Outline

1. Conditional independence
2. Undirected graphical models
3. Directed acyclic graphs
4. Faithfulness
Definition: A graph $\mathcal{G} = (V, E)$, $V = \{1, \ldots, p\}$ is a set of vertices (or nodes) and $E \subset V \times V$ is a set of edges.

- Undirected edge $i - j$: $(i, j) \in E \iff (j, i) \in E$.
- Directed edge $i \rightarrow j$: $(i, j) \in E \Rightarrow (j, i) \notin E$.
- Associate $V$ to random variables $X_i$ ($i = 1, \ldots, p$) with joint distribution $\mathbb{P}$. Then $(\mathcal{G}, \mathbb{P})$ is called a graphical model. Often use node $i$ and $X_i$ interchangeably.
- Use graph separation to represent conditional independence among $X_1, \ldots, X_p$. 
Conditional independence

Definition: If $X, Y, Z$ are three random variables with a joint distribution $\mathbb{P}$, we say $X \perp Y \mid Z$ if $\mathbb{P}(X \in A \mid Y, Z)$ is a function of $Z$ only for any measurable set $A$.

If they admit a joint density (or mass function) $f$, then

$$X \perp Y \mid Z \iff f_{XY\mid Z}(x, y \mid z) = f_{X\mid Z}(x \mid z)f_{Y\mid Z}(y \mid z).$$

Other equivalent conditions ($f$ as a generic symbol for densities):

- $f(x, y, z) = f(x, z)f(y, z)/f(z)$.
- $f(x \mid y, z) = f(x \mid z)$.
- $f(x, z \mid y) = f(x \mid z)f(z \mid y)$.
- $f(x, y, z) = h(x, z)k(y, z)$ for some $h, k$.
- $f(x, y, z) = f(x \mid z)f(y, z)$. 

Zhou, Q Graphical Models 3/29
Graphoid axioms (Pearl (1988), §3.1.2.)

CI statement defines a ternary relation: \( \langle X, Y \mid Z \rangle \) for \( X \perp Y \mid Z \). Suppose \( X, Y, Z, W \) are disjoint subsets of random variables from a joint distribution \( \mathbb{P} \). Then the CI relation satisfies

1. **(C1) symmetry:** \( \langle X, Y \mid Z \rangle \Rightarrow \langle Y, X \mid Z \rangle \);
2. **(C2) decomposition:** \( \langle X, YW \mid Z \rangle \Rightarrow \langle X, Y \mid Z \rangle \);
3. **(C3) weak union:** \( \langle X, YW \mid Z \rangle \Rightarrow \langle X, Y \mid ZW \rangle \);
4. **(C4) contraction:** \( \langle X, Y \mid Z \rangle \& \langle X, W \mid ZY \rangle \Rightarrow \langle X, YW \mid Z \rangle \).

If the joint density of \( \mathbb{P} \) is positive and continuous wrt a product measure, then

1. **(C5) intersection:** \( \langle X, Y \mid ZW \rangle \& \langle X, W \mid ZY \rangle \Rightarrow \langle X, YW \mid Z \rangle \).

In the above, \( YW \equiv Y \cup W \).
Conditional independence

Any ternary relation $\langle A, B | C \rangle$ that satisfies (C1) to (C4) is called a **semi-graphoid**. If (C5) also holds, then it is called a **graphoid**.

**Examples of graphoid:**

1. **Conditional independence of $P$** (positive and continuous).
2. **Graph separation in undirected graph:** $\langle X, Y | Z \rangle$ means nodes $Z$ separate $X$ and $Y$, i.e. $X \perp Z \perp Y$.
3. **Partial orthogonality:** Let $X, Y, Z$ be disjoint sets of linearly independent vectors in $\mathbb{R}^n$. $\langle X, Y | Z \rangle$ means $P_Z^\perp X$ is orthogonal to $P_Z^\perp Y$. Here $P_Z^\perp X = (I_n - P_Z)X$ is the residual after projecting $X$ onto span($Z$).

Graph separation provides an intuitive graphical interpretation for the CI axioms.
Undirected graphical models


Terminology for undirected graph $G = (V, E)$

- $i$ and $j$ are neighbors if $(i, j) \in E$; $\text{ne}(i)$ denotes the set of neighbors of $i$.

- A path of length $n$ from $i$ to $j$ is a sequence $a_0 = i, \ldots, a_n = j$ of distinct vertices so that $(a_{k-1}, a_k) \in E$ for all $k = 1, \ldots, n$.

- A subset $C \subset V$ separates $a$ and $b$ if all paths from $a$ to $b$ intersect $C$.

- $C$ separates $A$ and $B$ if $C$ separates $a$ and $b$ for every $a \in A$ and $b \in B$. Write $A - C - B$. 
Undirected graphical models

Markov properties on undirected graphs

Consider undirected graphical model \((G, \mathbb{P})\). We say \(\mathbb{P}\) satisfies

- (P) the pairwise Markov property wrt \(G\) if

\[(i, j) \notin E \Rightarrow i \perp j \mid V \setminus \{i, j\} := [V]_{ij};\]

- (L) the local Markov property wrt \(G\) if

\[(i, j) \notin E \Rightarrow i \perp j \mid \text{ne}(i);\]

- (G) the global Markov property wrt \(G\) if

\[A - C - B \Rightarrow A \perp B \mid C;\]
Factorization via cliques

- Complete subset and clique: A subset of $C \subset V$ is complete if the subgraph on $C$ is complete. A complete subset that is maximal (wrt $\subset$) is called a clique.

- (F) Factorization: $P$ factorizes according to $G$ if for every clique $A$, there exists $\psi_A(x_A) \geq 0$, such that the joint density of $P$ has the form

$$f(x) = \prod_{A \in C} \psi_A(x_A),$$

where $C$ is the set of cliques of $G$.

- Relations: (F) $\Rightarrow$ (G) $\Rightarrow$ (L) $\Rightarrow$ (P).

Examples.
When does $(F) \iff (G) \iff (L) \iff (P)$?

**Theorem 1**

If $\mathbb{P}$ has a positive and continuous density $f$ with respect to a product measure, then $(F) \iff (P)$.

- **Product measure**: (1) $X_j \in \mathbb{R}$, use Lebesgue measure; (2) $X_j$ finite discrete, use counting measure.
- **Conclusion implies $(F) \iff (G) \iff (L) \iff (P)$**.
- **Counter example**. Let $p = 5$, $X_1, X_5 \sim iid \text{ Bern}(0.5)$, $X_2 = X_1$, $X_4 = X_5$, and $X_3 = X_2X_4$. This defines $\mathbb{P}$. Let $G$ be a chain $E = \{(i, i+1) : i = 1, \ldots, 4\}$. Then $(L)$ holds but not $(G)$. Because density (probability mass function) is not positive on all possible values of $X_i$’s.

$(L)$: $X_2 \perp X_4 \mid (X_1, X_3)$ true; $(G)$: $X_2 \perp X_4 \mid X_3$ false!
Conditional independence graph (CIG).

- **Definition**: A CIG is a graphical model \((\mathcal{G}, \mathbb{P})\) such that \((P)\) holds. That is,

\[(i, j) \notin E \Rightarrow i \perp j \mid V \setminus \{i, j\} := [V]_{ij}.\]

- Sparser graph \(\mathcal{G}\) implies more conditional independence (CI) relations.
- One can always choose the minimal \(\mathcal{G}\) such that \((P)\) holds to be the CIG, i.e., replace \(\Rightarrow\) by \(\Leftrightarrow\).
- Estimate the structure of \(\mathcal{G}\) to detect CI relations, assuming we have observed iid data from \(\mathbb{P}\).
Gaussian graphical models (GGMs)

A CIG with $\mathbb{P} = \mathcal{N}_p(0, \Sigma)$, $\Sigma > 0$ (positive definite).

**Lemma 1**

Suppose $(X_1, \ldots, X_p) \sim \mathcal{N}_p(0, \Sigma)$ with $\Sigma > 0$ and let

$\Theta = (\theta_{jk})_{p \times p} = \Sigma^{-1}$. Then

$\theta_{jk} = 0 \iff X_j \perp X_k | X_{-\{j,k\}}$. \hspace{1cm} (1)

- $\Theta$ is called the precision matrix.
- (1) shows that GGM is constructed as

$\theta_{jk} = 0 \iff (j, k) \notin E$. \hspace{1cm} (2)
Partial correlation and neighborhood regression

- Partial correlation between $j$ and $k$ given $[V]_{jk}$:
  \[ \rho_{jk} = -\frac{\theta_{jk}}{\sqrt{\theta_{jj}\theta_{kk}}}. \]
  Correlation calculated from $\Sigma_{(j,k)\mid [V]_{jk}} = \text{Var}(j, k \mid [V]_{jk})$.

- Neighborhood regression, regress $X_j$ on $X_{\text{-}j}$:
  \[ X_j = \sum_{i \neq j} \beta_{ij} X_i + \varepsilon_j. \] (3)

Then $\beta_{kj} = -\frac{\theta_{jk}}{\theta_{jj}}$. (By symmetry $\beta_{jk} = -\frac{\theta_{kj}}{\theta_{kk}}$.)

- Thus, we have
  \[ (j, k) \notin E \iff \theta_{jk} = 0 \iff \rho_{jk} = 0 \iff \beta_{kj} = \beta_{jk} = 0. \] (4)
Learning GGMs: Given \( x_i \sim_{iid} \mathcal{N}_p(0, \Sigma), \ i = 1, \ldots, n \), estimate the structure of \( G \Leftrightarrow \text{supp}(\Theta) = \{(j, k): \theta_{jk} \neq 0\} \).

Also called covariance selection (Dempster 1972).

- **Log-likelihood**
  \[
  \ell(\Sigma) = -\frac{n}{2} \log \det(\Sigma) - \frac{1}{2} \text{tr}(S\Sigma^{-1}),
  \]
  where \( S = \sum_i x_i x_i^T \) is a \( p \times p \) matrix (sufficient statistic).

- \( \hat{\Sigma}^{MLE} = S/n \) (always exists).

- If \( n > p \), inverte \( \hat{\Sigma}^{MLE} \Rightarrow \hat{\Theta}^{MLE} = (\hat{\Sigma}^{MLE})^{-1} \).

Then obtain \( \hat{G} \) by thresholding:

\[
\hat{E} = \{(j, k): |\hat{\theta}_{jk}^{MLE}| > \tau\}.
\]
Regularized estimation under $\ell_1$ penalty (Yuan and Lin 2007; Friedman et al. 2008; Banerjee et al. 2008)

- Element-wise $\ell_1$ norm $\|\Theta\|_1 := \sum_{j<k} |\theta_{jk}|$.
- $\ell_1$ regularized estimate $\hat{\Theta} = \text{argmin}_{\Theta > 0} f(\Theta)$,

$$f(\Theta) = -\frac{2}{n} \ell(\Theta^{-1}) + \lambda \|\Theta\|_1$$

$$= - \log \det(\Theta) + \text{tr}(\hat{\Sigma}^{MLE} \Theta) + \lambda \|\Theta\|_1.$$  

- $f$ is convex, efficient algorithm.
- Well-defined for $p > n$.
- Sparse solution, $\hat{\theta}_{jk} = 0$ for some $(j, k)$. 

Undirected graphical models
Estimate $\mathcal{G}$ from $\hat{\Theta}$

- $\hat{E} = \{(j, k) : \hat{\theta}_{jk} \neq 0\}$, but needs very strong assumptions (irrepresentability) for $\mathbb{P}(\hat{E} = E_0) \to 1$.

- Thresholding $\hat{\Theta}$: $\hat{E} = \{(j, k) : |\hat{\theta}_{jk}| > \tau\}$. Weaker assumptions (RE, beta-min) for $\mathbb{P}(\hat{E} = E_0) \to 1$.

Choosing $\lambda$ by cross-validation, $\lambda_{CV}^*$, then $\mathbb{P}(\hat{E}(\lambda_{CV}^*) \supset E_0) \to 1$ under certain conditions (RE, beta-min).
Undirected graphical models

Estimate $G$ by neighborhood regression (Meinshausen and Bühlmann 2006)

- Apply model selection (e.g. lasso) for each neighborhood regression (3) ⇒ $\hat{\beta}_{jk}$ ($j, k = 1, \ldots, p$).
- Combine results to define $\hat{G}$, e.g.,

$$\hat{E} = \{(j, k) : \hat{\beta}_{jk} \neq 0, \hat{\beta}_{kj} \neq 0\}.$$

- Approximate $\hat{\Theta}$ if lasso is used in neighborhood regression.
Terminology for directed acyclic graph (DAG) $\mathcal{G} = (V, E)$

- If $i \rightarrow j$, then $i$ is a parent of $j$ and $j$ is a child of $i$; $\text{pa}(j)$ is the set of parents of $j$; $\text{ch}(i)$ is the set of children of $i$.
- If there is a path from $i$ to $j$, we say $i$ leads to $j$ and write $i \rightarrow j$.
  The ancestors $\text{an}(j) = \{i : i \rightarrow j\}$.
  The descendants $\text{de}(i) = \{j : i \rightarrow j\}$.
  The non-descendants $\text{nd}(i) = V \setminus (\text{de}(i) \cup \{i\})$.
- A *chain* of length $n$ from $i$ to $j$ is a sequence $a_0 = i, \ldots, a_n = j$ of distinct vertices so that $a_{k-1} \rightarrow a_k$ or $a_k \rightarrow a_{k-1}$ for all $k = 1, \ldots, n$. 
Directed acyclic graphs

- **$d$-separation:** A chain $\pi$ from $a$ to $b$ is said to be *blocked* by $S \subset V$, if the chain contains a vertex $\gamma$ such that either (1) or (2) holds:
  1. $\gamma \in S$ and the arrows of $\pi$ do *not* meet at $\gamma$ ($i \to \gamma \to j$).
  2. $\gamma \cup \text{de}(\gamma)$ not in $S$ and arrows of $\pi$ meet at $\gamma$ ($i \to \gamma \leftarrow j$).

Two subsets $A$ and $B$ are $d$-separated by $S$ is all chains from $A$ to $B$ are blocked by $S$.

- A topological sort of $G$ is an ordering $\sigma$, i.e., a permutation of $\{1, \ldots, p\}$, such that $j \in \text{an}(i)$ implies $j \prec i$ in $\sigma$. Due to acyclicity, every DAG has at least one sort.

- **Example $G$:** $1 \to 2 \to 3 \leftarrow 4$.
  $\{2\}$ $d$-separates 1 and 4; $\emptyset$ $d$-separates 1 and 4.
  $\sigma = (1, 2, 4, 3)$ or $(4, 1, 2, 3)$ or $(1, 4, 2, 3)$ are topological sorts.
Directed acyclic graphs

Markov properties on DAGs: We say a joint distribution $\mathbb{P}$

- (DF) admits a recursive factorization according to $\mathcal{G}$ if $\mathbb{P}$ has a density $f$ such that

$$f(x) = \prod_{j \in V} f_j(x_j \mid \text{pa}(j)), \quad (5)$$

where $f_j$ is the density for $[j \mid \text{pa}(j)]$.

- (DG) satisfies the directed global Markov property if

$$S \; d\text{-separates } A \text{ and } B \Rightarrow A \perp B \mid S;$$

- (DL) satisfies the directed local Markov property if

$$i \perp \text{nd}(i) \mid \text{pa}(i).$$

- (DP) satisfies the directed pairwise Markov property if for any $(i, j) \notin E$ with $j \in \text{nd}(i)$, $i \perp j \mid \text{nd}(i) \setminus \{j\}$. 

Directed acyclic graphs

Relations: (DF) ⇒ (DG) ⇒ (DL) ⇒ (DP).

**Theorem 2**

If $\mathbb{P}$ has a density $f$ with respect to a product measure, then (DF), (DG), and (DL) are equivalent.

Markov equivalence: Two DAGs are called Markov equivalent if they induce the same set of CI restrictions.

$Leftrightarrow$ Same skeleton and same $v$-structures (Verma and Pearl 1990).

Connections to Markov properties on undirected graphs:

- Moral graph $G^m$: add edges between all parents of a node in a DAG $G$ and delete directions.
- If $\mathbb{P}$ admits a recursive factorization according to $G$, then it factorizes according to $G^m$.
  That is, (DF) wrt $G$ ⇒ (F) wrt $G^m$ ⇒ (G), (L), (P) wrt $G^m$. 
Definition of Bayesian networks: Given $\mathbb{P}$ with density $f$ and an ordering $(\sigma(1), \ldots, \sigma(p))$, we factorize $f$

$$f(x) = \prod_{j=1}^{p} f(x_{\sigma(j)} \mid x_{\sigma(1)}, \ldots, x_{\sigma(j-1)})$$

$$= \prod_{j=1}^{p} f(x_{\sigma(j)} \mid x_{A_j}), \quad (6)$$

where $A_j \subseteq \{\sigma(1), \ldots, \sigma(j-1)\}$ is the minimum subset such that (6) holds. Then the DAG $G$ with $\text{pa}(\sigma(j)) = A_j$ for all $j \in V$ is a Bayesian network of $\mathbb{P}$.

CI: If $G$ is a BN of $\mathbb{P}$, then (DF) holds, so (DG), (DL), (DP) also hold.

Examples: Markov chains, HMMs, etc.
Parameterization: Given $\mathcal{G}$, to parameterize $[X_j \mid \text{pa}(j)]$ as in (5).

(1) Gaussian BNs

- Structural equations:

$$X_j = \sum_{i \in \text{pa}(j)} \beta_{ij} X_i + \varepsilon_j, \quad j = 1, \ldots, p.$$ 

Assume $\varepsilon_j \sim \mathcal{N}(0, \omega_j^2)$ and $\varepsilon_j \perp \text{pa}(j)$.

- Put $B = (\beta_{ij})$ and $\Omega = \text{diag}(\omega_1^2, \ldots, \omega_p^2)$. Then

$$X = B^T X + \varepsilon, \quad \varepsilon \sim \mathcal{N}_p(0, \Omega).$$

$\Rightarrow X \sim \mathcal{N}_p(0, \Theta^{-1})$, where $\Theta = (I_p - B)\Omega^{-1}(I_p - B)^T$ (Cholesky decomposition of $\Theta$); see van de Geer and Bühlmann (2013); Aragam and Zhou (2015).
(2) Discrete BNs

- Multinomial distribution: \( \theta_{km}^{(j)} = P(X_j = m \mid \text{pa}(j) = k) \). Parameter for \( X_j \) is a \( K \times M \) table:

\[
\left\{ \theta_{km}^{(j)} : \sum_m \theta_{km}^{(j)} = 1, \ k = 1, \ldots, K, \ m = 1, \ldots, M \right\}.
\]

\( K \): number of all possible combinations of \( \text{pa}(j) \). (Too many parameters if a node has many parents.)

- Multi-logit regression model (Gu et al. 2019): Use generalized linear model for \( [X_j \mid \text{pa}(j)] \).
Directed acyclic graphs

Structure learning

Given \( x_i \sim_{iid} \mathbb{P}, \, i = 1, \ldots, n \), estimate a BN \( \hat{G} \) for \( \mathbb{P} \). The sparser the \( \hat{G} \), the more CI relations learned from data.

- Score-based methods: Minimize a scoring function over DAGs; regularization to obtain sparse solutions.
- Constraint-based methods: Condition independence tests against \( X_i \perp X_j \mid X_S \) for all \( i, j, S \).
- Hybrid methods: First use constraint-based method to prune the search space, and then apply a score-based method to search for the optimal DAG.

See, e.g. Aragam and Zhou (2015) Section 1.2.
Directed acyclic graphs

Causal Bayesian networks

- Model causal relations among nodes: If $i \rightarrow j$, then $i$ is a causal parent of $j$.
- Causal relation defined by experimental intervention (Pearl 2000).
- If $\text{pa}(i)$ is fixed by intervention, then $i$ will not be affected by interventions on $V \setminus \{\text{pa}(i) \cup \{i\}\}$.
- If $j \in M$ are under intervention, then modify factorization

$$f(x) = \prod_{j \notin M} f_j(x_j | \text{pa}(j)) \prod_{j \in M} g_j(x_j), \quad (7)$$

where $g_j(\cdot)$ is the density of $X_j$ under intervention.
- $\ell_1$ regularization for learning causal BNs from intervention data (Fu and Zhou 2013).
Given a graphical model \((\mathcal{G}, P)\) where \(P\) satisfies, say \((G)\) or \((DG)\). Then graph separation \(\Rightarrow\) condition independence, but not \(\Leftarrow\). If \(P\) is faithful to \(\mathcal{G}\) then \(\Leftarrow\) holds as well. In this case, we have \(\Leftrightarrow\).

**Definition 1**

For a graphical model \((\mathcal{G}, P)\), we say the distribution \(P\) is faithful to the graph \(\mathcal{G}\) if for every triple of disjoint sets \(A, B, S \subseteq V\),

\[
A \perp B \mid S \iff S \text{ separates (d-separates) } A \text{ and } B.
\]

How likely is \(P\) faithful?

Gaussian graphs (undirected or DAGs), \(P\) is Gaussian.

- Given \(\mathcal{G}\), almost all parameter values will define a faithful \(P\).
- Counterexamples: Diamond graph, v-shape DAG.


References III


