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.

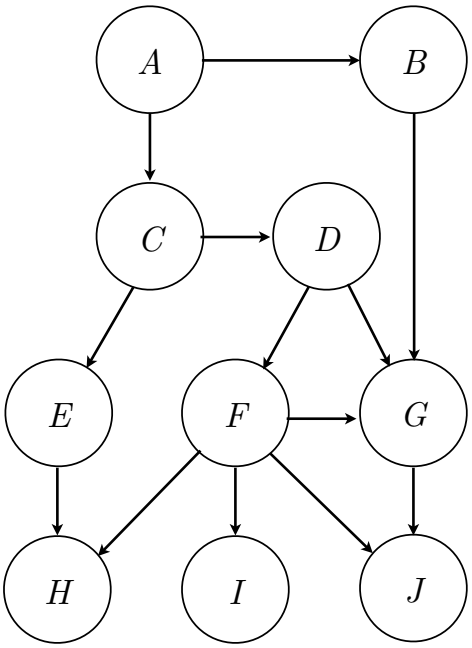
   Answer :

   As seen in class for DAGs, the implied factorization for any $p \in L(G)$ is given by the conditional probabilities of the nodes given their parents. Thus we have,

   $$p(x, y, z, t) = \prod_{v \in \{x,y,z,t\}} p(v|\pi_v) = p(x)p(y)p(z|x, y)p(t|z)$$

   Furthermore, since T is a descendent of Z, which is the child node in a V-structure, by Bayes's ball algorithm the statement $X \perp\!\!\!\perp Y|T$ does not hold for any $p \in L(G)$, since then the ball may "bounce back" up from T to either X or Y. Thus there exists some p such that $X \perp\!\!\!\perp Y|T$ is not true.

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) $B \perp I \mid J$?

Answer :

a) False, the trail C-A-B is not blocked.

b) True, all trails are blocked (A or V-structure).

c) False, the trail C-D-F-J-G-B is not blocked since J is observed in the V-structure.

d) True, all trails are again blocked (D now blocks the above path).

e) False, the trail C-D-G is not blocked.

f) False, the trail C-D-G is still not blocked.

g) True, all trails are now blocked by either B,D or H V-structure.

h) False, the trail C-E-H-F-G is no longer blocked when observing H in the V-structure.

i) True, since E blocks the path to H.

j) False, the trail B-G-J-F-I is not blocked since J is observed in the V-structure.

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: $X \to Z \leftarrow Y$. 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) $a > c$

Answer :

Given the canonical v-structure graph, the parameterization of any distribution which factors over the graph means that X and Y are marginally independent. This is the only restriction we have on constructing valid examples. Thus consider first the marginal probabilities for X and Y, which will be used to construct valid joint tables and thus valid conditional probability tables.

x	p(x)
0	0.8
1	0.2

y	p(y)
0	0.9
1	0.1

Table 1: Joint probability for X,Y,Z

	Y=0,Z=0	Y=1,Z=0	Y=0,Z=1	Y=1,Z=1
X = 0	0.5	0.01	0.22	0.07
X = 1	0.06	0.01	0.12	0.01

We can quickly check that this joint respects the marginal independence specified by the graph:

$$p_x(0)p_y(0) = 0.72 = p(X=0,Y=0,Z=0) + p(X=0,Y=0,Z=1) = p(X=0,Y=0)$$
$$p_x(0)p_y(1) = 0.08 = p(X=0,Y=1,Z=0) + p(X=0,Y=1,Z=1) = p(X=0,Y=1)$$
$$p_x(1)p_y(0) = 0.18 = p(X=1,Y=0,Z=0) + p(X=1,Y=0,Z=1) = p(X=1,Y=0)$$
$$p_x(1)p_y(1) = 0.02 = p(X=1,Y=1,Z=0) + p(X=1,Y=1,Z=1) = p(X=1,Y=1)$$

Thus this distribution factorizes over the graph. We now provide the resulting conditional probability table (with rounded probabilities to two decimals) for quantities b and c:

Table 2: P(X|Z)

	Z=0	Z=1
P(X = 0|Z)	0.88	0.69
P(X = 1|Z)	0.12	0.31

Table 3: P(X|Z,Y)

	Y=0,Z=0	Y=1,Z=0	Y=0,Z=1	Y=1,Z=1
P(X = 0|Z,Y)	0.89	0.5	0.65	0.88
P(X = 1|Z,Y)	0.11	0.5	0.35	0.12

Performing the table look-up we see that indeed,

$$a = P(X=1) = 0.2 > 0.12 = P(X=1|Y=1,Z=1) = c$$

(ii) $a < c < b$

We keep the same marginal distributions for X and Y as above, and present only the joint and conditional probability tables as the justification of independence being respected is the same as above.

Table 4: Joint probability for X,Y,Z

	Y=0,Z=0	Y=1,Z=0	Y=0,Z=1	Y=1,Z=1
X = 0	0.5	0.05	0.22	0.03
X = 1	0.06	0.01	0.12	0.01

Table 5: P(X|Z)

	Z=0	Z=1
P(X = 0|Z)	0.89	0.66
P(X = 1|Z)	0.11	0.34

Table 6: P(X|Z,Y)

	Y=0,Z=0	Y=1,Z=0	Y=0,Z=1	Y=1,Z=1
P(X = 0|Z,Y)	0.89	0.83	0.65	0.75
P(X = 1|Z,Y)	0.11	0.17	0.35	0.25

Again, we can see by table look-up that the given statement is now true with $b = 0.34, c = 0.25, a = 0.2$.

(iii) $b < a < c$.

We follow the same format again with different values of the joint for the final statement:

Table 7: Joint probability for X,Y,Z

	Y=0,Z=0	Y=1,Z=0	Y=0,Z=1	Y=1,Z=1
X = 0	0.22	0.07	0.5	0.01
X = 1	0.10	0.01	0.08	0.01

Table 8: P(X|Z)

	Z=0	Z=1
P(X = 0|Z)	0.72	0.85
P(X = 1|Z)	0.28	0.15

Table 9: P(X|Z,Y)

	Y=0,Z=0	Y=1,Z=0	Y=0,Z=1	Y=1,Z=1
P(X = 0|Z,Y)	0.69	0.88	0.86	0.5
P(X = 1|Z,Y)	0.31	0.12	0.14	0.5

Thus we see that here with $c = 0.5, a = 0.2, b = 0.15$, that the statement is verified given this distribution.

(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.

i. $a > c$

Here we have that Z = 0, given nothing else, is more likely than Z = 1. Additionally, though X is a more likely cause than Y, Y is a very strong causal variable for Z while X is a weak causal variable. Thus occurrence of Y = 1 and Z = 1 simultaneously suggests that Z = 1 was not "caused" by X = 1, and since Z = 1 is already marginally unlikely this reduces the likelihood of X = 1 with respect to its marginal. This is similar to the "alien abduction" and "watch broken" as causes for being late example given in class, where "watch broken" would be X=1.

ii. $b > c > a$

Here we only changed the probability for (X=0,Y=1,Z=1) and (X=0,Y=1,Z=0). In this case we have that occurrence of Y = 1 actually reduces the likelihood of Z = 1. So it can be seen as having more of a causal link with Z = 0 than Z = 1. Thus occurrence of Z=1,Y=1 is in fact an encouraging indicator for X = 1, since X still has the same causal link as before. This explains $c > a$. Now we also have $b > c$ since the causal link from X seems to be stronger when Y = 0. Indeed, when Y = 1 and X =1 both outcomes for Z are equally likely.

iii. $c > a > b$

Here we reversed the causal link somewhat by making X and Y causal links to the event Z = 0, which is also a priori more likely. Thus given Z = 1, it is natural that X = 1 is less likely than its marginal likelihood, giving $a > b$. The intuition for $c > a$ is that here Y and X have a similar and correlated link to Z,so co-occurence of $Y = 1, Z = 1$ means $X = 1$ is more likely as well.

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

Answer:

We will use the notation $\pi_i^{'}$ to denote the indices of the parent nodes of x_i in G'. Additionally, we note that edge (j, i) is also a covered edge in E'. Indeed, when flipping a covered edge (i, j) in E, we get that x_i is no longer a parent of x_j, so $\pi_j^{'} = \pi_i$ (subtract i from previous set of parents), and now $\pi_i^{'} = \pi_i \cup \{j\} = \pi_j^{'} \cup \{j\}$. Now, to show $L(G) = L(G')$ where G has the covered edge (i, j), take some $p \in L(G)$. Write $\{1, ..., n\}$ the set of node indices in V. We have

$$p \in L(G) \iff \tag{1}$$

$$p(x_v) = \prod_{t=1}^{n} p(x_t|x_{\pi_t}) \quad \text{(by factorization over G)} \tag{2}$$

$$= p(x_j|x_{\pi_j})p(x_i|x_{\pi_i}) \prod_{\substack{t=1 \\ t\neq i,j}}^{n} p(x_t|x_{\pi_t}) \tag{3}$$

$$= p(x_j|x_{\pi_i}, x_i)p(x_i|x_{\pi_i}) \prod_{\substack{t=1 \\ t\neq i,j}}^{n} p(x_t|x_{\pi_t}) \quad \text{(definition of } \pi_j\text{)} \tag{4}$$

$$= p(x_j, x_i|x_{\pi_i}) \prod_{\substack{t=1 \\ t\neq i,j}}^{n} p(x_t|x_{\pi_t}) \quad \text{(by chain rule on conditional probabilities)} \tag{5}$$

$$= p(x_i|x_{\pi_i}, x_j)p(x_j|x_{\pi_i}) \prod_{\substack{t=1 \\ t\neq i,j}}^{n} p(x_t|x_{\pi_t}) \quad \text{(by chain rule on conditional probabilities)} \tag{6}$$

$$= p(x_i|x_{\pi'_j}, x_j)p(x_j|x_{\pi'_j}) \prod_{\substack{t=1 \\ t\neq i,j}}^{n} p(x_t|x_{\pi'_t}) \quad \text{(equivalence in } G'\text{)} \tag{7}$$

$$= p(x_i|x_{\pi'_i})p(x_j|x_{\pi'_j}) \prod_{\substack{t=1 \\ t\neq i,j}}^{n} p(x_t|x_{\pi'_t}) \quad \text{(definition of } \pi'_i\text{)} \tag{8}$$

$$= \prod_{t=1}^{n} p(x_t|x_{\pi'_t}) \quad \text{(factorization over } G'\text{)} \tag{9}$$

$$\iff p \in L(G') \tag{10}$$

Thus $L(G) = L(G')$.

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 $\mathcal{L}(G) = \mathcal{L}(G')$.

Answer :

To show this, we first note a few things. First, since G' is constructed from G, we let π_i to denote the analogous node indices from G for G', as this is well defined. Given a definition for π_i, we also keep any topological orderings found on the nodes of G for G'. Furthermore, we can easily see that the maximal cliques within an undirected tree are of size at most 2 (exactly 2, assuming usual connectedness of trees). It is also easy to see that these cliques can be found by matching each child node with its (unique) parent node. Thus we can write the joint probability for G' as:

$$p(x_v) = \frac{1}{Z}\prod_{c\in\mathcal{C}}\psi(x_c) = \frac{1}{Z}\prod_{i=1}^{n}\psi_i(x_i, x_{\pi_i}) \tag{11}$$

where node indices follow any topological ordering obtained from G.

We also note that the potentials in equation 11 will repeat the root node, i.e. for each child x_c of a root x_r, the term $\psi(x_c, x_r)$ will appear as well as $\psi(x_r)$, but as we have seen in class for inference algorithms this is permissible without loss of generality for the underlying family of distributions, since it is simply a matter of redefining a joint potential on a root and its child. We are now ready for the proof. First, we show the easier direction $L(G) \subset L(G')$, assuming a topological ordering of the n nodes in G and G':

$$p \in L(G) \implies p(x_v) = \prod_{i=1}^{n} p(x_i|x_{\pi_i}) \tag{12}$$

Now since ψ_i's are arbitrary under the condition of positivity, we can simply choose $\psi_i(x_i, x_{\pi_i}) = p(x_i|x_{\pi_i})$, and so

$$\prod_{i=1}^{n} p(x_i|x_{\pi_i}) = \prod_{i=1}^{n}\psi_i(x_i, x_{\pi_i}) \tag{13}$$

$$= \frac{1}{Z}\prod_{i=1}^{n}\psi_i(x_i, x_{\pi_i}) \qquad (\mathrm{Z} = 1) \tag{14}$$

$$\implies p \in L(G') \tag{15}$$

Where it is easy to see that Z = 1 in equation 14 since the sums will be pushed into the conditional probability product terms when computing the normalization term and each such sum will be 1. So $L(G) \subset L(G')$. We now show the converse inclusion.

Lemma 1 *We can define the leaf plucking property for G'. Given n a leaf node in G' and $p \in G', p(x_1, ..., x_{n-1}) = \frac{1}{Z_{n-1}}\prod_{j=1}^{n-1}\psi_j(x_j, x_{\pi_j})$*

Proof

$$p(x_v) = p(x_1, ..., x_n) = \frac{1}{Z}\psi_n(x_n, x_{\pi_n})\prod_{j=1}^{n-1}\psi_j(x_j, x_{\pi_j})$$

Thus

$$p(x_1, ..., x_{n-1}) = \sum_{x_n} p(x_1, ..., x_n) = \frac{1}{Z}(\sum_{x_n}\psi_n(x_n, x_{\pi_n}))\prod_{j=1}^{n-1}\psi_j(x_j, x_{\pi_j})$$

But since $Z = \sum_{\mathbf{x}}(\prod_{i=1}^{n}\psi_i(x_i, x_{\pi_i})) = \prod_{i=1}^{n}\sum_{x_i}\psi_i(x_i, x_{\pi_i})$, then the last factor in Z cancels out and we get

$$p(x_1, ..., x_{n-1}) = \frac{1}{Z_{n-1}} \prod_{j=1}^{n-1} \psi_j(x_j, x_{\pi_j})$$

where $Z_{n-1} = \prod_{i=1}^{n-1} \sum_{x_i} \psi_i(x_i, x_{\pi_i})$ ■

Now, given a topological ordering of nodes and $p \in L(G')$, we have for a fixed i and by our above leaf-plucking property:

$$p(x_1, .., x_i) = \frac{1}{Z_i} \prod_{j=1}^{i} \psi_j(x_j, x_{\pi_j})$$

We can partition $\{x_1, ..., x_i\}$ into $\{x_i, x_{\pi_i}, x_A\}$, where x_A is the set of ancestors of x_i (not including parents). And so we get

$$p(x_i|x_{\pi_i}) = \frac{\sum_{x_A} p(x_i, x_{\pi_i}, x_A)}{\sum_{x_A} \sum_{x_i'} p(x_i', x_{\pi_i}, x_A)} = \frac{\frac{1}{Z_i}\psi_i(x_i, x_{\pi_i}) \sum_{x_A} \prod_{j<i} \psi_j(x_j, x_{\pi_j})}{\frac{1}{Z_i} \sum_{x_i'} \psi_i(x_i', x_{\pi_i'}) \sum_{x_A} \prod_{j<i} \psi_j(x_j, x_{\pi_j})} \tag{16}$$

$$= \frac{\psi_i(x_i, x_{\pi_i})}{\sum_{x_i'} \psi_i(x_i', x_{\pi_i'})} \tag{17}$$

Putting everything together to obtain the form of the full joint, we get

$$p(x_v) = \frac{1}{Z} \prod_{i=1}^{n} \psi_i(x_i, x_{\pi_i}) \tag{18}$$

$$= \frac{1}{Z} \prod_{i=1}^{n} p(x_i|x_{\pi_i}) \sum_{x_i} \psi_i(x_i, x_{\pi_i}) \qquad \text{(by equations 16 and 17)} \tag{19}$$

$$= \frac{(\prod_{i=1}^{n} p(x_i|x_{\pi_i}))(\prod_{i=1}^{n} \sum_{x_i} \psi_i(x_i, x_{\pi_i}))}{\prod_{i=1}^{n} \sum_{x_i} \psi_i(x_i, x_{\pi_i})} \qquad \text{(by the definition of Z)} \tag{20}$$

$$= \prod_{i=1}^{n} p(x_i|x_{\pi_i}) \implies p \in L(G) \tag{21}$$

Thus we also have $L(G') \subset L(G)$ and so $L(G') = L(G)$.

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.*

Answer :

The maximal cliques on this graph are $X_1 - X_2, X_2 - X_3, X_3 - X_4, X_4 - X_1$. Thus any distribution which factorizes according to G is proportional to the following potentials :

$$p(x_v) \propto \psi_1(x_1, x_2)\psi_2(x_2, x_3)\psi_3(x_3, x_4)\psi_4(x_4, x_1) \tag{22}$$

In particular, we note that any configuration with zero probability implies at least one of these potentials must be zero. Now, assume $p \in L(G)$. So p factorizes according to the above potentials. We will show that all these potentials must be strictly positive for all possible instantiations of the random variables (four for each potential). For each line below, the implication shows only *new* information with respect to the previous lines.

$$p(0,0,0,0) = \frac{1}{8} \implies \psi_1(0,0) > 0, \psi_2(0,0) > 0, \psi_3(0,0) > 0, \psi_4(0,0) > 0 \tag{23}$$
$$p(0,0,0,1) = \frac{1}{8} \implies \psi_3(0,1) > 0, \psi_4(1,0) > 0 \tag{24}$$
$$p(1,0,0,0) = \frac{1}{8} \implies \psi_1(1,0) > 0, \psi_4(0,1) > 0 \tag{25}$$
$$p(0,0,1,1) = \frac{1}{8} \implies \psi_2(0,1) > 0, \psi_3(1,1) > 0 \tag{26}$$
$$p(1,1,0,0) = \frac{1}{8} \implies \psi_1(1,1) > 0, \psi_2(1,0) > 0 \tag{27}$$
$$p(0,1,1,1) = \frac{1}{8} \implies \psi_1(0,1) > 0, \psi_2(1,1) > 0 \tag{28}$$
$$p(1,1,1,0) = \frac{1}{8} \implies \psi_3(1,0) > 0 \tag{29}$$
$$p(1,1,1,1) = \frac{1}{8} \implies \psi_4(1,1) > 0 \tag{30}$$

This shows that all 16 possible instantiations for the four potentials are strictly positive given p. But then $p(1,0,0,1) = 0$ implies that at least one of the above potentials must be zero. Contradiction. So $p \notin L(G)$.

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 ($X \perp\!\!\!\perp Z$ or $Y \perp\!\!\!\perp Z$)."

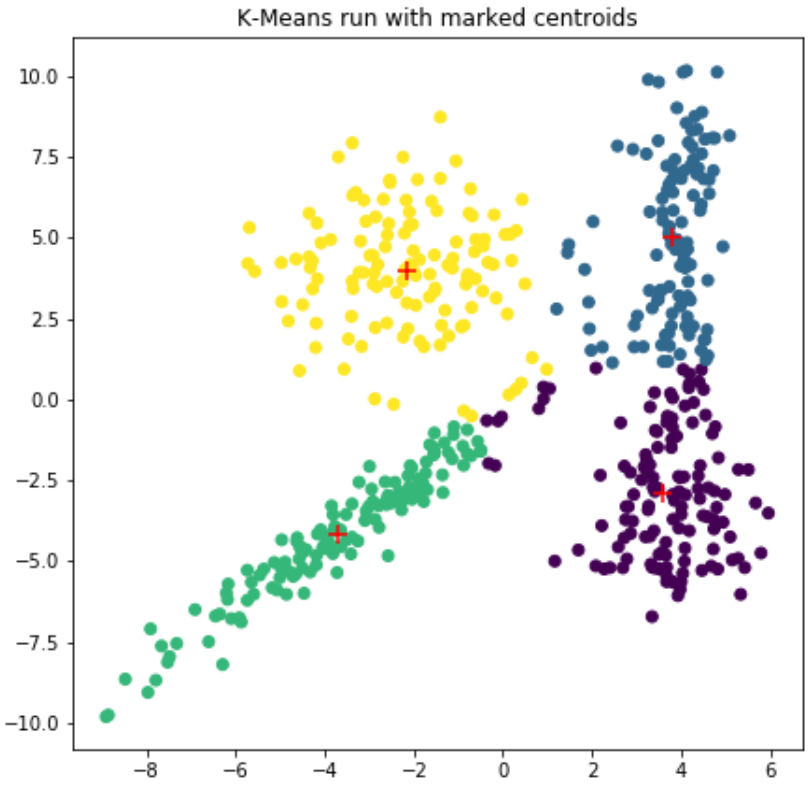
(a) Is this true if one assumes that Z is a binary variable? Prove or disprove.

(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).

Answer:
The implementation for the algorithms is found in the python file *asg3_module.py*, and the corresponding plots are in the Jupyter Notebook. Thus we simply present the generated data and clustering here.



In addition, we ran the algorithm several times and obtained the following values for the centers and distortion measures:

Run 1

Centroids: [[3.60401871 -2.88772669] [3.78809286 4.99905357] [-3.69444515 -4.16827091] [-2.14180002 3.97338429]]
Distortion measure: 3240.3817402445707

Run 2

Centroids: [[-2.14180002 3.97338429] [-3.72020481 -4.1849974] [3.78809286 4.99905357] [3.57429183 -2.880828]]
Distortion measure: 3240.1732663625185

Run 3

Centroids: [[-2.14180002 3.97338429] [-3.72020481 -4.1849974] [3.57429183 -2.880828] [3.78809286 4.99905357]]
Distortion measure: 3240.1732663625185

Run 4

Centroids: [[-3.72020481 -4.1849974] [3.57429183 -2.880828] [-2.14180002 3.97338429] [3.78809286 4.99905357]]
Distortion measure: 3240.1732663625185

Run 5

Centroids: [[-2.14180002 3.97338429] [3.78809286 4.99905357] [3.57429183 -2.880828] [-3.72020481 -4.1849974]]
Distortion measure: 3240.1732663625185

In general we see there is little variability between the achieved results, both in terms of total distortion measure and in terms of the centroids. In this run we have some initializations achieving the same centroid values. We can conclude from this that the data is robust to initialization choices for the centroids. This can be justified somewhat by the form of the data. Indeed, each cluster center is relatively well spread out from the other, giving us four distinct clusters in a two-dimensional space.

(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).

Answer:

We first derive the form of the M-step updates for the model. We have seen that the E-step consists of computing the posterior $\tau_{i,j} = p(z_{i,j} = 1|x_i, \theta)$, where we ignore the superscripts for the time steps to lighten notation. We have also seen that the M-step consists of performing maximum likelihood estimation on the *expected* complete log-likelihood, of the form

$$l(\theta, x)_{EC} = \sum_{i=1}^{n}\sum_{j=1}^{k} \tau_{i,j} \log p(x_i|\mu_j, \Sigma_j) + \tau_{i,j} \log \pi_j \tag{31}$$

which, in the case of an isotropic gaussian class-conditional (and d the dimension of the input data), is

$$l(\theta, x)_{EC} = \sum_{i=1}^{n}\sum_{j=1}^{k} \tau_{i,j}(-\frac{||(x_i - \mu_j)||^2}{2\sigma_j^2} - d\log\sigma_j - \frac{\log 2\pi}{2}) + \tau_{i,j} \log \pi_j \tag{32}$$

Let us now compute the MLE for parameters $\theta_j = (\mu_j, \sigma_j, \pi_j); j = 1, ..., k$, noting the constraint that $\sum_{j=1}^{k} \pi_j = 1$.

π_j: Eliminating constants w.r.t π_j, we get

$$\frac{\partial l(\theta, x)}{\partial \pi_j} = \frac{\partial}{\partial \pi_j} \sum_{i=1}^{n}\sum_{j=1}^{k} \tau_{i,j} \log \pi_j = \sum_{i=1}^{n} \frac{\tau_{i,j}}{\pi_j} \tag{33}$$

Given the constraint $g(\pi) = \sum_{j=1}^{k} \pi_j - 1 = 0$ and $\frac{\partial g(\pi)}{\pi_j} = 1$, by the Lagrange multipliers method we must have $\sum_{i=1}^{n} \frac{\tau_{i,j}}{\pi_j} = \lambda \implies \pi_j = \sum_{i=1}^{n} \frac{\tau_{i,j}}{\lambda}$.

But then the constraint gives

$$\sum_{j=1}^{k} \pi_j = 1 \implies \sum_{j=1}^{k}\sum_{i=1}^{n} \frac{\tau_{i,j}}{\lambda} = 1 \implies \lambda = \sum_{i=1}^{n}\sum_{j=1}^{k} \tau_{i,j} = \sum_{i=1}^{n} 1 = n \tag{34}$$

And so

$$\boxed{\hat{\pi_j}_{MLE} = \frac{\sum_{i=1}^{n} \tau_{i,j}}{n}} \tag{35}$$

μ_j: Eliminating constants w.r.t μ_j, we get

$$\nabla l(\theta, x)_{\mu_j} = \nabla_{\mu_j} \sum_{i=1}^{n} \sum_{j=1}^{k} -\tau_{i,j} \frac{||x_i - \mu_j||^2}{2\sigma_j^2} = \sum_{i=1}^{n} \tau_{i,j} \frac{(x_i - \mu_j)}{\sigma_j} \tag{36}$$

Setting equal to zero gives

$$\sum_{i=1}^{n} \tau_{i,j} x_i = \sum_{i=1}^{n} \tau_{i,j} \mu_j \implies \boxed{\hat{\mu_j}_{MLE} = \frac{\sum_{i=1}^{n} \tau_{i,j} x_i}{\sum_{i=1}^{n} \tau_{i,j}}} \tag{37}$$

σ_j: Eliminating constants w.r.t σ_j, we get

$$\frac{\partial l(\theta, x)}{\partial \sigma_j} = \frac{\partial}{\partial \sigma_j} \sum_{i=1}^{n} \sum_{j=1}^{k} -\tau_{i,j} \frac{||x_i - \mu_j||^2}{2\sigma_j^2} - \tau_{i,j} d \log \sigma_j \tag{38}$$

$$= \sum_{i=1}^{n} \tau_{i,j} \frac{||x_i - \mu_j||^2}{\sigma_j^3} - \frac{\tau_{i,j} d}{\sigma_j} \tag{39}$$

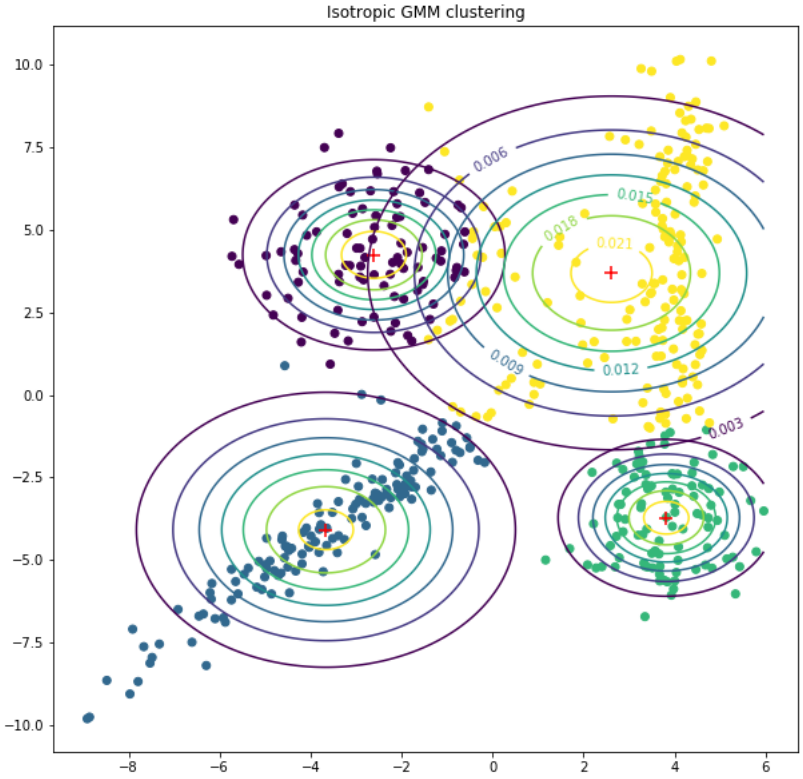
$$\tag{40}$$

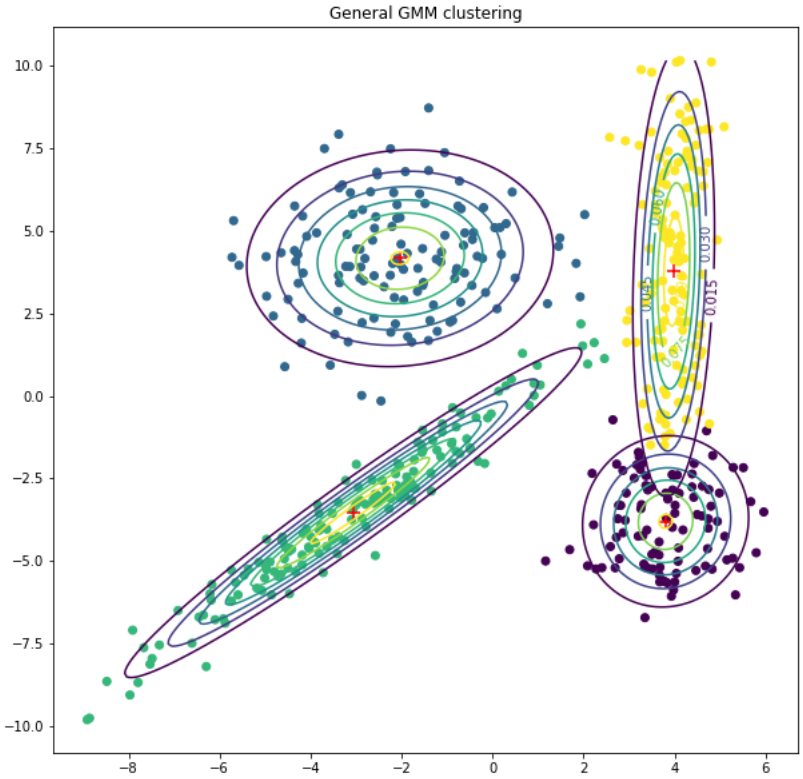
Setting to zero gives

$$\sum_{i=1}^{n} \tau_{i,j} \frac{||x_i - \mu_j||^2}{\sigma_j^3} = \frac{d \sum_{i=1}^{n} \tau_{i,j}}{\sigma_j} \implies \boxed{\hat{\sigma_j^2}_{MLE} = \frac{\sum_{i=1}^{n} \tau_{i,j} ||x_i - \mu_j||^2}{d \cdot \sum_{i=1}^{n} \tau_{i,j}}} \tag{41}$$

We now show the requested figures generated by the code for both question b an c). Means are marked with red crosses. See python file for implementation and notebook for all results.

Figure 1: Figures generated for questions b and c





(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).

Answer:
We can immediately see from the figures above that the general GMM performs much better than the one which assumes isotropic covariance. This is, of course, to be expected given that the different mixtures clearly have very different and non-isotropic covariance. This is obvious given the shape of the plotted clusters. Thus we would expect the normalized log-likelihood to be higher for the general GMM model than the isotropic one. We confirm this with the following results:

Normalized log-likelihood on training data

General GMM: -4.741765665602915
Isotropic GMM: -5.478460234659817

Normalized log-likelihood on test data

General GMM: -4.907123758599733
Isotropic GMM: -5.437695077856566

We see that our conclusion is validated : the general GMM model performs better on both the training and test sets, as expected. Comparing the plots with KMeans, it would seem that KMean clusters the data better than the isotropic GMM as well, which can be understood by the fact that KMeans is not constrained to model a circular covariance. Of course the general GMM is still better than KMeans as it is able to more or less fully capture the structure of the data, which is indeed a gaussian mixture.