Higher-order spectral clustering for geometric graphs

K. Avrachenkov [1]   A. Bobu [1]   M. Dreveton [1]

[1] Inria Sophia-Antipolis, France

LINCS seminar, 15 September 2021

Journal of Fourier Analysis and Applications, 27:22, 2021.
arXiv:2009.11353

Introduction: graph clustering

By now, graph clustering is a very established research area.

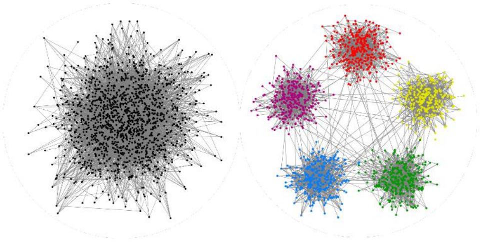
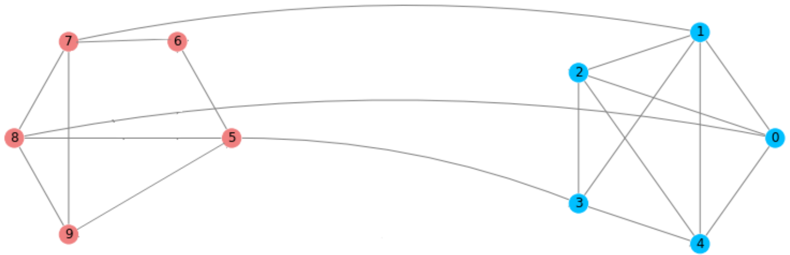


Figure: From [Abbe 2017]

Focus on graphs whose nodes have geometric attributes.
Restrict to 2 communities.

Contents

Intuition



Consider a graph $G = (V, E)$. Let $V = V_1 \sqcup V_2$. Then

$$\mathrm{Cut}(V_1, V_2) = \#(\text{edges between } V_1 \text{ and } V_2)$$

Our task then is to find

$$\arg\min_{V_1, V_2} \mathrm{Cut}(V_1, V_2)$$

given the fact that clusters V_1 and V_2 should be balanced.

Spectral clustering (SC)

Consider the vector $x = (x_i) \in \{-1, 1\}^n$ corresponding to the partition $V = V_1 \sqcup V_2$:

$$x_i = \begin{cases} 1, & \text{if } i \in V_1 \\ -1, & \text{if } i \in V_2 \end{cases}.$$

Take the adjacency matrix $A = (A_{ij})$, the diagonal matrix D, where $D_{ii} = \deg v_i = \sum_j A_{ij}$, and the graph Laplacian $L = D - A$. Then

$$\mathrm{Cut}(V_1, V_2) \;=\; \sum_{i \in V_1, j \in V_2} A_{ij} \;=\; \frac{1}{4} \sum_{i,j \in [n]} A_{ij}(x_i - x_j)^2 \;\propto\; x^T L x.$$

Continuous relaxation:

$$\underset{|V_1|=|V_2|=n/2}{\arg\min} \; \mathrm{Cut}(V_1, V_2) \;=\; \underset{\substack{x \in \{-1,1\}^n \\ x \perp 1_n}}{\arg\min} \; x^T L x \longrightarrow \underset{\substack{x \in \mathbf{R}^n \\ ||x||_2^2 = \sqrt{n} \\ x \perp 1_n}}{\arg\min} \; x^T L x$$

Eigenvectors of Laplacian matrix:

- First eigenvector of L is $v^{(1)} = (1, \ldots, 1)^T$ with $\lambda_1 = 0$;
- **Second eigenvector or Fiedler vector** $v^{(2)}$ provides the solution to the relaxed minimum cut problem;
- Cluster node i according to the sign of $v_i^{(2)}$.

Contents

Soft Geometric Block Model (SGBM)

Model parameters

number of nodes n, geometric dimension d and two measurables functions $F_{\mathrm{in}}, F_{\mathrm{out}} : \mathbb{T}^d \to [0,1]$.

Model definition

- Set of nodes $V = \{1, \ldots, n\}$;
- Each node i has random position X_i on the torus $\mathbb{T}^d$;
- Each node i gets randomly community label $\sigma_i \in \{-1, 1\}$;
- Each pair of nodes (i, j) is connected with probability

$$p_{ij} = \begin{cases} F_{\mathrm{in}}\left(X_i - X_j\right) & \text{if } \sigma_i = \sigma_j \\ F_{\mathrm{out}}\left(X_i - X_j\right) & \text{if } \sigma_i \neq \sigma_j \end{cases}$$

SGBM important particular cases

- An SGBM where $F_{\mathrm{in}}(x) = p_{\mathrm{in}}$ and $F_{\mathrm{out}}(x) = p_{\mathrm{out}}$ is an instance of Stochastic Block Model (SBM).

  Holland, P.W., Laskey, K.B., & Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks*.
- An SGBM where $F_{\mathrm{in}}(x) = 1(|x| \leq r_{\mathrm{in}})$, $F_{\mathrm{out}}(x) = 1(|x| \leq r_{\mathrm{out}})$ with $r_{\mathrm{in}} > r_{\mathrm{out}}$ is an instance of Geometric Block Model (GBM) introduced in

  Galhotra, S., Mazumdar, A., Pal, S., & Saha, B. (2018). The geometric block model. *Proceedings of AAAI*.
- Euclidean random graphes with known node locations have been studied in

  Abbe, E., Baccelli, F., & Sankararaman, A. (2021). Community detection on Euclidean random graphs. *Information and Inference*.

SGBM problem formulation

Inference problem

Estimate the latent node labeling σ given the observation of A (graph), and possibly the knowledge of $F_{\mathrm{in}}, F_{\mathrm{out}}$.

Specifically, defining the loss of an estimator $\widehat{\sigma}$ as

$$\ell(\sigma, \widehat{\sigma}) \;=\; \frac{1}{n} \min_{\pi \in \mathcal{S}_2} \sum_i 1(\sigma_i \neq \pi \circ \widehat{\sigma}_i),$$

we shall be interested in **weak consistency**

$$\forall \epsilon > 0 : \quad \lim_{n \to \infty} \mathbb{P}(\ell(\sigma, \widehat{\sigma}) > \epsilon) \;=\; 0,$$

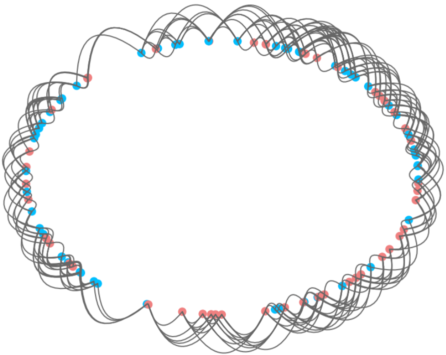
and **strong consistency**

$$\lim_{n \to \infty} \mathbb{P}(\ell(\sigma, \widehat{\sigma}) > 0) \;=\; 0.$$

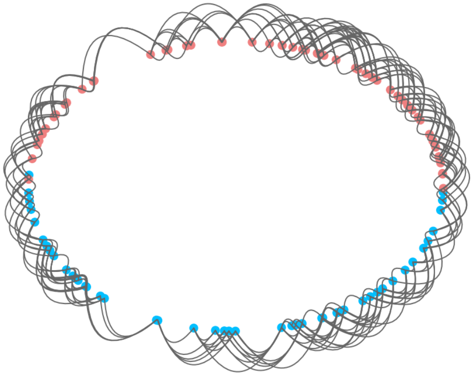
Example: GBM

Geometric Block Model

Consider $d = 1$ and
$F_{in}(x) = 1(|x| \leq r_{\mathrm{in}}), \quad F_{out}(x) = 1(|x| \leq r_{\mathrm{out}})$ with fixed $r_{\mathrm{in}} > r_{\mathrm{out}}$.

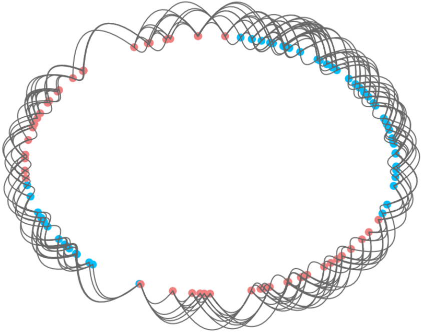


Spectral clustering on the SGBM (1)



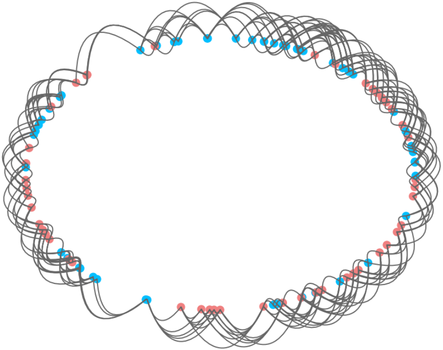
Fiedler vector produces geometric partitioning!

Spectral clustering on the SGBM (2)



The eigenvector v_4 associated with λ_4 (the fourth smallest eigenvalue) gives the partition into 4 regions.
The eigenvector v_6 divides the circle into 6 regions, and so on... Nothing useful?

Spectral clustering on the SGBM (3)



Then suddenly the eigenvector v_{10} gives 87% accuracy!

It appears that this eigenvector contains useful information about the true community structure.

How to choose the best eigenvector?

Suppose that nodes $V_1 = \{1, \ldots, n/2\}$ and $V_2 = \{n/2 + 1, \ldots, n\}$.
The ideal vector for recovery is then $v_* = (\underbrace{1, \ldots, 1}_{n/2}, \underbrace{-1, \ldots, -1}_{n/2})^T$.

Denote $\mu_{\text{in}} = \int_{\mathbf{T}^d} F_{in}(x)dx$ — average intra-cluster edge density

$\mu_{\text{out}} = \int_{\mathbf{T}^d} F_{out}(x)dx$ — average inter-cluster edge density.

v_* is an "approximate" eigenvector of $\mathbb{E}A$, associated to λ_* such that

$$\boxed{\lambda_* = \mathbb{E}\sum_{j=1}^{n/2} A_{ij} - \mathbb{E}\sum_{j=n/2+1}^{n} A_{ij} = \frac{(\mu_{in} - \mu_{out})n}{2}}$$

Higher-order spectral clustering algorithm

Higher-order spectral clustering algorithm (HOSC):

1. Calculate the eigenvalues of the adjacency matrix A;
2. Take the eigenvector $\tilde{v}$ associated with the eigenvalue $\tilde{\lambda}$ closest to $\lambda_* = (\mu_{\mathrm{in}} - \mu_{\mathrm{out}})n/2$;
3. Let $\widehat{\sigma}_i = \mathrm{sign}(\tilde{v}_i)$ for $i = 1, \ldots, n$.

Theorem (HOSC weak consistency)

In the GBM, for almost all choices of $(r_{\mathrm{in}}, r_{\mathrm{out}})$, we have with high probability $\widehat{\sigma}_i = \sigma_i$ for all but $o(n)$ nodes i.

Higher-order spectral clustering algorithm

Main steps of the proof:

1. Show that λ_* belong to the limiting spectrum;
2. Show that λ_* is isolated from other limiting eigenvalues;
3. Show that $\tilde{v} \approx v_* = (1, \ldots, 1, -1, \ldots, -1)^T$ when $\tilde{\lambda} \approx \lambda_*$.

Limiting spectrum of SGBM

For $k \in \mathbb{Z}^d$ and $F : \mathbf{T}^d \to \mathbb{R}$ we define the Fourier transform

$$\widehat{F}(k) \;=\; \int_{\mathbf{T}^d} F(x) e^{-2i\pi\langle k,x\rangle}\, dx$$

and assume that $F_{\mathrm{in}}(0), F_{\mathrm{out}}(0)$ are equal to the Fourier series of $F_{\mathrm{in}}(\cdot), F_{\mathrm{out}}(\cdot)$ evaluated at 0.

Limiting spectrum of SGBM

Theorem

Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of A, and

$$\mu_n \;=\; \sum_{i=1}^{n} \delta_{\lambda_i/n}$$

the spectral measure of the matrix $\frac{1}{n}A$. Then, for all Borel sets B with $\mu\left(\partial B\right) = 0$ and $0 \notin \bar{B}$, a.s.,

$$\lim_{n\to\infty} \mu_n(B) = \mu(B),$$

where μ is the following measure:

$$\mu \;=\; \sum_{k\in Z^d} \delta_{\frac{\widehat{F}_{\mathrm{in}}(k)+\widehat{F}_{\mathrm{out}}(k)}{2}} \;+\; \delta_{\frac{\widehat{F}_{\mathrm{in}}(k)-\widehat{F}_{\mathrm{out}}(k)}{2}}.$$

Limiting spectrum of SGBM

Good news: $\lambda_* = \frac{\mu_{\rm in} - \mu_{\rm out}}{2} n$ belongs to the limiting spectrum. (Recall $\mu_{\rm in} = \widehat{F}_{\rm in}(0)$ and $\mu_{\rm out} = \widehat{F}_{\rm out}(0)$.)

The proof is based on the moment method and is a generalization of

Bordenave, C. (2008). Eigenvalues of Euclidean random matrices. *Random Structures & Algorithms*, 33(4), 515-532.

to the block model, where we calculate

$$\mathbb{E}\mu_n(t^m) = \frac{1}{n^m}\sum_{i=1}^{n}\mathbb{E}\lambda_i^m = \frac{1}{n^m}\mathbb{E}\mathrm{Tr}A^m = \frac{1}{n^m}\mathbb{E}\sum_{\alpha\in[n]^m}\prod_{j=1}^{m}A_{i_j,i_{j+1}}$$

and then use Talagrand's concentration inequality and Borel-Cantelli lemma .

Separation of λ_*

Main steps of the proof:

1. Show that λ_* belong to the limiting spectrum;
2. Show that λ_* is isolated from other limiting eigenvalues;
3. Show that $\tilde{v} \approx v_* = (1, \ldots, 1, -1, \ldots, -1)^T$ when $\tilde{\lambda} \approx \lambda_*$.

Separation of λ_*

Proposition

Consider the adjacency matrix A of an SGBM and assume that:

$$\widehat{F}_{\mathrm{in}}(k) + \widehat{F}_{\mathrm{out}}(k) \neq \widehat{F}_{\mathrm{in}}(0) - \widehat{F}_{\mathrm{out}}(0), \qquad \forall k \in \mathbb{Z}^d,$$
$$\widehat{F}_{\mathrm{in}}(k) - \widehat{F}_{\mathrm{out}}(k) \neq \widehat{F}_{\mathrm{in}}(0) - \widehat{F}_{\mathrm{out}}(0), \qquad \forall k \in \mathbb{Z}^d \backslash \{0\}.$$

with $\widehat{F}_{\mathrm{in}}(0) \neq \widehat{F}_{\mathrm{out}}(0)$. *Then, the eigenvalue of A the closest to* $\frac{\widehat{F}_{\mathrm{in}}(0) - \widehat{F}_{\mathrm{out}}(0)}{2} n$ *is of multiplicity one. Moreover, there exists* $\epsilon > 0$ *such that for large enough n every other eigenvalue is at a distance at least* ϵn.

Remark In case of the GBM, we showed that the above conditions hold true for all but a zero Lebesgue measure set of parameters $r_{\mathrm{in}}, r_{\mathrm{out}}$.

Closeness of $\tilde{v}$ to v_*

Main steps of the proof:

1. Show that λ_* belong to the limiting spectrum;
2. Show that λ_* is isolated from other limiting eigenvalues;
3. Show that $\tilde{v} \approx v_* = (1, \ldots, 1, -1, \ldots, -1)^T$ when $\tilde{\lambda} \approx \lambda_*$.

Closeness of $\tilde{v}$ to v_*

The following result was very useful to demonstrate the closeness of $\tilde{v}$ to v_*.

Theorem (Kahan-Parlett-Jiang)

Let A be a real symmetric matrix. If $\widetilde{\lambda}$ is the eigenvalue of A closest to $\rho(v) = \frac{v^T A v}{v^T v}$, δ is the separation of ρ from the next closest eigenvalue of A and $\widetilde{v}$ is the eigenvector corresponding to $\widetilde{\lambda}$, then

$$|\sin \angle(v, \widetilde{v})| \leq \frac{\|Av - \rho v\|_2}{\|v\|_2 \delta}.$$

In our case, this leads to

$$\|v_* - \widetilde{v}\|_2 \leq \sqrt{2}\,|\sin \angle(v_*, \widetilde{v})| \leq \frac{C}{\sqrt{n/\log(n)}} \quad w.h.p.$$

Contents

Numerical experiments (1)

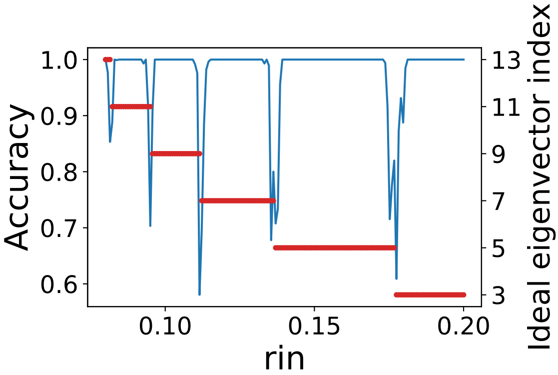


Figure: Evolution of accuracy (blue curve) with respect to $r_{\rm in}$, for a GBM with $n = 3000$ and $r_{\rm out} = 0.06$. The red curve show the index of the ideal eigenvector.

Numerical experiments (2)

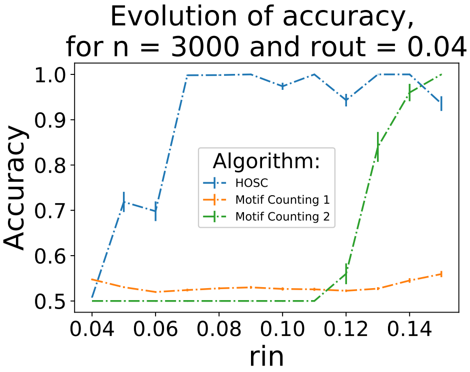


Figure: Accuracy obtained on 1-dimensional GBM for different clustering methods. Results are averaged over 50 realizations, and error bars show the standard error. Comparison with the methods of (Galhotra *et al*, 2018,2019).

Real data sets

- Wikivitals: links between wikipedia articles; cluster sizes (1715,1752)
- DBLP: co-authorship network between scientists; cluster sizes (6562,6764)

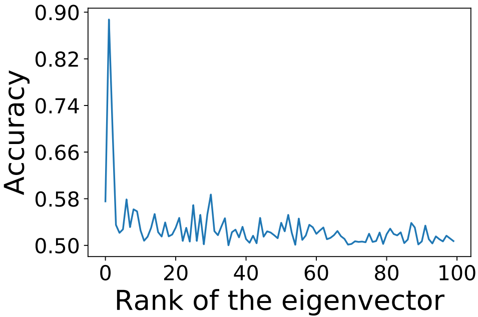


Figure: Accuracy per eigenvector rank: Wikivitals

Real data sets

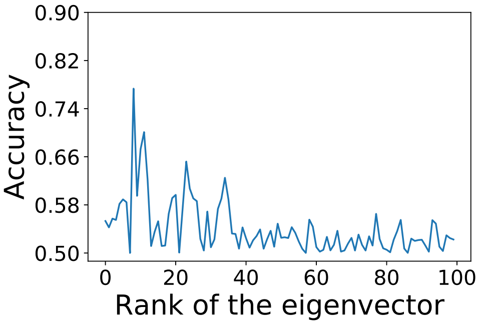


Figure: Accuracy per eigenvector rank: DBLP

Contents

Conclusion

Takeaway message:

If you use spectral clustering methods, check higher-order eigenvectors, they can be more effective!

Especially if you deal with geometric attributes.

Directions of further research

Future work:

- **More clusters**
  How to choose the eigenvector(s) if we have $K > 2$ clusters?
- **Sparse regime**
  The current proof does not work if the average degree is $o(n)$.
- **Weighted graphs**
  The results can easily be transferred to models with weighted edges instead of probability of edge appearance.
- **Model parameters**
  Is it possible to determine μ_{in} and μ_{out} from the observed graph?

Thank you for your attention!

Any questions?