Graphical Models Intro (2023)

- Suppose we have a collection of jointly distributed random variables
$$\underline{X} = [X_1, \dots, X_d]^T \quad \text{with joint pdf (or pmf)} \quad f_{\underline{X}}(x_1, \dots, x_d).$$

Questions

1. How can we compactly represent this joint distribution?
2. If the variables are observable, how can we learn such a compact representation from data?
3. How can we use this representation to infer the distribution for one subset of the variables given another in a reasonable amount of time?

- These are the questions to which the theory of graphical models aim to provide a general answer. We will see that the theory of graphical models contains a collection of theorems for answering these questions. As the theorems are model agnostic they can be applied broadly to problems in statical modeling and machine learning.
- In this module, we will prove these basic theorems and derive their consequences for the three questions above.

1 Representation

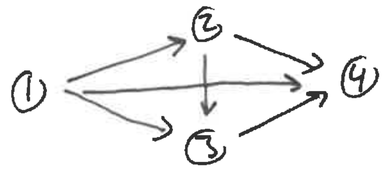
- The first goal of graphical models is to provide an informative representation of the joint distribution $f_{\underline{X}}(x_1, \dots, x_d)$.
- To do this we start with a rather uninformative representation that can be derived for any joint distribution.

Theorem 1 (Chain Rule) The pdf (pmf) for $\underline{X} = [X_1, \dots, X_d]^T$ satisfies
$$f_{\underline{X}}(x_1, \dots, x_d) = \prod_{i=1}^{d} f_{X_i \mid \underline{X}_{[i-1]}}(x_i \mid \underline{x}_{[i-1]})$$

$$\left([m] = \{1, \dots, m\} \quad \text{and} \quad \underline{X}_S = [X_i : i \in S]^T, \quad \underline{x}_S = [x_i : i \in S]^T \right)$$

- To represent the distribution we can draw a graph representing the dependencies specified by the conditional factors: $\left(f_{\underline{X}_i \mid \underline{X}_S}(x_i \mid \underline{x}_S) \Rightarrow j \to i \;\; \forall j \in S \right)$

- $\boxed{d = 4}$ $f_{\underline{X}}(x_1, x_2, x_3, x_4) = f_{\underline{X}_1}(x_1)\, f_{\underline{X}_2 \mid \underline{X}_1}(x_2 \mid x_1)\, f_{\underline{X}_3 \mid \underline{X}_{\{1,2\}}}(x_3 \mid x_1, x_2)\, f_{\underline{X}_4 \mid \underline{X}_{\{1,2,3\}}}(x_4 \mid x_1, x_2, x_3)$



- Pretty useless but still captures intuition for classic models:

$\boxed{\text{Example 1}}$ (Beta-Binomial) Consider the joint distribution $\underline{Z} = [\underline{X}, \Theta]^T$ where

$$\underline{X} \mid \Theta = \theta \sim \text{Bin}(n, \theta)$$
$$\Theta \sim \text{Beta}(\alpha, \beta).$$

The joint distribution is

$$f_{\underline{Z}}(x, \theta) = f_{\underline{X} \mid \Theta}(x \mid \theta)\, f_{\Theta}(\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}.$$

which we represent as $\Theta \longrightarrow \underline{X}$ ("$\underline{X}$ depends on Θ").

- The graph is not so informative yet. But it becomes increasingly useful as the model becomes increasingly complex.
- When building models for many observables with many parameters we tend to make use of conditional independence assumptions. The assumptions lead to more interesting/informative graph structure.

$\boxed{\text{Definition 1}}$ For $\underline{X} = [\underline{X}_1, \ldots, \underline{X}_d]^T$ and subsets $A, B, C \subseteq [d]$ disjoint with $A, B \neq \emptyset$, we say $\underline{X}_A \perp\!\!\!\perp \underline{X}_B \mid \underline{X}_C$ (~~with~~ $\underline{X}_A$ is independent of $\underline{X}_B$ given $\underline{X}_C$) if

$$f_{\underline{X}_A, \underline{X}_B \mid \underline{X}_C}(\underline{x}_A, \underline{x}_B \mid \underline{x}_C) = f_{\underline{X}_A \mid \underline{X}_C}(\underline{x}_A \mid \underline{x}_C)\, f_{\underline{X}_B \mid \underline{X}_C}(\underline{x}_B \mid \underline{x}_C)$$

for all $\underline{x}_A, \underline{x}_B, \underline{x}_C$ with $f_{\underline{X}_C}(\underline{x}_C) > 0$.

Theorem ② The following are equivalent:

(a) $\underline{X}_A \perp\!\!\!\perp \underline{X}_B \mid \underline{X}_C$

(b) $f_{\underline{X}_A \mid \underline{X}_B, \underline{X}_C}(\underline{x}_A \mid \underline{x}_B, \underline{x}_C) = f_{\underline{X}_A \mid \underline{X}_C}(\underline{x}_A \mid \underline{x}_C)$ for all $\underline{x}_A, \underline{x}_B, \underline{x}_C$

(c) $f_{\underline{X}_A \mid \underline{X}_B, \underline{X}_C}(\underline{x}_A \mid \underline{x}_B, \underline{x}_C) = f_{\underline{X}_A \mid \underline{X}_B, \underline{X}_C}(\underline{x}_A \mid \underline{x}_B', \underline{x}_C)$ for all $\underline{x}_A, \underline{x}_B, \underline{x}_B', \underline{x}_C$

Proof: (Exercise!)

- When our model makes certain conditional independence assumptions applying Theorem ② (b) to factors in our chain rule factorization can lead to more informative graph structure.

Example ② (A hierarchical model) Consider $\underline{Z} = [\Lambda, N, Y]^T$ defined by the three-stage hierarchy

$$Y \mid N = n \sim \text{Bin}(n, p) \qquad (p \text{ fixed})$$
$$N \mid \Lambda = \lambda \sim \text{Po}(\lambda)$$
$$\Lambda \sim \text{Gamma}(\alpha, \beta)$$

where we assume $Y \perp\!\!\!\perp \Lambda \mid N$.

By the chain rule,

$$f_{\Lambda, N, Y}(\lambda, n, y) = f_\Lambda(\lambda)\, f_{N \mid \Lambda}(n \mid \lambda)\, f_{Y \mid N, \Lambda}(y \mid n, \lambda)$$
$$= f_\Lambda(\lambda)\, f_{N \mid \Lambda}(n \mid \lambda)\, f_{Y \mid N}(y \mid n)$$

Theorem 2(b) + $Y \perp\!\!\!\perp \Lambda \mid N$

Get the graph

$$\Lambda \to N \to Y.$$

General Formula for getting a Graph Representation:

① Apply chain rule

② Apply available C.I. relations to shrink conditioning sets:

$$f_{Z_i \mid \underline{Z}_{[i-1]}}(x_i \mid \underline{x}_{[i-1]}) = f_{Z_i \mid \underline{Z}_{C_i}}(x_i \mid \underline{x}_{C_i}), \quad C_i \subseteq [i-1]$$

③ draw $j \to i \quad \forall j \in C_i, \ \forall i \in [d]$.

Example 3 (A Bayesian Hierarchical Model)

14 different laboratories have control groups of rats for a cancer study where laboratory i's control group contains n_i rats. Let Θ_i be the probability that a rat in control group i develops tumors. Given that the laboratories have similar practices and resources we expect some correlations between $\Theta_1, \ldots, \Theta_{14}$. So we treat them as a random sample from a Beta(α, β)-population. We have no good guess for the parameters α, β so we put a diffuse (but proper) prior on them with pdf $f_{A,B}(\alpha, \beta)$. We also assume the number of rats X_i in group i that develops tumors depends <u>only</u> on Θ_i.

We could write these CI assumptions explicitly, but it would be difficult to parse given their number. More easily we encode them in the pdf's factorization:

$$f_{\underline{X}, \underline{\Theta}, A, B}(\underline{x}, \underline{\theta}, \alpha, \beta) = f_{A,B}(\alpha, \beta) \prod_{i=1}^{14} f_{\Theta_i | A, B}(\theta_i | \alpha, \beta) \prod_{i=1}^{14} f_{X_i | \Theta_i}(x_i | \theta_i).$$

Even more easy to read is the associated graph:

$$\begin{tikzcd}
 & \Theta_1 \arrow[r] & X_1 \\
 & \Theta_2 \arrow[r] & X_2 \\
(A,B) \arrow[uur] \arrow[ur] \arrow[dr] & \vdots & \\
 & \Theta_{14} \arrow[r] & X_{14}
\end{tikzcd}$$

- Defining a graph as such according to a factorization always results in a DAG (Exercise!).
- The distribution together with its resulting DAG is a pair that we call a <u>DAG model</u>.

Definition (2) $\underline{X} = [X_1,\dots,X_d]^T$ is <u>Markov</u> to a DAG $G=([d],E)$ if
$$f_{\underline{X}}(x_1,\dots,x_d) = \prod_{i=1}^{d} f_{X_i \mid \underline{X}_{pa_G(i)}}(x_i \mid \underline{x}_{pa_G(i)}).$$

- The pair $(\underline{X}, G)$ is called a <u>DAG model</u>.
- The set $\mathcal{M}_F(G) = \{ \underline{X} : \underline{X}$ is Markov to $G \}$ is also called a <u>DAG model</u>.
- Sometimes $(\underline{X}, G)$ is called a <u>Bayesian network</u> since bayesian inference works backwards along the edges of the DAG.
- Sometimes $(\underline{X}, G)$ is called a <u>causal model</u>, although apriori there need not be any causal information in the model (there can be though!)
- In the Bayesian setting, using the graph representation can help us quickly identify ways to simplify posterior computations.

Example (4) Suppose we have a model where
$$T = Z + W, \qquad Z = X + Y,$$
$$X \mid \Lambda = \lambda \sim Po(\lambda), \quad Y \mid \Theta = \theta \sim Po(\theta), \quad W \mid \Gamma = \gamma \sim Po(\gamma), \text{ and}$$
Λ, Θ, Γ are independent and Gamma$(1,1)$-distributed.

- Rather than specifying a list of CI relations we instead say that $\underline{X} = [\Lambda, \Theta, \Gamma, X, Y, W, Z, T]^T$ is Markov to G:

$$\begin{tikzcd}
\Lambda \arrow[r] & X \arrow[dr] & & \\
& & Z \arrow[dr] & \\
\Theta \arrow[r] & Y \arrow[ur] & & T \\
\Gamma \arrow[r] & W \arrow[urr] & &
\end{tikzcd}$$

Given data $[\underline{Z}, \underline{W}, \underline{T}]^T = \mathbb{D}$ we want the posterior
$$f_{\Lambda,\Theta,\Gamma \mid \mathbb{D}}(\lambda, \theta, \gamma \mid \mathbb{D}).$$

Since all paths from Λ, Θ to Γ in G are "blocked" by Z, W, T it turns out that $\Lambda, \Theta \perp\!\!\!\perp \Gamma \mid Z, W, T$.

- This means we can break up our posterior computation in two

$$f_{\Lambda,\Theta,\Gamma|\mathbb{D}}(\lambda,\theta,\gamma|\mathbb{D}) = f_{\Lambda,\Theta|\mathbb{D}}(\lambda,\theta|\mathbb{D})\, f_{\Gamma|\mathbb{D}}(\gamma|\mathbb{D})$$

- Note that the CI relation $\Lambda,\Theta \perp\!\!\!\perp \Gamma \mid \underline{\mathbb{Z}}, W, T$ is not a relation that we assume when we assume $\underline{X}$ is Markov to G (Exercise: Check this!)
- Instead this relation is implied by our assumption.
- The fact that we can read CI relations implied by the Markov assumption is one useful property of DAG models!

Question S'pose $\underline{X}$ is Markov to G. What is the complete set of CI relations that we know hold in $\underline{X}$? (i.e. What CI relations are implied by the Markov assumption?)

- What does "implied" mean?

The Conditional Independence Axioms

(a) (Symmetry) $\underline{X}_A \perp\!\!\!\perp \underline{X}_B \mid \underline{X}_C \Rightarrow \underline{X}_B \perp\!\!\!\perp \underline{X}_A \mid \underline{X}_C$

(b) (Decomposition) $\underline{X}_A \perp\!\!\!\perp \underline{X}_{B\cup D} \mid \underline{X}_C \Rightarrow \underline{X}_A \perp\!\!\!\perp \underline{X}_B \mid \underline{X}_C$

(c) (Weak Union) $\underline{X}_A \perp\!\!\!\perp \underline{X}_{B\cup D} \mid \underline{X}_C \Rightarrow \underline{X}_A \perp\!\!\!\perp \underline{X}_B \mid \underline{X}_{C\cup D}$

(d) (Contraction) $\underline{X}_A \perp\!\!\!\perp \underline{X}_B \mid \underline{X}_{C\cup D}$ and $\underline{X}_A \perp\!\!\!\perp \underline{X}_D \mid \underline{X}_C \Rightarrow \underline{X}_A \perp\!\!\!\perp \underline{X}_{B\cup D} \mid \underline{X}_C$

(e) (Intersection) If $f_{\underline{X}}(\underline{x}) > 0 \ \forall \underline{x}$ then if $\underline{X}_A \perp\!\!\!\perp \underline{X}_B \mid \underline{X}_{C\cup D}$ and $\underline{X}_A \perp\!\!\!\perp \underline{X}_D \mid \underline{X}_{C\cup B} \Rightarrow \underline{X}_A \perp\!\!\!\perp \underline{X}_{B\cup D} \mid \underline{X}_C$.

Example 5 If $\underline{X} = [X_1, X_2, X_3, X_4]^T$ is Markov to

$$1 \to 2 \to 3 \to 4$$

then by definition we know $X_4 \perp\!\!\!\perp X_1, X_2 \mid X_3$ and $X_3 \perp\!\!\!\perp X_1 \mid X_2$.

Using the CI axioms we can show $X_1 \perp\!\!\!\perp X_4 \mid X_2$ (Exercise!)

Theorem 3 (Verma, Pearl) The complete set of CI relations implied by the Markov condition is

$$\mathcal{C}(G) = \{ \underline{X}_A \perp\!\!\!\perp \underline{X}_B \mid \underline{X}_C : A, B \text{ d-separated given } C \text{ in } G \}.$$

- By the end of the first two lectures we will have proven

Theorem 4 Let $\underline{X} = [X_1, \ldots, X_d]^T$ a distribution and $G = ([d], E)$ a DAG. The following are equivalent:

(a) $\underline{X}$ is Markov to G

(b) $\forall$ triples $A, B, C \subseteq [d]$ where A, B are d-separated given C in G we have $\underline{X}_A \perp\!\!\!\perp \underline{X}_B \mid \underline{X}_C$.

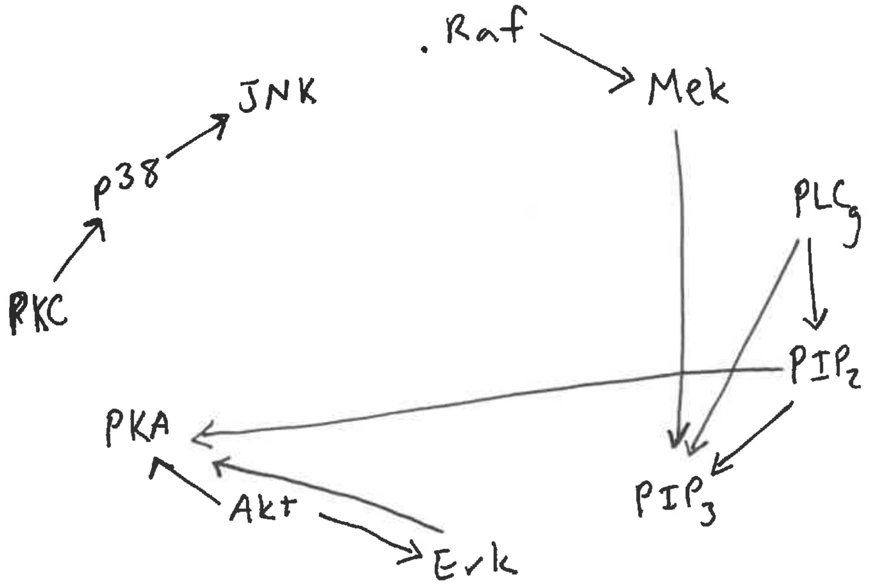
2 Structure Learning

- We may find ourselves having jointly distributed observed variables $\underline{X} = [X_1, \ldots, X_d]^T$, no idea how to parametrize the model, but would still like to learn a graphical structure that the distribution follows.
- This is the following unsupervised learning problem:

Problem Given a random sample (data) from $\underline{X} = [X_1, \ldots, X_d]^T$ can we find a "good" or "the best" DAG to which $\underline{X}$ is Markov?

- This problem is central in the field of causal inference, where it is called DAG structure learning or causal discovery.
- Applications are numerous and include fields like bioinformatics:

Example 6 (Protein Signaling Network) The abundances of 14 different phosophoproteins and phospholipids in primary human immune system cells were measured in 1755 individuals cells. Treating the joint distribution of these 14 different molecules as a multivariate normal distribution we can apply one of the causal discovery algorithms that we will see in this class to estimate a DAG to which the data-generating distribution is Markov:

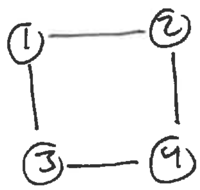


(PC Algorithm, $\alpha = 0.05$)

- The causal discovery algorithm used here relies on Theorem ④, using the characterization of the Markov property given in (b).
- To prove this theorem, and develop such algorithms, we will first prove similar theorems for <u>undirected graphical models</u>, which similarly provide useful representations of joint distributions;

Definition ③ A distribution $\underline{X} = [X_1, \dots, X_d]^T$ is <u>Markov</u> to an <u>undirected</u> graph $G = ([d], E)$ if $X_i \perp\!\!\!\perp X_j \mid \underline{X}_{[d] \setminus \{i,j\}}$ $\forall \{i,j\} \notin E$.

Example ⑦ $\underline{X} = [X_1, X_2, X_3, X_4]^T$ is Markov to



if $X_1 \perp\!\!\!\perp X_4 \mid X_2, X_3$ and $X_2 \perp\!\!\!\perp X_3 \mid X_1, X_4$.

- We can similarly learn undirected graph representations of a distribution.

Example ⑦ Returning to our molecules $\underline{X} = [X_1, \dots, X_{14}]^T$ from Example ⑥, we assume again that $\underline{X} \sim N(\mu, \Sigma)$ for some unknown mean vector $\mu \in \mathbb{R}^{14}$ and positive definite covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$.

- Since we assume a normal model we know that

$$X_i \perp\!\!\!\perp X_j \mid \underline{X}_{[14] \setminus \{i,j\}} \iff \mathrm{Cov}[X_i, X_j \mid \underline{X}_{[14] \setminus \{i,j\}}] = 0.$$

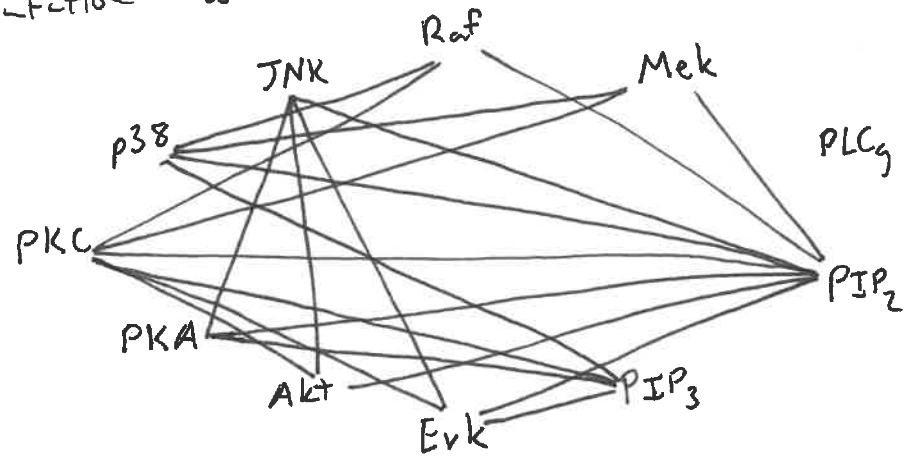
- To compute $\mathrm{Cov}[X_i, X_j \mid X_{[d]\setminus\{i,j\}}]$ we consider the covariance matrix for $[X_i, X_j]^T \mid X_{[14]\setminus\{i,j\}} = x_{[14]\setminus\{i,j\}} \sim N(\mu'_{\{i,j\}}, \Sigma'_{\{i,j\}})$

where

$$\mu'_{\{i,j\}} = \mu_{\{i,j\}} + \Sigma_{\{i,j\},[14]\setminus\{i,j\}} \Sigma^{-1}_{[14]\setminus\{i,j\},[14]\setminus\{i,j\}} (x_{[14]\setminus\{i,j\}} - \mu_{[14]\setminus\{i,j\}})$$

$$\Sigma'_{\{i,j\}} = \Sigma_{\{i,j\},\{i,j\}} - \Sigma_{\{i,j\},[14]\setminus\{i,j\}} \Sigma^{-1}_{[14]\setminus\{i,j\},[14]\setminus\{i,j\}} \Sigma_{[14]\setminus\{i,j\},\{i,j\}}.$$

- $\Sigma'_{\{i,j\}}$ is a 2×2 matrix whose off-diagonal entry is equal to $\mathrm{Cov}[X_i, X_j \mid X_{[d]\setminus\{i,j\}}]$.
- For each ~~pair~~ $\{i,j\} \in [14] \times [14]$ we can do a hypothesis test to check if this off-diagonal entry is (believably) zero.
- The result of these tests gives ~~use~~ the following undirected representation of relations in the data-generating distribution:

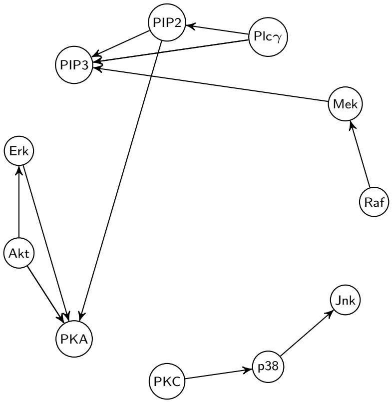


(Pearson Correlation, $\alpha = 0.0$

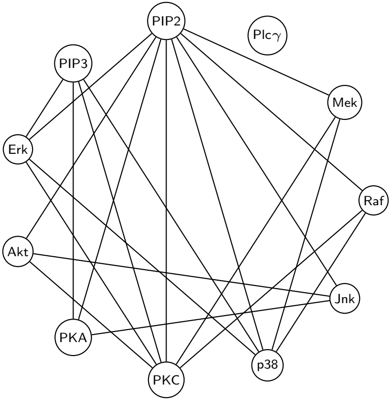
3 Inference with Graph Representations

- Lectures 5 and 6 will begin the discussion of how we can use a graph representation for inference (either a DAG or an undirected graph)
- We already saw a little example of how a graph may help us compute posterior distributions more efficiently.
- In these lectures we will look at exact inference algorithms for posterior computations that are motivated by the graph structure

- We will see that these inference algorithms have complexity bounds given by the structure of the graph.
- In module 2, you will study approximate inference algorithms that may be used even in cases where the complexity bounds we see in this module suggest that using the exact inference methods is infeasible.



Sachs data set estimated DAG;
PC algorithm ($\alpha = 0.05$).



Sachs data set estimated undirected graph;
Pairwise CI testing with Pearson Correlation Tests ($\alpha = 0.05$).