Assignment Three

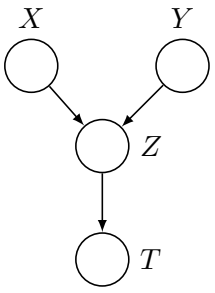
Matthew C. Scicluna
Département d'Informatique et de Recherche Opérationnelle
Université de Montréal
Montréal, QC H3T 1J4
matthew.scicluna@umontreal.ca

December 14, 2017

1 DGM

Given the following DGM G the implied factorization for any joint $p \in \mathcal{L}(G)$ is

$$p(X, Y, Z, T) = f_X(X) f_Y(Y) f_Z(Z; X, Y) f_T(T; Z)$$



It is not true that for any $p \in \mathcal{L}(G)$ $X \perp Y \mid T$. For a counterexample, take $X, Y \sim Bern\left(\frac{1}{2}\right)$, with $T = Z = X + Y$. Then,

$$P(X = 1, Y = 0 \mid Z = 1) = \frac{1}{2}$$

but

$$P(X = 1 \mid Z = 1) = P(Y = 0 \mid Z = 1) = \frac{1}{2}$$

Hence $X \not\perp Y \mid T$, for this $p \in \mathcal{L}(G)$, as needed.

2 d-Separation in DGM

(a) FALSE, consider the path $(C, A), (A, B)$

(b) TRUE

(c) FALSE, consider the path $(C, D), (D, G), (G, B)$

(d) TRUE

(e) FALSE, consider the path $(C, A), (A, B), (B, G)$

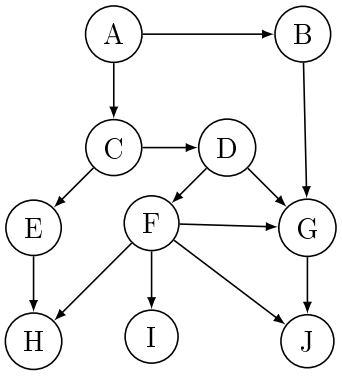
(f) FALSE, consider the path $(C, D), (D, G)$

(g) TRUE

(h) FALSE, consider the path $(C, E), (E, H), (H, F), (F, G)$

(i) TRUE

(j) FALSE, consider the path $(B, A), (A, C), (C, D), (D, F), (F, I)$



3 Positive interactions in-V-structure

Given X, Y, Z binary RV's with a joint distribution parameterized by $X \to Z \leftarrow Y$, with $a = P(X = 1), b = P(X = 1|Z = 1), c = P(X = 1|Z = 1, Y = 1)$. We notice that if we set $X \sim Bern\left(\frac{1}{2}\right)$, $Y \sim Bern\left(\frac{1}{2}\right)$ and if we denote $\alpha = P(Z = 1 \mid X = 1, Y = 1), \beta = P(Z = 1 \mid X = 1, Y = 0), \gamma = P(Z = 1 \mid X = 0, Y = 1)$ and $\delta = P(Z = 1 \mid X = 0, Y = 0)$ that we have, by cancellation of marginal probabilities of X and Y, that $c - b = \frac{\alpha}{\alpha+\gamma} - \frac{\alpha+\beta}{\alpha+\beta+\gamma+\delta}$. We then set $\alpha, \beta, \gamma, \delta$ accordingly to satisfy the inequality relations between a, b and c with fixed $a = \frac{1}{2}$.

(a) (i) $X \sim Bern\left(\frac{1}{2}\right)$, $Y \sim Bern\left(\frac{1}{2}\right)$ and $Z = 1 - X \oplus Y$. Then $a = \frac{1}{2}$ but $c = 0$

(ii) Again, we fix $X \sim Bern\left(\frac{1}{2}\right)$, $Y \sim Bern\left(\frac{1}{2}\right)$ and Z has the following probability table:

X	Y	P(Z=1)
0	0	0.1
0	1	0.8
1	0	1
1	1	1

Then $a = \frac{1}{2}$, $c = \frac{\frac{1}{2}}{0.8+0.2\frac{1}{2}} = \frac{5}{9}$ and $b = \frac{\frac{1}{2}}{0.83+0.17\frac{1}{2}} \approx 0.855$, and $a < c < b$.

(iii) Again, we fix $X \sim Bern\left(\frac{1}{2}\right)$, $Y \sim Bern\left(\frac{1}{2}\right)$ and Z has the following probability table:

X	Y	P(Z=1)
0	0	1
0	1	0.8
1	0	0
1	1	1

Then $a = \frac{1}{2}$, $c = \frac{1}{1.8}$ and $b = \frac{1}{2.8}$, and $b < a < c$.

(b) (i) The semantic here is that Z is a negated XOR gate for X and Y. Hence, knowing that both Y and Z are "on" means that X must not be.

(ii) Semantically, Z will certainly be "on" if X is, and probably will be "on" if Y is. So Z being "on" gives evidence for X, but having Y "on" as well can explain away Z (i.e. it is more likely that Z was caused by only Y).

(iii) Semantically, Z is likely to be "on" unless X is "on" and Y isn't. So if Z is "on", X is less likely to be, since Y would have to be "on" too.

4 Flipping a covered edge in a DGM

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 $E' = (E - \{(i, j)\}) \cup \{(j, i)\}$. Prove that $\mathcal{L}(G) = \mathcal{L}(G')$.

PROOF: Fix i, j, G and E. Denote π'_k as the set of parents for a node k under E'. We note that $\pi'_k = \pi_k$ for $k \neq i, j$ and $\pi'_i = \pi_i \cup \{j\}$ and $\pi'_j = \pi_i$.
For the forward direction we let $p \in \mathcal{L}(G)$. We want to show that $p(x_v) = \prod_{k=1}^{n} p(x_k|x_{\pi'_k})$
Notice that:

$$p(x_v) = \prod_{k \neq i,j} p(x_k|x_{\pi_k}) P(x_i|x_{\pi_i}) P(x_j|x_{\pi_i}, x_i) \tag{4.1}$$

$$= \prod_{k \neq i,j} p(x_k|x_{\pi_k}) P(x_i, x_j|x_{\pi_i}) \tag{4.2}$$

$$= \prod_{k \neq i,j} p(x_k|x_{\pi_k}) P(x_j|x_{\pi_i}) P(x_i|x_{\pi_i}, x_j) \tag{4.3}$$

$$= \prod_{k \neq i,j} p(x_k|x_{\pi'_k}) P(x_j|x_{\pi'_j}) P(x_i|x_{\pi'_i}) \tag{4.4}$$

$$= \prod_{k=1}^{n} p(x_k|x_{\pi'_k}) \tag{4.5}$$

As needed. For the reverse direction, let $p \in \mathcal{L}(G')$

$$p(x_v) = \prod_{k \neq i,j} p(x_k|x_{\pi'_k})P(x_i|x_{\pi'_j}, x_j)P(x_j|x_{\pi'_j}, x_i) \tag{4.6}$$

$$= \prod_{k \neq i,j} p(x_k|x_{\pi'_k})P(x_i|x_{\pi_i})P(x_j|x_{\pi_i}, x_i) \tag{4.7}$$

$$= \prod_{k \neq i,j} p(x_k|x_{\pi_k})P(x_i|x_{\pi_i})P(x_j|x_{\pi_j}) \tag{4.8}$$

And (4.5) and (4.8) complete the proof.

5 Equivalence of directed tree DGM with undirected tree UGM

Let G be a directed tree and G' be its corresponding undirected tree. Prove that $\mathcal{L}(G) = \mathcal{L}(G')$.

PROOF: For the forward direction we set the edge potentials $\psi_{(i,j)}(x_i, x_j) = P(x_j|x_i)$. For the node potentials we set $\psi_r(x_r) = P(x_r)$, where x_r is the root of the tree, and 1 for the rest. It is clear that

$$P(x_v) = \prod_{i \in V} P(x_i|x_{\pi_i}) = \psi_r(x_r) \prod_{(\pi_i,i) \in E} \psi_{(\pi_i,i)}(x_{\pi_i}, x_i)$$

We note that $Z = 1$ since

$$Z = \sum_{x_v} \psi_r(x_r) \prod_{(\pi_i,i) \in E} \psi_{(\pi_i,i)}(x_{\pi_i}, x_i)$$

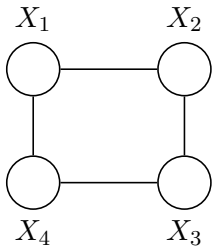
$$= \sum_{x_r} \psi_r(x_r) \prod_{(\pi_i,i) \in E} \sum_{x_i} \psi_{(\pi_i,i)}(x_{\pi_i}, x_i)$$

$$= \sum_{x_r} P(x_r) \prod_{(\pi_i,i) \in E} \sum_{x_i} P(x_i|x_{\pi_i}) = 1$$

For the reverse direction, we use that $p \in \mathcal{L}(G')$ implies that p satisfies the Global Markov property. This implies that $x_i \perp x_{nd(i)} \mid x_{\pi_i}$ for G. This is because any path from x_i to $x_{nd(i)}$ must pass through x_{π_i} since each node has at most one parent. Hence $p \in \mathcal{L}(G)$, as needed.

6 Hammersley-Clifford Counter example

Given $P(0,0,0,0) = P(1,0,0,0) = P(1,1,0,0) = P(1,1,1,0) = P(0,0,0,1) = P(0,0,1,1) = P(0,1,1,1) = P(1,1,1,1) = \frac{1}{8}$ and the following graph:



We want to show that $p \notin \mathcal{L}(G)$.

PROOF: Suppose $p \in \mathcal{L}(G)$, then there are ψ potentials such that

$$P(x_1, x_2, x_3, x_4) = \psi_{x_1,x_2}(x_1, x_2)\psi_{x_2,x_3}(x_2, x_3)\psi_{x_3,x_4}(x_3, x_4)\psi_{x_4,x_1}(x_4, x_1)$$

Notice that $P(0,1,0,0) = 0$ implies that at least one of the following must be 0: $\psi_{x_1,x_2}(0,1)$, $\psi_{x_2,x_3}(1,0)$, $\psi_{x_3,x_4}(0,0)$, $\psi_{x_4,x_1}(0,0)$

We notice that if $\psi_{x_1,x_2}(0,1) = 0$ then $P(0,1,1,1) = 0$ which contradicts that $P(0,1,1,1) = \frac{1}{8}$. Similarily, if $\psi_{x_2,x_3}(1,0) = 0$ then $P(1,1,0,0) = 0$, contradicting that it is $\frac{1}{8}$. The same reasoning shows why $\psi_{x_3,x_4}(0,0) \neq 0$. Finally, if $\psi_{x_4,x_1}(0,0) = 0$ then $P(0,0,0,0) = 0$, which again is a contradiction.

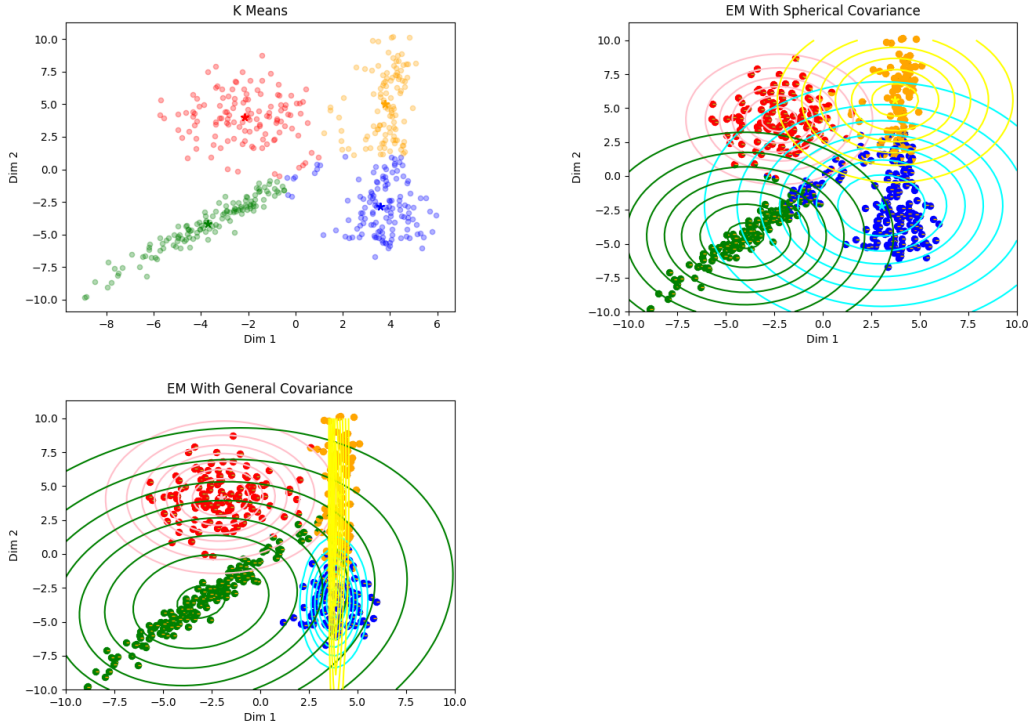


Figure 7.1: Data fit using K-Means (top left) and Mixture of Gaussians with spherical covariance (top right) and general covariance (bottom left).The data is colored with the class prediction and the contours of the distribution are plotted. The cluster centers are marked as stars.

7 Implementation: EM and Gaussian mixtures

7.1 K-means algorithm

We implement the K-means algorithm. The results were somewhat consistent between restarts, varying by around one unit. Typically the algorithm converged in around 10 steps and the ojective function (average squared distance from each points assigned cluster center) was around 6.5. To account for the discrepencies we ran averaged over a couple restarts and used this to initialize the EM Algorithm in the next section. The estimated cluster centers are presented in table 1. The results are presented in figure 1.

Table 7.1: Cluster centers for K-Means

Cluster 1	Cluster 2	Cluster 3	Cluster 4
$[-2.24.1]$	$[3.8, 5.1]$	$[-3.8, -4.3]$	$[3.3, -2.6]$

7.2 Gaussian mixture model with Spherical Covariance

The next model we trained the data on was a GMM with covariance restricted to be a multiple of the identity. We initialized the means using the centers found in K-Means, and the class covariances and marginals using the sample covariance and assignment proportions respectively. To find the M-step updates we optimize the following objective function:

$$\begin{aligned} f(q_{t+1}, \theta) &= \sum_{i=1}^{n}\sum_{j=1}^{k} \tau_{ij}^{t} log(P(x_i \mid \mu_j, \sigma_j)) + log\pi_j^t \\ &= \sum_{i=1}^{n}\sum_{j=1}^{k} \tau_{ij}^{t} \left[C - \frac{1}{2} log(\sigma_j^2) - \frac{1}{2\sigma_j^2}(x_i - \mu_j)^T(x_i - \mu_j) + log\pi_j^t \right] \end{aligned}$$

where $\theta = (\pi_j, \mu_j, \sigma_j)$ and C is a constant.
To find π_j subject to the contraint $\sum_{j=1}^{k} \pi_j = 1$ We use the method of Lagrange Multipliers. We look for any stationary points

$$\frac{\partial}{\partial \pi_j} f(q_{t+1}, \theta) + \lambda \left(1 - \sum_{j=1}^{k} \pi_j \right) = \frac{1}{\pi_j^t} \sum_{i=1}^{n} \tau_{ij}^t + \lambda = 0 \tag{7.1}$$

$$\Rightarrow \pi_j = \frac{\sum_{i=1}^{n} \tau_{ij}^t}{\lambda} \tag{7.2}$$

And substituting (7.2) and solving for λ yields

$$\frac{\partial}{\partial \lambda} f(q_{t+1}, \theta) + \lambda \left(1 - \sum_{j=1}^{k} \pi_j \right) = 1 - \sum_{j=1}^{k} \pi_j \tag{7.3}$$

$$= 1 - \frac{\sum_{j=1}^{k}\sum_{i=1}^{n} \tau_{ij}^t}{\lambda} = 0 \tag{7.4}$$

$$\Rightarrow \lambda = \sum_{j=1}^{k}\sum_{i=1}^{n} \tau_{ij}^t = n \tag{7.5}$$

Table 7.2: Parameter Values for GMM with Spherical Covariance

Cluster	π_k	μ_k	Σ_k
1	0.23	$\begin{bmatrix} -2.25 \\ 4.19 \end{bmatrix}$	$\begin{bmatrix} 2.44 & 0 \\ 0 & 2.44 \end{bmatrix}$
2	0.18	$\begin{bmatrix} 3.76 \\ 5.76 \end{bmatrix}$	$\begin{bmatrix} 2.91 & 0 \\ 0 & 2.91 \end{bmatrix}$
3	0.23	$\begin{bmatrix} -4.09 \\ -4.5 \end{bmatrix}$	$\begin{bmatrix} 3.7 & 0 \\ 0 & 3.7 \end{bmatrix}$
4	0.36	$\begin{bmatrix} 2.9 \\ -2.04 \end{bmatrix}$	$\begin{bmatrix} 4.83 & 0 \\ 0 & 4.83 \end{bmatrix}$

and so $\pi_j = \frac{\sum_{i=1}^{n} \tau_{ij}^t}{n}$. Solving for μ_j

$$\frac{\partial}{\partial \mu_j} f(q_{t+1}, \theta) = -\frac{1}{2\sigma_j^2} \sum_{i=1}^{n} \tau_{ij}^t (x_i - \mu_j) = 0 \tag{7.6}$$

$$\Rightarrow \sum_{i=1}^{n} \tau_{ij}^t x_i - \sum_{i=1}^{n} \tau_{ij}^t \mu_j = 0 \tag{7.7}$$

$$\Rightarrow \mu_j = \frac{\sum_{i=1}^{n} \tau_{ij}^t x_i}{\sum_{i=1}^{n} \tau_{ij}^t} \tag{7.8}$$

Finally, solving for σ_j

$$\frac{\partial}{\partial \sigma_j} f(q_{t+1}, \theta) = -\frac{1}{2\sigma_j^2} \sum_{i=1}^{n} \tau_{ij}^t + \frac{1}{2(\sigma_j^2)^2} \sum_{i=1}^{n} \tau_{ij}^t (x_i - \mu_j)^T (x_i - \mu_j) = 0 \tag{7.9}$$

$$\Rightarrow \sum_{i=1}^{n} \tau_{ij}^t = \frac{1}{\sigma_j^2} \sum_{i=1}^{n} \tau_{ij}^t (x_i - \mu_j)^T (x_i - \mu_j) \tag{7.10}$$

$$\Rightarrow \sigma_j^2 = \frac{\sum_{i=1}^{n} \tau_{ij}^t (x_i - \mu_j)^T (x_i - \mu_j)}{\sum_{i=1}^{n} \tau_{ij}^t} \tag{7.11}$$

Using these updates we get the parameters displayed in table 2.

7.3 Gaussian mixture model with General Covariance

The final model we trained the data on was a GMM without the covariance restriction. Again, we initialized the parameters as before. Using parameter updates from class, we get the parameter values shown in table 3.

Table 7.3: Parameter Values for GMM with General Covariance

Cluster	π_k	μ_k	Σ_k
1	0.25	$\begin{bmatrix} -2.03 \\ 4.2 \end{bmatrix}$	$\begin{bmatrix} 2.9 & 0.21 \\ 0.21 & 2.8 \end{bmatrix}$
2	0.24	$\begin{bmatrix} 4.0 \\ 4.3 \end{bmatrix}$	$\begin{bmatrix} 0.2 & 0.2 \\ 0.2 & 10 \end{bmatrix}$
3	0.31	$\begin{bmatrix} -3.0 \\ 3.5 \end{bmatrix}$	$\begin{bmatrix} 6.3 & 6.1 \\ 6.1 & 6.2 \end{bmatrix}$
4	0.206	$\begin{bmatrix} 3.8 \\ -3.6 \end{bmatrix}$	$\begin{bmatrix} 0.85 & 0.06 \\ 0.06 & 2.4 \end{bmatrix}$

Table 7.4: Log Likelihood for EM models

GMM Spherical Cov	GMM General Cov
-5.5 (-5.45)	-4.90 (-5.45)

7.4 results

We noticed that the results changed during each run, as was expected. For the shperical covariance model, we found that the variation in runs was greater, possibly because the model was not correct (since the data was certainly not from a Gaussian with this kind of covariance structure). So the optimums were certainly local (and so the algorithm was more sensitive to the K-means initializations). The general covariance structure looked like it was the true distribution, and so the optimum reached may have been a global. The normalized log-likelihoods are presented in table 4. Notice that the log likelihood for the spherical Covariance structure was smaller then the general structure since it was more restrictive, and that the test was smaller since the model was specifically trained to maximize the log likelihood of the training data.