Théorie des Matrices Aléatoires
Détection de Communautés dans les Graphes

Bianca Marin Moreno & Leonardo Martins Bianco

Mars 2022

1 Observations préliminaires

1.1 Décomposition de la matrice d'adjacence

C'est toujours possible de décomposer une matrice comme la somme de son espérance (qui est déterministe) et une matrice aléatoire de "résidus" :

$$\frac{1}{\sqrt{n}}A = \frac{1}{\sqrt{n}}\left(\mathbb{E}\left[A\right] + V\right).$$

Il faut alors démontrer les propriétés de l'énoncé. Etablissons d'abord une notation qui nous sera utile. Notons que

$$\mathbb{E}\left[A\right]_{ij} = \mathbb{E}\left[A_{ij}\right] = q_i q_j C_{ab} \tag{1}$$

Soit $Q = \operatorname{diag}(q)$ et $J \in \mathbb{R}^{n\times K}$ une matrice telle que $J_{ij} = \delta_{i\in\mathcal{C}_j}$. Soit aussi $C \in \mathbb{R}^{K\times K}$ la matrice des C_{ab}. Alors, definissons une matrice "augmentée" $C^A \in \mathbb{R}^{n\times n}$, $C^A := JCJ^t$. Cette matrice est telle que $\left(C^A\right)_{ij} = C_{ab}$ si $i \in \mathcal{C}_a$ et $j \in \mathcal{C}_b$. Alors nous pouvons écrire

$$\frac{1}{\sqrt{n}}A = \frac{1}{\sqrt{n}}\left(QC^AQ + V\right)$$

Nous pouvons aussi développer cela en fonction de la matrice M, car $C = \mathbb{1}_{K\times K} + M/\sqrt{n}$. En utilisant le fait facilement vérifiable que

$$J\mathbb{1}_{K\times K}J^t = \mathbb{1}_{n\times n}, \tag{2}$$

nous arrivons à

$$\frac{1}{\sqrt{n}}A = Q\left(\frac{\mathbb{1}_{n\times n}}{\sqrt{n}} + \frac{JMJ^t}{n}\right)Q + \frac{V}{\sqrt{n}}.$$

Rang de la partie déterministe. A un facteur $1/\sqrt{n}$ près, la partie déterministe vaut QC^AQ. En plus, Q a rang plein (car c'est diagonale avec $q_i > 0$), alors $\mathrm{rk}\left(QC^AQ\right) = \mathrm{rk}\left(C^A\right) = \mathrm{rk}\left(JCJ^t\right)$. Nous avons l'inégalité fondamentale pour toutes deux matrices A, B :

$$\mathrm{rk}\left(AB\right) \leq \min\left\{\mathrm{rk}\left(A\right), \mathrm{rk}\left(B\right)\right\},$$

alors, en utilisant le fait que $\mathrm{rk}\left(J\right) = \mathrm{rk}\left(J^t\right)$,

$$\mathrm{rk}\left(JCJ^t\right) \leq \min\left\{\mathrm{rk}\left(J\right), \mathrm{rk}\left(C\right)\right\}. \tag{3}$$

Or, les colonnes de J représentent les clusters du graphe. En particulier, si chaque point appartient à un seul cluster, alors une colonne n'est pas multiple d'autre et on voit que J a rang plein. Comme $J \in \mathbb{R}^{n\times K}$, alors en assumant que $n \geq K$, alors $\mathrm{rk}\left(J\right) = K$. Finalement, $C \in \mathbb{R}^{K\times K}$ implique $\mathrm{rk}\left(C\right) \leq K$. Alors l'équation 3 devient

$$\mathrm{rk}\left(JCJ^t\right) \leq K.$$

d'où le rang de la partie déterministe de $A/\sqrt{n}$ est au plus K.

Analyse de la partie aléatoire. Comme les entrées A_{ij} étaient indépendantes, alors $V = A - \mathbb{E}\left[A\right]$ a entrées indépendantes aussi, cas substraire des constantes ne modifie pas la dépendance des entrées. Nous vérifions facilement que V a entrées centrées :

$$\mathbb{E}\left[V_{ij}\right] = \mathbb{E}\left[A_{ij} - \mathbb{E}\left[A_{ij}\right]\right] = 0.$$

Une analyse similaire peut être réalisée pour la variance :

$$\mathbb{E}\left[V_{ij}^2\right] = \mathbb{E}\left[\left(A_{ij} - \mathbb{E}\left[A_{ij}\right]\right)^2\right] = \mathrm{Var}\left(A_{ij}\right) = q_iq_jC_{ab}\left(1 - q_iq_jC_{ab}\right) \tag{4}$$

Il nous sera utile (pour prendre la limite $n \to \infty$) de noter que cette équation peut être écrite, après une simple expansion des termes, comme

$$\mathbb{E}\left[V_{ij}^2\right] = q_iq_j\left(1 - q_iq_j\right) + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right).$$

1.2 Décomposition de la matrice de modularité

Etant donnée la décomposition de la matrice d'adjacence, on obtient une décomposition de la matrice de modularité $B/\sqrt{n} = A - qq^t$. Notons que $qq^t = Q\mathbb{1}_{n\times n}Q$, alors

$$\frac{B}{\sqrt{n}} = \frac{Q\left(C^A - \mathbb{1}_{n\times n}\right)Q}{\sqrt{n}} + \frac{V}{\sqrt{n}}.$$

Cette même expression peut être écrite en termes de M :

$$\frac{B}{\sqrt{n}} = Q\left(\frac{J\mathbb{1}_{K\times K}J^t}{\sqrt{n}} + \frac{JMJ^t}{n} - \frac{\mathbb{1}_{n\times n}}{\sqrt{n}}\right)Q + \frac{V}{\sqrt{n}}$$

En utilisant l'équation 2, on obtient

$$\frac{B}{\sqrt{n}} = \frac{QJMJ^tQ}{n} + \frac{V}{\sqrt{n}}. \tag{5}$$

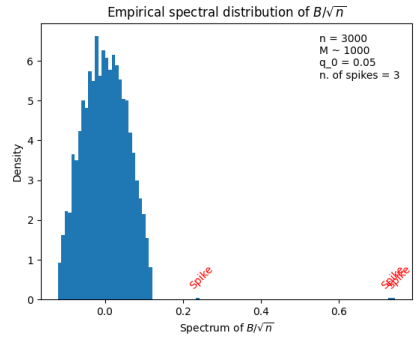
Cette équation nous permet de faire l'interprétation suivante : la matrice M représente la densité des clusters. Si une entrée $M_{ii} \gg 1$, alors le cluster i est très dense et donc "plus facile" à identifier. Comme dans les spiked models, cela peut être interprété comme un signal fort. Par contre, si $M_{ii} \ll 1$, alors le cluster i n'est pas trop dense et donc plus difficile à détecter. En analogie aux spiked models, cela serait un signal faible. La partie aléatoire $V/\sqrt{n}$, après due normalisation, a son spectre tendant vers une loi du semi-cercle, et cela représenterait du bruit.

1.3 Représentation graphique des valeurs et vecteurs propres

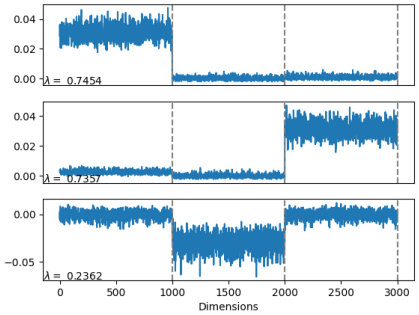
Représentons graphiquement le spectre de $B/\sqrt{n}$ pour des différents choix de M et q_i.

Premier cas. Nous prenons $q_i = q_0 = 0.05$ pour tout i. Cela signifie que tout noeud dans le graphe a le même prior de connectivité. On dénote par $M \sim \mu$ le fait d'avoir $M_{ii} \sim U(0, \mu), \forall i$. On prendra $n = 3000$ et $M \sim 100, 500, 1000$. Cela correspond à avoir l'ordre de M étant d'ordre égale à $\sqrt{n}$, entre $\sqrt{n}$ et n, égale à n, respectivement. On prendra aussi des communautés symmétriques, c'est-à-dire, le nombre de points dans chaque communauté sera le même en moyenne. Les spectres trouvés ainsi que leurs vecteurs propres sont organisés à la Figure 1. Nous observons que quand M est petite, alors le spectre n'a pas de spikes et ressemble à une loi du semi-cercle. Cela arrive car dans ce cas la partie aléatoire de $A/\sqrt{n}$ sera une matrice de Wigner. Lorsque M croît, on observe le nombre de spikes augmenter. Le nombre de valeurs propres isolées va jusqu'à trois, ce qui est d'accord avec l'observation d'avoir le rang de la partie déterministe de A inférieur ou égal à $K = 3$. Nous observons que quand M est suffisament grand pour que le spectre ait des spikes, alors les vecteurs propres s'approchent des vecteurs canoniques. Quand l'ordre de M diminut, alors les vecteurs propres deviennent moins distingués.

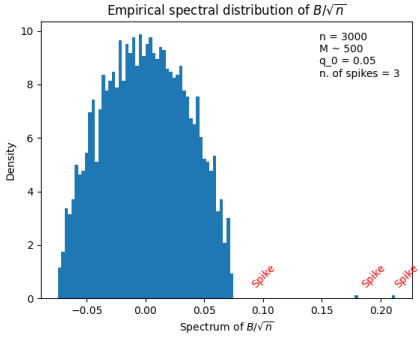
Deuxième cas. Maintenant nous considérons q_i uniforme autour de $q_0 = 0.5$, avec une fênetre de taille 0.25. Nous avons testé les valeurs $M \sim 1, 10, 50$ dans ce cas.



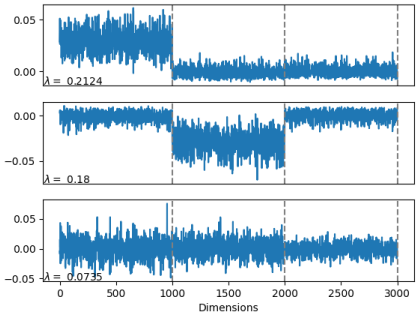
(a) $M \sim 1000$ (eigenvalues)



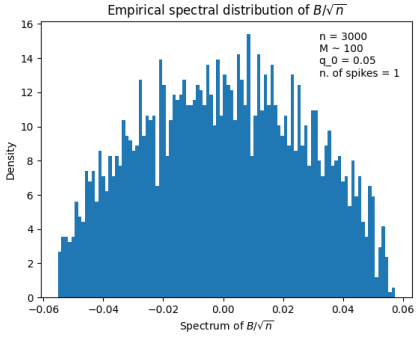
(b) $M \sim 1000$ (eigenvectors)



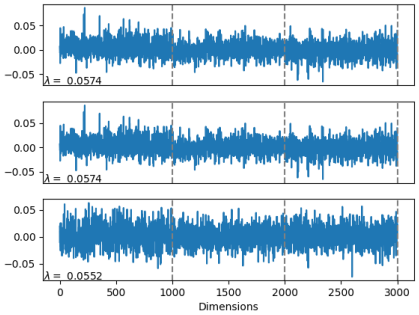
(c) $M \sim 500$ (eigenvalues)



(d) $M \sim 500$ (eigenvectors)



(e) $M \sim 100$ (eigenvalues)



(f) $M \sim 100$ (eigenvectors)

FIGURE 1 – Spectre et vecteurs propres dans le cas homogène pour des différentes valeurs de M

Le spectre observé, ainsi que les vecteurs propres associés, sont montrés à la Figure 2. Comme avant, M plus grand implique des spikes et des vecteurs propres plus définis et proches des vecteurs canoniques des classes.

Troisième cas. Maintenant nous considérons $q_i \in \left\{q^{(1)}, q^{(2)}\right\}$ pour deux valeurs très différentes : $q^{(1)} = 0.5$ et $q^{(2)} = 0.05$. Nous avons testé les valeurs $M \sim 10, 50, 100$ dans ce cas. Le spectre observé est montré à la Figure 3. Curieusement, la distribution limite ne semble pas être une loi du semi-cercle. Les vecteurs propres pour M grand oscillent autour de zéro dehors la classe correspondante à eux, mais sur la partie de la classe correcte ils sont très bruités.

2 Cas Homogène

Avant de commencer, notons que comme M est $\mathcal{O}(1)$ vis-à-vis n, alors $C_{ab} \to 1$ et par l'équation 4 nous avons (déjà utilisant $q_i = q_0 \forall i$) :

$$\mathrm{Var}\left(\frac{V}{\sqrt{n}}\right) = \frac{\mathrm{Var}\left(V\right)}{n} \to \frac{q_0^2\left(1 - q_0^2\right)}{n}$$

C'est-à-dire, pour utiliser la formule donnée pour la densité du semi-cercle, nous devons normaliser $V/\sqrt{n}$ par $\sigma = q_0\sqrt{1 - q_0^2}$.

2.1 Condition d'existence de valeurs propres isolées

Pour déterminer une condition d'existence de valeurs propres isolées, nous procédons comme dans le cours. Nous faisons l'expansion de l'équation caractéristique

$$\det\left(\frac{B}{\sqrt{n}} - \lambda I_n\right) = 0.$$

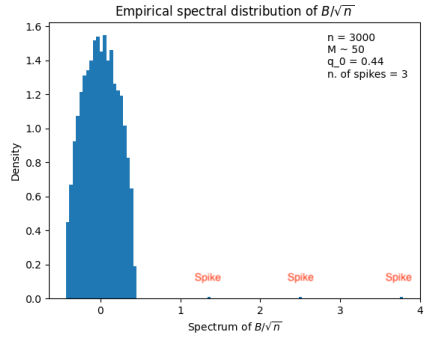
Développons cette expression en utilisant l'équation 5 :

$$\det\left(\frac{q_0^2 JMJ^t}{n} + \frac{V}{\sqrt{n}} - \lambda I_n\right) = 0$$

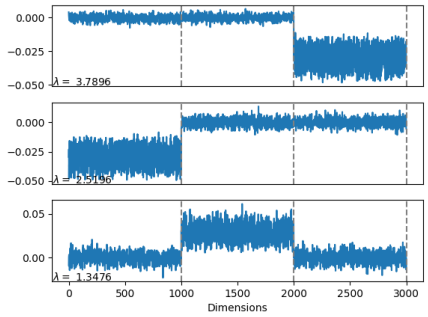
$$\therefore \quad \det\left(\frac{V}{\sqrt{n}} - \lambda I_n\right) \det\left(I_n + \frac{q_0^2 JMJ^t}{n}\left(\frac{V}{\sqrt{n}} - \lambda I_n\right)^{-1}\right) = 0.$$

Maintenant, en utilisant l'identité de Sylvester,

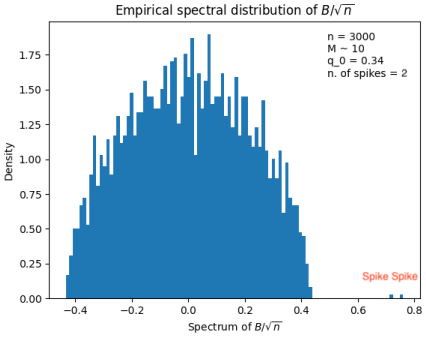
$$\det\left(\frac{V}{\sqrt{n}} - \lambda I_n\right) \det\left(I_n + \frac{q_0^2}{n} J^t\left(\frac{V}{\sqrt{n}} - \lambda I_n\right)^{-1} JM\right) = 0.$$



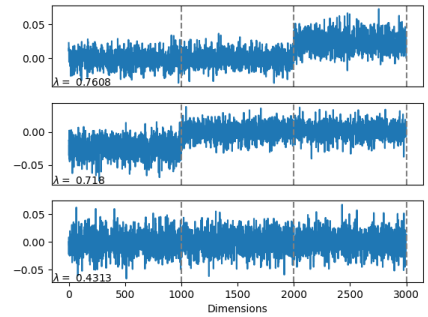
(a) $M \sim 50$ (eigenvalues)



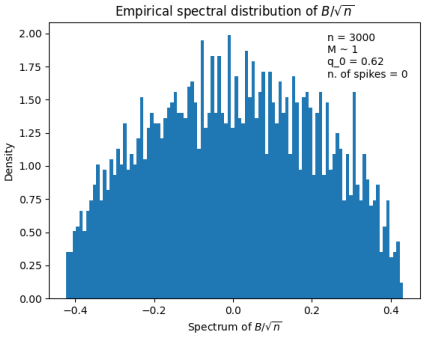
(b) $M \sim 50$ (eigenvectors)



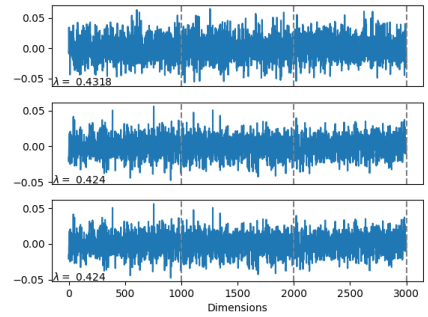
(c) $M \sim 10$ (eigenvalues)



(d) $M \sim 10$ (eigenvectors)

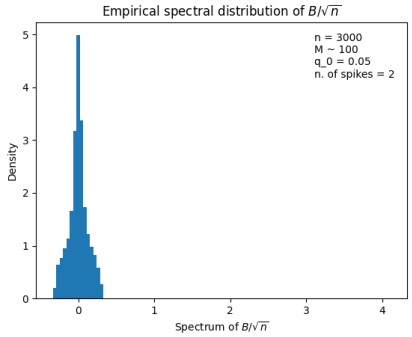


(e) $M \sim 1$ (eigenvalues)

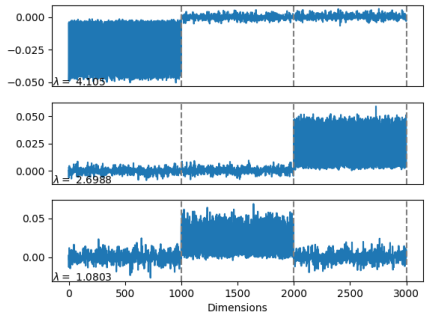


(f) $M \sim 1$ (eigenvectors)

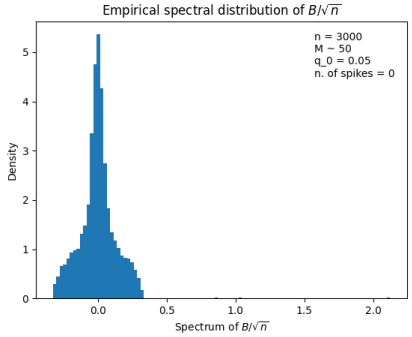
FIGURE 2 – Spectre dans le cas $q_i \sim U\,(0.25, 0.75)$, pour des différentes valeurs de M



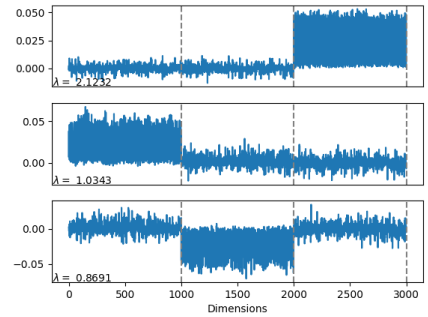
(a) $M \sim 100$ (eigenvalues)



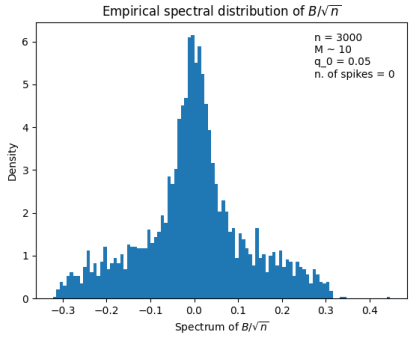
(b) $M \sim 100$ (eigenvectors)



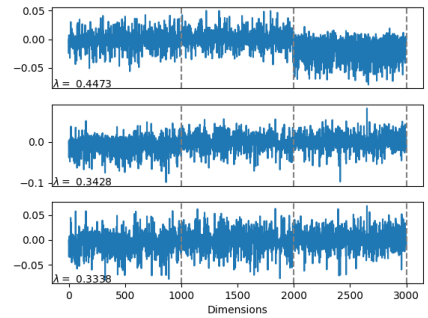
(c) $M \sim 50$ (eigenvalues)



(d) $M \sim 50$ (eigenvectors)



(e) $M \sim 10$ (eigenvalues)



(f) $M \sim 10$ (eigenvectors)

FIGURE 3 – Spectre dans le cas $q_i \in \{0.05, 0.5\}$, pour des différentes valeurs de M

C'est l'heure d'utiliser l'hypothèse de la convergence vers la transformée de Stieltjes du semi-cercle, en faisant attention au fait de normaliser V :

$$\begin{aligned} J^t\left(\frac{V}{\sqrt{n}} - \lambda I_n\right) &= J^t\left(\frac{\sigma V}{\sigma\sqrt{n}} - \sigma\frac{\lambda}{\sigma} I_n\right) \\ &= \frac{1}{\sigma} J^t\left(\frac{V}{\sigma\sqrt{n}} - \frac{\lambda}{\sigma} I_n\right)^{-1} J \\ &\to \frac{1}{\sigma} g_{sc}\left(\frac{\lambda}{\sigma}\right) J^t J \end{aligned} \tag{6}$$

D'où, en utilisant aussi le fait que le premier facteur n'aura pas de valeurs propres dehors le bulk (en normalisant V on aura convergence p.s. vers la loi du semi-cercle), on conclut (en utilisant $j_a^t j_a / n \to c_a$)

$$\prod_{i=1}^{K}\left(1 + \frac{q_0^2}{\sigma} M_{ii} c_i g_{sc}\left(\frac{\lambda}{\sigma}\right)\right) = 0.$$

Cherchons une condition pour que l'un des termes de ce produit, disons le i-ème, devienne zéro :

$$\left(1 + \frac{q_0^2}{\sigma} M_{ii} c_i g_{sc}\left(\frac{\lambda}{\sigma}\right)\right) = 0 \implies g_{sc}\left(\frac{\lambda}{\sigma}\right) = -\frac{1}{\frac{q_0^2 M_{ii} c_i}{\sigma}}. \tag{7}$$

Or, pour $\lambda = 2\sigma$ nous avons

$$g_{sc}(2) = -\frac{1}{2\pi}\int_{-2}^{2} \frac{\sqrt{4-x^2}}{2-x}\, dx = -\frac{1}{2\pi}\int_{-2}^{2} \sqrt{\frac{2+x}{2-x}}\, dx = -1.$$

En plus, $g_s c$ est croissante et $\lim_{\lambda\to+\infty} g_{sc}(\lambda) = 0$. Ainsi, pour que l'équation 7 ait une solution, il faut avoir

$$-1 \le -\frac{\sigma}{q_0^2 M_{ii} c_i} \le 0 \implies M_{ii} c_i \ge \frac{\sigma}{q_0^2} = \frac{\sqrt{1-q_0^2}}{q_0}.$$

2.2 Valeurs asymptotiques des valeurs propres isolées

Substituons l'expression trouvée à l'équation 7 dans l'équation canonique associée à g_{sc} :

$$-\frac{1}{\frac{q_0^2 M_{ii} c_i}{\sigma}} = -\frac{1}{\frac{\lambda}{\sigma} - \frac{1}{\frac{q_0^2 M_{ii} c_i}{\sigma}}} \implies \lambda = q_0^2 M_{ii} c_i + \frac{1-q_0^2}{M_{ii} c_i}. \tag{8}$$

Vérifions numériquement cette expression. Nous prenons un graphe de taille $n = 4000$ avec $k = 3$ communautés, et $q_0 = 0.05$. Le spectre est illustré dans la Figure 4.

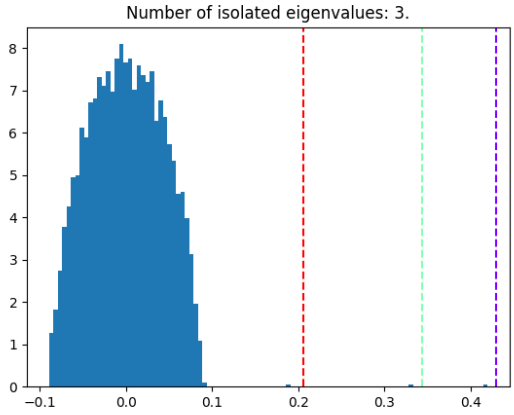


FIGURE 4 – Vérification numérique de l'expression des valeurs propres asymptotiques

2.3 Alignements asymptotiques

On determinera l'alignement asymptotique entre $\hat{u}_m$, un vecteur propre (empirique) isolé de $B/\sqrt{n}$, et le vecteur canonique j_a. Pour faire cela, nous évaluerons l'intégrale

$$\frac{1}{2\pi i} \oint_{\Gamma_m} \frac{1}{n_a} j_a^t \left(\frac{B}{\sqrt{n}} - z I_n \right)^{-1} j_a \, dz,$$

où le contour Γ_m est un cercle autour la limite asymptotique ρ_m pour la valeur propre isolée λ_m de $B/\sqrt{n}$ et sans aucun autre valeur propre à son intérieur. Notons d'abord que

$$\left(\frac{B}{\sqrt{n}} - z I_n \right)^{-1} = \sum_i \frac{\hat{u}_i \hat{u}_i^t}{\lambda_i - z}.$$

Alors, pour tous vecteurs a et b,

$$a^t \hat{u}_m \hat{u}_m^t b = -\frac{1}{2\pi i} \oint_{\Gamma_m} a^t \left(\frac{B}{\sqrt{n}} - z I_n \right)^{-1} b \, dz.$$

En particulier pour les vecteurs canoniques,

$$\frac{1}{n_a} j_a^t \hat{u}_m \hat{u}_m^t j_a = -\frac{1}{2\pi i} \oint_{\Gamma_m} \frac{1}{n_a} j_a^t \left(\frac{B}{\sqrt{n}} - z I_n \right)^{-1} j_a \, dz$$

Utilisons la formule d'inversion de Woodbury pour le terme inverse :

$$\begin{aligned}\left(\frac{B}{\sqrt{n}}-\lambda I_n\right)^{-1} &= \left(\frac{q_0^2 JMJ^t}{n}+\left(\frac{V}{\sqrt{n}}-zI_n\right)\right)^{-1}\\ &= Q(z)-Q(z)J\tilde{M}\left(I+JQ(z)J\tilde{M}\right)^{-1}JQ(z),\end{aligned}$$

où $Q(z)=\left(V/\sqrt{n}-zI_n\right)^{-1}$, et $\tilde{M}=q_0^2/nM$. Nous pouvons facilement manipuler cette expression à la forme

$$Q(z)-Q(z)J\left(\tilde{M}^{-1}+J^tQ(z)J\right)^{-1}J^tQ(z),$$

qui sera pratique car nous pourrons utiliser la convergence isotrope : si l'on fait attention à normaliser V par σ, nous savons que $J^tQ(z)J \to (1/\sigma)\, g_{sc}(z/\sigma)J^tJ$ (voir l'équation 6). En multipliant par $1/n_a$, j_a^t à gauche et par j_a à droite, nous avons

$$\frac{j_a^tQ(z)j_a}{n_a}-\frac{1}{n_a}j_a^tQ(z)J\left(\tilde{M}^{-1}+J^tQ(z)J\right)^{-1}J^tQ(z)j_a.$$

En intégrant cette expression, le premier terme devient zéro (car il est analytique à l'intérieur du contour), et l'autre devient (après une manipulation facile de multiplier et diviser par n_a à l'intérieur du terme inverse) :

$$\frac{1}{2\pi i}\oint_{\Gamma_m}\frac{1}{n_a^2}j_a^tQ(z)J\left(\frac{nM^{-1}}{n_aq_0^2}+\frac{J^tQ(z)J}{n_a}\right)^{-1}J^tQ(z)j_a\,dz \tag{9}$$

Or, analysons les pôles du terme intégré. Quand $n\to\infty$, il y aura une singularité quand quelque terme de la diagonale de l'inverse a une singularité à l'intérieur de Γ_m, c'est à dire, quand

$$\kappa_i(z)=\left(\frac{1}{c_aM_{ii}q_0^2}+\frac{1}{\sigma}g\left(\frac{z}{\sigma}\right)\frac{c_i}{c_a}\right)^{-1}=\frac{1}{h_i(z)}$$

a une singularité à l'intérieur de Γ_m. Nous pouvons vérifier en substituant l'équation 7 que cette singularité existera si, et seulement si, $i=m$ (en assumant que $M_{mm}\neq M_{ii}$ ou $c_i\neq c_m$). Analysons son ordre et résidu.

Montrons que ρ_m est un zéro de premier ordre pour $h_m(z)$. Or, $h_m(\rho_m)=0$ par la définition de ρ_m à l'équation 7, et $h_m'(\rho_m)\propto g_{sc}'(\rho_m/\sigma)\neq 0$ car $g_{sc}'(z)=g_{sc}^2(z)/\left(1-g_{sc}^2(z)\right)$ et aussi par l'équation 7 nous avons $g_{sc}(\rho_m/\sigma)\neq 0$. Alors ρ_m est un zéro de premier ordre pour $h_m(z)$ et donc un pôle de premier ordre pour $\kappa_m(z)$. Son résidu est alors $\lim_{z\to\rho_m}(z-\rho_m)\,\kappa_m(z)$. Assumons encore que $h_m(z)=(z-\rho_m)\,f_m(z)$ pour quelque fonction $f_m(z)$ avec $f_m(\rho_m)\neq 0$. Alors

$$h_m'(z)=f_m(z)+(z-\rho_m)\,f_m'(z)\implies f_m(z)=h_m'(z)-(z-\rho_m)\,f_m(z),$$

d'où, en substituant dans l'expression pour le résidu,

$$\lim_{z\to\rho_m}(z-\rho_m)\,\kappa_m(z)=\frac{1}{h'_m(\rho_m)}=c_a c_m\left(q_0^2 M_{mm}\right)^2-\sigma^2\frac{c_a}{c_m}=R_m,$$

où nous avons défini R_m. La matrice inverse de l'équation 9 devient alors $e_m R_m e_m^t$, où e_m est le m-ème vecteur canonique. Ainsi, l'intégrale de l'équation 9 devient, en utilisant le fait que $Je_m = j_m$:

$$\frac{1}{n_a^2}j_a^t Q(\rho_m)j_m R_m j_m^t Q(\rho_m)j_a\,dz\to\frac{1}{n_a^2}\frac{1}{\sigma^2}g_{sc}^2\left(\frac{\rho_m}{\sigma}\right)\langle j_m,j_a\rangle^2.$$

En utilisant le fait que $\langle j_m, j_a\rangle^2 = n_a^2\delta_{a=m}$ et en substituant l'expression pour g_{sc}, nous concluons

$$\frac{1}{n_a}j_a^t\hat{u}_m\hat{u}_m^t j_a=\delta_{a=m}\left(1-\frac{\sigma^2}{\left(c_m q_0^2 M_{mm}^2\right)^2}\right).$$

2.4 Vérification numérique

Vérifions numériquement ces alignements. Nous avons pris n allant de 100 jusqu'à 2000 pour une matrice $M \sim 100$ fixée. Le résultat est affiché à la Figure 5. Nous observons qu'une des convergences était lente. Cela se justifie quand on regarde les valeurs de M : $\mathrm{diag}\,(M) = (237.8, 785.7, 655.0)$. Alors la convergence lente probablement correspond à la communauté avec la valeur plus petite de M.

2.5 Proposition d'un algorithme de détection de communautés

Nous avons montré que les vecteurs propres liés aux valeurs propres isolées sont alignés avec les vecteurs des classes. Ainsi, une idée d'algorithme de détection de communautés consisterait à trouver les vecteurs propres isolés de la matrice $B/\sqrt{n}$ et à utiliser un algorithme de clustering de K classes, K-means par exemple, pour clusteriser les n lignes de la matrice des vecteurs propres isolés. Ainsi, l'algorithme proposé de détection de communauté basé sur le spectre consiste à

1. Identifier les valeurs propres isolées du spectre de $\frac{B}{\sqrt{n}}$. Extraire les vecteurs propres correspondants $V = [v_1, ..., v_m] \in \mathbb{R}^{n\times m}$ avec $m < K$.
2. Effectuer un K-means pour K classes sur les vecteurs lignes de V.

Un problème que nous pourrions rencontrer pour évaluer la performance de cet algorithme est qu'il peut produire des classes avec des labels permutées par rapport aux originales, donc une simple vérification de la précision ne serait pas efficace. Nous proposons deux façons de résoudre ce problème : la première consiste à prendre la plus grande

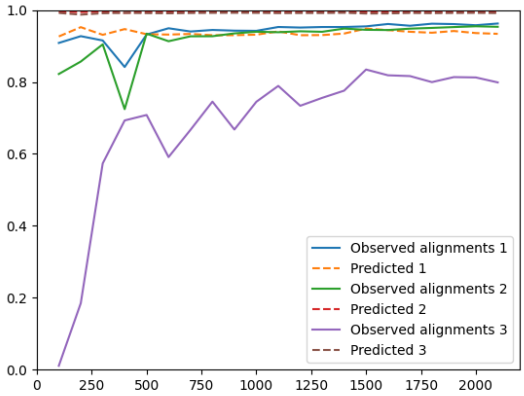


FIGURE 5 – Vérification numérique de l'alignment asymptotique entre les trois vecteurs propres associés aux spikes et les vecteurs canoniques des classes

précision pour toutes les permutations possibles des labels. La seconde consiste à utiliser l'indice de Rand, une fonction qui mesure la similarité des deux affectations en ignorant les permutations. Dans la suite, nous avons décidé d'utiliser l'indice de Rand pour mesurer les performances. Dans ce cas homogéne, on arrive a avoir en moyenne un indice de Rand égal à 0.92 pour $q_0 = 0.1$, $n = 3000$ et $M \sim 150$.

3 Cas Heterogène

3.1 Connectivités tirées indépendamment d'une loi bimodale

Nous autorisons maintenant différents degrés intrinsèques pour les nœuds du graphe. Les vecteurs propres de $\mathbb{E}[A]$ ne seront plus des combinaisons linéaires de $j_1, ..., j_K$, mais ils sont maintenant déformés par les poids $q_1, ..., q_n$. La variance de A est

$$\text{Var}\,(A_{i,j}) = q_i q_j (1 - q_i q_j) + O(n^{\frac{1}{2}}).$$

Ainsi, A est une matrice à entrées indépendantes, de moyenne nulle, et de variance $q_i q_j (1 - q_i q_j)$. Ceci implique que sa mesure spectrale, et celle de $B/\sqrt{n}$, est celle d'une matrice de Wigner déformée.

Dans la figure 3 nous avons tracé la densité du spectre de $B/\sqrt{n}$ dans un cas avec $q_1 = 0.05$ et $q_2 = 0.5$, différentes valeurs de M, et $n = 1000$. Alors on voit que le spectre est plus étalé qu'un demi-cercle dans le cas où les q_i sont tirées indépendamment d'une loi bimodale. En fonction de l'étalement des valeurs propres, nous pouvons avoir

plus ou moins de transition de phase pour les valeurs propres isolées dues aux communautés. Nous voyons que selon M nous avons un nombre différent de valeurs propres isolées. La mesure de la performance de notre algorithme de clustering basé sur le spectre dans ce cas nous donne un Rand index de ≈ 0.61. Ainsi, notre algorithme ne fonctionne plus.

3.2 Amélioration par renormalisation

3.2.1 Renormalisation de B

La première idée consiste à étaler B pour que son spectre soit plus proche d'un demi-cercle. En se basant sur la discussion du Chapitre 7 dans [1], en particulier le lemme 7.1, on remarque que pour des q_i pas trop grands, $q_i q_j(1 - q_i q_j) \simeq q_i q_j$ et $\frac{d_i}{\sqrt{\mathbf{d}^T \mathbb{1}_n}} \simeq q_i$, où $\mathbf{d}$ est le vecteur des degré du graphe. Ainsi, si D est la matrice diagonale des degrés du graphe, nous normalisons la matrice B en la multipliant par D^{-1} des deux côtés comme dans l'équation 10.

$$L = \frac{d^T \mathbb{1}_n}{\sqrt{n}} D^{-1} B D^{-1} \tag{10}$$

Ainsi, l'algorithme avec la renormalisation de B est alors :

1. Identifier les valeurs propres isolées du spectre de L dans l'équation 10, où $d^T \mathbb{1}_n$ est un facteur de normalisation. Extraire les vecteurs propres correspondants $V = [v_1, ..., v_m] \in \mathbb{R}^{n \times m}$ avec $m < K$.
2. Effectuer un K-means pour K classes sur les vecteurs lignes de V.

3.2.2 Renormalisation des vecteurs propres de B

Au lieu d'appliquer une renormalisation à la matrice B en la multipliant des deux côtés par D^{-1}, nous pouvons la multiplier par $D^{-\alpha}$ comme cela est fait dans l'équation 11 pour un certain α que nous aurions encore besoin d'optimiser.

$$L_\alpha = \frac{(d^T \mathbb{1}_n)^\alpha}{\sqrt{n}} D^{-\alpha} B D^{-\alpha} \tag{11}$$

D'après les discussions menées dans le Chapitre 7 de [1], on voit que les vecteurs propres dominants de L_α sont alignés avec une combinaison linéaire des vecteurs $Q^{1-\alpha} j_a$ pour $a = 1, ..., K$. Ainsi, nous devons multiplier les vecteurs propres de L_α par $Q^{\alpha-1}$ avant de réaliser le K-means. Comme dans la pratique il se peut que nous ne connaissions pas Q, nous normalisons plutôt les vecteurs propres en utilisant $D^{\alpha-1}$. Ainsi, l'algorithme avec la renormalisation des vecteurs propres de B est alors :

1. Identifier les valeurs propres isolées du spectre de L_α dans l'équation 11, pour un valeur de α. Extraire les vecteurs propres correspondants $V = [v_1, ..., v_m] \in \mathbb{R}^{n \times m}$ avec $m < K$.
2. Effectuer un K-means pour K classes sur les vecteurs lignes de $D^{\alpha-1}V$.

3.2.3 Résultats

Le tableau 1 compare les Rand indices obtenus en faisant des simulations avec le premier algorithme proposé pour le clustering (original), l'algorithme avec renormalisation de B (renorm. B), et celui avec renormalisation des vecteurs propres de B (renorm. eig. B) avec $\alpha = 0.5$. Nous prenons $q_1 = 0,05$, $q_2 = 0,5$, $n = 3000$ et $M \sim 150$. Nous voyons que la renormalisation de B et le vecteur propre de B augmentent les performances de l'algorithme dans le cas où le vecteur de connectivité est dessiné indépendamment par une loi bimodale.

Algo.	Adjus. Rand Index
original	0.61
renom. B	0.74
renom. eig. B	0.93

TABLE 1 – Comparaison des Rand indices pour des simulations utilisant l'algorithme original, et celles avec la renormalisation de B et la renormalisation des vecteurs propres de B pour $\alpha = 0.5$. Les paramètres sont $n = 3000$, $q_1 = 0.05$, $q_2 = 0.5$, et $M \sim 150$.

Références

[1] Zhenyu Liao Romain Couillet. *Random Matrix Theory for Machine Learning*, volume 1. 2022.