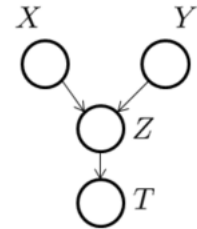


As usual, please hand in on paper form your derivations and answers to the questions. You can use any programming language for your source code (submitted on Studium as per the website instructions). All the requested figures should be printed on paper with clear titles that indicate what the figures represent.

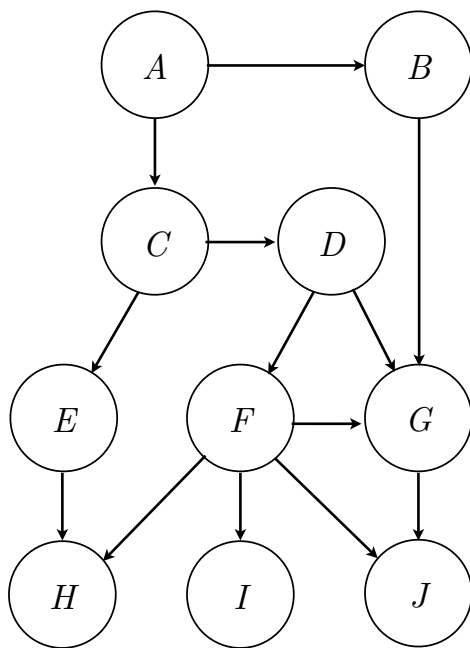
1. **DGM (5 points)**

Consider the directed graphical model G on the right. Write down the implied factorization for any joint distribution $p \in \mathcal{L}(G)$. Is it true that $X \perp\!\!\!\perp Y \mid T$ for any $p \in \mathcal{L}(G)$? Prove or disprove.



2. **d-separation in DGM (5 points)**

Indicate (yes or no) which conditional independence statements are true?



- (a) $C \perp B$?
- (b) $C \perp B \mid A$?
- (c) $C \perp B \mid A, J$?
- (d) $C \perp B \mid A, J, D$?
- (e) $C \perp G$?
- (f) $C \perp G \mid B$?
- (g) $C \perp G \mid B, D$?
- (h) $C \perp G \mid B, D, H$?
- (i) $C \perp G \mid B, D, H, E$?
- (j) $B \perp I \mid J$?

3. **Positive interactions in-V-structure (10 points)**

Let X, Y, Z be binary random variables with a joint distribution parametrized according to the graph: $X \rightarrow Z \leftarrow Y$. We define the following:

$$a := P(X = 1), \quad b := P(X = 1 \mid Z = 1), \quad c := P(X = 1 \mid Z = 1, Y = 1)$$

- (a) For all the following cases, provide examples of conditional probability tables (and calculate the quantities a, b, c), that render the statements true:
 - (i) $a > c$
 - (ii) $a < c < b$
 - (iii) $b < a < c$.
- (b) Think of X and Y as causes and Z as a common effect, for all the above cases (i, ii, et iii), summarize in a sentence or two why the declarations are true for your examples.

4. **Flipping a covered edge in a DGM (10 points)**

Let $G = (V, E)$ be a DAG. We say that a directed edge $(i, j) \in E$ is a *covered edge* if and only if $\pi_j = \pi_i \cup \{i\}$. Let $G' = (V, E')$, with $E' = (E \setminus \{(i, j)\}) \cup \{(j, i)\}$. Prove that $\mathcal{L}(G) = \mathcal{L}(G')$.

5. **Equivalence of directed tree DGM with undirected tree UGM (10 points)**

Let G be a directed tree and G' its corresponding undirected tree (where the orientation of edges is ignored). Recall that by the definition of a directed tree, G does not contain any v-structure. Prove that $\mathcal{L}(G) = \mathcal{L}(G')$.

6. **Hammersley-Clifford Counter example (10 points)**

In class, I mentioned that the strict positivity of the joint distribution was crucial in the Hammersley-Clifford theorem. Here is a counter-example that shows the problems when we have zero probabilities (it is example 4.4 in Koller & Friedman). Consider a joint distribution p over four binary random variables: X_1, X_2, X_3 and X_4 which gives probability $\frac{1}{8}$ to each of the following eight configurations, and probability zero to all others:

$$\begin{array}{cccc} (0, 0, 0, 0) & (1, 0, 0, 0) & (1, 1, 0, 0) & (1, 1, 1, 0) \\ (0, 0, 0, 1) & (0, 0, 1, 1) & (0, 1, 1, 1) & (1, 1, 1, 1). \end{array}$$

Let G be the usual four nodes undirected graph $X_1 - X_2 - X_3 - X_4 - X_1$. One can show that p satisfies the global Markov property with respect to this graph G because of trivial deterministic relationships. For example, if we condition on $X_2 = 0$ and $X_4 = 0$, then the only value of X_3 with non-zero probability is $X_3 = 0$, and thus $X_3 | X_2 = 0, X_4 = 0$ being a deterministic random variable, it is trivially conditionally independent to X_1 . By (painfully) going through all other possibilities, we get similar situations (for example $X_2 = 0$ and $X_4 = 1$ forces $X_1 = 0$, etc.). Prove that the distribution p *cannot* factorize according to G (and thus $p \notin \mathcal{L}(G)$). *Hint: argue by contradiction.*

7. **[BONUS]: bizarre conditional independence properties (10 bonus points)**

Let (X, Y, Z) be a random vector with a finite sample space. Consider the following statement:

“If $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp Y$ then $(X \perp\!\!\!\perp Z \text{ or } Y \perp\!\!\!\perp Z)$.”

- (a) Is this true if one assumes that Z is a binary variable? Prove or disprove.
- (b) Is the statement true in general? Prove or disprove.

8. **Implementation: EM and Gaussian mixtures (30 points)**

The file `EMGaussian.train` contains samples of data x_i where $x_i \in \mathbb{R}^2$ (one datapoint per row). The goal of this exercise is to implement the EM algorithm for some mixtures of K Gaussians in \mathbb{R}^d (here $d = 2$ and $K = 4$), for i.i.d. data. (NB: in this exercise, no need to prove any of the formula used in the algorithms except for question (b)).

- (a) Implement the K-means algorithm. Represent graphically the training data, the cluster centers, as well as the different clusters (use 4 colors). Try several random initializations and compare results (centers and the actual K-means objective values).

- (b) Consider a Gaussian mixture model in which the covariance matrices are proportional to the identity. Derive the form of the M-step updates for this model and implement the corresponding EM algorithm (using an initialization with K-means).
Represent graphically the training data, the centers, as well as the covariance matrices (an elegant way is to represent the ellipse that contains a specific percentage, e.g., 90%, of the mass of the Gaussian distribution).
Estimate and represent (e.g. with different colors or different symbols) the most likely latent variables for all data points (with the parameters learned by EM).
- (c) Implement the EM algorithm for a Gaussian mixture with general covariance matrices. Represent graphically the training data, the centers, as well as the covariance matrices. Estimate and represent (e.g. with different colors or different symbols) the most likely latent variables for all data points (with the parameters learned by EM).
- (d) Comment the different results obtained in earlier questions. In particular, compare the normalized log-likelihoods of the two mixture models on the training data, as well as on test data (in `EMGaussian.test`). (Here normalize the log-likelihood by the number of observations (rows) – it makes the number more manageable for comparison and puts it on the right scale).