical language of Bayesian networks in AI; d-separation underlies message-passing algorithms (belief propagation) for inference in graphical models. In econometrics, the DAG framework formalizes the concept of exogeneity and the conditions under which regression estimates have causal interpretations. In quantum mechanics, causal structure is formalized by quantum causal models extending DAGs to quantum channels.

Exercises

1. For the DAG $X_1 \to X_3 \leftarrow X_2$, $X_3 \to X_4$, determine all conditional independence relations implied by d-separation (i.e., enumerate all valid $A \perp_{\mathcal{G}} B \mid C$ statements).
2. Show that the Markov factorization $p(x_1, \ldots, x_p) = \prod_j p(x_j \mid \mathrm{pa}(x_j))$ implies the local Markov property $X_j \perp\!\!\!\perp \mathrm{nd}(X_j) \mid \mathrm{pa}(X_j)$.
3. Construct two DAGs on three variables that are Markov equivalent (same skeleton and v-structures) and one that is not. Identify the distinguishing conditional independence.
4. For the DAG $Z \to T \to Y$, $Z \to Y$, compute the interventional distribution $p(Y \mid do(T = t))$ using the truncated factorization and show it differs from $p(Y \mid T = t)$.
5. Prove that conditioning on a collider opens the path: for $X \to C \leftarrow Y$ with $X \perp\!\!\!\perp Y$ marginally, show that $X \not\perp\!\!\!\perp Y \mid C$ when the structural equations are linear with Gaussian noise.
6. Show that the adjacency matrix of a DAG is nilpotent (i.e., $A^p = 0$), and use this to prove that the factorization $p(x) = \prod_j p(x_j \mid \mathrm{pa}(x_j))$ defines a valid joint distribution (no circular dependencies).
7. Argue, structurally, why observational data alone can identify the Markov equivalence class but not the specific DAG, and explain what additional information (interventions, temporal ordering, or functional form assumptions) is needed to distinguish between Markov-equivalent DAGs.

Statistics · Module 11 — Causality · Node 11.2
Concepts: graph-theory, conditional-probability
Prerequisites: 1.7, 2.1
Enables: 11.3

hbar.university — own.your.cognition