1 Indépendances conditionnelles et factorisations

1. Pour le graphe donné, toute distribution $p \in \mathcal{L}(G)$ se factorise comme suit :

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

On peut trouver des distributions dans $\mathcal{L}(G)$ telles que X et Y sont dépendantes conditionnellement à T. Considérons en effet $T = Z$ et $Z = \mathbf{1}_{X=Y}$ avec X et Y indépendantes et dont les ensembles de valeurs possibles ne sont pas disjoints. On a alors la dépendance de X et Y sachant T. Par exemple, dans le cas où X et Y suivent toutes deux une loi de Bernouilli de paramètre $1/2$, on peut écrire :

$$P(X=0,Y=1|Z=1) = 0 \neq \begin{cases} P(X=0|Z=1) &= \frac{1}{2} \\ \times & \\ P(Y=1|Z=1) &= \frac{1}{2} \end{cases}$$

2. (a) Soit X et Y indépendantes, et indépendantes conditionnellement à Z, où Z est binaire.
Ecrivons $P(X=x,Y=y)$ de deux façons, la première en utilisant l'indépendance de X et Y (équation (1)) et la seconde en utilisant l'indépendance de X et Y conditionnellement à Z (équation (3)).

$$\begin{aligned} P(X=x,Y=y) =& P(X=x)P(Y=y) && (1)\\ =& \big(P(X=x|Z=0)P(Z=0) + P(X=x|Z=1)P(Z=1)\big) \\ & \times \big(P(Y=y|Z=0)P(Z=0) + P(Y=y|Z=1)P(Z=1)\big) \\ =& P(X=x|Z=0)P(Y=y|Z=0)P(Z=0)^2 \\ & + P(X=x|Z=0)P(Y=y|Z=1)P(Z=0)P(Z=1) \\ & + P(X=x|Z=1)P(Y=y|Z=0)P(Z=0)P(Z=1) \\ & + P(X=x|Z=1)P(Y=y|Z=1)P(Z=1)^2 && (2) \end{aligned}$$

$$\begin{aligned} P(X=x,Y=y) =& P(X=x,Y=y|Z=0)P(Z=0) + P(X=x,Y=y|Z=1)P(Z=1) && (3)\\ =& P(X=x|Z=0)P(Y=y|Z=0)P(Z=0) \\ & + P(X=x|Z=1)P(Y=y|Z=1)P(Z=1) && (4) \end{aligned}$$

En écrivant (2) − (4) = 0 et étant donné $P(Z=0) + P(Z=1) = 1$ on obtient :

$$\begin{aligned} 0 =& P(Z=0)P(Z=1)\big[- P(X=x|Z=0)P(Y=y|Z=0) + P(X=x|Z=0)P(Y=y|Z=1) \\ & + P(X=x|Z=1)P(Y=y|Z=0) - P(X=x|Z=1)P(Y=y|Z=1)\big] \end{aligned}$$

Si Z constant, l'implication est évidente car Z indépendant de toute variable.
Sinon, on a $P(Z=0)P(Z=1) \neq 0$ d'où :

$$\begin{aligned} 0 =& P(X=x|Z=0)\big(P(Y=y|Z=0) - P(Y=y|Z=1)\big) \\ & - P(X=x|Z=1)\big(P(Y=y|Z=0) - P(Y=y|Z=1)\big) \\ =& \big(P(X=x|Z=0) - P(X=x|Z=1)\big)\big(P(Y=y|Z=0) - P(Y=y|Z=1)\big) \end{aligned}$$

Au moins l'un des deux termes du produit est nul. Si $P(X=x|Z=1) = P(X=x|Z=0)$, on a $X \perp\!\!\!\perp Z$. Si $P(Y=y|Z=1) = P(Y=y|Z=0)$, on a $Y \perp\!\!\!\perp Z$. **On a donc bien démontré que : si X et Y sont indépendantes, et indépendantes conditionnellement à Z avec Z binaire, alors on a l'indépendance de X et Z ou celle de Y et Z (ou les deux).**

2. (b) Dans le cas général, prenons X et Y deux variables indépendantes telles que $P(X<0)>0$, $P(Y<0)>0$, $P(X>0)>0$, $P(Y>0)>0$ (par exemple suivant chacune une loi de Bernouilli de paramètre $1/2$ à valeurs dans $\{-1,1\}$). Et posons $Z = (\mathbf{1}_{X>0}\ ,\ \mathbf{1}_{Y>0})$. On a clairement la dépendance de X et Z, comme celle de Y et Z. Et en meme temps on a clairement $X \perp\!\!\!\perp Y|Z$. (Pour le justifier formellement, on peut écrire pour X, Y variables de Bernouilli : $\forall\, i,j \in \{-1,1\}\ , \forall k,l \in \{0,1\}, P(X=i,Y=j|Z=(k,l)) = \delta_{i,2k-1}\delta_{j,2l-1} = P(X=i|Z=(k,l))P(Y=j|Z=(k,l))$)

Dans le cas général, la propriété n'est donc pas vraie.

2 Factoriser des distributions dans un graphe (arbre)

1. Le détail des calculs, et une figure éclairante, sont données en section 4.1 (pour des raisons de place). On raisonne par équivalence :

$$\begin{aligned} p \in \mathcal{L}(G) &\Leftrightarrow \forall x, \quad p(x) = \prod_{k \in V} p(x_k | x_{\pi_G(k)}) \\ &\Leftrightarrow \forall x, \quad p(x) = p(x_i | x_{\pi_G(i)}) p(x_j | x_i, x_{\pi_G(i)}) \prod_{k \in V \setminus \{i,j\}} p(x_k | x_{\pi_G(k)}) \\ &\Leftrightarrow \forall x, \quad p(x) = p(x_i | x_j, x_{\pi_{G'}(j)}) p(x_j | x_{\pi_{G'}(j)}) \prod_{k \in V \setminus \{i,j\}} p(x_k | x_{\pi_{G'}(k)}) \\ &\Leftrightarrow p \in \mathcal{L}(G') \end{aligned}$$

2. Montrons $\mathcal{L}(G) \subset \mathcal{L}(G')$. Soit $p \in \mathcal{L}(G)$. p peut se factoriser comme suit : $\forall x, \ p(x) = \prod_{i \in V} p(x_i | x_{\pi_i}) = p(x_0) \prod_{(j,i) \in E} p(x_i | x_j)$ où 0 est l'indice de la racine de l'arbre et E l'ensemble des arêtes de l'arbre G.

Par conséquent, il suffit de définir les potentiels $\psi_0(x_0) = p(x_0)$ et $\psi_{ij}(x_i, x_j) = p(x_i | x_j) \ \ \forall (j,i) \in E$ pour pouvoir écrire : $p(x) = \psi(x_0) \prod_{(j,i) \in E} \psi_{ij}(x_i, x_j)$. Comme les cliques maximales d'un arbre non dirigé sont les paires de noeuds reliés par une arete, **cela prouve** $\boxed{p \in \mathcal{L}(G')}$.

Montrons l'inclusion réciproque $\mathcal{L}(G') \subset \mathcal{L}(G)$ Nous allons procéder par récurrence sur le nombre de noeuds de G'. Notre hypothèse de récurrence s'énonce donc comme suit : (H_n) : *Pour tout arbre non dirigé G' tel que $|V(G')| = n$, on a $\mathcal{L}(G') \subset \mathcal{L}(G)$.*

- (H_1) est bien évidemment vrai (arbre réduit à un noeud donc $\mathcal{L}(G') = \mathcal{L}(G)$).
- Supposons (H_n) et montrons (H_{n+1}). Soit $p \in \mathcal{L}(G')$ où $|V(G')| = n+1$. Désignons par l'indice $n+1$ une feuille de l'arbre et par l'indice n son parent.

$$p(x_1, ..., x_n, x_{n+1}) = \frac{1}{Z} \left[\prod_{\{i,j\} \in E \setminus \{n,n+1\}} \psi_{i,j}(x_i, x_j) \right] \psi_{n,n+1}(x_n, x_{n+1})$$

Supposons que x_n est tel que $\sum_{x_{n+1}} \psi_{n,n+1}(x_n, x_{n+1}) > 0$. Alors on peut écrire :

$$p(x_1, ..., x_n, x_{n+1}) = f(x_1, ...x_n) g(x_n, x_{n+1})$$

où on pose :

$$f(x_1, ...x_n) = \frac{1}{Z} \left[\prod_{\{i,j\} \in E \setminus \{n,n+1\}} \psi_{i,j}(x_i, x_j) \right] \sum_{x_{n+1}} \psi_{n,n+1}(x_n, x_{n+1})$$

et

$$g(x_n, x_{n+1}) = \frac{\psi_{n,n+1}(x_n, x_{n+1})}{\sum_{x_{n+1}} \psi_{n,n+1}(x_n, x_{n+1})}$$

.

Pour tout x_n vérifiant $\sum_{x_{n+1}} \psi_{n,n+1}(x_n, x_{n+1}) > 0$, $g(x_n, .)$ définit bien une probabilité.
Dans le cas où $\sum_{x_{n+1}} \psi_{n,n+1}(x_n, x_{n+1}) = 0$, on peut étendre la définition de g en posant $g(x_n, x_{n+1}) = \frac{1}{|S(x_{n+1})|}$ (où $S(x_{n+1})$ est l'ensemble des valeurs que peut prendre x_{n+1}) : $\sum_{x_{n+1}} g(x_n, x_{n+1}) = 1$ et on vérifie toujours $p(x_1, ..., x_n, x_{n+1}) = f(x_1, ...x_n) g(x_n, x_{n+1})$.
Par conséquent, on peut écrire : pour tout x_n, et donc pour tout $x = (x_1, ..., x_n)$:

$$p(x_1, ..., x_n, x_{n+1}) = p(x_{n+1} | x_n) f(x_1, ...x_n)$$

.

Enfin, $f \in \mathcal{L}(G'')$ où G'' est le graphe G' auquel on a retiré la feuille n. Par hypothèse de récurrence, $f \in \mathcal{L}(G)$, et donc :

$$p(x_1, ..., x_n, x_{n+1}) = p(x_{n+1} | x_n) \prod_{i \in \{1,...,n\}} p(x_i | x_{\pi(i)})$$

Ce qui montre $\boxed{p \in \mathcal{L}(G)}$ **et achève la récurrence.** L'inclusion réciproque est donc démontrée.

On conclut donc que $\mathcal{L}(G) = \mathcal{L}(G')$ si G est un arbre dirigé et G' son homologue non dirigé.

3 Implémentation - Modèles de mélange Gaussien

(a) Lorsqu'on lance 1000 fois l'algorithme de *kmeans*, on obtient environ 1% du temps le cas de la figure 4 (distortion de l'ordre de 1500) et le reste du temps le comportement attendu de la figure 1 (distortion de l'ordre de 1100). Selon l'initialisation de l'algorithme, on peut en effet aboutir à différents minimums locaux.

Comportement	Dataset	Distortion
"Bon" (Fig 1)	Train	1110
	Test	1090
"Mauvais" (Fig 4)	Train	1550
	Test	1460

(b) Le détail des calculs est donnée en partie 4.2 (raison de place). Le modèle s'écrit (notations introduites en 4.2) : $Z \sim \mathcal{M}(1, \pi_1, ...\pi_K)$; $X|Z = ((\delta(j,k))_{1\leq j\leq K}) \sim G(\mu_k, \sigma_k^2 I) \ \ \forall k \in [1, K]$

Notant $\Theta = \{\pi, \mu, \sigma^2\}$, la log-vraisemblance complète du modèle est, pour N observations i.i.d, :

$$l(Z, \Theta) = \sum_{i=1}^{N}\sum_{k=1}^{K} z_i^k \left[-\frac{d}{2}\log\sigma_k^2 + \log\pi_k - \frac{||x_i - \mu_k||^2}{2\sigma_k^2}\right] + cste$$

A l'itération $n+1$ de l'algorithme EM, on maximise la quantité $E_{Z\sim P(.|X,\Theta^n)}[l(Z,\Theta)]$

— On calcule donc $E_{Z\sim P(.|X,\Theta^n)}[l(Z,\Theta)]$:

$$E_{Z\sim P(.|X,\Theta^n)}[l(Z,\Theta)] = \sum_{i=1}^{N}\sum_{k=1}^{K} p_i^{k^{n+1}} \left[-\frac{d}{2}\log\sigma_k^2 + \log\pi_k - \frac{||x_i - \mu_k||^2}{2\sigma_k^2}\right] + cste$$

où $p_i^{k^{n+1}} = P(Z = z_i^k | X = x_i, \Theta^n) = C_{x_i}\frac{1}{(2\pi\sigma_k^{2^n})^{\frac{d}{2}}}\exp\left[-\frac{||x_i-\mu_k^n||^2}{2\sigma_k^{2^n}}\right]\pi_k^n$,avec C_{x_i} telle que $\sum_k p_i^{k^{n+1}} = 1$

— Ensuite on cherche Θ maximisant $E_{Z\sim P(.|X,\Theta^n)}[l(Z,\Theta)]$ sans oublier la contrainte $\sum_k \pi_k = 1$. On peut décomposer la fonction objectif comme suit : $E_{Z\sim P(.|X,\Theta^n)}[l(Z,\Theta)] = l_1(\pi) + l_2(\mu,\sigma) + cste$. En cherchant les maximums de l_1 et l_2, toutes deux concaves, on obtient donc les équations de mise à jour suivantes :

$$\pi_k^{n+1} = \frac{\sum_{i=1}^N p_i^{k^{n+1}}}{N}$$
$$\mu_k^{n+1} = \frac{\sum_{i=1}^N p_i^{k^{n+1}} x_i}{\sum_{i=1}^N p_i^{k^{n+1}}}$$
$$\sigma_k^{2^{n+1}} = \frac{1}{d}\frac{\sum_{i=1}^N p_i^{k^{n+1}} ||x_i - \mu_k^{n+1}||^2}{\sum_{i=1}^N p_i^{k^{n+1}}}$$

(c) Dans le cas général $(X|Z = (\delta(j,k))_{1\leq j\leq K} \sim G(\mu_k, \Sigma_k^2))$, on met à jour π et μ de la meme façon que dans le cas isotropique. Puis on met à jour Σ_k^2, la matrice de covariance, selon (cf slides de cours) :

$$\Sigma_k^{2^{n+1}} = \frac{\sum_{i=1}^N p_i^{k^{n+1}}(x_i - \mu_k^{n+1})(x_i - \mu_k^{n+1})^T}{\sum_{i=1}^n p_i^{k^{n+1}}}$$

Note : en partie 4.2, on détaille également le raisonnement menant à la représentation des matrices de covariances pour les deux modèles (comme cercles dans le cas isotropique, et comme ellipses dans le cas général).

(d) Dans la table suivante, on donne, pour chacun des deux modèles et sur chacun des jeux Train/Test, les log-vraisemblances finales "moyennes" (log-vraisemblance divisée par le nombre de données du jeu, ce afin de pouvoir comparer les résultats sur les jeux de Train et de Test).

— Comme attendu, on observe que, pour chacun des deux modèles de mélange, la log-vraisemblance obtenue en test est plus faible que celle en train.

— On observe également que les log-vraisemblances du modèle de mélange général en Train et en Test sont plus grandes que celles du modèle de mélange dans le cas isotropique. Les données en Train/Test semblent en effet avoir été générés depuis des gaussiennes de matrices de covariance non isotropiques (cf. Figures 2, 3) et il est donc logique que le modèle de mélange général donne de meilleurs résultats que le modèle isotropique. Pour une telle distribution de données, ce modèle génératif donne donc aussi de bien meilleurs résultats que KMeans (avec une intialisation par KMeans).

Modèle | Jeu de données	Train	Test
GMM isotropique	-5.29	-5.38
GMM général	-4.66	-4.82

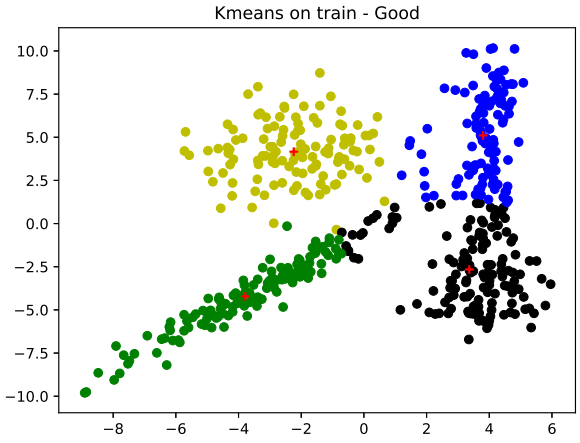

Figure 1 – Comportement « bon » du *kmeans* sur les données de *train*. Centres des clusters en rouge.

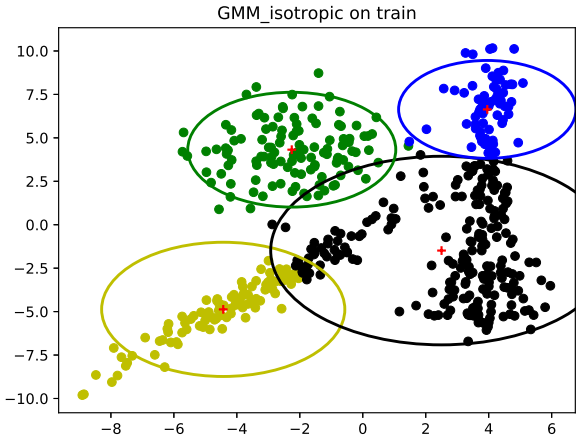

Figure 2 – Modèle de mélange gaussien isotropique, sur les données de *train*. Initialisation avec KMeans. Centres des clusters en rouge. Chaque cercle contient 90% de la masse de la distribution gaussienne correspondante.

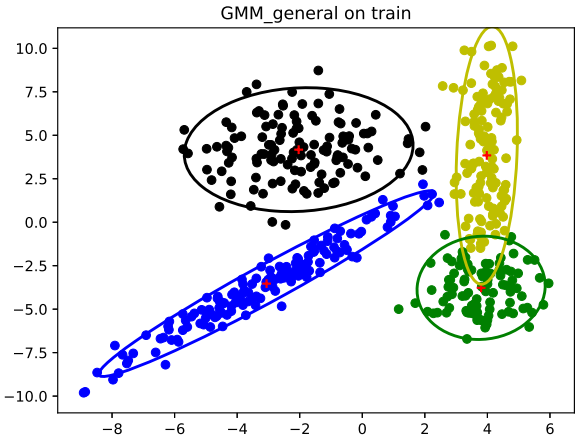

Figure 3 – Modèle de mélange gaussien général, sur les données de *train*. Initialisation avec KMeans. Centres des clusters en rouge. Chaque ellipse contient 90% de la masse de la distribution gaussienne correspondante.

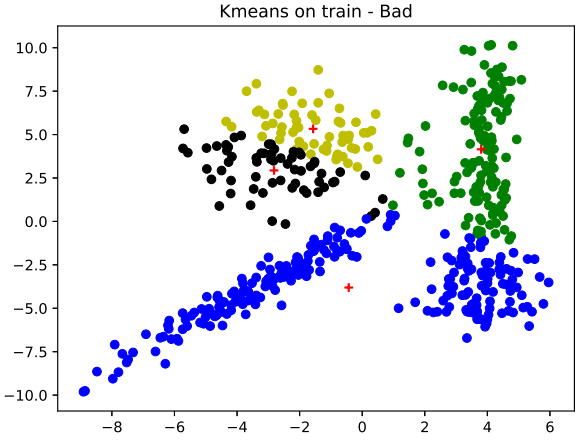

Figure 4 – Comportement « mauvais » du *kmeans* sur les données de *train*. Centres des clusters en rouge.

4 Détails des calculs/raisonnements

4.1 Exercice 2 : Question 1.

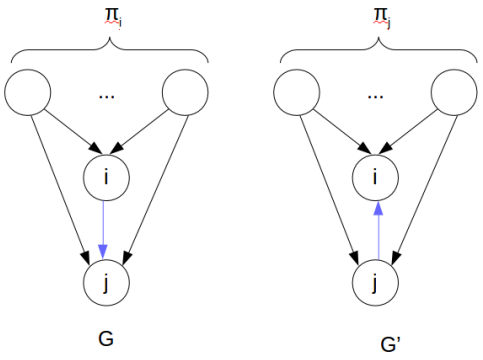

FIGURE 5 – Représentation des graphs G et G' et de l'arête converte.

Dans cette question, on notera $V(G) = V(G') = V$ l'ensemble des sommets de G (ou G').
Et on notera $\pi_G(i)$ (resp. $\pi_{G'}(i)$) les parents du noeud i dans le graphe G (resp. dans le graphe G'). On raisonne par équivalences :

$$
\begin{aligned}
& p \in \mathcal{L}(G) \\
\Leftrightarrow \forall x, \quad & p(x) = \prod_{k \in V} p(x_k | x_{\pi_G(k)}) \\
\Leftrightarrow \forall x, \quad & p(x) = p(x_i | x_{\pi_G(i)}) p(x_j | x_{\pi_G(j)}) \prod_{k \in V \setminus \{i,j\}} p(x_k | x_{\pi_G(k)}) \\
\Leftrightarrow \forall x, \quad & p(x) = p(x_i | x_{\pi_G(i)}) p(x_j | x_i, x_{\pi_G(i)}) \prod_{k \in V \setminus \{i,j\}} p(x_k | x_{\pi_G(k)}) && \text{(puisque i} \to \text{j est une arête couverte dans G)} \\
\Leftrightarrow \forall x, \quad & p(x) = \frac{p(x_i, x_{\pi_G(i)})}{p(x_{\pi_G(i)})} \frac{p(x_j, x_i, x_{\pi_G(i)})}{p(x_i, x_{\pi_G(i)})} \prod_{k \in V \setminus \{i,j\}} p(x_k | x_{\pi_G(k)}) \\
\Leftrightarrow \forall x, \quad & p(x) = \frac{p(x_j, x_i, x_{\pi_G(i)})}{p(x_{\pi_G(i)})} \prod_{k \in V \setminus \{i,j\}} p(x_k | x_{\pi_G(k)}) \\
\Leftrightarrow \forall x, \quad & p(x) = \frac{p(x_j, x_i, x_{\pi_G(i)})}{p(x_j, x_{\pi_G(i)})} \frac{p(x_j, x_{\pi_G(i)})}{p(x_{\pi_G(i)})} \prod_{k \in V \setminus \{i,j\}} p(x_k | x_{\pi_G(k)}) \\
\Leftrightarrow \forall x, \quad & p(x) = p(x_i | x_{\pi_G(i)}, x_j) p(x_j | x_{\pi_G(i)}) \prod_{k \in V \setminus \{i,j\}} p(x_k | x_{\pi_G(k)}) \\
\Leftrightarrow \forall x, \quad & p(x) = p(x_i | x_{\pi_{G'}(j)}, x_j) p(x_j | x_{\pi_{G'}(j)}) \prod_{k \in V \setminus \{i,j\}} p(x_k | x_{\pi_{G'}(k)}) && \text{(par construction de G', } \pi_G(i) = \pi_{G'}(j) \text{ et} \\
& && \pi_G(k) = \pi_{G'}(k) \ \ \forall k \in V \setminus \{i,j\}) \\
\Leftrightarrow \forall x, \quad & p(x) = p(x_i | x_{\pi_{G'}(i)}) p(x_j | x_{\pi_{G'}(j)}) \prod_{k \in V \setminus \{i,j\}} p(x_k | x_{\pi_{G'}(k)}) && \text{(par construction de G')} \\
\Leftrightarrow & \ p \in \mathcal{L}(G')
\end{aligned}
$$

4.2 Exercice 3 : (b)

On considère un modèle de mélange de gaussiennes dont les matrices de covariances sont proportionnelles à l'identité.

Considérons K gaussiennes, et désignons par X les observations et Z les variables latentes du modèle :

Z_n^k égale 1 si l'observation X_n est généré par la gaussienne numéro k, et 0 sinon. Notant $\mathcal{M}$ et G les lois multinomiales et gaussiennes, le modèle génératif peut s'écrire comme suit :

$Z \sim \mathcal{M}(1, \pi_1, ...\pi_K)$

$X|Z = ((\delta(j,k))_{1 \leq j \leq K}) \sim G(\mu_k, \sigma_k^2 I) \quad \forall k \in [1, K]$

Notons $\Theta = \{\pi, \mu, \sigma^2\}$, la vraisemblance complète du modèle s'écrit pour N observations i.i.d :

$$\begin{aligned}
\mathcal{L}(Z, \Theta) &= \prod_{i=1}^{N} P(Z = z_i, X = x_i|\Theta) \\
&= \prod_{i=1}^{N} P(Z = z_i|\pi) P(X = x_i|Z = z_i, \mu, \sigma) \\
&= \prod_{i=1}^{N} \prod_{k=1}^{K} \left[P(Z^k = z_i^k|\pi) P(X = x_i|Z^k = z_i^k, \mu, \sigma) \right]^{z_i^k} \\
&= \prod_{i=1}^{N} \prod_{k=1}^{K} \left[\pi_k \frac{1}{(2\pi)^{\frac{d}{2}} \left|\sigma_k^2 I\right|^{\frac{1}{2}}} \exp\left[-\frac{1}{2\sigma_k^2} (x_i - \mu_k)^T (x_i - \mu_k) \right] \right]^{z_i^k} \\
&= cste \times \prod_{k=1}^{K} \prod_{i=1}^{N} \frac{\pi_k^{z_i^k}}{(\sigma_k^2)^{z_i^k \frac{d}{2}}} \exp\left[-\frac{||x_i - \mu_k||^2}{2\sigma_k^2} \right]^{z_i^k}
\end{aligned}$$

Ainsi, la log-vraisemblance complète est :

$$l(Z, \Theta) = \sum_{i=1}^{N} \sum_{k=1}^{K} z_i^k \left[-\frac{d}{2} \log \sigma_k^2 + \log \pi_k - \frac{||x_i - \mu_k||^2}{2\sigma_k^2} \right] + cste$$

A l'itération $n+1$ de l'algorithme EM, on maximise la quantité $E_{Z \sim P(.|X,\Theta^n)}[l(Z, \Theta)]$.

— Calculons d'abord $E_{Z \sim P(.|X,\Theta)}[l(Z, \Theta)]$:
$E_{Z \sim P(.|X,\Theta)}[l(Z, \Theta)] = \sum_{i=1}^{N} \sum_{k=1}^{K} p_i^k \left[-\frac{d}{2} \log \sigma_k^2 + \log \pi_k - \frac{||x_i - \mu_k||^2}{2\sigma_k^2} \right] + cste$
en posant $p_i^k = P(Z = z_i^k|X = x_i)$, que l'on calcule comme suit :

$$\begin{aligned}
P(z^k|x) &= C_x P(x|z^k) P(z^k) \\
&= C_x \frac{1}{(2\pi\sigma_k^2)^{\frac{d}{2}}} \exp\left[-\frac{||x_i - \mu_k||^2}{2\sigma^2} \right] \pi_k
\end{aligned}$$

où C_x est une constante qui ne dépend que de x, que l'on connait implicitement en écrivant $\sum_k P(z^k|x) = 1$.

— Ensuite on cherche Θ maximisant $E_{Z \sim P(.|X,\Theta)}[l(Z, \Theta)]$ sans oublier la contrainte $\sum_k \pi_k = 1$.
On peut décomposer la fonction objectif comme suit :
$E_{Z \sim P(.|X,\Theta)}[l(Z, \Theta)] = l_1(\pi) + l_2(\mu, \sigma) + cste$
où :
$l_1(\pi) = \sum_{i=1}^{N} \sum_{k=1}^{K} p_i^k \log \pi_k$
$l_2(\mu, \sigma) = -\frac{d}{2} \sum_k \log \sigma_k^2 \sum_{i=1}^{N} p_i^k - \sum_{i=1}^{N} \sum_{k=1}^{K} p_i^k \frac{||x_i - \mu_k||^2}{2\sigma_k^2}$

l_1 est concave (somme de fonctions concaves). Par conséquent, si Lag est le lagrangien associé à la minimisation de l_1 sous la contrainte affine $\sum_k \pi_k = 1$, un point selle de Lag donne le nouvel estimateur π^* (dualité forte).
$Lag(\pi) = -\sum_{i=1}^{N} \sum_{k=1}^{K} p_i^k \log \pi_k + \lambda(\sum_k \pi_k - 1)$:
$\frac{\partial Lag}{\partial \pi_k}(\pi) = -\frac{1}{\pi_k} \sum_{i=1}^{N} p_i^k + \lambda$
D'où l'on obtient $-\sum_{i=1}^{N} p_i^k + \lambda \pi_k^* = 0$.
En sommant sur k, il vient : $\lambda = \sum_{i=1}^{N} \sum_{k=1}^{K} p_i^k = N$. Soit : $\pi_k^* = \frac{\sum_{i=1}^{N} p_i^k}{N}$

Par ailleurs, l_2 est également concave, la distribution gaussienne appartenant à la famille exponentielle.

$$\frac{\partial l_2}{\partial \mu_k}(\Theta) = \frac{1}{\sigma^2}\sum_{i=1}^{N} p_i^k(x_i - \mu_k) = \frac{1}{\sigma^2}\left[-\sum_{i=1}^{N} p_i^k x_i + \mu_k \sum_{i=1}^{N} p_i^k\right]$$
$$\frac{\partial l_2}{\partial \sigma_k^2}(\Theta) = -\frac{d}{2}\frac{1}{\sigma_k^2}\sum_{i=1}^{N} p_i^k + \frac{1}{2\sigma^4}\sum_{i=1}^{N}\sum_{k=1}^{K} p_i^k ||x_i - \mu_k||^2$$

On trouve comme point d'annulation de gradient de l_2, à savoir comme nouveaux estimateurs du maximum de vraisemblance μ^* et σ^{2^*} :

$$\mu_k^* = \frac{\sum_{i=1}^{N} p_i^k x_i}{\sum_{i=1}^{N} p_i^k}$$
$$\sigma_k^{2*} = \frac{1}{d}\frac{\sum_{i=1}^{N} p_i^k ||x_i - \mu_k^*||^2}{\sum_{i=1}^{N} p_i^k}$$

- Enfin, le critère qui a été utilisé pour terminaison de l'algorithme EM est un seuil (10^{-2}) sur la différence entre deux estimations consécutives de la log-vraisemblance réelle : $\sum_{i=1}^{N} \log \sum_{k=1}^{N} \pi_k d(x_i, \mu_k, \Sigma_k)$ où $d(., \mu, \Sigma)$ est la densité de la distribution gaussienne de moyenne μ et de matrice de covariance Σ.

Une fois achevée la convergence de l'algorithme EM, chaque observation x_i est assignée au cluster k, où k maximise la probabilité $p(Z^k = 1|X = x_i, \Theta)$.

Par ailleurs, afin de dessiner les frontières des zones contenant 90% de la masse des différentes distributions gaussiennes, nous avons écrit :

$$\begin{aligned}
&P(U \leq x) = 0.9, \;\; U \sim G(\mu_k, \sigma_k^2 I) \\
\Leftrightarrow\; &P(U \leq x - \mu_k) = 0.9, \;\; U \sim G(0, \sigma_k^2 I) \\
\Leftrightarrow\; &P(\sum_{l=0}^{d} U_l^2 \leq ||x - \mu_k||^2) = 0.9, \;\; U_l \sim G(0, \sigma_k^2), \;\; U_l \text{ indépendantes} \\
\Leftrightarrow\; &P(\sigma_k^2 V \leq ||x - \mu_k|^2) = 0.9, \;\; V \sim \chi^2(d) \\
\Leftrightarrow\; &||x - \mu_k||^2 = \sigma_k^2 F^{-1}_{\chi^2(d)}(0.9)
\end{aligned}$$

Ainsi, la zone contenant 90% de la masse de $G(\mu_k, \sigma_k^2 I)$ est le disque de centre μ_k et de rayon $\sqrt{\sigma_k^2 \; F^{-1}_{\chi^2(d)}(0.9)}$ avec $F_{\chi^2(d)}$ la fonction de répartition de la loi du $\chi^2(d)$.

4.3 Exercice 3 : (c)

De meme, dans le cas non-isotropique, afin de dessiner les frontières des zones contenant 90% de la masse des différentes distributions gaussiennes, nous avons écrit :

$$\begin{aligned}
&P(U \leq x) = 0.9, \;\; U \sim G(\mu_k, \Sigma_k^2) \\
\Leftrightarrow\; &P(U \leq x - \mu_k) = 0.9, \;\; U \sim G(0, \Sigma_k^2) \\
\Leftrightarrow\; &P(\sum_{l=0}^{d} U_l^2 \leq (x - \mu_k)^T {\Sigma_k^2}^{-1} (x - \mu_k)) = 0.9, \;\; U_l \sim G(0, 1), \;\; U_l \text{ indépendantes} \\
\Leftrightarrow\; &P(V \leq (x - \mu_k)^T {\Sigma_k^2}^{-1} (x - \mu_k)) = 0.9, \;\; V \sim \chi^2(d) \\
\Leftrightarrow\; &(x - \mu_k)^T {\Sigma_k^2}^{-1} (x - \mu_k) = F^{-1}_{\chi^2(d)}(0.9) \\
\Leftrightarrow\; &(x - \mu_k)^T A^{-1} (x - \mu_k) = 1, \;\; A = F^{-1}_{\chi^2(d)}(0.9) \; \Sigma_k^2
\end{aligned}$$

Ainsi, la zone contenant 90% de la masse de $G(\mu_k, \Sigma_k^2)$ est l'ellipse de centre μ_k, de demi-axes les deux valeurs propres de A, et d'orientation l'arc-tangente du rapport y/x, où (x, y) est un vecteur propre de A associé à sa plus grande valeur propre.