PGM - HW2 - Mehdi Boubnan & Amine Sadeq

Exercise 1: Conditional independence and factorizations

Question 1
The implied factorization for a joint distribution $p \in \mathcal{L}(\mathrm{G})$ is :

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

Let X and Y two iid random variables, $Z = X + Y$, $T = Z$ and p their joint distribution. It's clear that $p \in \mathcal{L}(\mathrm{G})$ (Indeed : $X \perp\!\!\!\perp Y$, $T \perp\!\!\!\perp X/Z$ and $T \perp\!\!\!\perp Y/Z$). Since T=Z=X+Y, $P_T = P_{X+Y} = P_X * P_Y$, which clearly shows that X and Y are not independent given T.

Question 2.a
Let (X, Y, Z) be three random variables on a finite space $\mathcal{E}$, with Z a binary random variable (z_1, z_2).
Let's assume that $X \perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp Y/Z$, therefore : $\begin{cases} p(x, y, z) = p(x, y/z)p(z) = p(x/z)p(y/z)p(z) \\ p(x, y) = p(x)p(y) \end{cases}$

$$p(x)p(y) = p(x, y) = \sum_z p(x, y, z) = \sum_z p(x/z)p(y/z)p(z)$$

$$p(y) = \sum_z p(y/z)\frac{p(x/z)p(z)}{p(x)} = \sum_z p(y/z)p(z/x) \quad \text{for } p(x) \neq 0$$

$$\text{we also have} \quad p(y) = \sum_z p(y/z)p(z)$$

We have then, for all x,y $\in \mathcal{E}$ such that $p(x) \neq 0$:

$$p(y/z_1)(p(z_1/x) - p(z_1)) = p(y/z_2)(p(z_2) - p(z_2/x)) = 0$$

If $p(z_1/x) - p(z_1) = 0$ then $p(z_2) - p(z_2/x) = 0$, so $p(z/x) = p(z)$ and $(p(z, x) = p(z/x)p(x) = p(z)p(x)$. Therefore $X \perp\!\!\!\perp Z$.

If $p(z_1/x_1) - p(z_1) \neq 0$ then **for all** $y \in \mathcal{E}$: $p(y/z_1) = p(y/z_2)\frac{p(z_2)-p(z_2/x)}{p(z_1/x)-p(z_1)}$ and by normalization $p(y/z_1) = p(y/z_2) = \lambda_y$, therefore $p(y) = \lambda_y p(z_1) + \lambda_y p(z_2) = \lambda_y = p(y/z)$ and finally $p(y, z) = p(y/z)p(z) = p(y)p(z)$. Therefore $Y \perp\!\!\!\perp Z$

Question 2.b

Let X,Y two independent random variables on a finite space $\mathcal{F}$.

We define Z a random variable on $\mathcal{F} \times \mathcal{F}$ such that Z=(X,Y).

$$\begin{aligned} P(X = x, Y = y/Z = (z_1, z_2)) &= \frac{P(X = x, Y = y, X = z_1, Y = z_2)}{P(X = z_1, Y = z_2)} = \frac{P(X = x, X = z_1)P(Y = y, y = z_2)}{P(X = z_1)P(Y = z_2)} \\ &= \frac{P(X = x/X = z_1)P(X = z_1)P(Y = y/Y = z_1)P(Y = z_1)}{P(X = z_1)P(Y = z_1)} \\ &= P(X = x/X = z_1)P(Y = y/Y = z_1) \end{aligned}$$

We also have :

$$P(X = x/Z = (z_1, z_2)) = P(X = x/X = z_1, Y = z_2) = P(X = x/X = z_1) \Rightarrow \text{X is dependent to Z}$$
$$P(Y = y/Z = (z_1, z_2)) = P(Y = y/X = z_1, Y = z_2) = P(Y = y/Y = z_2) \Rightarrow \text{Y is dependent to Z}$$

Therefore, even if $X \perp\!\!\!\perp Y/Z$ and $X \perp\!\!\!\perp Y$, X is dependent to Z and Y is dependent to Z.

We conclude that the statement is not true in general.

Exercise 2: Distribution factorizing in a graph

Question 1
Let's take G = (V,E) with $V = \{X_1, X_2, ..., X_n\}$ and $i \rightarrow j$ a covered edge.
Let G'=(V,E') with E' = $(E \backslash \{i \rightarrow j\}) \cup \{j \rightarrow i\}$
We can say that $\begin{cases} \forall k \in \{1,..,n\} \quad k \neq i \quad and \quad k \neq j \quad \pi_k^G = \pi_k^{G'} \\ \pi_i^{G'} = \pi_i^G \cup \{j\} \quad and \quad \pi_j^{G'} = \pi_i^G \end{cases}$
We now try to prove that $L(G) = L(G')$ by equivalences :

$$\begin{aligned} p \in L(G) &\iff p(x) = \prod_{k=1}^{n} p(x_k|\pi_k^G) \iff p(x) = p(x_i|\pi_i^G)p(x_j|\pi_i^G, x_i) \prod_{k=1,k\neq i,k\neq j}^{n} p(x_k|\pi_k^G) \\ &\iff p(x) = p(x_i|\pi_i^G)\frac{p(x_j, x_i|\pi_i^G)}{p(x_i|\pi_i^G)} \prod_{k=1,k\neq i,k\neq j}^{n} p(x_k|\pi_k^G) \iff p(x) = p(x_j, x_i|\pi_j^{G'}) \prod_{k=1,k\neq i,k\neq j}^{n} p(x_k|\pi_k^{G'}) \\ &\iff p(x) = p(x_i|\pi_j^{G'}, x_j)p(x_j|\pi_j^{G'}) \prod_{k=1,k\neq i,k\neq j}^{n} p(x_k|\pi_k^{G'}) \iff p(x) = p(x_i|\pi_i^{G'})p(x_j|\pi_j^{G'}) \prod_{k=1,k\neq i,k\neq j}^{n} p(x_k|\pi_k^{G'}) \\ &\iff p(x) = \prod_{k=1}^{n} p(x_k|\pi_k^{G'}) \iff p \in L(G') \end{aligned}$$

So we can conclude that L(G) = L(G')

Question 2
Let G be a directed tree and G' its corresponding undirected tree. Every vertex X_i in G has consequently only one parent π_i. The set of the maximum cliques in G' is therefore $C = \{c_i / c_i = (x_i, \pi_i)\}$.

$$p \in L(G) \iff p(x) = \prod_{i=1}^{n} f_i(x_i, \pi_i) \iff p(x) = \prod_{i=1}^{n} f_i(c_i) \iff p(x) = \prod_{c_i \in C} f_i(c_i)$$
$$with \quad fi \geq 0 \quad and \quad \forall i, \forall x_{\pi_i}, \quad \sum_{x_i} f_i(x_i, x_{\pi_i}) = 1$$

We know that to show $x \in L(G')$ we have to write $p(x) = \frac{1}{Z} \prod_{c_i \in C} \psi_i(c_i)$ with $Z = \sum_x \prod_{c_i \in C} \psi_i(c_i)$
Let's compute Z with $\psi_i = f_i$:

$$\begin{aligned} Z &= \sum_x \prod_{c_i \in C} f_i(c_i) = \sum_{x_1} \sum_{x_2} ... \sum_{x_n} \prod_{c_i \in C} f_i(c_i) = \sum_{x_1} \sum_{x_2} ... \sum_{x_n} \prod_{i=1}^{n} f_i(x_i, \pi_i) \\ &= \sum_{x_1} \sum_{x_2} ...[\sum_{x_n} f_n(x_n, \pi_n)] \prod_{i=1}^{n-1} f_i(x_i, \pi_i) \end{aligned}$$

For $p \in L(G)$ we have $\sum_{x_i} f_i(x_i, \pi_i) = 1$, so :

$$Z = \sum_{x_1} \sum_{x_2} ... \sum_{x_{n-1}} \prod_{i=1}^{n-1} f_i(x_i, \pi_i)$$

By iterating this computation, we get in the end $Z = 1$. We can then conclude :

$$p \in L(G) \iff p(x) = \frac{1}{Z} p(x) = \frac{1}{Z} \prod_{c_i \in C} f_i(c_i) \iff p \in L(G')$$

So we can conclude that L(G) = L(G')

Exercise 3: Implementation - Gaussian mixtures

Question 3.a : Kmeans

When starting with different random initialization, we notice that the KMeans algorithm converges to different distortion minimas (around 3240 and 6380). Indeed, the KMeans alogirthm is not a global optimization algorithm, and therefore converges to different local minimas depending on the initialization.

Question 3.b : M-step

The expectation of the complete loglikelihood is concave wrt π, μ and Σ^{-1}. We'll compute the derivatives :

$$(\pi, \mu, \Sigma^{-1}) = \theta = argmax_\theta \sum_{i=1}^{n}\sum_{j=1}^{k} \tau_i^j log(\pi_{i,t}) + \sum_{i=1}^{n}\sum_{j=1}^{k} \tau_i^j [log(\frac{1}{(2\pi)^{\frac{d}{2}}}) + \frac{1}{2}log(det(\Sigma_j^{-1})) - \frac{1}{2}(x_i - \mu_{j,t})^T \Sigma_j^{-1}(x_i - \mu_{j,t})]$$

`Estimate of` π_j:

Since the constraint $\sum_{j=1}^{k}\pi_j = 1$ and the E-step are $\mathcal{C}^1$ we'll use the lagrangian:

$$L(\pi, \lambda) = -\sum_{i=1}^{n}\sum_{j=1}^{k}\tau_i^j log(\pi_{i,t}) + \lambda(\sum_{j=1}^{k}\pi_{j,t} - 1)$$

$$\nabla_{\pi_{j,t}} L(\pi, \lambda) = -\sum_{i=1}^{n}\frac{\tau_i^j}{\hat{\pi}_{j,t+1}} + \lambda = 0 \iff \hat{\pi}_{j,t+1} = \frac{\sum_{i=1}^{n}\tau_i^j}{\lambda}$$

$$\sum_{j=1}^{k}\hat{\pi}_{j,t+1} = 1 \iff \sum_{j=1}^{k}\frac{\sum_{i=1}^{n}\tau_i^j}{\lambda} = 1 \iff \lambda = \sum_{i=1}^{n}\sum_{j=1}^{k}\tau_i^j = n$$

$$\boxed{\hat{\pi}_{j,t+1} = \frac{1}{n}\sum_{i=1}^{n}\tau_i^j}$$

`Estimate of` μ_j:

$$\nabla_{\mu_{j,t}}\mathcal{L} = 0 \iff \sum_{i=1}^{n}\tau_i^j\Sigma_j^{-1}(x_i - \hat{\mu}_{j,t+1}) = 0$$

$$\sum_{i=1}^{n}\tau_i^j(x_i - \hat{\mu}_{j,t+1}) = 0 \iff \boxed{\hat{\mu}_{j,t+1} = \frac{\sum_{i=1}^{n}\tau_i^j x_i}{\sum_{i=1}^{n}\tau_i^j}} \quad (\Sigma^{-1} \text{ is definite})$$

`Estimate of` Σ_j:

For the isotropic EM, covariance matrices are considered proportional to the identity. We have then : $\forall j$, $\Sigma_j = \alpha_j I$ with $\alpha_j > 0$, $\Sigma_j^{-1} = \frac{1}{\alpha_j}I$ and $det(\Sigma_j) = \alpha_j^d$. We'll then compute the estimates for each α_j, and therefore the derivative of the E-Step wrt to α_j :

$$\sum_{i=1}^{n}\tau_i^j(-\frac{d}{2\alpha_{j,t+1}} + \frac{1}{2\alpha_{j,t+1}^2}(x_i - \mu_{j,t+1})^T(x_i - \mu_{j,t+1})) = 0 \iff d\alpha_{j,t+1}\sum_{i=1}^{n}\tau_i^j = \sum_{i=1}^{n}\tau_i^j(x_i - \mu_{j,t+1})^T(x_i - \mu_{j,t+1})$$

$$\boxed{\alpha_{j,t+1} = \frac{\sum_{i=1}^{n}\tau_i^j(x_i - \mu_{j,t+1})^T(x_i - \mu_{j,t+1})}{d\sum_{i=1}^{n}\tau_i^j} = \frac{\sum_{i=1}^{n}\tau_i^j\left\|x_i - \mu_{j,t+1}\right\|^2}{d\sum_{i=1}^{n}\tau_i^j}}$$

Question 3.c : Covariance estimator

$$\boxed{\hat{\Sigma}_{j,t+1} = \frac{\sum_{i=1}^{n}\tau_i^j(x_i - \mu_{j,t+1})(x_i - \mu_{j,t+1})^T}{\sum_{i=1}^{n}\tau_i^j}}$$

Question 3.d : Comments

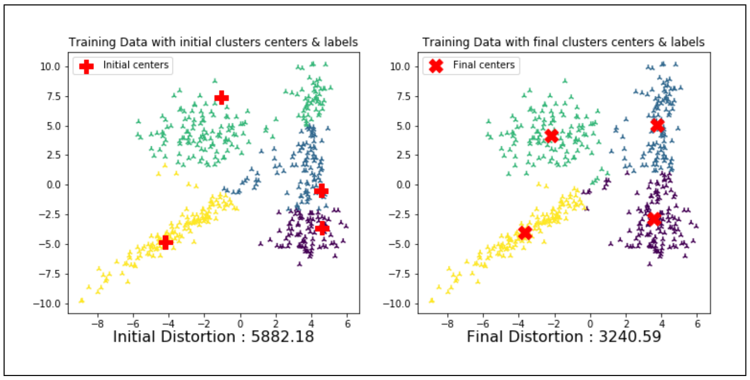
The train set and the test have the same number of observations (500), we can then compare the loglikelihoods for each model shown in the table on the right. We notice that the train set loglikelihoods are greater than the test set ones which is expected because the estimators were learned on the training data.

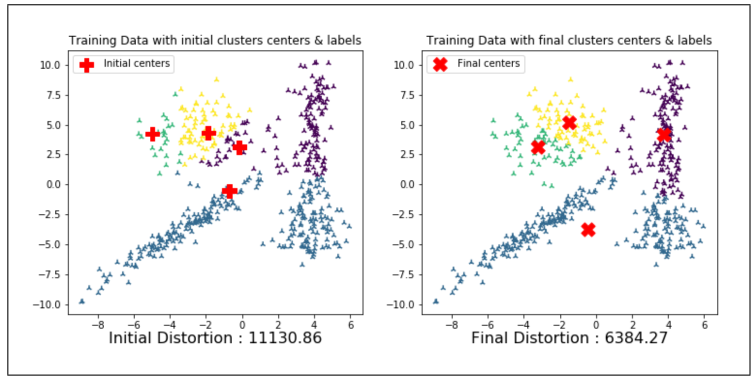
Loglikelihoods

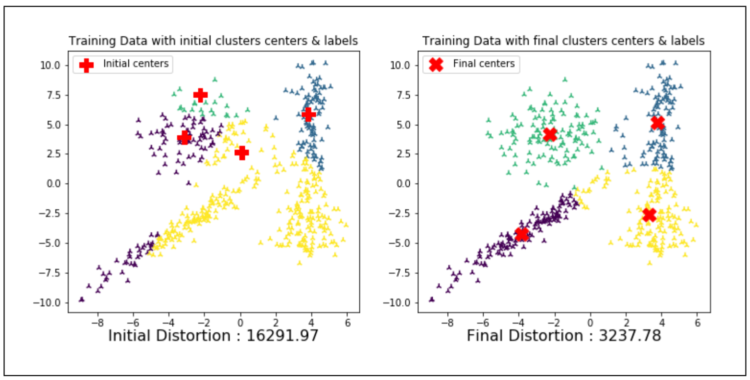
	Isotropic	General
Train set	-2681.9	-2392.72
Test set	-2724.75	-2462.69

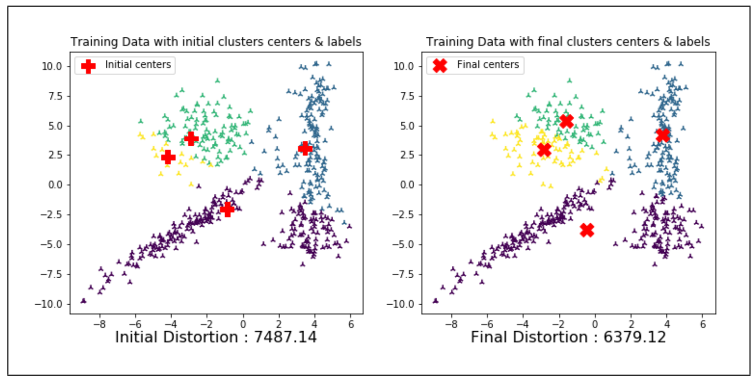
Figures

KMeans for different initializations









EM