Ecole Nationale Supérieur Paris-Saclay
Master Mathématiques Vision Apprentissage

Homework 2

DUCHEMIN Quentin & OREISTEIN Pierre

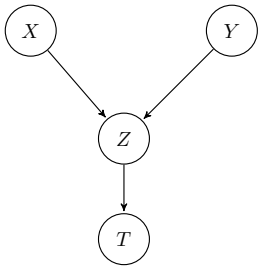
Cours: Probabilistic Graphical Models

October 2018

Indépendance conditionnelle et factorisation

Question 1

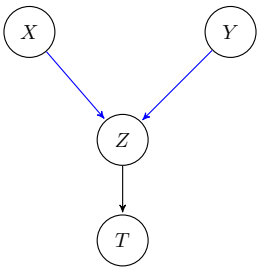
• Nous considérons le modèle graphique orienté donné par :



Pour toute distribution jointe $p \in \mathcal{L}(G)$, la factorisation associée au modèle graphique étudiée est :

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

• Ainsi, si on se réfère au théorème 1 (voir Annexes) vu en cours, on en déduit que $X \perp\!\!\!\perp Y \mid T$ si et seulement si toute chaîne dans le graphe reliant X à Y est bloquée. On considère alors la chaîne (X, Z, Y) mise en évidence en couleur sur le graphe suivant qui est la seule reliant X à Y:



Le noeud Z n'appartient pas à l'ensemble $C = \{T\}$ et (X, Z, Y) est une V-structure. Cependant, le noeud T est un descendant du noeud Z. Nous en déduisons que la chaîne (X, Z, Y) n'est pas bloquée par $C = \{T\}$. Cela implique que les noeuds X et Y ne sont pas d-séparés. Le théorème rappelé précédemment nous permet de conclure que **X et Y ne sont pas indépendantes conditionnellement à Z**.

Question 1.2.a

On considère que Z est une variable aléatoire binaire telle que $X \perp\!\!\!\perp Y|Z$ et $X \perp\!\!\!\perp Y$. On peut montrer en utilisant ces hypothèses que (pour le détail des calculs, on pourra se référer à l'annexe):

$$P(Z|X, Y) = \frac{P(Z|X)P(Z|Y)}{P(Z)}$$

De plus, si on note $\gamma = P(Z = 1)$, l'égalité $\sum_z p(z|y) = 1$ nous permet d'obtenir (avec les notations $x_i = P(Z = i|X)$ et $y_i = P(Z = i|Y)$) :

$$\frac{x_0 y_0}{1 - \gamma} + \frac{x_1 y_1}{\gamma} = 1$$

En utilisant le fait que $x_0 = 1 - x_1$, on obtient:

$$(\gamma - x_1)(\gamma - y_1) = 0$$

On en déduit alors que $\gamma = x_1$ ou $\gamma = y_1$. Autrement dit : $P(Z = 1) = P(Z = 1|X)$ ou $P(Z = 1) = P(Z = 1|Y)$. De plus vu que Z est binaire, légalité est aussi vraie pour $Z = 0$ (Voir calculs en annexes). Ainsi on a exactement que $X \perp\!\!\!\perp Z$ ou $Y \perp\!\!\!\perp Z$. **Ici, la propriété est vraie.**

Question 1.2.b

On considère la variable aléatoire $Z \sim \mathcal{U}(\{0, 1, 2, 3\})$ et les variables aléatoires X et Y définit telles que:

$$X = \mathbb{1}_{Z \in \{0,1\}}(Z) \qquad\qquad Y = \mathbb{1}_{Z \in \{0,2\}}(Z)$$

On peut alors montrer par le calcul que X et Y sont indépendantes. De même, on peut montrer que $X \perp\!\!\!\perp Y|Z$ (Pour les détails de calcul, voir l'annexe). Cependant, on peut aussi montrer que ni X ni Y n'est indépendant de Z. Ainsi, **la propriété générale est fausse.**

Factorisation de distributions dans un graphe

Question 2.1

Soit un DAG $G = (V, E)$. On considére deux sommets distincts de G notés i et j tel que $(i, j) \in E$ et tel que l'arête $i \to j$ soit *couverte*, i.e. $\pi_j = \pi_i \cup \{i\}$.
On note $G' = (V, E')$ le graphe tel que $E' = E \backslash \{i \to j\} \cup \{j \to i\}$.
Remarquons que les parents du noeud i dans le graphe G' (notés π'_i) sont $\pi_i \cup \{j\}$, et que les parents du noeud j dans le graphe G' (notés π'_j) sont $\pi_j \backslash \{i\} = \pi_i$ $(*)$.

- On peut montrer que G' est un DAG, *i.e.* G' est un graphe orienté sans cycle.

  Pour prouver le résultat, supposons par l'absurde que G' admette un cycle, alors ce cycle passe nécessairement par l'arête $j \to i$ car toutes les autres arêtes du graphes G' sont également présentes dans le graphe G (et G est acyclique).

  Il existe donc des sommets $v_1, \dots, v_K \in [\![1, n]\!] \backslash \{i, j\}$ (où $K \geq 1$ car $i \to j \notin E'$) deux à deux distincts tels que le chemin $j \to i \to v_1 \to \cdots \to v_K \to j$ soit dans G' et passant une seule fois par $j \to i$. v_K est donc un parent du noeud j dans G' ce qui signifie d'après $(*)$ que $v_K \in \pi_i$. Ainsi, le cycle $i \to v_1 \to \cdots \to v_K \to i$ est donc présent dans G, ce qui est absurde (car G est un DAG).

- Montrons que $\mathcal{L}(G) = \mathcal{L}(G')$.
  - Soit $p \in \mathcal{L}(G)$,

$$
\begin{aligned}
p(x) &= \prod_{l=1}^{n} p(x_l | x_{\pi_l}) \\
&= p(x_i | x_{\pi_i}) p(x_j | x_{\pi_j}) \times \underbrace{\prod_{l \notin \{i,j\}} p(x_l | x_{\pi_l})}_{= f(x)} \\
&= p(x_i | x_{\pi_i}) p(x_j | x_{\pi_i}, x_i) \times f(x) \\
&= p(x_i | x_{\pi_i}) \frac{p(x_j, x_i | x_{\pi_i})}{p(x_i | x_{\pi_i})} \times f(x) \\
&= p(x_j, x_i | x_{\pi_i}) \times f(x) \\
&= p(x_i | x_{\pi_i}, x_j) p(x_j | x_{\pi_i}) \times f(x) \\
&= p(x_i | x_{\pi'_i}) p(x_j | x_{\pi'_j}) \times \prod_{l \notin \{i,j\}} p(x_l | x_{\pi'_l})
\end{aligned}
$$

  où nous avons utilisé $(*)$ dans la dernière étape, i.e. :

$$
\pi'_l = \begin{cases} \pi_l & \text{si } l \notin \{i, j\} \\ \pi_i \cup \{j\} & \text{si } l = i \\ \pi_i & \text{si } l = j \end{cases}
$$

  La factorisation ainsi obtenue nous permet d'affirmer que la distribution p appartient à $\mathcal{L}(G')$.

  - Un calcul strictement analogue nous permet de montrer l'inclusion réciproque ie $\mathcal{L}(G') \subset \mathcal{L}(G)$.

Ainsi,

$$\boxed{\mathcal{L}(G) = \mathcal{L}(G').}$$

Question 2.2

Remarquons dans une premier temps que l'ensemble de cliques du graphe G' sont $C_i = \{(i, \pi_i),\ i \in V\}$ car G est un arbre orienté et donc vérifie: $|\pi_i| \leq 1$ pour tout i. On note $\mathcal{C}_{G'}$ l'ensemble de ces cliques.

• Soit $p \in \mathcal{L}(G)$: p se factorise sous la forme $p(x) = \prod_{j=1}^{n} p(x_j|x_{\pi_j})$ avec $|\pi_j| \leq 1$ car G est un arbre orienté (et donc ne possède pas de V-structure). En posant $\psi_j(x_j, x_{\pi_j}) = p(x_j|x_{\pi_j})$, nous obtenons que p peut se réécrire sous la forme :

$$p(x) = \prod_{j=1}^{n} \psi_j(x_j, x_{\pi_j}) = \prod_{C_j \in \mathcal{C}_{G'}} \psi_j(x_{C_j})$$

où les ψ_j sont des potentiels positifs (car une probabilité est positive) et le facteur de normalisation Z vaut 1.
Par définition de $\mathcal{L}(G')$, on a : $p \in \mathcal{L}(G')$. Ainsi : $\mathcal{L}(G) \subset \mathcal{L}(G')$.

• Montrons à présent l'inclusion inverse par récurrence sur la taille de l'arbre non orienté G'.
Soit $\mathcal{P}_n$ la propriété: "Pour tout arbre non orienté $G' = (V, E')$ avec $|V| \leq n$, et $G = (V, E)$ un arbre orienté associé à G', on a : $p \in \mathcal{L}(G') \implies p \in \mathcal{L}(G)$.

Initialisation : Cas $n = 1$
Pour $n = 1$, $\mathcal{L}(G')$ et $\mathcal{L}(G)$ représentent toutes les distributions de probabilités possibles et sont donc égaux.

Hérédité : Supposons la propriété $\mathcal{P}_n$ vraie pour un rang $n \geq 1$ fixé.
Soit $G' = (V, E')$ un arbre non orienté tel que $|V| = n + 1$ et G un arbre orienté associé à G'. Comme G est en particulier un DAG et que $n + 1 \geq 2$, il possède au moins un feuille qui n'est pas la racine que nous labelliserons $n + 1$ (quitte à renuméroter les noeuds du graphe G). Nous labellisons également son seul parent n.
Soit $p \in \mathcal{L}(G')$. Il existe donc des potentiels ψ_j tels que :

$$p(x) = \frac{1}{Z} \prod_{C_j \in \mathcal{C}_G} \psi_j(x_{C_j}) = \frac{1}{Z} \prod_{(i,j) \in E'} \psi_{(i,j)}(x_i, x_j)$$

On considère à présent l'arbre $\tilde{G}$ de taille n obtenu à partir de G en retirant la feuille $n + 1$. On note $\tilde{G}'$ l'arbre non orienté associé à $\tilde{G}$.
La probabilité marginale notée p_m de p de $(x_1, \ldots, x_n) = x_{1:n}$ s'écrit :

$$p_m(x_{1:n}) = \sum_{x_{n+1}} \frac{1}{Z} \prod_{(i,j) \in E'} \psi_{(i,j)}(x_i, x_j) = \frac{1}{Z}\ \psi_m(x_n) \prod_{(i,j) \in E' \setminus \{n,n+1\}} \psi_{(i,j)}(x_i, x_j)$$

où $\psi_m(x_n) = \sum_{x_{n+1}} \psi_{(n,n+1)}(x_n, x_{n+1})$.
Le calcul précédent montre que $p_m \in \mathcal{L}(\tilde{G}')$. Par hypothèse de récurrence $\mathcal{P}_n$, on en déduit que $p_m \in \mathcal{L}(G)$ et donc se factorise sous la forme : $p_m(x_{1:n}) = \prod_{i=1}^{n} q(x_i|x_{\pi_i})$.
En considérant à présent la fonction $q_{n+1}(x_{n+1}|x_n) = \psi_{(n,n+1)}(x_n, x_{n+1})/\psi_m(x_n)$ nous avons finalement que pour tout x :

$$p(x) = p_m(x_{1:n}) q_{n+1}(x_{n+1}|x_n) = \prod_{i=1}^{n} q(x_i|x_{\pi_i}) q_{n+1}(x_{n+1}|x_n)$$

Comme $\sum_{x_{n+1}} q_{n+1}(x_{n+1}|x_n) = 1$, et que x_n est le seul parent de x_{n+1}, nous avons montré que p appartient bien à $\mathcal{L}(G)$. Donc la propriété $\mathcal{P}_{n+1}$ est vraie.

Conclusion : D'après la principe de récurrence, pour tout $n \geq 1$, $\mathcal{P}_n$ est vraie. On a donc prouvé $\mathcal{L}(G') \subset \mathcal{L}(G)$.

En utilisant le résultat précédent on obtient donc le résultat désiré: Lorsque G est un arbre orienté:

$$\boxed{\mathcal{L}(G) = \mathcal{L}(G').}$$

Implémentation - Mélange de Gaussiennes

Question 3.a

Nous avons implémenté l'algorithme K-means et nous l'avons testé avec différentes initialisations aléatoires. Nous avons constaté que les résultats obtenus varient fortement suivant l'initialisation réalisée. L'algorithme K-means est réputé pour converger vers des minimaux locaux. Le problème considéré étant non convexe, il est possible que ces minimaux locaux soient non globaux, ce que nous pouvons observer sur l'exemple ici traité.

Question 3.b

On considère un mélange de gaussiennes dans lequel les matrices de covariances sont proportionnelles à l'identité. Notre modèle consiste à supposer que pour générer une nouvelle donnée, on observe la réalisation d'une variable aléatoire multinomiale $\mathcal{M}(1; \pi_1, \dots, \pi_K)$, puis que l'on tire une réalisation de la gaussienne dont le numéro correspondant à cette observation.
On considère que les K gaussiennes considérées ont pour moyennes $(\mu_k)_{k=1}^K \in \left(\mathbb{R}^2\right)^K$ et pour matrices de covariances $(\sigma_k^2 \mathbf{I_2})_{k=1}^K$. On notera θ l'ensemble des paramètres du modèles i.e. $\theta = (\pi_k, \mu_k, \sigma_k)_{k \in [\![1,K]\!]}$.
On considère les variables aléatoires $(Z_i^k)_{i \in [\![1,n]\!]}^{k \in [\![1,K]\!]}$ telles que :

$$\forall i \in [\![1,n]\!],\ \forall k \in [\![1,K]\!], \quad Z_i^k = \begin{cases} 1 & \text{si la donnée } i \text{ a été tirée selon la gaussienne n°}k \\ 0 & \text{sinon} \end{cases}$$

En se plaçant à une itération t de l'algorithme EM, l'étape E permet de calculer les probabilités a posteriori $(\gamma_i^k)^t = \mathbb{E}[Z_i^k | x_i, \theta_{t-1}]$. L'étape M consiste alors à calculer les paramètres du modèle permettant de maximiser l'approximation de la log-vraisemblance (donnée en calculant l'espérance de la log-vraisemblance par rapport aux variables aléatoires Z_i^k suivant une loi caractérisée par les $(\gamma_i^k)^t$).

π_k^t	μ_k^t	$(\sigma_k^2)^t$
$\frac{\sum_{i=1}^n (\gamma_i^k)^t}{n}$	$\frac{\sum_{i=1}^n (\gamma_i^k)^t x_i}{\sum_{j=1}^n (\gamma_j^k)^t}$	$\frac{\sum_{i=1}^n (\gamma_i^k)^t (x_i - \mu_k^t)^T (x_i - \mu_k^t)}{d \sum_{i=1}^n (\gamma_i^k)^t}$

Remarquons que dans le calcul des nouvelles valeurs de σ_k^2, nous utilisons les nouvelles valeurs des moyennes μ_k : μ_k^t (et non pas les μ_k^{t-1}). En cela, l'algorithme proposé est plutôt un ECM qu'un EM propre dit. Cette méthode assure une convergence plus rapide de l'algorithme.
Les détails des calculs sont donnés à la fin du rapport en Annexes.

Question 3.c

Dans le modèle général du mélange de gaussiennes, la seule différence par rapport au cas de la question précédente lors de l'étape M est l'actualisation des matrices de covariances. L'étape M à l'itération t de l'algorithme s'écrit pour tout $k \in [\![1,K]\!]$:

$$\Sigma_k^t = \frac{\sum_{i=1}^n (\gamma_i^k)^t (x_i - \mu_k^t)(x_i - \mu_k^t)^T}{\sum_{i=1}^n (\gamma_i^k)^t}.$$

Implémentation - Mélange de Gaussiennes

Question 3.d

• On peut déjà remarquer qu'on obtient des ellipses circulaires dans le cas de l'EM isotropique pour la représentation de 90% de la masse des gaussiennes. Ceci est logique car les matrices de covariances sont supposées proportionnelles à l'identité.
• On remarque également que les résultats obtenus pour la classification sont proches entre l'algorithme K-Means et l'algorithme EM Isotropique. Ceci est aussi logique car, comme les calculs le montre en annexe, on retrouve une sorte de K-means avec des noyaux gaussiens pour la log-vraisemblance de l'EM isotropique.
• Enfin, on constate que les valeurs des log-vraisemblance dépendent grandement de l'initialisation choisie. Cependant, si on normalise la valeur de la log-vraisemblance par le nombre de données, on constate que la valeur de la log-vraisemblance est toujours plus grand dans le cas du modèle général pour les données d'entraînement. Ceci semble logique car ce modèle est plus flexible, il est donc plus aisé d'approximer les données d'entraînement. Cependant, cela peut aussi entraîner plus facilement de l'over-fitting. Ainsi, en fonction de l'initialisation, la valeur de la log-vraisemblance du modèle isotropique peut être plus grande que celle du modèle général.

K-means

Figure 1 – Clustering donné par l'algorithme K-means (pour $K = 4$) sur les données d'entraînement pour une initialisation aléatoire des centroids.

EM Isotropique

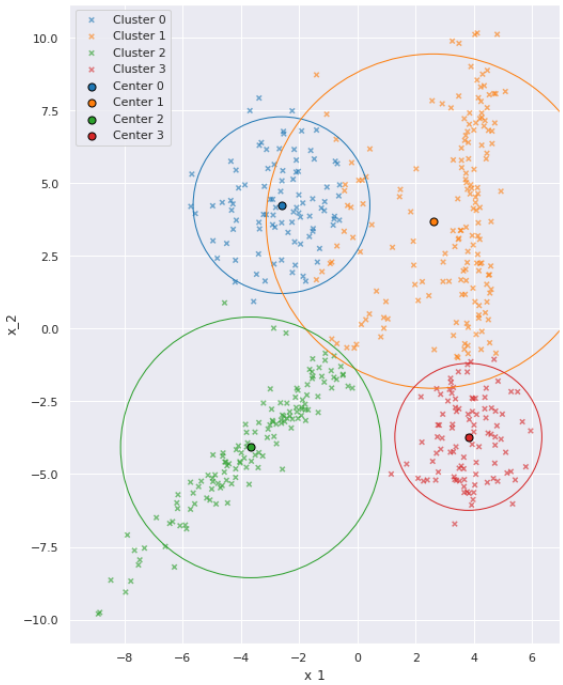


Figure 2 – Représentation des données d'entrainement. Les couleurs représentent le clustering donné par l'algorithm EM. Les ellipses contiennent 90% de la masse de la distribution de la gaussienne déterminée par l'EM.

EM Général

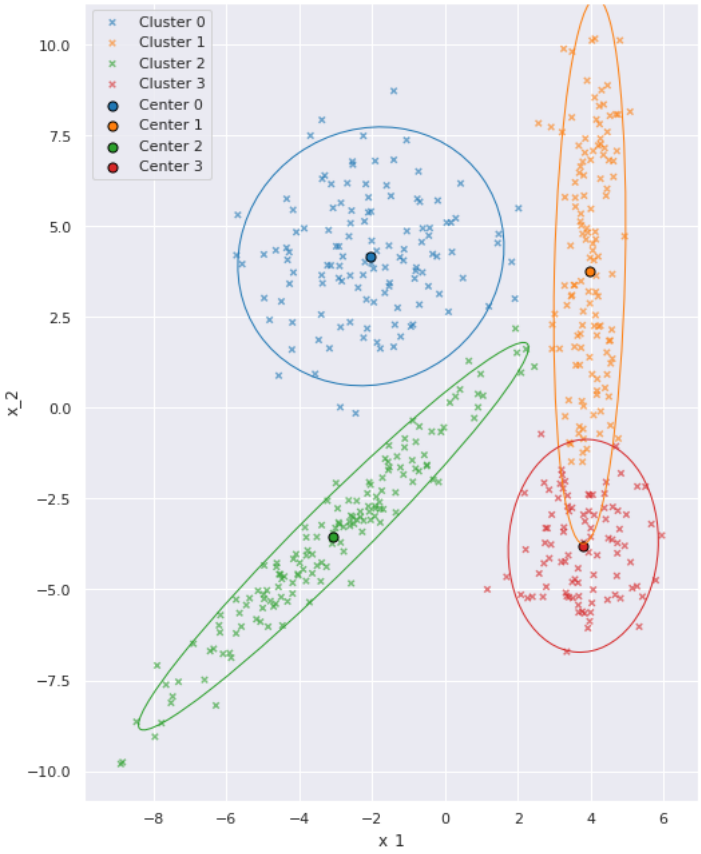


Figure 3 – Représentation des données d'entrainement. Les couleurs représentent le clustering donné par l'algorithm EM. Les ellipses contiennent 90% de la masse de la distribution de la gaussienne déterminée par l'EM.

Annexes

Question 1

Théorème 1
Soient $G = (\llbracket 1, n \rrbracket,\ E)$ un DAG et trois sous-ensembles disjoints de $\llbracket 1, n \rrbracket$ A, B et C.
Alors $X_A \perp\!\!\!\perp X_B | X_C$ si et seulement si pour tout $a \in A$ et $b \in B$, toute chaîne reliant a à b est bloquée par C.

Question 1.2.a

• Le premier résultat s'obtient de la manière suivante:

$$\begin{aligned} P(Z|X,Y) &= \frac{P(Z,X,Y)}{P(X,Y)} \\ &= \frac{P(X|Z,Y)P(Y,Z)}{P(X)P(Y)} && \text{par indépendance} \\ &= \frac{P(X|Z)P(Y,Z)}{P(X)P(Y)} && \text{par indépendance conditionnelle} \\ &= \frac{P(X,Z)P(Y,Z)}{P(Z)P(X)P(Y)} \\ &= \frac{P(Z|X)P(Z|Y)}{P(Z)} \end{aligned}$$

• Le deuxième résultat s'obtient via les calculs suivants:

$$\begin{aligned} P(Z=0|X,Y) + P(Z=1|X,Y) &= 1 \\ \frac{P(Z=0|X)P(Z=0|Y)}{1-\gamma} + \frac{P(Z=1|X)P(Z=1|Y)}{\gamma} &= 1 \\ \frac{x_0 y_0}{1-\gamma} + \frac{x_1 y_1}{\gamma} &= 1 \end{aligned}$$

Comme $x_0 = 1 - x_1$, nous pouvons réécrire l'égalité précédente sous la forme :

$$\gamma^2 - (x_1 + y_1)\gamma + x_1 y_1 = 0$$

$$\text{i.e. } (\gamma - x_1)(\gamma - y_1) = 0$$

• Pour le dernier résultat, on peut effectuer les calculs suivants:
Supposons par exemple que $P(Z = 1|X) = P(Z = 1)$. Montrons alors que nécessairement $P(Z = 0|X) = P(Z = 0)$.

$$\begin{aligned} P(Z=0|X) &= 1 - P(Z=1|X) \\ &= 1 - P(Z=1), && \text{en utilisant l'hypothèse} \\ &= P(Z=0) \end{aligned}$$

Le résultat s'obtient de la même manière pour Y. On vient donc de montrer que X ou Y est indépendant de tous événements de Z, ie X ou Y est indépendant de Z.

Question 1.2.b

On considère la variable aléatoire $Z \sim \mathcal{U}(\{0,1,2,3\})$ et les variables aléatoires X et Y définit telles que:

$$X = \mathbb{1}_{Z\in\{0,1\}}\ (Z) \qquad\qquad Y = \mathbb{1}_{Z\in\{0,2\}}\ (Z)$$

Avec $\mathbb{1}$ dénotant la fonction indicatrice. On peut déjà remarquer que comme Z suit une loi uniforme:

$$\forall z \in \{0,1,2,3\},\ \ \mathbb{P}(Z=z) = \frac{1}{4}$$

De même on peut remarquer facilement que:

$$\forall x \in \{0,1\},\ \ \mathbb{P}(X=x) = \frac{1}{2} \qquad \text{et} \qquad \forall y \in \{0,1\},\ \ \mathbb{P}(Y=y) = \frac{1}{2}$$

• Montrons que X et Y sont indépendantes. Soient $x, y \in \{0,1\}^2$:

$$\begin{aligned}
\mathbb{P}(X=x, Y=y) &= \begin{cases} \mathbb{P}(\{Z\in\{0,1\}\}\cap\{Z\in\{0,2\}\}) & \text{si } x=1 \text{ et } y=1 \\ \mathbb{P}(\{Z\in\{2,3\}\}\cap\{Z\in\{0,2\}\}) & \text{si } x=0 \text{ et } y=1 \\ \mathbb{P}(\{Z\in\{0,1\}\}\cap\{Z\in\{1,3\}\}) & \text{si } x=1 \text{ et } y=0 \\ \mathbb{P}(\{Z\in\{2,3\}\}\cap\{Z\in\{1,3\}\}) & \text{si } x=0 \text{ et } y=0 \end{cases} \\
&= \begin{cases} \mathbb{P}(Z=0) & \text{si } x=1 \text{ et } y=1 \\ \mathbb{P}(Z=2) & \text{si } x=0 \text{ et } y=1 \\ \mathbb{P}(Z=1) & \text{si } x=1 \text{ et } y=0 \\ \mathbb{P}(Z=3) & \text{si } x=0 \text{ et } y=0 \end{cases} \\
&= \frac{1}{4}
\end{aligned}$$

Or grâce aux résultats précédents sur X et Y on a, $\forall x, y \in \{0,1\}^2; \mathbb{P}(X=x) * \mathbb{P}(Y=y) = \frac{1}{2} * \frac{1}{2} = \frac{1}{4}$
Ainsi, on vient de prouver que X et Y sont bien indépendantes.

• Montrons maintenant que $X \perp\!\!\!\perp Y|Z$. Soient $x, y \in \{0,1\}^2, z \in \{0,1,2,3\}$,

$$\begin{aligned}
\mathbb{P}(X=x, Y=y|Z=z) &= \begin{cases} \mathbb{P}(\{Z\in\{0,1\}\}\cap\{Z\in\{0,2\}\} \mid Z=z) & \text{si } x=1 \text{ et } y=1 \\ \mathbb{P}(\{Z\in\{2,3\}\}\cap\{Z\in\{0,2\}\} \mid Z=z) & \text{si } x=0 \text{ et } y=1 \\ \mathbb{P}(\{Z\in\{0,1\}\}\cap\{Z\in\{1,3\}\} \mid Z=z) & \text{si } x=1 \text{ et } y=0 \\ \mathbb{P}(\{Z\in\{2,3\}\}\cap\{Z\in\{1,3\}\} \mid Z=z) & \text{si } x=0 \text{ et } y=0 \end{cases} \\
&= \begin{cases} \mathbb{P}(Z=0 \mid Z=z) & \text{si } x=1 \text{ et } y=1 \\ \mathbb{P}(Z=2 \mid Z=z) & \text{si } x=0 \text{ et } y=1 \\ \mathbb{P}(Z=1 \mid Z=z) & \text{si } x=1 \text{ et } y=0 \\ \mathbb{P}(Z=3 \mid Z=z) & \text{si } x=0 \text{ et } y=0 \end{cases} \\
&= \begin{cases} \mathbb{1}_{z=0} & \text{si } x=1 \text{ et } y=1 \\ \mathbb{1}_{z=2} & \text{si } x=0 \text{ et } y=1 \\ \mathbb{1}_{z=1} & \text{si } x=1 \text{ et } y=0 \\ \mathbb{1}_{z=3} & \text{si } x=0 \text{ et } y=0 \end{cases}
\end{aligned}$$

De même calculons $\mathbb{P}(X=x|Z=z) * \mathbb{P}(Y=y|Z=z)$:

$$\begin{aligned}
\mathbb{P}(X=x|Z=z) * \mathbb{P}(Y=y|Z=z) &= \begin{cases} \mathbb{P}(\{Z\in\{0,1\}\} \mid Z=z) * \mathbb{P}(\{Z\in\{0,2\}\}|Z=z) & \text{si } x=1 \text{ et } y=1 \\ \mathbb{P}(\{Z\in\{2,3\}\} \mid Z=z) * \mathbb{P}(\{Z\in\{0,2\}\}|Z=z) & \text{si } x=0 \text{ et } y=1 \\ \mathbb{P}(\{Z\in\{0,1\}\} \mid Z=z) * \mathbb{P}(\{Z\in\{1,3\}\}|Z=z) & \text{si } x=1 \text{ et } y=0 \\ \mathbb{P}(\{Z\in\{2,3\}\} \mid Z=z) * \mathbb{P}(\{Z\in\{1,3\}\}|Z=z) & \text{si } x=0 \text{ et } y=0 \end{cases} \\
&= \begin{cases} \mathbb{1}_{z\in\{0,1\}} * \mathbb{1}_{z\in\{0,2\}} & \text{si } x=1 \text{ et } y=1 \\ \mathbb{1}_{z\in\{2,3\}} * \mathbb{1}_{z\in\{0,2\}} & \text{si } x=0 \text{ et } y=1 \\ \mathbb{1}_{z\in\{0,1\}} * \mathbb{1}_{z\in\{1,3\}} & \text{si } x=1 \text{ et } y=0 \\ \mathbb{1}_{z\in\{2,3\}} * \mathbb{1}_{z\in\{1,3\}} & \text{si } x=0 \text{ et } y=0 \end{cases}
\end{aligned}$$

$$= \begin{cases} \mathbb{1}_{z=0} & \text{si } x=1 \text{ et } y=1 \\ \mathbb{1}_{z=2} & \text{si } x=0 \text{ et } y=1 \\ \mathbb{1}_{z=1} & \text{si } x=1 \text{ et } y=0 \\ \mathbb{1}_{z=3} & \text{si } x=0 \text{ et } y=0 \end{cases}$$
$$= \mathbb{P}(X=x, Y=y|Z=z)$$

On vient donc de montrer que $X \perp\!\!\!\perp Y|Z$.

Montrons maintenant que ni X ni Y n'est indépendant de Z. Soit $x \in \{0,1\}, z \in \{0,1,2,3\}$:

$$\mathbb{P}(X=x \mid Z=z) = \begin{cases} \mathbb{P}(\{Z \in \{0,1\}\} \mid Z=z) & \text{si } x=1 \\ \mathbb{P}(\{Z \in \{2,3\}\} \mid Z=z) & \text{si } x=0 \end{cases}$$
$$= \begin{cases} \mathbb{1}_{z\in\{0,1\}} & \text{si } x=1 \\ \mathbb{1}_{z\in\{2,3\}} & \text{si } x=0 \end{cases}$$

Or, $\forall x \in \{0,1\}$: $\mathbb{P}(X=x) = \frac{1}{2}$. Ainsi si $z=0$ et $x=0$, $\frac{1}{2} = \mathbb{P}(X=x) \neq \mathbb{P}(X=x \mid Z=z) = 0$. On en déduit que X n'est pas indépendant de Z.

De la même manière on montre que Y n'est pas indépendant de Z en prenant par exemple $z=0$ et $y=0$.

On vient donc de prouver le résultat.

Question 3

Détails des calculs de l'exercice 3.b

On considère un mélange de gaussiennes dans lequel les matrices de covariances sont proportionnelles à l'identité. Notre modèle consiste à supposer que pour générer une nouvelle donnée, on observe la réalisation d'une variable aléatoire multinomiale $\mathcal{M}(1;\pi_1,\dots,\pi_K)$, puis que l'on tire une réalisation de la gaussienne dont le numéro correspondant à cette observation.
On considère que les K gaussiennes considérées ont pour moyennes $(\mu_k)_{k=1}^K \in \left(\mathbb{R}^2\right)^K$ et pour matrices de covariances $(\sigma_k^2\mathbf{I_2})_{k=1}^K$. On notera θ l'ensemble des paramètres du modèles i.e. $\theta = (\pi_k, \mu_k, \sigma_k)_{k\in[\![1,K]\!]}$.
On considère les variables aléatoires $(Z_i^k)_{i\in[\![1,n]\!]}^{k\in[\![1,K]\!]}$ telles que :

$$\forall i \in [\![1,n]\!],\ \forall k \in [\![1,K]\!], \quad Z_i^k = \begin{cases} 1 & \text{si la donnée } i \text{ a été tirée selon la gaussienne n°}k \\ 0 & \text{sinon} \end{cases}$$

La log-vraisemblance s'écrit :

$$\begin{aligned} \log(VS) &= \log\left(\prod_{i=1}^n p(x_i)\right) \\ &= \log\left(\prod_{i=1}^n \prod_{k=1}^K \left(p(x_i|Z_i^k=1)\mathbb{P}(Z_i^k=1)\right)^{Z_i^k}\right) \\ &= \sum_{i=1}^n \sum_{k=1}^K Z_i^k \left(\log(p(x_i|Z_i^k=1)) + \log(\mathbb{P}(Z_i^k=1))\right) \\ &= \sum_{i=1}^n \sum_{k=1}^K Z_i^k \left(-\log(2\pi\sigma_k^2) - \frac{(x_i-\mu_k)^T(x_i-\mu_k)}{2\sigma_k^2} + \log(\mathbb{P}(Z_i^k=1))\right) \end{aligned}$$

• On se place à une itération t de l'algorithme EM. On suppose que l'étape E vient d'être réalisée, i.e. que nous venons de calculer les probabilités a posteriori:

$$
\begin{aligned}
\gamma_i^k &= \mathbb{E}[Z_i^k | x_i, \theta_{t-1}] \\
&= \mathbb{P}(Z_i^k = 1 | x_i, \theta_{t-1}) \\
&= \frac{p(x_i | Z_i^k = 1, \theta_{t-1}) \mathbb{P}(Z_i^k = 1 | \theta_{t-1})}{\mathbb{P}(x_i | \theta_{t-1})} \\
&= \frac{p(x_i | Z_i^k = 1, \theta_{t-1}) \mathbb{P}(Z_i^k = 1 | \theta_{t-1})}{\sum_{l=1}^K \mathbb{P}(x_i | Z_l^k = 1, \theta_{t-1}) \mathbb{P}(Z_l^k = 1 | \theta_{t-1})} \\
&= \frac{\bar{\pi}_k \left(2\pi\sigma_k^2\right)^{-1} \exp\left(-\frac{(x_i - \bar{\mu}_k)^T (x_i - \bar{\mu}_k)}{2\bar{\sigma}_k^2}\right)}{\sum_{l=1}^K \bar{\pi}_l \left(2\pi\sigma_l^2\right)^{-1} \exp\left(-\frac{(x_i - \bar{\mu}_l)^T (x_i - \bar{\mu}_l)}{2\bar{\sigma}_l^2}\right)}
\end{aligned}
$$

où nous avons noté $\theta_{t-1} = (\bar{\pi}_k, \bar{\mu}_k, \bar{\sigma}_k)_{k \in [\![1,K]\!]}$ pour plus de lisibilité.

L'étape M consiste alors à maximiser la log-vraisemblance des observations par rapport aux paramètres $(\pi_k, \mu_k, \sigma_k)_k$. La problème d'optimisation étant séparable en k, on cherche donc à calculer pour tout $k \in [\![1, K]\!]$:

$$
\underset{\pi_k, \mu_k, \sigma_k}{argmax} \sum_{i=1}^n \gamma_i^k \left(-\log(2\pi\sigma_k^2) - \frac{(x_i - \mu_k)^T (x_i - \mu_k)}{2\sigma_k^2} + \log(\pi_k) \right)
$$

Nous cherchons donc à maximiser sur un ouvert convexe non vide une fonction concave. La condition nécessaire et suffisante d'optimisalité s'écrit alors en annulant le gradient du lagrangien de notre problème. Nous obtenons alors :

- En dérivant par rapport à σ_k:

$$
\sum_{i=1}^n -\frac{2\gamma_i^k}{\sigma_k} + \gamma_i^k \frac{(x_i - \mu_k)^T (x_i - \mu_k)}{\sigma_k^3} = 0
$$

$$
\boxed{\sigma_k^2 = \frac{\sum_{i=1}^n \gamma_i^k (x_i - \mu_k)^T (x_i - \mu_k)}{2 \sum_{i=1}^n \gamma_i^k}}
$$

- En dérivant par rapport à μ_k:

$$
\sum_{i=1}^n \gamma_i^k \left(-2x_i + 2\mu_k\right) = 0
$$

$$
\boxed{\mu_k = \frac{\sum_{i=1}^n \gamma_i^k x_i}{\sum_{j=1}^n \gamma_j^k}}
$$

- En dérivant par rapport à π_k le lagrangien:

  Notant λ le multiplicateur de Lagrange associé à la contrainte $\sum_{k=1}^K \pi_k = 1$, on a :

$$
\begin{aligned}
\frac{\sum_{i=1}^n \gamma_i^k}{\pi_k} - \lambda &= 0 \\
\lambda \pi_k &= \sum_{i=1}^n \gamma_i^k
\end{aligned}
$$

La contrainte $\sum_{k=1}^{K} \pi_k = 1$ fournit alors la relation $\lambda = \sum_{k=1}^{K} \sum_{i=1}^{n} \gamma_i^k = n$, d'où:

$$\pi_k = \frac{\sum_{i=1}^{n} \gamma_i^k}{n}$$