Directed Acyclic Graphs

Qing Zhou

UCLA Department of Statistics

Stats 212 Graphical Models
Lecture Notes

Outline

1. DAGs and terminology
2. *d*-separation
3. Markov properties
4. Parameterizations
5. Overview of topics
6. Chain graphs

DAGs and terminology

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

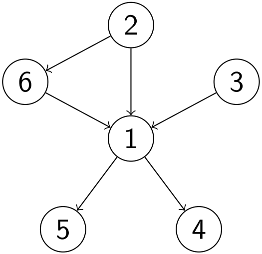
- $E = \{(i, j) : i \to j\}$ (all edges are directed).
- If $i \to j$, then i is a parent of j and j is a child of i; pa(j) is the set of parents of j; ch(i) is the set of children of i.
- A *path* of length n from i to j is a sequence $a_0 = i, \ldots, a_n = j$ of distinct vertices so that $(a_{k-1}, a_k) \in E$ for all $k = 1, \ldots, n$, i.e. $i \to a_1 \to \cdots \to a_{n-1} \to j$.
- An n-cycle is a path of length n with the modification that $i = j$. A cycle is directed if it contains a directed edge.
- DAG: (i) all edges are directed; (ii) has no directed cycles.

DAGs and terminology

- If there is a path from i to j, we say i leads to j and write $i \mapsto j$.
  The ancestors $\text{an}(j) = \{i : i \mapsto j\}$.
  The descendants $\text{de}(i) = \{j : i \mapsto j\}$.
  The non-descendants $\text{nd}(i) = V \setminus (\text{de}(i) \cup \{i\})$.
- A topological sort of $\mathcal{G}$ over p vertices 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.

DAGs and terminology

Example:

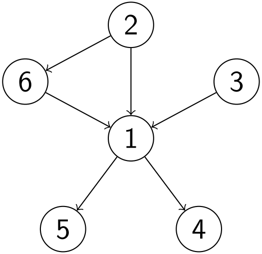


- $\text{pa}(1) = \{2, 3, 6\}$, $\text{ch}(1) = \{4, 5\}$.
- Path: $2 \to 6 \to 1 \to 4$, $3 \to 1 \to 5$.
  $2 \to 6 \to 1 \leftarrow 3$ is *not* a path.
- $\text{an}(4) = \{2, 6, 3, 1\}$
  $\text{de}(6) = \{1, 4, 5\}$, $\text{nd}(6) = \{2, 3\}$.
- topological sorts: $(2, 6, 3, 1, 4, 5)$,
  $(3, 2, 6, 1, 5, 4)$, etc.

d-separation

- 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 $k = 1, \ldots, n$. Example: $i \leftarrow a_1 \to a_2 \to \cdots \to a_{n-1} \leftarrow j$.
- 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 $i \leftarrow \gamma \to j$). ($\gamma$ is a non-collider.)
  2. $\gamma \cup \text{de}(\gamma)$ not in S and arrows of π meet at γ ($i \to \gamma \leftarrow j$). ($\gamma$ is a collider.)
- Two subsets A and B are d-separated by S is all chains from A to B are blocked by S.

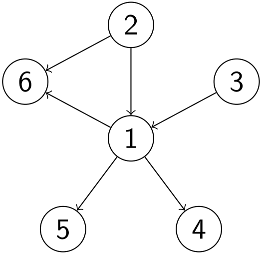
Example:



- chain $2 \to 6 \to 1 \to 4$ has no collider and is blocked by $\{1\}, \{6\}$, or $\{1, 6\}$.
- chain $2 \to 6 \to 1 \leftarrow 3$ has a collider (node 1), and thus is blocked by $\varnothing$. But this chain is *not* blocked by $\{1\}$ or any node in de(1)$= \{4, 5\}$, i.e. the chain is d-connected given $\{1\}, \{4\}$ or $\{5\}$.
- Find S to d-separate 2 and 4: $S = \{1\}$, $S = \{1, 6\}$.
- Find S to d-separate 3 and 6: $S = \varnothing$, $S = \{2\}$, $S \neq$ any subset of $\{1, 4, 5\}$.

d-separation

Example (flip the edge between 1 and 6)



Find S to d-separate 3 and 6:

1. To block $3 \to 1 \to 6$, must include $1 \in S$.
2. But 1 is a collider in $3 \to 1 \leftarrow 2 \to 6$, given node 1 this chain is d-connected.
3. Thus, to block $3 \to 1 \leftarrow 2 \to 6$, must include $2 \in S$.
4. $S = \{1, 2\}$ d-separates 3 and 6.

Markov properties

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

$$f(x) = \prod_{j \in V} f_j(x_j \mid \mathrm{pa}(j)), \tag{1}$$

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

- (DG) satisfies the directed global Markov property if for any disjoint (A, B, S),

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

Markov properties

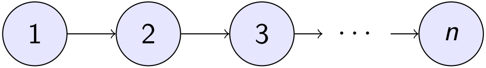
- (DL) satisfies the directed local Markov property if $i \perp \text{nd}(i) \mid \text{pa}(i)$ for all $i \in V$.
- (DP) satisfies the directed pairwise Markov property if for any $(i, j) \notin E$ with $j \in \text{nd}(i)$, $i \perp j \mid \text{nd}(i) \setminus \{j\}$.

Relations: (DF) $\Rightarrow$ (DG) $\Rightarrow$ (DL) $\Rightarrow$ (DP).

Theorem 1

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

Example: Markov chain



$\mathrm{pa}(i) = i - 1, i = 2, \ldots, n.$
(DF) holds:

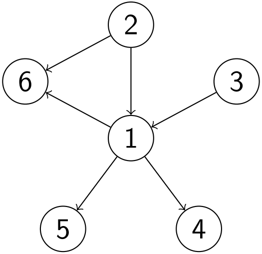
$$\mathbb{P}(X_1, \ldots, X_n) = \mathbb{P}(X_1)\mathbb{P}(X_2 \mid X_1) \cdots \mathbb{P}(X_n \mid X_{n-1}).$$

Thus, (DG) holds: For any $i < j < k$, j d-separates i and k and therefore,

$$X_i \perp X_k \mid X_j.$$

Markov properties

Example: Suppose $f(x_1, \ldots, x_6)$ factorizes according to $\mathcal{G}$.



1. (DG): $\{1, 2\}$ *d*-separates 3 and 6
$\Rightarrow X_3 \perp X_6 \mid \{X_1, X_2\}$.
(DL): pa(6)$=\{1, 2\}$ and $3 \in$ nd(6)
$\Rightarrow X_3 \perp X_6 \mid \{X_1, X_2\}$.
2. (DG): 2 and 3 are *d*-separated by $\varnothing$, thus $X_2 \perp X_3$.
$X_2 \perp X_3 \mid X_5$? False, because 5 is a descendant of a collider 1.
3. (DL): pa(4)$=\{1\}$ and node 4 has no descendant. Thus
$X_4 \perp \{X_2, X_3, X_6, X_5\} \mid X_1$.

Markov properties

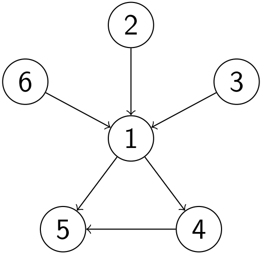
Connections to Markov properties on undirected graphs:

- Moral graph $\mathcal{G}^m$: add edges between all parents of a node in a DAG $\mathcal{G}$ and then ignoring edge orientations. The resulting undirected graph is the moral graph of $\mathcal{G}$.
- 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 $\mathcal{G}^m$.
- S d-separates A and B in $\mathcal{G} \Leftrightarrow S$ separates A and B in $(\mathcal{G}_{\text{An}(A \cup B \cup S)})^m$.
  If $\text{pa}(i) \subset A$ for all $i \in A$, then the subset A is an ancestral set.
  For a subset A of nodes, $\text{An}(A)$ is the smallest ancestral set containing A.
  For a DAG, $\text{An}(A)$ is A and the ancestors of A.

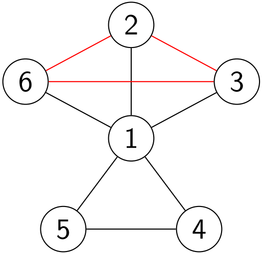
Markov properties

DAG and its moral graph:

DAG $\mathcal{G}$



Moral graph $\mathcal{G}^m$

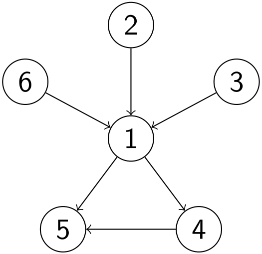


In the moral graph $\mathcal{G}^m$, red edges added between all parents of node 1.

Markov properties

d-separation from moral graphs:

DAG $\mathcal{G}$



- 2 and 3 are d-separated by $\varnothing$.
  $\text{An}(\{2,3\}) = \{2,3\}$

  $(\mathcal{G}_{\{2,3\}})^m$: 

- 2 and 3 are not d-separated by 5.
  $\text{An}(\{2,3,5\}) = \{1,2,3,4,5,6\}$

  In $\mathcal{G}^m$, 2 and 3 are not separated by 5.

Markov properties

Markov equivalence:

Definition 1 (Markov equivalence)

Two DAGs are called Markov equivalent if they imply the same set of *d*-separations.

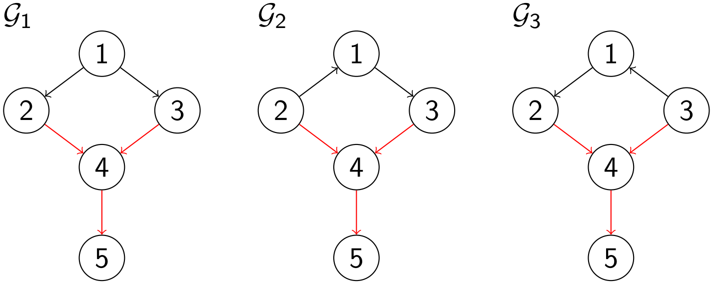
A *v*-structure is a triplet $\{i, j, k\} \subseteq V$ of the form $i \to k \leftarrow j$: i and j are nonadjacent; k is called an *uncovered collider*.

Theorem 2 (Verma and Pearl (1990))

Two DAGs are Markov equivalent if and only if they have the same skeleton and the same v-structures.

Markov properties

Markov equivalence, examples: $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3$ are equivalent DAGs.



Red: compelled edges, same orientation in all equivalent DAGs.
Black: reversible edges, either direction occurs in at least one equivalent DAG.

Markov properties

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

$$f(x) = \prod_{j=1}^{p} f(x_{\sigma(j)} \mid x_{\sigma(1)}, \ldots, x_{\sigma(j-1)})$$
$$= \prod_{j=1}^{p} f(x_{\sigma(j)} \mid x_{A_j}), \quad (2)$$

  where $A_j \subset \{\sigma(1), \ldots, \sigma(j-1)\}$ is the minimum subset such that (2) 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.

Parameterizations

Parameterization: Given $\mathcal{G}$, to parameterize $[X_j \mid \text{pa}(j)]$ as in (1).

(1) Gaussian BNs

- Structural equations:

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

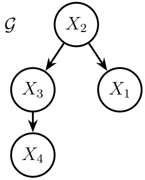
  Assume $\varepsilon_j \sim \mathcal{N}(0, \omega_j^2)$ and $\varepsilon_j \perp \text{pa}(j)$.

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

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

  $\Rightarrow X \sim \mathcal{N}_p(0, \Theta^{-1})$, 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).

Parameterizations



$$B_0 = \begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \\ X_4 \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & X_3 & X_4 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 \\ \beta^{\mathbf{0}}_{\mathbf{21}} & 0 & \beta^{\mathbf{0}}_{\mathbf{23}} & 0 \\ 0 & 0 & 0 & \beta^{\mathbf{0}}_{\mathbf{34}} \\ 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

$$\pi = [\mathbf{4}, \mathbf{1}, \mathbf{3}, \mathbf{2}]$$

$$P_\pi = \begin{pmatrix} 0 & 0 & 0 & \mathbf{1} \\ \mathbf{1} & 0 & 0 & 0 \\ 0 & 0 & \mathbf{1} & 0 \\ 0 & \mathbf{1} & 0 & 0 \end{pmatrix}$$

$$B_\pi = P_\pi B_0 P_\pi^\top = \begin{array}{c} \\ X_4 \\ X_1 \\ X_3 \\ X_2 \end{array} \begin{array}{c} \begin{array}{cccc} X_4 & X_1 & X_3 & X_2 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \beta^{\mathbf{0}}_{\mathbf{34}} & 0 & 0 & 0 \\ 0 & \beta^{\mathbf{0}}_{\mathbf{21}} & \beta^{\mathbf{0}}_{\mathbf{23}} & 0 \end{pmatrix} \end{array}$$

Ye et al. (2021)

- An example DAG $\mathcal{G}$ and its coefficient matrix $B_0 = (\beta^0_{ij})_{4\times 4}$.
- π is a reversed topological sort: $(2, 3, 1, 4)$ is a sort.
- B_π permutes columns and rows of B_0 according to π, and is strictly lower triangular. Similarly define Θ_π and Ω_π.
- $\Theta_\pi = (I - B_\pi)\Omega_\pi^{-1}(I - B_\pi)^{\mathsf{T}}$: Cholesky decomposition.

Parameterizations

(2) Discrete BNs

- Multinomial distribution: $\theta_{km}^{(j)} = \mathbb{P}(X_j = m \mid \mathrm{pa}(j) = k)$. Parameter for $[X_j \mid \mathrm{pa}(j)]$ 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 $[X_j \mid \mathrm{pa}(j)]$.

Parameterizations

Faithfulness:
Given a DAG model $(\mathcal{G}, \mathbb{P})$ where $\mathbb{P}$ satisfies, say (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 2

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

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

Parameterizations

How likely is $\mathbb{P}$ faithful?
Gaussian DAGs.

- Given a DAG $\mathcal{G}$, consider all $B = (\beta_{ij})$ such that $\beta_{ij} \neq 0 \Leftrightarrow i \to j$. Almost all such B and Ω will define a joint distribution $\mathbb{P}$ that is faithful to $\mathcal{G}$.
- Counterexamples: The parameters (β_{ij}) satisfy additional equality constraints that define CI in $\mathbb{P}$ not implied by any d-separation in $\mathcal{G}$.
- For example, path coefficients cancel from i to j. Then $X_i \perp X_j$ but the nodes i and j are not d-separated by $\varnothing$.

Overview of topics

Causal inference

- Model causal relations among nodes: If $i \to j$, 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

$$f(x) = \prod_{j \notin M} f_j(x_j \mid \text{pa}(j)) \prod_{j \in M} g_j(x_j), \tag{3}$$

where $g_j(\bullet)$ is the density of X_j under intervention.

Overview of topics

Structure learning

Given $x_i \sim_{iid} \mathbb{P}$ defined by a DAG $\mathcal{G}$, estimate the DAG $\widehat{\mathcal{G}}$.
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.

Chain graphs

Reference: Lauritzen (1996) §3.2.3

A chain graph on V may contain two types of edges, undirected $(i - j)$ and directed $i \to j$.

- Partition $V = V_1 \cup \cdots \cup V_T$.
- All edges between vertices in the same V_t are undirected.
- All edges between two different subsets V_s, V_t $(s < t)$ are directed and pointing from V_s to V_t.

Special cases: undirected graphs ($T = 1$) and DAGs ($|V_t| = 1$ for all t).

Applications:

- Represent a larger class of distributions.
- Represent Markov equivalence class of a DAG.

Chain graphs

Connectivity components:

- A *path* 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 $k = 1, \ldots, n$.
- If there is a path from i to j, we say i leads to j and write $i \mapsto j$.
- If $i \mapsto j$ and $j \mapsto i$, then we say i and j connect, write $i \leftrightarrow j$.
- The equivalence class $[i] := \{j \in V : i \leftrightarrow j\}$ defined by connectivity is a connectivity component of $\mathcal{G}$.
- Examples:
  1. If $i - j - k$, then $i \leftrightarrow k$ and $i, j, k \in [i]$.
  2. For a DAG, every connectivity component consists of a single node.

Chain graphs

Characterizations of a chain graph:

- Have no directed cycles.
- Its connectivity components (called chain components) induce undirected subgraphs.

To find chain components:

1. Remove all directed edges;
2. Take connectivity components.

Chain graphs

Markov properties on chain graphs:

- Boundary $\mathrm{bd}(i) = \mathrm{pa}(i) \cup \mathrm{ne}(i)$.
- Ancestors $\mathrm{an}(j) = \{i : i \mapsto j, j \not\mapsto i\}$.
- Descendants $\mathrm{de}(i) = \{j : i \mapsto j, j \not\mapsto i\}$.
- Non-descendants $\mathrm{nd}(i) = V \setminus (\mathrm{de}(i) \cup \{i\})$.
- If $\mathrm{bd}(i) \subset A$ for all $i \in A$, then A is an ancestral set.
- Moral graph:
  (1) For each chain component C, add undirected edges between $\mathrm{pa}(C) = \cup_{i \in C} \mathrm{p}a(i)$;
  (2) ignore all edge directions.

Chain graphs

Markov properties on a chain graph $\mathcal{G}$: A joint distribution $\mathbb{P}$

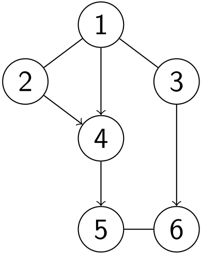
- satisfies the local chain Markov property if $i \perp \text{nd}(i) \mid \text{bd}(i)$ for all $i \in V$.
- satisfies the global chain Markov property if for any disjoint (A, B, S),

$$S \text{ separates } A \text{ and } B \text{ in } (\mathcal{G}_{\text{An}(A \cup B \cup S)})^m \Rightarrow A \perp B \mid S.$$

Unify Markov properties for undirected graphs and DAGs.

Chain graphs

Example chain graph: $V_1 = \{1, 2, 3\}, V_2 = \{4\}, V_3 = \{5, 6\}$.

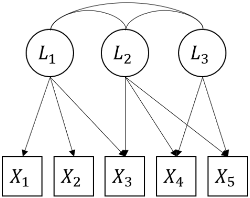


- Chain components: V_1, V_2, V_3.
- Paths: $2 \mapsto 3$, $3 \mapsto 2$, $1 \mapsto 5$, $5 \not\mapsto 1$.
- $\text{bd}(1) = \{2, 3\}$, $\text{bd}(4) = \{1, 2\}$, $\text{bd}(5) = \{4, 6\}$
- $\text{de}(3) = \{4, 5, 6\}$, $\text{de}(5) = \varnothing$.
- Local Markov property:
  $5 \perp \{1, 2, 3\} \mid \{4, 6\}$.
- Global Markov property:
  $2 \perp 3 \mid 1$, from $(\mathcal{G}_{\{1,2,3\}})^m = 2 - 1 - 3$
  $2 \not\perp 3 \mid \{1, 6\}$, from $\mathcal{G}^m$
  $1 \perp 6 \mid \{3, 4\}$, from $\mathcal{G}^m$
  $\mathcal{G}^m$: add $3 - 4$.

Chain graphs

Example application: Factor analysis.

- $V = L \cup X$
  $L = (L_1, \ldots, L_d)$ (latent factors)
  $X = (X_1, \ldots, X_p)$ (observed variables)
- $L \sim \mathcal{N}(0, \Phi)$ (oblique factor analysis)
- $X_j = \beta_j^\mathsf{T} L + \varepsilon_j$, $j = 1, \ldots, p$.



Other applications, see Lauritzen and Richardson (2002).

References I

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

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.

Steffen L Lauritzen and Thomas S. Richardson. Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64: 312–361, 2002.

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

References II

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.

Q. Ye, A.A. Amini, and Qing Zhou. Optimizing regularized cholesky score for order-based learning of Bayesian networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:3555–3572, DOI: 10.1109/TPAMI.2020.2990820, 2021.