Conditional independence and undirected graphical models

Kaie Kubjas, 17.03.2023
Agenda

• Conditional independence
• Different ways to associate a statistical model with an undirected graph
  • Using conditional independence statements
  • Using factorizations of functions
• The relations between the different definitions
Graphical models example

* genes A, B, C
* Relationships
  - A regulates C
  - B regulates C

BIOLOGY

GRAPH

PROBABILISTIC MODEL

$P(A,B,C) = P(A)P(B)P(C|A,B)$

Genes ↔ Vertices ↔ Random variables
Relationships ↔ Edges ↔ Statistical dependencies
Correlation vs causation

- Genes regulated as $X \rightarrow Y \rightarrow Z$

- $X$ and $Z$ are correlated, but do not interact directly
Examples

- Gene association network
- Stock exchange
- Markov chains
- Hidden Markov models: DNA sequence alignment
Nonnegative matrix factorizations

- The set of $m \times n$ probability matrices of nonnegative rank at most $r$ corresponds to the graphical model associated with the graph above, where $Z$ is a hidden variable.

- $X$ and $Y$ are observed variables taking values in $[m]$ and $[n]$ respectively.

- $Z$ is a hidden variable taking values in $[r]$. 
Conditional independence
## Soccer vs hair example

<table>
<thead>
<tr>
<th>soccer \ hair</th>
<th>bald</th>
<th>short</th>
<th>medium</th>
<th>lots</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 hours/wk</td>
<td>17</td>
<td>43</td>
<td>66</td>
<td>125</td>
</tr>
<tr>
<td>2 hours/wk</td>
<td>23</td>
<td>65</td>
<td>63</td>
<td>101</td>
</tr>
<tr>
<td>5 hours/wk</td>
<td>41</td>
<td>110</td>
<td>68</td>
<td>79</td>
</tr>
<tr>
<td>10 hours/wk</td>
<td>34</td>
<td>81</td>
<td>42</td>
<td>45</td>
</tr>
</tbody>
</table>

Is watching soccer independent from the length of the hair?
Soccer vs hair example

- Random variables $X$ and $Y$ with outcomes in $[m]$ and $[n]$

  Joint probabilities are recorded in the matrix $P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mn} \end{pmatrix}$.

- $X$ and $Y$ are independent if and only if $P = \begin{pmatrix} P(X = i) \\ \vdots \\ P(Y = j) \end{pmatrix}$, i.e., if and only if $P$ is of rank 1.
Soccer vs hair example

- This matrix is far from being rank one.

- If we slightly perturb the original matrix, we get

\[
\begin{pmatrix}
14 & 45 & 66 & 125 \\
23 & 65 & 62 & 100 \\
42 & 110 & 68 & 80 \\
31 & 80 & 44 & 45 \\
\end{pmatrix}
= \begin{pmatrix}
4 & 20 & 56 & 120 \\
3 & 15 & 42 & 90 \\
2 & 10 & 28 & 60 \\
1 & 5 & 14 & 30 \\
\end{pmatrix}
+ \begin{pmatrix}
10 & 25 & 10 & 5 \\
20 & 50 & 20 & 10 \\
40 & 100 & 40 & 20 \\
30 & 75 & 30 & 15 \\
\end{pmatrix}
\]

\[
= \begin{pmatrix}
4 \\
3 \\
2 \\
1 \\
\end{pmatrix}
[1 5 14 30] + \begin{pmatrix}
1 \\
2 \\
4 \\
3 \\
\end{pmatrix}
[10 25 10 5]
\]
Soccer vs hair example

\[
\begin{pmatrix}
14 & 45 & 66 & 125 \\
23 & 65 & 62 & 100 \\
42 & 110 & 68 & 80 \\
31 & 80 & 44 & 45
\end{pmatrix}
= \begin{pmatrix}
4 & 20 & 56 & 120 \\
3 & 15 & 42 & 90 \\
2 & 10 & 28 & 60 \\
1 & 5 & 14 & 30
\end{pmatrix}
+ \begin{pmatrix}
10 & 25 & 10 & 5 \\
20 & 50 & 20 & 10 \\
40 & 100 & 40 & 20 \\
30 & 75 & 30 & 15
\end{pmatrix}
\]

= \begin{pmatrix}
4 \\
3 \\
2 \\
1
\end{pmatrix}
\begin{pmatrix}
1 & 5 & 14 & 30
\end{pmatrix}
+ \begin{pmatrix}
2 \\
4 \\
3
\end{pmatrix}
\begin{pmatrix}
10 & 25 & 10 & 5
\end{pmatrix}

• This is a size-2 nonnegative factorization!

• There could be one summand for women and one summand for men.

• This means that watching soccer is independent of the length of the hair given gender.

• This is an example of a conditional independence statement.
Setup

- Random vector $X = (X_1, \ldots, X_m)$
  
  - $X$ takes values in a Cartesian product space $\mathcal{X} = \prod_{i=1}^{m} \mathcal{X}_i$

- Examples:
  
  - Weight $X_1$ and height $X_2$ of a person: $\mathcal{X} = \mathbb{R} \times \mathbb{R}$
  
  - Two results $X_1, X_2$ of rolls of a die: $\mathcal{X} = [6] \times [6]$
Setup

We assume that either:

- $X$ has density $f(x) = f(x_1, \ldots, x_n)$ that is continuous on $\mathcal{X}$, or

- $\mathcal{X}$ is a finite set and then $f(x)$ is the joint distribution $\mathbb{P}(X = x)$
Setup

• Given $A \subseteq [m] := \{1,2,\ldots,m\}$, let

  • $X_A = (X_a)_{a \in A}$

  • $\mathcal{X}_A = \prod_{a \in A} \mathcal{X}_a$

• Given a partition $A_1 \mid \cdots \mid A_k$ of $[m]$, let $f(x_{A_1}, \ldots, x_{A_k})$ denote $f$ with some variables grouped together
Def: Let $A \subseteq [m]$. The marginal density $f_A(x_A)$ of $X_A$ is obtained by integrating out $x_{[m] \setminus A}$

$$f_A(x_A) := \int_{x_{[m] \setminus A}} f(x_A, x_{[m] \setminus A}) dx_{[m] \setminus A}$$

for all $x_A$.

For discrete random variables, we replace the integral with sum:

$$f_A(x_A) := \sum_{x_{[m] \setminus A}} f(x_A, x_{[m] \setminus A})$$
Marginalization example

The corresponding joint probability matrix is

\[
\begin{pmatrix}
0.0170 & 0.0419 & 0.0659 & 0.1248 \\
0.0230 & 0.0649 & 0.0629 & 0.1008 \\
0.0409 & 0.1098 & 0.0679 & 0.0788 \\
0.0339 & 0.0808 & 0.0419 & 0.0449
\end{pmatrix}
\]

The marginal probability \( P_A(x_A) \) of \( X_A \) is

\[
\begin{pmatrix}
0.2495 \\
0.2515 \\
0.2974 \\
0.2016
\end{pmatrix}
\]
Conditioning

Def: Let $A, B \subseteq [m]$ be pairwise disjoint subsets and let $x_B \in \mathcal{X}_B$. The conditional density of $X_A$ given $X_B = x_B$ is defined as

$$f_{A|B}(x_A | x_B) := \begin{cases} \frac{f_{A\cup B}(x_A, x_B)}{f_B(x_B)} & \text{if } f_B(x_B) > 0, \\ 0 & \text{otherwise} \end{cases}$$
Conditioning example

We start with the same joint probability matrix

\[
\begin{pmatrix}
0.0170 & 0.0419 & 0.0659 & 0.1248 \\
0.0230 & 0.0649 & 0.0629 & 0.1008 \\
0.0409 & 0.1098 & 0.0679 & 0.0788 \\
0.0339 & 0.0808 & 0.0419 & 0.0449
\end{pmatrix}
\]

The conditional density of \( X \) given \( Y = \text{bald} \) is

\[
\begin{pmatrix}
0.1478 \\
0.2000 \\
0.3565 \\
0.2957
\end{pmatrix}
\]

This is obtained by taking the first column of the joint probability matrix and dividing by 0.1148 (the sum of the entries in the first column)
Def: Let $A, B, C \subseteq [m]$ be pairwise disjoint subsets. We say that $X_A$ is conditionally independent of $X_B$ given $X_C$ if and only if

$$f_{A\cup B \mid C}(x_A, x_B \mid x_C) = f_{A \mid C}(x_A \mid x_C)f_{B \mid C}(x_B \mid x_C)$$

for all $x_A, x_B, x_C$.

- The notation $X_A \perp\!\!\!\!\!\!\perp X_B \mid X_C$ (or $A \perp\!\!\!\!\!\!\perp B \mid C$) denotes that the random vector $X$ satisfies the conditional independence (CI) statement that $X_A$ is conditionally independent of $X_B$ given $X_C$. 
Soccer vs hair example

- $X$ = random variable for the length of the hair
- $Y$ = random variable for how much soccer one watches
- $Z$ = random variable for the gender

In the soccer vs hair example: $X \perp Y \mid Z$
Marginal independence

- A statement of the form $X_A \perp \perp X_B := X_A \perp \perp X_B | X_{\emptyset}$ is called a marginal independence statement.

- It corresponds to the factorization of densities

\[
f_{A \cup B}(x_A, x_B) = f_A(x_A)f_B(x_B).
\]

- This is the same as the independence of random variables.
Conditional independence axioms

• Suppose a random vector $X$ satisfies a set of conditional independence statements. Which other conditional independence relations must the same random vector satisfy?

• There are some easy conditional independence implications, which are called the conditional independence axioms or conditional independence rules.
Conditional independence axioms

Prop: Let $A, B, C, D \subseteq [m]$ be pairwise disjoint subsets. Then

- **(symmetry)** $X_A \perp X_B \mid X_C \implies X_B \perp X_A \mid X_C$

- **(decomposition)** $X_A \perp X_{B \cup D} \mid X_C \implies X_A \perp X_B \mid X_C$

- **(weak union)** $X_A \perp X_{B \cup D} \mid X_C \implies X_A \perp X_B \mid X_{C \cup D}$

- **(contraction)** $X_A \perp X_B \mid X_{C \cup D}$ and $X_A \perp X_D \mid X_C \implies X_A \perp X_{B \cup D} \mid X_C$
Prop (Intersection axiom): Suppose that $f(x) > 0$ for all $x \in \mathcal{X}$. Then

$$X_A \perp X_B \mid X_{CUD} \text{ and } X_A \perp X_C \mid X_{BUD} \implies X_A \perp X_{BUC} \mid X_D.$$ 

- The condition $f(x) > 0$ for all $x$ is stronger than necessary.

- For discrete random variables, precise conditions can be given which guarantee that the intersection axiom holds. This is done using algebra!
Undirected graphical models
Graphs

- Graph $G = (V, E)$
- Nodes or vertices $V$
- Edges $E \subseteq V \times V$
- A graph is undirected if $(u, v) \in E$ implies that $(v, u) \in E$
- Corresponding random vector $X = (X_v : v \in V)$
  - $X$ takes values in $\mathcal{X}_V = \prod_{v \in V} \mathcal{X}_v$
Graphical models

In the graphical model associated to a graph $G$:

- an edge $(u, v)$ of the graph $G$ expresses some sort of dependence between the vertices $u$ and $v$;

- a non-edge $(u, v)$ of the graph $G$ expresses some sort of conditional independence between the vertices $u$ and $v$. 
• A path between vertices $u$ and $w$ in a graph $G$ is a sequence of vertices $u = v_1, v_2, \ldots, v_k = w$ such that each $(v_{i-1}, v_i) \in E$.

• A pair of vertices $a, b \in V$ is separated by a set of vertices $C \subseteq V \setminus \{a, b\}$ if every path from $a$ to $b$ contains a vertex in $C$.

• Let $A, B, C$ be disjoint subsets of $V$. Then $A$ and $B$ are separated by $C$, if $a$ and $b$ are separated by $C$ for any $a \in A$ and $b \in B$. 
Poll: Let $G$ be a graph with nodes $\{1,2,3,4\}$ and edges $(1,2), (2,3), (2,4), (3,4)$. Which of the following sets are separators for the nodes 1 and 4?

1. $\{2\}$
2. $\{3\}$
3. $\{2,3\}$
4. $\{1,2,3,4\}$
Poll: Let $G$ be a graph with nodes $\{1,2,3,4\}$ and edges $(1,2), (2,3), (2,4), (3,4)$. Which of the following sets are separators for the nodes 1 and 4?

1. $\{2\}$ - Correct
2. $\{3\}$
3. $\{2,3\}$ - Correct
4. $\{1,2,3,4\}$
Pairwise Markov property

Let $G = (V, E)$ be an undirected graph.

Def: The pairwise Markov property associated to $G$ consists of all conditional independence statements $X_u ⊥ ⊥ X_v | X_{V \setminus \{u, v\}}$, where $(u, v)$ is not an edge of $G$. It is denoted $G_{\text{pairs}}$.

Example: The pairwise Markov property associated to $G$ is:

1. $\{1 \perp \perp 3 | (2,4), 1 \perp \perp 4 | (2,3)\}$
2. $\{1 \perp \perp 3 | 2, 1 \perp \perp 4 | 2\}$
3. $\{1 \perp \perp 3 | (2,4)\}$
4. $\{1 \perp \perp 4 | (2,3)\}$
Let $G = (V, E)$ be an undirected graph.

**Def:** The **pairwise Markov property** associated to $G$ consists of all conditional independence statements $X_u \perp X_v | X_{V \setminus \{u, v\}}$, where $(u, v)$ is not an edge of $G$.

**Example:** The pairwise Markov property associated to $G$ is:

1. $\{1 \perp 3 | (2, 4), 1 \perp 4 | (2, 3)\}$ - Correct
2. $\{1 \perp 3 | 2, 1 \perp 4 | 2\}$
3. $\{1 \perp 3 | (2, 4)\}$
4. $\{1 \perp 4 | (2, 3)\}$
Pairwise Markov property

- Let $G = (V, E)$ be an undirected graph.
- We consider random vectors $X_V = (X_v)_{v \in V}$ taking values in $\mathcal{X}_V = \prod_{v \in V} \mathcal{X}_v$.
- Moreover, we assume that the probability distributions of the random vectors belong to a statistical model.
  - Example 1: $|V|$-dimensional multivariate normal distributions
  - Example 2: Discrete random vectors on on a fixed discrete set
- The first statistical model that we consider consists of all such random vectors that also satisfy the pairwise Markov property relative to $G$. 
Multivariate normal distribution

Let $PD_m$ be the set of $m \times m$ symmetric positive definite matrices.

**Def:** Suppose $\mu \in \mathbb{R}^m$ and $\Sigma \in PD_m$. Then a random vector $X = (X_1, \ldots, X_m)$ is distributed according to the multivariate normal distribution $\mathcal{N}_m(\mu, \Sigma)$ if it has the density function

$$
\phi_{\mu, \Sigma}(y) = \frac{1}{(2\pi)^{m/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (y - \mu)^T \Sigma^{-1} (y - \mu) \right\}.
$$
Prop: The conditional independence statement $X_A \perp\!\!\!\perp X_B \mid X_C$ holds for a multivariate normal random vector $X \sim \mathcal{N}(\mu, \Sigma)$ if and only if the submatrix $\Sigma_{A \cup C, B \cup C}$ of the covariance matrix $\Sigma$ has rank $\#C$. 

Multivariate Gaussian random variables

- The CI statement $X_u \perp\!\!\!\perp X_v \mid X_{V\setminus\{u,v\}}$ is equivalent to the matrix $\Sigma_{V\setminus\{u\}, V\setminus\{v\}}$ having rank $|V\setminus\{u, v\}|$ or equivalently $\det(\Sigma_{V\setminus\{u\}, V\setminus\{v\}}) = 0$.

- This is equivalent to $(\Sigma^{-1})_{u,v} = 0$.

- The pairwise Markov property holds for a Gaussian distribution if and only if the entries of the concentration matrix corresponding to non-edges are zero.
Multivariate Gaussian random variables

Which form do the concentration matrices of a Gaussian distribution obeying the pairwise Markov property have?

1. \[
\begin{pmatrix}
  k_{11} & 0 & k_{13} & k_{14} \\
  0 & k_{22} & 0 & 0 \\
  k_{13} & 0 & k_{33} & 0 \\
  k_{14} & 0 & 0 & k_{44}
\end{pmatrix}
\]

2. \[
\begin{pmatrix}
  k_{11} & k_{12} & 0 & 0 \\
  k_{12} & k_{22} & k_{23} & k_{24} \\
  0 & k_{23} & k_{33} & k_{34} \\
  0 & k_{24} & k_{34} & k_{44}
\end{pmatrix}
\]
Multivariate Gaussian random variables

Which form do the concentration matrices of a Gaussian distribution obeying the pairwise Markov property have?

1. \[
\begin{pmatrix}
k_{11} & 0 & k_{13} & k_{14} \\
0 & k_{22} & 0 & 0 \\
k_{13} & 0 & k_{33} & 0 \\
k_{14} & 0 & 0 & k_{44}
\end{pmatrix}
\]

2. \[
\begin{pmatrix}
k_{11} & k_{12} & 0 & 0 \\
k_{12} & k_{22} & k_{23} & k_{24} \\
0 & k_{23} & k_{33} & k_{34} \\
0 & k_{24} & k_{34} & k_{44}
\end{pmatrix}
\]

- Correct
Def: The global Markov property associated to $G$ consists of all conditional independence statements $X_A \perp X_B \mid X_C$ for all disjoint sets $A$, $B$, and $C$ such that $C$ separates $A$ and $B$ in $G$. It is denoted $\mathcal{C}_{\text{global}}$.

Example: The global Markov property associated to $G$ is:

1. $\{1 \perp (3,4) \mid 2\}$

2. $\{1 \perp 3 \mid (2,4), 1 \perp 4 \mid (2,3)\}$

3. $\{1 \perp 3 \mid (2,4), 1 \perp 4 \mid (2,3), 1 \perp (3,4) \mid 2\}$
Def: The global Markov property associated to $G$ consists of all conditional independence statements $X_A \perp X_B | X_C$ for all disjoint sets $A$, $B$, and $C$ such that $C$ separates $A$ and $B$ in $G$.

Example: The global Markov property associated to $G$ is:

1. $\{1 \perp (3,4) | 2\}$
2. $\{1 \perp 3 | (2,4), 1 \perp 4 | (2,3)\}$
3. $\{1 \perp 3 | (2,4), 1 \perp 4 | (2,3), 1 \perp (3,4) | 2\}$ - Correct
Global Markov property

- Let $G = (V, E)$ be an undirected graph.
  - We consider random vectors $X_V = (X_v)_{v \in V}$ taking values in $\mathcal{X}_V = \prod_{v \in V} \mathcal{X}_v$.
- Moreover, we assume that the probability distributions of the random vectors belong to a statistical model.
  - Example 1: $|V|$-dimensional multivariate Gaussian distributions.
  - Example 2: Discrete random vectors on discrete sets.
- The second statistical model that we consider consists of all such random vectors that also satisfy the global Markov property relative to $G$. 
Markov properties

- It always holds $\mathcal{C}_{\text{pairs}} \subseteq \mathcal{C}_{\text{global}}$.

**Example:**

- $\mathcal{C}_{\text{pairs}} = \{ 1 \perp 3 \mid (2,4), 1 \perp 4 \mid (2,3) \}$
- $\mathcal{C}_{\text{global}} = \mathcal{C}_{\text{pairs}} \cup \{ 1 \perp (3,4) \mid 2 \}$
Prop (Intersection axiom): Suppose that $f(x) > 0$ for all $x$. Then

$$X_A \perp X_B \mid X_{C \cup D} \text{ and } X_A \perp X_C \mid X_{B \cup D} \implies X_A \perp X_{B \cup C} \mid X_D.$$ 

• The condition $f(x) > 0$ for all $x$ is stronger than necessary.
Markov properties

**Theorem:** If the distribution \( P \) of a random vector \( X \) satisfies the intersection axiom, then \( P \) obeys the pairwise Markov property for \( G \) if and only if it obeys the global Markov property for \( G \).
Multivariate Gaussian random variables

For multivariate Gaussian random variables with non-singular covariance matrix, the density function is strictly positive.

⟹ the intersection axiom holds

⟹ the Markov properties are equivalent in this class of distributions
Next we want to characterize all the distributions that satisfy the Markov properties for a given graph.

Hammersley-Clifford theorem relates the implicit description of a graphical model through Markov properties to a parametric description.
Let $G = (V, E)$ be an undirected graph.

A subset of vertices $C \subseteq V$ is a **clique** if $(i, j) \in E$ for all $i, j \in C$.

The set of **maximal cliques** of $G$ is denoted $\mathcal{C}(G)$.

For each $C \in \mathcal{C}(G)$, we introduce a continuous nonnegative **potential function** $\phi_C : \mathcal{X}_C \to \mathbb{R}_{\geq 0}$. 
Maximal cliques

Example: Which are maximal cliques of $G$?

1. \{1\}
2. \{1,2\}
3. \{1,2,3\}
4. \{2,3,4\}
Maximal cliques

Example: Which are maximal cliques of $G$?

1. $\{1\}$ - Correct

2. $\{1,2\}$ - Correct

3. $\{1,2,3\}$

4. $\{2,3,4\}$ - Correct
Factorization property

**Def:** The distribution of $X$ factorizes according to the graph $G$ if its probability density function $f(x)$ can be written as

$$f(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}(G)} \phi_C(x_C),$$

where $\phi_C$ are some potential functions and $Z < \infty$ is the normalizing constant.
Factorization property

\[ f(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}(G)} \phi_C(x_C) \]

Example: A distribution factorizes according to \( G \) if its density \( f(x) \) can be written as

\[ f(x) = \frac{1}{Z} \phi_{12}(x_1, x_2) \phi_{234}(x_2, x_3, x_4). \]
Factorization property

• Let \( G = (V, E) \) be an undirected graph.

  • We consider random vectors \( X_V = (X_v)_{v \in V} \) taking values in \( \mathcal{X}_V = \prod_{v \in V} \mathcal{X}_v \).

• Moreover, we assume that the probability distributions of the random vectors belong to a statistical model.

  • Example 1: \( |V| \)-dimensional multivariate Gaussian distributions.

  • Example 2: Discrete random vectors on discrete sets.

• The third statistical model that we consider consists of all such random vectors whose distributions factorize according to \( G \).
Theorem (Hammersley-Clifford): A distribution with positive and continuous density $f$ satisfies the pairwise Markov property on the graph $G$ if and only if it factorizes according to $G$.

- The Gaussian case is completely covered by the Hammersley-Clifford theorem.
- All distributions on a discrete space are considered continuous.
- What happens in the discrete case?
• Lauritzen “Graphical Models”

• Maathuis, Drton, Lauritzen, Wainwright “Handbook of Graphical Models”

• Koller and Friedman “Probabilistic Graphical Models”