TP 2 - Théorie des matrices aléatoires

António Loison, Florian Le Bronnec

March 20, 2022

1 Observations préliminaires

1. Si on suppose les q_i fixés, alors nous avons :

$$\mathbb{E}\left[\mathbf{A}_{ij}\right] = q_i q_j \mathbf{C}_{ab}, \tag{1}$$

Nous pouvons alors réinterpréter cela en introduisant la matrices $\mathbf{J}$ des vecteurs canoniques des classes $\mathcal{C}_i$:

$$\mathbb{E}\left[\mathbf{A}_{ij}\right] = q_i q_j \left[\mathbf{J}\mathbf{C}\mathbf{J}^*\right]_{ij}, \tag{2}$$

donc l'espérance de la matrice $\mathbf{A}$ peut s'écrire :

$$\mathbb{E}\left[\mathbf{A}\right] = \mathbf{diag}(q_i)\mathbf{J}\mathbf{C}\mathbf{J}^*\mathbf{diag}(q_i). \tag{3}$$

Et l'on a directement que $\mathrm{rg}\left(\mathbb{E}\left[\mathbf{A}\right]\right) \leq K$.

Nous pouvons maintenant décomposer $\dfrac{1}{\sqrt{n}}\mathbf{A}$:

$$\frac{1}{\sqrt{n}}\mathbf{A} = \frac{1}{\sqrt{n}}\left[\begin{array}{ccc} \underbrace{\mathbf{A} - \mathbb{E}\left[\mathbf{A}\right]}_{\substack{\text{à entrées centrées} \\ \text{et indépendantes.}}} & + & \underbrace{\mathbb{E}\left[\mathbf{A}\right]}_{\mathrm{rg} \leq K.} \end{array}\right]. \tag{4}$$

Les variances des entrées de $\mathbf{A}$ sont celles de variables aléatoires suivant une loi de Bernoulli :

$$\mathbb{V}\left[\frac{1}{\sqrt{n}}\mathbf{A}_{ij}\right] = \frac{1}{n} q_i q_j \mathbf{C}_{ab}(1 - q_i q_j \mathbf{C}_{ab}), \tag{5}$$

$$= \frac{1}{n} q_i q_j (1 - q_i q_j) + \mathcal{O}\left(\frac{1}{n\sqrt{n}}\right). \tag{6}$$

2. Pour la matrice, $\mathbf{B}$, en remarquant que $\mathbf{J}\mathbf{1}_K\mathbf{1}_K^*\mathbf{J}^* = \mathbf{J}\mathbf{J}^* = \mathbf{1}_n\mathbf{1}_n^*$, il vient directement que:

$$\mathbb{E}\left[\mathbf{B}\right] = \mathbf{diag}(q_i)\mathbf{J}\frac{1}{\sqrt{n}}\mathbf{M}\mathbf{J}^*\mathbf{diag}(q_i). \tag{7}$$

Et donc, comme pour $\mathbf{A}$, nous avons la décomposition suivante:

$$\frac{1}{\sqrt{n}}\mathbf{B} = \frac{1}{\sqrt{n}}\left[\begin{array}{ccc} \underbrace{\mathbf{B} - \mathbb{E}[\mathbf{B}]}_{\text{à entrées centrées et indépendantes.}} & + & \underbrace{\mathbb{E}[\mathbf{B}]}_{\text{rg} \leq K.} \end{array}\right]. \tag{8}$$

Et puisque les entrées de $\mathbf{B}$ correspondent à celles de $\mathbf{A}$ à une constante près, son profil de variance est le même que pour $\mathbf{A}$.

3. Les différentes observations sont présentées pour différentes valeurs de $\mathbf{M}$, pour différents q sur les figs. 1, 3 and 5 et la répartition dans l'espace déterminé par les vecteurs propres dominants sur les figs. 2, 4 and 6.

   Suite à nos observations nous pouvons faire quelques remarques sur l'influence des différents paramètres :

   - Tout d'abord le paramètre ayant le plus d'influence sur la forme de répartition des valeurs propres est le vecteur $(q_i)_{1 \leq i \leq N}$.
   - On observe notamment un cas intéressant lorsque $\forall i \in [1, N]$, $q_i = q_0$. On observe une mesure spectrale empirique qui se rapproche d'une loi du demi-cercle.
   - Dans les autres cas, lorsque les q_i ne sont pas constants, la forme est plus difficile à caractériser et semble s'éloigner davantage de celle d'une matrice de Wigner. Il semble que plus les q_i sont dispersés, plus la loi du demi-cercle semble dégénérée et moins les classes semblent facilement séparables dans l'espace des vecteurs propres dominants.
   - On observe des valeurs propres isolées sous certaines conditions. En effet, la décomposition de $\mathbf{B}$ (8) fait clairement penser à un modèle avec perturbations. On peut alors postuler l'existence d'un critère à dépasser pour observer les valeurs propres isolées.
   - On observe au plus K valeurs propres, avec K le nombre de classes.
   - Ce sont les valeurs de $\mathbf{M}$ qui semblent déterminer principalement l'apparition de spikes. Plus celles-ci sont grandes, plus les spikes ont des chances d'apparaître (en fonction de la répartition de chaque classe).
   - En effet, on peut empiriquement se dire que plus l'intéraction entre classes est forte, plus on s'éloigne d'un modèle purement aléatoire où la distribution asymptotique des valeurs propres serait celle d'une matrice de Wigner.
   - On observe également des motifs intéressants dans l'allure des vecteurs propres, qui ressemblent fortement aux vecteurs indicateurs de classes. Comme précédemment, ce schéma semble le plus visible lorsque les q_i ne sont pas dispersés.

4. En ce qui concerne les vecteurs propres, ces derniers semblent être de la même forme que les vecteurs indicatifs des classes. On peut imaginer un algorithme spectrale de détection de communauté basé sur la projection des vecteurs de la matrice $\frac{1}{\sqrt{n}}\mathbf{B}$ sur les directions principales.

   En effet, si ces vecteurs ont des formes proches des vecteurs de classe, on peut s'attendre à ce que ces dernières soient bien séparées selon ces axes.

On observe cependant en pratique que la séparation selon les directions principales est optimale lorsque les q_i sont constants. En effet, dans ce cas seule les informations de classes vont venir perturber la matrice, alors que lorsque les q_i sont différents, chaque noeud pourra perturber de manière plus ou moins forte le processus, réduisant ainsi l'accessibilité aux informations sur les classes.

2 Cas homogène

1. D'après question 2, $\frac{1}{\sqrt{n}}\mathbf{B}$ peut s'écrire sous la forme :

$$\frac{1}{\sqrt{n}}\mathbf{B} = \mathbf{X}_n + \frac{1}{\sqrt{n}}\mathbf{V}_n, \tag{9}$$

avec $\mathbf{X}_n \in \mathbb{R}^{n\times n}$ à entrées indépendantes, de moyenne nulle et de variance convergeant vers $\frac{1}{n}\sigma^2 = \frac{1}{n}q_0^2(1-q_0^2)$.

Nous allons maintenant établir une condition pour avoir des valeurs propres isolées pour $\frac{1}{\sqrt{n}}\mathbf{B}$.

Soit λ une valeur propre asymptotique isolée de $\frac{1}{\sqrt{n}}\mathbf{B}$. λ est donc isolée du spectre asymptotique de $\mathbf{X}_n$, presque sûrement à support dans $[-2\sigma, 2\sigma]$.

$$\begin{aligned} 0 &= \det\left(\mathbf{X}_n + \frac{1}{\sqrt{n}}\mathbf{V}_n - \lambda\mathbf{I}\right), \\ &= \det\left(\mathbf{X}_n + \frac{1}{n}q_0^2\mathbf{J}\mathbf{M}\mathbf{J}^* - \lambda\mathbf{I}_n\right), \\ &= \det\left(\mathbf{X}_n - \lambda\mathbf{I}_n\right)\det\left(\mathbf{I}_n + \frac{1}{n}q_0^2\mathbf{J}\mathbf{M}\mathbf{J}^*\mathbf{Q}(\lambda)\right) \quad \text{où} \quad \mathbf{Q}(z) = \left(\mathbf{X}_n - z\mathbf{I}_n\right)^{-1}, \end{aligned} \tag{10}$$

Puisque λ est une valeur propre asymptotique isolée, à partir d'un certain rang $\det\left(\mathbf{X}_n - \lambda\mathbf{I}_n\right) \neq 0$. Et en utilisant l'identité de Silvester :

$$0 = \det\left(\mathbf{I}_K + \frac{1}{n}q_0^2\mathbf{M}\mathbf{J}^*\mathbf{Q}\mathbf{J}\right). \tag{11}$$

Puis :

$$(\mathbf{J}^*\mathbf{Q}(\lambda)\mathbf{J})_{ab} = j_a^*\mathbf{Q}(\lambda)j_b \xrightarrow[n\to\infty]{\boldsymbol{a.s}} \frac{1}{\sigma}g_{sc}(\lambda/\sigma)j_a^*j_b = \begin{cases} 0 & \text{si} \quad a \neq b, \\ \frac{1}{\sigma}g_{sc}(\lambda/\sigma)n_a & \text{sinon.} \end{cases} \tag{12}$$

Donc nous pouvons exprimer le limite presque sûre du déterminant de (11), en utilisant le fait que $\frac{n_k}{n} \xrightarrow[n\to\infty]{} c_k$:

$$\det\left(\mathbf{I}_K + \frac{1}{n}q_0^2\mathbf{M}\mathbf{J}^*\mathbf{Q}\mathbf{J}\right) \xrightarrow[n\to\infty]{\boldsymbol{a.s}} \prod_{k=1}^{K}\left(1 + \frac{1}{\sigma}g_{sc}(\lambda)c_k m_k q_0^2\right). \tag{13}$$

λ doit donc annnuler le produit (13). La conditions qui déterminent l'existence de valeurs propres asymptotiques isolées est alors :

$$\exists k \in [\![1, K]\!],\ c_k m_k q_0^2 g_{sc}(\lambda/\sigma) = -\sigma \iff g_{sc}(\lambda/\sigma) = -\frac{\sigma}{c_k m_k q_0^2}. \tag{14}$$

Étudions maintenant plus en détails la condition (14).

Soit $k \in [\![1, K]\!]$ correspondant de la condition (14).

Lors du DM précédent nous avons montré que l'expression de $g_{sc}(z)$ comme solution de l'équation $X^2 + zX + 1 = 0$ menait à l'expression suivante :

$$g_{sc}(z) = \frac{-z + \sqrt{z^2 - 4}}{2}. \tag{15}$$

Cette fonction est dérivable, de dérivée $z \longmapsto -\frac{1}{2} + \frac{z}{\sqrt{z^2-4}}$. Son tableau de variation est donc le suivant :

x	$-\infty$		-2		2		$+\infty$
$g_{sc}(x)$	$+\infty$	$\searrow$	1		-1	$\nearrow$	0

Ainsi, (14) a une solution si et seulement si :

$$\begin{cases} c_k m_k q_0^2 > \sigma & \text{si} \quad m_k > 0, \\ \text{ou} & \\ c_k \left|m_k\right| q_0^2 < \sigma & \text{si} \quad m_k < 0. \end{cases} \tag{16}$$

2. Pour les valeurs des solutions, nous allons utiliser (14) avec le résultat rappelé par l'énoncé : $g_{sc}(z) = -\dfrac{1}{z + g_{sc}(z)}$.

   Nous avons au plus K valeur propres isolées, et celles qui respectent la condition (16) ont leur valeur donnée par:

$$\frac{\lambda_k}{\sigma} = \frac{1}{\sigma} c_k m_k q_0^2 + \frac{\sigma}{c_k m_k q_0^2} \iff \lambda_k = c_k m_k q_0^2 + \frac{\sigma^2}{c_k m_k q_0^2}. \tag{17}$$

3. Nous avons :

$$\left(\frac{1}{\sqrt{n}}\mathbf{B} - z\mathbf{I}_n\right)^{-1} = \sum_{i=1}^{n} \frac{u_i u_i^*}{\lambda_i - z}, \tag{18}$$

   avec λ_i les valeurs propres de $\mathbf{B}$ et u_i les vecteurs propres associés. Soit λ_k une valeur propre presque sûrement asymptotiquement isolée, u_k son vecteur propre associé et j_a le vecteur canonique d'une classe.

Soit Γ_k un contour entourant λ_k et laissant les autres à l'extérieur. Grâce à (18), nous avons :

$$\frac{1}{n_a}(j_a^* u_k)^2 = \frac{1}{n_a} j_a^* u_k u_k^* j_a = -\frac{1}{2\pi \boldsymbol{i}} \oint_{\Gamma_k} \frac{1}{n_a} j_a^* \left(\frac{1}{\sqrt{n}} \mathbf{B} - z \mathbf{I}_n \right)^{-1} j_a \, \mathrm{d}z. \tag{19}$$

En utilisant l'identité de Woodbury, nous avons :

$$\begin{aligned} \frac{1}{n_a} j_a^* \left(\frac{1}{\sqrt{n}} \mathbf{B} - z \mathbf{I}_n \right)^{-1} j_a &= \frac{1}{n_a} j_a^* \left(\mathbf{W}_n + \frac{1}{n} q_0^2 \mathbf{J} \mathbf{M} \mathbf{J}^* - z \mathbf{I}_n \right)^{-1} j_a, \\ &= \frac{1}{n_a} j_a^* \mathbf{Q} j_a - \frac{1}{n_a} j_a^* \mathbf{Q} \mathbf{J} \left(\mathbf{I}_K + \frac{1}{n} q_0^2 \mathbf{M} \mathbf{J}^* \mathbf{Q} \mathbf{J} \right)^{-1} \frac{1}{n} q_0^2 \mathbf{M} \mathbf{J}^* \mathbf{Q} j_a. \end{aligned} \tag{20a}$$

(20b)

D'après (12) :

$$\left(\mathbf{I}_K + \frac{1}{n} q_0^2 \mathbf{M} \mathbf{J}^* \mathbf{Q} \mathbf{J} \right)^{-1} \mathbf{M} \xrightarrow[n \to \infty]{\boldsymbol{a.s}} \mathbf{diag} \left(\frac{m_k}{1 + \frac{1}{\sigma} g_{sc}(z/\sigma) c_k q_0^2 m_k} \right) \in \mathbb{R}^{K \times K}, \tag{21a}$$

$$\mathbf{J}^* \mathbf{Q} j_a \xrightarrow[n \to \infty]{\boldsymbol{a.s}} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \frac{1}{\sigma} n_a g_{sc}(z/\sigma) \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^K. \tag{21b}$$

Nous avons donc :

$$\begin{aligned}\frac{1}{n_a} j_a^* \left(\frac{1}{\sqrt{n}} \mathbf{B} - z\mathbf{I}_n \right)^{-1} j_a &= \frac{1}{\sigma} g_{sc}(z/\sigma) - \frac{g_{sc}^2(z/\sigma) c_a m_a q_0^2}{\sigma^2} \times \frac{1}{1 + \frac{1}{\sigma} g_{sc}(z/\sigma) c_a m_a q_0^2}, \\ &= \frac{1}{\sigma} g_{sc}(z/\sigma) - \frac{g_{sc}^2(z/\sigma)}{\sigma} \times \frac{1}{\frac{\sigma}{c_a m_a q_0^2} + g_{sc}(z/\sigma)}, \\ &= \frac{1}{\sigma} g_{sc}(z/\sigma) - \frac{g_{sc}^2(z/\sigma)}{\sigma} \times \frac{1}{g_{sc}(z/\sigma) - g_{sc}(\lambda_a/\sigma)}.\end{aligned} \tag{22a}$$

(22b)

(22c)

Nous pouvons alors utiliser la formule des résidus :

$$\frac{1}{n_a} j_a^* u_k u_k^* j_a = \lim_{z \to \lambda_k} (z - \lambda_k) \left(\frac{g_{sc}^2(z/\sigma)}{\sigma} \times \frac{1}{g_{sc}(z/\sigma) - g_{sc}(\lambda_a/\sigma)} - \frac{1}{\sigma} g_{sc}(z/\sigma) \right), \tag{23a}$$

$$= \begin{cases} 0 & \text{if} \quad k \neq a, \\ \dfrac{g_{sc}^2(\lambda_a/\sigma)}{g_{sc}'(\lambda_a/\sigma)} & \text{if} \quad k = a. \end{cases} \tag{23b}$$

Et puisque $g_{sc}' = \dfrac{g_{sc}^2}{1 - g_{sc}^2}$, nous avons :

$$\frac{1}{n_a} j_a^* u_a u_a^* j_a = 1 - g_{sc}^2(\lambda_a/\sigma), \tag{24a}$$

$$= 1 - \frac{\sigma^2}{c_a^2 m_a^2 q_0^4}, \quad \text{d'après (14).} \tag{24b}$$

Pour résumer :

$$\frac{1}{n_a} j_a^* u_k u_k^* j_a = \begin{cases} 0 & \text{if} \quad k \neq a, \\ 1 - \dfrac{\sigma^2}{c_a^2 m_a^2 q_0^4} & \text{if} \quad k = a. \end{cases} \tag{25}$$

4. Nous avons comparé d'abord les positions asymptotiques théoriques et les positions observées des valeurs propres sur la fig. 7 et l'allure des vecteurs propres avec les vecteurs indicateurs des classes sur la fig. 8. On remarque pour un nombre relativement élevé d'observations ($N = 3000$) l'adéquation entre théorie et pratique est bien marquée.

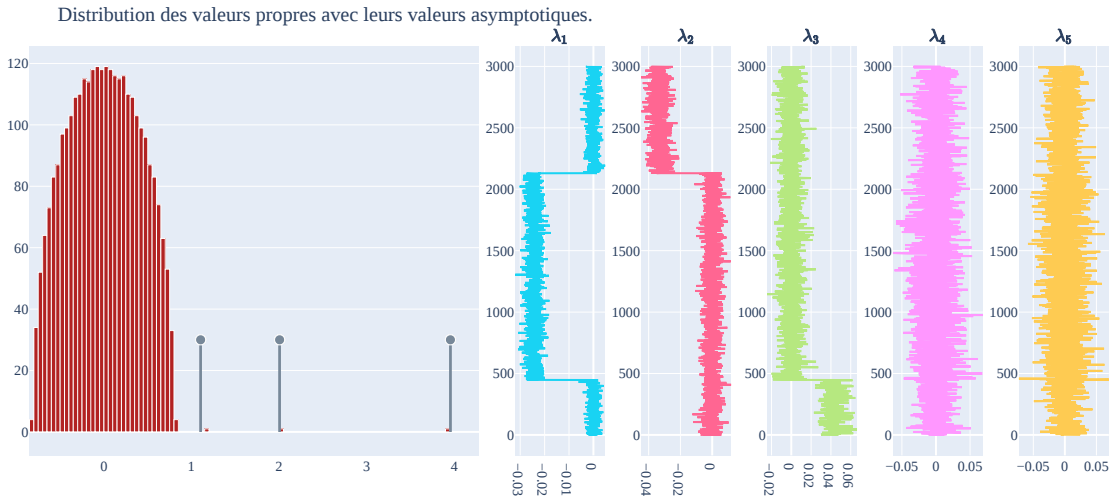

Figure 7: Observations des vecteurs propres associés aux valeurs propres isolées et comparaison entre l'alignement réel et l'alignement asymptotique théorique.

Figure 8: Observations des vecteurs propres associés aux valeurs propres isolées et comparaison entre l'alignement réel et l'alignement asymptotique théorique.

5. Nous pouvons déduire de cela un algorithme de détection de communautés.

   (i) Obtenir la matrice $\frac{1}{\sqrt{n}}B$ à partir de la matrice d'adjacence du graph.

   (ii) Détecter le nombre de de valeurs propres isolées. Si K n'est pas trop grand on peut facilement le faire visuellement. Sinon, on peut regarder la distance moyenne entre deux valeurs propres successives et chercher un coude dans cette courbe.

(iii) Nous obtenons ainsi au plus K directions selon lesquelles les données seront bien dispersés et les clusters bien répartis, grâce aux valeurs des alignements calculés plus haut.

(iv) Lancer un algorithme de clustering classique comme K-Means dans ce nouvel espace de dimension.

Le problème avec cet algorithme est qu'en pratique nous ne connaissons pas la valeur q_0. Pour l'estimer grossièrement nous pouvons utiliser la moyenne empirique des connexions de chaque noeud. En effet, si $i \in [\![1, N]\!]$ correspond à un noeud dans la classe a :

$$\mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n}\mathbf{A}_{ij}\right] = q_0^2 + \frac{1}{\sqrt{n}}c_a m_a = q_0^2 + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right), \tag{26}$$

et les $\mathbf{A}_{ij}$ étant indépendants, la variance de cette moyenne empirique sera en $\mathcal{O}\left(\frac{1}{n}\right)$, ce qui nous fournira un estimateur correct de q_0^2.

3 Cas hétérogène

1. Sur la fig. 9 nous observons un cas où les classes seront difficilement séparables par l'algorithme.

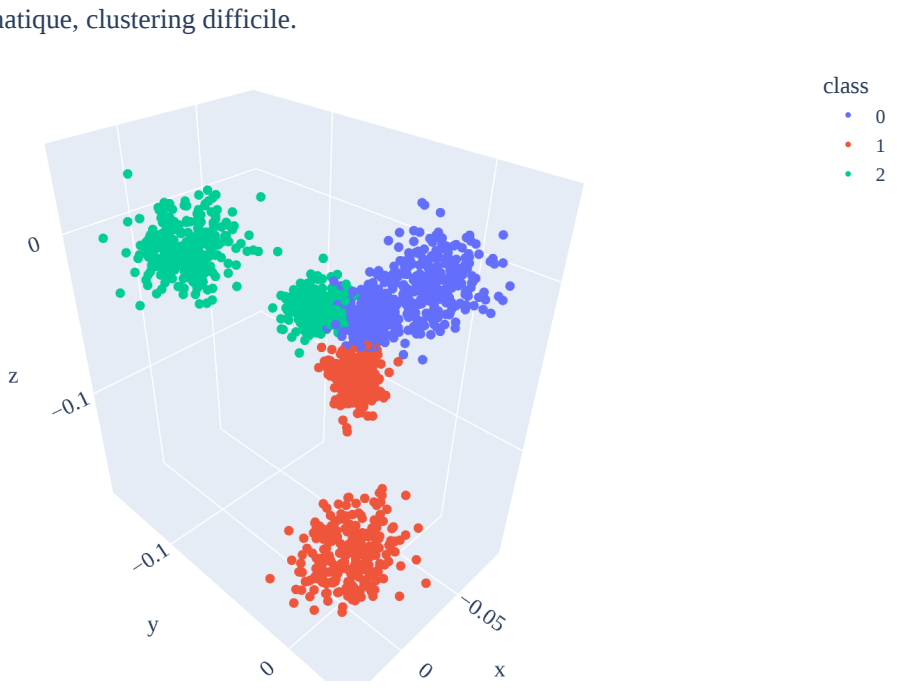

Figure 9: Situation où le clustering est délicat. $\mathbf{M}$ diagonale et les $q_i \in \{0.1, 0.4\}$.

2. Nous proposons pour ces algorithmes d'essayer de retrouver un cas où la distribution va se

rapprocher de celle d'une loi du demi-cercle. Pour cela, on peut remarquer que :

$$\mathbb{V}\left[\mathbf{B}_{ij}\right] = q_i q_j \mathbf{C}_{ab}(1 - q_i q_j \mathbf{C}_{ab}), \tag{27a}$$
$$= q_i q_j \mathbf{C}_{ab} - (q_i q_j)^2 \mathbf{C}_{ab}, \tag{27b}$$
$$\approx q_i q_j \mathbf{C}_{ab}, \tag{27c}$$

Si les q_i ne sont pas très grands. Et donc renormaliser la matrice $\mathbf{B}$ de la manière suivante :

$$\widetilde{\mathbf{B}} = \mathbf{diag}\left(\frac{1}{\sqrt{q_i}}\right) \mathbf{B} \mathbf{diag}\left(\frac{1}{\sqrt{q_i}}\right), \tag{28}$$

permettra d'obtenir une matrice avec une variance similaire à celle d'une matrice générée avec les q_i constants.

Cette renormalisation est en fait équivalente à renormaliser à les vecteurs propres de $\frac{1}{\sqrt{n}}\mathbf{B}$, puisque la décomposition spectrale de $\widetilde{\mathbf{B}}$ sera également mutlipliée à droite et à gauche par ces matrices diagonales.

En pratique, nous observons bien que le profil de répartition des valeurs propres se rapproche d'une distribution du demi-cercl (fig. 10), et lorsque qu'on les visualise en trois dimensions, ils sont effectivement peut-être un peu mieux séparables, fig. 11.

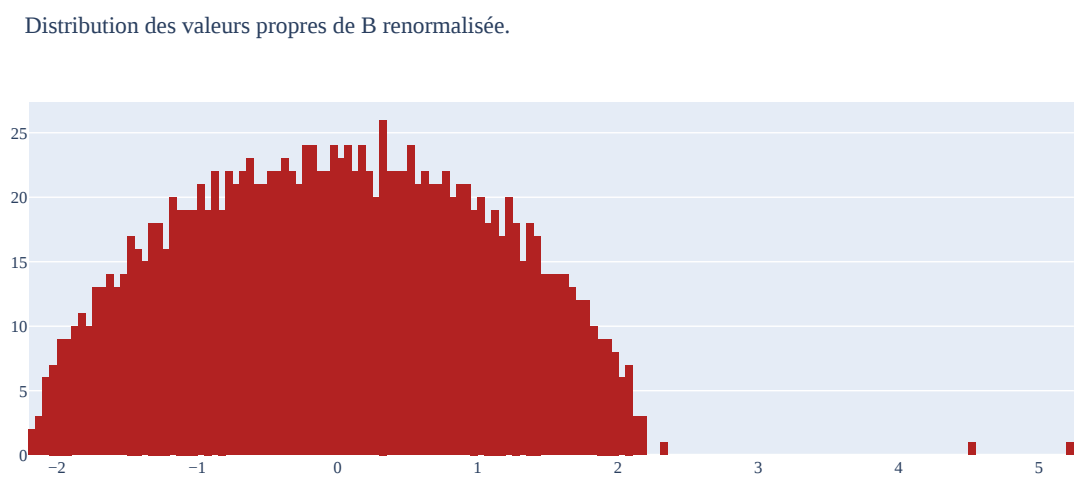

Figure 10: $\widetilde{\mathbf{B}}$ renormalisée. $\mathbf{M}$ diagonale et les $q_i \in \{0.1, 0.4\}$.

Répartition pour B renormalisée.

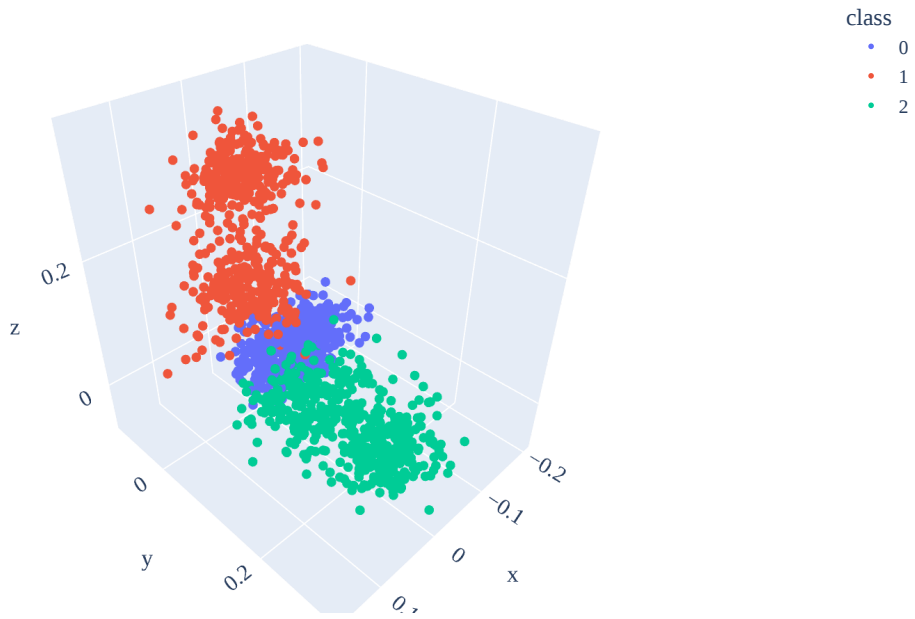

Figure 11: $\widetilde{\mathbf{B}}$ renormalisée. $\mathbf{M}$ diagonale et les $q_i \in \{0.1, 0.4\}$.

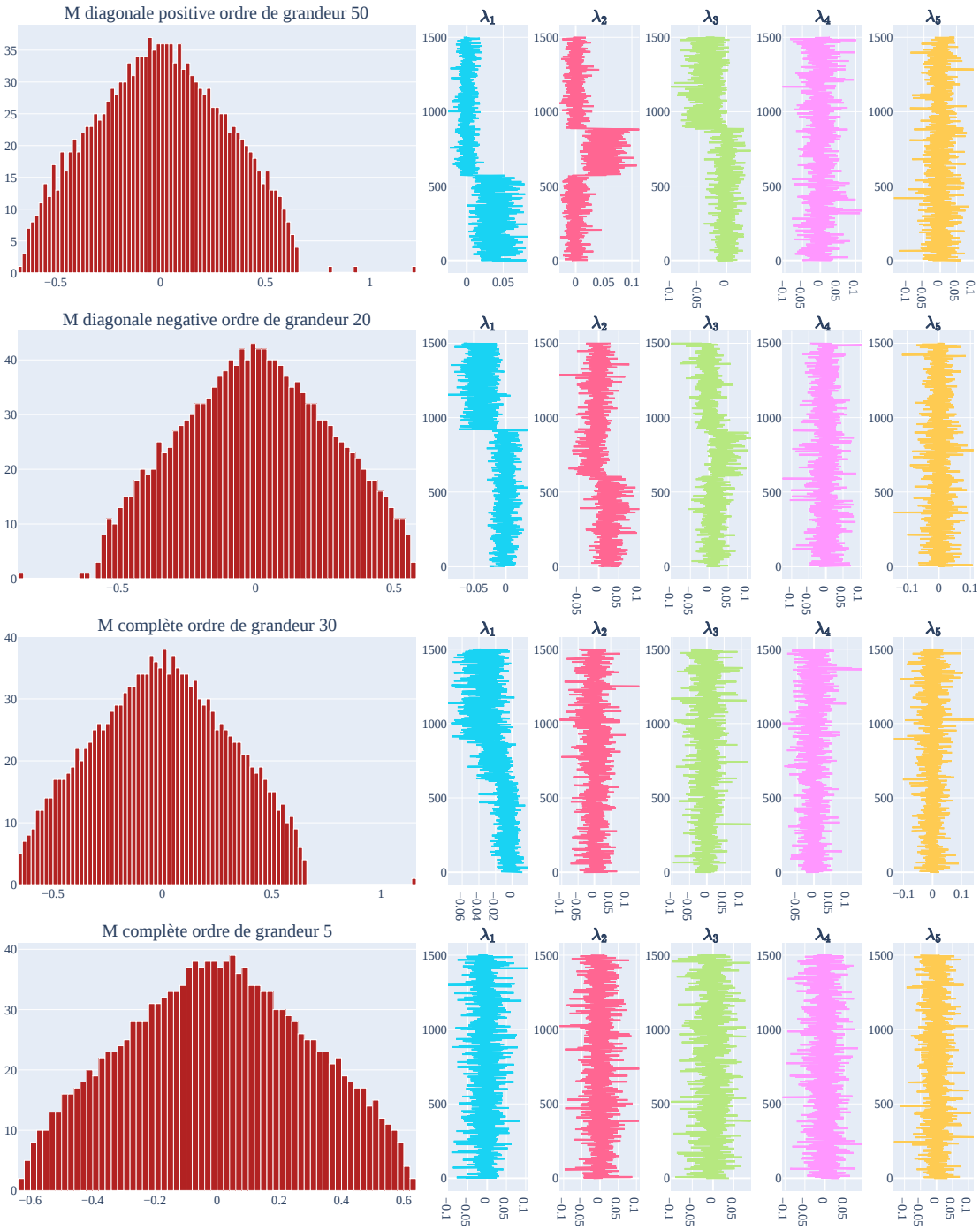

Figure 1: Observations lorsque les q_i sont répartis uniformément entre $[0.1, 0.5]$.

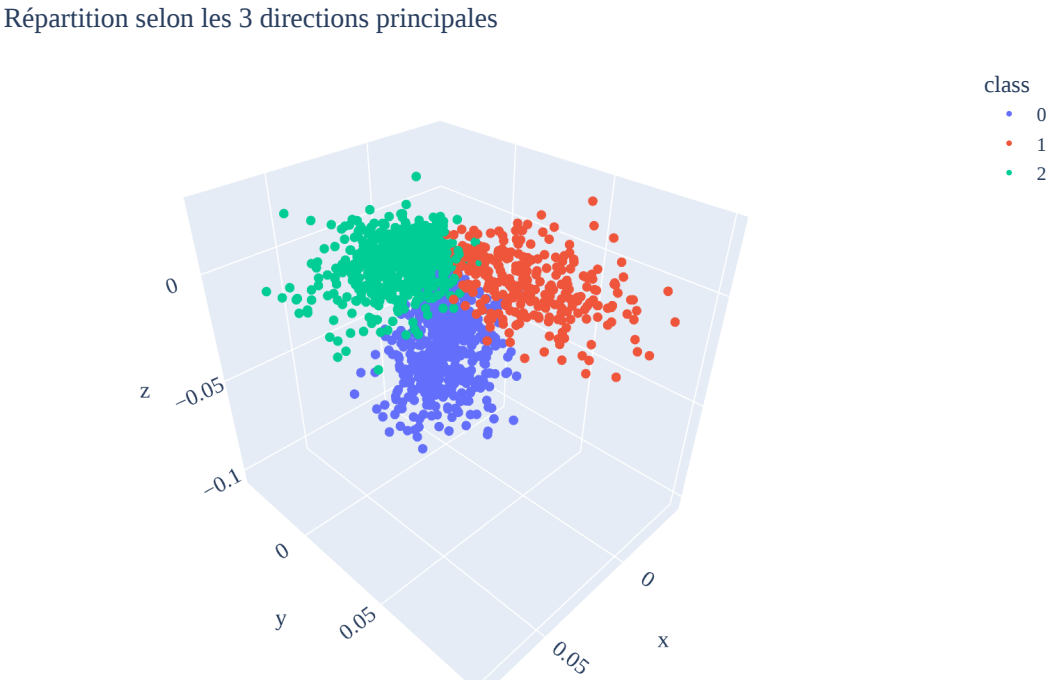

Figure 2: Répartition des classes lorsque les q_i sont répartis uniformément entre $[0.1, 0.5]$.

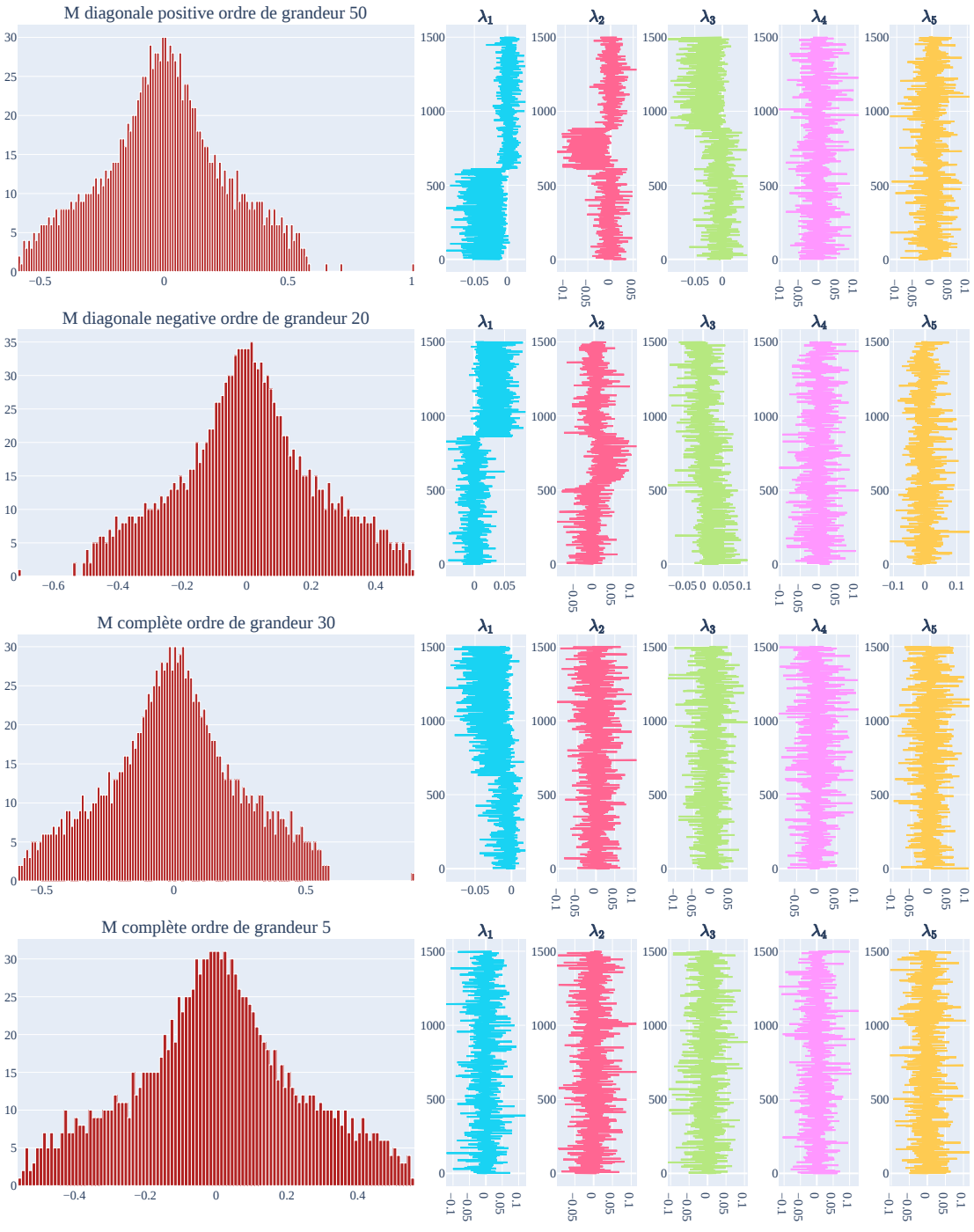

Figure 3: Observations lorsque $q_i \in \{0.1, 0.4\}$.

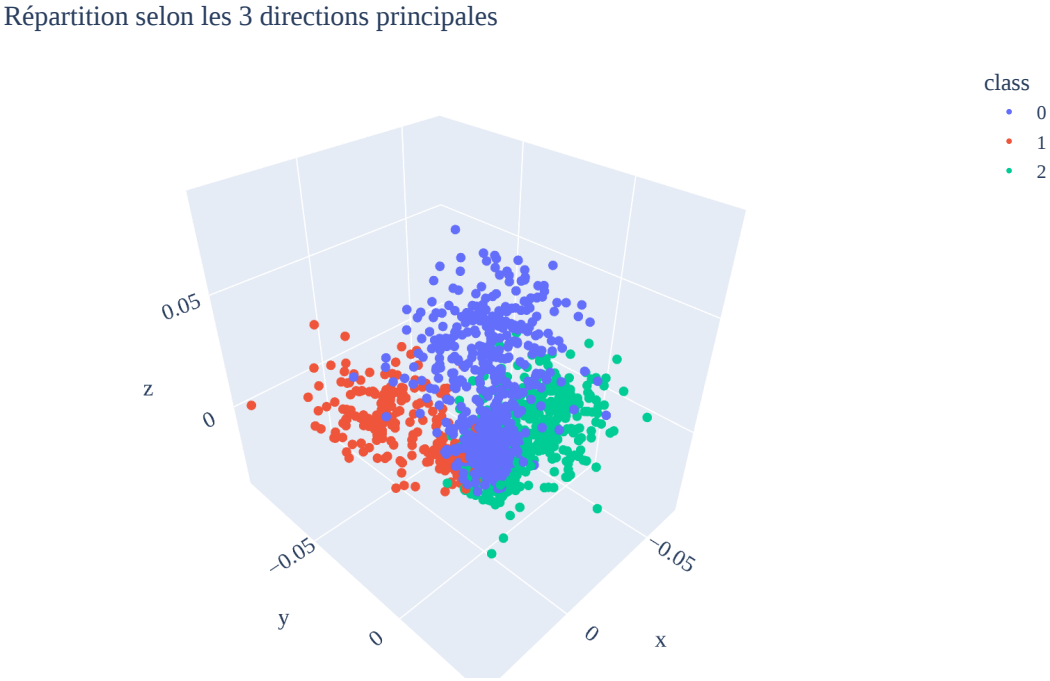

Figure 4: Répartition des classes lorsque $q_i \in \{0.1, 0.4\}$.

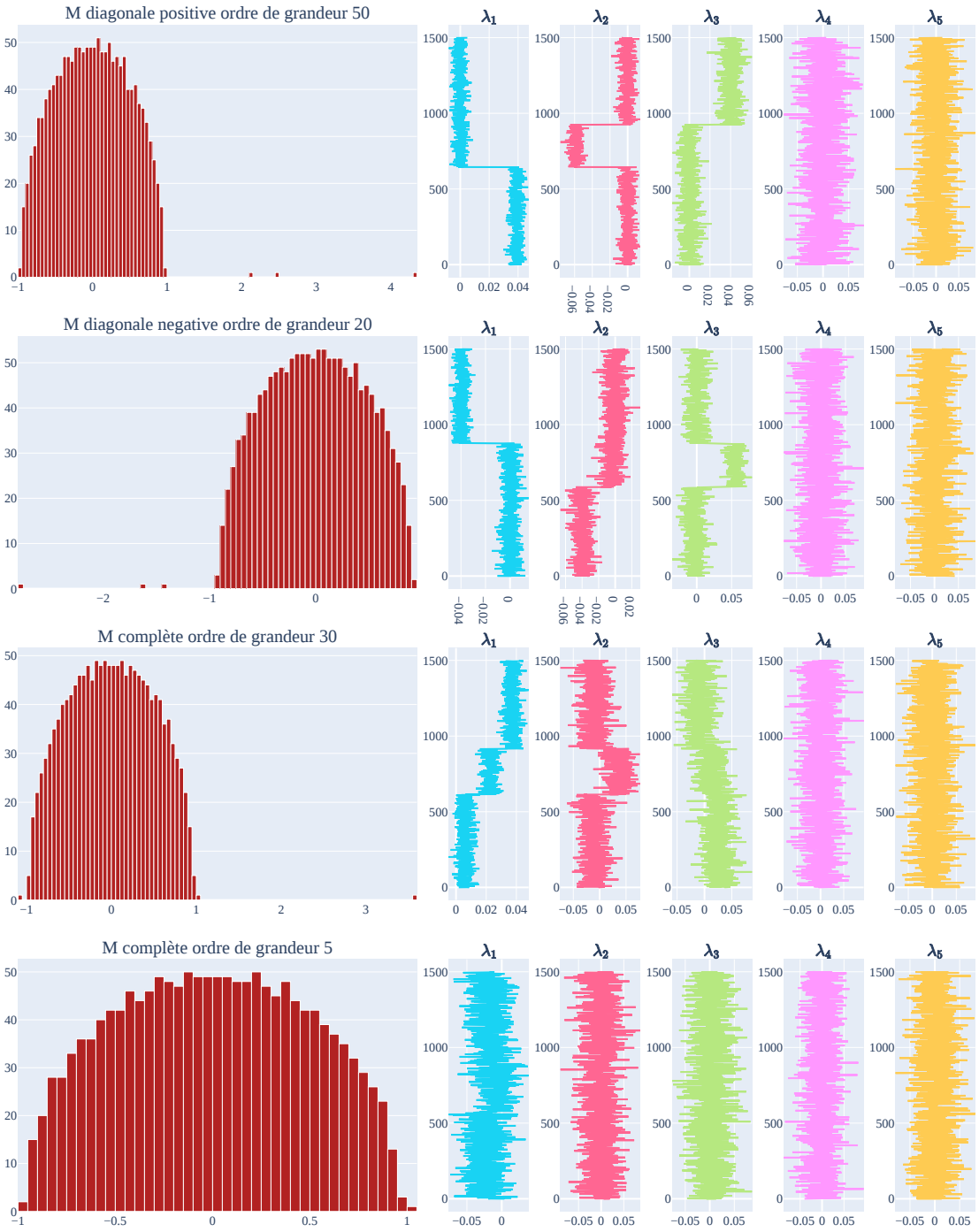

Figure 5: Observations lorsque $q_i = 0.6$, une valeur constante. On observe une distribution qui ressemble énormément à la loi du demi-cercle.

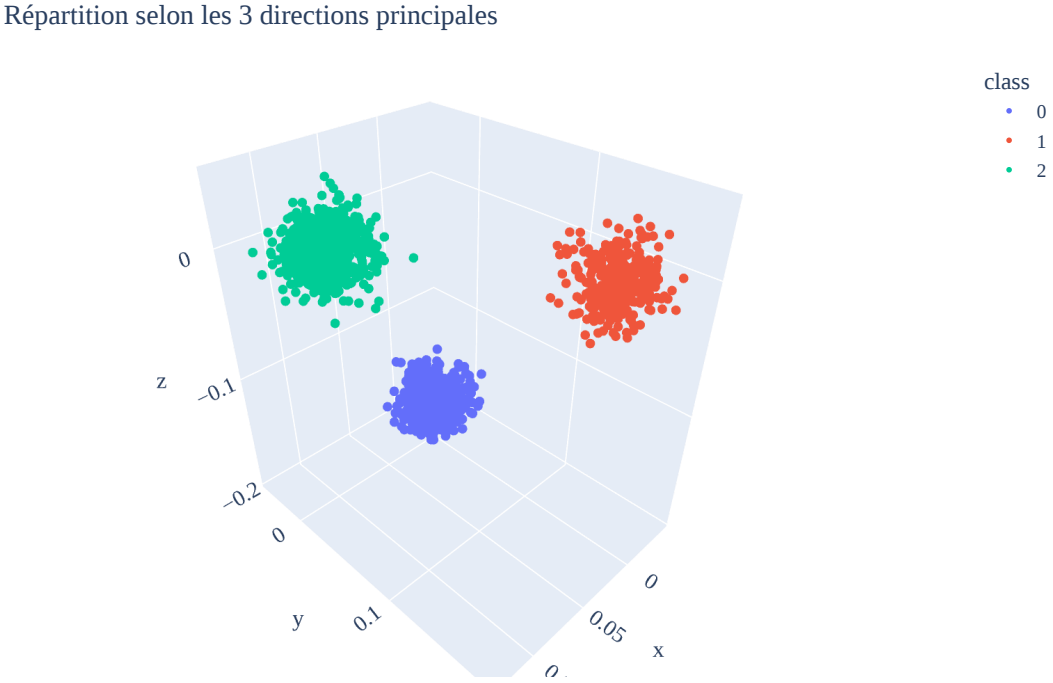

Figure 6: Répartition des classes lorsque $q_i = 0.6$. On remarque que points de différentes classes sont facilement séparables les uns des autres dans l'espace à trois dimensions des vecteurs propres isolés.