IITM-CS6730 : Probabilistic Graphical Models — Release Date: Feb 9, 2020

Assignment 2 — Due Date : Feb 18, 23:59

Roll No: EE16B026 — Name: Muqeeth

Collaborators (if any): Gowri Shankar(EE16B021)

References (if any): https://www.cs.princeton.edu/ bee/courses/scribe/lec-09-09-2013.pdf

- Use LaTeX to write-up your solutions (in the solution blocks of the source LaTeX file of this assignment), and submit the resulting single pdf file at GradeScope by the due date. (Note: As always, **no late submissions** will be allowed, other than one-day late submission with 10% penalty! Within GradeScope, indicate the page number where your solution to each question starts, else we won't be able to grade it!).
- Collaboration is encouraged, but all write-ups must be done individually and independently, and mention your collaborator(s) if any. Same rules apply for codes written for any programming assignments (i.e., write your own code; we will run plagiarism checks on codes).
- If you have referred a book or any other online material for obtaining a solution, please cite the source. Again don't copy the source *as is* - you may use the source to understand the solution, but write-up the solution in your own words.

1. (4 points) [APPLYING D-SEP PRECISELY!] Let's clarify some implications of d-sep sketched in class. Let X, Y be two **non-adjacent** r.v.s in a BN G, and let X, Y have at least two common ancestors and at least two common descendants. Note that X and Y can either be in an ancestor-descendant relation (denoted *adr*), or not (denoted *non-adr*). Answer if each statement below is True/False.
   (a) (1 point) For adr case, $(X \perp Y \mid \text{all common ancs.})$ holds (i.e., this CI is in $I(G)$ for any G).
   sol: True, Assuming every node is ancestor of itself.
   (b) (1 point) For adr case, $(X \perp Y \mid Pa_X \cup Pa_Y)$ doesn't hold in some G.
   sol: False
   (c) (1 point) For non-adr case, whether $(X \perp Y \mid \text{a common anc. } A)$ holds (or not) can depend on which common anc. is chosen as A.
   sol: True
   (d) (1 point) For non-adr case, whether $(X \perp Y \mid \text{a common desc. } D)$ holds (or not) can depend on which common desc. is chosen as D.
   sol: False

2. (5 points) [ISING MEETS BOLTZMANN] Consider a MN H over n nodes (such as a grid graph, with $i \sim j$ indicating edges (X_i, X_j) in H). Ising model defines a joint distbn. over ± 1-valued r.v.s X as $P(x) = \frac{1}{Z'} \exp\left\{-\sum_{i\sim j} a_{ij} x_i x_j - \sum_i b_i x_i\right\}$, whereas Boltzmann machine defines a joint distbn. over $0/1$**-valued** r.v.s. using the same form as $P(x) = \frac{1}{Z} \exp\left\{-\sum_{i\sim j} c_{ij} x_i x_j - \sum_i d_i x_i\right\}$, with the result being a single non-zero contribution of each edge (X_i, X_j)'s potential when $X_i = X_j = 1$).

(a) (2 points) Show how any Ising model can be reformulated as a Boltzmann machine (with -1 mapped to 0) by showing how the node/edge potentials and partition function of both models are related.
sol:
Ising model is defined as $P_X(x) = \frac{1}{Z'} \exp\left\{-\sum_{i\sim j} a_{ij}x_i x_j - \sum_i b_i x_i\right\}$
Boltzmann machine is defined as $P_Y(y) = \frac{1}{Z} \exp\left\{-\sum_{i\sim j} c_{ij}y_i y_j - \sum_i d_i y_i\right\}$
For mapping Ising model and Boltzmann, we have $x_i = 2y_i - 1$
Here we consier $a_{ij} = 0$ if there is no edge between node i and node j.

$$
\begin{aligned}
P_Y(y) &= P_X(2y-1) \\
&= \frac{1}{Z'} \exp\left\{-\sum_{(i,j)\in E} a_{ij}(2y_i-1)(2y_j-1) - \sum_{i\in V} b_i(2y_i-1)\right\} \\
&= \frac{1}{Z'} \exp\left\{-\sum_{(i,j)\in E} 4a_{ij}y_i y_j + 2\sum_{(i,j)\in E} a_{ij}(y_i+y_j) - \sum_{(i,j)\in E} a_{ij} - 2\sum_{i\in V} b_i y_i + \sum_{i\in V} b_i\right\} \\
&= \frac{1}{Z'} \exp\left\{-\sum_{(i,j)\in E} a_{ij} + \sum_{i\in V} b_i\right\} \exp\left\{-\sum_{(i,j)\in E} 4a_{ij}y_i y_j + \sum_i \sum_j a_{ij}(y_i+y_j) - 2\sum_{i\in V} b_i y_i\right\} \\
&= \frac{1}{Z'} \exp\left\{-\sum_{(i,j)\in E} a_{ij} + \sum_{i\in V} b_i\right\} \exp\left\{-\sum_{(i,j)\in E} 4a_{ij}y_i y_j + \sum_i \sum_j 2a_{ij}y_i - 2\sum_{i\in V} b_i y_i\right\} \\
&= \frac{1}{Z'} \exp\left\{-\sum_{(i,j)\in E} a_{ij} + \sum_{i\in V} b_i\right\} \exp\left\{-\sum_{(i,j)\in E} 4a_{ij}y_i y_j - \sum_{i\in V}(\sum_{j\in V} 2a_{ij} - 2b_i)y_i\right\} \qquad (1)
\end{aligned}
$$

We know, $P_Y(y) = \frac{1}{Z} \exp\left\{-\sum_{i\sim j} c_{ij}y_i y_j - \sum_i d_i y_i\right\}$
comparing it with equation 1, we get

$$
\begin{aligned}
\frac{1}{Z} &= \frac{1}{Z'} \exp\left\{-\sum_{(i,j)\in E} a_{ij} + \sum_{i\in V} b_i\right\} \\
c_{ij} &= 4a_{ij} \\
d_i &= \sum_{j\in V} 2a_{ij} - 2b_i
\end{aligned}
$$

(b) (2 points) Derive the conditional distribution $P(X_i \mid Nbrs_{X_i})$ for the Boltzmann machine, and express it as a "neuron activation model" (i.e., in terms of the function $sigmoid(z) = e^z/(1+e^z)$).
sol:

$P(x) = \frac{1}{Z} \exp\left\{-\sum_{i\sim j} c_{ij}x_i x_j - \sum_i d_i x_i\right\}$
Here $\mathcal{X}$ is set of all nodes
By using defination of markov blanket of X_l, we have $P(X_l \mid Nbrs_{X_l}) = P(X_l \mid \mathcal{X} - X_k)$

$$\begin{aligned} P(X_l \mid Nbrs_{X_l}) &= P(X_l \mid \mathcal{X} - X_k) \\ &= \frac{P(X_k, \mathcal{X} - X_k)}{P(\mathcal{X} - X_k)} \\ &= \frac{P(X_k, \mathcal{X} - X_k)}{\sum_{X_k} P(\mathcal{X})} \end{aligned}$$

X_l is 0/1-valued r.v

$$\begin{aligned} P(X_l = 1 \mid Nbrs_{X_l}) &= \frac{P(X_l = 1, \mathcal{X} - X_l)}{\sum_{X_l} P(\mathcal{X})} \\ &= \frac{P(X_l = 1, \mathcal{X} - X_l)}{P(X_l = 0, \mathcal{X} - X_l) + P(X_l = 1, \mathcal{X} - X_l)} \\ &= \frac{1}{1 + \frac{P(X_l=0, \mathcal{X}-X_l)}{P(X_l=1, \mathcal{X}-X_l)}} \end{aligned}$$

$$\begin{aligned} \frac{P(X_l = 0, \mathcal{X} - X_l)}{P(X_l = 1, \mathcal{X} - X_l)} &= \frac{exp(-\sum_{i\sim j, i,j\neq l} c_{ij}x_i x_j - \sum_{i,i\neq l} d_i x_i)}{exp(-\sum_j c_{lj}x_j - b_l) exp(-\sum_{i\sim j, i,j\neq l} c_{ij}x_i x_j - \sum_{i,i\neq l} d_i x_i)} \\ &= exp(\sum_j c_{lj}x_j + b_l) \\ &= exp(z) \end{aligned}$$

Here, $z = \sum_j c_{lj}x_j + b_l$

$$\begin{aligned} P(X_l = 1 \mid Nbrs_{X_l}) &= \frac{1}{1 + exp(z)} \\ P(X_l = 0 \mid Nbrs_{X_l}) &= \frac{1}{1 + exp(-z)} \\ &= \frac{e^z}{1 + e^z} \end{aligned}$$

(c) (1 point) Is Ising model in the exponential family of distributions? Explain using the defn. of exponential family seen in the last assignment.
sol:
$P(x) = \frac{1}{Z'} \exp\left\{-\sum_{i,j\in E} a_{ij}x_i x_j - \sum_i b_i x_i\right\}$
Here consider, $a_{ij} = \theta_{ij} = \theta_e$, $b_i = \theta_{v_i}$. $v \in (v_1, v_2, ..v_V)$ and $e \in (e_1, e_2, ..e_E)$

Z' is function of $\theta_{v_1}, \theta_{v_2}, ..\theta_{v_V}, \theta_{e_1}, \theta_{e_2}, .., \theta_{e_E}$. For notational ease, $Z' = f(\theta_v, \theta_e)$

$$\begin{aligned} P(x) &= \frac{1}{Z'} \exp\left\{ -\sum_{i\sim j} a_{ij} x_i x_j - \sum_i b_i x_i \right\} \\ &= \frac{1}{f(\theta_v, \theta_e)} \exp\left\{ -\sum_{i,j\in E} \theta_{ij} x_i x_j - \sum_{v\in V} \theta_v x_i \right\} \\ &= \exp\left\{ -\sum_{i,j\in E} \theta_{ij} x_i x_j - \sum_{v\in V} \theta_v x_i - \ln(f(\theta_v, \theta_e)) \right\} \end{aligned}$$

Comparing it with the exponential family, $p(x|\theta) = h(x) \exp\left\{\eta(\theta)^\top t(x) - a(\theta)\right\}$, we get,

$$\begin{aligned} h(x) &= 1 \\ t(x) &= [x_v \ \ (v \in V) \ \ , x_i x_j \ \ ((i,j) \in E)] \\ \eta(\theta) &= [\theta_{v_1}, \theta_{v_2}, ..\theta_{v_V}, \theta_{e_1}, \theta_{e_2}, .., \theta_{e_E}] \\ a(\theta) &= \ln(f(\theta_v, \theta_e)) \end{aligned}$$

Therefore, Ising model is in exponential family distributions.

3. (7 points) [(OVER)-PARAMETERIZATIONS]

   (a) (2 points) Consider a complete graph MN over 3 binary r.v.s parameterized either as a single clique potential or as a product of 3 pairwise potentials. The former requires 8 parameters to write down the clique potential table, whereas the latter requires 12 parameters to write down the pairwise potentials. Can you explain how to resolve the apparent contradiction between these two different parameter counts?
   sol:
   If we have N binary random variables forming a clique, In order to represent the joint over these N r.v's we need 2^n parameters.

   If we consider only the edge potentials, we need 4 parameters for each edge and taking all edges we need $4\binom{n}{2} = 2n(n-1)$ parameters.

   For n>=6, $2^n > 2n(n-1)$. A clique potential over nodes sometimes cannot be broken down to product of edge potentials. Therefore, a clique over all nodes need more parameters.

   For n=3, 8 parameters are enough to specify the distribution. Among 12 parameters, some are dependent.

   (b) (2 points) Consider a MN comprising among all other factors, two specific factors $\phi_1(A, B)$ and $\phi_2(B, C)$. Will changing these factor parameterizations as below change the joint distribution?

Justify your answer.

$$\begin{aligned}\phi_1(A, B = b) &:= \phi_1(A, B = b)\,\lambda_b \\ \phi_2(B = b, C) &:= \phi_2(B = b, C)\,(1/\lambda_b),\end{aligned}$$

where λ_1 and λ_2 are any positive constants.
sol:
The joint probability distribution is product of local factors.

$$\phi_1(A, B = b)\ \ \phi_2(B = b, C) = \phi_1(A, B = b)\,\lambda_b\ \ \phi_2(B = b, C)\,(1/\lambda_b)$$

Hence, the joint probability distribution does not change given λ_b is positive.

(c) (3 points) Consider the following factorisation of a distribution P over 4 r.v.s that take values in $\{1, 2, \ldots, a\}$

$$P\,(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4) = \frac{1}{Z}\psi_1\,(x_1, x_2)\,\psi_2\,(x_2, x_3)\,\psi_3\,(x_3, x_4)\,\psi_4\,(x_1, x_4)\,\psi_5\,(x_2, x_4)$$

Let $\mathcal{H}$ be the Markov network for the factorisation above. If Q is a distribution that factorizes according to $\mathcal{H}$, what can be said about the form of the distribution Q? How many parameters would be required to represent such a distribution Q? Compare this with the number of parameters required to represent the distribution P. (You may assume a is large enough that $a - 1 \approx a$ and a^k is negligible when compared to a^{k+1}.)
sol:
The markov network $\mathcal{H}$ has two cliques (X_2, X_3, X_4) ,(X_1, X_2, X_4). Therefore number of parmaters required are 2 * (a * a *a) = $2a^3$ for distribution Q which factorizes according to $\mathcal{H}$.

Distribution P has 5 cliques of edges. Therfore, number of parameters required are $5a^2$.

Taking the ratio of number of parameters of Q and number of parameters of P, we have.

$$\frac{2a^3}{5a^2} = \frac{2a}{5} \qquad >> 1$$

Thus, distribution Q requires more parameters than P.

4. (9 points) [I-EQUIVALENCE AND POSITIVE FRIENDS]

(a) (2 points) How many graphs are I-equivalent to the simple directed chain $X_1 \to X_2 \cdots \to X_n$?
sol:
There are n graphs that are I-equivalent to the given directed chain. For every node, the right edges should point toward the right and left edges should point toward the left.

(b) (2 points) Show a BN G (if any) that is I-equivalent to the Diamond Graph MN H (Student Misconception Example with $I(H) = \{(A \perp C \mid B, D), (B \perp D \mid A, C)\}$), when extra variables are allowed in the BN. That is, you are allowed a few extra variables W in the BN G that are always unobserved s.t. the set of dsep-implied CIs in G over the X variables alone (which holds in the marginal distbn. $P(X) = \sum_w P(X, W)$) is identical to $I(H)$.
sol:
No, we cannot have such a BN G. We know that there is no BN G′, where $I(G') = I(H)$ for given diamond graph H

Assume such BN G exists,and the set of dsep-implied CIs over the X variables alone in graph G of X,W = $I(H)$

We know, the set of dsep-implied CIs over the X variables alone in graph G of X,W will be equal to dsep-implied CIs in graph G′ of X, $I(G')$ when W is marginalized

This means graph G′ exists where $I(G') = I(H)$. But we know it cannot. Therefore, we cannot have such a BN G.

(c) (2 points) Let $\mathscr{P}$ be a positive distribution, then prove that the Markov Blanket $MB_{\mathscr{P}}(X)$ is unique. [Hint: What happens if two different sets U_1 and U_2 are each minimal $MB_P(X)$?]
sol:
Let $\mathcal{X}$ be set of all the random variables in the distribution. Consider two different sets U_1 and U_2 which are minimal $MB_p(X)$
Let us define disjoint sets as follows:

- $A = U_1 \setminus U_2$
- $C = U_2 \setminus U_1$
- $B = U_1 \cap U_2$
- $P = \mathcal{X} \setminus (A \cup B \cup C \cup X)$

U_1 is minimal $MB_p(X) \Rightarrow X \perp P, C \mid A, B$ (1)
U_2 is minimal $MB_p(X) \Rightarrow X \perp P, A \mid B, C$ (2)
Using Decomposition on (1) , (2) respectively. We get
$X \perp A \mid B, C$ (3)
$X \perp C \mid A, B$ (4)
Given positive distribution, construced sets are disjoint. Using intersection on (3) and (4). we get
$X \perp C, A \mid B$ (5)
Using weak union on (1) we get
$X \perp P \mid B, A, C$ (6)
Using contraction on (5) and (6) we get
$X \perp A, P, C \mid B$

$B = U_1 \cap U_2$ now becomes minimal. For U_1 and U_2 to be minimal they have to be equal. So, markov blanket is unique.

(d) (3 points) If a DAG G has no immoralities, then prove that G is chordal (i.e, the underlying undirected graph is chordal). Does such DAGs G always have a perfect MN I-map?
sol:
The undirected graph H is constructed by dropping the orientation of edges in DAG G.
Given DAG G has no immoralities, we have to prove H is chordal.
If H has no loops with size >4. it is chordal
Take any loop with size >4 in H if it exists. Correspondingly, we will have a loop in DAG G.
Since a loop in DAG G should not have a cycle, there will be atleast one v-edge in the loop.
The parents of v-edge in the loop will be connected by an edge since DAG G has no immoralities. This edge will serve as a chord for the loop in graph H.
We have shown for every loop in H, there is a chord joining non-adjacent vertices. So, H is chordal.

If there are no immoralities in DAG G, the independencies captured by dsep and Hsep are just by blocking paths/trails between two nodes with some observed nodes and so they become the same.Therefore for DAG G, we will have a perfect MN I-map. ie I(G) = I(H)

(e) (3 points) [BONUS] Solve Exercise 4.1 from [KF] book related to the non-positive distribution example in Page 116. It will show how Hammersley-Clifford theorem doesn't hold for this distribution.
sol:
The given undirected graph H is:

$$X_1 - X_2 - X_3 - X_4 - X_1$$

Assume the distribution factorizes w.r.t H, then we have:

$$P_H(\mathcal{X}) = \frac{1}{Z}\phi_1(X_1, X_2)\phi_2(X_2, X_3)\phi_3(X_3, X_4)\phi_4(X_4, X_1)$$

constructing $\phi_1(X_1, X_2)$ from given distribution, we will have 4 non-zero entries. Same holds for other ϕ_i
Let us take (0,0,1,0) data point. P(0,0,1,0) = 0 from given distribution. But,

$$\begin{aligned} P_H(0,0,1,0) &= \frac{1}{Z}\phi_1(X_1 = 0, X_2 = 0)\phi_2(X_2 = 0, X_3 = 1)\phi_3(X_3 = 1, X_4 = 0)\phi_4(X_4 = 0, X_1 = 0) \\ &= \frac{2 * 1 * 1 * 1}{Z} \\ &> 0 \end{aligned}$$

Therefore $P_H(\mathcal{X}) \neq P(\mathcal{X})$

5. (8 points) [GAUSS EYES MARKOV] A multivariate normal distribution can be viewed either as a BN or as a MN as pointed in our Friday reading. Here we will motivate the latter. Let $\mathbf{\Sigma}$ be the covariance matrix of a set of p variables X. Consider the partial covariance matrix $\Sigma_{a|b} := \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$ between the two subsets of variables $X_a = (X_1, X_2)$ consisting of the first two, and X_b the rest. This is the covariance matrix between these two variables, after linear adjustment for all the rest. In the Gaussian distribution, this is the covariance matrix of the conditional distribution of $X_a|X_b$, which is also multivariate Gaussian. Define $\Theta = \Sigma^{-1}$.

   (a) (2 points) Show that $\Sigma_{a|b} = (\Theta_{aa})^{-1}$.
   sol:
   Let us define Σ and Σ^{-1} as follows: $\Sigma = \begin{bmatrix}\Sigma_{aa} & \Sigma_{ab}\\ \Sigma_{ba} & \Sigma_{bb}\end{bmatrix} = \begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix}$
   $\Sigma^{-1} = \begin{bmatrix}\Theta_{aa} & \Theta_{ab}\\ \Theta_{ba} & \Theta_{bb}\end{bmatrix} = \begin{bmatrix}B_{11} & B_{12}\\ B_{21} & B_{22}\end{bmatrix}$
   $\Sigma\Sigma^{-1} = \begin{bmatrix}I & 0\\ 0 & I\end{bmatrix}$
   $\begin{bmatrix}A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22}\\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{12}B_{22}\end{bmatrix} = \begin{bmatrix}I & 0\\ 0 & I\end{bmatrix}$
   Using the above equations, we get
   $A_{11}B_{11} + A_{12}B_{21} = I$ (1)
   $A_{11}B_{12} + A_{12}B_{22} = O$ (2)
   $A_{21}B_{11} + A_{22}B_{21} = O$ (3)
   $A_{21}B_{12} + A_{12}B_{22} = I$ (4)
   From equation (2) , (3)
   $B_{12} = -A_{11}^{-1}A_{12}B_{22}$
   $B_{21} = -A_{22}^{-1}A_{21}B_{11}$
   From equation (1)
   $A_{11}B_{11} + A_{12}B_{21} = I$
   $A_{11}B_{11} + A_{12}(-A_{22}^{-1}A_{21}B_{11}) = I$
   $(A_{11} - A_{12}A_{22}^{-1}A_{21}B_{11})B_{11} = I$
   $(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})\Theta_{aa} = I$
   $(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}) = (\Theta_{aa})^{-1}$
   $\Sigma_{a|b} = (\Theta_{aa})^{-1}$

   (b) (3 points) Show that if any off-diagonal element of Θ is zero, then the partial covariance between the corresponding variables is zero. Does this suggest a (minimal I-map) MN H for this distribution, with an edge between any variable pair exhibiting non-zero partial covariance? Why or why not?
   sol:

$$\Sigma_{a|b} = \Theta_{aa}^{-1}$$

   If Θ_{aa} is diagonal, then $\Sigma_{a|b}$ is diagonal. ie if $i, j (i \neq j)$ entry in Θ_{aa} then i, j entry in $\Sigma_{a|b}$ is

zero. which implies partial covariance X_i and X_j is zero.

(c) (1 point) Can any multivariate joint distbn. in the exponential family be expressed as a MN distbn. P? If so, specify the (local) factors in this Markov Network MN and their scopes (using terms in the exponential family defn. given in the last assignment).
sol: The multivariate joint distribution is given as:
$f_X(x) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$
Consier $\Theta = \Sigma^{-1}$, $\Theta\mu = \eta$

$$\begin{aligned} f_X(x) &= \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} \exp(-\frac{1}{2}(x^T\Theta x - 2x^T\Theta\mu + \mu^T\Theta\mu)) \\ &\propto \exp(-\frac{1}{2}x^T\Theta x + (\Theta\mu)^T x) \\ &\propto \exp(-\frac{1}{2}x^T\Theta x + (\eta)^T x) \end{aligned}$$

The terms with only variable X_i: $-\frac{1}{2}\Theta_{ii}x_i^2 + \eta_i x_i$
The terms that involve variable X_i, X_j:

$$-\frac{1}{2}(\Theta_{ij}x_i x_j + \Theta_{ji}x_j x_i) = \Theta_{ij}x_i x_j$$

The exponential in $f_X(x)$ has the sum of node potential given by $-\frac{1}{2}\Theta_{ii}x_i^2 + \eta_i x_i$ and edge potentials given by $\Theta_{ij}x_i x_j$. Therefore, multivariate joint distribution in the exponential family can be expressed is expressed as MN distribution.

(d) (2 points) Express a multivariate normal distribution as an exponential family in terms of its natural parameters: $\Theta = \Sigma^{-1}$ and $\eta := \Sigma^{-1}\mu$, and thereby show that the MN for this exponential family is identical to the MN derived in part (b) above.
sol:
The multivariate Gaussian distribution is given by,

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

Consier $\Theta = \Sigma^{-1}$, $\Theta\mu = \eta$

$$\begin{aligned} f_X(x) &= \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} \exp(-\frac{1}{2}(x^T\Theta x - 2x^T\Theta\mu + \mu^T\Theta\mu)) \\ &= \frac{1}{\sqrt{(2\pi)^n}} \exp(-\frac{1}{2}x^T\Theta x + (\Theta\mu)^T x - \frac{1}{2}\mu^T\Theta\mu - \frac{1}{2}\ln(|\Sigma|)) \\ &= \frac{1}{\sqrt{(2\pi)^n}} \exp(-\frac{1}{2}tr(\Theta x x^T) + (\eta)^T x - \frac{1}{2}\eta^T\Theta^{-1}\eta - \frac{1}{2}\ln(|\Sigma|)) \end{aligned}$$

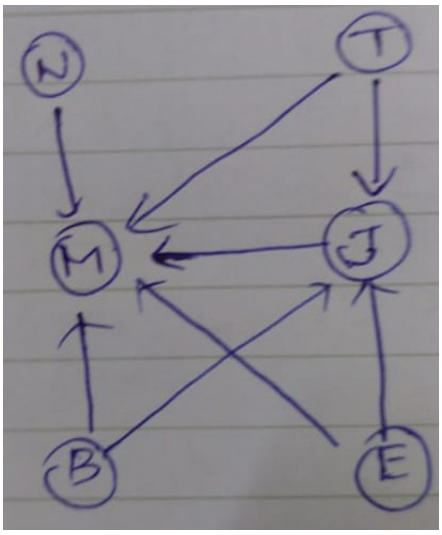
Comparing with exponential family $p(x|\theta) = h(x)\exp\left\{\eta'(\theta)^\top t(x) - a(\theta)\right\}$, we get

$$h(x) = \frac{1}{\sqrt{(2\pi)^n}}$$
$$t(x) = \begin{bmatrix} x \\ stackbycolumn(xx^\top) \end{bmatrix}$$
$$\eta'(\theta) = \begin{bmatrix} \eta \\ -\frac{1}{2} stackbyrow(\Theta) \end{bmatrix}$$
$$a(\theta) = \frac{1}{2}(\eta^\top \Theta^{-1}\eta + ln(|\Sigma|))$$

6. (7 points) [MINIMAL BN I-MAPS] Solve Exercise 3.11 from [KF] book (first two parts and the additional part below).

   (a) (3 points) Give the minimal I-map for the marginal distribution over all r.v.s except *Alarm* in the Burglary Alarm Network.
   sol:
   consider the ordering B,E,T,N,J,M,A
   P(B,E,T,N,J,M,A) = P(B)P(E | B)P(T | E,B)P(N | T,E,B)P(J | N,T,E,B)P(M | J,N,T,E,B)P(A | M,J,N,T,E,B)

$$\begin{aligned} P(B,E,T,N,J,M) &= \Sigma_A P(B,E,T,N,J,M,A) \\ &= P(B)P(E|B)P(T|E,B)P(N|T,E,B)P(J|N,T,E,B) \\ P(M|J,N,T,E,B)&\Sigma_A P(A|M,J,N,T,E,B) \\ &= P(B)P(E)P(T)P(N)P(J|B,E,T)P(M|B,E,T,N,J) \end{aligned}$$



(b) (3 points) Give an algorithm that outputs a minimal I-map G' as in part (a), but when removing (marginalizing out) a r.v. in a general DAG G.
(Hint: Recall that the minimal BN I-map is not always unique and depends on the node ordering; employ a node ordering that simplifies your algorithm by adding as few edges as possible to G to obtain G').
sol:
suppose we are marginalizing node A in G. We remove node A, all edges connected to A to get G'. Let $X_1, X_2, ..X_{n-1}$ be the toplogical order of G'.

We consider ordering $X_1, X_2, ...X_{n-1}, A$

$$\Sigma_A P(\mathcal{X}) = P(X_1)P(X2|X_1)....P(X_{n-1}|X_{n-2}, ..X_1)\Sigma_A P(A|X_{n-1}, ..X_1)$$
$$P(X_1, X_2, ..X_{n-1}) = P(X_1)P(X2|X_1)....P(X_{n-1}|X_{n-2}, ..X_1)$$

Constructing graph G' : For each X_j we pick some minimal subset U of $(X_1....X_{j-1})$ to be X_j's parents in G'. U satisfies the following property. $P(X_j|U) = P(X_j|X_1, X_2, ..X_{j-1})$

(c) (1 point) What happens if your algorithm in part (b) is applied to remove the class variable Y in the Naive Bayes model $P(Y, X) = P(Y)\prod_{i=1}^{p} P(X_i|Y)$?
sol:
The ordering chosen is $X_1, X_2, X_3, ...X_p, Y$

$$\Sigma_Y P(\mathcal{X}, \mathcal{Y}) = P(X_1)P(X2|X_1)....P(X_p|X_{p-1}, ..X_1)\Sigma_Y P(Y|X_p, ..X_1)$$
$$P(X_1, X_2, ..X_p) = P(X_1)P(X2|X_1)....P(X_P|X_{p-1}, ..X_1)$$

In the resulting graph G', every node X_i will have parents from $(X_1, X_2, ...X_{i-1})$