Théorie des matrices aléatoires, devoir maison 2

Guillaume Houry guillaume.houry@live.fr
Dorian Gailhard dorian.gailhard@telecom-paris.fr

June 27, 2023

1 Observations préliminaires

1. Soit i, j les indices de deux noeuds. On a :

$$\mathbb{E}\left[A_{ij}|q\right] = C_{ab}q_iq_j.$$

En notant $Q \in \mathbb{R}^{n\times K}$ la matrice $Q_{ia} = q_i\mathbb{1}_{i\in\mathcal{C}_a}$, cette relation s'écrit alors

$$\mathbb{E}\left[A_{ij}|q\right] = (QCQ^*)_{i,j},$$

soit

$$\mathbb{E}\left[A|q\right] = QCQ^*.$$

On peut alors décomposer

$$\boxed{\frac{1}{\sqrt{n}}A = \frac{1}{\sqrt{n}}\left(A - QCQ^*\right) + \frac{1}{\sqrt{n}}QCQ^* = U' + V'}$$

avec $\mathbb{E}\left[U'\middle|\, q\right] = 0$ et V' de rang au plus K. Finalement, conditionnellement à q, les coefficients de A sont indépendants (car lois de Bernoulli de coefficients entièrement déterminés par q). La matrice QCQ^* étant elle-même entièrement déterminée par q, on en déduit que les coefficients de U' sont indépendants conditionnellement à q. Finalement, les variances des coefficients de U' sont des variances de Bernoulli :

$$\mathrm{Var}(U'_{ij}|q) = \frac{1}{n}C_{ab}q_iq_j(1 - C_{ab}q_iq_j).$$

2. De même,

$$\mathbb{E}\left[B_{ij}|q\right] = C_{ab}q_iq_j - (qq^*)_{ij} = \frac{1}{\sqrt{n}}M_{ab}q_iq_j,$$

car $C_{ab} = 1 + M_{ab}/\sqrt{n}$. Par conséquent, on obtient

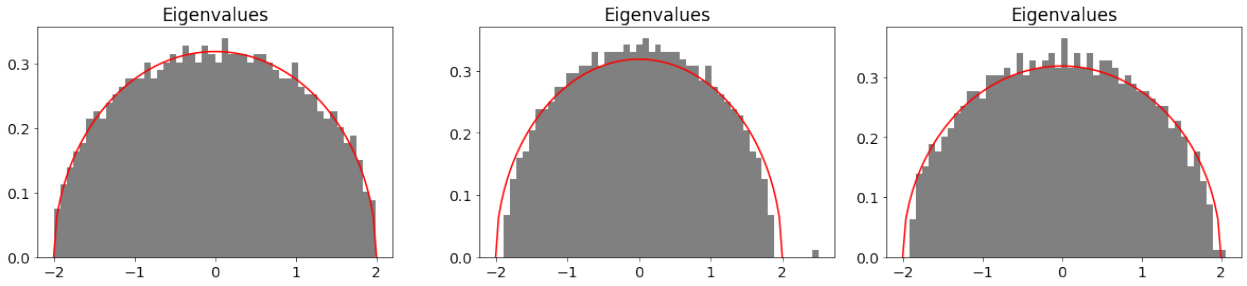
$$\boxed{\frac{1}{\sqrt{n}}B = \frac{1}{\sqrt{n}}\left(B - \frac{1}{\sqrt{n}}QMQ^*\right) + \frac{1}{n}QMQ^* = U + V}$$

le premier terme étant d'espérance conditionnelle nulle et le second de rang au plus K. De même que précédemment, les coefficients de U sont tous indépendants conditionnellement aux q et l'on peut calculer leur variance:

$$\mathrm{Var}(U_{ij}|q) = \frac{1}{n}C_{ab}q_iq_j\left(1 - C_{ab}q_iq_j\right) \sim \frac{1}{n}q_iq_j(1 - q_iq_j),$$

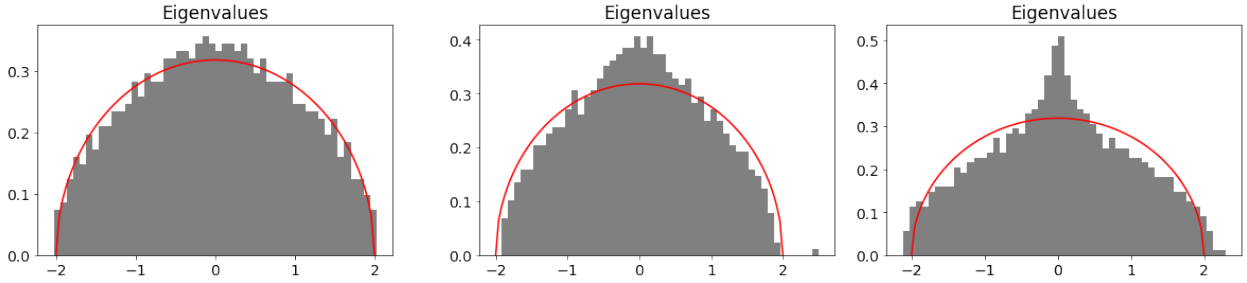
car $C_{ab} \to 1$.

3. La répartition des valeurs propres dans les trois cas demandées a été normalisée et centrée. La loi du demi-cercle a été représentée en rouge dans les figures qui suivent.



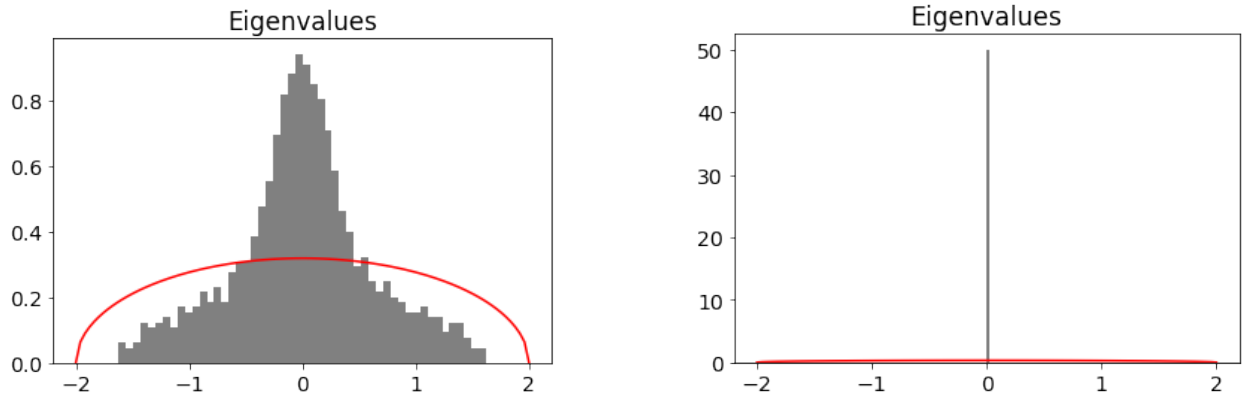
$M = 0_K$ $M = 1_K$ $M = \frac{1}{2}1_K + \frac{1}{2}I_K$ (i.e. $M_{ij} = 0.5 + 0.5\delta_{ij}$).

Figure 1: Cas homogène : $\forall i, q_i = q_0 = 0.9$



$M = 0_K$, l'étalement est pris à 0.1 $M = 1_K$, l'étalement est pris à 0.1 $M = 1_K$, l'étalement est pris à 0.3

Figure 2: q_i uniforme autour de $q_0 = 0.9$



$M = 1_K, q_1 = 0.1, q_2 = 0.7, \mathbb{P}(q_i = q_1) = 0.3$ $M = 1_K, q_1 = 0, q_2 = 1, \mathbb{P}(q_i = q_1) = 0.5$

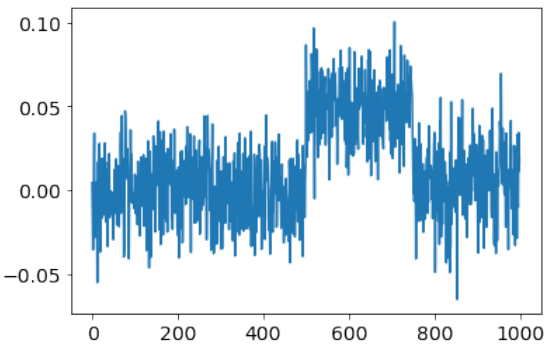
Figure 3: q_i bimodal : $q_i \in \{q_1, q_2\}$.

Lorsque les q_i sont égaux, la répartition des valeurs propres reste très proche de la loi du demi-cercle. Lorsqu'on autorise un étalement autour de q_0, le spectre se contracte progressivement vers un dirac en 0, avec une contraction d'autant plus forte que l'étalement est large. Le même phénomène a lieu lorsque $q_i \in \{q_1, q_2\}$, le dirac est d'ailleurs visible sur la dernière figure où

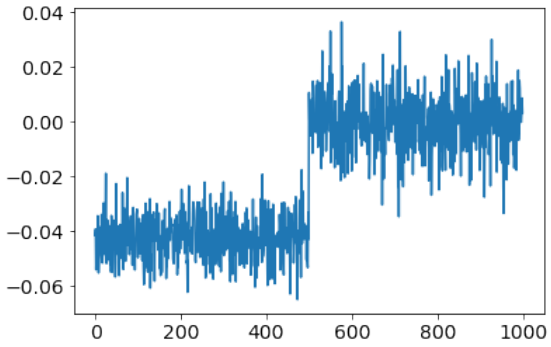
$q_1 = 0$ et $q_2 = 1$. En effet, dans ce dernier cas, toutes les colonnes non nulles sont égales: la matrice B est donc de rang 1, d'où une concentration des valeurs propres en 0.

Dans le cas où $M = 1_K$, on observe des valeurs propres isolées hors du bulk qui sont absentes lorsque $M = 0$ ou que M prend des valeurs plus petites. Cela montre donc l'existence d'un lien entre les valeurs de M et le spectre de B.

4. La figure 4 représente la valeur des coefficients des deux vecteurs propres de plus grande valeur dans un cas où il existe des vecteurs propres isolés. On se rend compte que ces vecteurs propres décrivent des plateaux successifs, avec un bruit résiduel sur chaque composante. Ces plateaux correspondent en fait aux différentes classes des noeuds du graphe. On peut alors imaginer effectuer un clustering grâce à l'information contenue dans ces vecteurs propres (en interprétant la i-ème ligne de la concaténation des vecteurs comme information du i-ème échantillon).



Pour l'avant-dernière valeur propre



Pour la dernière valeur propre

Figure 4: Allure de différents vecteurs propres associés à des valeurs isolées, cas $M = 3I_K$ et $q_0 = 0.9$.

Une méthode de clustering par analyse spectrale de B est bien adaptée lorsque les q_i sont tous égaux, car seules les valeurs de M font varier la probabilité de connections des différents échantillons : les probabilités de liaisons entre les noeuds ne dépendent que des classes auxquels ils appartiennent. En revanche, lorsque les q_i sont hétérogènes, les deux termes intervenant dans les probabilités de la matrice d'adjacence (les q_i et la matrice M) interfèrent, et l'algorithme de clustering spectral risque de confondre les contributions des deux termes. On risque donc de mal retrouver les classes réelles de noeuds.

2 Cas Homogène

1. Dans le cas homogène, la variance asymptotique de chaque coefficient U de la matrice non perturbée vaut $\frac{q_0^2(1-q_0^2)}{n}$. Par conséquent, la mesure spectrale de $\frac{1}{q_0\sqrt{1-q_0^2}}U$ converge la loi du demi-cercle, car ses coefficients sont indépendants, d'espérance nulle et de variance convergeant uniformément vers $1/n$. . En particulier, les valeurs propres de U sont asymptotiquement incluses dans l'intervale $[-2\tilde{q_0}, 2\tilde{q_0}]$ où $\tilde{q_0} = q_0\sqrt{1-q_0^2}$.

Soit $\lambda \in \mathbb{R} \setminus [-2\tilde{q_0}, 2\tilde{q_0}]$. Pour que λ soit valeur propre de $\frac{1}{\sqrt{n}}B$ sans être valeur propre de la

matrice sans perturbation U, il faut que

$$\det\left(\frac{1}{\sqrt{n}}B - \lambda I\right) = \det\left(U - \lambda I + V\right) = 0,$$

et

$$\det\left(U - \lambda I\right) \neq 0.$$

On peut alors factoriser le premier déterminant par le second pour obtenir la condition

$$\det\left(I + \left(U - \lambda I\right)^{-1} V\right) = 0.$$

Or, $V = \frac{1}{n} QMQ^*$. Dans le cas homogène, $Q = q_0 J$ où $J \in \mathbb{R}^{n \times K}$ est l'indicatrice des classes de chaque noeud. La condition devient

$$\det\left(I + \frac{q_0^2}{n}\left(U - \lambda I\right)^{-1} JMJ^*\right) = \det\left(I + \frac{q_0^2}{n} J^*\left(U - \lambda I\right)^{-1} JM\right) = 0,$$

la permutation provenant de l'identité de Silvester.

Le théorème de Wigner isotrope nous assure que presque sûrement, pour tout $\tilde{\lambda}$,

$$\frac{1}{n} J^* \left(\frac{1}{\tilde{q_0}} U - \tilde{\lambda} I\right)^{-1} J - g_{sc}(\tilde{\lambda}) \frac{1}{n} J^* J \to 0.$$

En particulier, pour $\tilde{\lambda} = \lambda / \tilde{q_0}$, cela donne:

$$\frac{1}{n} J^* \left(U - \lambda I\right)^{-1} J - \frac{g_{sc}(\lambda / \tilde{q_0})}{\tilde{q_0}} \frac{1}{n} J^* J \to 0.$$

Or, $\frac{1}{n} J^* J$ est la matrice $\operatorname{diag}\left(\frac{|\mathcal{C}_1|}{n}, \ldots, \frac{|\mathcal{C}_K|}{n}\right)$ qui converge vers $\operatorname{diag}(c_1, \ldots, c_K)$. Par conséquent, pour que λ soit une valeur propre asymptotique de $\frac{1}{\sqrt{n}} B$, il faut que

$$\det\left(I + \frac{q_0^2}{\tilde{q_0}} g_{sc}(\lambda / \tilde{q_0}) \cdot \operatorname{diag}(c_1, \ldots, c_K) M\right) = 0.$$

M étant une matrice diagonale par hypothèse, la condition devient:

$$\prod_{k=1}^{K} \left(1 + \frac{q_0^2}{\tilde{q_0}} g_{sc}(\lambda / \tilde{q_0}) c_k M_{kk}\right) = 0.$$

Il faut donc qu'il existe un λ et une classe k satisfaisant

$$1 + \frac{q_0^2}{\tilde{q_0}} g_{sc}(\lambda / \tilde{q_0}) c_k M_{kk} = 0,$$

soit

$$g_{sc}(\lambda / \tilde{q_0}) = -\frac{\tilde{q_0}}{q_0^2 c_k M_{kk}}.$$

Or, pour $z > 2$, $g_{sc}(z) = \frac{1}{2}\left(-z + \sqrt{z^2-4}\right)$ et pour $z < -2$, $g_{sc}(z) = \frac{1}{2}\left(-z - \sqrt{z^2-4}\right)$. Une simple anaylse de fonction montre que g_{sc} prend ses valeurs dans $]-1,1[$ sur $\mathbb{R} \setminus [-2,2]$.

On en déduit la condition suivante pour l'existence d'une valeur propre isolée de $\frac{1}{\sqrt{n}}B$: il faut qu'il existe une classe k telle que

$$\frac{q_0^2}{\tilde{q_0}} c_k |M_{kk}| > 1,$$

soit:

$$\boxed{\frac{q_0}{\sqrt{1-q_0^2}} c_k |M_{kk}| > 1.} \tag{1}$$

2. Si la condition précédente est vérifiée, on a l'existence de λ_k tel que $g = g_{sc}(\lambda_k/\tilde{q_0}) = -\frac{\tilde{q_0}}{q_0^2 c_k M_{kk}}$. Vu que $g = -\frac{1}{\lambda_k/\tilde{q_0}+g}$, on a $\lambda_k/\tilde{q_0} = -g - \frac{1}{g}$, ce qui nous donne une formule explicite pour la valeur propre correspondante :

$$\lambda_k/\tilde{q_0} = \frac{q_0^2 c_k M_{kk}}{\tilde{q_0}} + \frac{\tilde{q_0}}{q_0^2 c_k M_{kk}};$$

$$\boxed{\lambda_k = q_0^2 c_k M_{kk} + \frac{1-q_0^2}{c_k M_{kk}}.} \tag{2}$$

Pour vérifier expérimentalement ces résultats, on se place dans un modèle très simple à 2 classes telles que $c_2 = 1 - c_1$ et $M_{22} = 0$. Les paramètres à ajuster sont donc q_0, M_{11} et c_1. On peut tester la valeur λ_1 d'apparition de la valeur propre associée à la classe 1 ; les simulations sont représentées sur la figure 5.

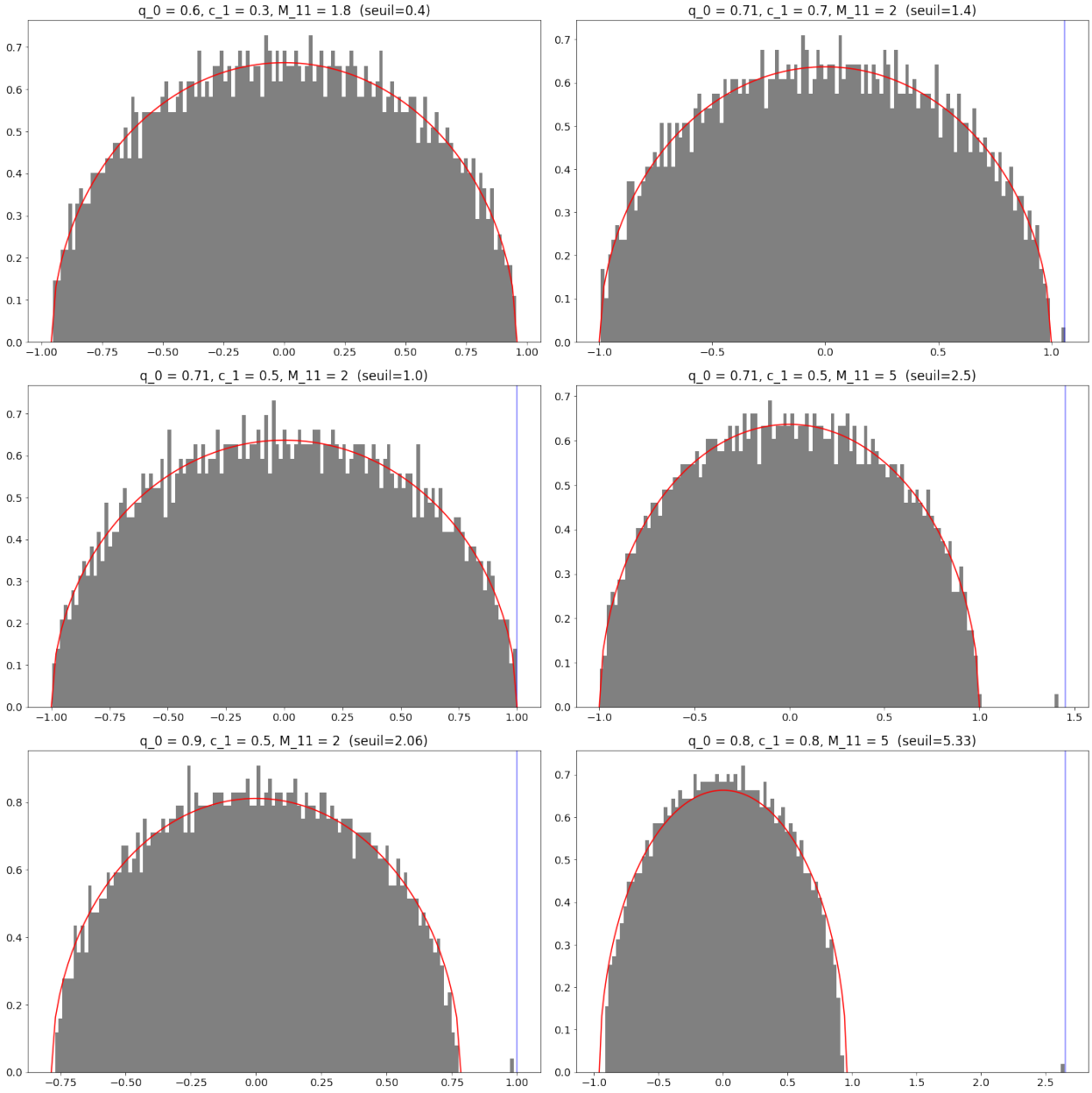


Figure 5: Simulations obtenues dans le cas homogène, pour différentes valeurs de q_0, c_1 et M_{11}, et $n = 1800$. La distribution des valeurs propres est représentées en gris. La courbe rouge représente la distribution théorique du bulk. La ligne bleue identifie la valeur asymptotique calculée dans l'équation 2. Enfin, le seuil indiqué dans le titre de chaque graphe correspond à l'équation 1. On constate que des valeurs propres isolées apparaissent dès que le seuil dépasse 1, et ces valeurs propres sont proches de la valeur asymptotique attendue.

3. Soit u_k le vecteur propre associé à la valeur propre isolée λ_k précédente. Soit Γ un contour complexe orienté positivement autour de λ_k qui n'encercle aucune autre valeur propre. Par formule de Cauchy intégrale, il vient que :

$$\frac{1}{n_k} j_k^* u_k u_k^* j_k = \frac{1}{2i\pi n_k} \oint_\Gamma j_k^* \left(\frac{1}{\sqrt{n}} B - I \right)^{-1} j_k dz,$$

avec $n_k = \|j_k\|^2 = |\mathcal{C}_k|$. Cela nous donne donc une formule pour calculer l'alignement entre le vecteur propre u_k et le vecteur indicateur j_k.

Or, $\left(\frac{1}{\sqrt{n}} B - zI \right)^{-1} = \left(U - zI + \frac{q_0^2}{n} JMJ^* \right)^{-1}$. De plus, d'après l'identité de Woodbury:

$$\left(U - zI + \frac{q_0^2}{n} JMJ^* \right)^{-1} = (U-zI)^{-1} - \frac{q_0^2}{n} (U-zI)^{-1} JM \left(I + \frac{q_0^2}{n} J^* (U-zI)^{-1} JM \right)^{-1} J^* (U-zI)^{-1}.$$

Tout d'abord, $j_k^* (U-zI)^{-1} j_k$ est une fonction holomorphe sur toute la surface encerclée par Γ, car Γ n'encercle aucune valeur propre de U. Par conséquent,

$$\oint_\Gamma j_k^* (U-zI)^{-1} j_k dz = 0.$$

De plus, on sait que $\frac{1}{n} J^* (U-zI)^{-1} J \to \mathrm{diag}(c_1, \ldots, c_K) \frac{g_{sc}(z/\tilde{q}_0)}{\tilde{q}_0}$ et $\frac{1}{\sqrt{n_k n}} j_k^* (U-zI)^{-1} J \to c_k \frac{g_{sc}(z/\tilde{q}_0)}{\tilde{q}_0} e_k$ presque sûrement, avec $e_k \in \mathbb{R}^K$ valant 1 sur la k^e coordonnée et 0 ailleurs. Par conséquent, le terme

$$\frac{1}{n_k} \frac{q_0^2}{n} j_k^* (U-zI)^{-1} JM \left(I + \frac{q_0^2}{n} J^* (U-zI)^{-1} JM \right)^{-1} J^* (U-zI)^{-1} j_k$$

converge presque sûrement vers

$$q_0^2 c_k \left(\frac{g_{sc}(z/\tilde{q}_0)}{\tilde{q}_0} \right)^2 \cdot e_k^* M \left(I + \frac{q_0^2}{n} \mathrm{diag}(c_1, \ldots, c_K) M \frac{g_{sc}(z/\tilde{q}_0)}{\tilde{q}_0} \right)^{-1} e_k.$$

La matrice $M \left(I + \frac{q_0^2}{n} \mathrm{diag}(c_1, \ldots, c_K) M \right)^{-1}$ étant diagonale, cette quantité se réécrit

$$\frac{q_0^2 c_k M_{kk} \left(g_{sc}(z/\tilde{q}_0)/\tilde{q}_0 \right)^2}{1 + q_0^2 c_k M_{kk} g_{sc}(z/\tilde{q}_0)/\tilde{q}_0} = \frac{1}{\tilde{q}_0} \frac{g_{sc}(z)^2}{g_{sc}(z/\tilde{q}_0) - g_{sc}(\lambda_k/\tilde{q}_0)},$$

car $g_{sc}(\lambda_k/\tilde{q}_0)/\tilde{q}_0 = -\frac{1}{q_0^2 c_k M_{kk}}$.

Finalement, l'alignement asymptotique des vecteurs propres avec k_k vérifie :

$$\frac{1}{n_k} j_k^* u_k u_k^* j_k = \frac{1}{2i\pi} \oint_\Gamma \frac{1}{\tilde{q}_0} \frac{g_{sc}(z/\tilde{q}_0)^2}{g_{sc}(z/\tilde{q}_0) - g_{sc}(\lambda_k/\tilde{q}_0)} dz + o(1).$$

Or,

$$\lim_{z \to \lambda_k} (z - \lambda_k) \frac{1}{\tilde{q}_0} \frac{g_{sc}(z/\tilde{q}_0)^2}{g_{sc}(z/\tilde{q}_0) - g_{sc}(\lambda_k/\tilde{q}_0)} = \frac{g_{sc}(\lambda_k/\tilde{q}_0)^2}{g_{sc}'(\lambda_k/\tilde{q}_0)} = 1 - g_{sc}^2(\lambda_k/\tilde{q}_0).$$

La formule des résidus permet alors de conclure:

$$\boxed{\frac{1}{n_k} j_k^* u_k u_k^* j_k = 1 - g_{sc}^2(\lambda_k / \tilde{q_0}) + o(1) = 1 - \frac{1 - q_0^2}{q_0^2 c_k^2 M_{kk}^2} + o(1).} \tag{3}$$

En particulier, l'alignement se rapproche 1 lorsque M_{kk} augmente.

4. On reprend le dispositif expérimental précédent, en mesurant cette fois l'alignement du vecteur propre de plus grande valeur avec le vecteur indicateur j_i. Les résultats sont représentés sur la figure 6.

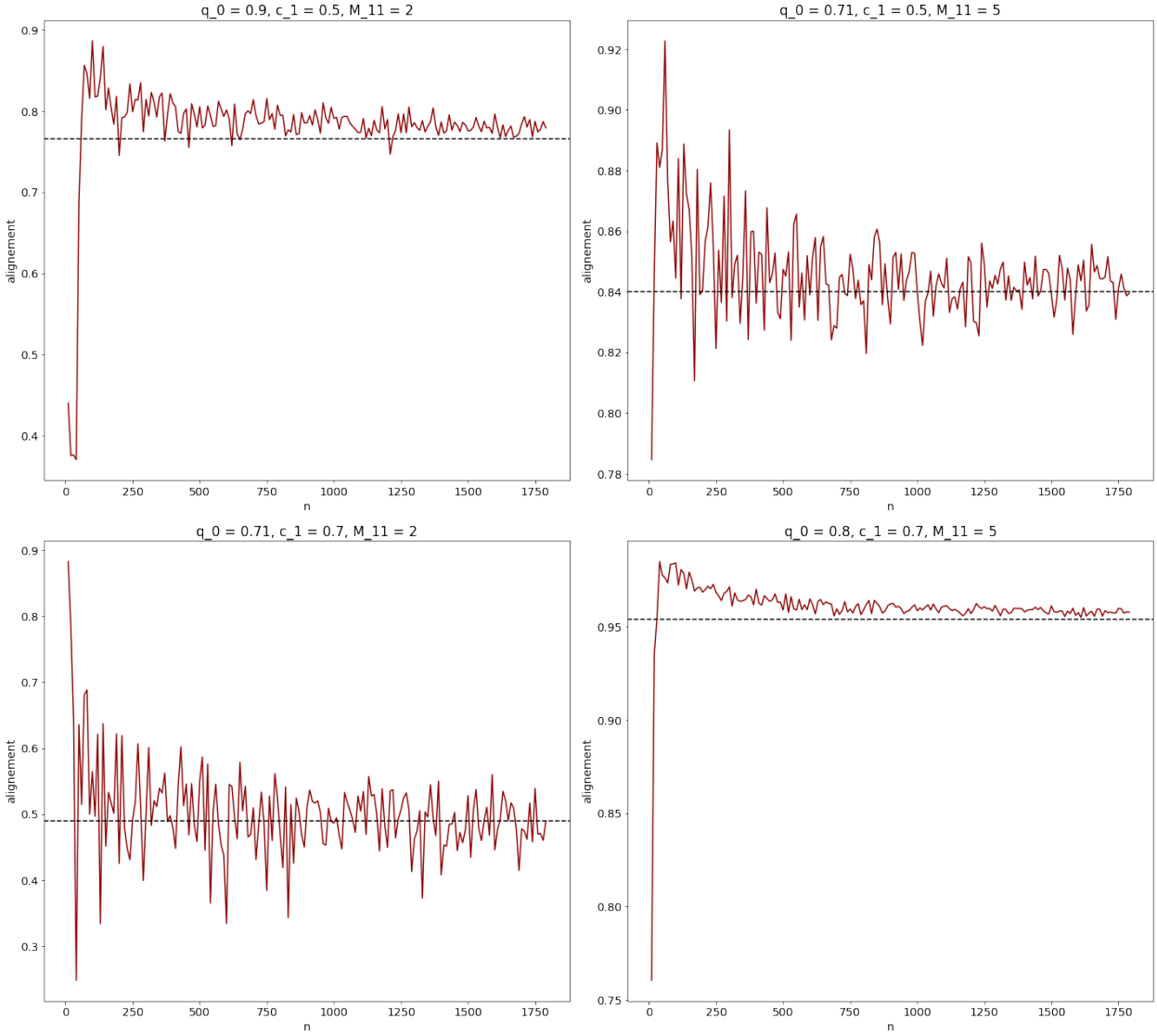


Figure 6: Evolution de l'alignement en fonction de n, pour différentes valeurs de q_0, c_1 et M_{11}. La ligne en pointillée correspond à la valeur limite de l'équation 3. Les alignements convergent bien vers leur valeur théorique.

5. On a montré que lorsque les bonnes conditions sont réunies, les vecteurs propres de $\frac{1}{\sqrt{n}}B$ de plus grande valeur s'alignent avec les vecteurs caractéristiques des différentes classes. On peut donc utiliser ces vecteurs propres pour identifier les communautés dans un graphe donné.

Pour cela, étant donné une matrice d'adjacence A, on calcule la matrice de modularité $B = A - q_0^2 \mathbf{1}\mathbf{1}^T$. Si q_0 n'est pas connu, on peut l'estimer au premier ordre comme la proportion de coefficients non nuls de A. On calcule ensuite les K premiers vecteurs propres de B (dans l'ordre de valeur propre décroissante en valeur absolue). On peut ensuite utiliser un algorithme de clustering linéaire comme KMeans pour identifier les communautés à partir de ces vecteurs propres. Le code correspondant est donnée à la figure 7.

```python
def clustering(A, K=2):
  """
    A = matrice d'adjacence du graphe
    K = nombre de classes à identifier
  """

  n = A.shape[0]

  B = A - np.mean(A)
  eigen, vectors = sp.linalg.eigh(B/np.sqrt(n), subset_by_index=[n-K+1, n-1])

  classes = KMeans(n_clusters=K, n_init='auto').fit_predict(vectors)
  return classes
```

Figure 7: Un algorithme de détection de communautés basé sur les valeurs propres de la matrice de modularité du graphe.

Pour évaluer les performances de l'algorithme, il faut maintenant définir la métrique à mesurer. La difficulté est que dans le cas du clustering, on détecte des ensembles de communautés et non des labels. Il y n'a donc pas de correspondance explicite entre les classes prédites par l'algorithmes et les classes réelles à prédire. On définira donc la précision de l'algorithme comme la prédiction maximale pour toutes les associations possibles entre communautés prédites et communautés réelles.

Les résultats sur différents modèles de données aléatoires sont présentés figure 8. On constate notamment que l'existence de valeurs propres isolées hors du bulk n'est pas une condition suffisante pour retrouver les communautés à partir des vecteurs propres.

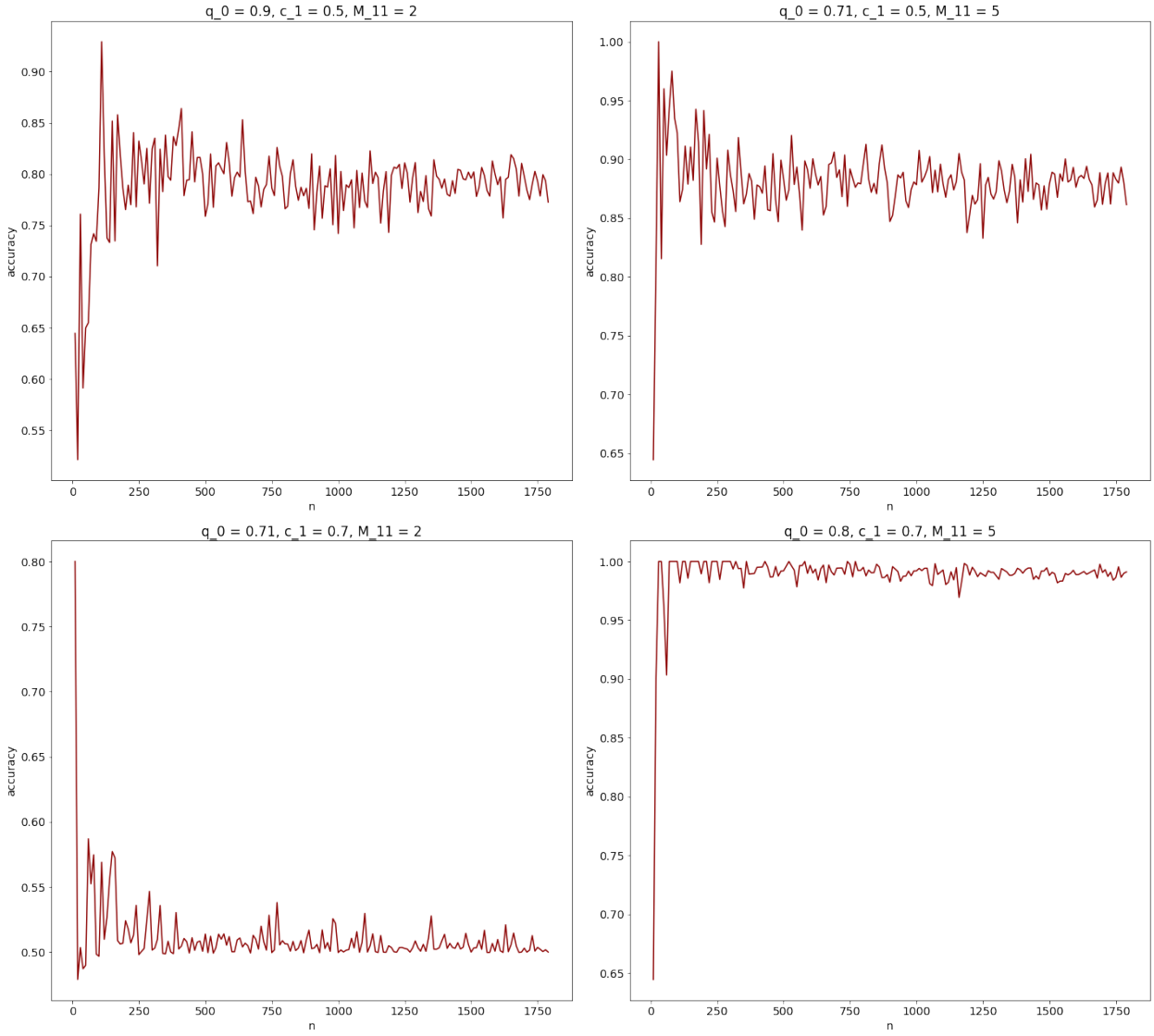


Figure 8: Précision du clustering pour les modèles de matrices d'adjacence utilisés précédemment, en fonction de n. Les modèles utilisés ici dépassent tous le seuil de l'équation 1. Les performances varient grandement selon le modèle, allant de 0.5 (clustering aléatoire) à 1 (détection parfaite des communautés).

3 Cas Hétérogène

1. Reprenons le modèle expérimental de la partie précédente, et choisissons $q_0 = 0.8$, $c_1 = 0.7$, $M_{11} = 5$ (situation qui donnait les meilleurs résultats pour la détection de communauté). Cette fois, on fixe une autre valeur q_1, et on attribue de manière équiprobable la valeur q_0 ou q_1 à chaque q_i, selon des tirages iid indépendant de la communauté de chaque i :

$$\forall i, \quad q_i \sim q_0\alpha_i + q_1(1-\alpha_i), \quad \alpha_i \sim \text{Ber}(0.5).$$

La figure 9 montre les performances du clustering en fonction de la valeur de q_1. On voit que

l'algorithme proposé précédemment ne fonctionne plus lorsque q_1 prend des valeurs trop différentes de q_0, montrant les limites de l'algorithme dans le cas hétérogène.

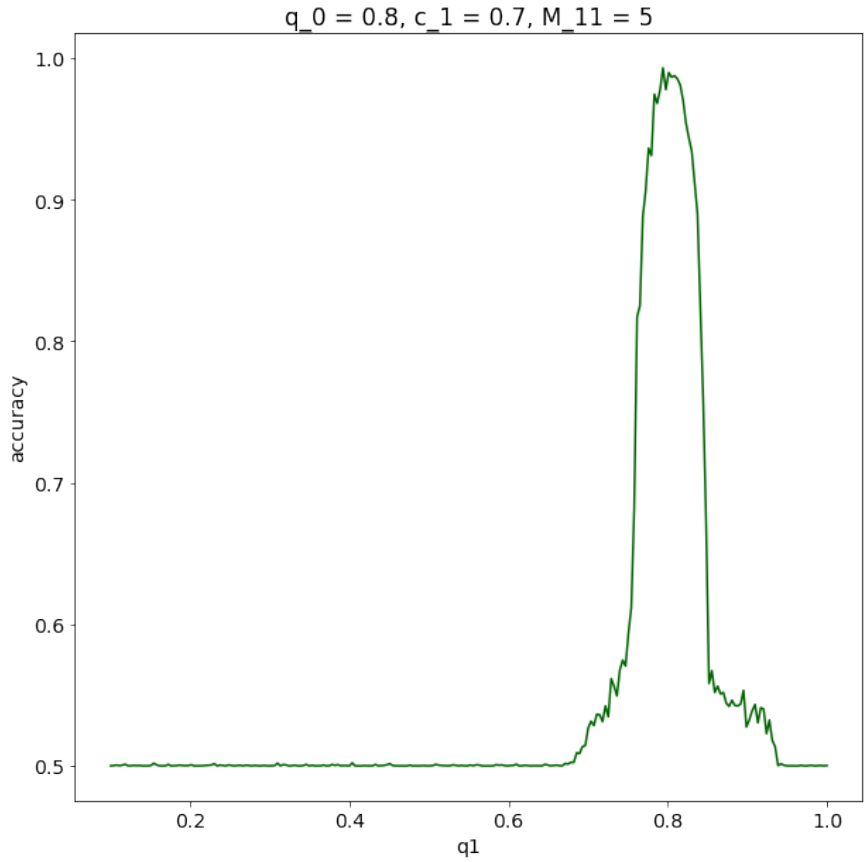


Figure 9: Résultats de l'algorithme précédent sur un modèle hétérogène en fonction de q_1. Lorsque q_1 s'éloigne trop de q_0, le clustering ne fonctionne plus.

2. Pour corriger ce problème, deux approches sont possible.

La première approche consiste à diviser toutes les lignes et les colonnes de B par les valeurs de q correspondantes :

$$\tilde{B}_i, j = \frac{1}{q_i q_j} B_{ij}.$$

Cela revient à appliquer l'algorithme de clustering initial à la matrice normalisée

$$\tilde{B} = DBD,$$

avec $D = \text{diag}(q_1^{-1}, \ldots, q_n^{-1})$.

La seconde revient à effectuer la normalisation au niveau des vecteurs propres. C'est-à-dire qu'avant d'appliquer KMeans, on remplace chaque vecteur propre u_k de B par:

$$\tilde{u_k} = Du_k.$$

Les résultats sont présentés figure 10. Cette fois, les résultats sont bien plus robustes aux hétérogénéités de q. Néanmoins, lorsque q_0 est proche de q_1, les performances obtenues sont plus faibles que la méthode précédente. On peut interprété cette perte de précision au fait que les q_i utilisés dans les algorithmes normalisés sont en fait des estimateurs statistiques des valeurs théoriques de ces q_i. Par conséquent, lorsque tous les q_i théoriques sont égaux, la normalisation ajoute une source de bruit aux données qui n'était pas présente dans l'algorithme naïf.

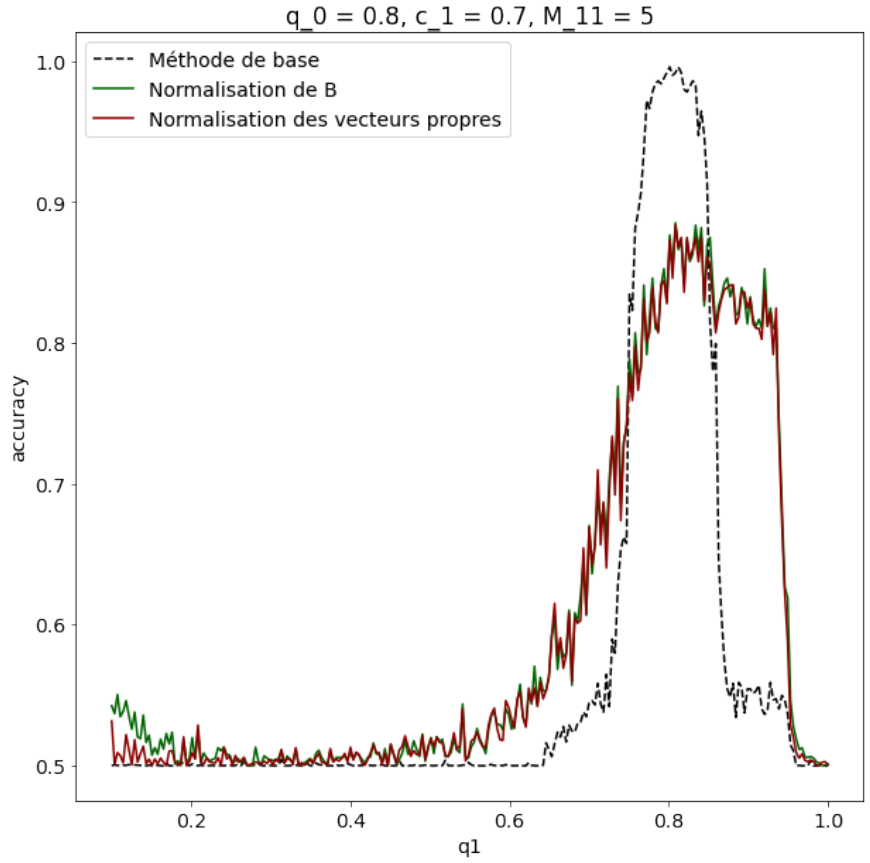


Figure 10: Résultats de les algorithmes de clustering corrigés sur le même modèle que figure 9.