Master MVA
Probabilistic Graphical Models

Homework 2: Classification and Graph Theory

Victor Busa
victor.busa@ens-paris-saclay.fr

November 10, 2017

1 Conditional independence and factorizations

1. We want to prove that: $X \perp\!\!\!\perp Y \mid Z$ iff $p(x|y,z) = p(x,z)$ for all pairs (y,z) such that $p(y,z) > 0$, using only:

$$\begin{aligned}
&\text{a)} \quad p(x,y|z) = p(x|z)p(y|z) \\
&\text{b)} \quad p(x,y) = p(x|y)p(y) \\
&\text{c)} \quad \sum_x p(x|y) = 1
\end{aligned}$$

Proof.

⇒ Suppose $p(x,y|z) = p(x|z)p(y|z)$, let's show that $p(x|y,z) = p(x|z)$ for (y,z) s.t $p(y,z) \overset{(\Delta)}{>} 0$:

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

⇐ Suppose $p(x|y,z) \overset{(E)}{=} p(x|z)$ and $p(y,z) \overset{(\Delta)}{>} 0$ and let's show that $p(x,y|z) = p(x|z)p(y|z)$:

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

Note: We can divide by $p(z)$, because, by (Δ):

$$p(y,z) > 0 \Rightarrow p(z) \overset{(c)}{=} p(z) \sum_y p(y|z) \overset{(b)}{=} \sum_y p(y,z) > 0$$

□

2. Let $p \in \mathcal{L}(G)$, then we can write:

$$p(t,z,x,y) = p(t|z)p(z|x,y)p(x)p(y)$$

moreover, we have $X \not\perp\!\!\!\perp Y \mid T$ because we have the d-separation property. Indeed, T is observed and is a child of a V-structure separating X from Y.

3.

(a) We want to prove that, assuming Z is a **binary variable**:

$$\left.\begin{array}{l}(1)\ p(x,y|z) = p(x|z)\\(2)\ p(x,y) = p(x)p(y)\end{array}\right\} \Rightarrow \Big(p(x,z) = p(x)p(z)\Big) \vee \Big(p(y,z) = p(y)p(z)\Big)$$

We notice that we have either $p(x,z) = p(x)p(z)$ or $p(y,z) = p(y)p(z)$, hence we might think, that, at the end of our demonstration we would have something like:

$$\big[p(x,z) - p(x)p(z)\big]\,\big[p(y,z) - p(y)p(z)\big] = 0$$

In order to have something of this form, one good idea is to use 2 different expressions of a certain probability. Moreover, we will need to use that Z is a **binary variable**, that is to say: $p(z=1) = 1 - p(z=0)$.

Notations:

$$p(z=0) \triangleq p$$
$$p(z=0|X) \triangleq p_0(X)$$

Using our idea, we can write $p(x,y)$ in 2 different ways:

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

$$p(x,y) \overset{(2)}{=} p(x)p(y) = \sum_z p(x,z)p(z) \sum_{z'} p(y|z')p(z') \tag{E'}$$

Equating both relations and developing, we have:

$$
\begin{aligned}
& (E) = (E') \\
\Leftrightarrow \quad & p_0(x)p_0(y)p + p_1(x)p_1(y)(1-p) = \Big(p_0(x)p + p_1(x)(1-p)\Big)\Big(p_0(y)p + p_1(y)(1-p)\Big) \\
\Leftrightarrow \quad & p_0(x)p_0(y)p + \cancel{p_1(x)p_1(y)} - \cancel{p_1(x)p_1(y)p} \\
& = \quad p_0(x)p_0(y)p^2 + p_0(x)p_1(y)p - p_0(x)p_1(y)p^2 + p_1(x)p_0(y)p \\
& - p_1(x)p_0(y)p^2 + \cancel{p_1(x)p_1(y)} - \cancel{2}p_1(x)p_1(y)p + p_1(x)p_1(y)p^2 \\
\Leftrightarrow \quad & p_0(x)p_0(y)p(p-1) - p_0(x)p_1(y)p(p-1) - p_1(x)p_0(y)p(p-1) + p_1(x)p_1(y)p(p-1) = 0 \\
\Leftrightarrow \quad & p(p-1)\Big(p_0(x)p_0(y) - p_0(x)p_1(y) - p_1(x)p_0(y) + p_1(x)p_1(y)\Big) = 0 \\
\Leftrightarrow \quad & p(p-1)\Big(p_1(y) - p_0(y)\Big)\Big(p_0(x) - p_1(x)\Big) = 0
\end{aligned}
$$

Will will consider 3 cases: $p = 0$, $p = 1$ and $0 < p < 1$:

$$
\begin{array}{rl}
p = 0\text{:} & p(z=0) = 0 \quad \text{and} \quad p(z=1) = 1 \quad \text{and} \quad p(z=1) = 1 \Rightarrow p(z=1|x) = 1 = p(z=1) \\
p = 1\text{:} & p(z=0) = 1 \quad \text{and} \quad p(z=1) = 0 \quad \text{and} \quad p(z=0) = 1 \Rightarrow p(z=0|x) = 1 = p(z=0) \\
\text{if } 0 < p < 1\text{:} & \Big(p_1(y) - p_0(y)\Big)\Big(p_0(x) - p_1(x)\Big) = 0 \quad \text{i.e} \quad p_1(y) = p_0(y) \quad \text{or} \quad p_0(x) = p_1(x)
\end{array}
$$

Hence, if Z is a **binary variable**, the proposition
"If $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp Y$ then ($X \perp\!\!\!\perp Y$ or $Y \perp\!\!\!\perp Z$)" is **true**

(b) I guess this statement is true in general.

2 Distributions factorizing in a graph

1. Let's prove that if $G = (V, E)$ is a DAG then we have $\mathcal{L}(G) = \mathcal{L}(G')$ where, if $i \to j$ is a covered edge, then we define $E' = (E \backslash \{i \to j\}) \bigcup \{i \to j\}$. To prove this assertion, it suffices to prove that (Figure 2.3), if $i \to j$ is a covered edge, then:

$$
p(x_i|x_{\pi_i})p(x_j|x_{\pi_j}) = p(x_i|x_{\pi_i})p(x_j|x_{\pi_i}, x_i) = p(x_j|x_{\pi_i})p(x_i|x_{\pi_i}, x_j) \tag{P}
$$

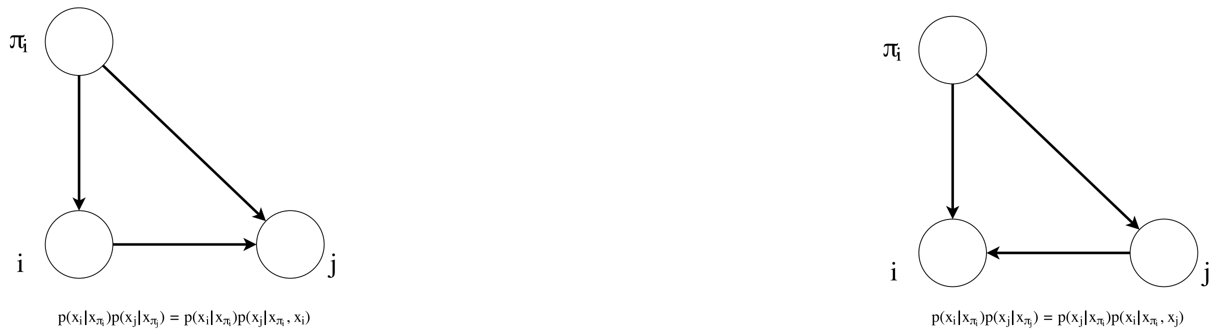
π_i
i
j
$p(x_i|x_{\pi_i})p(x_j|x_{\pi_j}) = p(x_i|x_{\pi_i})p(x_j|x_{\pi_i}, x_i)$
π_i
i
j
$p(x_i|x_{\pi_i})p(x_j|x_{\pi_j}) = p(x_j|x_{\pi_i})p(x_i|x_{\pi_i}, x_j)$

Figure 2.1: Edge (i, j) is covered

Figure 2.2: Edge (i, j) is reversed

Figure 2.3: Covered edge and its reverse

because, in this case $p(x)$ can be written either as (1) or (2).

$$\begin{aligned}
p(x) &= \prod_k p(x_k|x_{\pi_k}) \\
&= p(x_i|x_{\pi_i})p(x_j|x_{\pi_j}) \prod_{\substack{k\neq i \\ k\neq j}} p(x_k|x_{\pi_k}) \\
&= p(x_i|x_{\pi_i})p(x_j|x_{\pi_i}, x_i) \prod_{\substack{k\neq i \\ k\neq j}} p(x_k|x_{\pi_k}) \qquad (1) \\
&= p(x_i|x_j, x_{\pi_i})p(x_j|x_{\pi_i}) \prod_{\substack{k\neq i \\ k\neq j}} p(x_k|x_{\pi_k}) \qquad (2)
\end{aligned}$$

(1) corresponds to $p \in \mathcal{L}(G)$ and (2) corresponds to $p \in \mathcal{L}(G')$. Which proves that $\mathcal{L}(G) = \mathcal{L}(G')$. So let's prove (P):

Proof.

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

□

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

Proof.
To prove that $\mathcal{L}(G) = \mathcal{L}(G')$, we have to express the factorization of p over $\mathcal{L}(G)$ as a factorization of p over $\mathcal{L}(G')$. A directed tree is a DAG in which all the nodes have exactly **1 parent** beside the root node that has **no parent**. If we let r be the root node, n be the number of nodes in the tree and $E = \{(i, j), i \to j\}$ be the set of all directed edges, we have:

$$p(x) = \prod_{i=1}^{n} p(x_i|x_{\pi_i}) = p(x_r) \prod_{(i,j)\in E} p(x_j|x_i)$$

Indeed, (i, j) being a directed edge, and G being a directed tree, $\{i\} = \pi_j$ is the unique parent of j. If, now, we define:

$$
\begin{aligned}
\psi(x_r) &= p(x_r) \\
\psi(x_i) &= 1 \quad \forall i \neq r \\
\psi(x_i, x_j) &= p(x_j|x_i) \\
Z &= 1
\end{aligned}
$$

then we notice that:

$$
\begin{aligned}
p(x) &= p(x_r) \prod_{(i,j)\in E} p(x_j|x_i) \\
&= \frac{1}{Z}\left(\prod_{i\in V}\psi(x_i) \prod_{(i,j)\in E} \psi(x_i, x_j)\right)
\end{aligned}
$$

which is the factorization of p over G' the undirected tree construct from G, so we have proved that $\mathcal{L}(G) \subset \mathcal{L}(G')$

To prove that $\mathcal{L}(G') \subset \mathcal{L}(G)$, we can notice that in an undirected tree there is a unique path between every pair of nodes. If we select an arbitrary node to be the root node we can always direct all the edges in the graph to point from the root to the leaf nodes. Moreover every edge in a tree correspond to a two-node potential function that we can replace without loss of generality by a conditional probability distribution provided a certain normalization constant. This conditional probability will only depend on the unique parent of each node as the constructed tree is an directed tree. If the undirected tree doesn't factorize over the directed tree provided our arbitrary choice of the root node, we can, without loss of generality pick another node as the root node. Hence we can replace any undirected tree by N different directed tree where N is the number of nodes of the tree. □

3 Entropy and Mutual Information

1.

(a) $H(X) \geq 0$ because $\forall x \in \mathcal{X}, \quad 0 \leq p(x) \leq 1$, and hence, $log\, p_X(x) \leq 0$, and so:

$$
H(X) = -\sum_{x\in\mathcal{X}} \underbrace{p_X(x) log p_X(x)}_{\leq 0} \geq 0
$$

Furthermore, $H(X) = 0 \Rightarrow \forall x \in \mathcal{X}$, $p_X(x) log\, p_X(x) = 0$, i.e $\forall x \in \mathcal{X}$, $p_X(x) = 0$ or $p_X(x) = 1$. Yet, we also have that, $\sum\limits_{x\in\mathcal{X}} p_X(x) = 1$. Hence, $\exists\, x$ s.t $p_X(x) = 1$. Which proves that X takes one value (hence is constant) with probability 1.

(b) We can compute $D(p \parallel q)$ as follow:

$$D(p \parallel q) = \sum_{x \in \mathcal{X}} p(x) log \frac{p(x)}{q(x)} = -H(X) - \sum_{x \in \mathcal{X}} p(x)\, log q(x)$$

If we let q be the uniform distribution on $\mathcal{X}$, we have: $\sum\limits_{x \in \mathcal{X}} p(x) log\, q(x) = \sum\limits_{x \in \mathcal{X}} p(x) log \frac{1}{|\mathcal{X}|} = -log|\mathcal{X}|$, because $\sum\limits_{x \in \mathcal{X}} p_X(x) = 1$, so finally we have:

$$D(p \parallel q) = -H(X) + log|\mathcal{X}|$$

(c) Hence, using question (b) and the fact that $D(p||q) \geq 0$ (seen in class) we have :

$$H(X) \leq log|\mathcal{X}| \triangleq log(k)$$

Where we used the notation: $k \triangleq |\mathcal{X}|$

2.

(a) Let's define $I(X_1, X_2) = \sum\limits_{(x_1,x_2) \in \mathcal{X}_1 \times \mathcal{X}_2} p_{1,2}(x_1, x_2) log \frac{p_{1,2}(x_1,x_2)}{p_1(x_1) p_2(x_2)}$. $I(X_1, X_2) \geq 0$ because we can write: $I(X_1, X_2) = D(p_{1,2}(X_1, X_2) \parallel p_1(X_1) p_2(X_2)) \geq 0$ (seen in class).

(b) Let's prove that $I(X_1, X_2)$ can be expressed as a function of $H(X_1)$, $H(X_2)$ and $H(X_1, X_2)$, where $H(X_1, X_2)$ is the entropy of $X = (X_1, X_2)$

Proof.

$$\begin{aligned}
\sum_{(x_1,x_2) \in \mathcal{X}_1 \times \mathcal{X}_2} p_{1,2}(x_1, x_2) log \frac{p_{1,2}(x_1, x_2)}{p_1(x_1) p_2(x_2)} &= \sum_{(x_1,x_2) \in \mathcal{X}_1 \times \mathcal{X}_2} p_{1,2}(x_1, x_2) log\, p_{1,2}(x_1, x_2) - \\
&\quad \sum_{(x_1,x_2) \in \mathcal{X}_1 \times \mathcal{X}_2} p_{1,2}(x_1, x_2) log\, p_1(x_1) p_2(x_2) \\
&= \sum_{(x_1,x_2) \in \mathcal{X}_1 \times \mathcal{X}_2} p_{1,2}(x_1, x_2) log\, p_{1,2}(x_1, x_2) - \\
&\quad \sum_{(x_1,x_2) \in \mathcal{X}_1 \times \mathcal{X}_2} p_{1,2}(x_1, x_2) log\, p_1(x_1) - \\
&\quad \sum_{(x_1,x_2) \in \mathcal{X}_1 \times \mathcal{X}_2} p_{1,2}(x_1, x_2) log\, p_2(x_2)
\end{aligned}$$

furthermore, we have:

$$\sum_{(x_1,x_2)\in\mathcal{X}_1\times\mathcal{X}_2} p_{1,2}(x_1,x_2) log\, p_2(x_2) = \sum_{x_2\in\mathcal{X}_2} log\, p_2(x_2) \sum_{x_1\in\mathcal{X}_1} p_{1,2}(x_1,x_2)$$

$$= \sum_{x_2\in\mathcal{X}_2} log\, p_2(x_2) p_2(x_2) = -H(X_2)$$

And in the same way,

$$\sum_{(x_1,x_2)\in\mathcal{X}_1\times\mathcal{X}_2} p_{1,2}(x_1,x_2) log\, p_1(x_1) = -H(X_1)$$

so, finally, we can rewrite $I(X_1, X_2)$ as:

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

□

(c) Using the previous questions we have that: $I(X_1, X_2) = -H(X_1, X_2) + H(X_1) + H(X_2) \geq 0$, i.e

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

And we have equality when $X_1 \perp\!\!\!\perp X_2$, i.e the maximal entropy is such that: $p_{1,2}(x_1, x_2) = p_1(x_1)p_2(x_2)$

4 Implementation - Gaussian mixtures

(a) See code for the implementation of K-means. The Figure 4.1 displays the histograms of distortion for 500 random initializations. The Figure 4.2 displays the clusters and distortion for a certain random initialization. The Figure 5.1 displays two more clusters and distortion curve for 2 more random initialization.

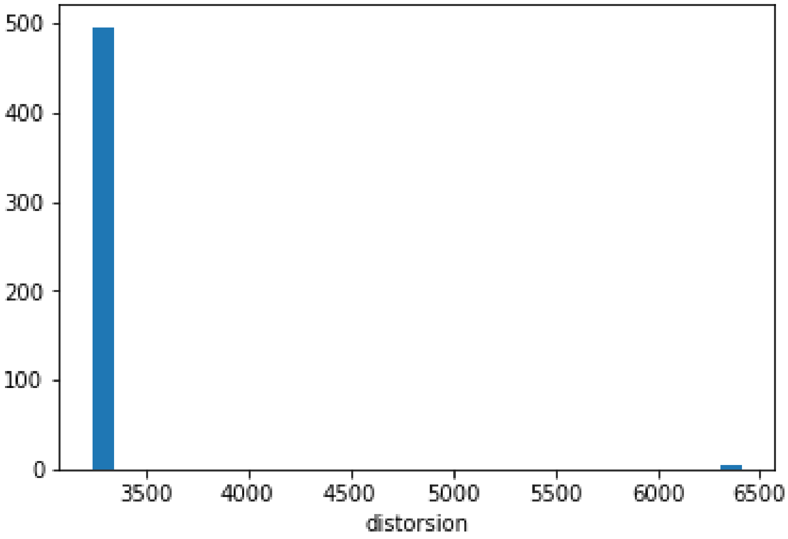

Figure 4.1: histograms of the distortion for 500 random initialization

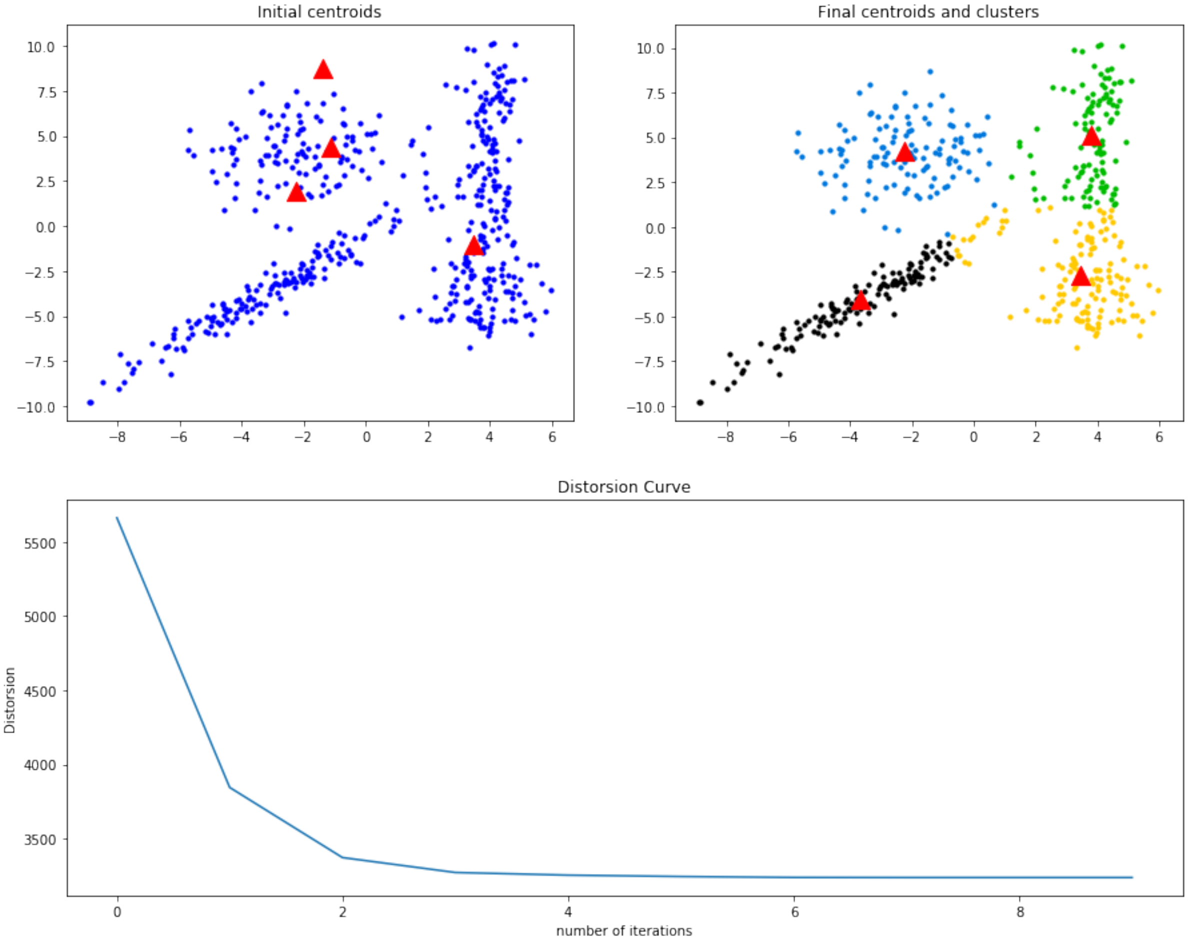

Figure 4.2: clusters and distortion for one random initializations

We can see that, in our experiment, K-means converges to the same near optimal centroids **in**

most of the case $^{4.1}$ $^{5.2}$. Yet we've also noticed that for some bad random initialization K-means doesn't converge to a good nearly optimal solution. Hence we must carefully check that K-means has converged to a nearly optimal solution.

(b) Let (x_i, z_i) be a couple, for $i \in \{1, \dots, n\}$ with $x_i \in \mathbb{R}^2$, $z_i \sim \mathcal{M}(1, \pi_1, ...\pi_K)$ and $(x_i|z_i = k) \sim \mathcal{N}(\mu_k, \Sigma_k)$. Here, the parameters are: $\theta = (\pi, \mu, \Sigma)$. To derive the **E**xpectation Step and the **M**aximization Step, we will write the complete log-likelihood at the time step **t**, and we note $\ell_{c,t}$:

$$\begin{aligned}\ell_{c,t} = log\ p(x, z; \theta_t) &= \sum_{i=1}^{n} log\ p(x_i, z_i; \theta_t) \\ &= \sum_{i=1}^{n} log\ p(z_i; \theta_t) p(x_i|z_i; \theta_t) \\ &= \sum_{i=1}^{n} log\ p(z_i; \theta_t) + \sum_{i=1}^{n} log\ p(x_i|z_i; \theta_t) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} \delta_{i,k} log\ p(\pi_{k,t}) + \sum_{i=1}^{n}\sum_{k=1}^{K} \delta_{i,k} log\ \mathcal{N}(x_i, \mu_{k,t}, \Sigma_{k,t})\end{aligned}$$

E-Step: We need to write the Expectation of $\ell_{c,t}$ w.r.t $(Z|X)$. i.e:

$$\begin{aligned}\mathbb{E}_{(Z|X)}[\ell_{c,t}] &= \mathbb{E}\left[\sum_{i=1}^{n}\sum_{k=1}^{K} \delta_{i,k} log\ p(\pi_{k,t}) + \sum_{i=1}^{n}\sum_{k=1}^{K} \delta_{i,k} log\ \mathcal{N}(x_i, \mu_{k,t}, \Sigma_{k,t})\right] \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{E}_{(Z|X)}[\delta_{i,k}] log\ p(\pi_{k,t}) + \sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{E}_{(Z|X)}[\delta_{i,k}] log\ \mathcal{N}(x_i, \mu_{k,t}, \Sigma_{k,t})\end{aligned}$$

So, we only need to compute $\mathbb{E}_{(Z|X)}[\delta_{i,k}]$. But $\mathbb{E}_{(Z|X)}[\delta_{i,k}] = p(z_i = k|x_i; \theta_t)$. So we need to compute $p(z_i = k|x_i; \theta_t)$. As seen in class, we have the relation:

$$\begin{aligned}p(z_i = k|x_i; \theta_t) &= \frac{\pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k)}{\sum_{k'} \pi_{k'} \mathcal{N}(x_i|\mu_{k'}, \Sigma_{k'})} \\ &\triangleq \tau_i^k\end{aligned}$$

M-Step: We need to maximize the following quantity with respect to $\theta_t = (\pi_t, \mu_t, \Sigma_t)$:

$$\sum_{i=1}^{n}\sum_{k=1}^{K} \tau_i^k log\ p(\pi_{k,t}) + \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_i^k log\ \mathcal{N}(x_i, \mu_{k,t}, \Sigma_{k,t}) \tag{1}$$

The calculation of the maximization of (1) w.r.t $\theta_t = (\pi_t, \mu_t, \Sigma_t)$ has been seen in class. The only difference in the isotropic case is the calculation of $\Sigma_{k,t}$ (Σ w.r.t the k^{th} cluster at time step t). In the Isotropic case we have $\Sigma_{k,t} = \sigma_{k,t} I$, such that:

$$\mathcal{N}(x_i|u_{k,t},\Sigma_{k,t}) = \frac{1}{(2\pi)^{d/2}\sigma_{k,t}^{d/2}} exp\left(\frac{-1}{2\sigma_{k,t}}(x_i-\mu_{k,t})^\intercal(x_i-\mu_{k,t})\right)$$

As we want to maximize σ_k at time step $t+1$, we need to compute the value of $\sigma_{k,t+1}$ for which $\nabla_{\sigma_k}\left(\sum\limits_{i=1}^{n}\sum\limits_{k=1}^{K}\tau_i^k log\,\mathcal{N}(x_i|u_{k,t},\Sigma_{k,t})\right)=0$:

$$\nabla_{\sigma_k}\left(\sum_{i=1}^{n}\sum_{k=1}^{K}\tau_i^k(C-log(\sigma_k^{d/2})-\frac{1}{2\sigma_k}(x_i-\mu_k)^\intercal(x_i-\mu_k))\right)=0$$

$$\Leftrightarrow \sum_{i=1}^{n}-\frac{d}{2\sigma_k}\tau_i^k+\frac{1}{2\sigma_k^2}\tau_i^k(x_i-\mu_k)^\intercal(x_i-\mu_k)=0$$

$$\Leftrightarrow \sigma_k=\frac{1}{d}\frac{\sum\limits_{i=1}^{n}\tau_i^k\|x_i-\mu_k\|_2^2}{\sum\limits_{i=1}^{n}\tau_i^k}$$

So finally, in the isotropic case, the update becomes:

$$\sigma_{k,t+1}=\frac{1}{d}\frac{\sum\limits_{i=1}^{n}\tau_{i,t}^k\left\|x_i-\mu_{k,t}\right\|_2^2}{\sum\limits_{i=1}^{n}\tau_{i,t}^k}$$

The Figure 4.3 displays the clusters, the centroids and the 4 gaussians centered on the centroids. We can see that this model doesn't fit well the data as the clusters are approximated by circles due to the isotropic covariance.

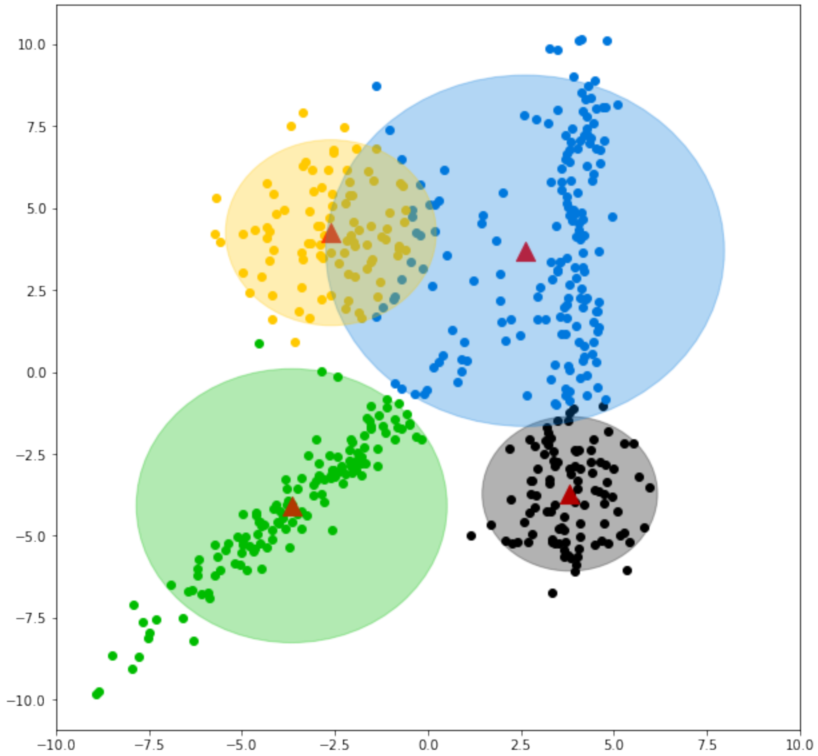

Figure 4.3: EM with isotropic covariance and $K = 4$

(c) The Figure 4.4 displays the clusters, the centroids and the 4 gaussians centered on the centroids. We can see that this model fits the data quite well. Indeed the data seems to have been generated by 4 Gaussians with different non isotropic covariance. Hence, as we removed the assumption that the covariance is isotropic, the clusters can be any kind of ellipses and that is why this model performs well.

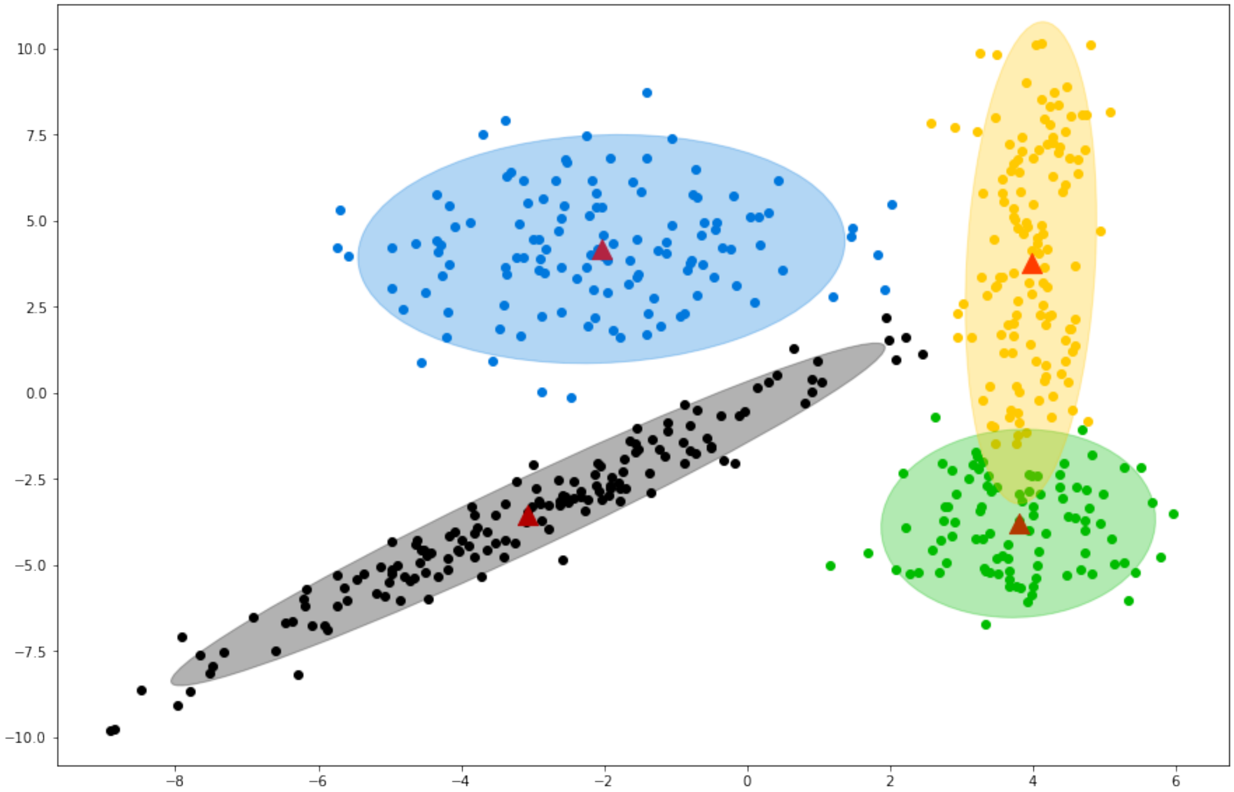

Figure 4.4: EM with isotropic covariance and $K = 4$

(d) The results are summarized in Table 1 below

	log-likelihood	
model	Training set	Testing set
Isotropic	-2639.56	-2614.60
General	-2327.72	-2408.97

Table 1: Log-likelihood of the two mixture models for the Training and Testing set

As expected, the log-likelihood of the general mixture model is smaller than the log-likelihood of the isotropic model. Hence, the general model outperforms the isotropic model. We can also notice that the log-likelihood on the Testing set is slightly lower than on the Training set for the isotropic case, which is quite unusual. This can be easily explains by the fact that the data have surely been randomly generated using a mixture of Gaussians

5 Annexe

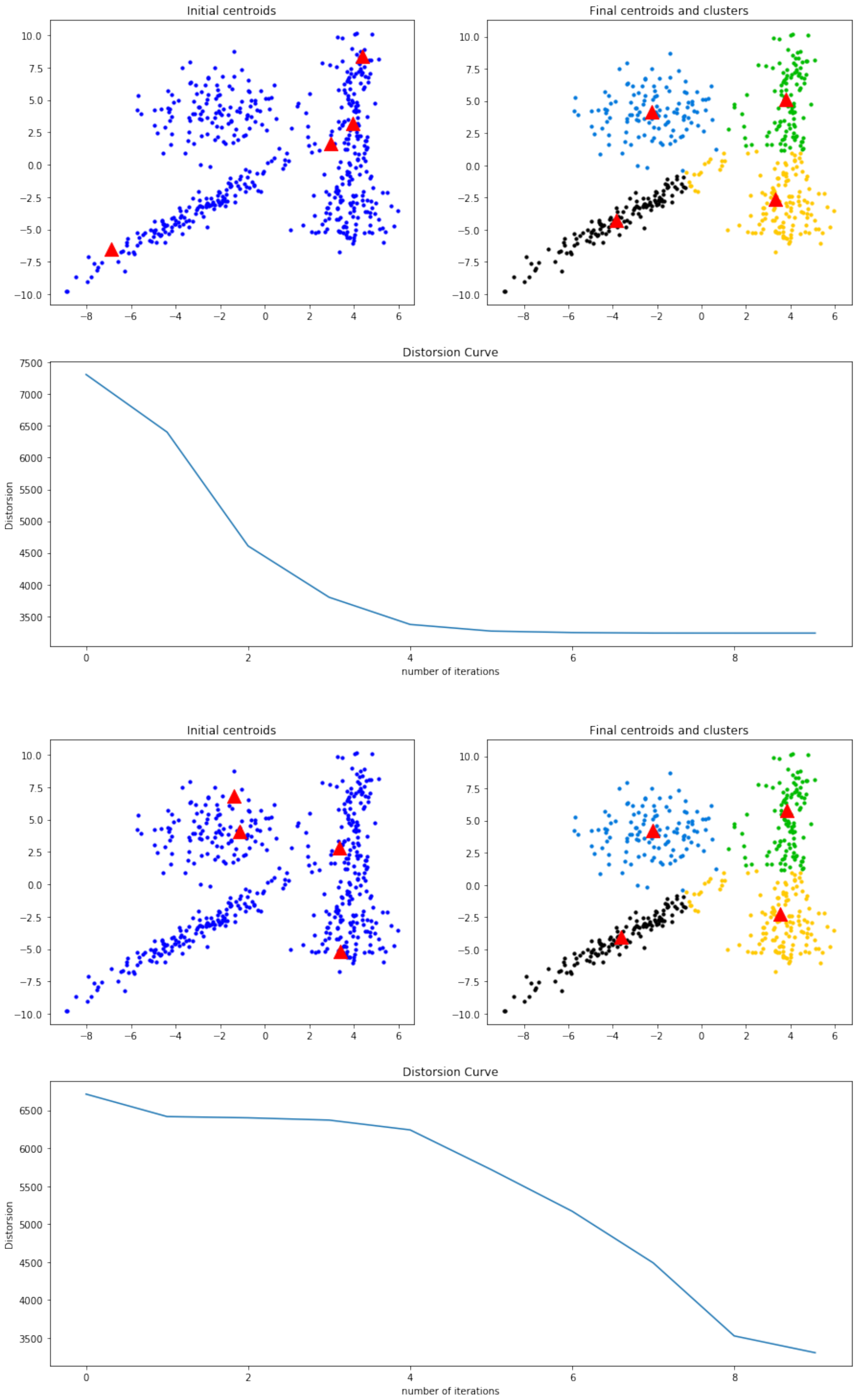

Figure 5.1: clusters and distortion for 2 more random initialization

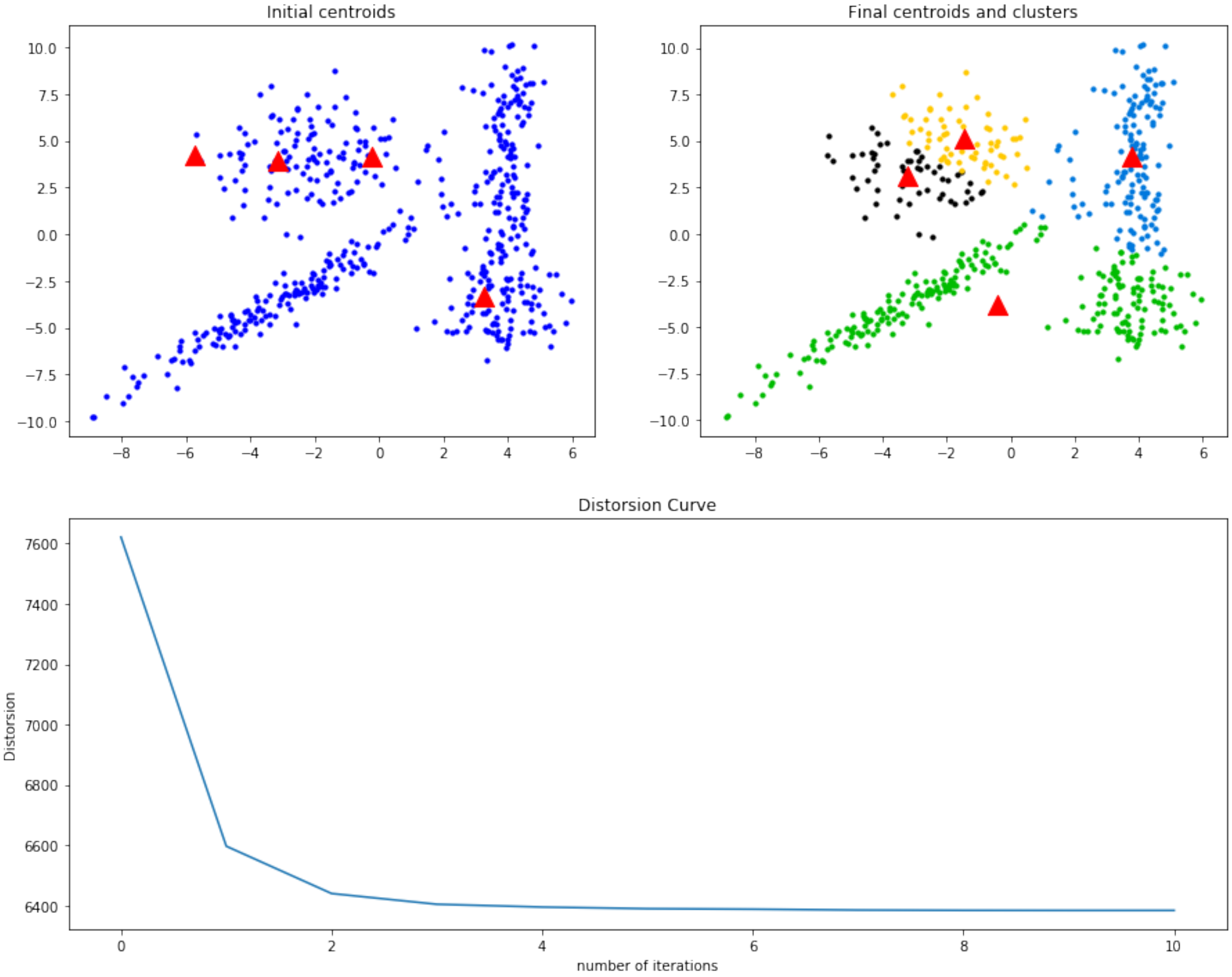

Figure 5.2: clusters and distortion for a bad random initialization