Graphical models - HWK 2

Yifan WANG
Xiao CHU

1 Conditional independence and factorizations

1. Prove that $X \perp\!\!\!\perp Y|Z$ if and only if $p(x|y,z) = p(x|z)$ for all pairs (y,z), such that $p(y,z) > 0$

($\Rightarrow$) If $X \perp\!\!\!\perp Y|Z$, then by definition of conditional independence we have $\forall(x,y,z), p(x,y|z) = p(x|z)p(y|z)$. For all pairs (y,z) that $p(y,z) > 0$, we have $p(y,z) = p(y|z)p(z) > 0$, so $p(z) > 0$ and $p(y|z) > 0$.

$$p(x|y,z) \overset{(1)}{=} \frac{p(x,y,z)}{p(y,z)} \overset{(2)}{=} \frac{p(x,y|z)p(z)}{p(y|z)p(z)} \tag{1}$$

$$\overset{(3)}{=} \frac{p(x,y|z)}{p(y|z)} \overset{(4)}{=} \frac{p(x|z)p(y|z)}{p(y|z)} \tag{2}$$

$$\overset{(5)}{=} p(x|z) \tag{3}$$

(1) and (2) hold because of the definition of conditional probability. (3) holds because $p(z) > 0$. (4) holds because $X \perp\!\!\!\perp Y|Z$. (5) holds because $p(y|z) > 0$.

($\Leftarrow$) If for all pairs (y,z), such that $p(y,z) > 0$, we have $p(x|y,z) = p(x|z)$, then we have $\forall(x,y,z)$

$$p(x|y,z) \overset{(1)}{=} \frac{p(x,y,z)}{p(y,z)} \overset{(2)}{=} \frac{p(x,y|z)p(z)}{p(y|z)p(z)} \tag{4}$$

$$\overset{(3)}{=} \frac{p(x,y|z)}{p(y|z)} \overset{(4)}{=} p(x|z) \tag{5}$$

(1), (2) and (3) hold because of the same reason as above. (4) is our assumption. Thus we have $\forall(x,y,z), p(x,y|z) = p(x|z)p(y|z)$, which is equivalent to $X \perp\!\!\!\perp Y|Z$.

2. Is it true that $X \perp\!\!\!\perp Y|T$ for any $p \in \mathcal{L}(G)$? Prove or disprove.

No, it's not true.$Z \notin \{T\}$, (X,Y,Z) is a v-structure but T (descendant of Z) is in $\{T\}$, so the chain from X to Y are not blocked at Z, thus the chain from X to Y is not blocked, and $\{X\}$ and $\{Y\}$ are not d-separated by T. Therefore X and Y are not conditionally independent, and we can find a distribution $p \in \mathcal{L}(G)$ for which $X \perp\!\!\!\perp Y|T$ doesn't hold.

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

It's true. Z is a binary variable, so $p(Z=0) = 1 - p(Z=1)$. For all pairs (x,y), we can express $p(x,y)$ in two ways:

$$\begin{aligned} p(x,y) =& p(x,y,Z=0)+p(x,y,Z=1) \\ =& p(x|Z=0)p(y|Z=0)p(Z=0)+p(x|Z=1)p(y|Z=1)p(Z=1) \\ =& \frac{p(x,Z=0)p(y,Z=0)}{p(Z=0)}+\frac{p(x,Z=1)p(y,Z=1)}{p(Z=1)} \end{aligned}$$

and

$$\begin{aligned} p(x,y) =& p(x)p(y) \\ =& (p(x,Z=0)+p(x,Z=1))(p(y,Z=0)+p(y,Z=1)) \\ =& \sum_{i\in\{0,1\}}\sum_{j\in\{0,1\}} p(x,Z=i)p(y,Z=j) \end{aligned}$$

These two expressions should be equivalent, so their difference should be 0.

$$\begin{aligned} &\frac{p(x,Z=0)p(y,Z=0)}{p(Z=0)}+\frac{p(x,Z=1)p(y,Z=1)}{p(Z=1)}-\sum_{i\in\{0,1\}}\sum_{j\in\{0,1\}} p(x,Z=i)p(y,Z=j) \\ &= p(x,Z=0)(\frac{p(y,Z=0)p(Z=1)}{p(Z=0)}-p(y,Z=1))+p(x,Z=1)(\frac{p(y,Z=1)p(Z=0)}{p(Z=1)}-p(y,Z=0)) \\ &= 0 \end{aligned}$$

$\Rightarrow$

$$\begin{aligned} &\frac{p(x,Z=0)}{p(Z=0)}(\frac{p(y,Z=0)}{p(Z=0)}-\frac{p(y,Z=1)}{p(Z=1)})+\frac{p(x,Z=1)}{p(Z=1)}(\frac{p(y,Z=1)}{p(Z=1)}-\frac{p(y,Z=0)}{p(Z=0)}) \\ &= (p(x|Z=0)-p(x|Z=1))(p(y|Z=0)-p(y|Z=1))=0 \end{aligned}$$

So we have $p(x|Z=0)=p(x|Z=1)$ or $p(y|Z=0)=p(y|Z=1)$. We can deduce from the former that

$$\begin{aligned} &p(x|Z=0)=p(x|Z=1)\Rightarrow p(x,Z=1)p(Z=0)=(p(x)-p(x,Z=1))p(Z=1) \\ &\Rightarrow p(x,Z=1)=p(x)p(Z=1) \end{aligned}$$

and

$$\begin{aligned} &p(x|Z=0)=p(x|Z=1)\Rightarrow (p(x)-p(x,Z=0))p(Z=0)=p(x,Z=0)p(Z=1) \\ &\Rightarrow p(x,Z=0)=p(x)p(Z=0) \end{aligned}$$

Since Z is a binary variable, we deduce that $X \perp\!\!\!\perp Z$. Similarly, we can deduce from the later that $Y \perp\!\!\!\perp Z$. So if $X \perp\!\!\!\perp Y|Z$ and $X \perp\!\!\!\perp Y$, then $X \perp\!\!\!\perp Z$ or $Y \perp\!\!\!\perp Z$.

3.(b) Is the statement true in general? Prove or disprove.

We assume that $Z \in \{0,...,n\}$ without loss of generality.

$$\begin{aligned} p(x,y) =& \sum_{i=1}^{n} p(x,y,Z=i) \\ =& \sum_{i=1}^{n} \frac{p(x,Z=i)p(y,Z=i)}{p(Z=i)} \end{aligned}$$

and

$$
\begin{aligned}
p(x,y) =& p(x)p(y) \\
=& (\sum_{i=1}^{n} p(x, Z=i))(\sum_{i=1}^{n} p(y, Z=i)) \\
=& \sum_{i=1}^{n}\sum_{j=1}^{n} p(x, Z=i)p(y, Z=j)
\end{aligned}
$$

The difference of these two expressions shall be 0.

$$
\begin{aligned}
& \sum_{i=1}^{n} \frac{p(x, Z=i)p(y, Z=i)}{p(Z=i)} - \sum_{i=1}^{n}\sum_{j=1}^{n} p(x, Z=i)p(y, Z=j) = 0 \\
\Rightarrow & \\
& \sum_{i=1}^{n} p(x, Z=i)(\frac{p(y, Z=i)}{p(Z=i)} - \sum_{j=1}^{n} p(y, Z=j)) = 0 \\
\Rightarrow & \\
& \sum_{i=1}^{n} p(x|Z=i)p(y|Z=i)p(Z=i)(\sum_{j\neq i} p(Z=j)(1 - \frac{p(y|Z=j)}{p(y|Z=i)})) = 0
\end{aligned}
$$

2 Distributions factorizing in a graph

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

We note π_i^G set of parents of node i in DAG G, and $\pi_i^{G'}$ set of parents of node i in DAG G'. To prove $\mathcal{L}(G) = \mathcal{L}(G')$, we just need to prove that all distributions that factorize in G also factorize in G'. Let p be a distribution that factorize in G, we have $p(x) = \prod_{i=1}^{n} p(x_i|x_{\pi_i^G})$, and we want to prove we also have $p(x) = \prod_{i=1}^{n} p(x_i|x_{\pi_i^{G'}})$. Since $E' = (E \backslash \{i \to j\}) \cup \{j \to i\}$, we only need to prove that $p(x_i|x_{\pi_i^G})p(x_j|x_{\pi_j^G}) = p(x_i|x_{\pi_i^{G'}})p(x_j|x_{\pi_j^{G'}})$.

We have $\pi_j^G = \pi_i^G \cup \{i\}$, and according to the relation between E' and E, we have $\pi_i^{G'} = \pi_i^G \cup j$ and $\pi_j^{G'} = \pi_j^G \backslash i$. We can deduce that $\pi_j^{G'} = \{i\} \cup \pi_i^G \backslash \{i\} = \pi_i^G$, $\pi_i^{G'} = \pi_i^G \cup \{j\} = \pi_j^{G'} \cup \{j\}$ and $\pi_i^{G'} \backslash \{j\} = \pi_j^{G'}$.

$$
\begin{aligned}
& p(x_i|x_{\pi_i^G})p(x_j|x_{\pi_j^G}) \\
=& p(x_i|x_{\pi_i^{G'} \backslash \{j\}})p(x_j|x_{\pi_j^{G'} \cup \{i\}}) \\
=& \frac{p(x_i, x_{\pi_j^{G'}})}{p(x_{\pi_j^{G'}})} \frac{p(x_j, x_{\pi_j^{G'} \cup \{i\}})}{p(x_{\pi_j^{G'} \cup \{i\}})}
\end{aligned}
$$

Since we have $p(x_j, x_{\pi_j^{G'} \cup \{i\}}) = p(x_i, x_j, x_{\pi_j^{G'}}) = p(x_i, x_{\pi_j^{G'} \cup \{j\}}) = p(x_i, x_{\pi_i^{G'}})$ and $p(x_{\pi_j^{G'} \cup \{i\}}) = p(x_i, x_{\pi_j^{G'}})$, the above formula can be expressed as:

$$
\begin{aligned}
&p(x_i|x_{\pi_i^G})p(x_j|x_{\pi_j^G}) \\
=&\frac{p(x_i, x_{\pi_i^{G'}\setminus\{j\}})}{p(x_{\pi_i^{G'}\setminus\{j\}})}\frac{p(x_j, x_{\pi_j^{G'}\cup\{i\}})}{p(x_{\pi_j^{G'}\cup\{i\}})} \\
=&\frac{p(x_i, x_{\pi_j^{G'}})}{p(x_{\pi_j^{G'}})}\frac{p(x_{\pi_i^{G'}})}{p(x_{\pi_j^{G'}}, x_i)}\frac{p(x_i, x_{\pi_i^{G'}})}{p(x_{\pi_i^{G'}})} \\
=&\frac{p(x_j, x_{\pi_j^{G'}})}{p(x_{\pi_j^{G'}})}\frac{p(x_i, x_{\pi_i^{G'}})}{p(x_{\pi_i^{G'}})} \\
=&p(x_j|x_{\pi_j^{G'}})p(x_i|x_{\pi_i^{G'}})
\end{aligned}
$$

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

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

G is a directed tree without any v-structure, so its symmetrized graph $\hat{G}$ is the same as its moralized graph $\bar{G}$, and $\mathcal{L}(G) = \mathcal{L}(\bar{G}) = \mathcal{L}(\hat{G})$. G' is the corresponding undirected tree of G, which is by definition the symmetrized graph of G, so we have $\mathcal{L}(\hat{G}) = \mathcal{L}(G')$. Therefore we have $\mathcal{L}(G) = \mathcal{L}(G')$.

3 Entropy and Mutual Information

1.(a) Prove that the entropy H(X) is greater than or equal to zero, with equality if and only if X is a constant with probability 1.

By definition, the entropy is:

$$H(X) = -\sum_{x\in\chi} p_X(x) log p_X(x) \quad with \quad p_X(x) = P(X = x)$$

As $0 \leqslant p_X(x) \leqslant 1$ for all $x \in \chi$, we can get that $p_X(x)logp_X(x) \leqslant 0$ for all x and that $p_X(x)logp_X(x) = 0$ if and only if $p_X(x) = 0$ or $p_X(x) = 1$.

Thus we can deduce that $H(X) = -\sum_{x\in\chi} p_X(x)logp_X(x) \geqslant 0$. The equality is obtained if and only if $p_X(x)logp_X(x) = 0$ for all $x \in \chi$, which means that $p_X(x) = 0$ or $p_X(x) = 1$ for all $x \in \chi$. Besides, we should have $\sum_{x\in\chi} p_X(x) = 1$, so the equality is obtained if and only if for one x, $p_X(x) = 1$ and for the other x, $p_X(x) = 0$, which means that X is a constant with probability 1.

Therefore, we have proved that the entropy H(X) is greater than or equal to zero, with equality if and only if X is a constant with probability 1.

1.(b)What is the relation between the Kullback-Leibler divergence $D(p_X||q)$ and the entropy H(X) of the distribution $\mathbf{p}_X$?

As q follows the uniform distribution on χ and $|\chi| = k$, we have that $q(x) = \frac{1}{k}$.

Thus we can get:

$$
\begin{aligned}
D(p_X||q) &= \sum_{x\in\chi} p_X(x) log\frac{p_X(x)}{q(x)} \\
&= \sum_{x\in\chi} p_X(x) logp_X(x) + \sum_{x\in\chi} p_X(x) logk \\
&= \sum_{x\in\chi} p_X(x) logp_X(x) + log(k) \\
&= -H(X) + log(k)
\end{aligned}
$$

Therefore, we obtain that $\boxed{D(p_X||q) = -H(X) + log(k)}$.

1.(c) Deduce an upper bound on the entropy that depends on k.

According to the result in question 1(b), we have that for q the uniform distribution on χ:

$$H(X) = -D(p_X||q) + logk$$

As for all pairs of distributions (p, q), we have $D(p||q) \geqslant 0$ and that $D(p||q) = 0$ if and only if p = q.

Here we can deduce that $\boxed{H(X) \leqslant logk}$ and that $H(X) = logk$ if and only if $p_X(x) = \frac{1}{k}$.

2.(a) Prove that $I(X_1, X_2) \geqslant 0$.

Method 1:

The mutual information I(X1, X2) is defined as:

$$\begin{aligned} I(X_1, X_2) &= \sum_{(x_1,x_2)\in\chi_1\times\chi_2} p_{1,2}(x_1, x_2) log \frac{p_{1,2}(x_1, x_2)}{p_1(x_1)p_2(x_2)} \\ &= - \sum_{(x_1,x_2)\in\chi_1\times\chi_2} p_{1,2}(x_1, x_2) log \frac{p_1(x_1)p_2(x_2)}{p_{1,2}(x_1, x_2)} \end{aligned}$$

According to Jensen's Inequality, for φ a convex function, we have: $\varphi(\mathrm{E}[X]) \leq \mathrm{E}[\varphi(X)]$

Here, $\varphi(x) = -log(x)$ is a convex function and $\sum_{(x_1,x_2)\in\chi_1\times\chi_2} p_{1,2}(x_1, x_2) = 1$. So by applying Jensen's inequality, we have:

$$I(X_1, X_2) \geqslant -log[\sum_{x_1,x_2} p_{1,2}(x_1, x_2) \frac{p_1(x_1)p_2(x_2)}{p_{1,2}(x_1, x_2)}] = -log(\sum_{x_1,x_2} p_1(x_1)p_2(x_2)) = 0$$

Therefore we have proved that $\boxed{I(X_1, X_2) \geqslant 0}$.

Method 2:

As by definition,

$$I(X_1, X_2) = \sum_{x_1,x_2} p_{1,2}(x_1, x_2) log \frac{p_{1,2}(x_1, x_2)}{p_1(x_1)p_2(x_2)}$$

$$D(p||q) = \sum_{x\in\chi} p(x) log \frac{p(x)}{q(x)}$$

Thus we can express $I(X_1, X_2)$ by:

$$I(X_1, X_2) = D(p_{1,2}(x_1, x_2)||p_1(x_1) \times p_2(x_2))$$

As for all pairs of distributions (p, q), we have $D(p||q) \geqslant 0$ and that $D(p||q) = 0$ if and only if p = q.

Here, we can deduce that $\boxed{I(X_1, X_2) \geqslant 0}$ and the equality is obtained if and only if $p_{1,2}(x_1, x_2) = p_1(x_1) \times p_2(x_2)$ which means X_1 and X_2 are independent.

2.(b) Show that I(X1, X2) can be expressed as a function of H(X1), H(X2) and H(X1, X2).

According to the definition of entropy and mutual Information, we can get that:

$$\begin{aligned}
I(X_1, X_2) &= \sum_{x_1,x_2} p_{1,2}(x_1, x_2) log \frac{p_{1,2}(x_1, x_2)}{p_1(x_1)p_2(x_2)} \\
&= \sum_{x_1,x_2} p_{1,2}(x_1, x_2) log \frac{p_{1,2}(x_1, x_2)}{p_1(x_1)} - \sum_{x_1,x_2} p_{1,2}(x_1, x_2) log p_2(x_2) \\
&= \sum_{x_1,x_2} p_1(x_1) p(x_2|x_1) log p(x_2|x_1) - \sum_{x_1,x_2} p_{1,2}(x_1, x_2) log p_2(x_2) \\
&= \sum_{x_1} p_1(x_1) \left(\sum_{x_2} p(x_2|x_1) log p(x_2|x_1) \right) - \sum_{x_2} log p_2(x_2) \left(\sum_{x_1} p_{1,2}(x_1, x_2) \right) \\
&= -\sum_{x_1} p_1(x_1) H(X_2|X_1 = x_1) - \sum_{x_2} (log p_2(x_2)) p_2(x_2) \\
&= -H(X_2|X_1) + H(X_2)
\end{aligned}$$

Besides, we have:

$$\begin{aligned}
H(X_2|X_1) &= \sum_{x_1,x_2} p_{1,2}(x_1, x_2) log \frac{p_1(x_1)}{p_{1,2}(x_1, x_2)} \\
&= -\sum_{x_1,x_2} p_{1,2}(x_1, x_2) log p_{1,2}(x_1, x_2) + \sum_{x_1,x_2} p_{1,2}(x_1, x_2) log p_1(x_1) \\
&= H(X_1, X_2) + \sum_{x_1} p_1(x_1) log p_1(x_1) \\
&= H(X_1, X_2) - H(X_1)
\end{aligned}$$

Therefore, we can deduce that:

$$\begin{aligned}
I(X_1, X_2) &= -H(X_2|X_1) + H(X_2) \\
&= -\left(H(X_1, X_2) - H(X_1)\right) + H(X_2) \\
&= H(X_1) + H(X_2) - H(X_1, X_2)
\end{aligned}$$

Thus we can get that $\boxed{I(X_1, X_2) = H(X_1) + H(X_2) - H(X_1, X_2)}$.

2.(c) What is the joint distribution p1,2 of maximal entropy with given marginals p1 and p2?

From question 2(b), we know that:

$$H(X_1, X_2) = H(X_1) + H(X_2) - I(X_1, X_2)$$

As we have proved that $I(X_1, X_2) \geqslant 0$ and the equality is obtained if and only if $p_{1,2}(x_1, x_2) = p_1(x_1) \times p_2(x_2)$ which means X_1 and X_2 are independent in the question 2(a). Here, we can deduce that:

$$H(X_1, X_2) \leqslant H(X_1) + H(X_2)$$

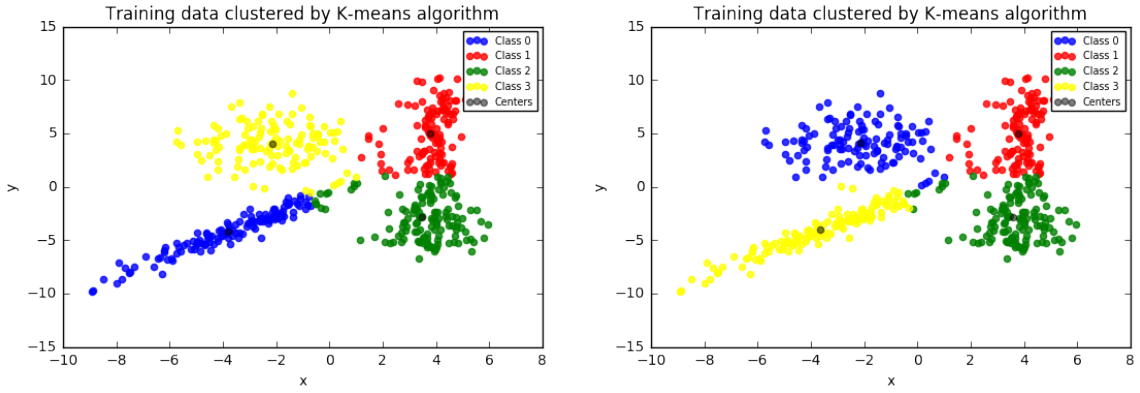
The equality is obtained if and only if $I(X_1, X_2) = 0$, which means $p_{1,2}(x_1, x_2) = p_1(x_1) \times p_2(x_2)$.

Thus the joint distribution p1,2 of maximal entropy is $\boxed{p_{1,2}(x_1, x_2) = p_1(x_1) \times p_2(x_2)}$.

4 Implementation - Gaussian mixtures

4.(a) Implement the K-means algorithm. Try several random initializations and compare results (centers and distortion measures).

We implement the K-means algorithm several times and we represent graphically two clustering results as shown in the Figure 2 as an example.



(a) Clustering result 1

(b) Clustering result 2

Figure 1: Clustering results of training data with K-means algorithm

It can be observed from Figure 1 that the cluster centers in these two clustering results are almost the same although there is a little difference. Generally speaking, the form of the four clusters obtained by two random initializations are very similar though for some points (especially for points at the edge of classes), the clustering results maybe different.

We define the distortion J(, z) by:

$$J(\mu, z) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_i^k \left\| x_i - \mu_k \right\|^2$$

where $_k$ is the center of the cluster k and $z_i^k = 1$ if x_i belongs to the cluster k, $z_i^k = 0$ otherwise.

We measure also the distortion and the number of steps, and we find that K-means converges after several number of steps and the final distortion obtained by different initializations are very closed but not identical.

Therefore, we can say that K-means converges in a finite number of steps. The convergence could be local and the result depends on the initialization. So in practice it is necessary to try several restarts of the algorithm with a random initialization to have chances to obtain a better solution.

4.(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.

For Gaussian mixture model, we have:

$$p(x) = \sum_{k=1}^{K} \pi_k N(x; \mu_k, \Sigma_k)$$

where $N(x; \mu_k, \Sigma_k) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_k|}} exp(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k))$, and for this data set d=2.

Suppose that $z_n \sim \prod_{k=1}^{K} \pi_k^{1(z_n=k)}$, we have: $x_n | z_n, (\mu_k, \Sigma_k)_k \sim \prod_{k=1}^{K} N(x_n; \mu_k, \Sigma_k)^{1(z_n=k)}$.

We apply the EM algorithm to the Gaussian mixture model.

E-Step:

Suppose that $\theta = (\pi_{1:K}, \mu_{1:K}, \Sigma_{1:K})$, we want to calculate:

$$L_t(\theta) = E[logp(X_{1:N}, Z_{1:N}|\theta)]_{p(Z_{1:N}|X_{1:N}, \theta_t)}$$

As we have:

$$\begin{aligned} logp(X_{1:N}, Z_{1:N}|\theta) &= logp(X_{1:N}|Z_{1:N}, \theta) + logp(Z_{1:N}|\theta) \\ &= \sum_{n=1}^{N} logp(x_n|z_n, \theta) + \sum_{n=1}^{N} logp(z_n|\theta) \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} 1(z_n = k) logN(x_n; \mu_k, \Sigma_k) + \sum_{n=1}^{N}\sum_{k=1}^{K} 1(z_n = k) log\pi_k \end{aligned}$$

Therefore, we have:

$$\begin{aligned} L_t(\theta) &= \sum_{n=1}^{N}\sum_{k=1}^{K} E[1(z_n = k)] logN(x_n; \mu_k, \Sigma_k) + \sum_{n=1}^{N}\sum_{k=1}^{K} E[1(z_n = k)] log\pi_k \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} p(z_n = k|x_n, \theta_t) logN(x_n; \mu_k, \Sigma_k) + \sum_{n=1}^{N}\sum_{k=1}^{K} p(z_n = k|x_n, \theta_t) log\pi_k \end{aligned}$$

We define $\gamma_k^{(t)}(x_n)$ as:

$$\begin{aligned} \gamma_k^{(t)}(x_n) &= p(z_n = k|x_n, \theta_t) \\ &= \frac{p(x_n|z_n = k, \theta_t) p(z_n|\theta_t)}{p(x_n|\theta_t)} \\ &= \frac{\pi_k^{(t)} N(x_n; \mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{i=1}^{K} \pi_i^{(t)} N(x_n; \mu_i^{(t)}, \Sigma_i^{(t)})} \end{aligned}$$

Here, as the covariance matrices are proportionalto the identity, we can suppose that: $\Sigma_k = \alpha_k^2 I_2$.
Thus we have:

$$\boxed{\gamma_k^{(t)}(x_n) = \frac{\pi_k^{(t)} \frac{1}{2\pi(\alpha_k^{(t)})^2} exp(-\frac{1}{2(\alpha_k^{(t)})^2}(x_n - \mu_k^{(t)})^T(x_n - \mu_k^{(t)}))}{\sum_{i=1}^{K} \pi_i^{(t)} \frac{1}{2\pi(\alpha_i^{(t)})^2} exp(-\frac{1}{2(\alpha_i^{(t)})^2}(x_n - \mu_i^{(t)})^T(x_n - \mu_i^{(t)}))}}$$

We need to estimate the values of α_i. The log likelyhood of the data in cluster k is (we neglect (t) for simplicity)

$$l_k = \sum_{z_i = k} -\frac{1}{2} log((2\pi)^d \alpha_k^4) - \frac{1}{2\alpha_k^2}(x_i - \mu_k)^2$$

By setting its derivative to zero, we get the maximum likelihood estimation of α^2 $\hat{\alpha}^2_{MLE} = \frac{\sum_{z_i=k}(x_i - \mu_k)^2}{2n_k}$ where $N_k = N \times \pi_k$ is the number of samples in cluster k.
Thus we can deduce that:

$$\boxed{L_t(\theta) = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma_k^{(t)}(x_n) logN(x_n; \mu_k^{(t)}, \Sigma_k^{(t)}) + \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma_k^{(t)(t)}(x_n) log\pi_k^{(t)}}$$

where $\Sigma_k^{(t)} = (\alpha_k^{(t)})^2 I_2$

M-Step:

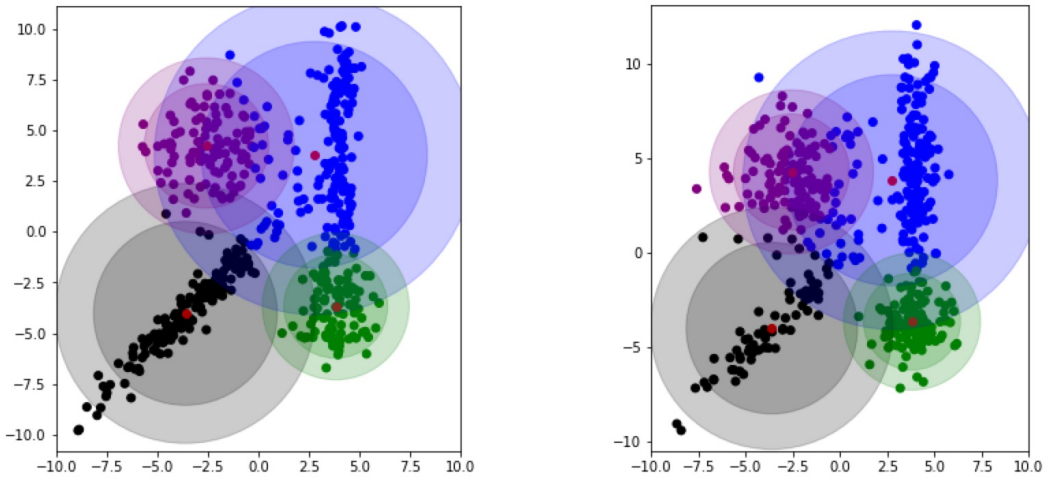
$\theta_{t+1} = (\pi_{1:K}^{(t+1)}, \mu_{1:K}^{(t+1)}, \Sigma_{1:K}^{(t+1)}) = \text{argmax}_\theta L_t(\theta)$

Therefore, we can deduce that:

$$\begin{cases} \frac{\partial L_t(\theta)}{\partial \pi_k} = 0 & \Rightarrow \boxed{\pi_k^{(t+1)} = \frac{1}{N}\sum_{n=1}^{N} \gamma_k^{(t)}(x_n)} \\ \frac{\partial L_t(\theta)}{\partial \mu_k} = 0 & \Rightarrow \boxed{\mu_k^{(t+1)} = \frac{\sum_{n=1}^{N} x_n \gamma_k^{(t)}(x_n)}{\sum_{n=1}^{N} \gamma_k^{(t)}(x_n)}} \\ \frac{\partial L_t(\theta)}{\partial \alpha_k} = 0 & \Rightarrow \boxed{(\alpha_k^{(t+1)})^2 = \frac{\sum_{n=1}^{N} (x_n - \mu_k^{(t)})^2 \gamma_k^{(t)}(x_n)}{2\sum_{n=1}^{N} \gamma_k^{(t)}(x_n)}} \end{cases}$$

We implement this EM algorithm by using an initialization with K-means.

The training data, the centers, as well as the covariance matrices are presented graphically in the Figure 2a. Here, we represent the ellipse that contains respectively 90% and 99% percentage of the mass of the Gaussian distribution. We apply this GMM model to the test data to estimate our model and the result is shown in the Figure 2b. The four classes are presented by different colors and the centers are shown in red points.



(a) Training data (b) Test data

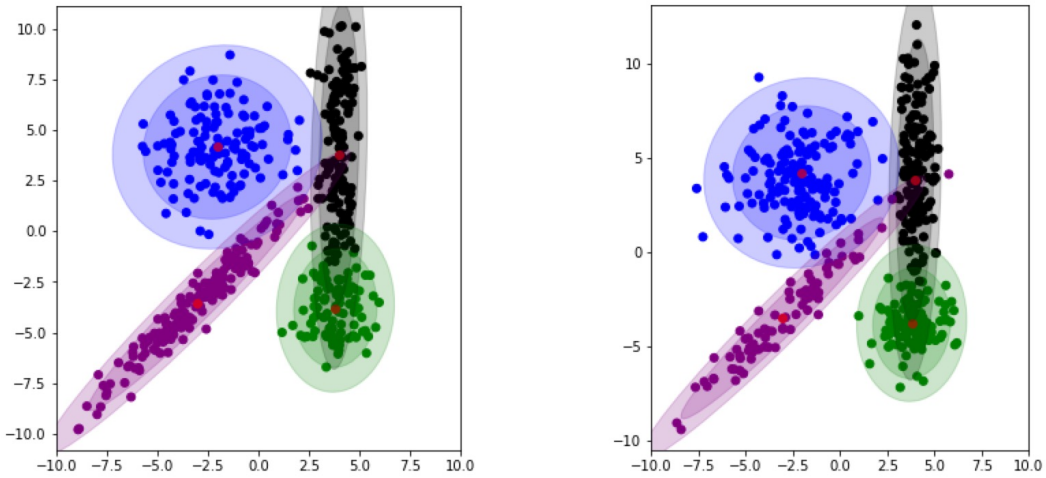
Figure 2: Clustering results with GMM (Spherical covariance matrice)

4. (c) Implement the EM algorithm for a Gaussian mixture with general covariance.

To implement the EM algorithm for a Gaussian mixture with general covariance, we just need to modify the update of covariance.

$$\Sigma_k^{(t)} = \frac{\sum_{n=1}^{N} (x_n - \mu_k^{(t)})(x_n - \mu_k^{(t)})^T \gamma_k^{(t)}(x_n)}{\sum_{n=1}^{N} \gamma_k^{(t)}(x_n)}$$

The training data, the centers, as well as the covariance matrices are presented graphically in the Figure 3a. Here, we represent the ellipses that contain respectively 90% and 99% percentage of the mass of the multivariate Gaussian distribution. We apply this GMM model to the test data to estimate our model and the result is shown in the Figure 3b. The four classes are presented by different colors and the centers are shown in red points.



(a) Training data (b) Test data

Figure 3: Clustering results with GMM (General covariance matrice)

(d) Comment the different results obtained in earlier questions. In particular, compare the log-likelihoods of the two mixture models on the training data, as well as on test data.

To better compare the log-likelihoods, we measure the average log-likelihood which is $\frac{L_t(\theta)}{N}$ for both spherical covariance matrix and general covariance matrix as showed in the Table 1.

$\frac{log-likelihood}{N}$	Training data	Test data
GMM(Spherical covariance matrice)	-5.47	-5.43
GMM(General covariance matrice)	-4.74	-4.91

Table 1: Average log-likelihood of the two mixture model

Intuitively, we find that GMM with general covariance matrix gives us better clusters than GMM with spherical covariance matrix. The log-likelihood metrics also demonstrate that the former has a better performance as its average log-likelihood is bigger.

Both the training and test dataset contain two clusters that looks very different from spheres, so if we use a spherical covariance matrix, the data points in the cluster far from the center will be more likely to be incorrectly clustered. In this case, a general covariance matrix can better describe the distribution of the data points.