Chapter 6
Introduction to Graphical Models

Qing Zhou

UCLA Department of Statistics

Stats 201C Advanced Modeling and Inference
Lecture Notes

Outline

Conditional independence

Definition: If X, Y, Z are three random variables, 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

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

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

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

Conditional independence

CI in statistical inference (Dawid 1979):

- Sufficient and ancillary statistics: Suppose $X \mid \Theta \sim P_\Theta$.
  1. $T = T(X)$ is a sufficient statistic for Θ if $X \perp \Theta \mid T$.
  2. $S = S(X)$ is an ancillary statistic if $S \perp \Theta$.

  Example: $X = (X_1, \ldots, X_n) \mid \mu, \sigma^2 \sim \mathcal{N}(\mu, \sigma^2)$. Then
  $T_1 = \sum_i X_i$ is sufficient for μ;
  $T_2 = \sum_i (X_i - \bar{X})^2$ is ancillary for μ.
- Model selection: $Y = X\beta + \varepsilon$. If $\text{supp}(\beta) = S$, then $Y \perp (X \setminus X_S) \mid X_S$.
- Parameter identification: $X \mid \Theta, \Phi \sim P_{(\Theta,\Phi)}$. If $X \perp \Phi \mid \Theta$, then Φ is not identifiable.
  Example: Gaussian linear model $Y = X\beta + \varepsilon$ with X not having full column rank. Let $\Theta = X\beta \in \text{col}(X)$ and $\Phi = \beta - X^- X\beta$ (X^- is a g-inverse of X; $XX^-X = X$). Then $X\Phi = 0$, i.e. $\Phi \in \text{null}(X)$. Thus $Y \perp \Phi \mid (\Theta, \sigma^2)$, i.e. Φ is not identifiable. Note $\dim(\Theta) + \dim(\Phi) = \dim(\beta)$.

Conditional independence

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

(C1) symmetry: $\langle X, Y \mid Z \rangle \Rightarrow \langle Y, X \mid Z \rangle$;

(C2) decomposition: $\langle X, YW \mid Z \rangle \Rightarrow \langle X, Y \mid Z \rangle$;

(C3) weak union: $\langle X, YW \mid Z \rangle \Rightarrow \langle X, Y \mid ZW \rangle$;

(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 $\mathbb{P}$ wrt a product measure is positive and continuous, then

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

In the above, $YW := Y \cup W$.

Conditional independence

Any ternary relation $\langle A, B \mid C \rangle$ 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 $\mathbb{P}$ (positive and continous).
2. Graph separation in undirected graph: $\langle X, Y \mid Z \rangle$ means nodes Z separate X and Y, i.e. $X - Z - Y$.
3. Partial orthogonality: Let X, Y, Z be disjoint sets of linearly independent vectors in $\mathbb{R}^n$. $\langle X, Y \mid Z \rangle$ 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.

Conditional independence

Example application of CI in causal inference:

- Treatment X, outcome Y. Let I indicates each individual, $I = 1, \ldots, n$. Want to test if $Y \perp X \mid I$ (untestable).
- Suppose $Z = Z(I)$ is a set of sufficient covariates such that $Y \perp I \mid (X, Z)$. Then

$$Y \perp X \mid I \Leftrightarrow Y \perp X \mid Z \text{ (testable based on data)} \quad (1)$$

- Proof outline:
  Note $Y \perp X \mid I \Leftrightarrow Y \perp X \mid (I, Z)$ because $Z = Z(I)$.
  $\Leftarrow$: Sufficient set and RHS of (1) imply $Y \perp (I, X) \mid Z$ by (C4) and thus $Y \perp X \mid (I, Z)$ by (C3).
  $\Rightarrow$: Sufficient set and LHS ($Y \perp X \mid (I, Z)$) imply $Y \perp (X, I) \mid Z$ by (C5) and thus $Y \perp X \mid Z$ by (C2).

Conditional independence

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 $i - j$: $(i, j) \in E \Leftrightarrow (j, i) \in E$.
- Directed edge $i \to j$: $(i, j) \in E \Rightarrow (j, i) \notin E$.
- 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 $X_1, \ldots, X_p$.

Undirected graphical models

Reference: Lauritzen (1996), chapters 2 and 3.

Terminology for undirected graph $\mathcal{G} = (V, E)$

- 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 $(\mathcal{G}, \mathbb{P})$. We say $\mathbb{P}$ satisfies

- (P) the pairwise Markov property wrt $\mathcal{G}$ if

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

- (L) the local Markov property wrt $\mathcal{G}$ if

$$(i, j) \notin E \Rightarrow i \perp j \mid \mathrm{ne}(i);$$

- (G) the global Markov property wrt $\mathcal{G}$ if

$$A - C - B \Rightarrow A \perp B \mid C;$$

Undirected graphical models

Factorization via cliques

- Clique and maximal clique: A subset of $C \subset V$ is a clique if the subgraph on C is complete. A clique that is maximal (wrt $\subset$) is called a maximal clique.
- (F) Factorization: $\mathbb{P}$ factorizes according to $\mathcal{G}$ if for every maximal clique A, there exists $\psi_A(x_A) \geq 0$, such that the joint density of $\mathbb{P}$ has the form

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

  where $\mathcal{C}$ is the set of cliques of $\mathcal{G}$.
- Relations: (F) $\Rightarrow$ (G) $\Rightarrow$ (L) $\Rightarrow$ (P).

Examples.

Undirected graphical models

When does (F) $\Leftrightarrow$ (G) $\Leftrightarrow$ (L) $\Leftrightarrow$ (P)?

Theorem 1

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

- Product measure: (1) $X_j \in \mathbb{R}$, use Lebesgue measure; (2) X_j finite discrete, use counting measure.
- Conclusion implies (F) $\Leftrightarrow$ (G) $\Leftrightarrow$ (L) $\Leftrightarrow$ (P).
- Counter example. Let $p = 5$, $X_1, X_5 \sim_{iid}$ Bern(0.5), $X_2 = X_1$, $X_4 = X_5$, and $X_3 = X_2 X_4$. This defines $\mathbb{P}$. Let $\mathcal{G}$ 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!

Undirected graphical models

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 $\mathbb{P}$.

Undirected graphical models

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

$$\theta_{jk} = 0 \Leftrightarrow X_j \perp X_k \mid X_{-\{j,k\}}. \tag{2}$$

- Θ is called the precision matrix.
- (2) shows that GGM is constructed as

$$\theta_{jk} = 0 \Leftrightarrow (j, k) \notin E. \tag{3}$$

Undirected graphical models

Partial correlation and neighborhood regression

- Partial correlation between j and k given $[V]_{jk}$:
  $\rho_{jk} = -\theta_{jk}/\sqrt{\theta_{jj}\theta_{kk}}$.
  Correlation calculated from $\Sigma_{(j,k)|[V]_{jk}} = \mathrm{Var}(j, k \mid [V]_{jk})$.
- Neighborhood regression, regress X_j on X_{-j}:

$$X_j = \sum_{i \neq j} \beta_{ij} X_i + \varepsilon_j. \tag{4}$$

  Then $\beta_{kj} = -\theta_{jk}/\theta_{jj}$. (By symmetry $\beta_{jk} = -\theta_{kj}/\theta_{kk}$.)
- Thus, we have

$$(j, k) \notin E \Leftrightarrow \theta_{jk} = 0 \Leftrightarrow \rho_{jk} = 0 \Leftrightarrow \beta_{kj} = \beta_{jk} = 0. \tag{5}$$

Undirected graphical models

Learning GGMs: Given $x_i \sim_{iid} \mathcal{N}_p(0, \Sigma)$, $i = 1, \ldots, n$, estimate

$$\text{the structure of } \mathcal{G} \Leftrightarrow \text{supp}(\Theta) = \{(j,k) : \theta_{jk} \neq 0\}.$$

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 $S = \sum_i x_i x_i^{\mathsf{T}}$ is a $p \times p$ matrix (sufficient statistic).
- $\hat{\Sigma}^{\text{MLE}} = S/n$ (always exists).
- If $n > p$, inverte $\hat{\Sigma}^{\text{MLE}} \Rightarrow \hat{\Theta}^{\text{MLE}} = (\hat{\Sigma}^{\text{MLE}})^{-1}$.
  Then obtain $\widehat{\mathcal{G}}$ by thresholding: $\widehat{E} = \{(j,k) : |\hat{\theta}_{jk}^{\text{MLE}}| > \tau\}$.

Undirected graphical models

Regularized estimation under ℓ_1 penalty (Yuan and Lin 2007; Friedman et al. 2008; Banerjee et al. 2008)

- Element-wise ℓ_1 norm $\|\Theta\|_1 := \sum_{j<k} |\theta_{jk}|$.
- ℓ_1 regularized estimate $\hat{\Theta} = \text{argmin}_{\Theta > 0} f(\Theta)$,

$$f(\Theta) = -\frac{2}{n}\ell(\Theta^{-1}) + \lambda\|\Theta\|_1$$
$$= -\log\det(\Theta) + \text{tr}(\hat{\Sigma}^{\text{MLE}}\Theta) + \lambda\|\Theta\|_1.$$

- 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 $\hat{\Theta}$

- $\widehat{E} = \{(j,k) : \hat{\theta}_{jk} \neq 0\}$, but needs very strong assumptions (irrepresentability) for $\mathbb{P}(\widehat{E} = E_0) \to 1$.
- Thresholding $\hat{\Theta}$: $\widehat{E} = \{(j,k) : |\hat{\theta}_{jk}| > \tau\}$. Weaker assumptions (RE, beta-min) for $\mathbb{P}(\widehat{E} = E_0) \to 1$.

Choosing λ by cross-validation, λ^*_{CV}, then $\mathbb{P}(\widehat{E}(\lambda^*_{CV}) \supset E_0) \to 1$ under certain conditions (RE, beta-min).

Estimate $\mathcal{G}$ by neighborhood regression (Meinshausen and Bühlmann 2006)

- Apply model selection (e.g. lasso) for each neighborhood regression (4) $\Rightarrow \hat{\beta}_{jk}$ $(j, k = 1, \ldots, p)$.
- Combine results to define $\widehat{\mathcal{G}}$, e.g.,

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

- Approximate $\hat{\Theta}$ if lasso is used in neighborhood regression.

Directed acyclic graphs

Terminology for directed acyclic graph (DAG) $\mathcal{G} = (V, E)$

- If $i \to j$, 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
  $i \longmapsto j$.
  The ancestors $\text{an}(j) = \{i : i \longmapsto j\}$.
  The descendants $\text{de}(i) = \{j : i \longmapsto j\}$.
  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 $a_{k-1} \to a_k$ or
  $a_k \to a_{k-1}$ for all $k = 1, \ldots, n$.

Directed acyclic graphs

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

  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 $\mathcal{G}$ is an ordering σ, i.e., a permutation of $\{1, \ldots, p\}$, such that $j \in \text{an}(i)$ implies $j \prec i$ in σ. Due to acyclicity, every DAG has at least one sort.
- Example $\mathcal{G}$: $\quad 1 \to 2 \to 3 \leftarrow 4$.
  $\{2\}$ *d*-separates 1 and 4; $\varnothing$ *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

$$f(x) = \prod_{j \in V} f_j(x_j \mid \mathrm{pa}(j)), \tag{6}$$

  where f_j is the density for $[j \mid \mathrm{pa}(j)]$.
- (DG) satisfies the directed global Markov property if

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

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

Directed acyclic graphs

Relations: (DF) $\Rightarrow$ (DG) $\Rightarrow$ (DL) $\Rightarrow$ (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 $\mathcal{G}^m$: add edges between all parents of a node in a DAG $\mathcal{G}$ and delete directions.
- If $\mathbb{P}$ admits a recursive factorization according to $\mathcal{G}$, then it factorizes according to $\mathcal{G}^m$.
  That is, (DF) wrt $\mathcal{G} \Rightarrow$ (F) wrt $\mathcal{G}^m \Rightarrow$ (G), (L), (P) wrt $\mathcal{G}^m$.

Directed acyclic graphs

- Definition of Bayesian networks: Given $\mathbb{P}$ with density f and an ordering $(\sigma(1), \ldots, \sigma(p))$, we factorize f

$$
\begin{aligned}
f(x) &= \prod_{j=1}^{p} f(x_{\sigma(j)} \mid x_{\sigma(1)}, \ldots, x_{\sigma(j-1)}) \\
&= \prod_{j=1}^{p} f(x_{\sigma(j)} \mid x_{A_j}), \qquad (7)
\end{aligned}
$$

  where $A_j \subset \{\sigma(1), \ldots, \sigma(j-1)\}$ is the minimum subset such that (7) holds. Then the DAG $\mathcal{G}$ with $\mathrm{pa}(\sigma(j)) = A_j$ for all $j \in V$ is a Bayesian network of $\mathbb{P}$.
- CI: If $\mathcal{G}$ is a BN of $\mathbb{P}$, then (DF) holds, so (DG), (DL), (DP) also hold.
- Examples: Markov chains, HMMs, etc.

Directed acyclic graphs

Parameterization: Given $\mathcal{G}$, to parameterize $[X_j \mid \text{pa}(j)]$ as in (6).

(1) Gaussian BNs

- Linear structural equation model (SEM):

$$X_j = \sum_{i \in \text{pa}(j)} \beta_{ij} X_i + \varepsilon_j, \qquad j = 1, \dots, p. \tag{8}$$

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

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

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

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

Directed acyclic graphs

(2) Discrete BNs

- Multinomial distribution: $\theta_{km}^{(j)} = \mathbb{P}(X_j = m \mid \text{pa}(j) = k)$. Parameter for $[X_j \mid \text{pa}(j)]$ 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 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 $\widehat{\mathcal{G}}$ for $\mathbb{P}$.
The sparser the $\widehat{\mathcal{G}}$, 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 DAG model:

- Model causal relations among nodes: If $i \to j$, then i is a causal parent (direct cause) of j.
- Causal relation defined by experimental intervention (Pearl 2000): Force X to some fixed value x, which we denote by $do(X = x)$ or $do(x)$ for short.
- Effect of $do(x_i)$: to replace the SEM for X_i by $X_i = x_i$ and substitute $X_i = x_i$ in the other SEMs for $X_j, j \neq i$. See Eq (8).
- The causal effect of X on Y is defined by the mapping $x \mapsto \mathbb{P}[Y \mid do(X = x)] \equiv \mathbb{P}(Y \mid do(x))$.
  1. linear SEM: Causal effect $\frac{\partial \mathbb{E}(Y|do(x))}{\partial x}$.
  2. Treatment $(X = 1)$ vs control $(X = 0)$: Causal effect $\mathbb{E}(Y \mid do(X = 1)) - \mathbb{E}(Y \mid do(X = 0))$.

Faithfulness

Given a graphical model $(\mathcal{G}, \mathbb{P})$ where $\mathbb{P}$ satisfies, say (G) or (DG). Then graph separation $\Rightarrow$ condition independence, but not $\Leftarrow$. If $\mathbb{P}$ is faithful to $\mathcal{G}$ then $\Leftarrow$ holds as well. In this case, we have $\Leftrightarrow$.

Definition 1

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

$$A \perp B \mid S \Leftrightarrow S \text{ separates } (d\text{-separates}) \ A \text{ and } B.$$

How likely is $\mathbb{P}$ faithful?

Example: Gaussian graphs (undirected or DAGs), $\mathbb{P}$ is Gaussian.

- Given $\mathcal{G}$, almost all parameter values will define a faithful $\mathbb{P}$.
- Counterexamples: The parameters, Θ or (β_{ij}), satisfy additional equality constraints that define CI in $\mathbb{P}$ not implied by any separation in $\mathcal{G}$.

References I

Bryon Aragam and Qing Zhou. Concave penalized estimation of sparse Gaussian Bayesian networks. *Journal of Machine Learning Research*, 16:2273–2328, 2015.

Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516, 2008.

A.P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 41:1–31, 1979.

Arthur P Dempster. Covariance selection. *Biometrics*, 28(1): 157–175, 1972.

References II

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the Graphical Lasso. *Biostatistics*, 9(3):432–441, 2008.

Fei Fu and Qing Zhou. Learning sparse causal Gaussian networks with experimental intervention: Regularization and coordinate descent. *Journal of the American Statistical Association*, 108 (501):288–300, 2013.

Jiaying Gu, Fei Fu, and Qing Zhou. Penalized estimation of directed acyclic graphs from discrete data. *Statistics and Computing*, 29:161–176, 2019.

Steffen L. Lauritzen. *Graphical Models*. Oxford University Press, 1996. ISBN 0-19-852219-3.

References III

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

Judea Pearl. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann, 1988.

Judea Pearl. *Causality: Models, reasoning and inference*. Cambridge Univ Press, 2000.

Sara van de Geer and Peter Bühlmann. ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *The Annals of Statistics*, 41(2):536–567, 2013.

T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, pages 220–227, 1990.

Ming Yuan and Yi Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.