Probabilistic Graphical Models

Devoir maison 2 :

Othmane SAYEM
Pirashanth RATNAMOGAN
Promotion MVA 2018

Professeurs :
M. Francis Bach
M. Guillaume Obozinski

11/10/2017

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 $p(y,z) > 0$

We will prove the equivalence in two steps.

- $X \perp\!\!\!\perp Y|Z \Rightarrow p(x|y,z) = p(x|z)$ for all pairs (y,z) such that $p(y,z) > 0$:

Let us consider a pair (y,z) in Y and Z value set such that $p(y,z) > 0$, and an element x of X value set. By definition we have that :

$$\begin{aligned} p(x|y,z) &= \frac{p(x,y,z)}{p(y,z)} \\ &= \frac{p(x,y|z)p(z)}{p(y|z)p(z)} \quad \text{with } p(x,y|z) \underset{\text{hyp}}{=} p(x|z)p(y|z) \\ &= p(x|z) \end{aligned}$$

- $X \perp\!\!\!\perp Y|Z \Leftarrow p(x|y,z) = p(x|z)$ for all pairs (y,z) such that $p(y,z) > 0$:

Let us consider a pair (y,z) in Y and Z value set such that $p(y,z) > 0$, and an element x of Xvalue set. We have that

$$\begin{aligned} p(x,y|z) =& \frac{p(x,y,z)}{p(z)} \\ \underset{\text{chain rule}}{=}& \frac{p(x|y,z)p(y|z)p(z)}{p(z)} \quad \text{with } p(x|y,z) \underset{\text{hyp}}{=} p(x|z) \\ =& p(x|z)p(y|z) \end{aligned}$$

That's exactly the definition of $X \perp\!\!\!\perp Y|Z$

We have hence proved that

$$\boxed{\text{X}\perp\!\!\!\perp Y|Z \text{ if and only if } p(x|y,z) = p(x|z) \text{ for all pairs } (y,z) \text{ such that } p(y,z) > 0}$$

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

For any joint distribution $p \in \mathcal{L}(G)$ we have

$$p(x,y,z,t) = p(x)p(y)p(z|y,x)p(t|z) \quad \forall x,y,z,t$$

The bayes ball algorithm let us understand that $X \not\perp\!\!\!\perp Y|T$ for any $p \in \mathcal{L}(G)$. We will now consider an example to show that the assertion is false.

We will consider a special case of our graphical model where X, Y, Z, T are binary variables. We will moreover give numerical values for the probability distributions of each data respecting the graphical model.

$$p(X=1)=0.5 \quad p(Y=1)=0.2$$

p(Z=1|X,Y)	X=0	X=1
Y=0	0.01	0.8
Y=1	0.8	0.95

p(T=1|Z)	Z=0	Z=1
	0.2	0.7

With those numerical values we have that :

$$\begin{aligned}
p(X=1,Y=1,T=1) &= p(X=1,Y=1,T=1,Z=0)+p(X=1,Y=1,T=1,Z=1)\\
&= p(X=1)p(Y=1)p(Z=0|Y=1,X=1)p(T=1|Z=0)\\
&+ p(X=1)p(Y=1)p(Z=1|Y=1,X=1)p(T=1|Z=1)\\
&= 0.001+0.0665\\
&= 0.0675
\end{aligned}$$

$$\begin{aligned}
p(Y=1,T=1) &= p(X=1,Y=1,T=1,Z=0)+p(X=1,Y=1,T=1,Z=1)\\
&+ p(X=0,Y=1,T=1,Z=0)+p(X=0,Y=1,T=1,Z=1)\\
&= p(X=1)p(Y=1)p(Z=0|Y=1,X=1)p(T=1|Z=0)\\
&+ p(X=1)p(Y=1)p(Z=1|Y=1,X=1)p(T=1|Z=1)\\
&+ p(X=0)p(Y=1)p(Z=0|Y=1,X=0)p(T=1|Z=0)\\
&+ p(X=0)p(Y=1)p(Z=1|Y=1,X=0)p(T=1|Z=1)\\
&= 0.001+0.0665+0.004+0.0560\\
&= 0.1275
\end{aligned}$$

$$\begin{aligned}
p(X=1,T=1) &= p(X=1,Y=1,T=1,Z=0)+p(X=1,Y=1,T=1,Z=1)\\
&+ p(X=1,Y=0,T=1,Z=0)+p(X=1,Y=0,T=1,Z=1)\\
&= p(X=1)p(Y=1)p(Z=0|Y=1,X=1)p(T=1|Z=0)\\
&+ p(X=1)p(Y=1)p(Z=1|Y=1,X=1)p(T=1|Z=1)\\
&+ p(X=1)p(Y=0)p(Z=0|Y=0,X=1)p(T=1|Z=0)\\
&+ p(X=1)p(Y=0)p(Z=1|Y=0,X=1)p(T=1|Z=1)\\
&= 0.001+0.0665+0.0160+0.2240\\
&= 0.3075
\end{aligned}$$

$$\begin{aligned}
p(T=1) &= p(X=1,Y=1,T=1,Z=0)+p(X=1,Y=1,T=1,Z=1)\\
&+p(X=0,Y=1,T=1,Z=0)+p(X=0,Y=1,T=1,Z=1)\\
&+p(X=1,Y=0,T=1,Z=0)+p(X=1,Y=0,T=1,Z=1)\\
&+p(X=0,Y=0,T=1,Z=0)+p(X=0,Y=0,T=1,Z=1)\\
&=p(X=1)p(Y=1)p(Z=0|Y=1,X=1)p(T=1|Z=0)\\
&+p(X=1)p(Y=1)p(Z=1|Y=1,X=1)p(T=1|Z=1)\\
&+p(X=0)p(Y=1)p(Z=0|Y=1,X=0)p(T=1|Z=0)\\
&+p(X=0)p(Y=1)p(Z=1|Y=1,X=0)p(T=1|Z=1)\\
&+p(X=1)p(Y=0)p(Z=0|Y=0,X=1)p(T=1|Z=0)\\
&+p(X=1)p(Y=0)p(Z=1|Y=0,X=1)p(T=1|Z=1)\\
&+p(X=0)p(Y=0)p(Z=0|Y=0,X=0)p(T=1|Z=0)\\
&+p(X=0)p(Y=0)p(Z=1|Y=0,X=0)p(T=1|Z=1)\\
&=0.001+0.0665+0.004+0.0560+0.0160+0.2240+0.0792+0.0028\\
&=0.4495
\end{aligned}$$

That allows us to calculate :

$$\begin{aligned}
p(X=1,Y=1|T=1) &= \frac{p(X=1,Y=1,T=1)}{p(T=1)}\\
&=0.1502\\
p(X=1|T=1) &= \frac{p(X=1,T=1)}{p(T=1)}\\
&=0.6841\\
p(Y=1|T=1) &= \frac{p(Y=1,T=1)}{p(T=1)}\\
&=0.2836
\end{aligned}$$

And we finaly have that

$$p(X=1,Y=1|T=1)=0.1502\neq 0.1940=p(X=1|T=1)p(Y=1|T=1)$$

So for the given graphical model we have that

$$\boxed{\text{It is false that X} \perp\!\!\!\perp Y|T \text{ for any } p\in\mathcal{L}(G)}$$

(3) Let (X,Y,Z) be a r.v. on a finite space. Consider the following statement : "If $X \perp\!\!\!\perp Y|Z$ 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

Considering Z as a binary variable, we will write $\overline{z}=\begin{cases}0 & \text{if } z=1\\ 1 & \text{if } z=0\end{cases}$.

Using the hypothesis of independence of X and Y we have for all (x, y, z) that :

$$\begin{aligned}
p(x,y) &= p(x)p(y) \\
&= (p(x|z)p(z) + (1-p(z))p(x|z))(p(y|z)p(z) + (1-p(z))p(y|z) \\
&= p(x|z)p(y|z)p^2(z) + p(x|\overline{z})p(y|\overline{z})(1-p(z))^2 \\
&+ (p(x|z)p(y|\overline{z}) + p(y|z)p(x|\overline{z}))(1-p(z))p(z)
\end{aligned}$$

We can also write using the hypothesis that $X \perp\!\!\!\perp Y|Z$: for all (x, y, z)

$$\begin{aligned}
p(x,y) &= p(x,y|z)p(z) + p(x,y|\overline{z})(1-p(z)) \\
&= p(x|z)p(y|z)p(z) + p(x|\overline{z})p(y|\overline{z})(1-p(z))
\end{aligned}$$

Subtracting the two equations of $p(x, y)$ we get the equation :

$$\begin{aligned}
p(x|z)p(y|z)p(z)(1-p(z)) + p(x|\overline{z})p(y|\overline{z})(1-p(z))p(z) \\
-(p(x|z)p(y|\overline{z}) + p(y|z)p(x|\overline{z}))(1-p(z))p(z) = 0
\end{aligned} \tag{0.1}$$

If Z is deterministic the assertion is obviously true, hence we will consider that $0 < p(z) < 1 \quad \forall z$

Hence 0.1 becomes

$$\begin{aligned}
& p(x|z)p(y|z) + p(x|\overline{z})p(y|\overline{z}) - p(x|z)p(y|\overline{z}) - p(y|z)p(x|\overline{z}) = 0 \\
\Rightarrow \quad & (p(x|\overline{z}) - p(x|z))(p(y|z) - p(y|\overline{z})) = 0 \\
\Rightarrow \quad & (p(x|\overline{z}) = p(x|z)) \text{ or } (p(y|z) = p(y|\overline{z})) \\
\Rightarrow \quad & X \perp\!\!\!\perp Z \text{ or } Y \perp\!\!\!\perp Z
\end{aligned}$$

The last implication is given because $(p(x|\overline{z}) = p(x|z))$ implies that $(p(x|\overline{z}) = p(x|z) = p(x)$ using the summation rule. Hence $\forall z \quad (p(x|z) = p(x))$ or $(p(y|z) = p(y))$. That allows us to conclude :

$$\boxed{\text{If X}\perp\!\!\!\perp Y|Z \text{ and } X \perp\!\!\!\perp Y \text{ and } Z \text{ is a binary variable then } (X \perp\!\!\!\perp Z \text{ or } Y \perp\!\!\!\perp Z)}$$

(b) Is the statement true in general 1 ? Prove or disprove.

Now we will consider the general case of the previous statement where Z is not a binary variable. We will note $\mathcal{Z} = (z_i)_{i \in \mathbb{N}}$ the space where Z takes value. We will also generalize the notation that we have introduced before for a given z we will define $\overline{z}$ such that

$$\{Z = \overline{z}\} = \{or(Z = z') \quad \forall z' \in \mathcal{Z} \backslash \{z\}\} = \{Z \neq z\} = \overline{\{Z = z\}}$$

We have obviously that for all z $p(Z = \overline{z}) = 1 - p(z)$ by construction. Using this formulation we can actually compute the same calculus that we considered in the previous question, for all (x, y), and for all $z_i \in \mathcal{Z}$ we will have

$$\begin{aligned}
p(x,y) &= p(x)p(y) \\
&= p(x|z_i)p(y|z_i)p^2(z) + p(x|\overline{z_i})p(y|\overline{z_i})(1-p(z_i))^2 \\
&+ (p(x|z_i)p(y|\overline{z_i}) + p(y|z_i)p(x|\overline{z_i}))(1-p(z_i))p(z_i) \\
&= p(x,y|z_i)p(z_i) + p(x,y|\overline{z_i})(1-p(z_i)) \\
&= p(x|z_i)p(y|z_i)p(z_i) + p(x|\overline{z_i})p(y|\overline{z_i})(1-p(z_i))
\end{aligned}$$

All this formulations for all x, y will lead as previously (no deterministic variables) to

$$\begin{aligned}
& (p(x|\overline{z_i}) - p(x|z_i))(p(y|z_i) - p(y|\overline{z_i})) = 0 && \forall z_i \in \mathcal{Z} \\
\Rightarrow \quad & (p(x|\overline{z_i}) = p(x|z_i)) \text{ or } (p(y|z_i) = p(y|\overline{z_i})) && \forall z_i \in \mathcal{Z} \\
\Rightarrow \quad & (p(x|z_i) = p(x)) \text{ or } (p(y|z_i) = p(y)) && \forall z_i \in \mathcal{Z} \\
& \Rightarrow \exists \mathcal{Z}_x \text{ and } \mathcal{Z}_y \quad s.a \quad \forall z_x \in \mathcal{Z}_x \quad p(x|z_x) = p(x), \forall z_y \in \mathcal{Z}_y \quad p(y|z_y) = p(y) && \mathcal{Z}_x \cup \mathcal{Z}_y = \mathcal{Z}
\end{aligned}$$

But using the previous notation and our hypothesis **we can't actually prove** that ($\mathcal{Z}_x = \mathcal{Z}$ or $\mathcal{Z}_y = \mathcal{Z}$) and this condition is needed to prove that ($X \perp\!\!\!\perp Z$ or $Y \perp\!\!\!\perp Z$).

With the previous consideration, we will construct a counter-example to prove that the assertion is false. To do so, we will consider a special case of the following graphical model :

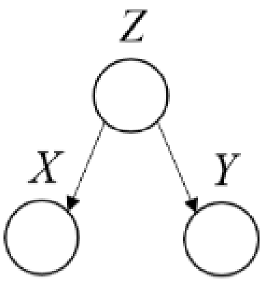

Figure 1: Common latent "cause"

This graph could be factorized as :

$$p(x, y, z) = p(z)p(x|z)p(y|z) \quad \forall x, y, z$$

Using previously introduced notation we will consider that $\mathcal{Z} = \{z_1, z_2, z_3, z_4\}$. We consider that value space of cardinal 4 because we have shown that if $\mathcal{Z}$ is binary the assertion is true, and we could easily show that if the value space of Z is of cardinal 3 the assertion is true as well. We will fix later $i \in \{1, .., 4\} \quad p(z_i)$ and in the following we will do as if this data was a given positive data that sums to 1 . Then we will set X and Y binary distributions taking respectively value in $\{x_0, x_1\}$ and $\{y_0, y_1\}$. We will use $(\alpha_0, \beta_0) \neq 0$ which are fixed real values that are fixed such as the conditional probabilities defined bellow are all between 0 and 1.

We will then define X and Y following :

$$\begin{aligned}
p(x_0|z_1) &= p(x_0), & p(x_0|z_2) &= p(x_0) \\
p(x_0|z_3) &= p(x_0) + \frac{\alpha_0}{p(z_3)}, & p(x_0|z_4) &= p(x_0) - \frac{\alpha_0}{p(z_4)} \\
p(y_0|z_1) &= p(y_0) + \frac{\beta_0}{p(z_1)}, & p(y|z_2) &= p(y_0) - \frac{\beta_0}{p(z_2)} \\
p(y_0|z_3) &= p(y_0), & p(y_0|z_4) &= p(y_0) \\
p(x_1|z_1) &= p(x_1), & p(x_1|z_2) &= p(x_1) \\
p(x_1|z_3) &= p(x_1) - \frac{\alpha_0}{p(z_3)}, & p(x_1|z_4) &= p(x_1) + \frac{\alpha_0}{p(z_4)} \\
p(y_1|z_1) &= p(y_1) - \frac{\beta_0}{p(z_1)}, & p(y_1|z_2) &= p(y_1) + \frac{\beta_0}{p(z_2)} \\
p(y_1|z_3) &= p(y_1), & p(y_1|z_4) &= p(y_1)
\end{aligned} \tag{0.2}$$

Using this formulation we have by construction that :

$$\forall z \in \mathcal{Z} \quad p(x_1|z) + p(x_0|z) = p(x_0) + p(x_1) = 1$$

$$\forall z \in \mathcal{Z} \quad p(y_1|z) + p(y_0|z) = p(y_0) + p(y_1) = 1$$

Moreover we have that

$$\begin{aligned}
p(x_0, y_0) &= \sum_i p(x_0, y_0, z_i) \\
&= \sum_i p(z_i)p(x_0|z_i)p(y_0|z_i) \\
&= p(x_0)(p(y_0) + \frac{\beta_0}{p(z_1)})p(z_1)) + p(x_0)(p(y_0) - \frac{\beta_0}{p(z_2)})p(z_2) \\
&+ p(y_0)(p(x_0) + \frac{\alpha_0}{p(z_3)})p(z_3)) + p(y_0)(p(x_0) - \frac{\alpha_0}{p(z_4)})p(z_4) \\
&= (p(z_1) + p(z_2) + p(z_3) + p(z_4))p(x_0)p(y_0) + \beta_0 p(x_0) - \beta_0 p(x_0) + \alpha_0 p(y_0) - \alpha_0 p(y_0) \\
&= p(x_0)p(y_0)
\end{aligned}$$

$$\begin{aligned}
p(x_1, y_1) &= \sum_i p(x_1, y_1, z_i) \\
&= \sum_i p(z_i)p(x_1|z_i)p(y_1|z_i) \\
&= p(x_1)(p(y_1) - \frac{\beta_0}{p(z_1)})p(z_1)) + p(x_1)(p(y_1) + \frac{\beta_0}{p(z_2)})p(z_2) \\
&+ p(y_1)(p(x_1) - \frac{\alpha_0}{p(z_3)})p(z_3)) + p(y_1)(p(x_1) + \frac{\alpha_0}{p(z_4)})p(z_4) \\
&= (p(z_1) + p(z_2) + p(z_3) + p(z_4))p(x_1)p(y_1) + \beta_0 p(x_1) - \beta_0 p(x_1) + \alpha_0 p(y_1) - \alpha_0 p(y_1) \\
&= p(x_1)p(y_1)
\end{aligned}$$

$$\begin{aligned}
p(x_0,y_1) &= \sum_i p(x_0,y_1,z_i) \\
&= \sum_i p(z_i)p(x_0|z_i)p(y_1|z_i) \\
&= p(x_0)(p(y_1) - \frac{\beta_0}{p(z_1)})p(z_1)) + p(x_0)(p(y_1) + \frac{\beta_0}{p(z_2)})p(z_2) \\
&+ p(y_1)(p(x_0) + \frac{\alpha_0}{p(z_3)})p(z_3)) + p(y_1)(p(x_0) - \frac{\alpha_0}{p(z_4)})p(z_4) \\
&= (p(z_1)+p(z_2)+p(z_3)+p(z_4))p(x_0)p(y_1) + \beta_0 p(x_0) - \beta_0 p(x_0) + \alpha_0 p(y_1) - \alpha_0 p(y_1) \\
&= p(x_0)p(y_1)
\end{aligned}$$

$$\begin{aligned}
p(x_1,y_0) &= \sum_i p(x_1,y_0,z_i) \\
&= \sum_i p(z_i)p(x_1|z_i)p(y_0|z_i) \\
&= p(x_1)(p(y_0) + \frac{\beta_0}{p(z_1)})p(z_1)) + p(x_1)(p(y_0) - \frac{\beta_0}{p(z_2)})p(z_2) \\
&+ p(y_0)(p(x_1) - \frac{\alpha_0}{p(z_3)})p(z_3)) + p(y_0)(p(x_1) - \frac{\alpha_0}{p(z_4)})p(z_4) \\
&= (p(z_1)+p(z_2)+p(z_3)+p(z_4))p(x_1)p(y_0) + \beta_0 p(x_1) - \beta_0 p(x_1) + \alpha_0 p(y_0) - \alpha_0 p(y_0) \\
&= p(x_1)p(y_0)
\end{aligned}$$

That proves that $\forall x,y \quad p(x,y) = p(x)p(y)$,that means that $X \perp\!\!\!\perp Y$. We also have because of the graphical model that we have used that for all (x,y,z) :

$$\begin{aligned}
p(x,y|z) &= \frac{p(x,y,z)}{p(z)} \\
&= \frac{p(z)p(x|z)p(y|z)}{p(z)} \\
&= p(x|z)p(y|z)
\end{aligned}$$

Hence we have that $X \perp\!\!\!\perp Y|Z$

Finally we have that $X \perp\!\!\!\perp Y|Z$ and $X \perp\!\!\!\perp Y$ but if we take $p(z_1), p(z_2), p(z_3), p(z_4) > 0$. We can using our expressions in **0.2** see that $p(x|z_3) \neq p(x)$, $p(x|z_4) \neq p(x)$ and that $p(y|z_1) \neq p(y)$,$p(y|z_2) \neq p(y)$. That shows that we have in this case we have $X \not\perp\!\!\!\perp Z$ and $Y \not\perp\!\!\!\perp Z$. That allows us to prove that the statement is false in the general case.

$$\boxed{\text{The statement : "If X}\perp\!\!\!\perp Y|Z \text{ and } X \perp\!\!\!\perp Y \text{ then } (X \perp\!\!\!\perp Z \text{ or } Y \perp\!\!\!\perp Z)\text{" }\textbf{is false}}$$

2 - Distributions factorizing in a graph

(a) Let $G = (V, E)$ be a DAG. We say that an edge $i \to j$ in E is a covered edge if and only if $\pi_j = \pi_i \cup \{i\}$; let $G' = (V, E')$, with $E' = (E - \{i \to j\})$ $\{j \to i\}$. Prove that $L(G) = L(G')$.

Let us show first that G' is a DAG. Suppose we have a cycle C in the new graph G' : since the only change that happened was switching the orientation of the edge $i \to j$, and since G is a DAG, then the cycle C in G' contains necessarily the edge $j \to i$.
Since the only parents of j in G' are $\{\pi_i\}$, then the edge $\pi_i \to j \to i$ is necessarily contained in the cycle C. Thus, there exists other vertices x_j such that the cycle C is written :$\pi_i \to j \to i \to ...x_j... \to \pi_i$. However, this means that the vertices π_i, i and x_j form a cycle in G, which is not possible since it is a DAG. Hence, we conclude that G' is a DAG.

Let $p(x) \in L(G)$:

$$\forall x, \qquad p(x) = \prod_{i=1}^{n} p(x_i|x_{\pi_i})$$

We denode by π'_i and π'_j the parents of the node i and j respectively in the graph G'. Since $i \to j$ is a covered edge in G then $i \to j$ is a covered edge in G', which means that :

$$\pi'_i = \pi_i \cup \{j\} \quad ; \qquad \pi'_j = \pi_i \quad ; \qquad \pi'_l = \pi_l \qquad \forall l \notin \{i, j\}$$

By using Bayes rule, we show that :

$$\begin{aligned}
p(x_j|x_{\pi_j}) &= p(x_j|x_{\pi_i}, x_i) \\
&= \frac{p(x_j, x_{\pi_i}, x_i)}{p(x_{\pi_i}, x_i)} \\
&= \frac{p(x_i|x_{\pi_i}, x_j) \times p(x_{\pi_i}, x_j)}{p(x_{\pi_i}, x_i)} \\
&= \frac{p(x_i|x_{\pi_i}, x_j) \times p(x_{\pi_i}, x_j)}{p(x_{\pi_i}, x_i)} \times \frac{p(x_{\pi_i})}{p(x_{\pi_i})} \\
&= p(x_i|x_{\pi_i}, x_j) \times \frac{p(x_j|x_{\pi_i})}{p(x_i|x_{\pi_i})} \\
&= p(x_i|x_{\pi'_i}) \times \frac{p(x_j|x_{\pi'_j})}{p(x_i|x_{\pi_i})}
\end{aligned}$$

We conclude then :

$$\boxed{p(x_j|x_{\pi_j}) \times p(x_i|x_{\pi_i}) = p(x_j|x_{\pi'_j}) \times p(x_i|x_{\pi'_i})}$$

Since for all other vertices we have $\pi'_l = \pi_l$, then we conclude that :

$$\forall x, \qquad p(x) = \prod_{k=1}^{n} p(x_k|x_{\pi'_k})$$

which means that $p \in L(G')$, and $\boxed{L(G) \subset L(G')}$.

To prove the other inclusion, we just have to consider the graph $G'' = (V, E'')$ such that $E'' = (E' - \{j \to i\}) \cup \{i \to j\}$. So by applying what we have just proved, we have the inclusion : $L(G') \subset L(G'')$. However, and by definition, $G = G''$, then $L(G') \subset L(G'') = L(G)$. We conclude finally that : $\boxed{L(G) = L(G')}$.

2. 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 L(G) = L(G').

Let $p(x) \in L(G)$, since G is a directed tree, each node has only one parent and there is only one node without parents : the root x_r. Thus, p factorizes G means :

$$p(x) = \prod_{v \in V} p(x_v | x_{\pi_v}) = p(x_r) \prod_{v \in V-r} p(x_v | x_{\pi_v})$$

Furthermore, for each node v, π_v is only a singleton. Thus, we can write p(x) in the following form :

$$p(x) = p(x_r) \prod_{(i,j) \in \{E\}} p(x_i | x_j)$$

we define the potential functions :

$$\psi(x_r) = p(x_r) \qquad \psi(x_i, x_j) = p(x_i | x_j) \qquad \forall (i,j) \in E$$

we re-write the probability distribution using these potentials :

$$p(x) = \psi(x_r) \prod_{(i,j) \in \{E\}} \psi(x_i, x_j)$$

To show that p factorizes the symmetrized graph G', we need to prove that p is of the form :

$$p(x) = \frac{1}{Z} \prod_{c \in C} \psi_c(x_c) \qquad \textit{where C is the set of all cliques in G'}$$

Since in our case, we have an undirected tree,all the possible cliques are either individual nodes or cliques 2 nodes. p should take the form :

$$p(x) = \frac{1}{Z} \prod_{v \in V} \psi_v(x_v) \times \prod_{(i,j) \in E'} \psi_{i,j}(x_i, x_j)$$

We already proved that

$$p(x) = \psi(x_r) \prod_{(i,j) \in \{E\}} \psi(x_i, x_j)$$

So by taking : $\psi(x_v) = 1 \quad \forall v \in V - \{r\}$ and since E=E', we conclude that :

$$p(x) = \frac{1}{Z} \prod_{v \in V} \psi_v(x_v) \times \prod_{(i,j) \in E'} \psi_{i,j}(x_i, x_j)$$

Thus, we just proved that p factorizes G'.

Let p a distribution that factorizes the undirected tree G', and let G a corresponding directed tree to G'. Since $p \in L(G')$, p has the following form :

$$\begin{aligned} p(x) &= \frac{1}{Z} \prod_{c \in C_{max}} \psi_c(x_c) \qquad \textit{where Cmax is the set of maximal cliques in G'} \\ &= \frac{1}{Z} \prod_{(i,j) \in E'} \psi(x_i, x_j) \end{aligned}$$

In the chosen oriented graph G, we denote π_i the set of parents of node i. This set is a singleton since the graph is an oriented tree, and we can write $p(x_i|x_{\pi_i})$ as the following :

$$\begin{aligned} p(x_i|x_{\pi_i}) &= \frac{p(x_i, x_{\pi_i})}{p(x_{\pi_i})} \\ &= \frac{\displaystyle\sum_{j \notin (i,\pi_i)} p(x)}{\displaystyle\sum_{j \notin (\pi_i)} p(x)} \\ &= \frac{\sum_{x_1,,x_{\pi_i-1},x_{\pi_i+1},.,x_{i-1},x_{i+1}..x_n} \quad \prod_{(l,j) \in E'} \psi(x_l, x_j)}{\sum_{x_1,.,x_{i-1},x_{i+1}..x_n} \quad \prod_{(l,j) \in E'} \psi(x_l, x_j)} \end{aligned}$$

By using the chain rule, since the parents of each node i are individual nodes, we can write :

$$p(x) = p(x_r) \prod p(x_i|x_{\pi_i})$$

We were able then to reconstruct the conditional probabilities from the potentials of p. Thus, we showed that p factorizes the oriented graph G. We have then $L(G) = L(G')$.

3 - Entropy and Mutual Information

Let X be a discrete random variable on a finite space $\mathbb{X}$ with $|\mathbb{X}| = K$. We define the entropy H(X) as following :

$$H(X) = -\sum_{x \in \mathbb{X}} p_X(x) \log p_X(x) \qquad \textit{with} \quad p_X(x) := \mathbb{P}(X = x)$$

(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 of the probability distribution, we have $0 \leq p_X(x) = \mathbb{P}(X = x) \leq 1$ for all $x \in \mathbb{X}$. Thus we conclude that :

$$-p_X(x) \log p_X(x) \geq 0 \qquad \forall x \in \mathbb{X}$$

Which implies, by summing over x in $\mathbb{X}$:

$$\boxed{H(X) = -\sum_{x\in\mathbb{X}} p_X(x)\log p_X(x) \geq 0}$$

Since all terms of the sum are greater or equal to zero, the equality $H(X) = 0$ is equivalent to :

$$p_X(x)\log p_X(x) = 0 \qquad \forall x \in \mathbb{X}$$

The function $z \to z\log z$ has been extended by continuity in 0 so that $0\log(0) = 0$, which means that the previous equation implies :

$$p_X(x) = 1 \quad \text{or} \quad p_X(x) = 0 \qquad \forall x \in \mathbb{X}$$

Using the fact that $\sum_{x\in\mathbb{X}} \mathbb{P}(X = x) = 1$ we deduce that, there exists a unique $x_0 \in \mathbb{X}$ such that :

$$\mathbb{P}(X = x_0) = p_X(x_0) = 1$$

Thus, we conclud that there is an equality $H(X) = 0$ if and only if X is constant with probability 1.

(b) What is the relation between the Kullback-Leibler divergence $D(p_X||q)$ and the entropy $H(X)$ of the distribution p_X ?

We recall that $q(X)$ is the uniform distribution on $\mathbb{X}$, which means that $q(X = x) = \frac{1}{K}$ for all x in $\mathbb{X}$. This implies that :

$$\begin{aligned} D(p_X||q) &= \sum_x p_X(x)\log\frac{p_X(x)}{q(x)} \\ &= \sum_x p_X(x)\log(p_X(x)) - \sum_x p_X(x)\log(q(x)) \\ &= \sum_x p_X(x)\log(p_X(x)) - \log(\frac{1}{K})\sum_x p_X(x) \\ &= -H(X) + \log(K) \end{aligned}$$

Thus the relation between $H(X)$ and $D(p_X||q)$: $\boxed{D(p_X||q) = -H(X) + \log(K)}$

(c) Deduce an upper bound on the entropy that depends on K

We have already proved in class that for all pairs of probability distributions (p,q) we have : $D(p||q) \geq 0$. By applying that to our case, we have $D(p_X||q) \geq 0$. By applying the result of the previous question, we conclude the following upper bound on the entropy :

$$\boxed{\log(K) \geq H(X)}$$

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

The mutual information $I(X_1, X_2)$ is defined as :

$$\begin{aligned} I(X_1, X_2) &= \sum_{(x_1,x_2)\in\mathbb{X}_1\times\mathbb{X}_2} p_{1,2}(x_1, x_2) \log \frac{p_{1,2}(x_1, x_2)}{p_1(x_1)p_2(x_2)} \\ &= \sum_{(x_1,x_2)\in\mathbb{X}_1\times\mathbb{X}_2} (\frac{p_{1,2}(x_1, x_2)}{p_1(x_1)p_2(x_2)}) \log \frac{p_{1,2}(x_1, x_2)}{p_1(x_1)p_2(x_2)} p_1(x_1)p_2(x_2) \\ &= \mathbb{E}_{(X_1,X_2)\sim q}\left[\frac{p(X_1, X_2)}{q(X_1, X_2)} \log \frac{p(X_1, X_2)}{q(X_1, X_2)}\right] \end{aligned}$$

Where we defined the probability distribution q as following : $q(x_1, x_2) = p_1(x_1)p_2(x_2)$ for $(x_1, x_2) \in \mathbb{X}_1 \times \mathbb{X}_2$.

If there exists x_1 such that $p_1(x_1) = 0$ or x_2 such that $p_2(x_2) = 0$, then $I(X_1, X_2) = +\infty$, which means $I(X_1, X_2) \geq 0$. Thus, We can suppose for the rest of the proof that $q(X_1, X_2) > 0$.

By using the convexity of the function $z \to z\log(z)$, we can apply the Jensen's inequality :

$$\mathbb{E}_q\left[\frac{p(X_1, X_2)}{q(X_1, X_2)} \log \frac{p(X_1, X_2)}{q(X_1, X_2)}\right] \geq \mathbb{E}_q\left[\frac{p(X_1, X_2)}{q(X_1, X_2)}\right] \log \mathbb{E}_q\left[\frac{p(X_1, X_2)}{q(X_1, X_2)}\right]$$

We explicit the second term of the inequality :

$$\begin{aligned} \mathbb{E}_q\left[\frac{p(X_1, X_2)}{q(X_1, X_2)}\right] &= \sum_{(x_1,x_2)\in\mathbb{X}_1\times\mathbb{X}_2} q_(x_1, x_2)\frac{p_{1,2}(x_1, x_2)}{q_(x_1, x_2)} \\ &= \sum_{(x_1,x_2)\in\mathbb{X}_1\times\mathbb{X}_2} p_{1,2}(x_1, x_2) \\ &= 1 \end{aligned}$$

By replacing this term in the Jensen's inequality, we have proved that :

$$\boxed{I(X_1, X_2) = \mathbb{E}_q\left[\frac{p(X_1, X_2)}{q(X_1, X_2)} \log \frac{p(X_1, X_2)}{q(X_1, X_2)}\right] \geq 0}$$

(b) Show that $I(X_1, X_2)$ can be expressed as a function of $H(X_1)$, $H(X_2)$ and H(X_1, X_2)

$$\begin{aligned}
I(X_1, X_2) &= \sum_{(x_1,x_2)\in\mathbb{X}_1\times\mathbb{X}_2} p_{1,2}(x_1, x_2)\log\frac{p_{1,2}(x_1, x_2)}{p_1(x_1)p_2(x_2)} \\
&= \sum_{(x_1,x_2)\in\mathbb{X}_1\times\mathbb{X}_2} p_{1,2}(x_1, x_2)(\log p_{1,2}(x_1, x_2) - \log p_1(x_1) - \log p_2(x_2)) \\
&= -H(X_1, X2) - \sum_{(x_1,x_2)} p_{1,2}(x_1, x_2)\log p_1(x_1) - \sum_{(x_1,x_2)} p_{1,2}(x_1, x_2)\log p_2(x_2) \\
&= -H(X_1, X2) - \sum_{x_1}\log p_1(x_1)\sum_{x_2} p_{1,2}(x_1, x_2) - \sum_{x_2}\log p_2(x_2)\sum_{x_2} p_{1,2}(x_1, x_2) \\
&= -H(X_1, X2) - \sum_{x_1} p_1(x_1)\log p_1(x_1) - \sum_{x_2} p_2(x_2)\log p_2(x_2) \\
&= -H(X_1, X2) + H(X_1) + H(X_2)
\end{aligned}$$

$$\boxed{I(X_1, X_2) = H(X_1) + H(X_2) - H(X_1, X2)}$$

(c) What is the joint distribution $p_{1,2}$ of maximal entropy with given marginals p_1 and p_2 ?

We want to solve the following problem :

$$\max_{p_{1,2}} H(X_1, X_2) = \max_{p_{1,2}} H(X_1)+H(X_2)-I(X_1, X2) = H(X_1)+H(X_2)-\min_{p_{1,2}} I(X_1, X2)$$

We have already proved that $I(X_1, X2) \geq 0$, and since for a joint distribution $q(x_1, x_2) = p_1(x_1)p_2(x_2)$, we have $I_q(X_1, X_2) = 0$. Then :

$$\min_{p_{1,2}} I(X_1, X2) = 0 \qquad \text{and} \qquad p^*_{1,2}(x_1, x_2) = p_1(x_1)p_2(x_2)$$

Which implies the maximal entropy :

$$\boxed{\max_{p_{1,2}} H(X_1, X2) = H(X_1) + H(X_2) \qquad \text{and} \qquad p^*_{1,2}(x_1, x_2) = p_1(x_1)p_2(x_2)}$$

4 - Implementation - Gaussian mixtures

(a) Implement the K-means algorithm. Represent graphically the training data, the cluster centers, as well as the different clusters. Try several random initial- izations and compare results (centers and distortion measures).

We have implemented the K-means algorithm using python. You can find below one of the outcome of our implementation in the training dataset. The white crosses correspond to cluster centers, black points to data positions, and colors allows to assign each part of the space to a certain cluster.

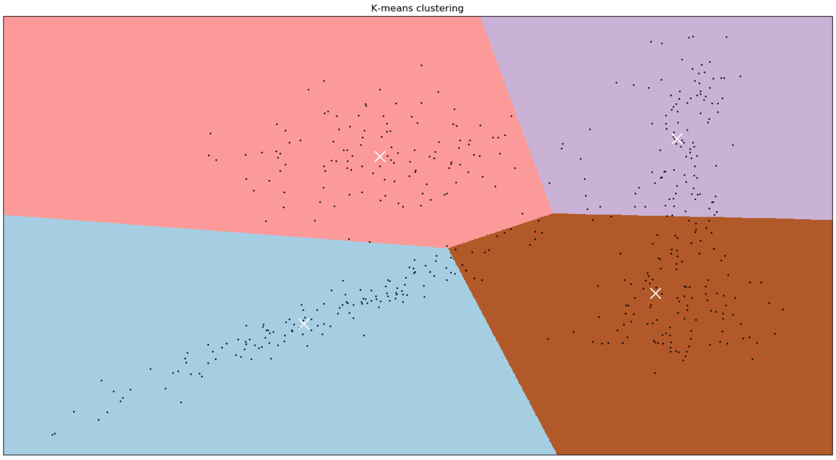

Figure 2: K-Means outcome for the training dataset

We have computed K-Means 5 times with random initializations. You can find below the centers and the distortion mesure that we obtained.

As we can see the initialization play an important part in the algorithm performances. That's why in practice people use an aggregation of multiple call to K-Means with various random initializations.

	Cluster Center positions				Distortion
	μ_1	μ_2	μ_3	μ_4	
Init 1	$\begin{bmatrix}-3.79\\-4.24\end{bmatrix}$	$\begin{bmatrix}-2.24\\4.13\end{bmatrix}$	$\begin{bmatrix}3.80\\5.03\end{bmatrix}$	$\begin{bmatrix}3.36\\-2.71\end{bmatrix}$	1108.037
Init 2	$\begin{bmatrix}3.80\\5.10\end{bmatrix}$	$\begin{bmatrix}-3.82\\-4.27\end{bmatrix}$	$\begin{bmatrix}3.33\\-2.64\end{bmatrix}$	$\begin{bmatrix}-2.24\\4.13\end{bmatrix}$	1109.42
Init 3	$\begin{bmatrix}3.79\\5.00\end{bmatrix}$	$\begin{bmatrix}-2.16\\4.11\end{bmatrix}$	$\begin{bmatrix}3.60\\-2.89\end{bmatrix}$	$\begin{bmatrix}-3.64\\-4.05\end{bmatrix}$	1102.55
Init 4	$\begin{bmatrix}3.36\\-2.66\end{bmatrix}$	$\begin{bmatrix}-3.78\\-4.22\end{bmatrix}$	$\begin{bmatrix}-2.24\\4.16\end{bmatrix}$	$\begin{bmatrix}3.80\\5.10\end{bmatrix}$	1108.46
Init 5	$\begin{bmatrix}3.57\\-2.88\end{bmatrix}$	$\begin{bmatrix}-2.14\\3.97\end{bmatrix}$	$\begin{bmatrix}-3.72\\-4.18\end{bmatrix}$	$\begin{bmatrix}3.79\\5.0\end{bmatrix}$	1103.92

Table 1: K-Means centers and distortion for different random initializations

(b) Consider a Gaussian mixture model in which the covariance matrices are pro- portional 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 certain percentage, e.g., 90%, of the mass of the Gaussian distribution). Estimate and represent (e.g. with different colors or different symbols) the latent variables for all data points (with the parameters learned by EM).

In this section we will consider Gaussian Mixture Models with K Gaussians with covariance matrices with the form

$$\Sigma_j = \sigma_j^2 I_d \quad \forall j \in \{1, .., K\}$$

Using the notations that we have introduced in class, we will calculate MLE from the general log-likelihood for Gaussian Mixture Models. Indeed, we need to compute this MLE to compute the M-Step of EM algorithm, we have that :

$$\begin{aligned} L(\theta) &= \sum_{i=1}^{n}\sum_{j=1}^{k} \tau_i^j \log(\pi_{j,t}) + \sum_{i=1}^{n}\sum_{j=1}^{k} \tau_i^j [\log(\frac{1}{(2\pi)^{k/2}} + \log(\frac{1}{det(\Sigma_{j,t})^{1/2}} - \frac{1}{2}(x_i - \mu_{j,t})^T \Sigma_{j,t}^{-1}(x_i - \mu_{j,t})] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{k} \tau_i^j \log(\pi_{j,t}) + \sum_{i=1}^{n}\sum_{j=1}^{k} \tau_i^j [\log(\frac{1}{(2\pi)^{k/2}} + \log(\frac{1}{\sigma_{j,t}^2}) - \frac{1}{2\sigma_{j,t}^2}(x_i^1 - \mu_{j,t}^1)^2 - \frac{1}{2\sigma_{j,t}^2}(x_i^2 - \mu_{j,t}^2)^2] \end{aligned}$$

Where we use x_j^i to designate the ith component of x_j. As in the general case, the sum is separated into two terms independent along the variables and the expression of π_t is exactly the same :

$$\boxed{\pi_{j,t+1} = \frac{1}{n}\sum_i^n \tau_i^j}$$

The covariances doesn't appear in μ_t calculations and we have

$$\boxed{\mu_{j,t+1} = \frac{\sum_i^n i\tau_i^j}{\sum_i^n \tau_i^j}}$$

Finally using the same trick as the one used to compute the MLE of the variance in the univariate gaussian (we derivate over $\frac{1}{\sigma^2}$). We have that

$$\boxed{\sigma_{j,t+1}^2 = \frac{1}{2\sum_i^n \tau_i^j}\sum_i^n \tau_i^j((x_i^2 - \mu_{j,t}^2)^2 + (x_i^1 - \mu_{j,t}^1)^2)}$$

With our training dataset, initializing with K-means, waiting until the convergence of the log-likelihood (69 iterations, convergence threshold at 10^{-5}) we estimate those parameters :

$$\pi_1 = 0.20 \quad \pi_2 = 0.17 \quad \pi_3 = 0.27 \quad \pi_4 = 0.37$$

$$\mu_1 = \begin{bmatrix} 3.82 \\ -3.72 \end{bmatrix} \quad \mu_2 = \begin{bmatrix} -2.61 \\ 4.25 \end{bmatrix} \quad \mu_3 = \begin{bmatrix} -3.66 \\ -4.08 \end{bmatrix} \quad \mu_4 = \begin{bmatrix} 2.61 \\ 3.70 \end{bmatrix}$$

$$\sigma_1^2 = 1.39 \quad \sigma_2^2 = 2.00 \quad \sigma_3^2 = 4.36 \quad \sigma_4^2 = 7.16$$

$$L(\theta) = -2639.56$$

You can find below the graphical representation of our algorithm (one color for each clusters, a circle represents the ellipse that contains 90% of the Gaussian mass distribution, latent variable for each data point is given by its representation with the color of the cluster in which it belongs).

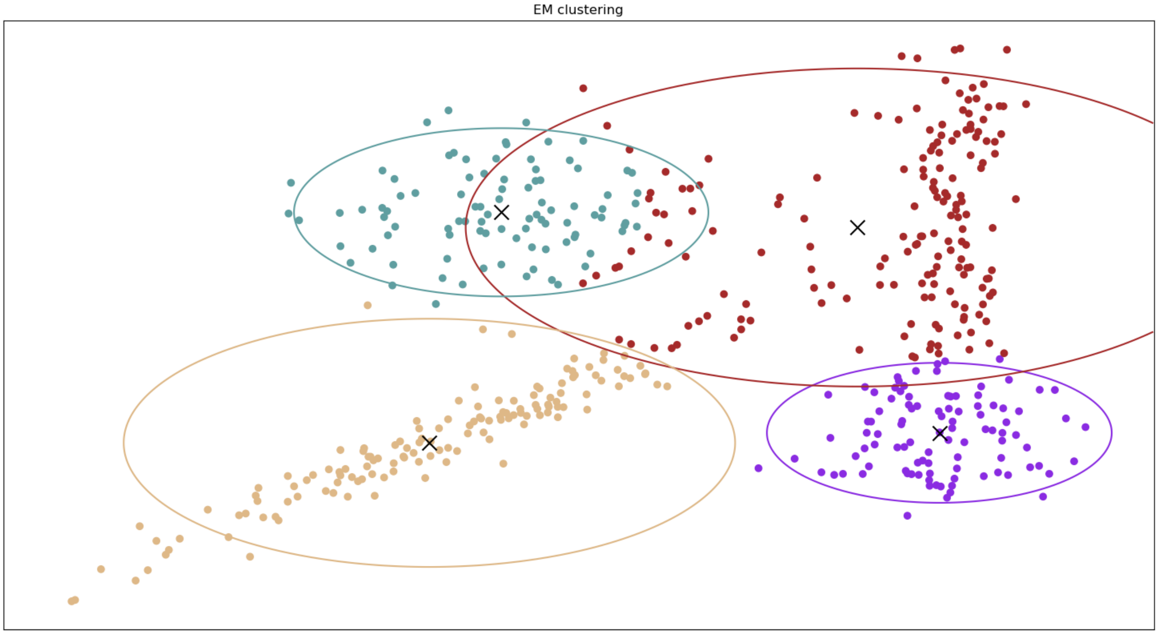

Figure 3: Isotropic Gaussian Mixture Output (training set)

(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 latent variables for all data points (with the parameters learned by EM).

We compute the EM algorithm for a Gaussian mixture with general covariance matrices using the methods that we have seen in class. We get this time :

$$\pi_1 = 0.26 \quad \pi_2 = 0.31 \quad \pi_3 = 0.18 \quad \pi_4 = 0.25$$

$$\mu_1 = \begin{bmatrix} 3.98 \\ 3.78 \end{bmatrix} \quad \mu_2 = \begin{bmatrix} -3.06 \\ -3.53 \end{bmatrix} \quad \mu_3 = \begin{bmatrix} 3.80 \\ -3.80 \end{bmatrix} \quad \mu_4 = \begin{bmatrix} -2.03 \\ 4.17 \end{bmatrix}$$

$$\Sigma_1 = \begin{bmatrix} 0.21 & 0.29 \\ 0.29 & 12.22 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 6.24 & 6.05 \\ 6.05 & 6.18 \end{bmatrix} \quad \Sigma_3 = \begin{bmatrix} 0.92 & 0.06 \\ 0.06 & 1.87 \end{bmatrix} \quad \Sigma_4 = \begin{bmatrix} 2.90 & 0.21 \\ 0.21 & 2.76 \end{bmatrix}$$

$$L(\theta) = -2327.7156969115204$$

We get the previous outcome with 39 iterations with a convergence threshold at 10^{-5}.

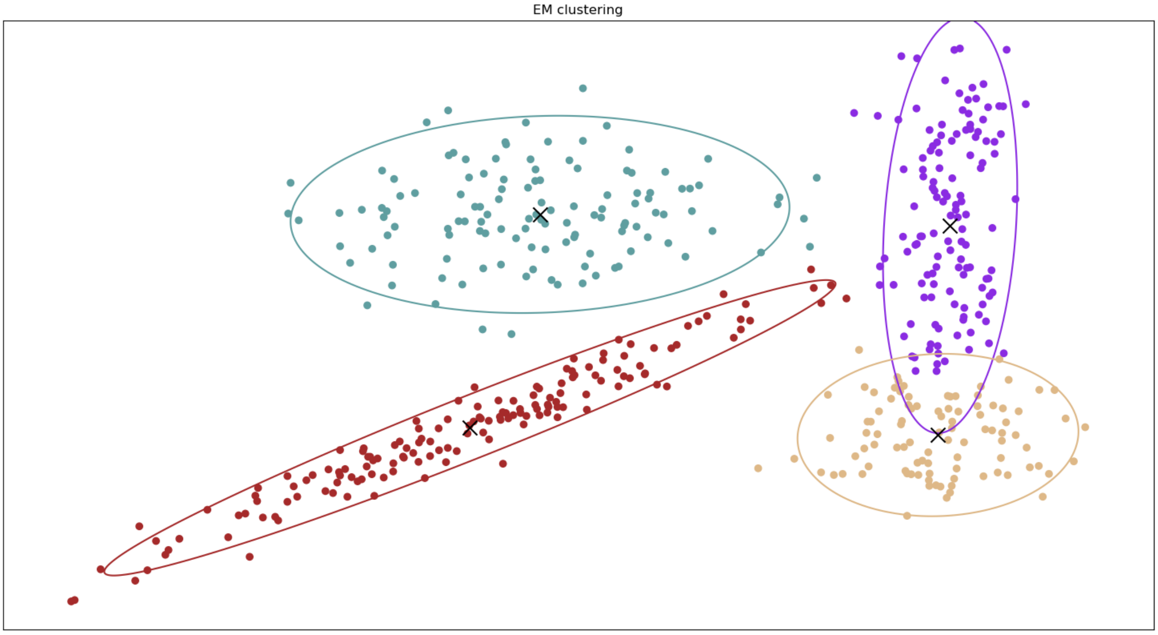

Figure 4: Gaussian Mixture Output (training set)

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

$$L_{test}^{isoGauss} = -2614.6018544036833$$
$$L_{test}^{Gauss} = -2408.985315802382$$

The Gaussian Mixture Model really suits well the data. You can see below the graphical outcome when we train the data with the train set and test the data with the test set.

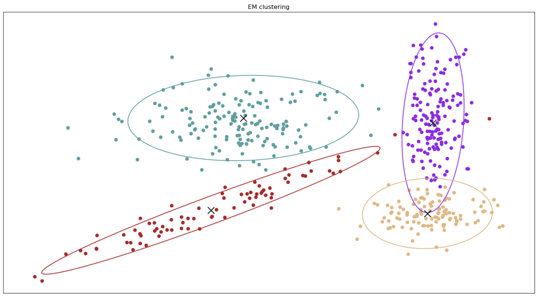

Figure 5: GM Output (test set)

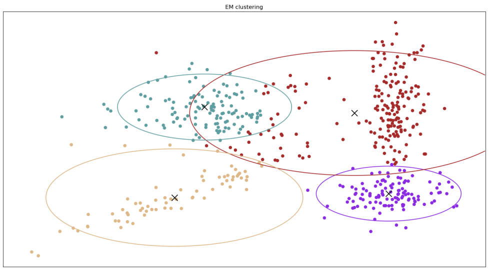

Figure 6: Iso GM Output (test set)