# 0.1 Resuming the article of Rohe

## 0.1.1 Notations, conventions, definitions.

*Motivation*

When studying the relationship between SBMs and spectral clustering algorithms, it is natural to ask when is the spectral clustering algorithm capable of recovering the community partitions of graphs generated under some SBM. The seminal work of [**?**] was among the first to study this question. Figure 1 is a schematic diagram linking the results presented in it. Lemma 3.1. will be of particular interest for connections with probabilistic approaches.

*General latent space models*

It is important to notice that all the results concerning the first objective in Figure 1 are valid for latent space models, which is a class of models more general than the SBM.

**Definition 1** (Latent space model)**.** For i.i.d. random vectors $z_1, \dots, z_n \in \mathbb{R}^k$ and random adjacency matrix $A \in \{0,1\}^{n \times n}$, let $\mathbb{P}(A_{ij}|z_i, z_j)$ be the probability mass function of $A_{ij}$ conditioned on $z_i, z_j$. If a probability distribution on $A$ has the conditional dependence relationships

$$\mathbb{P}(A|z_1, \dots, z_n) = \prod_{i<j} \mathbb{P}(A_{ij}|z_i, z_j),$$

and $\mathbb{P}(A_{ii} = 0) = 1$ for all $i$, then it is called an *undirected latent space model*.

*Choice of Laplacian*

They use the matrix $L = D^{-1/2}AD^{-1/2}$ as Laplacian. This is justified, since this matrix has the same eigenvectors as the more common normalized Laplacian $\tilde{L} = I - L$, and the eigenvectors are the only thing that matters in the spectral clustering algorithm. However, due care should be taken when translating their results to those obtained when using the unnormalized Laplacian. The population adjacency matrix is defined as $\mathscr{A} := \mathbb{E}[A|Z^\star]$, and the population degree matrix is the diagonal matrix with diagonal entries $\mathscr{D}_{ii} = \sum_k \mathscr{A}_{ik}$. This allows the definition of the population Laplacian $\mathscr{L} = \mathscr{D}^{-1/2}\mathscr{A}\mathscr{D}^{-1/2}$. Their work consists of two parts.

## 0.1.2 First part: convergence of eigenvectors

The first part consists in showing that the eigenvectors of the empirical Laplacian converge in some sense to the eigenvectors of the population Laplacian. This would be immediate if these matrices converged in Frobenius norm, i.e., if $\|L - \mathscr{L}\|_F \to 0$. However, they do not converge in such a norm, and so a "detour" needs to be made in order to achieve this result. They show instead that the *squared* version of these matrices converge in Frobenius norm, and then show directly that this implies that *up to a rotation* the eigenvectors of $L$ converge to those of $\mathscr{L}$.

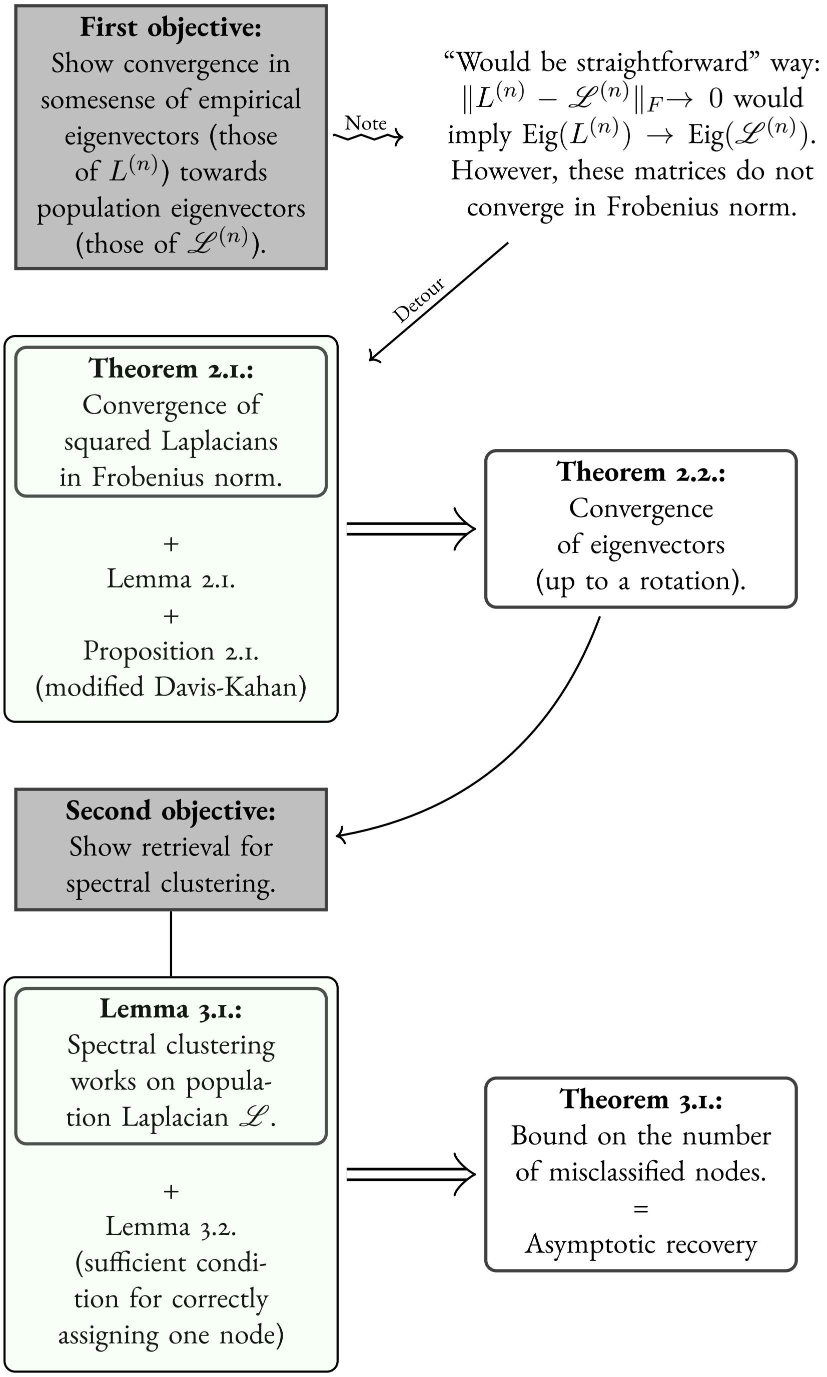
**First objective:** Show convergence in somesense of empirical eigenvectors (those of $L^{(n)}$) towards population eigenvectors (those of $\mathscr{L}^{(n)}$).
Note
"Would be straightforward" way: $\|L^{(n)} - \mathscr{L}^{(n)}\|_F \to 0$ would imply $\mathrm{Eig}(L^{(n)}) \to \mathrm{Eig}(\mathscr{L}^{(n)})$. However, these matrices do not converge in Frobenius norm.
Detour
**Theorem 2.1.:** Convergence of squared Laplacians in Frobenius norm.
+
Lemma 2.1.
+
Proposition 2.1. (modified Davis-Kahan)
**Theorem 2.2.:** Convergence of eigenvectors (up to a rotation).
**Second objective:** Show retrieval for spectral clustering.
**Lemma 3.1.:** Spectral clustering works on population Laplacian $\mathscr{L}$.
+
Lemma 3.2. (sufficient condition for correctly assigning one node)
**Theorem 3.1.:** Bound on the number of misclassified nodes.
=
Asymptotic recovery

Figure 1: Diagram of results in [**?**].

### 0.1.3 Second part: retrieval for spectral clustering

The results of this part focus on the special case of an SBM, and shows the asymptotic consistency when using the spectral clustering algorithm to estimate the community assignment clatent variables $Z^\star$. It starts with the following lemma, which shows that applying the algorithm to the expected Laplacian given the assignments recovers precisely the partitions of the SBM. Notice that this fact is non-asymptotic.

**Lemma 1.** *Consider the Stochastic Blockmodel with $k$ blocks,*

$$\mathscr{A} = Z\Gamma Z^t \in \mathbb{R}^{n\times n} \textit{ for } \Gamma \in \mathbb{R}^{k\times k} \textit{ and } Z \in \{0,1\}^{n\times k},$$

*and let $\mathscr{L}$ be the expected Laplacian given the true assignments $Z^\star$. Then, there exists a matrix $\mu \in \mathbb{R}^{k\times k}$ such that the eigenvectors of $\mathscr{L}$ corresponding to the nonzero eigenvalues are the columns of the $Z\mu$. Furthermore,*

$$z_i\mu = z_j\mu \iff z_i = z_j, \tag{1}$$

*where $z_i$ is the $i$-th row of $Z$.*

*Remark.* The equivalence in Equation 1 means that rows $i$ and $j$ of $Z\mu$ are equal if, and only if, the corresponding rows of $Z$ are equal, that is, if nodes $i$ and $j$ belong to the same community. Since there are $k$ communities, this implies that there can be at most $k$ unique rows in the matrix $Z\mu$ of eigenvectors of $\mathscr{L}$. Spectral clustering applies $k$-means to these vectors, and thus these become precisely the centroids of $k$-means (since one is applying $k$-means to at most $k$ different vectors). The rows of $Z\mu$ will then obviously be attributed to the centroid they are equal to, and by the equivalence in Equation 1, this implies that spectral clustering perfectly identifies the clusters in the expected Laplacian $\mathscr{L}$.

*Proof.* Here is a sketch of the proof, which is quite algebraic and not much motivated.

1. Factor $\mathscr{L}$ as $\mathscr{L} = Z\Gamma_L Z^t$ for some matrix $\Gamma_L \in \mathbb{R}^{k\times k}$.

2. Consider now the (different) matrix $(Z^tZ)^{1/2}\Gamma_L(Z^tZ)^{1/2}$: this is the decomposition given for $\mathscr{L}$ under the change $Z \to (Z^tZ)^{1/2}$, which can be thought of as a "square matrix version" of $Z$.

3. Show that $(Z^tZ)^{1/2}\Gamma_L(Z^tZ)^{1/2}$ is symmetric and positive-definite, implying the spectral decomposition $(Z^tZ)^{1/2}\Gamma_L(Z^tZ)^{1/2} = V\Lambda V^t$.

4. Multiply the spectral decomposition on both sides by $(Z^tZ)^{-1/2}Z^t$, revealing that

   $$Z\Gamma_L Z^t = \mathscr{L} = (Z\mu)\Lambda(Z\mu)^t \tag{2}$$

   for $\mu := (Z^tZ)^{-1/2}V$.

5. Together with the fact that $(Z\mu)^t(Z\mu) = I_k$, where $I_k$ is the $k \times k$ identity, Equation 2 is precisely the eigenvector equation for $\mathscr{L}$. This shows that the columns of $Z\mu$ are the eigenvectors of $\mathscr{L}$ associated to the non-zero eigenvalues.

6. Finally, the equivalence is a direct consequence of the fact that $\mu$ is invertible:

$$\det(\mu) = \det((Z^t Z)^{-1/2}) \det(V) > 0.$$

□