

Seminar in Directed Graphical Models and Causality

1. Conditional Independence and Directed Acyclic Graphs

Grigor Keropyan

Technical University of Munich

August 12, 2021

Outline

1. (Conditional) Independence
2. Properties of Conditional Independence
3. Directed Acyclic Graphs (DAG)
4. Markov Properties for DAGs
5. Factorization according to DAG
6. Exercises

Densities

- ▶ Let $X \in \mathbf{R}^m$ and $Y \in \mathbf{R}^n$ be random vectors, where $m, n \in \mathbf{N}$.
- ▶ Assume $f(x, y)$ is the joint density function of (X, Y) with respect to product measure $\lambda = \lambda_X \otimes \lambda_Y$, where λ_X and λ_Y are measures in \mathbf{R}^m and \mathbf{R}^n , respectively.
- ▶ Marginal distributions P^X and P^Y have densities

$$f_X(x) = \int f(x, y) d\lambda_Y(y) \quad \text{and} \quad f_Y(y) = \int f(x, y) d\lambda_X(x)$$

Conditional Densities/Distributions

Definition

Conditional density of X given $Y = y$ is

$$f_{X|Y}(x|y) = \begin{cases} \frac{f(x,y)}{f_Y(y)} & \text{if } f_Y(y) > 0 \\ \text{any density } f_0(x) & \text{otherwise} \end{cases}$$

The **conditional distribution** of X given $Y = y$ is

$$P^{X|Y=y}(A) \equiv P(X \in A|Y = y) := \int_A f(x|y) d\lambda_X(x) \quad \forall A \in \mathbf{R}^m \text{ Borel set.}$$

Independence

- ▶ X and Y are called **independent** if

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

$\forall A, B$ Borel sets and write $X \perp\!\!\!\perp Y$.

- ▶ The following is a characterization of independence

$$X \perp\!\!\!\perp Y \iff f(x, y) = f_X(x)f_Y(y) \quad [\lambda - a.e.]$$

Conditional independence

- ▶ Let Z be random vector in \mathbf{R}^k , where $k \in \mathbf{N}$.

Definition

X and Y are called **conditionally independent** given Z if

$$P(X \in A, Y \in B | Z = z) = P(X \in A | Z = z)P(Y \in B | Z = z) \quad [P^Z - a.e]$$

$\forall A, B$ Borel sets and write $X \perp\!\!\!\perp Y | Z$.

Question 1

Given a joint density of random vector (X, Y, Z) as

$$f(x, y, z) = \frac{1}{C}(x - z)^4 x^2 y^6 (y - z)^8,$$

where constant C ensures that we have a valid density. Is the following relation true?

$$X \perp\!\!\!\perp Y|Z.$$

Properties of Conditional Independence

- ▶ Assume $f(x, y, z)$ is the joint density of (X, Y, Z) with respect to product measure $\lambda = \lambda_X \otimes \lambda_Y \otimes \lambda_Z$.

Lemma

The followings are equivalent and the equations hold $P^{X,Y,Z}$ -a.e.:

1. $X \perp\!\!\!\perp Y|Z$
2. $f(x, y|z) = f(x|z)f(y|z)$
3. $f(x|y, z) = f(x|z)$
4. $f(x, y, z) = \frac{f(x, z)f(y, z)}{f(z)}$
5. $f(x, y, z) = g(x, z)h(y, z)$ for some measurable functions g and h
6. $f(x|y, z) = g(x, z)$ for some measurable function g

Proof of Lemma

1 \iff 2: If $X \perp\!\!\!\perp Y|Z$, then

$$\begin{aligned}P(X \in A, Y \in B|Z = z) &= P(X \in A|Z = z)P(Y \in B|Z = z) \\&= \int_A f(x|z) d\lambda_X(x) \int_B f(y|z) d\lambda_Y(y) \\&= \int_{A \times B} f(x|z)f(y|z) d(\lambda_X \otimes \lambda_Y)(x, y)\end{aligned}$$

where A and B are arbitrary Borel sets. So, $f(x, y|z) = f(x|z)f(y|z)$ almost surely. If $f(x, y|z) = f(x|z)f(y|z)$ a.s., then

$$\begin{aligned}P(X \in A, Y \in B|Z = z) &= \int_{A \times B} f(x, y|z) d(\lambda_X \otimes \lambda_Y)(x, y) = \int_{A \times B} f(x|z)f(y|z) d(\lambda_X \otimes \lambda_Y)(x, y) \\&= \int_A f(x|z) d\lambda_X(x) \int_B f(y|z) d\lambda_Y(y) \\&= P(X \in A|Z = z)P(Y \in B|Z = z).\end{aligned}$$

Proof of Lemma

2 \iff 3:

$$\begin{aligned} f(x, y|z) = f(x|z)f(y|z) &\iff \frac{f(x, y, z)}{f(z)} = \frac{f(x, z)f(y, z)}{f(z)f(z)} \\ &\iff \frac{f(x, y, z)}{f(y, z)} = \frac{f(x, z)}{f(z)} \iff f(x|y, z) = f(x|z), \end{aligned}$$

where we are considering all the cases when the denominator is not zero and the equations hold almost surely. From the definition of conditional density the zero cases are trivial.

3 \iff 4:

$$f(x|y, z) = f(x|z) \iff \frac{f(x, y, z)}{f(y, z)} = \frac{f(x, z)}{f(z)} \iff f(x, y, z) = \frac{f(x, z)f(y, z)}{f(z)}.$$

Proof of Lemma

3 \implies 6: Denote $g(x, z) := f(x|z)$.

6 \implies 5: Denoting $h(y, z) := f(y, z)$ we have $f(x, y, z) = f(x|y, z)f(y, z) = g(x, z)h(y, z)$.

5 \implies 4: We have

$$\begin{aligned}\frac{f(x, z)f(y, z)}{f(z)} &= \frac{\int f(x, y, z) d\lambda_Y(y) \int f(x, y, z) d\lambda_X(x)}{\int f(x, y, z) d(\lambda_X \otimes \lambda_Y)(x, y)} \\ &= \frac{g(x, z)h(y, z) \int h(y, z) d\lambda_Y(y) \int g(x, z) d\lambda_X(x)}{\int g(x, z) d\lambda_X(x) \int h(y, z) d\lambda_Y(y)} \\ &= g(x, z)h(y, z) = f(x, y, z).\end{aligned}$$



Question 1 (now should be easy)

Given a joint density of random vector (X, Y, Z) as

$$f(x, y, z) = \frac{1}{C}(x - z)^4 x^2 y^6 (y - z)^8,$$

where constant C ensures that we have a valid density. Is the following relation true?

$$X \perp\!\!\!\perp Y|Z.$$

General Properties of Conditional Independence

(C1) "Symmetry":

$$X \perp\!\!\!\perp Y|Z \iff Y \perp\!\!\!\perp X|Z.$$

(C2) "Decomposition":

$$X \perp\!\!\!\perp Y|Z \implies h(X) \perp\!\!\!\perp Y|Z \text{ for any measurable function } h.$$

In particular, $(X, W) \perp\!\!\!\perp Y|Z \implies X \perp\!\!\!\perp Y|Z$.

(C3) "Weak union":

$$X \perp\!\!\!\perp Y|Z \implies X \perp\!\!\!\perp Y|(Z, h(X)) \text{ for any measurable function } h.$$

In particular, using also (C2) we obtain $(X, W) \perp\!\!\!\perp Y|Z \implies X \perp\!\!\!\perp Y|(Z, W)$.

(C4) "Contraction":

$$X \perp\!\!\!\perp Y|Z \text{ and } X \perp\!\!\!\perp W|(Y, Z) \iff X \perp\!\!\!\perp (W, Y)|Z.$$

Proof of (C1) and (C2)

(C1): For all Borel sets A, B and for all values of z we have

$$P(X \in A, Y \in B | Z = z) = P(X \in A | Z = z)P(Y \in B | Z = z)$$

(C2): For all Borel sets A, B and for all values of z we have

$$\begin{aligned} P(h(X) \in A, Y \in B | Z = z) &= P(X \in h^{-1}(A), Y \in B | Z = z) \\ &= P(X \in h^{-1}(A) | Z = z)P(Y \in B | Z = z) \\ &= P(h(X) \in A | Z = z)P(Y \in B | Z = z) \end{aligned}$$

So, $h(X) \perp\!\!\!\perp Y | Z$.

Proof of (C3) and (C4)

(C3): The proof is only for last equation when we have densities

$$f(x|y, z, w) = \frac{f(x, w|y, z)}{f(w|y, z)} = \frac{f(x, w|z)}{f(w|y, z)} = \frac{f(x, w|z)}{f(w|z)} = f(x|w, z)$$

So, $X \perp\!\!\!\perp Y|(Z, W)$.

(C4): If $X \perp\!\!\!\perp Y|Z$ and $X \perp\!\!\!\perp W|(Y, Z)$, then

$$P^{X|(W,Y,Z)=(w,y,z)} = P^{X|(Y,Z)=(y,z)} = P^{X|Z=z} \quad [P^{(W,Y,Z)} - a.s.]$$

So, $X \perp\!\!\!\perp (W, Y)|Z$. Now if $X \perp\!\!\!\perp (W, Y)|Z$ from (C3) we have $X \perp\!\!\!\perp W|Y, Z$ and from (C2) we have $X \perp\!\!\!\perp Y|Z$.

Intersection "Axiom"

From (C4) we have

$$X \perp\!\!\!\perp (W, Y)|Z \implies X \perp\!\!\!\perp W|(Y, Z) \text{ and } X \perp\!\!\!\perp (W, Y)|Z \implies X \perp\!\!\!\perp Y|(W, Z).$$

(C5) "Intersection":

Assume that we have a joint density $f(x, y, w, z)$ with respect to $\lambda = \lambda_X \otimes \lambda_Y \otimes \lambda_W \otimes \lambda_Z$ such that $f(y, w, z) > 0$ [$\lambda - a.e.$]. Then,

$$X \perp\!\!\!\perp (W, Y)|Z \iff X \perp\!\!\!\perp W|(Y, Z) \text{ and } X \perp\!\!\!\perp Y|(W, Z)$$

Proof of (C5)

From previous slide we need only the reverse implication \Leftarrow . From the Lemma we have

$$f(x, y, w, z) = \frac{f(x, w, z)f(y, w, z)}{f(w, z)} = \frac{f(x, y, z)f(y, w, z)}{f(y, z)}$$

Since $f(y, w, z) > 0$ almost surely we have

$$\frac{f(x, w, z)}{f(w, z)} = \frac{f(x, y, z)}{f(y, z)} \implies f(x, w, z)f(y, z) = f(x, y, z)f(w, z).$$

From the marginalization we have

$$f(x, w, z)f(z) = f(x, w, z) \int f(y, z) d\lambda_Y(y) = \int f(x, y, z)f(w, z) d\lambda_Y(y) = f(x, z)f(w, z).$$

So, from the Lemma we have $X \perp\!\!\!\perp W|Z$. Using (C4) with $X \perp\!\!\!\perp Y|(W, Z)$ we obtain $X \perp\!\!\!\perp (W, Y)|Z$. □

Terminology and Notation for DAGs

Definition

A graph $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ consists of a finite set of nodes \mathbf{V} and edges $\mathcal{E} \subseteq \mathbf{V} \times \mathbf{V}$ of ordered pairs of distinct nodes.

- ▶ Given a set of random variables $\mathbf{X} = (X_1, \dots, X_p)$, $\mathbf{V} := \{1, \dots, p\}$ and a graph $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ we associate every random variable X_j with node $j \in \mathbf{V}$.
- ▶ The joint distribution of \mathbf{X} is denoted by $P^{\mathbf{X}}$ and marginal distribution of \mathbf{X}_j by $P^{\mathbf{X}_j}$.
- ▶ A graph $\mathcal{G}_1 = (\mathbf{V}_1, \mathcal{E}_1)$ is called a **subgraph** of \mathcal{G} if $\mathbf{V}_1 \subseteq \mathbf{V}$ and $\mathcal{E}_1 \subseteq \mathcal{E}$.
- ▶ If \mathcal{G}_1 is a subgraph of \mathcal{G} we write $\mathcal{G}_1 \leq \mathcal{G}$ and if $\mathcal{E}_1 \neq \mathcal{E}$ we say \mathcal{G}_1 is **proper subgraph** of \mathcal{G} .

Example 1

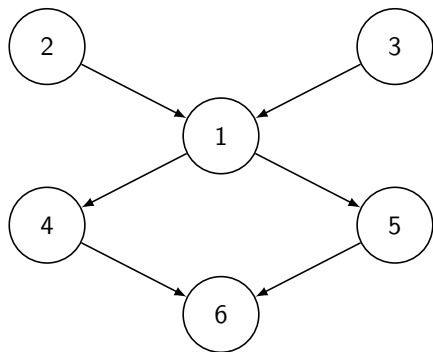


Figure 1: Graph $\mathcal{G} = (\{1, 2, 3, 4, 5, 6\}, \{(2, 1), (3, 1), (1, 4), (1, 5), (4, 6), (5, 6)\})$.

Terminology and Notation for DAGs

- ▶ A node i is called a child of j if $(j, i) \in \mathcal{E}$ and is called a parent, if $(i, j) \in \mathcal{E}$.
- ▶ If $(i, j) \in \mathcal{E}$ we also write $i \rightarrow j$.
- ▶ Children of j is denoted by $\mathbf{CH}_j^{\mathcal{G}} := \{i \in \mathbf{V} : (j, i) \in \mathcal{E}\}$ and parents of j by $\mathbf{PA}_j^{\mathcal{G}} := \{i \in \mathbf{V} : (i, j) \in \mathcal{E}\}$.
- ▶ Two nodes i and j are called **adjacent** if $(j, i) \in \mathcal{E}$ or $(i, j) \in \mathcal{E}$ and if both holds we say the edge between i and j is **undirected**, otherwise **directed**.
- ▶ A graph is called **complete** if every two nodes are adjacent. **Cliques** of a graph \mathcal{G} are the maximal complete subgraphs of \mathcal{G} (here maximal in a sense of set inclusion).
- ▶ A **path** in \mathcal{G} is a sequence of distinct nodes j_1, \dots, j_n such that j_k and j_{k+1} are adjacent $\forall k = 1, \dots, n-1$ and $n \geq 2$. If $j_k \rightarrow j_{k+1} \forall k = 1, \dots, n-1$ path is called **directed** from j_1 to j_n .

Example 2

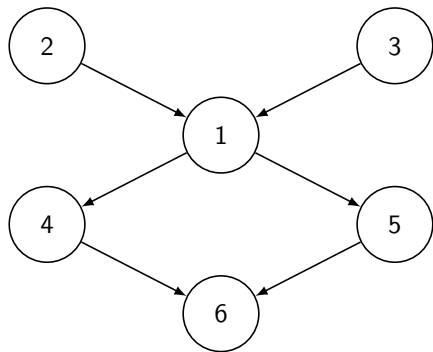


Figure 2: Graph $\mathcal{G} = (\{1, 2, 3, 4, 5, 6\}, \{(2, 1), (3, 1), (1, 4), (1, 5), (4, 6), (5, 6)\})$.

Some (directed) paths are

$$1 \rightarrow 4 \rightarrow 6 \leftarrow 5, \quad 3 \rightarrow 1 \rightarrow 4 \rightarrow 6, \quad 5 \rightarrow 6$$

Terminology and Notation for DAGs

- ▶ We say j is a **descendant** of i if there is a directed path from i to j and denote all the descendants of j by $\mathbf{DE}_j^{\mathcal{G}}$ and all non-descendants by $\mathbf{ND}_j^{\mathcal{G}}$. Note that descendants and non-descendants do not contain the node.
- ▶ j_k is called a **collider** in the path if $j_{k-1} \rightarrow j_k$ and $j_{k+1} \rightarrow j_k$.
- ▶ \mathcal{G} is called a **Partially Directed Acyclic Graph (PDAG)** if there is no directed cycle, i.e., if there is no pair (i, j) such that there are directed paths from i to j and from j to i .
- ▶ \mathcal{G} is called **Directed Acyclic Graph (DAG)** if all edges are directed and there is no cycle in \mathcal{G} .

Terminology and Notation for DAGs

- ▶ Three nodes i, j, k are called **immorality** or **v-structure** if one of them, say j is a child of the others and these parents are not adjacent: $i \rightarrow j, k \rightarrow j$ and $(k, i) \notin \mathcal{E}, (i, k) \notin \mathcal{E}$.
- ▶ The **skeleton** of graph \mathcal{G} is the set of all edges without taking the direction into account, that is all (i, j) such that $i \rightarrow j$ or $j \rightarrow i$.

Example 3

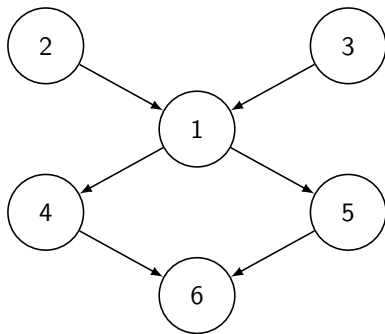


Figure 3: Graph $\mathcal{G} = (\{1, 2, 3, 4, 5, 6\}, \{(2, 1), (3, 1), (1, 4), (1, 5), (4, 6), (5, 6)\})$.

- ▶ Descendants of node 1 are $\{4, 5, 6\}$.
- ▶ 6 is a collider in the path $1 \rightarrow 4 \rightarrow 6 \leftarrow 5$
- ▶ 4, 5, 6 is a v-structure

Local Markov Property

Definition

The joint distribution $P^{\mathbf{X}}$ of \mathbf{X} is said to be **Local Markov with respect to the DAG \mathcal{G}** if

$$\forall v \in \mathbf{V}: \quad v \perp\!\!\!\perp \mathbf{V} \setminus \{\{v\} \cup \mathbf{PA}_v^{\mathcal{G}} \cup \mathbf{DE}_v^{\mathcal{G}}\} \mid \mathbf{PA}_v^{\mathcal{G}}.$$

Example 4

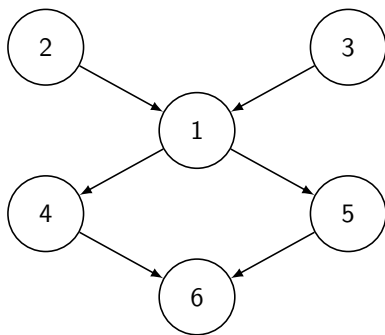


Figure 4: Graph $\mathcal{G} = (\{1, 2, 3, 4, 5, 6\}, \{(2, 1), (3, 1), (1, 4), (1, 5), (4, 6), (5, 6)\})$.

- From Local Markov Property we have $\{5\} \perp\!\!\!\perp \{2, 3, 4\} \mid \{1\}$.

Definition

In a DAG $\mathcal{G} = (\mathbf{V}, \mathcal{E})$, a path between i and j is **blocked** by $\mathbf{S} \subsetneq \mathbf{V}$ ($i, j \notin \mathbf{S}$) whenever there is a node k in the path and one of the following holds:

1. $k \in \mathbf{S}$ and k is not a collider in the path, or
2. $k \notin \mathbf{S}$ and k is a collider in the path and $\forall I \in \mathbf{DE}_k^{\mathcal{G}} \implies I \notin \mathbf{S}$.

Definition

Given disjoint subsets $\mathbf{A}, \mathbf{B}, \mathbf{C}$, we say \mathbf{A} and \mathbf{B} are **d-separated** by \mathbf{C} if every path between nodes in \mathbf{A} and \mathbf{B} is blocked by \mathbf{C} .

Example 5

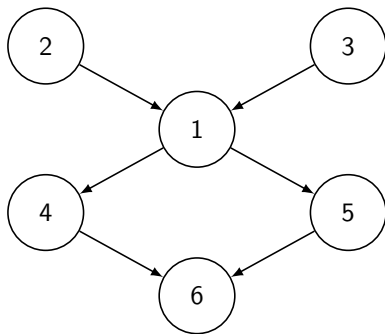


Figure 5: Graph $\mathcal{G} = (\{1, 2, 3, 4, 5, 6\}, \{(2, 1), (3, 1), (1, 4), (1, 5), (4, 6), (5, 6)\})$.

- Are $\{2\}$ and $\{4, 6\}$ d-separated by $\{1\}$?
- Are $\{2\}$ and $\{3\}$ d-separated by $\{1\}$?

Markov Property and Faithfulness

Definition

The joint distribution $\mathcal{L}(\mathbf{X})$ of \mathbf{X} is said to be **(Global) Markov with respect to the DAG \mathcal{G}** if

$$\mathbf{A}, \mathbf{B} \text{ d-sep. by } \mathbf{C} \implies \mathbf{A} \perp\!\!\!\perp \mathbf{B} | \mathbf{C}.$$

for all disjoint sets $\mathbf{A}, \mathbf{B}, \mathbf{C} \subseteq \mathbf{V}$.

Definition

The joint distribution $\mathcal{L}(\mathbf{X})$ is said to be **faithful to the DAG \mathcal{G}** if

$$\mathbf{A}, \mathbf{B} \text{ d-sep. by } \mathbf{C} \iff \mathbf{A} \perp\!\!\!\perp \mathbf{B} | \mathbf{C}.$$

for all disjoint sets $\mathbf{A}, \mathbf{B}, \mathbf{C} \subseteq \mathbf{V}$.

Example 6

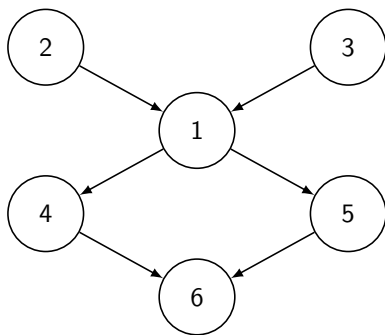


Figure 6: Graph $\mathcal{G} = (\{1, 2, 3, 4, 5, 6\}, \{(2, 1), (3, 1), (1, 4), (1, 5), (4, 6), (5, 6)\})$.

- From Global Markov Property we have $\{2, 3\} \perp\!\!\!\perp \{4, 5, 6\} \mid \{1\}$.

Factorization

Definition

Let \mathcal{G} be a DAG. The joint distribution $P^{\mathbf{x}}$ factorizes according to \mathcal{G} if the joint density has the following form

$$f(x) = \prod_{v \in \mathbf{V}} f(x_v | x_{\mathbf{PA}_v^{\mathcal{G}}})$$

Example 7

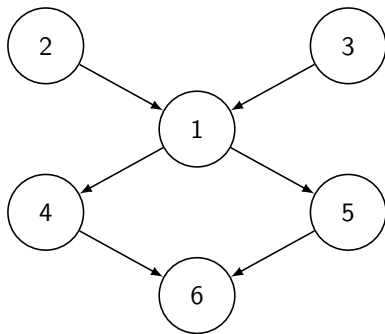


Figure 7: Graph $\mathcal{G} = (\{1, 2, 3, 4, 5, 6\}, \{(2, 1), (3, 1), (1, 4), (1, 5), (4, 6), (5, 6)\})$.

- If the joint distribution f factorizes according to the above graph then

$$f(x) = f(x_2)f(x_3)f(x_1|x_2, x_3)f(x_4|x_1)f(x_5|x_1)f(x_6|x_4, x_5).$$

Exercises

1. Given $(X, W) \perp\!\!\!\perp (Y, T) | Z$. Show that $X \perp\!\!\!\perp Y | (Z, W, T)$.
2. In Figure 7 write all possible Local Markov Properties with respect to the graph.
3. In Figure 7 write two Global Markov Properties with respect to the graph which are not Local Markov Property with respect to the graph.
4. Let \mathcal{G} be a DAG and A, B any non adjacent nodes. Prove that there is a set of nodes \mathbf{S} such that A and B are d-separated given \mathbf{S} .
5. Given a DAG $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ and any non adjacent nodes L and W in \mathbf{V} . Then, for any set of nodes \mathbf{R} in \mathbf{V} such that $\mathbf{R} \subset \mathbf{ND}_W^{\mathcal{G}}$

L, W d-sep. by $\mathbf{S} \cup \mathbf{R}$,

where $\mathbf{S} := \mathbf{PA}_L^{\mathcal{G}} \cup \mathbf{PA}_W^{\mathcal{G}}$.

References

Some of the statements and proofs I have taken from Prof. Dr. Mathias Drton lecture in "Graphical Models in Statistics" at TUM.