Compte-rendu du devoir 2

Vincent Matthys

1 Indépendance conditionnelle et factorisations

1.1

$$\begin{aligned} X \perp\!\!\!\perp Y \mid Z &\Leftrightarrow p(x,y \mid z) = p(x \mid z)p(y \mid z) \qquad \forall x,y,z \quad \text{t.q.} \quad p(z) > 0 \\ &\Leftrightarrow p(x,y,z) = p(x \mid z)p(y \mid z)p(z) \qquad \forall x,y,z \quad \text{t.q.} \quad p(z) > 0 \\ &\Leftrightarrow \frac{p(x,y,z)}{p(y,z)} = p(x \mid z)\frac{p(y \mid z)p(z)}{p(y,z)} \qquad \forall x,y,z \quad \text{t.q.} \quad p(y,z) > 0 \\ &\Leftrightarrow p(x \mid y,z) = p(x \mid z) \qquad \forall x,y,z \quad \text{t.q.} \quad p(y,z) > 0 \end{aligned} \tag{1}$$

1.2

Etant donné le modèle graphique orienté G :

$$p \in \mathcal{L}(G) \Leftrightarrow \forall x,y,z,t \quad p(x,y,z,t) = p(x)p(y)p(z \mid x,y)p(t \mid z) \tag{2}$$

1.3

Non traité.

2 Distributions factorisant sur un graphe

2.1

Non traité.

2.2

Non traité.

3 Entropie et information mutuelle

3.1 Entropie

3.1.a)

Avec les conventions définies :

$$p_X(x) = \mathbb{P}(X = x) < 1 \Rightarrow p_X(x)\log(p_X(x)) < 0 \Rightarrow -\sum_{x\in\mathcal{X}} p_X(x)\log(p_X(x)) = H(X) \geq 0$$
$$H(X) = 0 \Rightarrow \forall x \in \mathcal{X},\, p_X(x)\log(p_X(x)) = 0 \Rightarrow \forall x \in \mathcal{X},\, p_X(x) = 1 \Rightarrow \text{X is constant with probability 1} \tag{3}$$

3.1.b)

Par définition de la Kullback-Leibler divergence il vient :

$$\begin{aligned} D(p_X \parallel q) &= \sum_{x\in\mathcal{X}} p_X(x)\log\frac{p_X(x)}{q(x)} \\ &= -\sum_{x\in\mathcal{X}} p_X(x)\log q(x) - H(X) \\ &= -\sum_{x\in\mathcal{X}} p_X(x)\log\frac{1}{|X|} - H(X) \quad \text{puisque} \quad \forall x \in \mathcal{X},\, q(x) = \frac{1}{k} = \frac{1}{k} \\ &= \log k \sum_{x\in\mathcal{X}} p_X(x) - H(X) \\ &= \log k - H(X) \quad \text{car} \quad \sum_{x\in\mathcal{X}} p_X(x) = 1 \end{aligned} \tag{4}$$

3.1.c)

Avec les équations (3) et (4), on a directement

$$\log k - H(X) = D(p_X \parallel q) \leq \log k \tag{5}$$

3.2 Information mutuelle

3.2.a)

Par définition de l'information mutuelle :

$$\begin{aligned} I(X_1, X_2) &= \sum_{(x_1,x_2)\in\mathcal{X}_1\times\mathcal{X}_2} p_{1,2}(x_1,x_2)\log\frac{p_{1,2}(x_1,x_2)}{p_1(x_1)\,p_2(x_2)} \\ I(X_1, X_2) &= D(p_{1,2} \parallel p_1p_2) \quad \text{par définition de la Kullback-Leibler divergence} \end{aligned} \tag{6}$$

Or la Kullback-Leibler divergence est positive pour toute paire $(p_{1,2}, p_1p_2)$ de distributions, donc $I(X_1, X_2) \geq 0$.

3.2.b)

Toujours avec la défintion de l'information mutuelle :

$$
\begin{aligned}
I(X_1, X_2) &= \sum_{(x_1,x_2)\in\mathcal{X}_1\times\mathcal{X}_2} p_{1,2}(x_1, x_2) \log \frac{p_{1,2}(x_1, x_2)}{p_1(x_1)\, p_2(x_2)} \\
I(X_1, X_2) &= \sum_{(x_1,x_2)\in\mathcal{X}_1\times\mathcal{X}_2} p_{1,2}(x_1, x_2) \log p_{1,2}(x_1, x_2) \\
&\quad - \sum_{(x_1,x_2)\in\mathcal{X}_1\times\mathcal{X}_2} p_{1,2}(x_1, x_2) \log \left(p_1(x_1)p_2(x_2)\right) \\
I(X_1, X_2) &= -H(X_1, X_2) \\
&\quad - \sum_{x_1\in\mathcal{X}_1} \left(\sum_{x_2\in\mathcal{X}_2} p_{1,2}(x_1, x_2) \right) \log \left(p_1(x_1)\right) \\
&\quad - \sum_{x_2\in\mathcal{X}_2} \left(\sum_{x_1\in\mathcal{X}_1} p_{1,2}(x_1, x_2) \right) \log \left(p_2(x_2)\right) \\
I(X_1, X_2) &= -H(X_1, X_2) \\
&\quad - \sum_{x_1\in\mathcal{X}_1} p_1(x_1) \log \left(p_1(x_1)\right) \\
&\quad - \sum_{x_2\in\mathcal{X}_2} p_2(x_2) \log \left(p_2(x_2)\right) \\
I(X_1, X_2) &= H(X_1) + H(X_2) - H(X_1, X_2)
\end{aligned}
\tag{7}
$$

Et ainsi l'information mutuelle peut s'écrire uniquement à partir des entropies de X_1, X_2 et de (X_1, X_2).

3.2.c)

D'après l'éqiation (7), on peut réecrire :

$$
\begin{aligned}
&H(X_1, X_2) - (H(X_1) + H(X_2)) = -I(X_1, X_2) \\
&H(X_1, X_2) - (H(X_1) + H(X_2)) \leq 0 \quad \text{puisque, } I(X_1, X_2) \geq 0
\end{aligned}
\tag{8}
$$

D'après l'équation (8), pour p_1 et p_2 données, l'entropie maximale correspond à $I(X_1, X_2) = 0$, c'est-à-dire $D(p_{1,2} \parallel p_1 p_2) = 0$ d'après l'équation (4). Or la Kullback-Leibler divergence s'anulle si et seulement si les distributions sont identiques. On a alors $p_{1,2} = p_1 p_2$, ce qui correspond à $X_1 \perp\!\!\!\perp X_2$.

4 Gaussian Mixtures

4.1 K-means

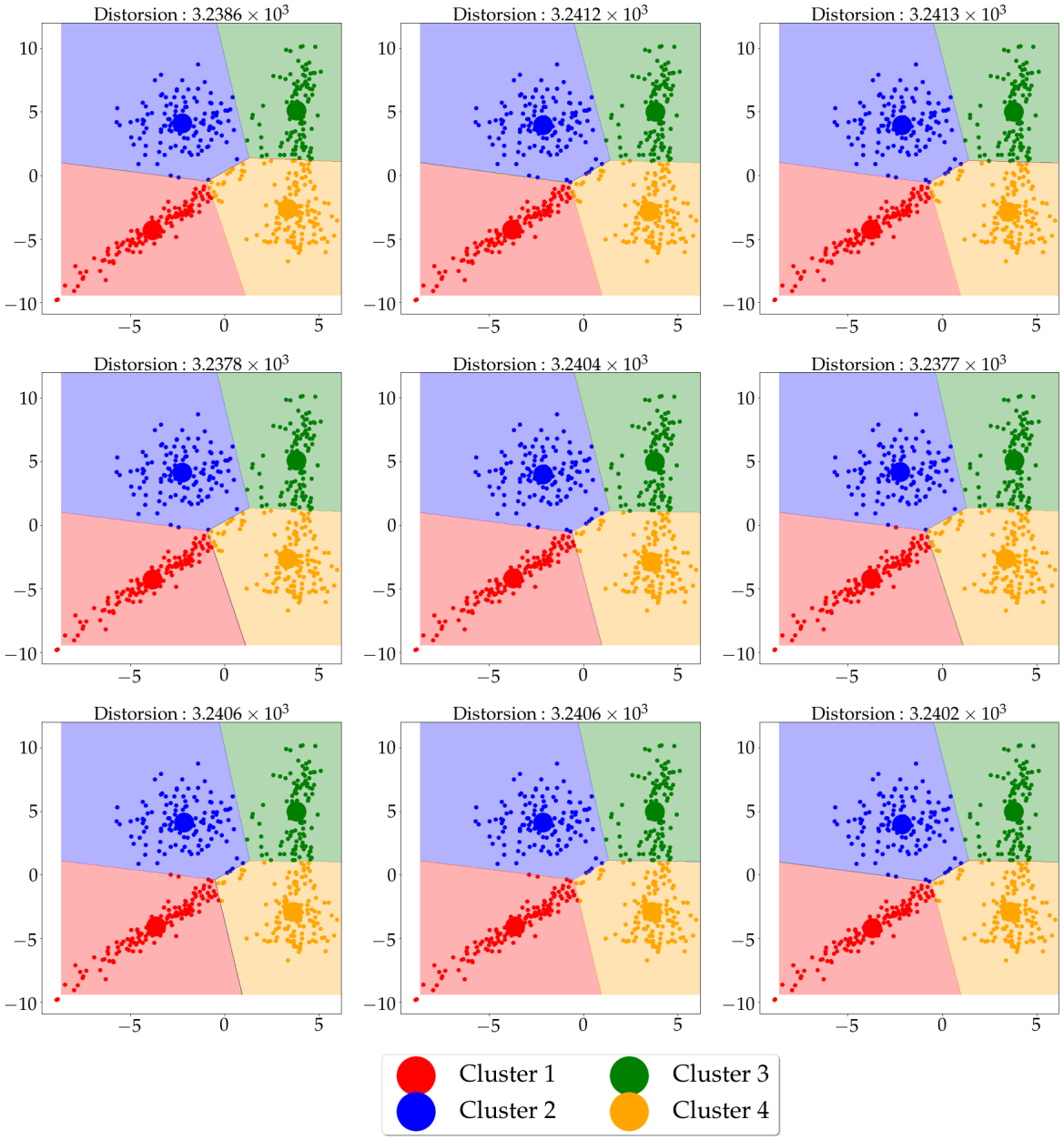

FIGURE 1 – Visualisation du clustering obtenu par k-means des données d'entrainement sur 9 initialisations aléatoires des centroïdes

En figure 1 sont présentés les résultats du clustering des données d'entraînement *EMGaussian.data* avec 9 initialisations aléatoires distinctes des centroïdes. La condition d'arrêt est remplie quand la variation de distorsion est inférieure à 1. On a utilisé l'angle des centroïdes finaux afin d'artificiellement repérer par les mêmes couleurs et mêmes chiffres les clusters situés dans des zones similaires du plan. On constate que la distorsion de ces 9 initialisations différentes ne diffère que par le 3^{e} chiffre significatif, avec une moyenne de 3.239 10^3 et avec un écart-type 1.335. Les minima locaux de cette distorsion ont donc des valeurs très semblables sur ces itérations, et les positions des centroïdes finaux sont regroupées dans la table 1. On

Centroïde	$\mu_x \pm \sigma_x$	$\mu_y \pm \sigma_y$
1	-3.747 ± 0.068	-4.193 ± 0.081
2	-2.177 ± 0.044	4.060 ± 0.078
3	3.794 ± 0.007	5.038 ± 0.048
4	3.483 ± 0.107	-2.795 ± 0.105

TABLE 1 – Positions des centroïdes sur les 9 initialisations aléatoires

peut constater que les variations autour des positions moyennes sont de l'ordre du dixième dans le pire des cas, correspondant à une variation en distance de l'ordre de 0.3 %. Néanmoins on constate que les points à la frontière des clusters ont tendance à basculer, tantôt vers l'un ou l'autre des clusters, notamment à la frontière des clusters jaune bleu et rouge, où les points proches de $(0, 0)$ sont facilement attribué par l'humain au cluster 1, rouge, qui présente une forme très allongée, différente des autres formes de clusters. Or la clusterisation par k-means ne permet pas de prendre en compte cette spécificité de forme. En définitif, les clusters trouvés par k-means sont, pour ce jeu de données d'entraînement, très stables pour la majorité des initialisations aléatoires et permette de séparer en première approximation, indépendamment de la forme du cluster, les points suivant un consensus proche de celui de l'homme, à l'exception de quelques points centraux.

4.2 Mixture de gaussiennes isotropiques

Dans le cas d'une mixture de gaussiennes isotropiques, on peut réecrire l'étape de maximisation suivant :

$$\begin{aligned}
\boldsymbol{\theta}^{(t+1)} &= \arg\max_{\boldsymbol{\theta}} \left(\mathbb{E}_{\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta}^{(t)}}\right) \\
\boldsymbol{\theta}^{(t+1)} &= \arg\max_{\boldsymbol{\theta}} \left(\sum_{n=1}^{N} \sum_{k=1}^{4} q_{nk}^{(t)} \left(\log \pi_k - \frac{1}{2} \log |\Sigma_k| - \frac{1}{2} \left(x_i - \mu_k\right)^\intercal \Sigma_k^{-1} \left(x_i - \mu_k\right) \right) \right)
\end{aligned} \tag{9}$$

où $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ et $q_{nk}^{(t)} = \mathbb{P}\left(z_n = k | x_n, \boldsymbol{\theta}^{(t)}\right)$ avec z_n variable latente associée à x_n.
On peut alors séparer la maximisation suivant $\boldsymbol{\pi}$ et suivant $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$), de sorte que :

$$\begin{aligned}
\boldsymbol{\pi}^{(t+1)} &= \arg\max_{\boldsymbol{\pi}} \left(\sum_{n=1}^{N} \sum_{k=1}^{4} q_{nk}^{(t)} \log \pi_k \right) \quad \text{avec} \quad \forall n, \sum_{k=1}^{4} q_{nk} = 1 \\
(\boldsymbol{\mu}^{(t+1)}, \boldsymbol{\Sigma}^{(t+1)}) &= \arg\max_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \left(\sum_{n=1}^{N} \sum_{k=1}^{4} q_{nk} \left(-\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} \left(x_i - \mu_k\right)^\intercal \Sigma_k^{-1} \left(x_i - \mu_k\right) \right) \right)
\end{aligned} \tag{10}$$

Les deux problèmes de minimisation en (10) sont différentiables et concaves (le deuxième étant analogue au calcul du MLE pour une distribution normale) et on a :

$$\begin{aligned}
\frac{\partial \left(\sum_{n=1}^{N} \sum_{k=1}^{4} q_{nk}^{(t)} \log \pi_k^{(t+1)} \right)}{\partial \pi_k^{(t+1)}} &= \sum_{n=1}^{N} q_{nk}^{(t)} - \lambda \pi_k^{(t+1)} = 0 \quad \text{avec} \quad \lambda \in \mathbb{R} \\
\pi_k^{(t+1)} &\propto \sum_{n=1}^{N} q_{nk}^{(t)} \\
\pi_k^{(t+1)} &= \frac{\sum_{n=1}^{N} q_{nk}^{(t)}}{\sum_{k=1}^{4} \sum_{n=1}^{N} q_{nk}^{(t)}} = \frac{1}{N} \sum_{n=1}^{N} q_{nk}^{(t)}
\end{aligned} \tag{11}$$

$$\begin{aligned}\frac{\partial\left(\sum_{n=1}^{N}\sum_{k=1}^{4} q_{nk}^{(t)}\left(-\frac{1}{2}\left(x_i-\mu_k\right)^{\intercal}\Sigma_k^{-1}\left(x_i-\mu_k\right)\right)\right)}{\partial\mu_k} &= \sum_{n=1}^{N} q_{nk}^{(t)}\left(\Sigma_k^{-1}(x_i-\mu_k^{(t+1)})\right)=0\\ \mu_k^{(t+1)} &= \frac{\sum_{n=1}^{N} q_{nk}^{(t)} x_i}{\sum_{n=1}^{N} q_{nk}^{(t)}}\end{aligned} \tag{12}$$

Avec l'hypothèse $\boldsymbol{\Sigma_k} = \sigma_k^2 \boldsymbol{I}$ d'isotropie, il vient :

$$\begin{aligned}\frac{\partial\left(\sum_{n=1}^{N}\sum_{k=1}^{4} q_{nk}^{(t)}\left(-\frac{1}{2}\log\sigma_k^4-\|x_i-\mu_k\|^2/2\sigma_k^2\right)\right)}{\partial\sigma_k^2} &= \sum_{n=1}^{N} q_{nk}\left(-\frac{1}{\sigma_k^2}+\frac{1}{2\sigma_k^4}\left(x_i-\mu_k\right)^{\intercal}\left(x_i-\mu_k\right)\right)\\ &= 0\\ \sigma_k^2 &= \frac{\sum_{n=1}^{N} q_{nk}\left(x_i-\mu_k\right)^{\intercal}\left(x_i-\mu_k\right)}{2\sum_{n=1}^{N} q_{nk}}\end{aligned} \tag{13}$$

En figure 2 sont présentés les résultats de la mixture de gaussiennes isotropiques sur les mêmes données d'entraînement qu'en figure 1. La condition d'arrêt est atteinte quand la variation de la vraisemblance est inférieure à 10^{-3}. La même convention d'angle a été utilisé pour identifier les clusters délimités dans les zones semblables. La table regroupant les positions des centres est également présentée en table 2. On constate que la différence des positions des centres de masse par rapport à la position des centres de K-means est de l'ordre de grandeur de l'écart-type sur toutes les initialisations, de sorte qu'on peut conclure que la modélisation n'apporte pas de manière significative de changement dans les posistions des centres. En revanche, on constate que plusieurs extrema locaux de vraisemblance sont trouvés. En particulier, il peut arriver que tous les points proches de $(0, 0)$ soit plutôt attribués au cluster 4 qu'au 3 autres (du moins, ces points peuvent avoir une variable lattente plus grande pour le cluster 4). On peut également observer des étalement de masse des gaussiennes différentes suivant les initialisations. Enfin, la vraisemblance moyenne pour cette mixture de gaussiennes isotropiques est de $(-2.642 \pm 0.003)\ 10^3$.

Centroïde	$\mu_x \pm \sigma_x$	$\mu_y \pm \sigma_y$	$\|\Delta\mu$(kmeans-iso)$\|$
1	-3.991 ± 0.400	-4.417 ± 0.411	0.256
2	-2.441 ± 0.166	4.278 ± 0.038	0.342
3	3.252 ± 0.630	5.040 ± 1.437	0.542
4	3.229 ± 0.656	-2.712 ± 1.101	0.267

TABLE 2 – Positions des centres de masse de chaque gaussienne représentée en figure 2, et norme de la différence des positions moyennes des centres pour k-means et des centres de masse des gaussiennes isotropiques.

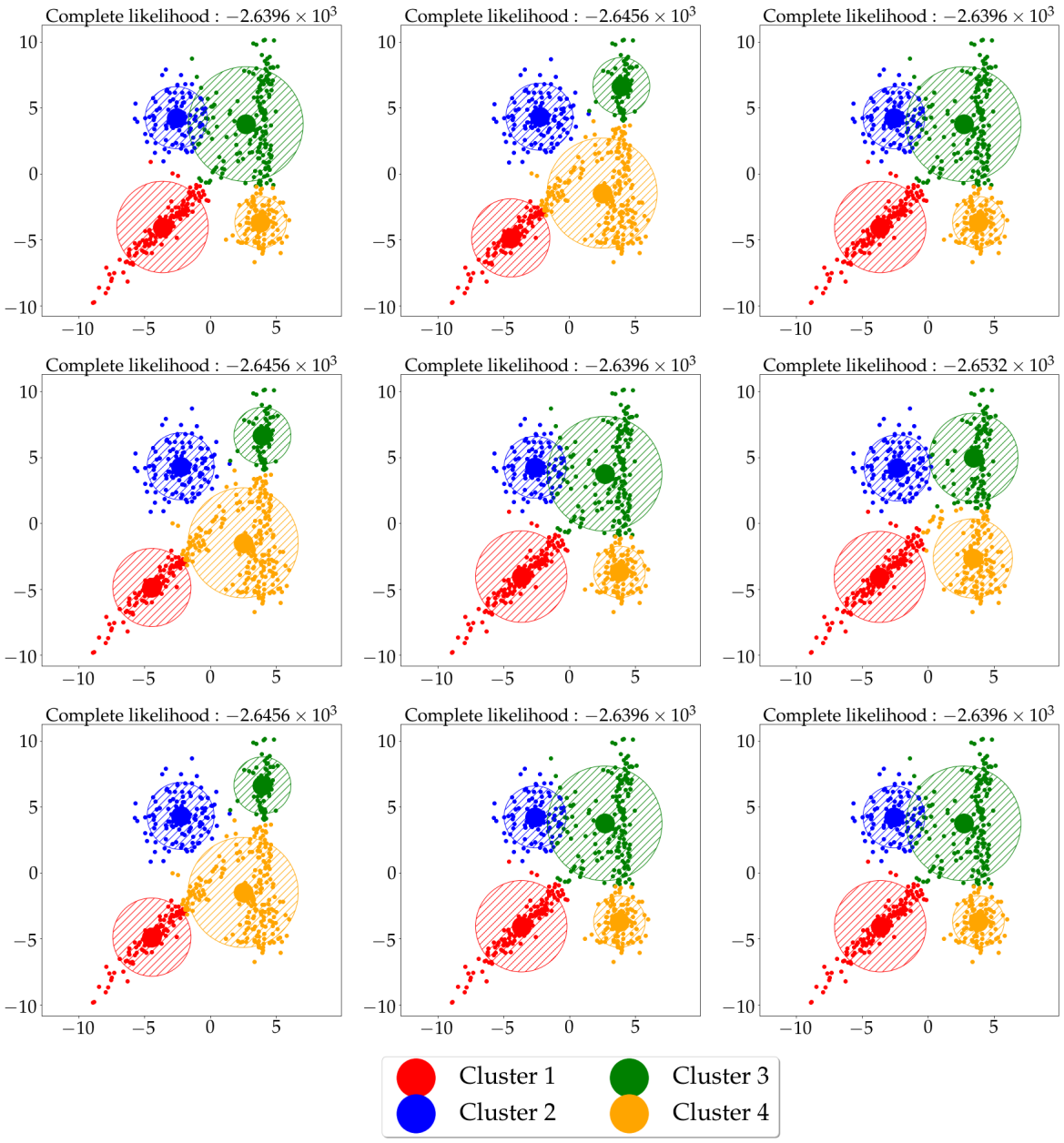

FIGURE 2 – Visualisation du clustering obtenu par mixtures de gaussiennes avec des covariances isotropiques avec 9 initialisations par K-means des données d'entrainement. Chaque couleur représente un noeud gaussien. Le maxima des variables latentes de chaque point a été représenté, ainsi que la zone contenant 90 % de la masse de chaque gaussienne, en hachuré.

4.3 Mixture de gaussiennes générales

Sans l'hypothése d'isotropie, on a alors :

$$\begin{aligned} \frac{\partial \left(\sum_{n=1}^{N} \sum_{k=1}^{4} q_{nk}^{(t)} \left(-\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x_i - \mu_k)^\intercal \Sigma_k^{-1} (x_i - \mu_k) \right) \right)}{\partial \sigma_k^2} &= 0 \\ \Sigma_k &= \frac{\sum_{n=1}^{N} q_{nk} (x_i - \mu_k) (x_i - \mu_k)^\intercal}{\sum_{n=1}^{N} q_{nk}} \end{aligned} \tag{14}$$

En figure 3 sont présentés les résultats de la mixture de gaussiennes avec covariance générale sur les mêmes données d'entraînement qu'en figure 1. La condition d'arrêt est atteinte quand la variation de la vraisemblance est inférieure à 10^{-3}. La même convention d'angle a été utilisé pour identifier les clusters délimités dans les zones semblables. La table regroupant les positions des centres est également présentée en table 3. On constate que les différences des positions des centres de masse entre le modèle isotropique et le modèle général sont largement supérieures, jusqu'à un facteur 4, comparées aux différences entre le passage du K-means à la mixture de gaussienne isotropique. La vraisemblance moyenne pour cette mixture de gaussiennes générales est de $(-2.327 \pm 0.000)\ 10^3$ et est inférieure à la vraisemblance obtenue pour la mixture de gaussiennes isotropiques. On constate, d'autre part, une très grande stabilité des positions des centres de masses, corroborée par la stabilité de la vraisemblance, pouvant mettre en évidence qu'un unique maximum a été atteint (qui n'est peut-être pas global). En l'occurence, sur ces données d'entraînement, les points en $(0, 0)$ sont systématiquement attribués au cluster 1, du moins, en prenant le maximum des variables latentes comme attribution. Ce modèle permet de rendre compte des formes diverses des noeuds gaussiens, dont les distributions en masse sont tantôt allongés, tantôt regrouopées.

Centroïde	$\mu_x \pm \sigma_x$	$\mu_y \pm \sigma_y$	$\|\Delta\mu(\text{iso-gen})\|$
1	-3.060 ± 0.000	-3.532 ± 0.000	1.100
2	-2.034 ± 0.000	4.173 ± 0.000	0.420
3	3.981 ± 0.000	3.838 ± 0.005	1.406
4	3.800 ± 0.000	-3.775 ± 0.000	1.207

TABLE 3 – Positions des centres de masse de chaque gaussienne représentée en figure 3, et norme de la différence des positions moyennes des centres de masses pour la mixture isotropique et générale.

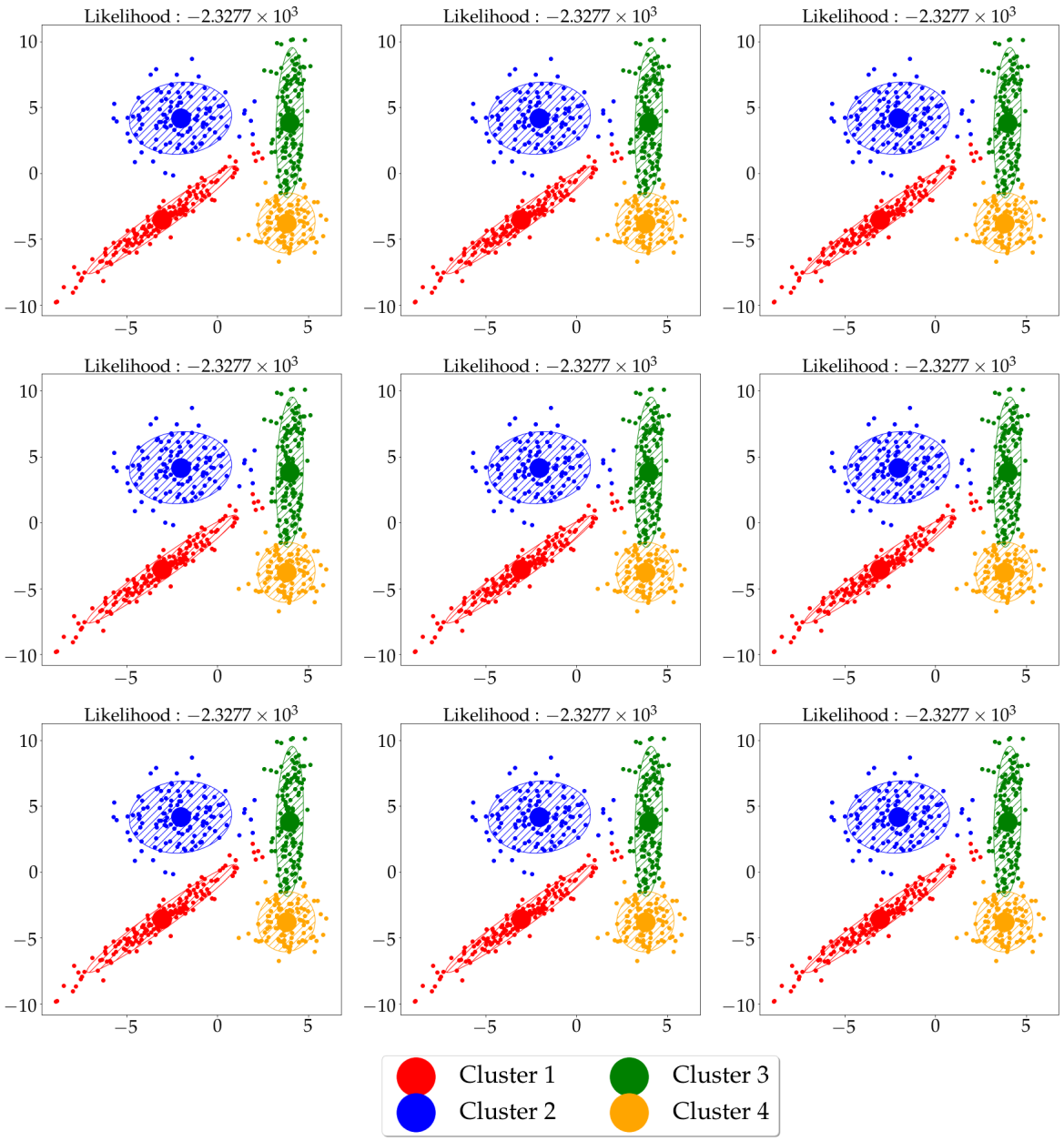

FIGURE 3 – Visualisation du clustering obtenu par mixtures de gaussiennes avec covariance générale avec 9 initialisations par K-means des données d'entrainement. Chaque couleur représente un noeud gaussien. Le maxima des variables latentes de chaque point a été représenté, ainsi que la zone contenant 90 % de la masse de chaque gaussienne, en hachuré.

4.4 Comparaison des résultats

En figures 4, 5 sont présentés les résultats des deux mixtures de gaussiennes sur les données de test. On constate encore la grande stabilité en vraisemblance et en position des centre de masse, pour la mixture de gaussiennes générales, tandis qu'on observe toujours une instabilité en position et en forme des distribution pour la mixture de gaussiennes isotropiques. Pour cette dernière, la vraisemblance moyenne est de $(-2.593 \pm 0.037)\ 10^3$, inférieure à la vraisemblance de la mixture de gaussiennes générales, de $(-2.360 \pm 0.002)\ 10^3$, rendant bien compte de la modélisation plus adaptée du modèle général, avec des masses très allongées pour les clusters 1 et 3, qui ne sont pas bien rendues par des gaussiennes isotropiques.

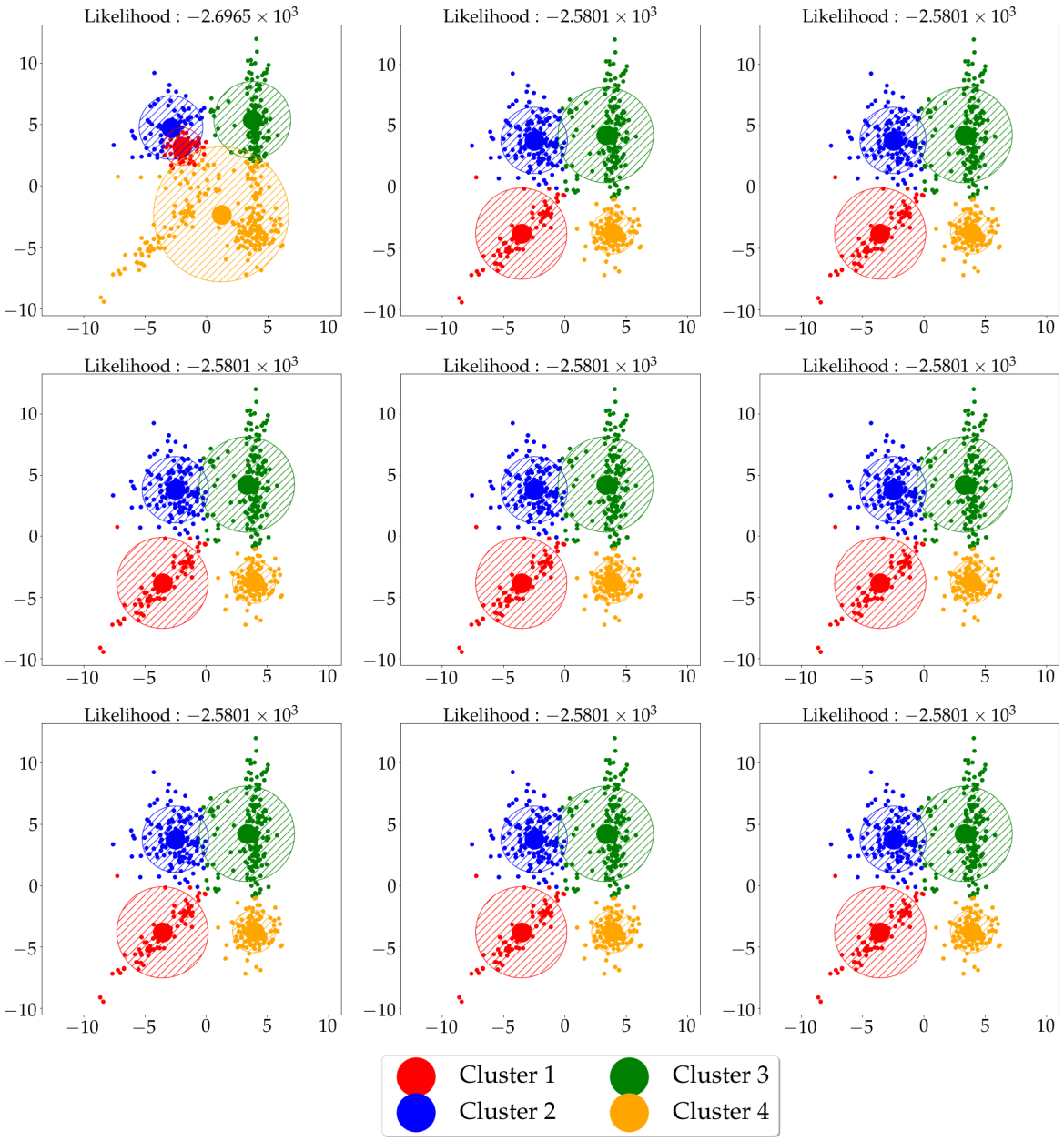

FIGURE 4 – Visualisation du clustering obtenu par mixtures de gaussiennes avec des covariances isotropiques avec 9 initialisations par K-means des données de test. Chaque couleur représente un noeud gaussien. Le maxima des variables latentes de chaque point a été représenté, ainsi que la zone contenant 90 % de la masse de chaque gaussienne, en hachuré.

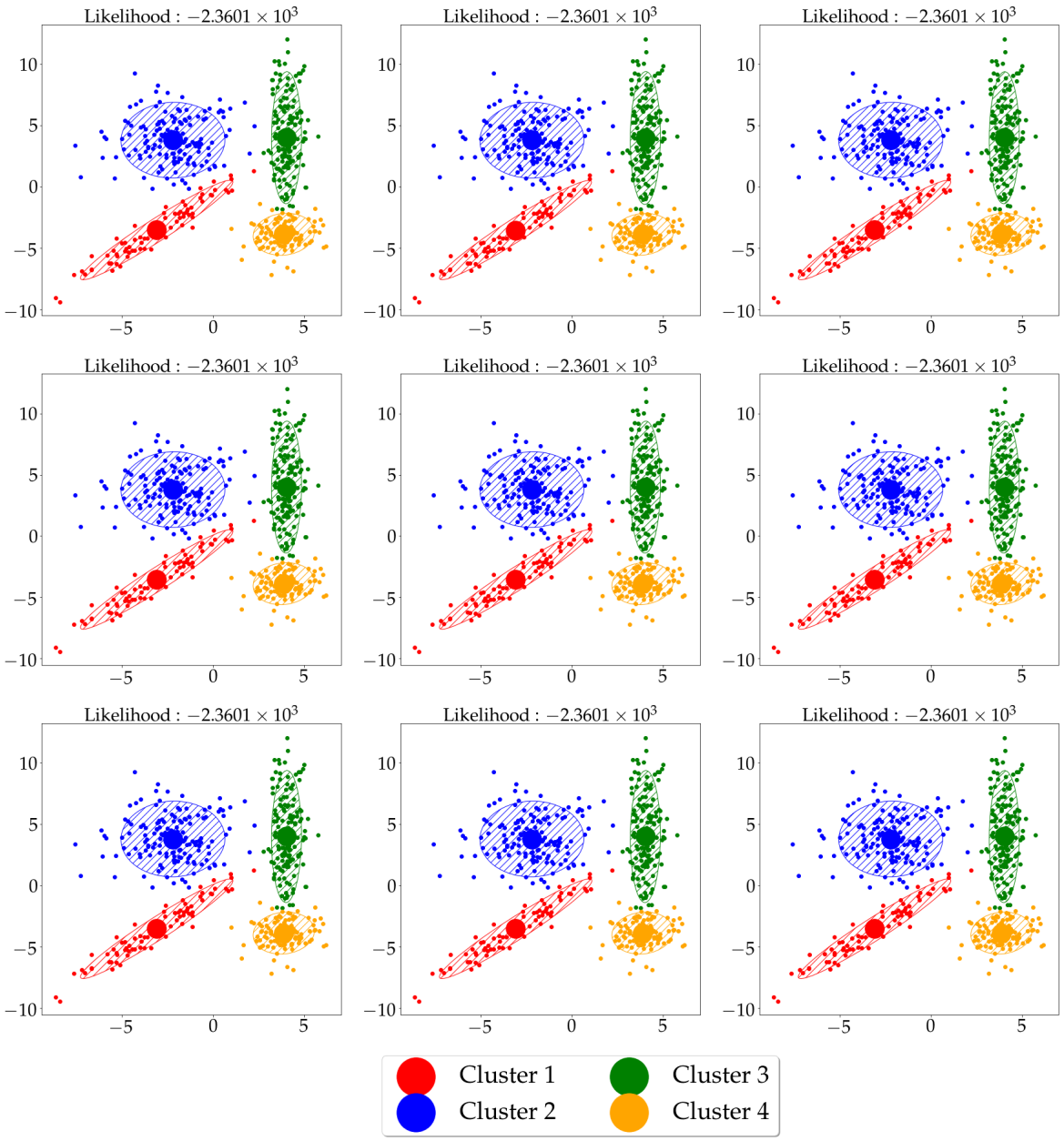

FIGURE 5 – Visualisation du clustering obtenu par mixtures de gaussiennes avec covariance générale avec 9 initialisations par K-means des données de test. Chaque couleur représente un noeud gaussien. Le maxima des variables latentes de chaque point a été représenté, ainsi que la zone contenant 90 % de la masse de chaque gaussienne, en hachuré.