Random Matrix Theory

Essakine Amer
amer.essakine@ens-paris-saclay.fr

March 17, 2025

For readability, all the figures are in the annex and are referenced in their corresponding questions

Preliminary observations

1. We can write that :

$$\frac{1}{\sqrt{n}}A = \frac{1}{\sqrt{n}}\mathbb{E}[A] + \frac{1}{\sqrt{n}}(A - \mathbb{E}[A]) = G + Y$$

Therefore, to conclude we need to show that conditionally on q_i's that M has rank at most K and W a random matrix with independent zero-mean entries with a variance profiles. We have that conditionally on q_i's:

$$E[A_{i,j}|q] = q_i q_j \mathbb{E}[C_{a,b}] = \sum_{a=1}^{K}\sum_{b=1}^{K} q_i q_j \mathbb{E}[C_{a,b}] \mathbb{1}_{\{i\in\mathcal{C}_a\}} \mathbb{1}_{\{j\in\mathcal{C}_b\}}$$

Following the last equality, we consider the matrix $C = (\mathbb{E}[C_{a,b}])_{1\leq a,b\leq K} \in \mathcal{M}_K$ and the matrix $J \in \mathcal{M}_{n,K}$, which encodes the individuals in each community. More specifically, for $1 \leq a \leq K$ and $1 \leq i \leq n$, the entry $J_{a,i}$ is defined as follows:

$$J_{a,i} = \begin{cases} 1, & \text{if } i \in \mathcal{C}_a, \\ 0, & \text{otherwise.} \end{cases}$$

Finally, let $Q = diag(q)$, Then we have that $\frac{1}{\sqrt{n}}\mathbb{E}[A] = QJCJ^*Q$. It follows that

$$rank(\frac{1}{\sqrt{n}}\mathbb{E}[A]) \leq min(rank(C), rank(J)) \leq K$$

Now for the second matrix. It is a random matrix whose entries have a zero mean. They are independent conditionally on q_i's since $A_{i,j}$ are independent conditionally on

q_i's. Now, for the variance :

$$\begin{aligned} Var(Y_{i,j}|q) &= \frac{1}{n}\mathbb{E}[A_{i,j}^2|q] \\ &= \frac{1}{n}q_iq_jC_{a,b}(1-q_iq_jC_{a,b}) \\ &= \frac{1}{n}q_iq_j(1-q_iq_j)+\mathcal{O}(\frac{1}{n\sqrt{n}}) \end{aligned}$$

2. For the matrix B, we have a similar decomposition. Conditionally on the q_i's:

$$\frac{1}{\sqrt{n}}B = \frac{1}{\sqrt{n}}\mathbb{E}[B] + \frac{1}{\sqrt{n}}(B-\mathbb{E}[B]) = H + X$$

   And we have that

$$\frac{1}{\sqrt{n}}\mathbb{E}[B] = \frac{1}{\sqrt{n}}(\mathbb{E}[A]-qq^*) = \frac{1}{\sqrt{n}}(QJCJ^*Q-qq^*) = \frac{1}{n}JMJ^*$$

   Where M is the matrix whose entries are given by $M_{a,b}$, we conclude that the first term is a matrix of rant at most n by the same reasoning as the previous question. The second term is a random matrix with independent zero-mean entries and has the same variance profile as in the case of A.

$$Var(X_{i,j}|q) = \frac{1}{n}q_iq_j(1-q_iq_j)+\mathcal{O}(\frac{1}{n\sqrt{n}})$$

3. We can see the spectrum distribution in figures [1][2][3][4]We observe that the histograms of the eigenvalues of $\frac{1}{\sqrt{n}}B$ resemble the Wigner semicircle law. This is because $\frac{1}{\sqrt{n}}B$ can be decomposed into a low-rank deterministic part (of rank at most K) plus a random matrix with zero-mean entries. Consequently, we expect up to K spikes in the spectrum:

   - **First case** ($q_i = 0.5$): The distribution is very close to the semicircle law. As M increases, the influence of the low-rank part grows, and more spikes separate from the bulk distribution.
   - **Second case** (q_i around 0.5 with small spread 0.15 or larger spread 0.4):
     - With a small spread, the distribution still closely follows the semicircle law.
     - With a larger spread, the spectrum diverges more from the semicircle shape.

     In both scenarios, increasing M leads to more pronounced spikes that move farther from the bulk.
   - **Last case** ($q_i \in \{0.1, 0.9\}$): Here, the distribution departs more significantly from the semicircle shape and appears somewhat split into two parts. As M increases, additional spikes become visible.

   In conclusion, the range of the q_i values primarily affects the shape of the bulk distribution, while M controls the appearance of spikes, with larger M producing more distinct outliers.

4. - As we can see from the visualization of the eigenvectors [1][2][3][4], the spectral community detection algorithm—based on projecting nodes onto the eigenvector space of the matrix $\frac{1}{\sqrt{n}}B$—appears to be more efficient when the q_i values are constant. In this homogeneous setting, the clustering is clearly visible in the eigenvector representation, and the performance improves as the scale of M increases.
   - When the q_i values vary, the strategy is less effective, as indicated by the eigenvector plots. This is likely because each node perturbs the community signal differently, making it more challenging to recover the underlying class structure compared to the case with constant q_i values and uniform perturbation across classes.

Homogeneous Case

1. By the question 2., we can decompose B as :

$$B = H + X_n = \frac{q_0}{n} JMJ^* + X_n$$

Let λ be a real that is not an eigenvalue of X, we can write then :

$$\begin{aligned}
\det(\frac{1}{\sqrt{n}}B - \lambda I_n) &= \det(X - \lambda I_n + \frac{q_0^2}{n} JMJ^*) \\
&= \det(X - \lambda I_n)(I_n + (X - \lambda I_n)^{-1} \frac{q_0^2}{n} JMJ^*) \\
&= \det(X - \lambda I_n) det(I_n + Q(\lambda) \frac{q_0^2}{n} JMJ^*)
\end{aligned}$$

Where Q is the resolvent of X. We now use the Sylvester's determinant identity which states in particular that $\det(I + AB) = \det(I + BA)$, we have

$$\det(\frac{1}{\sqrt{n}}B - \lambda I_n) = \det(X - \lambda I_n) \det(I_K + \frac{q_0^2}{n} MJ^* Q(\lambda) J)$$

Now by the first admitted result in the TP (Wigner's theorem) : Asymptotically, the eigenvalues of X_n are not isolated. If we consider λ an isolated eigenvalue (supposing it exists), then :

$$\det(I_K + \frac{q_0^2}{n} MJ^* Q(\lambda) J) = 0$$

Next, we will use the result on the convergence of the Stieltjes transform. We need to normalize X before, we know that

$$Var(X) = q_0^2(1 - q_0^2)\frac{1}{n} + \mathcal{O}(\frac{1}{n\sqrt{n}})$$

Hence, we apply the result to $\frac{X}{q_0\sqrt{1-q_0^2}}$, to simplify the notation we denote $\sigma = q_0\sqrt{1 - q_0^2}$
We have that :

$$JQ(\lambda)J^* \to \frac{1}{\sigma} g_{sc}(\frac{\lambda}{\sigma}) J^* J$$

Now, we remark that J^*J is a diagonal matrice whose diagonal is exactly $(|\mathcal{C}_1|, ..., |\mathcal{C}_k|)$. Then using the fact that $\frac{|\mathcal{C}_i|}{n} \to c_i$, we finally find that (since M is a diagonal matrix) :

$$\det(I_K + \frac{q_0^2}{n} M J^* Q(\lambda) J) \to \prod_{i=1}^{K} (1 + \frac{c_i M_{i,i} q_0^2}{\sigma} g_{sc}(\frac{\lambda}{\sigma}))$$

To have an isolated eigenvalue, then at least for one $i \in \{1, ..., K\}$ we must have that

$$1 + \frac{c_i M_{i,i} q_0^2}{\sigma} g_{sc}(\frac{\lambda}{\sigma}) = 0 \iff g_{sc}(\frac{\lambda}{\sigma}) = -\frac{\sigma}{c_i M_{i,i} q_0^2}$$

But we know that

$$g_{sc} = \frac{-z + \sqrt{z^2 - 4}}{2}$$

(which can be proven the equation $g_{sc}(z)^2 + z g_{sc}(z) + 1 = 0$ and we take the solution that maps the upper imaginary half plane to itself).
Since, B is symmetric then all its eigenvalues are real, and since $g_{sc}(\mathbb{R}\ [-2, 2]) = [-1, +\infty[$. So the equation, have a solution if and only if (by distinguishing cases accroding to if $M_{i,i}$ is positive or not):

$$\text{There exists } i \in \{1, ..., K\} \text{ such that : } \begin{cases} \sigma \leq q_0^2 c_i M_{i,i} & \text{if } M_{i,i} \geq 0, \\ \sigma \geq q_0^2 c_i |M_{i,i}| & \text{if } M_{i,i} < 0. \end{cases}$$

Or more compactly :

$$(\sigma - q_0^2 c_i |M_{i,i}|) M_{i,i} \leq 0$$

2. Since g_{sc} verify the equation $g = \frac{1}{g+z}$, then we have :

$$\frac{\sigma}{q_0^2 c_i M_{i,i}} = \frac{1}{\frac{\lambda}{\sigma} - \frac{\sigma}{q_0^2 c_i M_{i,i}}}$$

   Solving this yield :

$$\lambda_i = \frac{\sigma^2}{q_0^2 c_i M_{i,i}} + q_0^2 c_i M_{i,i}$$

   In figure [5], we find the numerical verification of the theoretical isolated eigenvalues $(K = 3, N = 1500)$ and M a random diagonal matrix with large entries

3. Let $a \in \{1, .., K\}$ and denote λ its associated isolated eigenvalue (if it exists). Let us consider the function

$$f_n(z) = \frac{1}{n_a} j_a^* (\frac{1}{\sqrt{n}} B - z I_n)^{-1} j_a$$

   By the Woodbury matrix identity

$$\begin{aligned} f_n(z) &= \frac{1}{n_a} j_a^* (Q(\lambda) - Q(\lambda) J (I_K + \frac{q_0^2}{n} M J^* Q(\lambda) J)^{-1} \frac{q_0^2}{n} M J^* Q(\lambda)) j_a \\ &= \frac{1}{n_a} j_a^* Q(\lambda) j_a - \frac{1}{n_a} j_a^* Q(\lambda) J (I_K + \frac{q_0^2}{n} M J^* Q(\lambda) J)^{-1} \frac{q_0^2}{n} M J^* Q(\lambda) j_a \end{aligned}$$

The first term converges pointwise to :

$$\frac{1}{n_a} j_a^* Q(\lambda) j_a \to \frac{1}{\sigma} g_{sc}(\frac{z}{\sigma})$$

For the second term, we first have from the previous question :

$$(I_K + \frac{q_0^2}{n} M J^* Q(\lambda) J)^{-1} M \to diag(\frac{M_{a,a}}{1 + \frac{c_a M_{a,a} q_0^2}{\sigma} g_{sc}(\frac{z}{\sigma})})$$

We also multiplied by M which is a diagonal matrix, next :

$$J^* Q(\lambda) j_a \to \begin{bmatrix} 0 \\ 0 \\ \frac{n_a}{\sigma} g_{sc}\left(\frac{z}{\sigma}\right) \\ 0 \\ 0 \end{bmatrix}$$

Therefore :

$$(I_K + \frac{q_0^2}{n} M J^* Q(\lambda) J)^{-1} M J^* Q(\lambda) j_a \to \begin{bmatrix} 0 \\ 0 \\ \frac{q_0^2}{n} \frac{n_a M_{a,a}}{\sigma + q_0^2 c_a M_{a,a} g_{sc}(\frac{z}{\sigma})} g_{sc}(\frac{z}{\sigma}) \\ 0 \\ 0 \end{bmatrix}$$

And finally :

$$\frac{1}{n_a} j_a^* Q(\lambda) J (I_K + \frac{q_0^2}{n} M J^* Q(\lambda) J)^{-1} M J^* Q(\lambda) j_a \to \frac{q_0^2 c_a M_{a,a}}{\sigma^2 + \sigma q_0^2 c_a M_{a,a} g_{sc}(\frac{z}{\sigma})} g_{sc}(\frac{z}{\sigma})^2$$

Hence regrouping all terms :

$$\begin{aligned} f_n(z) &\to \frac{1}{\sigma} g_{sc}(\frac{z}{\sigma}) - \frac{q_0^2 c_a M_{i,i}}{\sigma^2 + \sigma q_0^2 c_i M_{i,i} g_{sc}(\frac{z}{\sigma})} g_{sc}(\frac{z}{\sigma})^2 \\ &\to \frac{1}{\sigma} g_{sc}(\frac{z}{\sigma}) - \frac{1}{\frac{\sigma}{q_0^2 c_a M_{a,a}} + g_{sc}(\frac{z}{\sigma})} \frac{g_{sc}(\frac{z}{\sigma})^2}{\sigma} \\ &\to \frac{1}{\sigma} g_s c(\frac{z}{\sigma}) - \frac{1}{g_{sc}(\frac{z}{\sigma}) - g_{sc}(\frac{\lambda}{\sigma})} \frac{g_{sc}(\frac{z}{\sigma})^2}{\sigma} \end{aligned}$$

Conisder $\Gamma_a = \{z \in \mathbb{C} \ , |z - \lambda| = \epsilon\}$, where ϵ is small enough such that there are no other eigenvalues in Γ_a. We have that f_n converges point-wise almost surely on Γ_a, and we also have for any $z \in \Gamma_a$:

$$|f(z)| \le \frac{\|j_a\|}{n_a} \frac{1}{|z - \lambda|} \le \frac{1}{\epsilon}$$

Usint the inequality on the resolvant, the function $z \to \frac{1}{\epsilon}$ is integrable on Γ_a. Hence :

$$\begin{aligned}\frac{1}{2\pi i}\oint f_n(z)\,dz &\to \lim_{z\to\lambda}(z-\lambda)\frac{1}{g_{sc}(\frac{z}{\sigma})-g_{sc}(\frac{\lambda}{\sigma})}\frac{g_{sc}(\frac{z}{\sigma})^2}{\sigma} \\ &\to \frac{g_{sc}(\frac{\lambda}{\sigma})^2}{g'_{sc}(\frac{\lambda}{\sigma})}\end{aligned}$$

Where we used the Residue formula on the point-wise limit of f_n, the integral of the first term since the function is analytic on Γ_a, and the second term have one singularity which is λ and of order 1.
We also know that $g'_{sc} = \frac{g_{sc}^2}{1-g_{sc}^2}$ which gives that the integral converges to $1 - g_{sc}(\frac{\lambda}{\sigma})^2$
On the other hand, we can also express f_n as

$$f_n(z) = \frac{1}{n_a}\sum_{i=1}^{n}\frac{j_a^* u_i u_i^* j_a}{\lambda_i - z}$$

Where u_i are the eigenvectors of $\frac{1}{\sqrt{n}}B$. Integrating this equality over Γ_a gives by the Residue theorem :

$$\frac{1}{2\pi i}\oint f_n(z)\,dz = \frac{1}{n_a}(j_a^* u)^2$$

Hence :

$$\frac{1}{n_a}(j_a^* u_a)^2 = 1 - g_{sc}(\frac{\lambda}{\sigma})^2 = 1 - \frac{\sigma^2}{q_0^4 c_a^2 M_{i,i}^2}$$

Finally, if we take $i \neq a \in \{1, ..., K\}$ and that the contour Γ_i we will find in the same way that

$$\frac{1}{n_a}(j_a^* u_i)^2 = 0$$

Hence the final result for the alignments :

$$\frac{1}{n_a}(j_a^* u_i)^2 = \begin{cases} 1 - \frac{\sigma^2}{q_0^4 c_a^2 M_{i,i}^2} & \text{if } i = a \\ 0 & \text{if } i \neq a. \end{cases}$$

4. We find in figure [6] the results of the numerical verication of the alignements as N grows. We can see that the empirical alignements for each community converges to the theoretical one as N goes to infinity in accordance with the result we found in the previous question.

5. Inspired from the previous questions, we propose the following algorithm :

   - Given the indicator matrix J, q_0 and the matrix M generate the scaled modularity matrix $\frac{1}{\sqrt{N}}B$
   - Extract the K eigenvectors $u_1, ..., u_k$ corresponding to the most K isolated eigenvalues of the scaled modularity matrix. Each row of the matrix $[u_1, ..., u_k]$ represents the embedding of a node in a K-dimensional space. This K directions represents where the data is dispersed the best thanks to the values of alignments calculated

- Apply a clustering algorithm on the K-dimensional space defined by the rows of $[u_1, \dots, u_K]$ to partition the nodes into K communities.

A problem that we could encouter when evaluating the performance of the algorithm is that the algorithm may assign labels that are a permutation of the ground truth labels. This label mismatch can lead to misleading performance metrics if the labels are not aligned properly. To resolve this, we can apply the Hungarian algorithm to optimally match the predicted labels with the ground truth before computing the evaluation metrics.

Now to estimate the error, consider an isolated eigenvalue λ. And we want to bound the gap between λ and an eigenvalue inside the support, we have by the proposition 2.13 in the course that :

$$|\lambda - \tilde{\lambda}| \leq \| \frac{q_0^2}{n} JMJ^* \|$$

Where $\tilde{\lambda}$ is an eigenvalue of X which lies eventually inside the semi cirle law support. Now since the entries of $\frac{q_0^2}{n} JMJ^*$ are $\mathcal{O}(\frac{1}{n\sqrt{n}})$, we obtain by bounding the Frobienus norm for example and using the equivalence of norm that the error is:

$$|\lambda - \tilde{\lambda}| = \mathcal{O}(\frac{1}{\sqrt{n}})$$

Heterogeneous Case

1. When the q_i's are heterogeneous, they introduce variability in the eigenvalues and eigenvectors. In particular, nodes with larger q_i's tend to dominate the spectrum, causing the corresponding eigenvectors to be biased towards these nodes. As a result, the extracted spectral embedding does not accurately reflect the true community structure.
   To showcase this, we compared the spectral clustering algorithm with $N = 1500$ and having 5 classes, we also take $q_0 = 0.6$, and in the heterogeneous case (with q_i's taking values randomly between 0.7 and 0.3), the results are in [7][8], and in this table we have the metric evaluation The misclassification rate is calculated after applying the

	Homogeneous	Heterogeneous
Misclassification rate	0.022	0.250
Adjusted Rand Index	0.956	0.557

   Table 1: Clustering performance under homogeneous and heterogeneous q_i's.

   Hungarian algorithm and then matching the true labels and predicted ones. We see that the error increase and the Rand index decrease in the heteregenous case and the model is no longer able to seperate clusters

2. To solve the problems presented in the previous question, we suggest two algorithms

(a) The first algorithm based on the normalization of B in order to migitate the influence of high or low q_i's. We consider the diagonal matrix $D = diag(q_1, ..., q_n)$ and instead of B we use:

$$\tilde{B} = D^{-\frac{1}{2}} B D \frac{1}{2}$$

. This normalization reduces the effect of the variability of the weights q_i's and helps improve the clustering.

(b) The second method consists of computing the isolated eigenvectors of $\frac{1}{\sqrt{n}} B$ then nomralizing them, so instead of clustering the rows of U, we normalize before clustering :

$$\tilde{u}_i = \frac{u_i}{\|u_i\|}$$

This normalization aims to remove the effect of the variability of the weights q_i's from the eigenvectors.

We tested both methods as well as the one in question 4 for $N = 1500$ and having 5 classes, we took M a full matrix taken randomly and scaled by 50 and for q_i's values taken randomly between 0.2 and 0.8, we find the result of the simulation in [9]. We can see that the normalization improves the clusterings :

Method	Misclassification Rate	Adjusted Rand Index
Standard Spectral Clustering	0.303	0.638
Normalization of B	0.047	0.889
Normalization of Eigenvectors	0.030	0.928

Table 2: Performance of spectral clustering methods under heterogeneous q_i's.

Annexe

Preliminary obeservations

Homogeneous Case

0.1 Heterogeneous Case

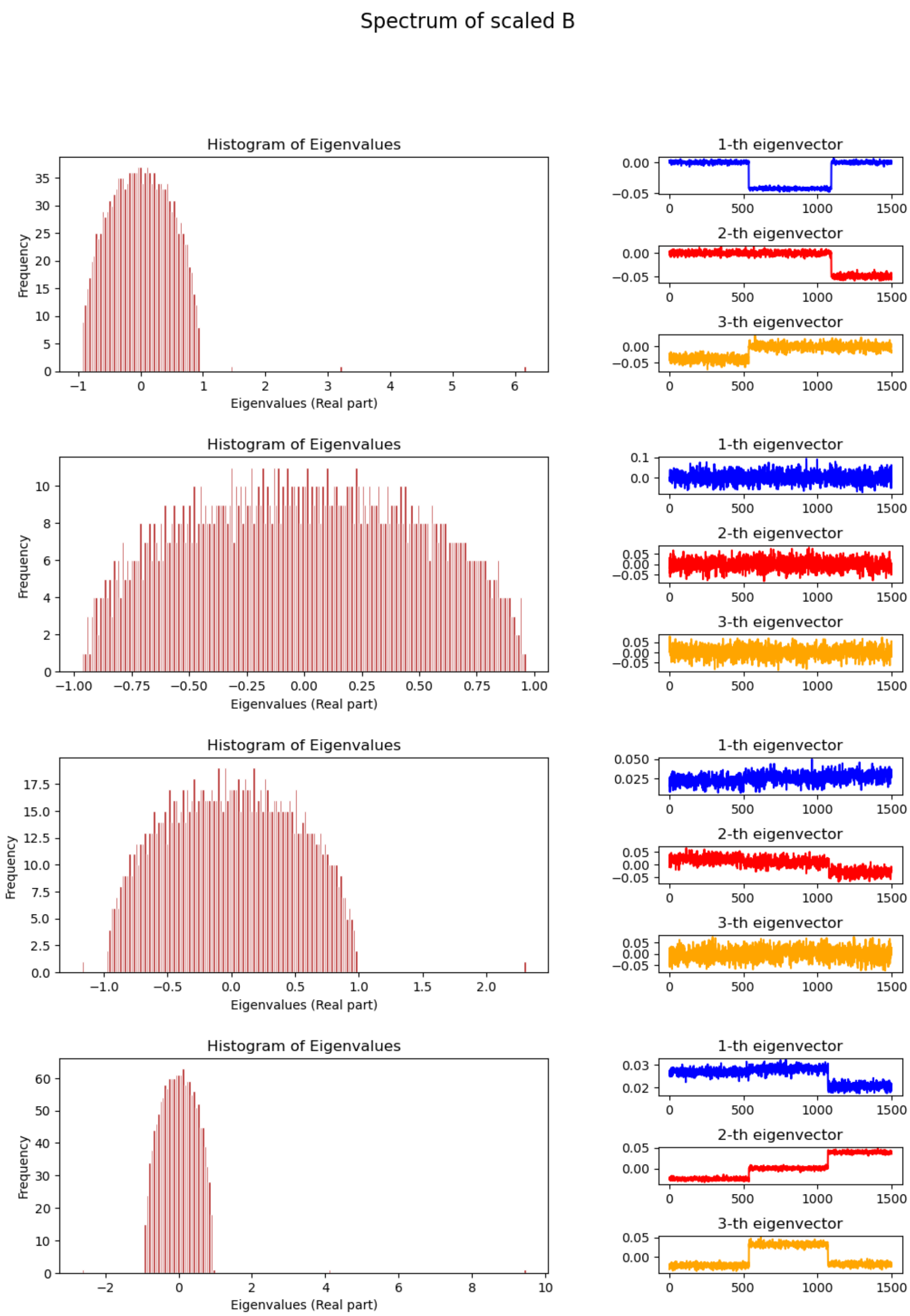

Figure 1: The spectrum distribution and the representation of the first three eigenvectors for the case $q_i = 0.6$

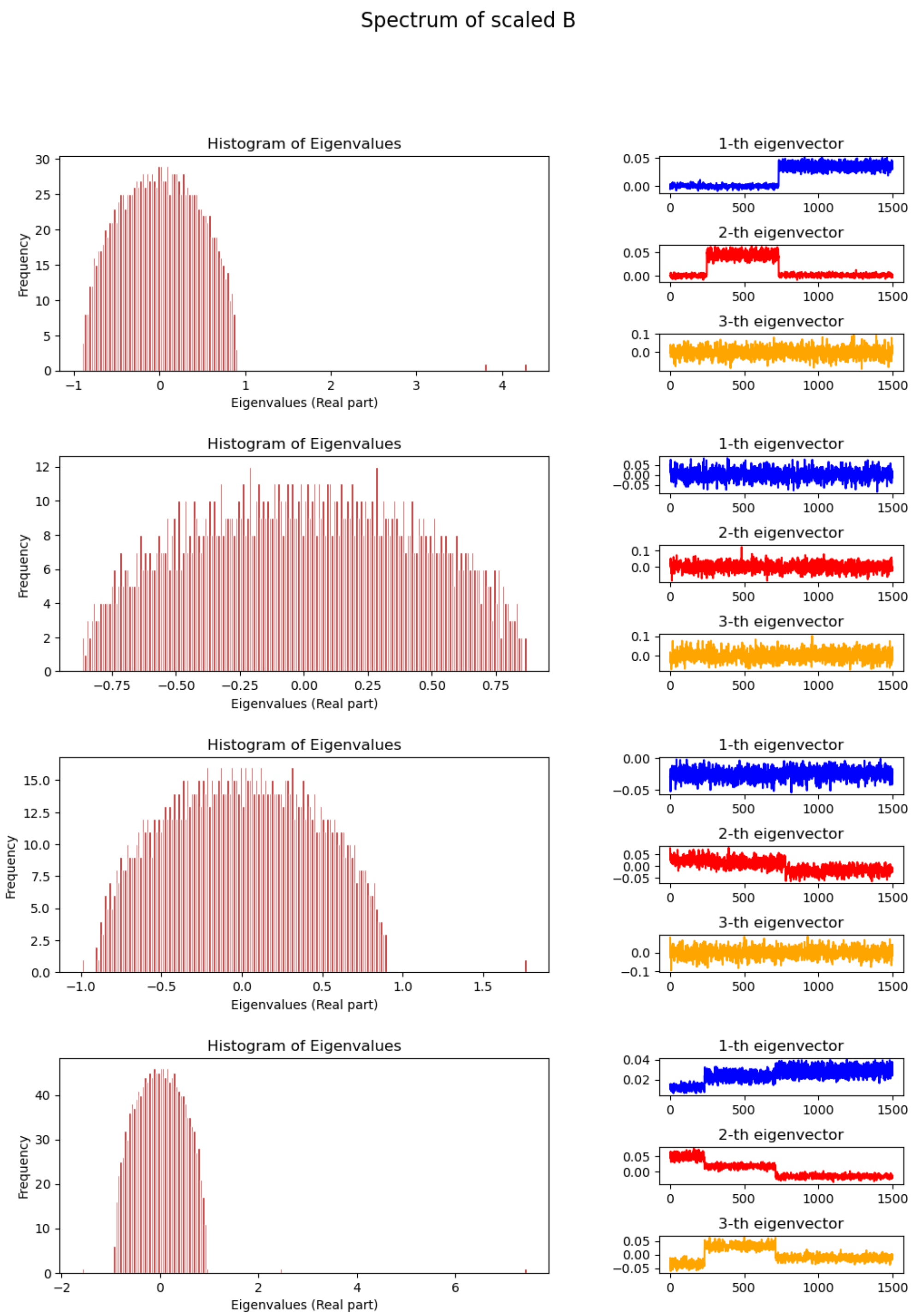

Figure 2: The spectrum distribution and the representation of the first three eigenvectors for the case q_i taking values in [0.4,0.6]

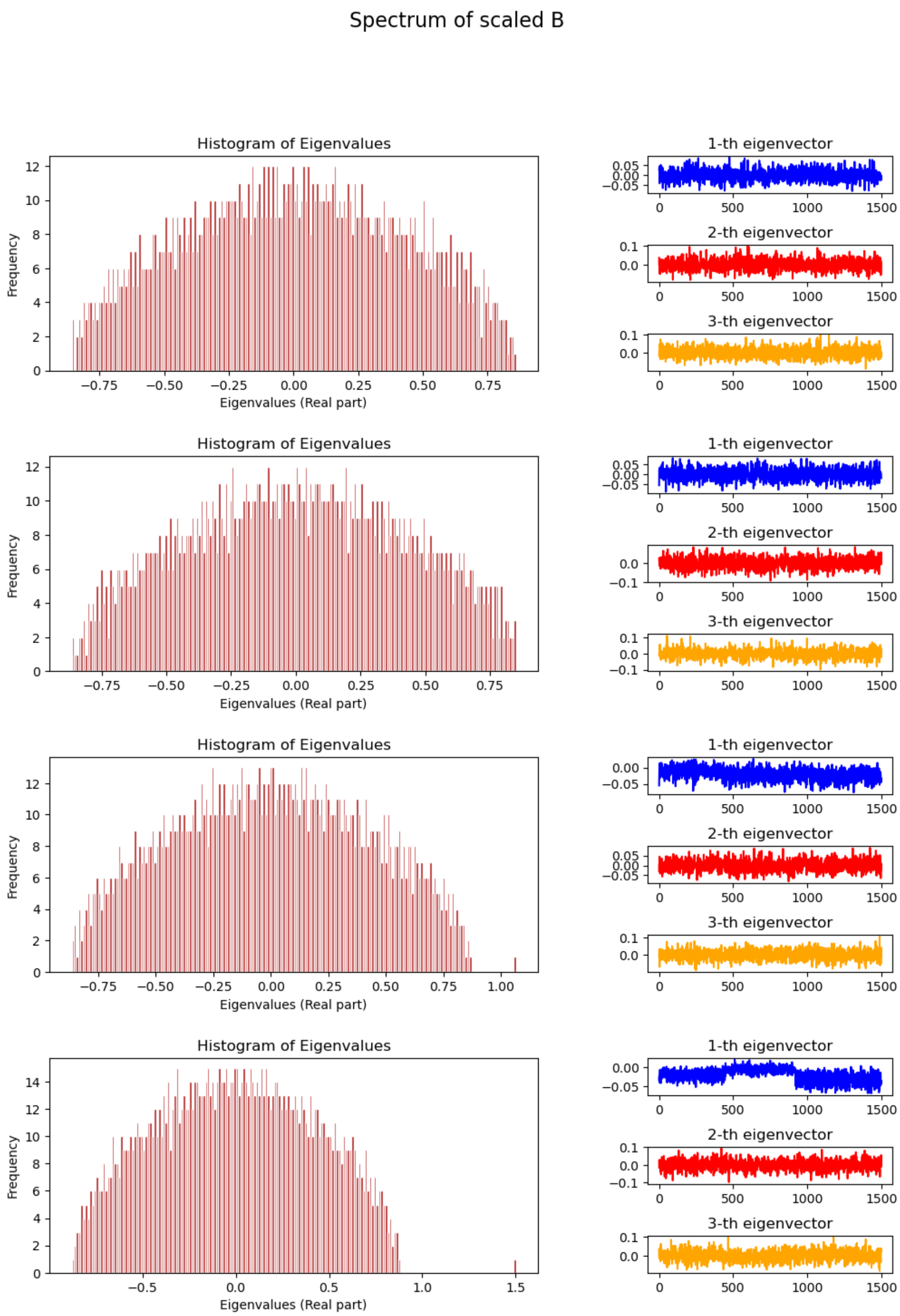

Figure 3: The spectrum distribution and the representation of the first three eigenvectors for the case q_i taking values in [0.1,0.9]

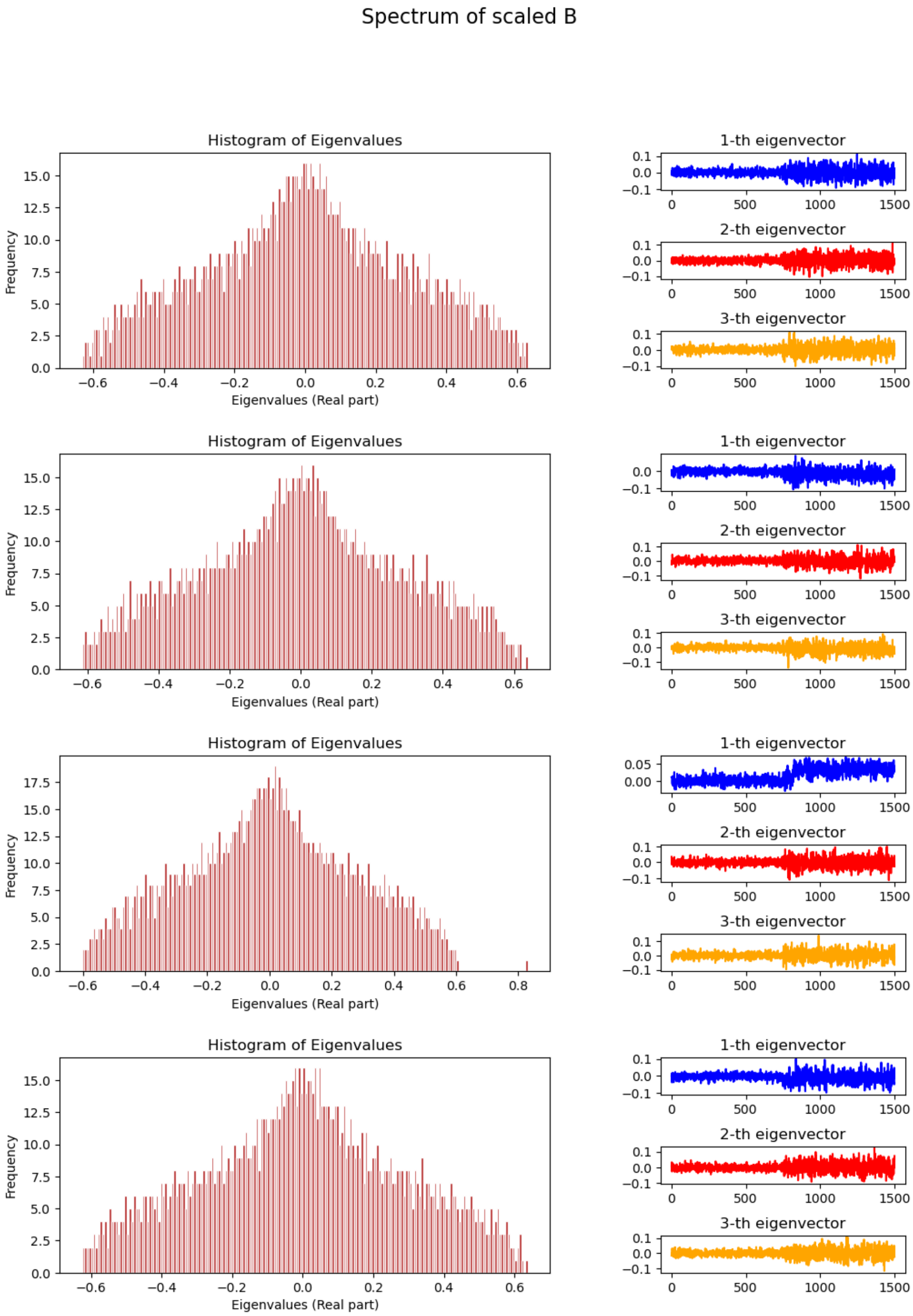

Figure 4: The spectrum distribution and the representation of the first three eigenvectors for the case q_i taking values in 0.,0.7

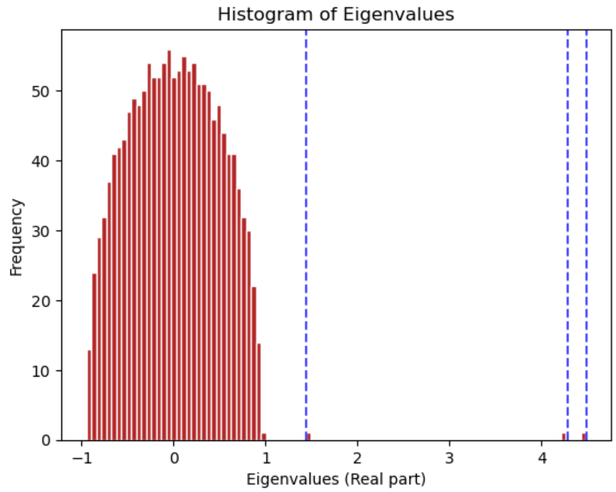

Figure 5: Numerical verification of the asymptotic position of isolated eigenvalues with $n = 3000$

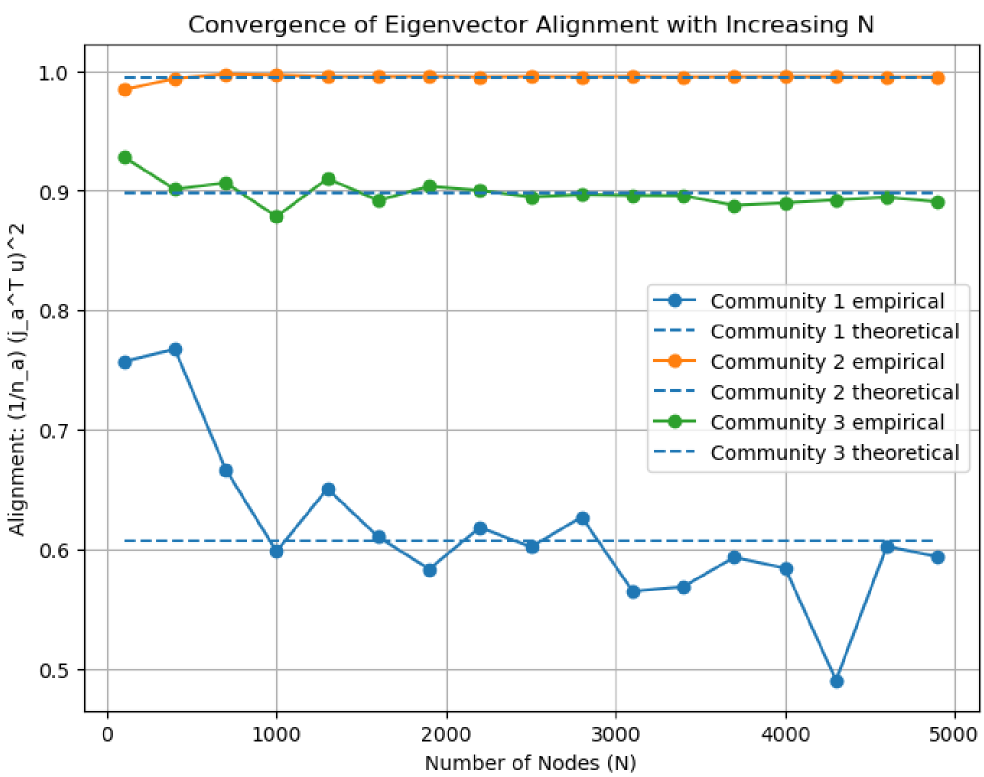

Figure 6: Numerical verification of the asymptotic values of alignements with n ranging from 100 to 3000 with iteration of 200

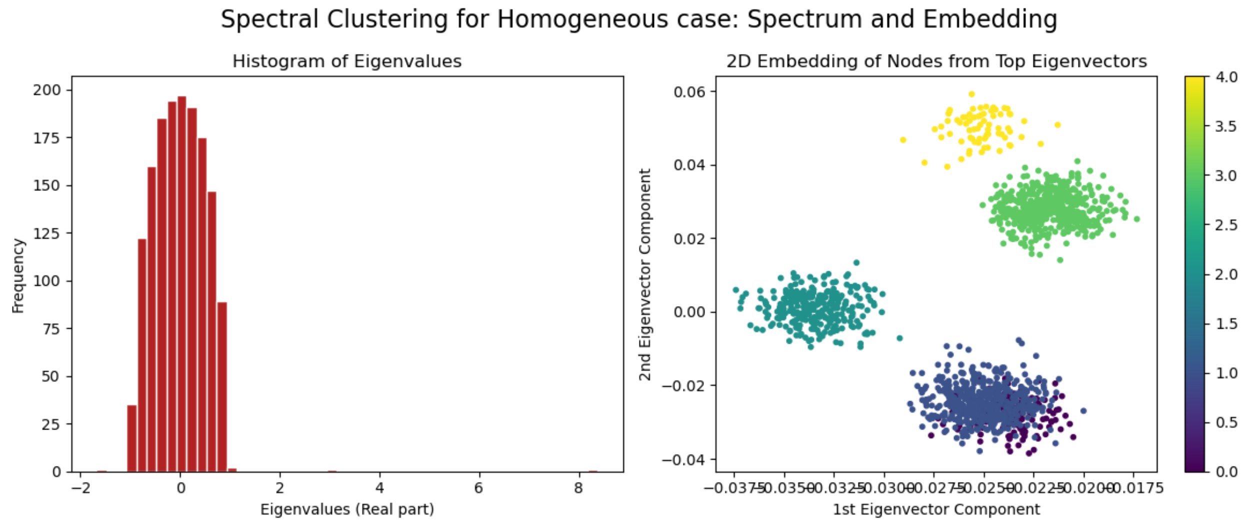

Figure 7: The result of spectral clustering on homogeneous case: 2D embedding of the nodes based on the top eigenvectors, colored by their ground truth communities. we can see the algorithm is able to well cluster the communities

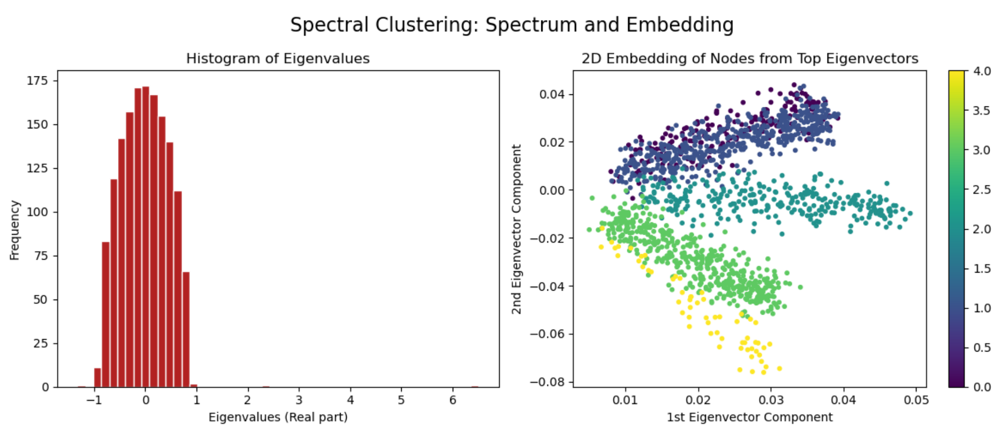

Figure 8: For the heterogeneous case, we can see that the overlap among clusters in the scatter plot indicates that the spectral embedding fails to clearly separate the communities.

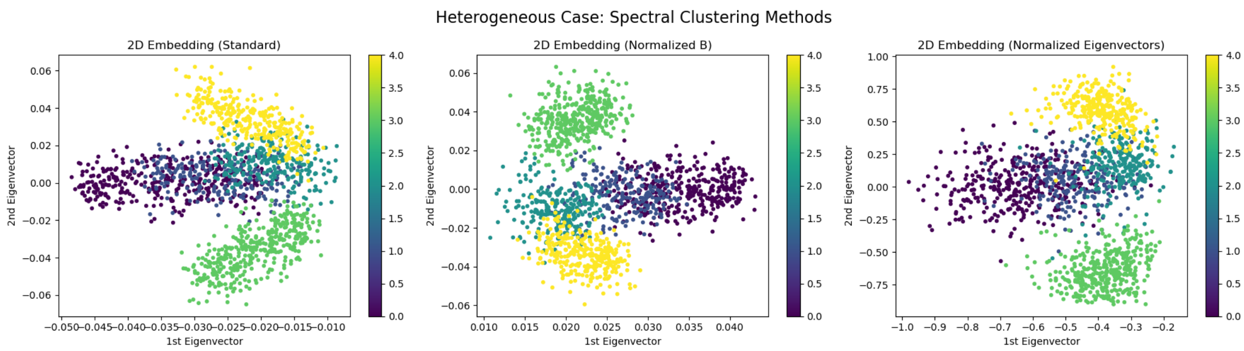

Figure 9: Left: Standard spectral embedding, which shows significant overlap between clusters. Middle: Embedding after normalizing the matrix B, improving the separation among clusters. Right: Embedding after normalizing the eigenvectors themselves, which also yields clearer cluster boundaries.