Graphical Model

Songpan Yang, Jiaqi Hu

School of Mathematics Sun Yat-sen University
School of Gifted Young University of Science and Technology of China

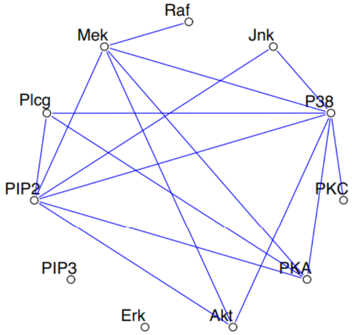
Sep, 1, 2020

Table of Contents

Table of Contents

Graphical model

- Why we choose Graphical model?
- To understand the joint distribution of the entire set of random variables.
- **graph** $G = (V, E)$: $V = \{1, ..., d\}$ $\quad E = \{(s, t) : s, t \in V\}$



Graphical model

- **Adjacency matrix**: $G(s,t) = 1$ if $(s,t) \in E$.
- **Neighbors**: $\text{nbr}(s) \triangleq \{t : G(s,t) = 1 \vee G(t,s) = 1\}$.
- **Degree**: The **degree** of a node is the number of neighbors.
- **Cycle or loop** : Get back to where we started by following edges.
- **Clique**: A subset of vertices that are all joined by edges.
- **Maximal clique**: A clique can't be made any larger without losing the clique property.

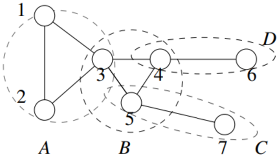


Table of Contents

Factorization

Definition

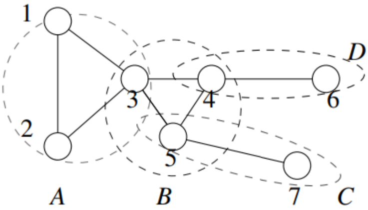
The random vector $(X_1, \ldots, X_d)$ factorizes according to the graph G if its density function p can be represented as:

$$p(x_1, \ldots, x_d) \propto \prod_{C \in \mathfrak{C}} \psi_C(X_C)$$

for some collection of clique compatibility functions: $\psi_C : \mathcal{X}^C \to [0, \infty)$

- $\mathfrak{C}$ is the set of all cliques in G
- $\forall C \in \mathfrak{C}$, ψ_C is clique compatibility functions of $X_C := (x_j, j \in C)$

Factorization



- Subsets A and B are 3-cliques
- Subsets C and D are 2-cliques
- Then, the function is:

$$p(x_1, \ldots, x_7) \propto \psi_{123}(x_1, x_2, x_3)\, \psi_{345}(x_3, x_4, x_5)\, \psi_{46}(x_4, x_6)\, \psi_{57}(x_5, x_7)$$

Examples: Markov chain factorization

- The Markov Chain:

$$p(x_1, \ldots, x_d) = p(x_1)\, p(x_2 \mid x_1)\, (x_d \mid x_{d-1})$$

- The vertex-based functions:

$$\begin{aligned} \psi_1(x_1) &= p(x_1) \\ \psi_j(x_j) &= 1 \quad \text{for all } j = 2, \ldots, d \end{aligned}$$

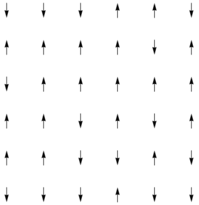
- The edge-based functions:

$$\psi_{j,j+1}(x_j, x_{j+1}) = p(x_{j+1} \mid x_j)$$

Examples: Ising model

- A mathematical model of ferromagnetism in statistical mechanics.
- Let $s_i \in \{+1, -1\}, \quad (i = 1, \ldots, d)$ be the spins.
- The model can shown below.
- $E_{\{s_i\}}$ is the energy under $\{s_i\}$.
- J is a constant, H is external magnetic field strength.

$$E_{\{s_i\}} = -J \sum_{\langle i,j \rangle} s_i s_j - H \sum_{i}^{d} s_i$$



Examples: Ising model

- The new state of s_i can be:

$$s_i(t+1) = \begin{cases} s_i' & \text{with probability } \mu \\ s_i(t) & \text{with probability } 1-\mu \end{cases}$$

- And

$$\mu = \min\left\{\exp\left(E\left(s_i(t)\right) - E\left(s_i'\right)\right)/(kT), 1\right\}$$

- Then we can get the Boltzmann distribution:

$$p\left(\{s_i\}\right) = \frac{1}{Z}\exp\left(-\frac{E_{\{s_i\}}}{kT}\right)$$

Examples: Ising model

- Consider a vector $X = (X_1, \ldots, X_d)$, $X_j \in \{0, 1\}$
- Given an undirected graph $G = (V, E)$

$$p(x_1, \ldots, x_d; \theta^*) = \frac{1}{Z(\theta^*)} \exp\left\{\sum_{j \in V} \theta_j^* x_j + \sum_{(j,k) \in E} \theta_{jk}^* x_j x_k\right\}$$

- The parameter θ_j^* is associated with vertex $j \in V$.
- The parameter θ_{jk}^* is associated with edge $(j, k) \in E$.
- The quantity $Z(\theta^*)$ is normalization.

$$Z(\theta^*) = \sum_{x \in \{0,1\}^d} \exp\left\{\sum_{j \in V} \theta_j^* x_j + \sum_{(j,k) \in E} \theta_{jk}^* x_j x_k\right\} \tag{1}$$

Examples: Multivariate Gaussian factorization

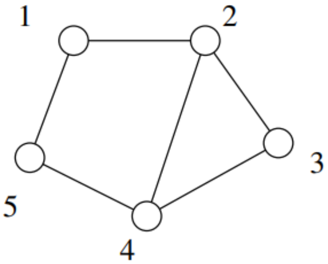
- For a non-degenrate Gaussian distribution with zero mean.
- The precision matrix: $\mathbf{\Theta}^* = \mathbf{\Sigma}^{-1}$
- The density can be written as:

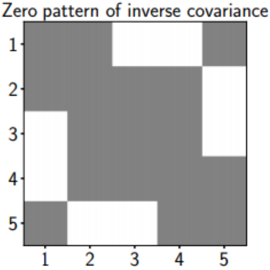
$$p\left(x_1, \ldots, x_d; \Theta^*\right) = \frac{\sqrt{\det\left(\Theta^*\right)}}{(2\pi)^{d/2}} e^{-\frac{1}{2} x^{\mathrm{T}} \Theta^* x}$$

- Expanding the quadratic form:

$$e^{-\frac{1}{2} x^{\mathrm{T}} \mathbf{\Theta}^* x} = \exp\left(-\frac{1}{2} \sum_{(j,k) \in E} \Theta^*_{jk} x_j x_k\right) = \prod_{(j,k) \in E} \underbrace{e^{-\frac{1}{2} \Theta^*_{jk} x_j x_k}}_{\psi_{jk}\left(x_j, x_k\right)}$$

- Zero-mean Gaussian distribution can be factorized in terms of functions on edges, or cliques of size two





Conditional independence

Definition

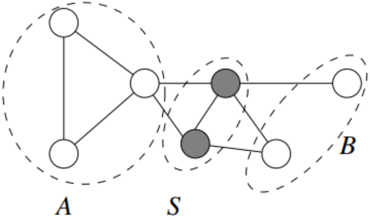
A random vector $X = (X_1, \ldots, X_d)$ is Markov with respect to a graph G if, for all vertex cutsets S breaking the graph into disjoint pieces A and B, the conditional independence statement $\mathrm{X}_A \perp \mathrm{X}_B \mid \mathrm{X}_S$ holds.

Vertex cutset

- For $G = (V, E)$, a subset S of V ($S \subseteq V$), Remove S.
- The subgraph $G = (V \backslash S)$ with vertex set $V \backslash S$ and residual edge set:

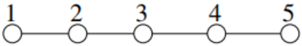
$$E(V \backslash S) := \{(j, k) \in E \mid j, k \in V \backslash S\}$$

- The set S is vertex cutset if $G = (V \backslash S)$ consists of two or more disconnected non-empty components.



Examples: Markov chain conditional independence

- In the Markov chain, each vertex $j \in \{2, 3, \ldots, d-1\}$ is non-trivial cuset.
- Break the graph into the past $P = \{1, 2, \ldots, j-1\}$ and the future $F = \{j+1, \ldots, d\}$.
- The past X_P and future X_F are conditionally independent given the presetn X_j.



Examples: Neighborhood-based cutsets

- $\forall j \in V$, the neighborhood set is $\mathcal{N}(j) := \{k \in V \mid (j,k) \in E\}$
- $\mathcal{N}(j)$ is always vertex cutset, $A = \{j\}$, $B = V \backslash (\mathcal{N}(j) \cup \{j\})$ are two disjoint components
- The choice of vertex cutset plays an important role in our discussion of neighborhood-based methods for graphical model selection.

Hammersley–Clifford equivalence

Theorem

For a given undirected graph and any random vector $X = (X_1, \ldots, X_d)$ with strictly positive density p, the above two properties are equivalent

Here we show the factorization property implies the Markov property.

- Suppose the factorization holds. We need to show $X_A \perp X_B | X_S$
- Define $\mathfrak{C}_A := \{C \in \mathfrak{C} \mid C \cap A \neq \emptyset\}$
- Define $\mathfrak{C}_B := \{C \in \mathfrak{C} \mid C \cap B \neq \emptyset\}$
- Define $\mathfrak{C}_S := \{C \in \mathfrak{C} \mid C \subseteq S\}$
- $\mathfrak{C} = \mathfrak{C}_A \cup \mathfrak{C}_S \cup \mathbb{C}_B$

Hammersley–Clifford equivalence

- we may write

$$p\left(x_A, x_S, x_B\right) = \frac{1}{Z} \underbrace{\left[\prod_{C \in \mathbb{C}_A} \psi_C\left(x_C\right)]\right.}_{\Psi_A(x_A, x_S)} \underbrace{\left[\prod_{C \in \mathbb{C}_S} \psi_C\left(x_C\right)\right]}_{\Psi_S(x_S)}] \underbrace{\left[\prod_{C \in \mathbb{C}_B} \psi_C\left(x_C\right)\right]}_{\Psi_B(x_B, x_S)}$$

- Define

$$Z_A\left(x_S\right) := \sum_{x_A} \Psi_A\left(x_A, x_S\right) \quad \text{and} \quad Z_B\left(x_S\right) := \sum_{x_B} \Psi_B\left(x_B, x_S\right)$$

- then

$$p\left(x_S\right) = \frac{Z_A\left(x_S\right) Z_B\left(x_S\right)}{Z} \Psi_S\left(x_S\right)$$

$$p\left(x_A, x_S\right) = \frac{Z_B\left(x_S\right)}{Z} \Psi_A\left(x_A, x_S\right) \Psi_S\left(x_S\right)$$

Hammersley–Clifford equivalence

for any x_S for which $p(x_S) > 0$

$$\frac{p(x_A, x_S, x_B)}{p(x_S)} = \frac{\frac{1}{Z}\Psi_A(x_A, x_S)\Psi_S(x_S)\Psi_B(x_B, x_S)}{\frac{Z_A(x_S)Z_B(x_S)}{Z}\Psi_S(x_S)}$$

$$= \frac{\Psi_A(x_A, x_S)\Psi_B(x_B, x_S)}{Z_A(x_S)Z_B(x_S)}$$

$$\frac{p\left(x_A, x_S\right)}{p\left(x_S\right)} = \frac{\frac{Z_B(x_S)}{Z}\Psi_A\left(x_A, x_S\right)\Psi_S\left(x_S\right)}{\frac{Z_A(x_S)Z_B(x_S)}{\Psi}_S\left(x_S\right)} = \frac{\Psi_A\left(x_A, x_S\right)}{Z_A\left(x_S\right)}$$

$$\frac{p\left(x_B, x_S\right)}{p\left(x_S\right)} = \frac{\frac{Z_A(x_S)}{Z}\Psi_B\left(x_B, x_S\right)\Psi_S\left(x_S\right)}{\frac{Z_A(x_S)Z_B(x_S)}{\Psi}_S\left(x_S\right)} = \frac{\Psi_B\left(x_B, x_S\right)}{Z_B\left(x_S\right)}$$

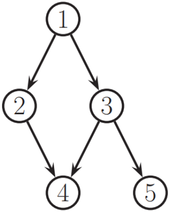
Table of Contents

Terminology

Directed graphical models(**DGM**):
$G = (\mathcal{V}, \mathcal{E}), \mathcal{E} = \{1, \cdots, d\}, \mathcal{E} = \{(s,t) : s, t \in \mathcal{V}\}$

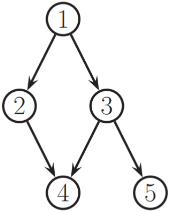
$$G(s,t) = \begin{cases} 1 & s \to t \ is \ an \ edge \\ 0 & s \to t \ is \ not \ an \ edge \end{cases}$$

- **Parent**, $\mathrm{pa}(s) \triangleq \{t : G(t,s) = 1\}$.
- **Child**, $\mathrm{ch}(s) \triangleq \{t : G(s,t) = 1\}$.
- **Family**, $\mathrm{fam}(s) \triangleq \{s\} \bigcup pa(s)$.
- **Ancestors**, the set of nodes that connect to t via a trail, $\mathrm{anc}(t) \triangleq \{s : s \rightsquigarrow t\}$.
- **Descendants**, the set of nodes that can be reached via trails from s, $\mathrm{desc}(s) \triangleq \{t : s \rightsquigarrow t\}$.
- **Neighbors**, $\mathrm{nbr}(s) \triangleq \{t : G(s,t) = 1 \vee G(t,s) = 1\}$.



Terminology

- **Cycle or loop**, the path that start from a node and return to the node.
- **DAG**, directed acyclic graph, is a directed graph with no directed cycles.
- **Bayesian networks**, a directed graph with no directed cycles.
- **Topological ordering**, for a DAG, topological ordering is a numbering of nodes such that parents have lower numbers than their children.
- **Clique**, a clique is a set of nodes that are all neighbors of each other.
- **maximal clique**, a clique which cannot be made any larger without losing the clique property.

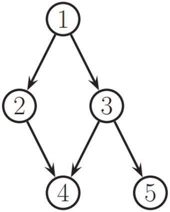


DGM

ordered Markov property: a node only depends on its immediate parents, not on all predecessors in the ordering,

$$x_s \perp x_{\text{pred}}(s) \setminus \text{pa}(s) \mid x_{\text{pa}}(s)$$

pred(s) are the predecessors of nodes **in the ordering**.



DGM

In general, the joint distribution of a DGM(no circle) can be expressed as:

$$p\left(x_{1:d} \mid G\right) = \prod_{t=1}^{d} p\left(x_t \mid x_{\mathrm{pa}(t)}\right)$$

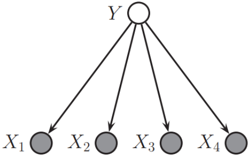
Proof.

$$\begin{aligned} p\left(x_{1:d} \mid G\right) &= p(x_1)p(x_2|x_1)p(x_3|x_2,x_1)\cdots p(x_d|x_1 : d-1) \\ &= \prod_{t=1}^{d} p(x_t|x_{1:t-1}) = \prod_{t=1}^{d} p(x_t|pa(x_t), x_{1:t-1} \backslash pa(x_t)) \\ &= \prod_{t=1}^{d} p(x_t|pa(x_t)) \end{aligned}$$

$\square$

Example: Naive Bayes classifier

The naive Bayes assumption: the features are conditionally independent.



We can write the formula below.

$$p(Y,X) = p(Y)\prod_{j=1}^{d} p\left(X_j \mid Y\right)$$

Example: Markov chain

- **Frst order Markov chain** is a sequence $x_1, x_2, \cdots,$ of random variables satistying the rule of conditonal independence:

$$P(x_n = i_n | x_{n-1} = i_{n-1}, \cdots, x_1 = i_1) = P(x_n = i_n | x_{n-1} = i_{n-1})$$

- **Second order Markov chain**: x_n dependent on x_{n-1}, x_{n-2}, satistying the rule of conditonal independence: $P(x_n = i_n | x_{n-1} = i_{n-1}, \cdots, x_1 = i_1) = P(x_n = i_n | x_{n-1} = i_{n-1}, x_{n-2} = i_{n-2})$

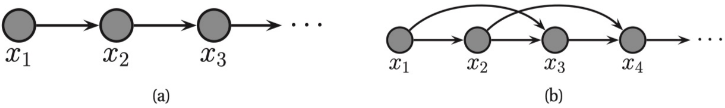


Figure 10.3 A first and second order Markov chain.

Example: Hidden Markov model

Let $\{X_n\}, \{Z_n\}$ be discrete-time stochastic processes, the pair (X_n, Z_n) is a hidden markov model if:

- $\{Z_n\}$ is a markov process and is not directly observable(hidden).
- $P(X_n \in A | Z_1 = z_1, \cdots, Z_n = z_n) = P(X_n \in A | Z_n = z_n)$

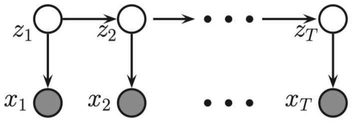


Figure 10.4 A first-order HMM.

In graphical model, blank nodes are not observed, the nodes filled with color are observed variables.

Example: Directed Gaussian graphical models

Suppose all the variables are real-valued, and the CPD have the following form:

$$p\left(x_t \mid \mathrm{x}_{\mathrm{pa}(t)}\right) = \mathcal{N}\left(x_t \mid \mu_t + \mathrm{w}_t^T \mathrm{x}_{\mathrm{pa}(t)}, \sigma_t^2\right)$$

Multiplying all these CPDs and get a **Gaussian Bayes net**

$$p(\mathrm{x}) = \mathcal{N}(\mathrm{x} \mid \boldsymbol{\mu}, \Sigma)$$

We want to derive $\boldsymbol{\mu}$ and Σ

$$x_t = \mu_t + \sum_{s \in \mathrm{pa}(t)} w_{ts}\left(x_s - \mu_s\right) + \sigma_t z_t$$

where $z_t \sim \mathcal{N}(0,1)$, σ_t is the conditional standard deviation of x_t given its parents.

w_{ts} is the strength of the $s \to t$ edge. μ_t is the local mean.

Example: Directed Gaussian graphical models

let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$, $\mathsf{S} \triangleq \mathrm{diag}(\boldsymbol{\sigma})$, we have

$$(\mathsf{x} - \boldsymbol{\mu}) = \mathsf{W}(\mathsf{x} - \boldsymbol{\mu}) + \mathsf{Sz}$$

Then

$$\mathsf{Sz} = (\mathsf{I} - \mathsf{W})(\mathsf{x} - \boldsymbol{\mu})$$

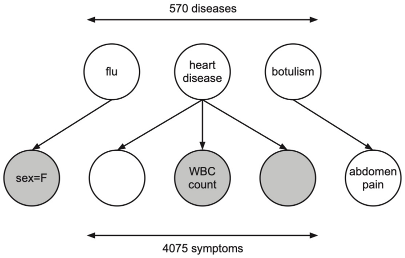
$$\mathsf{Sz} = \begin{pmatrix} 1 & & & & \\ -w_{21} & 1 & & & \\ -w_{32} & -w_{31} & 1 & & \\ \vdots & & & \ddots & \\ -w_{d1} & -w_{d2} & \ldots & -w_{d,d-1} & 1 \end{pmatrix} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \\ \vdots \\ x_d - \mu_d \end{pmatrix}$$

$$\begin{aligned} \boldsymbol{\Sigma} &= \mathrm{cov}[\mathsf{x}] = \mathrm{cov}[\mathsf{x} - \boldsymbol{\mu}] \\ &= \mathrm{cov}[\mathsf{USz}] = \mathsf{US}\,\mathrm{cov}[\mathsf{z}]\mathsf{SU}^T = \mathsf{US}^2\mathsf{U}^T \end{aligned}$$

where $\mathsf{U} = (\mathsf{I} - \mathsf{W})^{-1}$

Quick medical reference (QMR) network is a medical diagnosis network. The QMR model is a bipartite graph structure, with **diseases (causes)** at the top and **symptoms or findings** at the bottom. All nodes are binary. We can write the distribution as follows:

$$p(\mathrm{v}, \mathrm{h}) = \prod_s p(h_s) \prod_t p\left(v_t \mid \mathrm{h}_{\mathrm{pa}(t)}\right)$$



h_s represent the **hidden nodes** (diseases), and v_t represent the **visible nodes** (symptoms).

Conditional independence properties of DGMs

At the heart of any graphical model is a set of **conditional indepence (CI)** assumptions. For example, we write $\mathrm{x_A} \perp_G \mathrm{x_B} | \mathrm{x_C}$ if A is independent of B given C in the graph G.
Let $I(G)$ be the set of all such CI statements encoded by the graph.
P is a probability distribution, $I(P)$ is the set of all CI statements that hold for distribution P.

- **I-map**: independence map, if $I(G) \subseteq I(P)$. That is, for each $X, Y, Z \subseteq V$, if

$$X \perp_G Y | Z \Longrightarrow X \perp_P Y | Z$$

- **D-map**: dependence map, if $I(P) \subseteq I(G)$. That is,

$$X \perp_P Y | Z \Longrightarrow X \perp_G Y | Z$$

- **P-map**: perfect map, if $I(G) = I(P)$, that is,

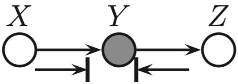
$$X \perp_G Y | Z \Longleftrightarrow X \perp_P Y | Z$$

d-separation

We said a path P is d-separation by a set of nodes E iff(if and only if) at least one of the following three conditions hold:

- P contains a **chain**, $X \rightarrow Y \rightarrow Z$, where $Y \in E$. In this case $X \perp Z | Y$

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

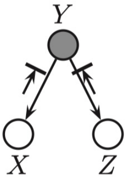


(a)

- P contains a tent, $X \swarrow Y \searrow Z$, where $Y \in E$. In this case $X \perp Z | Y$

d-separation

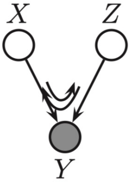
$$p(x,z|y) = \frac{p(x,y,z)}{p(y)} = \frac{p(y)p(x|y)p(z|y)}{p(y)} = p(x|y)p(z|y)$$



(b)

- P contains a $v-structure$, $X \searrow Y \swarrow Z$, where m is not in E and nor any descendant of m. Condition on Y, X and Z are not independent.

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



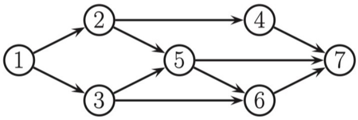
(c)

Bayes ball algorithm

Bayes ball algorithm (Shachter 1998) is a simple way to see if A is d-separated from B given E.

1. shade the nodes of E
2. place "balls" at each node in A, and found a path to B.

If we can not find a path, we say A is not d-separated from B given E.
If we can find a path we say A is d-separated from B given E.



- $x_2 \perp x_6 | x_5$, since the only path $x_2 \to x_5 \to x_6$ is blocked by x_5.
- x_2 is not $\perp x_6 | x_5, x_6$, since through path $x_2 \to x_4 \to x_7 \to x_6$, we can reach x_6.