IFT 6269, Homework 3

Adel Nabli, ID: 20121744

28 Octobre 2018

1 DGM

Given the directed graphical model G, we can write that $\forall p \in \mathcal{L}(G),\ \forall\, x_V,\, p(x_V) = p(x)p(y)p(z|x,y)p(t|z)$. It is **not** true that $\forall p \in \mathcal{L}(G),\ X \perp\!\!\!\perp Y|T$. Indeed, the path from X to Y is **not** blocked by T: we have a *V structure* in Z but T is a descendant of Z.

We can confirm that by calculation:

- First, given the graph's "*explaining away*" structure, we have $X \perp\!\!\!\perp Y$.
- On the one hand, $p(x,y|t) = \dfrac{p(t|x,y)p(x,y)}{p(t)} = p(t|x,y)\dfrac{p(x)p(y)}{p(t)}$
- On the other hand, $p(x|t)p(y|t) = \dfrac{p(x,t)}{p(t)}\dfrac{p(y,t)}{p(t)} = \dfrac{p(t|y)p(y)p(t|x)p(x)}{p(t)p(t)} = \dfrac{p(t|y)p(t|x)}{p(t)}\dfrac{p(x)p(y)}{p(t)}$.
- But $p(t|x,y) \neq \dfrac{p(t|y)p(t|x)}{p(t)}$, thus $p(x,y|t) \neq p(x|t)p(y|t)$.

2 d-separation in DGM

To assess all the statements, we use the *Bayes ball algorithm.* We "throw" a ball from the first vertex and see if it can reach the second one.

a) False (the ball pass through A)

b) True (the ball is blocked in A, J and H)

c) False (the ball can bounce back on J and reach B)

d) True (the ball is blocked in A, D and H)

e) False (the ball can reach G through D)

f) False (the ball can reach G through D)

g) True (the ball is blocked in B, D and H)

h) False (the ball can bounce back on H and reach G)

i) True (the ball is blocked in B, D and E)

j) False (the ball can bounce back on J)

3 Positive interactions in-V-structure

We have X, Y, Z three binary random variables with a joint distribution parametrized according to the graph $X \to Z \leftarrow Y$. Thus, $\forall\, x_V \in \{0,1\}^3,\ p(x_V) = p(x)p(y)p(z|x,y)$.
We also want to have $\sum_x p(x) = \sum_y p(y) = \sum_{x_V} p(x_V) = 1$.

a) To make the task of finding examples simpler, we created examples that respect $P(X=0) = P(Y=0) = P(Z=0) = \frac{1}{2}$, making the marginal $p(z|x,y)$ proportional to the joint $p(x,y,z)$ (indeed, $\forall\, x,y \in \{0,1\}^2,\ p(x) = p(y) = \frac{1}{2}$ thus, $\forall\, z,\ p(z|x,y) = 4p(x,y,z)$).

That way, we have:

- $a = P(X=1) = \frac{1}{2}$
- $b = P(X=1|Z=1) = \frac{P(X=1,Z=1)}{P(Z=1)} = 2\sum_{y\in\{0,1\}} P(X=1,Y=y,Z=1)$
- $c = P(X=1|Z=1,Y=1) = \frac{P(X=1,Y=1,Z=1)}{\sum_{x\in\{0,1\}} P(X=x,Y=1,Z=1)}$

Using, those formulas, we now can check whether those three examples respect the conditions or not:

1 We want $c < a$:

$P(Z=1\|\bullet\bullet)$	$X=1$	$X=0$
$Y=1$	0.2	0.5
$Y=0$	0.5	0.8

$P(Z=0\|\bullet\bullet)$	$X=1$	$X=0$
$Y=1$	0.8	0.5
$Y=0$	0.5	0.2

Thus, $c = \frac{0.2}{0.2+0.5} \simeq 0.2857 < a = 0.5$.

2 We want $a < c < b$:

$P(Z=1\|\bullet\bullet)$	$X=1$	$X=0$
$Y=1$	0.7	0.6
$Y=0$	0.4	0.3

$P(Z=0\|\bullet\bullet)$	$X=1$	$X=0$
$Y=1$	0.3	0.4
$Y=0$	0.6	0.7

Thus, $a = 0.5 < c = \frac{0.7}{0.7+0.6} \simeq 0.538 < b = 2\frac{0.7+0.4}{4} = 0.55$.

3 We want $b < a < c$:

$P(Z=1\|\bullet\bullet)$	$X=1$	$X=0$
$Y=1$	0.6	0.4
$Y=0$	0.3	0.7

$P(Z=0\|\bullet\bullet)$	$X=1$	$X=0$
$Y=1$	0.4	0.6
$Y=0$	0.7	0.3

$$\text{Thus, } b = 2\frac{0.3+0.6}{4} = 0.45 < a = 0.5 < c = \frac{0.6}{0.6+0.4} = 0.6.$$

b) 1 Here, $Y=1$ will generally cause $Z=0$, and $X=0$ pushes Z to value 1. Thus, knowing $Y=1$ **and** $Z=1$ will make the belief $X=1$ less probable than without this information (the belief $X=0$ is stronger with this information).

2 Here, $Y=1$ will generally cause $Z=1$, and $X=1$ pushes Z to value 1. Thus, knowing only $Z=1$ gives very good reasons to believe $X=1$, but knowing $Z=1$ **and** $Y=1$ makes this belief a bit weaker because $Y=1$ already causes $Z=1$.

3 Here, $X=1$ will generally cause $Z=0$, so if we know $Z=1$, it will make the belief $X=1$ less probable than without this information. But here, the event $Y=1$ taken alone affect as much the value of Z as the event $Y=0$, except that combined with the knowledge of $Z=1$, $Y=1$ skew the distribution in a way that $X=1$ is then the most probable option.

4 Flipping a covered edge in a DGM

We know that $\mathcal{L}(G)=\mathcal{L}(G') \Leftrightarrow \mathcal{L}(G') \subseteq \mathcal{L}(G)$ and $\mathcal{L}(G) \subseteq \mathcal{L}(G')$.

- Let's show that $\forall\, p \ \in \mathcal{L}(G),\ p \ \in \mathcal{L}(G')$:

$$p \in \mathcal{L}(G) \Leftrightarrow \forall\, x_V,\ p(x_V) = \prod_{k\in V} p(x_k|x_{\pi_k}) = \Big[\prod_{k\in V\backslash\{i,j\}} p(x_k|x_{\pi_k})\Big] p(x_i|x_{\pi_i})p(x_j|x_{\pi_j})$$

But, as we know that (i,j) is a *covered edge*, we have $\pi_j = \{i\}\cup\pi_i$, and then:

$$\begin{aligned} p(x_i|x_{\pi_i})p(x_j|x_{\pi_j}) = p(x_i|x_{\pi_i})p(x_j|x_{\pi_i},x_i) &= \frac{\cancel{p(x_i,x_{\pi_i})}}{p(x_{\pi_i})}\frac{p(x_j,x_i,x_{\pi_i})}{\cancel{p(x_i,x_{\pi_i})}} \\ &= \frac{p(x_i|x_j,x_{\pi_i})p(x_j,x_{\pi_i})}{p(x_{\pi_i})} \\ &= p(x_i|\underbrace{x_j,x_{\pi_i}}_{x_{\pi_j}})p(x_j|x_{\pi_i}) \end{aligned}$$

Thus, $p \ \in \mathcal{L}(G')$ and we showed $\mathcal{L}(G) \subseteq \mathcal{L}(G')$.

- For the other way around, we can show the exact same way that $\mathcal{L}(G') \subseteq \mathcal{L}(G)$ and then we have proven $\mathcal{L}(G)=\mathcal{L}(G')$.

5 Equivalence of directed tree DGM with undirected tree UGM

- Let's E_U be the set of edges in the undirected tree and E_D the set of edges in the directed one. We can write $E_U = \{\{i,j\}\in V/i - j\}$ and $E_D = \{(i,j)\in V/i\to j\}$.
- It's clear that there is a bijection between E_U and E_D ($\{i,j\}=\{j,i\}$ in E_U).
- But we know that in a directed tree, each vertex has at most one parent: $\forall\, i \ \in V,\ |\pi_i| \le 1$.
- Thus, in the undirected tree G', the set of cliques $\mathcal{C}$ equals to E_U, which means that there is a bijection between $\mathcal{C}$ and E_D: the cliques are exactly the sets of 2 vertices that are in the couples of E_D.

- Thus, $\forall\ p\ \in \mathcal{L}(G)$, we can build a potential ψ such that $\forall (i,j)\ \in\ E_D,\ p(x_j|x_i) = \psi_c(x_i|x_j)$, we can define $Z = 1$ and have:

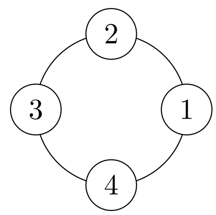
$$\forall\ p\ \in\ \mathcal{L}(G),\ p(x) = \prod_{(i,j)\in E_D} p(x_j|x_i) = \frac{1}{Z} \prod_{c\,=\{i,j\}\in\mathcal{C}} \psi_c(x_i, x_j)$$

  This means that $\mathcal{L}(G) \subseteq \mathcal{L}(G')$

- For the other way, we can obtain the conditional probabilities by taking the potentials and renormalizing them properly, which leads to $\mathcal{L}(G') \subseteq \mathcal{L}(G)$.

6 Hammersley-Clifford Counter example

Let G be the undirected graph:



Suppose that $p \in \mathcal{L}(G)$, i.e $\exists\ \psi$ s.t $\forall x,\ p(x) = \frac{1}{Z} \prod_{c\in\mathcal{C}} \psi_c(x_c)$ with $Z = \sum_x \prod_{c\in\mathcal{C}} \psi_c(x_c)$.

- Here, we have $\mathcal{C} = E = \{\{1,2\},\{2,3\},\{3,4\},\{4,1\}\}$, thus $\forall x,\ p(x) = \frac{1}{Z} \prod_{\{i,j\}\in E} \psi_{\{i,j\}}(x_i, x_j)$ with $Z = \sum_x \prod_{\{i,j\}\in E} \psi_{\{i,j\}}(x_i, x_j)$

- Let's call A the set of 8 elements of $\{0,1\}^4$ on which p is non null. We have that $\forall\ a \in A,\ p(a) = \frac{1}{8}$ and $\forall\ b \in \{0,1\}^4 \backslash A,\ p(b) = 0$.

- Z being a constant, it means that $\lambda : x \mapsto \prod_{\{i,j\}\in E} \psi_{\{i,j\}}(x_i, x_j)$ is **constant non null** on A and is null otherwise.

- Thus, because $(1,1,0,1)$ is **not** in A, we have:

$$\lambda(1,1,0,1) = \psi_{12}(1,1)\psi_{23}(1,0)\psi_{34}(0,1)\psi_{41}(1,1) = 0$$

- But, we know that $\lambda(1,1,1,1) \neq 0$, which leads to $\psi_{12}(1,1) \neq 0$ and $\psi_{41}(1,1) \neq 0$.
  We also have $\lambda(1,1,0,0) \neq 0 \Rightarrow \psi_{23}(1,0) \neq 0$ and $\lambda(0,0,1,1) \neq 0 \Rightarrow \psi_{34}(0,1) \neq 0$.
  CONTRADICTION

Thus, we showed that $p \notin \mathcal{L}(G)$.

7 Bizarre conditional independence properties

We have a random vector (X, Y, Z) with a finite sample space such that $X \perp\!\!\!\perp Y|Z$ and $X \perp\!\!\!\perp Y$.

a) We suppose that Z is a binary variable. Let's define the constant $p = p(Z = 0) = 1 - p(Z = 1)$.

- On the one hand, because $X \perp\!\!\!\perp Y$, we have:

$$\begin{aligned} p(x,y) = p(x)p(y) &= \Big[p(x, Z = 0) + p(x, Z = 1)\Big]\Big[p(y, Z = 0) + p(y, Z = 1)\Big] \\ &= \underbrace{\Big[p(x|Z = 0)p + p(x|Z = 1)(1 - p)\Big]}_{\alpha}\underbrace{\Big[p(y|Z = 0)p + p(y|Z = 1)(1 - p)\Big]}_{\beta} \end{aligned}$$

- On the other hand, because $X \perp\!\!\!\perp Y|Z$, we have:

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

By writing $p(x,y) = p(x,y)$, we derive the following equations:

$$\begin{aligned} &(E) \quad \alpha p(y|Z = 0) + \beta p(x|Z = 1) - p(x|Z = 1)p(y|Z = 0) - \alpha\beta = 0 \\ &(E) \Leftrightarrow \Big(\alpha - p(x|Z = 0)\Big)\Big(p(y|Z = 0) - \beta\Big) = 0 \\ &(E) \Leftrightarrow (1 - p)^2\Big(p(x|Z = 1) - p(x|Z = 0)\Big)\Big(p(y|Z = 0) - p(y|Z = 1)\Big) = 0 \end{aligned}$$

Thus, we have three cases to consider:

$$\begin{cases} (1): \quad (1 - p) = 0 \\ (2): \quad p(x|Z = 1) - p(x|Z = 0) = 0 \\ (3): \quad p(y|Z = 0) - p(y|Z = 1) = 0 \end{cases}$$

(1) If $P(Z = 0) = 1$, the random variable Z is **constant almost surely** and then the statement "$X \perp\!\!\!\perp Z$ or $Y \perp\!\!\!\perp Z$" is True.

(2) We have:

$$p(x|Z = 1) = p(x|Z = 0) \Rightarrow p(x|Z = 1)p(Z = 0)p(Z = 1) = p(x|Z = 0)p(Z = 0)p(Z = 1)$$

But we know that $p(x, Z = 0) = p(x, Z = 0 \cup Z = 1) - p(x, Z = 1)$, thus:

$$\begin{aligned} p(x|Z = 1) = p(x|Z = 0) &\Rightarrow p(x, Z = 1)p(Z = 0) = \Big[p(x) - p(x, Z = 1)\Big]p(Z = 1) \\ &\Rightarrow p(x, Z = 1)\Big[\underbrace{p(Z = 0) + p(Z = 1)}_{=1}\Big] = p(x)p(Z = 1) \end{aligned}$$

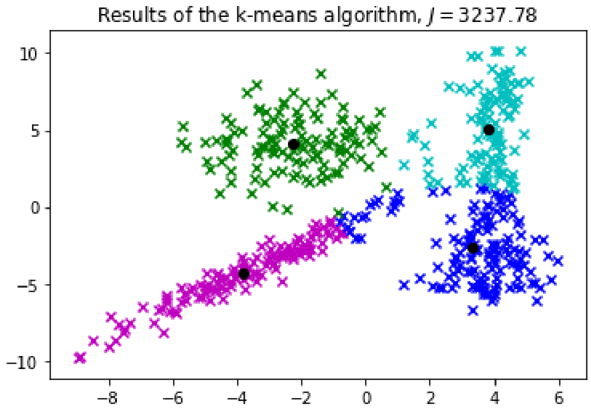
Using $p(x, Z = 1) = p(x) - p(x, Z = 0)$, we also show that $p(x, Z = 0) = p(x)p(Z = 0)$. Thus $(2) \Rightarrow X \perp\!\!\!\perp Z$.

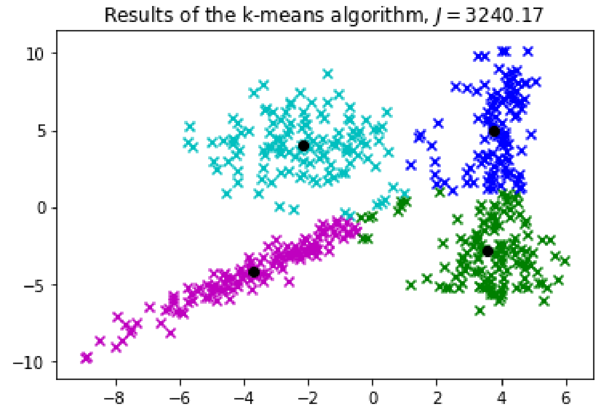
(3) We use the same method to show that $(3) \Rightarrow Y \perp\!\!\!\perp Z$

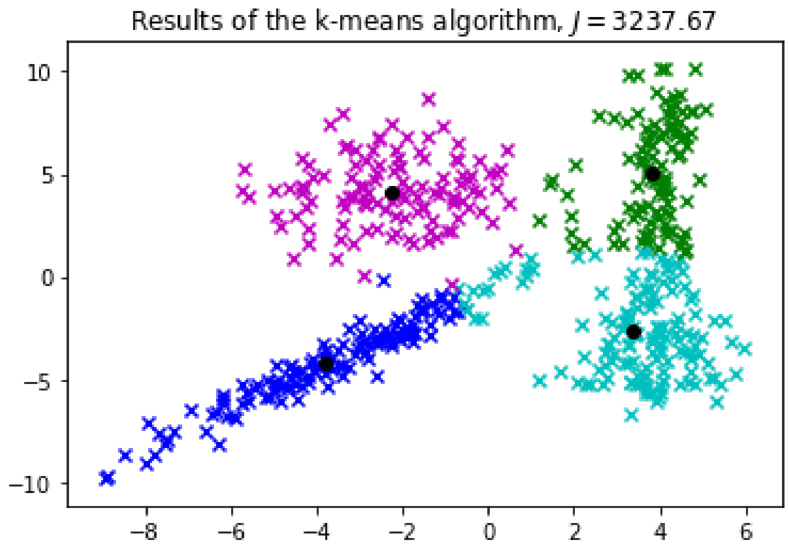
Finally, we showed that $X \perp\!\!\!\perp Y|Z$ and $X \perp\!\!\!\perp Y \Rightarrow (X \perp\!\!\!\perp Z$ or $Y \perp\!\!\!\perp Z)$

8 Implementation: EM and Gaussian mixtures

a) With three different random initialization, that's the results we get:







b) Let z be the hidden variables and x be the observed data. We make the assumption that the $x_i \in \mathbb{R}^d$, $i \in \{1, ..., N\}$ are i.i.d. We suppose that the $z_i \sim \mathcal{M}(\pi_1, ...\pi_K)$ and $(x_i \mid z_i = k) \sim \mathcal{N}(\mu_k, \Sigma_k)$. We define $\theta = (\pi, \mu, \Sigma)$. We are considering the case $\Sigma_k = \sigma_k^2 \mathbb{I}_d$.
We want to find $\underset{\theta}{\operatorname{argmax}}\, \mathbb{E}_{Z|X}(l_{c,t})$ where $l_{c,t}$ is the complete log-likelihood.
Let's call $\mathbb{E}_{Z|X}(l_{c,t}) = f(\theta^t)$ and $p_{\theta^t}(z_i = k \mid x_i) = \tau_i^k$. We have:

$$f(\theta^t) = \sum_{i=1}^{N}\sum_{k=1}^{K} \tau_i^k \log(\pi_k^t) + \sum_{i=1}^{N}\sum_{k=1}^{K} \tau_i^k \Big[- \log(2\pi)\frac{d}{2} - d\log(\sigma_k) - \frac{1}{2\sigma_k^2}(x_i - \mu_k^t)^T(x_i - \mu_k^t)\Big]$$

Thus:

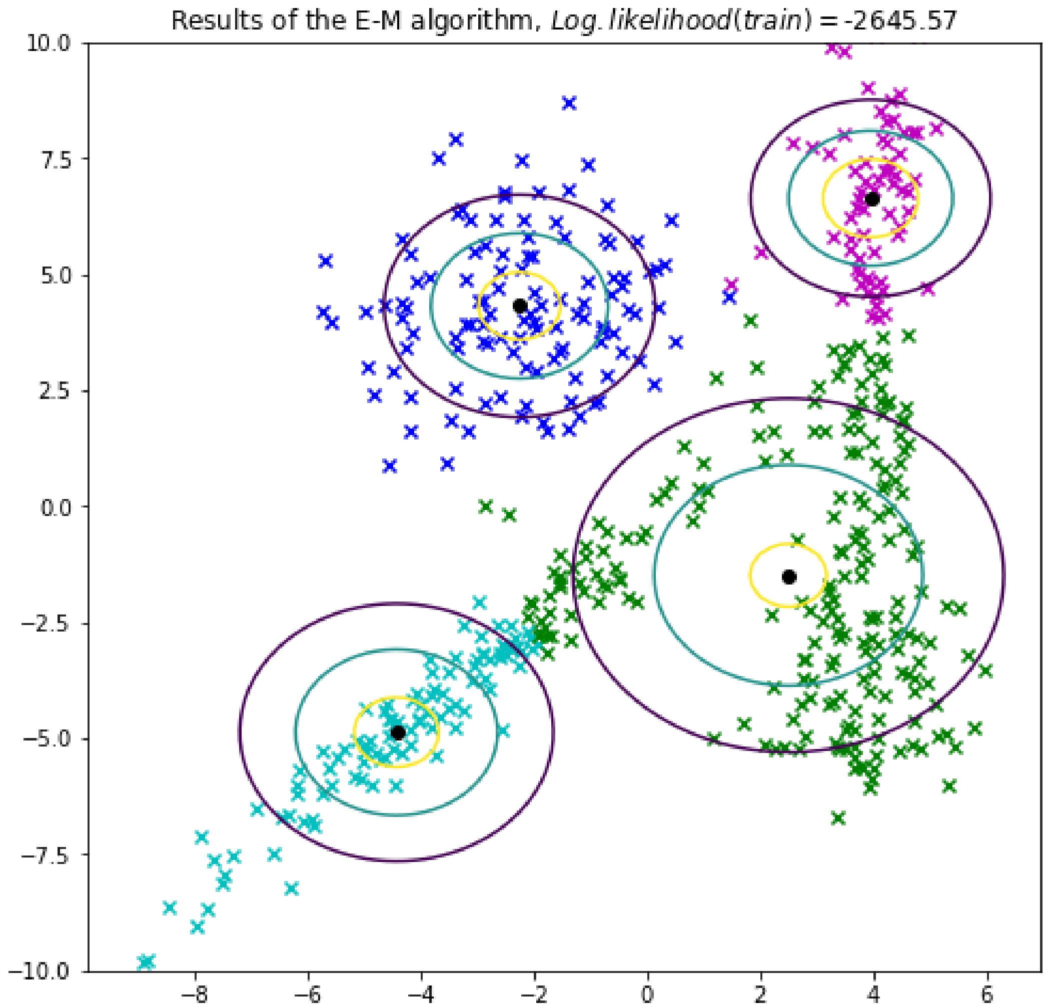
$$\forall k \in [\![1, K]\!], \frac{\partial f}{\partial \sigma_k} = \sum_{i=1}^{N} \tau_i^k \Big[- \frac{d}{\sigma_k} + \frac{(x_i - \mu_k^t)^T(x_i - \mu_k^t)}{\sigma_k^3}\Big]$$

$$\frac{\partial f}{\partial \sigma_k} = 0 \Rightarrow \frac{1}{\sigma_k^2}\sum_{i=1}^{N} \tau_i^k (x_i - \mu_k^t)^T(x_i - \mu_k^t) = d\sum_{i=1}^{N}\tau_i^k$$

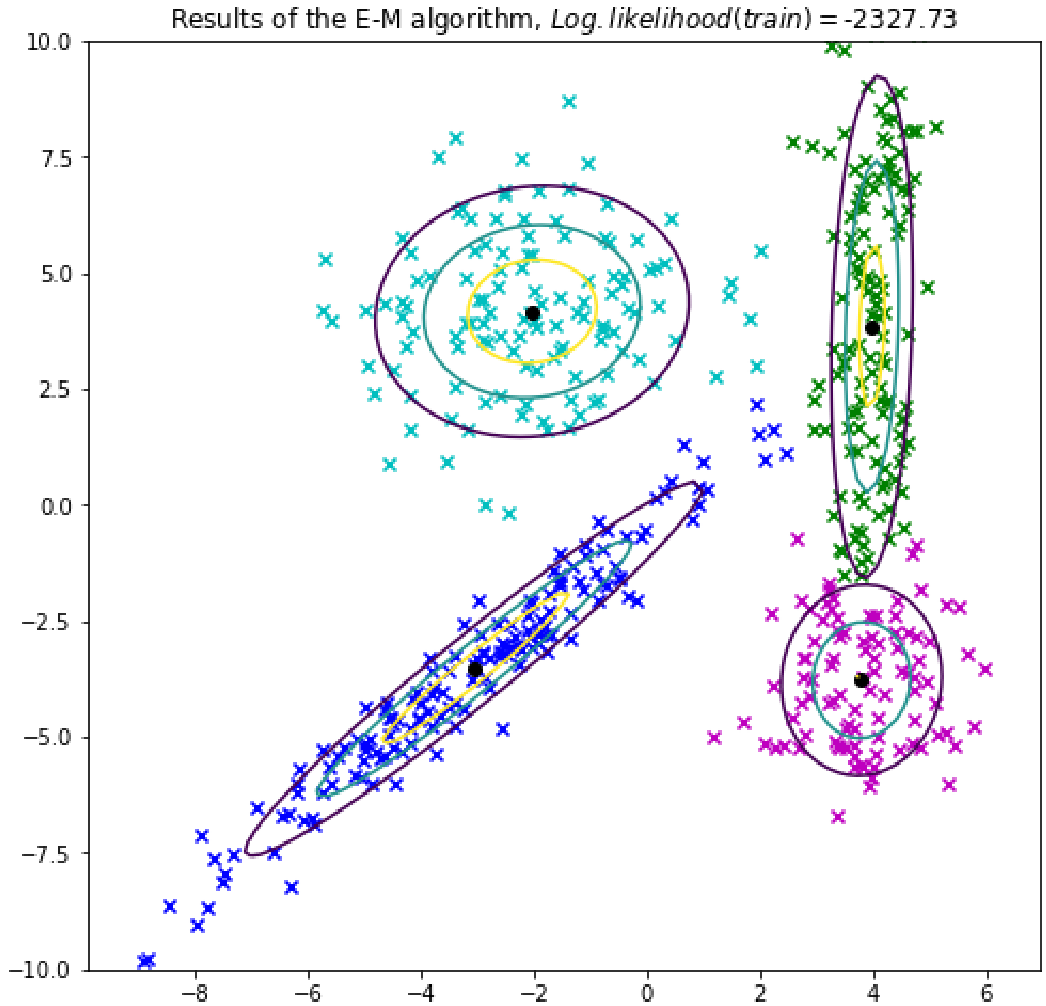
$$\Rightarrow \sigma_k^2 = \frac{\sum_{i=1}^{N} \tau_i^k (x_i - \mu_k^t)^T(x_i - \mu_k^t)}{d\sum_{i=1}^{N}\tau_i^k}$$

Which is a maximum. The other parameters not being influenced by the "change" $\Sigma_k = \sigma_k^2 \mathbb{I}_d$, we can directly use the general formulas given in the class.

After implementation (*cf Notebook*), this is what we get for the isotropic gaussians:



c) For the general case, we get the following results:



d) The formula for the *log-likelihood* is given by:

$$\log p_\theta(x) = \sum_{i=1}^{N} \log \sum_{z_i} p_\theta(x_i, z_i) = \sum_{i=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k)$$

This is what we get on the *training* and *test* sets after normalization:

Models / Sets	**Train set**	**Test set**
isotropic gaussian	-5.29	-5.384
general gaussian	-4.654	-4.818

Normalized log-likelihood on both the training and test sets for the models tested

As we could have predicted, the "general" model works best on both the training and test sets (*log-likelihood* higher). Indeed, from the disposition of the data, it seems that the "true" underlying model is a mixture of gaussians which are not necessarily isotropic. Both models seem to generalize well and don't seem to overfit: even if there performance are better on the training data, the *log-likelihood* on the test data isn't much lower (for example, we have better results on the test data of the general mixture than on the trainig set with the isotropic model).