

Homework 1 - Introduction to Probabilistic Graphical Models

kipngeno koech - bkoech

February 13, 2025

1 Bayesian Networks

1. Consider a simple Markov Chain structure $X \rightarrow Y \rightarrow Z$, where all variables are binary. You are required to:
 - (a) Write a code (using your preferred programming language) that generates a distribution (not necessarily a valid BN one) over the 3 variables.
[in the notebook]
 - (b) Write a code that verifies whether a distribution is a valid BN distribution.
[in the notebook]
 - (c) Using your code, generate 10000 distributions and compute the fraction of distributions that are valid BN distributions.
[in the notebook]
2. Given the following Bayesian Network

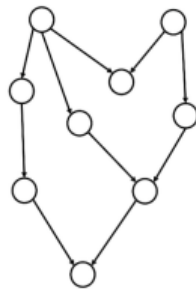


Figure 1: A Bayesian network.

Figure 1: Bayesian Network

- (a) Propose a topological ordering of this graph
In Figure 2, the topological ordering is:
 - i. $A \rightarrow B \rightarrow C \rightarrow D \rightarrow E \rightarrow F \rightarrow G \rightarrow H \rightarrow I$
 - ii. $A \rightarrow B \rightarrow C \rightarrow E \rightarrow D \rightarrow F \rightarrow G \rightarrow I \rightarrow H$
- (b) Let \mathbf{X} be a random vector that is Markov with respect to the graph. We assume that the random variables X_i are binary. Write all the local conditional independence

X_A has no parents, so no independence condition applies here.

X_B has no parents, so no independence condition applies here.

X_C is conditionally independent of all other nodes given its parent X_A :

$$X_C \perp \{X_B, X_D, X_E, X_F, X_G, X_H, X_I\} \mid X_A$$

X_D is conditionally independent of all other nodes given its parent X_A :

$$X_D \perp \{X_B, X_C, X_E, X_F, X_G, X_H, X_I\} \mid X_A$$

X_E is conditionally independent of all other nodes given its parents X_C and X_D :

$$X_E \perp \{X_A, X_B, X_F, X_G, X_H, X_I\} \mid \{X_C, X_D\}$$

X_F is conditionally independent of all other nodes given its parent X_B :

$$X_F \perp \{X_A, X_C, X_D, X_E, X_G, X_H, X_I\} \mid X_B$$

X_G is conditionally independent of all other nodes given its parents X_E and X_F :

$$X_G \perp \{X_A, X_B, X_C, X_D, X_H, X_I\} \mid \{X_E, X_F\}$$

X_H is conditionally independent of all other nodes given its parent X_D :

$$X_H \perp \{X_A, X_B, X_C, X_E, X_F, X_G, X_I\} \mid X_D$$

X_I is conditionally independent of all other nodes given its parents X_G and X_H :

$$X_I \perp \{X_A, X_B, X_C, X_D, X_E, X_F\} \mid \{X_G, X_H\}$$

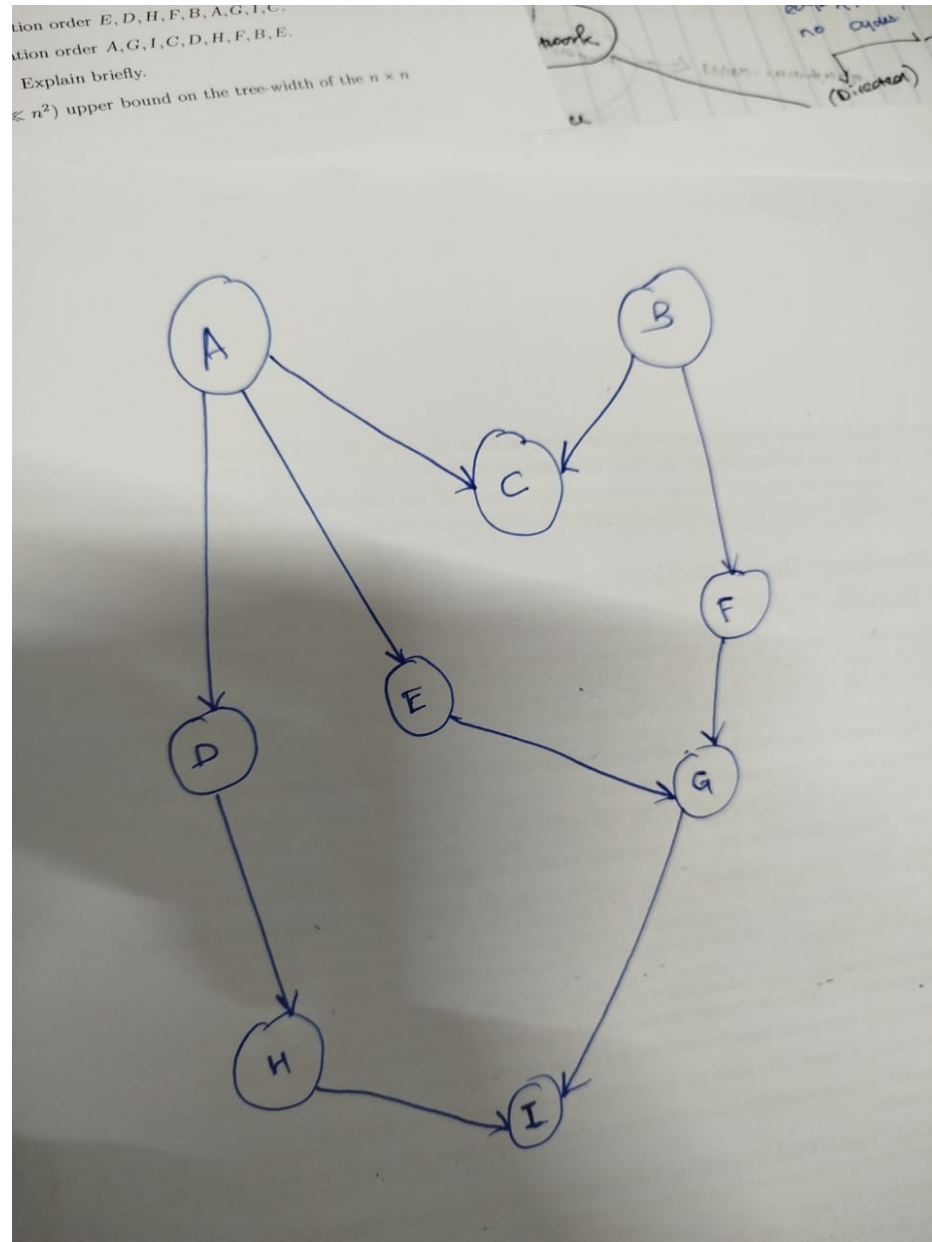


Figure 2: Bayesian Network

3. State True or False, and briefly justify your answer within 3 lines. The statements are either direct consequences of theorems in Koller and Friedman (2009, Ch. 3), or have a short proof. In the follows, P is a distribution and G is a BN structure.

(a) If $A \perp B \mid C$ and $A \perp C \mid B$, then $A \perp B$ and $A \perp C$. (Suppose the joint distribution of A, B, C is positive.) (This is a general probability question not related to BNs.)

- **False.** Whilst local independence implies global independence, global independence does not imply local independence.

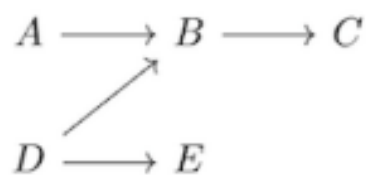


Figure 2: A Bayesian network.

Figure 3: Bayesian Network

(b) In Figure 2, $E \perp C \mid B$

True. E is conditionally independent of C given B because we have conditioned on the node B .

(c) in Figure 2, $A \perp E \mid C$

False. A is not conditionally independent of E given C because there is a path from A to E that is not blocked by C .

In figure 3, Recall the definitions of local and global independences of G and independences of P .

$$I_l(G) = \{(X \perp \text{NonDescendants}_G(X) \mid \text{Parents}_G(X))\} \quad (1)$$

$$I(G) = \{(X \perp Y \mid Z) : \text{d-separated}_G(X, Y \mid Z)\} \quad (2)$$

$$I(P) = \{(X \perp Y \mid Z) : P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)\} \quad (3)$$

(d) In Figure 3, relation 1 is true.

(e) In Figure 3, relation 2 is true.

(f) In Figure 3, relation 3 is true.

(g) If G is an I-map for P , then P may have extra conditional independencies than G .

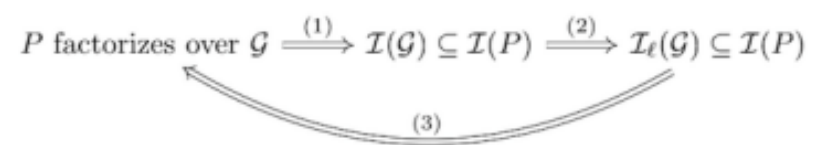


Figure 3: Some relations in Bayesian networks.

- (h) Two BN structures G_1 and G_2 are I-equivalent if they have the same skeleton and the same set of v-structures.
- (i) If G_1 is an I-map of distribution P , and G_1 has fewer edges than G_2 , then G_2 is not a minimal I-map of P .
- (j) The P-map of a distribution, if it exists, is unique.

2 Markov Networks

Let $\mathbf{X} = (X_1, \dots, X_d)$ be a random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The partial correlation matrix \mathbf{R} of \mathbf{X} is a $d \times d$ matrix where each entry $R_{ij} = \rho(X_i, X_j \mid \mathbf{X}_{-ij})$ is the partial correlation between X_i and X_j given the $d - 2$ remaining variables \mathbf{X}_{-ij} . Let $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ be the inverse covariance matrix of \mathbf{X} .

We will prove the relation between \mathbf{R} and $\boldsymbol{\Theta}$, and furthermore how $\boldsymbol{\Theta}$ characterizes conditional independence in Gaussian graphical models.

$$\begin{pmatrix} \Theta_{ii} & \Theta_{ij} \\ \Theta_{ji} & \Theta_{jj} \end{pmatrix} = \begin{pmatrix} \text{Var}[e_i] & \text{Cov}[e_i, e_j] \\ \text{Cov}[e_i, e_j] & \text{Var}[e_j] \end{pmatrix}^{-1}$$

for any $i, j \in [d], i \neq j$. Here e_i is the residual resulting from the linear regression of X_{-ij} to X_i , and similarly e_j is the residual resulting from the linear regression of X_{-ij} to X_j .

1. (10 points) Show that

$$R_{ij} = -\frac{\Theta_{ij}}{\sqrt{\Theta_{ii}\Theta_{jj}}}$$

We start from the given matrix equation:

$$\begin{pmatrix} \Theta_{ii} & \Theta_{ij} \\ \Theta_{ji} & \Theta_{jj} \end{pmatrix} = \begin{pmatrix} \text{Var}[e_i] & \text{Cov}[e_i, e_j] \\ \text{Cov}[e_i, e_j] & \text{Var}[e_j] \end{pmatrix}^{-1}.$$

let us find the determinant of th matrix on the right hand side:

$$\det \begin{pmatrix} \text{Var}[e_i] & \text{Cov}[e_i, e_j] \\ \text{Cov}[e_i, e_j] & \text{Var}[e_j] \end{pmatrix} = \text{Var}[e_i]\text{Var}[e_j] - \text{Cov}[e_i, e_j]^2.$$

Using the determinant, let us find the inverse of the matrix on the right hand side:

$$\frac{1}{\text{Var}[e_i]\text{Var}[e_j] - \text{Cov}[e_i, e_j]^2} \begin{pmatrix} \text{Var}[e_j] & -\text{Cov}[e_i, e_j] \\ -\text{Cov}[e_i, e_j] & \text{Var}[e_i] \end{pmatrix}$$

Let us push the determinant into the matrix:

$$\begin{pmatrix} \frac{\text{Var}[e_j]}{\text{Var}[e_i]\text{Var}[e_j] - \text{Cov}[e_i, e_j]^2} & -\frac{\text{Cov}[e_i, e_j]}{\text{Var}[e_i]\text{Var}[e_j] - \text{Cov}[e_i, e_j]^2} \\ -\frac{\text{Cov}[e_i, e_j]}{\text{Var}[e_i]\text{Var}[e_j] - \text{Cov}[e_i, e_j]^2} & \frac{\text{Var}[e_i]}{\text{Var}[e_i]\text{Var}[e_j] - \text{Cov}[e_i, e_j]^2} \end{pmatrix}$$

We can now equate the two matrices:

$$\begin{pmatrix} \Theta_{ii} & \Theta_{ij} \\ \Theta_{ji} & \Theta_{jj} \end{pmatrix} = \begin{pmatrix} \frac{\text{Var}[e_j]}{\text{Var}[e_i]\text{Var}[e_j] - \text{Cov}[e_i, e_j]^2} & -\frac{\text{Cov}[e_i, e_j]}{\text{Var}[e_i]\text{Var}[e_j] - \text{Cov}[e_i, e_j]^2} \\ -\frac{\text{Cov}[e_i, e_j]}{\text{Var}[e_i]\text{Var}[e_j] - \text{Cov}[e_i, e_j]^2} & \frac{\text{Var}[e_i]}{\text{Var}[e_i]\text{Var}[e_j] - \text{Cov}[e_i, e_j]^2} \end{pmatrix}$$

To show that $R_{ij} = -\frac{\Theta_{ij}}{\sqrt{\Theta_{ii}\Theta_{jj}}}$, we need to show that:

$$\begin{aligned} R_{ij} &= -\frac{\Theta_{ij}}{\sqrt{\Theta_{ii}\Theta_{jj}}} \\ R_{ij} &= -\frac{-\frac{\text{Cov}[e_i, e_j]}{\text{Var}[e_i]\text{Var}[e_j] - \text{Cov}[e_i, e_j]^2}}{\sqrt{\frac{\text{Var}[e_i]}{\text{Var}[e_i]\text{Var}[e_j] - \text{Cov}[e_i, e_j]^2} \frac{\text{Var}[e_j]}{\text{Var}[e_i]\text{Var}[e_j] - \text{Cov}[e_i, e_j]^2}}} \\ R_{ij} &= -\frac{-\frac{\text{Cov}[e_i, e_j]}{\text{Var}[e_i]\text{Var}[e_j] - \text{Cov}[e_i, e_j]^2}}{\sqrt{\frac{\text{Var}[e_i]\text{Var}[e_j]}{(\text{Var}[e_i]\text{Var}[e_j] - \text{Cov}[e_i, e_j]^2)^2}}} \\ R_{ij} &= -\frac{-\frac{\text{Cov}[e_i, e_j]}{\text{Var}[e_i]\text{Var}[e_j] - \text{Cov}[e_i, e_j]^2}}{\frac{\sqrt{\text{Var}[e_i]\text{Var}[e_j]}}{\text{Var}[e_i]\text{Var}[e_j] - \text{Cov}[e_i, e_j]^2}} \end{aligned}$$

But we know that:

$$\begin{aligned} \Theta_{ij} &= -\frac{\text{Cov}[e_i, e_j]}{\text{Var}[e_i]\text{Var}[e_j] - \text{Cov}[e_i, e_j]^2} \\ R_{ij} &= -\frac{-\text{Cov}[e_i, e_j]}{\sqrt{\text{Var}[e_i]\text{Var}[e_j]}} \\ R_{ij} &= \frac{\text{Cov}[e_i, e_j]}{\sqrt{\text{Var}[e_i]\text{Var}[e_j]}} \end{aligned}$$

let us rearrange this equation:

$$\text{Cov}[e_i, e_j] = R_{ij} \sqrt{\text{Var}[e_i]\text{Var}[e_j]}$$

Then we know:

$$\begin{aligned} \theta_{ij} &= -\frac{\text{Cov}[e_i, e_j]}{\text{Var}[e_i]\text{Var}[e_j] - \text{Cov}[e_i, e_j]^2} \\ \theta_{ij} &= -\frac{R_{ij} \sqrt{\text{Var}[e_i]\text{Var}[e_j]}}{\text{Var}[e_i]\text{Var}[e_j] - \text{Cov}[e_i, e_j]^2} \end{aligned}$$

let us have R_{ij} in the LHS:

$$R_{ij} = -\frac{\theta_{ij}\text{Var}[e_i]\text{Var}[e_j] - \text{Cov}[e_i, e_j]^2}{\sqrt{\text{Var}[e_i]\text{Var}[e_j]}}$$

we know θ_{ii} and Θ_{jj} as:

$$\theta_{ii} = \frac{\text{Var}[e_i]}{\text{Var}[e_i]\text{Var}[e_j] - \text{Cov}[e_i, e_j]^2}$$

$$\theta_{jj} = \frac{\text{Var}[e_j]}{\text{Var}[e_i]\text{Var}[e_j] - \text{Cov}[e_i, e_j]^2}$$

let us multiply the two:

$$\theta_{ii}\theta_{jj} = \frac{\text{Var}[e_i]\text{Var}[e_j]}{(\text{Var}[e_i]\text{Var}[e_j] - \text{Cov}[e_i, e_j]^2)^2}$$

we the proceed to find the square root of the product:

$$\sqrt{\theta_{ii}\theta_{jj}} = \frac{\sqrt{\text{Var}[e_i]\text{Var}[e_j]}}{\text{Var}[e_i]\text{Var}[e_j] - \text{Cov}[e_i, e_j]^2}$$

Let us rearrange the equation:

$$\text{Var}[e_i]\text{Var}[e_j] - \text{Cov}[e_i, e_j]^2 = \frac{\sqrt{\text{Var}[e_i]\text{Var}[e_j]}}{\sqrt{\theta_{ii}\theta_{jj}}}$$

We can then proceed and replace this in the equation for R_{ij} :

$$R_{ij} = -\frac{\theta_{ij} \frac{\sqrt{\text{Var}[e_i]\text{Var}[e_j]}}{\sqrt{\theta_{ii}\theta_{jj}}}}{\sqrt{\text{Var}[e_i]\text{Var}[e_j]}}$$

This simplifies to:

$$R_{ij} = -\frac{\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}}$$

We have therefore shown that $R_{ij} = -\frac{\Theta_{ij}}{\sqrt{\Theta_{ii}\Theta_{jj}}}$.

2. (15 points) From the above result and the relation between independence and correlation, we know

$$\Theta_{ij} = 0 \iff R_{ij} = 0 \implies X_i \perp X_j \mid X_{-ij}$$

Note the last implication only holds in one direction. Now suppose $X \sim N(\mu, \Sigma)$ is jointly Gaussian. Show that $R_{ij} = 0 \implies X_i \perp X_j \mid X_{-ij}$.

$$R_{ij} = -\frac{\Theta_{ij}}{\sqrt{\Theta_{ii}\Theta_{jj}}}.$$

If $R_{ij} = 0$, then $\Theta_{ij} = 0$.

For a jointly Gaussian distribution, the inverse covariance matrix $\Theta = \Sigma^{-1}$ encodes conditional independence. Specifically:

$$\Theta_{ij} = 0 \iff X_i \perp X_j \mid X_{-ij}.$$

This is because the off-diagonal elements of Θ represent the conditional dependence between variables after accounting for all other variables.

Since $R_{ij} = 0$ implies $\Theta_{ij} = 0$, it follows that:

$$X_i \perp X_j \mid X_{-ij}.$$

3 Exact Inference - Variable Elimination

Reference materials for this problem:

- Jordan textbook Ch. 3, available at <https://people.eecs.berkeley.edu/~jordan/prelims/chapter3.pdf>
- Koller and Friedman (2009, Ch. 9 and Ch. 10)

3.1 Variable elimination on a grid [10 points]

Consider the following Markov network:

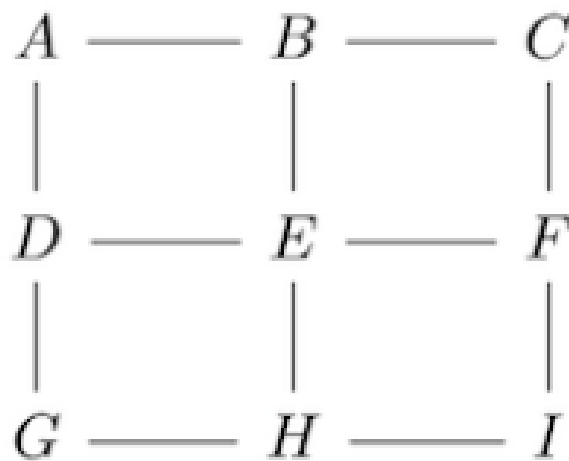


Figure 4: Markov Network

We are going to see how tree-width, a property of the graph, is related to the intrinsic complexity of variable elimination of a distribution

- (5 points) Write down largest clique(s) for the elimination order $E, D, H, F, B, A, G, I, C$.
 We start by eliminating E , its neighbors are D, F, H, B , so the clique here is $\{D, F, H, B\}$. size: **4**
 We then eliminate D , its neighbors are A, B, F, G, H , so the clique here is: $\{A, B, F, G, H\}$. size: **5**
 We then eliminate H , its neighbors are G, A, I, B, F , so the clique here is: $\{G, A, I, B, F\}$. size: **5**
 We then eliminate F , its neighbors are A, B, C, G, I , so the clique here is: $\{A, B, C, G, I\}$. size: **5**
 We then eliminate B , its neighbors are A, C, G, I , so the clique here is: $\{A, C, G, I\}$. size: **4**
 We then eliminate A , its neighbors are C, G, I , so the clique here is: $\{C, G, I\}$. size: **3**
 We then eliminate G , its neighbors are C, I , so the clique here is: $\{C, I\}$. size: **2**
 We then eliminate I , its neighbors are C , so the clique here is: $\{C\}$. size: **1**
 The largest clique(s) for the elimination order $E, D, H, F, B, A, G, I, C$ is **5**
- (5 points) Write down largest clique(s) for the elimination order $A, G, I, C, D, H, F, B, E$.
 We start by eliminating A , its neighbors are D, B , so the clique here is: $\{D, B\}$. size: **2**
 We then eliminate G , its neighbors are D, H , so the clique here is: $\{D, H\}$. size: **2**
 We then eliminate I , its neighbors are H, F , so the clique here is: $\{H, F\}$. size: **2**
 We then eliminate C , its neighbors are F, B , so the clique here is: $\{F, B\}$. size: **2**
 We then eliminate D , its neighbors are H, B, E , so the clique here is: $\{H, B, E\}$. size: **3**
 We then eliminate H , its neighbors are F, B, E , so the clique here is: $\{F, B, E\}$. size: **3**
 We then eliminate F , its neighbors are B, E , so the clique here is: $\{B, E\}$. size: **2**
 We then eliminate B , its neighbors are E , so the clique here is: $\{E\}$. size: **1**
 The largest clique(s) for the elimination order $A, G, I, C, D, H, F, B, E$ is **3**
- (5 points) Which of the above ordering is preferable? Explain briefly.
 The second ordering is preferable because it results in a smaller largest clique size (3) compared to the first ordering (5). A smaller clique size generally leads to more efficient computations in variable elimination algorithms.
- (10 points) Using this intuition, give a reasonable ($\ll n^2$) upper bound on the tree-width of the $n \times n$ grid.
 The tree-width of an $n \times n$ grid is at most $2n - 2$. This is because the grid can be decomposed into a tree structure where each node has at most $2n - 2$ neighbors. This upper bound is reasonable and much smaller than n^2 , which would be the case for a fully connected graph.